# Exploration and Contextualization through Interaction and Concepts

**Marc Michiel Bron**

# Exploration and Contextualization through Interaction and Concepts

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in
de Agnietenkapel
op donderdag 28 november 2013, te 14:00 uur

door

Marc Michiel Bron

geboren te Utrecht

**Promotiecommissie**

Promotor:
Prof. dr. M. de Rijke

Copromotoren:
Dr. K. Balog
Dr. F. Nack

Overige leden:
Prof. dr. L. W. M. Bod
Prof. dr. L. Hardman
Dr. D. Kelly
Prof. dr. ir. A. P. de Vries

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research was supported by Netherlands Organisation for Scientific Research (NWO) under project number 640.004.802.

SIKS Dissertation Series No. 2013-43
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Cover by Jiyin He and Karen Weening, based on ’Books’ by brokenview:
http://www.flickr.com/photos/44824500@N05/4447528376 licensed under CC BY-NC-SA 3.0.
Printed by Off Page, Amsterdam

ISBN: 978-94-6182-374-8

# Acknowledgements

I started the work described in this thesis four and a half years ago because I liked figuring out how "stuff" works. I soon learned that that is not necessarily the same as doing research. Since then I have learned a great many things from many people which has allowed me to write this thesis. Here I would like to thank the people who have played a major role in this process.

First of all I would like to thank my supervisor Maarten de Rijke. He taught me how to write papers, how to view a problem from different angles, and how to clearly explain the "stuff" I was working on to others. I would also like to thank my two co-promoters, i.e., Krisztian Balog and Frank Nack, who have each guided me through a different stage of my PhD. Working with Krisztian felt like trying to jump on a running train, but he helped me on board and showed me how to write my first paper. About halfway through my PhD I decided to head in a different direction and this is where Frank helped me to get my bearings and publish in a new field.

I am happy that Arjen de Vries, Diane Kelly, Lynda Hardman, and Rens Bod agreed to serve as members of my PhD committee as their diverse areas of expertise reflect the interdisciplinary nature of the thesis.

I am grateful to all my co-authors and I would like to especially thank Jasmijn. I learned as much from our discussions as our misunderstandings. Some of the experiments in this thesis would not have been possible without the help of Beeld en Geluid. Lotte and Wietske were a great help and provided us with anything we needed. A big thank you also goes to Andrei, Bart, and Justin who programmed the experimental interfaces used in our experiments.

It is difficult to describe what it is like to work at ILPS but from the moment I started there I felt like I became part of a family. Edgar, Manos, Simon, Wouter, Bouke, Katja. Thank you for all the great times we shared, this experience would not have been the same without you.

I spent three months in San Francisco and I thank Bruce, Merrillee, Ellen and all the others members of the OCLC San Mateo office for the food, cat sitting, art festivals and other outings that made my internship so much fun. Also a big thanks to Loredana for helping me out with a place to stay while I was there.

Of course there was also life outside of work where my friends provided the necessary distractions. Playing football was one of my favorite ways of clearing my head and I want to thank my teammates of JOS wgm 4 for providing an interesting contrast to the research environment.

Some people have known me from long before I started my PhD. Isidoor, Don, Aziz, Niels, Peterpaul, whenever we met we had fun like in the old days and nothing seemed to have changed.

Last but not least, I would like to thank my parents, Gwen and Thijs, for always supporting me, my brothers Maarten and Jochem for being brothers and agreeing to be my paranymphs, and Jiyin, my love, who was there at my side, always.

# Contents

# 1 Introduction

The philosophy behind the empirical research paradigm is to gain knowledge through observations: a researcher arrives at a theory through previous observations or experiences; based on this theory predictive statements or hypotheses are formulated about the phenomenon under consideration; and finally these hypotheses are verified through rigorous experimentation and analysis of results. Today's technology enables the production, recording and storage of an unprecedented number of observations [140]. On the one hand this provides an opportunity for researchers to investigate phenomena that were difficult or impossible to study before, e.g., the discovery of archeological sites through the analysis of satellite imagery [192]. On the other hand the number of observations goes beyond the limits of what researchers are able to process manually.

The abundance of available observations has not only had an impact on researchers but also on research itself. Fields of study, however, differ in the way they have adapted their research methods and tools to utilize the availability of these observations. For some fields this transition has come more naturally than for others. Consider, for example, the natural sciences versus the humanities. Sub-disciplines of the natural sciences such as earth sciences, computational biology, and astronomy, have readily adopted computationally intensive methods, as they study immense collections of observations consisting of signals recorded by radars, sensors, or produced by simulations [168]. This type of observation lends itself well for analyses through data mining and visualization techniques to discover patterns and develop insights into the phenomenon under study.

In contrast, the traditional objects of study in the humanities have always been analogue records such as books, letters, and photographs [151]. These objects are studied using analytical, critical, and interpretative approaches instead of computational methods. However, driven by continuous digitization, electronic production, and electronic storage of records the humanities is turning into a data intensive discipline as well, as witnessed by the birth of the digital humanities [147, 300]. As the introduction of new technology and information sources is changing the way humanities researchers work and the questions they seek to answer [60, 99, 243, 324], a new challenge arises for the development of tools and algorithms that support new practices as well as traditional ones using new types of information. This observation has been made before, for example, by Menzel [241] when he called to attention the information challenge posed by the vast increase in scientific publications in the 1940s and 1950s. Later, others noted the influence of the photocopier on the research behavior of historians [89, 290] or investigated the use of library databases by humanities scholars [44, 351]. Calls for the continued study of the research behavior and development of tools also originate from within the humanities. In 2003 Unsworth [336] observed that ten years of tool development for the humanities had yet to result in portable, reliable, and extensible tools. He advocated building a toolkit that supports the *scholarly primitives* of humanities researchers. A 2009 call to action invited more behavioral research in the humanities as the discipline evolved and tools were not keeping up [60]. In this thesis we take a fresh look at some of the challenges faced by humanities researchers in the big data era. We follow a user centered approach and develop information retrieval tools and algorithms to support humanities researchers to cope with the increasing number of digital records available in archives, digital libraries, and on the web.

Particular challenges for humanities researchers raised by the abundance of available material are to gain insight in which materials to consider for a study and once chosen to obtain a holistic view of the research topic. To address these challenges we focus on the following two themes: *exploration* and *contextualization*. In the first part of the thesis we focus on developing and evaluating novel information retrieval tools that allow for richer interactions to support exploration of collections of digital records. In the humanities the approach to research is interpretative in nature. In order to shape their questions researchers embed themselves in material and allow themselves to be guided through their knowledge, intuitions, and interests [89]. In this process exploration is key to arrive at a broad overview of a research topic. In the second part of the thesis we focus on developing and evaluating algorithms that enable contextualization. The term contextualization as used in this thesis refers to the discovery of additional information that a researcher needs to interpret the material under study [6, 89, 123]. For example, when studying changes in society over time by examining news broadcasts, the dominance of reports about crime would suggest that society is unsafe and degenerate. Understanding the news production environment, however, provides an explanation in that crime is covered constantly because of its entertainment value [6].

## 1.1 Research Outline and Questions

The central questions in the first part of the thesis, which is devoted to exploration, are: how do researchers in the humanities use existing information retrieval tools to search for research material in digital collections and can information retrieval tools with richer means of interaction be designed to provide better support for exploration? The search behavior of researchers and of other users of digital libraries has long been the subject of study, see Chapter 2 for an overview. The rapid changes in technology and available digital materials, however, are changing the research practices of humanities researchers [60, 99, 243]. We investigate these changes by zooming in on a particular discipline that considers itself part of both the humanities and the social sciences: *media studies*. Media studies is a field that is diverse in the objects studied, technologies used, and methodologies applied. This makes media studies researchers an interesting group of subjects to study. Moreover, findings for this group are potentially relevant for other humanities disciplines as well. To gain a better understanding of the research practices of this group of researchers we first ask:

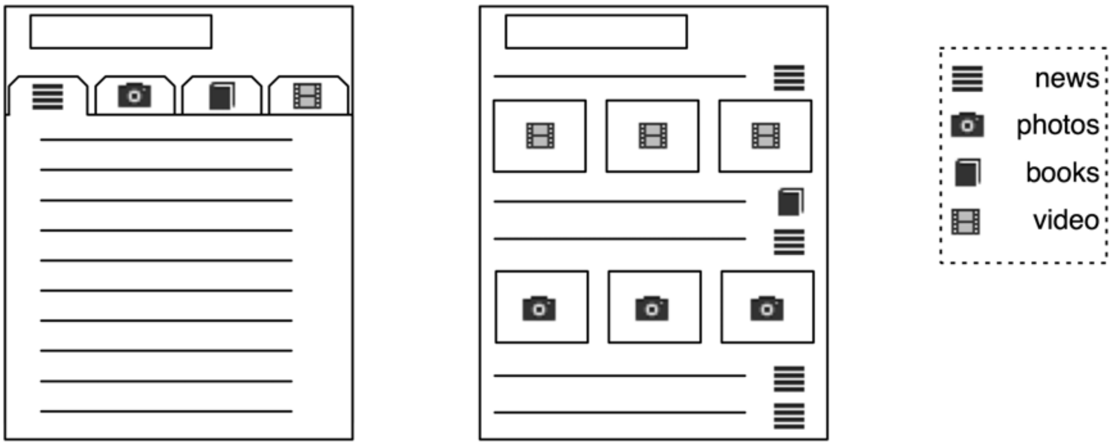


**Figure 1.1:** Two types of aggregated search display: tabbed (left) and blended (right).

**RQ 1.** What does the research process of media studies researchers look like?

**a.** Can we identify sequences of activities in research projects of media studies researchers and how does the resulting model compare to other models of the humanities research cycle?

**b.** Do research questions of media studies researchers change during research projects, and can we identify factors that influence this change?

**c.** Which information needs and information gathering challenges do media studies researchers face during research projects?

One observation from our analysis of the research process is that media studies researchers require material originating from various sources, e.g., multiple archives, and of varying modality, e.g., photos, transcripts, and videos. As these sources are rarely accessible through a single search engine, seeking for information across these sources requires researchers to sequentially go through each of them individually. Aggregated search interfaces are a solution to this problem; they provide users with an overview of the results from various sources (verticals) by collecting and presenting information from multiple collections. Two general types of aggregated search interfaces exist: *tabbed* and *blended* [240]. Tabbed interfaces provide access to each source in separate tabs, while blended interfaces combine multiple sources into a single result page, see Figure 1.1. We investigate whether users engaging in multiple search sessions during a research project benefit from alternative result presentation methods, and ask:

**RQ 2.** How do master students conducting a media studies research project use alternative aggregated search result presentation methods?

**a.** Do media studies students switch between tabbed and blended display types during a multi-session search task and what is the motivation to switch between display types?

**b.** Do changes in media studies students' information need across sub-tasks influence preference for a particular display type?

**c.** What other factors are related to changes in display preference during a multi-session search task?

Another observation from our analysis of the research process of media studies researchers is the importance of exploration and the effect this has on the research questions they ask. Exploratory search tools support various ways of exploring collections, e.g., through filtering by facets, visualizations, or relevance feedback; see [148, 164, 305, 361] for overviews of systems and their capabilities. Work on exploratory search systems, however, focuses on supporting exploration in a general setting. Few studies have considered how exploratory search systems can support researchers in discovering alternative views and trends in the data during exploration. We propose to extend the traditional exploratory search system design in two ways: we incorporate two side-by-side versions of an exploratory search interface in a single display to create a *subjunctive interface* in which multiple queries can be explored simultaneously [223]. Second, we add visualizations that allow for contrasting and comparing the characteristics of the result sets. Given this subjunctive exploratory search interface, we ask:

**RQ 3.** Does the ability to make comparisons support media studies researchers in exploring a collection of television broadcast metadata descriptions?

**a.** Does a subjunctive exploratory search interface better support media studies researchers in a complex exploratory search task than a standard exploratory search interface?

**b.** Does the subjunctive exploratory search interface better support media studies researchers in refining a research question than a standard exploratory search interface?

**c.** Does the increase in complexity caused by the inclusion of additional features affect the usability of the subjunctive interface as compared to a standard exploratory search interface?

Once a humanities researcher has explored the available material and settled on a particular topic, the next challenge is to obtain a holistic view of the concepts involved in the research topic and the relationships among them. The second part of the thesis is centered around methods to enable contextualization. We broadly define contextualization as the discovery of additional information that a researcher needs to interpret the topic under study. The need for contextualization is not restricted to people involved in research. To support contextualization on the web, hyperlinks are added to web pages, which enable navigation to related information. Another type of application that provides contextual information are news recommendation systems, which suggest news articles relevant to the interests of a particular user [167, 233, 276]. A more recent development is the incorporation of additional markup such as micro-formats, schema's, or RDF to parts of web pages in order to create a machine understandable Web of Data [53, 55]. In principle, this allows tools to automatically discover and retrieve background information relevant to a user's information need. Google's *knowledge graph biography*[1] is an example of an application that utilizes structured information to provide background information about concepts.

[1] http://googleblog.blogspot.nl/2012/05/introducing-knowledge-graph-things-not.html

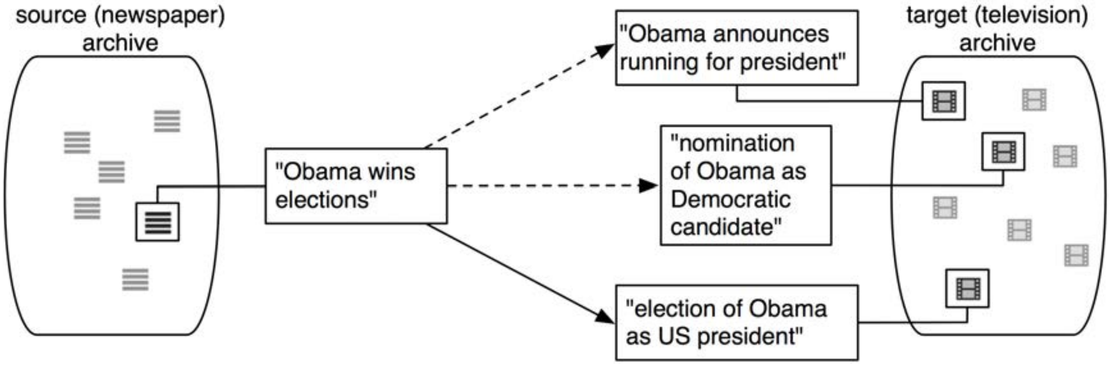

source (newspaper) archive
"Obama wins elections"
"Obama announces running for president"
"nomination of Obama as Democratic candidate"
"election of Obama as US president"
target (television) archive

**Figure 1.2:** Schematic view of the linking archives task. A representation of a record from a source archive is connected to representations of records in the target archive that cover the same (solid arrow) or related (dashed arrows) events.

These applications do not provide users with a ranked list of search results, such as common in a web search setting. Instead they aim to satisfy a user's information need directly by aggregating or extracting specific pieces of information, e.g., a list of recommended news articles or a short biography with the birth date, names of co-authors, and works of a particular author. Several sub-fields have emerged within the field of information retrieval that seek to address these types of task that go beyond traditional document retrieval, see Chapter 6 for an overview. To be able to build upon, and compare to, the methods emerging from these sub-fields we move away from the user centered approach of the first part of the thesis. Instead, we adopt a system centered approach and in the second part of the thesis we evaluate methods using tasks and test collections that are abstractions of problems faced by humanities researchers. A benefit of this abstraction is that it reduces the experimental overhead inherent to user based testing and allows us to explore a wider range of contextualization methods. A disadvantage is that the methods developed may not be able to be incorporated directly in real world applications.

One way to enable contextualization is to automatically create links between archives with records that describe a particular event. For example, to reconstruct the 2008 election of Barack Obama, a media studies researcher may supplement the chain of events described by records in a newspaper archive with those from a television archive. In order to provide this type of access, archives and libraries organize material according to events, themes, and entities, by adding subject headings and category descriptors as metadata to records. However, metadata descriptions vary across archives in the formats and level of description depending on the institution, region, and type of the record. We define the task of *linking archives* as follows: given a representation of a record and its associated metadata from one archive, connect it to the representation of a record and its metadata in another archive, where that record describes the same event or a related event, see Figure 1.2. Given this task, we investigate a method for connecting a record from a newspaper archive to records in a television archive that discuss the same or related events and ask:

**RQ 4.** How can we automatically generate links from a record in a newspaper archive with a rich textual representation to records in a television archive that tend to have sparse textual representations?

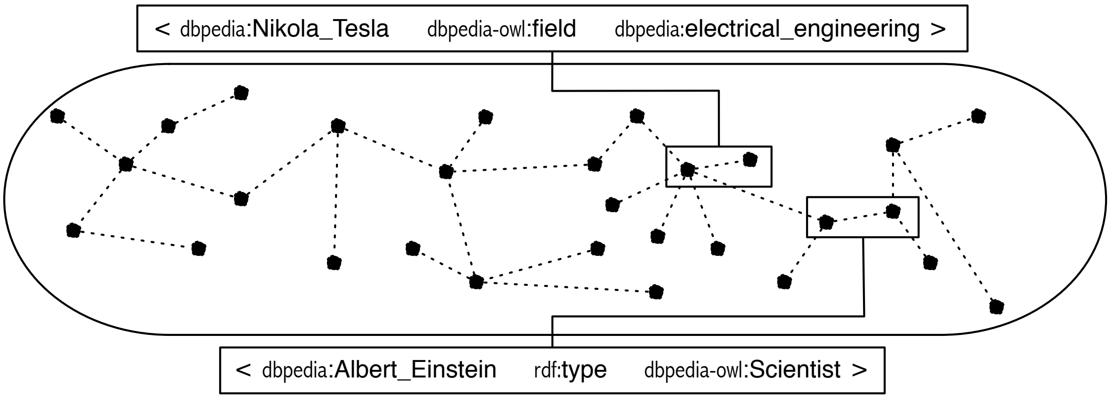

< dbpedia:Nikola_Tesla dbpedia-owl:field dbpedia:electrical_engineering >
< dbpedia:Albert_Einstein rdf:type dbpedia-owl:Scientist >

**Figure 1.3:** Schematic view of the data used in the example-based entity finding task. Nodes indicate entities or objects, dashed lines represent predicates.

**a.** Does expanding sparse record representations with text from other sources improve linking performance?

**b.** What effect does modeling reduced versions, e.g., by selecting informative terms of the original richly represented records from the source archive, have on linking performance?

In the linking archives task we investigated contextualization of records exclusively devoted to a particular type of concept, i.e., events, but archival records may also be organized around other concepts, e.g., entities or themes. The particular way in which archival material is organized, however, does not necessarily align with the research topic of a humanities researcher. In order to gather relevant material and to be able to complete the information necessary for interpretation, a humanities researcher first needs to identify the concepts relevant to his/her research topic [98]. For example, a historian investigating the relation between science and religion may discover archival records dealing with Albert Einstein's opinions about religion and be interested to complement this with information about other scientists with religious views, e.g., Nikola Tesla.

This type of access asks for a contextualization method that provides the user with a set of related concepts. To investigate such a method we define the example-based entity finding task as follows: find a group of target entities that all have the same relationship with a particular concept in common, e.g., scientists with religious views, given a number of examples, e.g., Albert Einstein and Nikola Tesla.

The Web of Data is an interconnected network of objects that contains information from a multitude of knowledge bases and information repositories [53]. This type of structured data provides information about entities and the relationships among them [55] and has the potential to be helpful in entity-oriented search tasks. In this setting, entities are represented by a unique identifier (URI) and relations are encoded as (RDF) triples consisting of: (i) a *subject*, which is an entity; (ii) a *predicate*, which is a relation; and (iii) an *object*, which is either another entity or a property of an entity encoded as text, see Figure 1.3. We investigate a specific instantiation of the example-based entity finding task where entities are represented by URIs and relations to other entities by triples in the Web of Data and ask:

**RQ 5.** How can we exploit the structural information available in the Web of Data to find a set of entities, that all have the same relationship with a particular concept in common, based on a number of example entities?

**a.** Is a structure-based method that uses examples competitive when compared against a text-based approach?

**b.** Does the performance of text- and structure-based methods depend on the quality and the number of examples that are given?

**c.** Can a hybrid method automatically balance between the two approaches in a query-dependent manner?

So far, our methods to discover related concepts have relied on structured data, however, structured data is not always available. For example, a political studies researcher investigating the downfall of Bill Clinton at the end of his presidential years may start by identifying Clinton's supporters and opponents. Relations such as supporters and opponents are generally not encoded in knowledge bases except for the most prominent of figures.

We now let go of the structured data requirement and explore the utility of unstructured data to find contextual information for an entity. One of the problems introduced by moving away from structured data is that entities are no longer grounded, i.e., uniquely identifiable, by their relations to other entities or objects. Instead, entities are discovered as surface forms in text and an additional normalization step is required to uniquely identify these entities. Another challenge is that relations are no longer specified by predicates but occur as parts of passages surrounding the surface forms of entities. As a first step we turn to Wikipedia as a corpus that provides a middle ground between structured data and noisy content on the Web. Each page in Wikipedia describes and uniquely identifies an entity. Additionally, categories are associated with Wikipedia pages that provide coarse relations between entities. We utilize these properties and aim to discover not just the surface forms of entities, but also their corresponding Wikipedia pages. The categories are used as an initial filter to reduce the number of entities considered. With these requirements in mind we investigate a method to address the related entity finding (REF) task, i.e., given a (source) entity, e.g., Bill Clinton, a free text description of a relation, e.g., Clinton's political opponents during his presidential years, and the type of the (target) entities, e.g., persons, finds related entities or which the specified relationship with the source entity holds. We ask:

**RQ 6A.** How can we find related entities for which the specified relation with a given source entity holds and that satisfy the target type constraint?

**a.** How do different measures for computing co-occurrence affect the recall of a pure co-occurrence based related entity finding model?

**b.** Can a type filtering approach based on Wikipedia categories successfully be applied to related entity finding to improve precision without hurting recall?

**c.** Can recall and precision be enhanced by combining the co-occurrence model with a context model, so as to ensure that source and target entities engage in the right relation?

We then go beyond the relatively clean setting provided by Wikipedia and investigate the behavior of the method developed above on a large web corpus. This adds another challenge in that pages no longer uniquely identify entities. Instead, we set out to find homepages for the surface forms of related entities and ask:

**RQ 6B.** How can we find related entities and their homepages in a web corpus?

**a.** Does the use of a larger corpus improve estimations of co-occurrence and context models?

**b.** Is the initial focus on Wikipedia a sensible approach; can it achieve performance comparable to other approaches?

**c.** Can our basic framework effectively incorporate additional heuristics in order to be competitive with other state-of-the-art approaches?

The research questions as outlined above deal with ideas and technology in various stages of maturity. In particular, in the first part of our investigations we focus on supporting research practices by extending existing information retrieval tools with richer means of interaction. The user centered investigation of these ideas in the context of the research processes in media studies is possible as the information retrieval tools and interaction methods are familiar to the research community considered. In the second part of the thesis novel techniques are investigated to support research practices. These techniques, however, are developed and evaluated on abstractions of existing tasks, and applications incorporating these methods still lack the technological maturity for adoption within a research practice. Therefore, incorporation into tools of the techniques from the latter parts of the thesis and user centered evaluation thereof is left as future work.

## 1.2 Main Contributions

The main contributions of the thesis are as follows.

- **Analysis of the research process of media studies researchers.**

  We provide a detailed analysis of the research cycle of media studies researchers during a research project and find that it is consistent with existing models of information behavior. It extends previous work in that it is more detailed in identifying the influence of information gathering and analysis activities on the research questions developed during a project.

- **Analysis of the factors that affect the research questions of media studies researchers.**

  We identify factors that lead to changes in the research questions of media studies researchers during the research process and illustrate the importance of exploration and contextualization. We add to existing work by providing a detailed picture of the various data sources and technologies used during a research project and how availability and methods of access play a role in changing the questions researchers ask.

- **Design of two complementary studies to investigate the search behavior of media studies students with multiple types of aggregated search display during a multi-session research task.**

  We describe two exploratory studies on the use of aggregated search result displays during a multi-session research task: a longitudinal study and a laboratory study. The longitudinal study is aimed at observing the use of various display types in a naturalistic setting, while the laboratory study provides insight in the factors that influence display use in the first study. This is the first work on implementing this type of design to evaluate aggregated search interfaces in the setting of multi-session search tasks.

- **Analysis of media studies students' preferences for aggregated search display type during a multi-session research task.**

  We identify the motivations behind switching behavior and changes in display preference of students using multiple aggregated search displays during a multi-session research task. We add to the existing body of work by providing motivations for switching behavior and changes in display preference not previously observed in studies evaluating aggregated search interfaces or studies evaluating interfaces for complex search tasks.

- **Design of a subjunctive exploratory search interface.**

  We introduce an interface that supports the generation and comparison of multiple views on search topics based on a side-by-side display. Each side features an exploratory search component to retrieve and refine a result set for a particular query. An additional visualization component allows for the comparison of characteristics of the result sets obtained with each side.

- **Evaluation of the support a subjunctive exploratory search interface provides for exploration.**

  We conduct a user study to compare the effectiveness in supporting exploration provided by a traditional exploratory search interface with that provided by a subjunctive exploratory search interface. We contribute to work on exploratory search by demonstrating the effectiveness of the subjunctive interface in supporting exploration. Additionally, we analyze the affect of this type of exploration on the research question of media studies researchers.

- **An algorithm to automatically generate links between records with different representations and originating from different archives.**

  We introduce a method that generates links between records from a newspaper archive with a rich textual representation to records from an audio-visual archive with relative sparse representations. We extend existing work by demonstrating the effectiveness of term selection and document expansion to overcome issues arising from sparsity and topic drift.

- **Analysis of the effectiveness of variations of the linking method on linking records describing the same event and records describing related events.**

We investigate the effectiveness of document expansion, term selection, and time restrictions to boost performance of linking to records describing the same and related events. These findings add to the body of work on tracking events.

- **Analysis of the effectiveness of text-based, structure-based, and hybrid methods for example-based entity finding in the Web of Data.**

  We investigate the sensitivity of text-based, structure-based, and hybrid methods for example-based entity finding, when provided with sets of example entities of varying size and composition. This analysis shows the sensitivity of existing methods to specific sets of examples and relevance judgements.

- **A hybrid example-based entity finding method that is able to more effectively combine text-based and structure-based approaches on a per query basis than a hybrid method optimized for a batch of queries.**

  We introduce a hybrid method that uses example entities to make a decision about using a text-based method to find entities, a structure-based method, or a combination of both methods. We extend existing work by demonstrating the robustness of this approach to example entity sets of varying composition and quality.

- **A transparent architecture for related entity finding.**

  We introduce a transparent architecture for a system aimed at tackling the related entity finding task with detailed descriptions of its components and demonstrate ways to extend the system with additional heuristics. This system provides an alternative to other systems used to address the related entity finding task in which design choices and combinations of individual components are more ad-hoc.

- **An analysis of the effectiveness of each of the components of a related entity finding system.**

  We provide a detailed analysis of the effectiveness of each of the components of our related entity finding system. We add to existing work by showing the benefits of using Wikipedia versus a web corpus, the utility of various co-occurrence methods for candidate selection, and how incorporating various heuristics in a related entity finding system affects performance.

## 1.3 Thesis Overview

This thesis is divided into two parts. The first part, consisting of Chapters 2, 3, 4, and 5, describes the research process of media studies researchers and the development and evaluation of tools to support them. The second part, consisting of Chapters 6, 7, 8, and 9, is motivated by, but somewhat independent of the first part and describes work on algorithms to support discovery of information that provides background knowledge for a research topic.

Chapter 2 introduces background on search behavior, research practices in the humanities, and interactive information retrieval. Chapter 3 then goes into a study of the research process of media studies researchers. The observations in this chapter about the

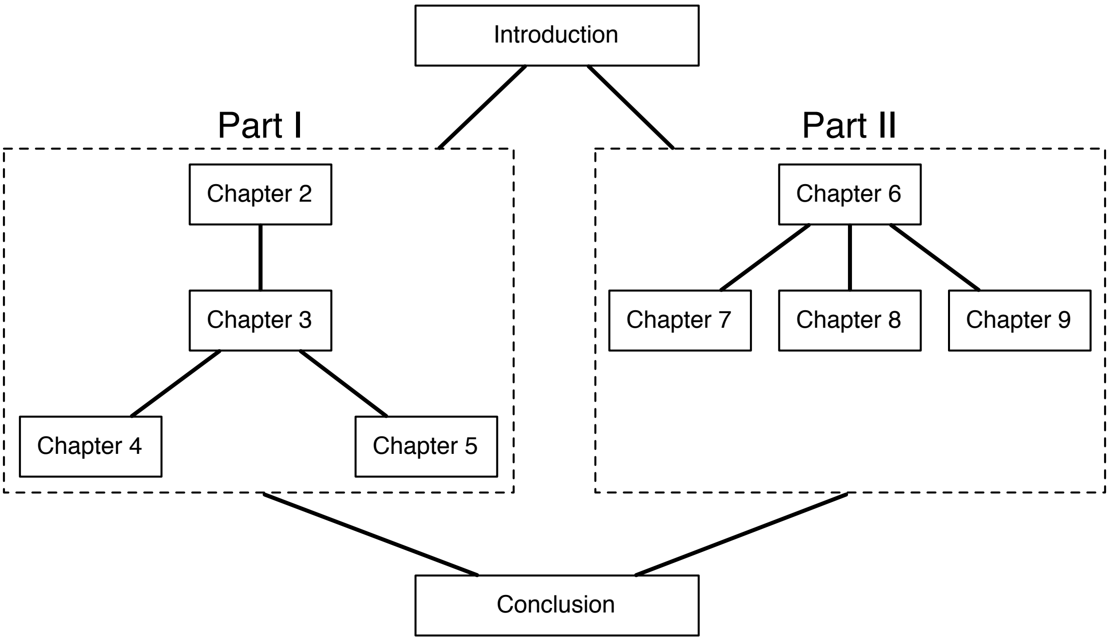


**Figure 1.4:** Overview of the organization of the thesis. Lines indicate where chapters provide necessary background for subsequent parts.

need for exploration and contextualization during the information gathering processes of media studies researchers across various sources motivates several of the chapters in this thesis. Specifically, Chapter 4 describes two studies into the use of aggregated search displays by media studies students during a multi-session search task. Chapter 5 proposes a novel interface, i.e., a subjunctive exploratory search interface that allows comparisons to be made between search results in a side-by-side display, and describes the results of a user study that evaluates the effectiveness of the subjunctive interface in supporting exploration.

The second part of the thesis starts with an overview of the field of information retrieval. Chapter 6 highlights methods that go beyond traditional document retrieval and methods that use structured data to improve search, i.e., from the area of semantic search. The next three chapters each describe an alternative method to enable the discovery of information related to a topic, i.e., methods to enable contextualization. Chapter 7 describes a method to automatically generate links between records with rich textual representations in a newspaper archive, to records in an audio-visual archive with sparse annotations. Chapter 8 introduces a method to find related entities based on examples exploiting the rich structure provided by the Web of Data. Chapter 9 describes an architecture of a system to find related entities in a web corpus based on a (source) entity and a free text description of a relation that must hold between the source and target entities. Finally, Chapter 10 summarizes our answers to the research questions raised in Chapter 1 and provides an outlook on future work. An overview of the organization of the thesis is provided in Figure 1.4.

The prototypes of the interfaces that we use in our studies in Chapter 4 and 5 as well as the test collections we create to evaluate our methods in Chapter 8 and 9 are made available. Additional information about these interfaces and resources is provided in Appendix A.

## 1.4 Origins

The work presented in this thesis builds on a number of publications; below we list the work that forms the basis for each chapter.

- Chapter 3: "Media Studies Research in the Data-Driven Age: How Research Questions Evolve" by Bron, van Gorp, and de Rijke, submitted to JASIST;
- Chapter 4: "Aggregate Search Interface Preferences for Multi-session Search Tasks" by Bron, van Gorp, Nack, Baltussen, and de Rijke, published as full paper at SIGIR 2013 [72];
- Chapter 5: "A Subjunctive Exploratory Search Interface to Support Media Studies Researchers" by Bron, van Gorp, Nack, de Rijke, Vishneuski, and de Leeuw, published as full paper at SIGIR 2012 [70];
- Chapter 7: "Linking Archives Using Document Enrichment and Term Selection" by Bron, Huurnink, and de Rijke, published as full paper at TPDL 2011 [67];
- Chapter 8: "Example Based Entity Finding in the Web of Data" by Bron, Balog, and de Rijke, published as full paper at ECIR 2013 [71]
- Chapter 9: "Ranking Related Entities: Components and Analyses" by Bron, Balog, and de Rijke, published as full paper at CIKM 2010 [65].

Indirectly, the thesis also builds on work on related but separate topics. They are provided below as pointers for further reading. In particular, related to Part I is work on user studies:

- "Exploratory Search in an Audio-Visual Archive: Evaluating a Professional Search Tool for Non-Professional Users" by Bron, van Gorp, Nack, and de Rijke, published as full paper at EuroHCIR 2011 [68];

the development of interfaces

- "Ingredients for a User Interface to Support Media Studies Researchers in Data Collection" by Bron, Nack, de Rijke, and van Gorp, published as full paper at EuroHCIR 2012 [69],
- "AVResearcher: Exploring Audiovisual Metadata" by Huurnink, Bronner, Bron, van Gorp, de Goede, and Wees, published as demo paper at DIR 2013 [174];

and log analysis

- "Characterizing Stages of a Multi-Session Complex Search Task through Direct and Indirect Query Modifications" by He, Bron, and de Vries, published as poster at SIGIR 2013 [162].

Specifically related to Part II is work on entity search:

- "Category-based Query Modeling for Entity Search" by Balog, Bron, and de Rijke, published as full paper at ECIR 2010 [32],
- "Query Modeling for Entity Search Based on Terms, Categories and Examples" by Balog, Bron, and de Rijke, published as journal paper in TOIS 2011 [35];

and entity linking

- "Learning semantic query suggestions" by Meij, Bron, Hollink, Huurnink, and de Rijke, published as full paper at ISWC 2009 [237],
- "Mapping Queries to the Linking Open Data Cloud: A Case Study Using DBpedia" by Meij, Bron, Hollink, Huurnink, and de Rijke, published as journal paper in Web Semantics: Science, Services and Agents on the World Wide Web 2011 [238].

Other valuable experiences have been gained by participations in various evaluation campaigns such as the TREC entity [27, 28, 66], INEX entity [30, 33], and the TAC knowledge base population and entity linking [153, 202] evaluation tracks.

Work not directly related to the thesis helped shape a broader understanding of the information seeking and information retrieval landscape: on diversity [160], implicit relevance judgements [169, 173], and reasoning in ontologies [194].

# Part I

# Exploration through Interaction

# 2 Background on Information Behavior

The work in this thesis on developing tools and algorithms to support humanities researchers draws on background material from several areas. In particular, it builds on work from the field of information behavior, which is the study of human interaction with information, and on work in information retrieval, which concerns the development of efficient and effective systems to search and manage information. In this chapter we provide an overview of research from three areas in information behavior and cover *models of information seeking behavior*, *information needs and habits in the humanities*, and *work on interactive information retrieval*, see Figure 2.1. Interactive information retrieval has roots both in information behavior and information retrieval. Our discussion here focuses on interfaces that support particular types of user interaction with retrieval systems and not on the underlying retrieval algorithms. In Chapter 6 we provide further background on traditional information retrieval systems, tasks that go beyond document retrieval, and algorithms that support semantic search.

We start with a brief overview of the origins of the field of information behavior in Section 2.1. We then discuss work dealing with various models of information behavior in Section 2.2, which provides the theoretical background for the investigation of research behavior in the humanities. In Section 2.3 we cover work dealing with information needs, habits, and tools in the humanities, which provides insights in the practical issues involved in specific areas of humanities research. These parts provide the nec-

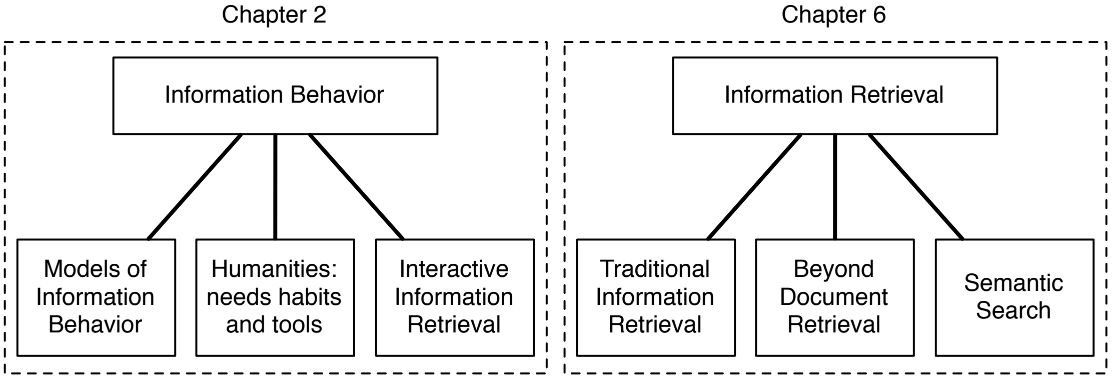

Chapter 2
Information Behavior
Models of Information Behavior
Humanities: needs habits and tools
Interactive Information Retrieval
Chapter 6
Information Retrieval
Traditional Information Retrieval
Beyond Document Retrieval
Semantic Search

**Figure 2.1:** Overview of the order in which background material for the work in this thesis is provided.

essary background for part I of the thesis and for Chapter 3 in particular. Finally, in Section 2.4 we discuss work from the area of interactive information retrieval. We focus on studies that investigate factors related to the search task, the user, and the interface, which informs the design of our experiments in Chapter 4 and 5.

The vastness of the body of literature that constitutes the field of information behavior prohibits the inclusion of a complete review of the work within the scope of this thesis. We refer the reader to overviews in the ARIST series [105, 119, 242, 261] for earlier work on information behavior. For a more recent review see [88].

## 2.1 A Brief History of Information Behavior

Interacting with information is a part of human nature. It allows us to learn about our environment and make decisions beneficial to our survival. In the early days of human evolution, information was passed along through speech and signaling [212]. Later, written language systems developed that allowed asynchronous exchange of information through symbols on physical objects such as clay tablets and papyrus scrolls [367]. As the production of these records became cheaper and more widespread a new problem emerged: finding a piece of information within newly emerging and expanding collections of records.

To facilitate the retrieval and storage of records, libraries were created and librarians started developing new ways of organizing and indexing these collections of records. One of the first known instances of a tool to support search is an index dating back to 300 BC [260, 265]. An index provides a mapping from representative keywords to records and enables the lookup of records through searching by keyword. Through the ages the creation and maintenance of indexes remained a manual effort. This changed halfway during the 20th century when the topic of automatic indexing received increased attention. Driven by the vision of automatic systems that support information access [79], researchers set out to create more efficient and effective information retrieval systems. This development gave rise to two new areas of study: information retrieval, which concerns the study of systems [251] and information science, which concerns the theory and practice of information and its relation to the world [61, 146, 297].

During the 1950s work in information science focused on identifying scientists' use of and satisfaction with information systems in order to derive requirements for the development and improvement of information systems [242]. In the years that followed the focus started to broaden from the delivery of information by a system to a user, i.e., retrieving, to include the theory of information, communication, and interaction between information and users [3]. In this thesis we do not address theoretical studies of information or social studies of inter human communication. Instead, we focus on human behavior in relation to information, an area known as *information behavior* [365].

Meanwhile, the number of studies on the subject of information behavior rapidly increased and a 1986 review of the information science literature [119] brought to the attention a split between empirical and theoretical studies that divided the research field. The more prevalent empirical studies of that time perceived users as passive receivers of objective information independent of the search environment [119]. These studies typically investigated the information used by a particular group of users to arrive at a set

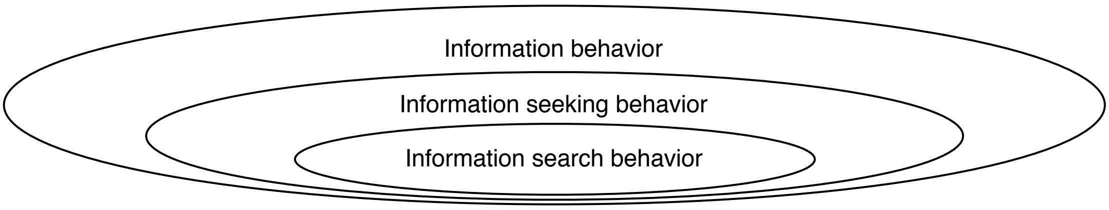


**Figure 2.2:** Wilson [364]'s nested model of studies focusing on particular sub-sets of information behavior.

of requirements for tools and search strategies. This led to numerous studies investigating increasingly diverse groups of users interacting with different systems and engaging in different tasks [105].

In contrast, theoretical studies in information behavior sought to bring order to the disarray of empirical studies and set out to develop models that incorporate and explain findings regarding specific groups of users. These models moved away from the idea of a static user unaffected by the information he/she encounters. As a result, a user's cognitive states and the ways in which these are affected by the user's interaction with information were considered as factors in models of information behavior. We adopt this cognitive view on information interaction and in the rest of the thesis will consider information as a thing [76], i.e., recorded knowledge that is collected and processed, thereby transforming the information into a user's knowledge [176].

## 2.2 General Models in Information Behavior

In this section we provide an overview of generally accepted theoretical models of information behavior and identify where we place our investigation of the research behavior of humanities researchers against this background.

An early version of a general model of information behavior introduced two new notions: (i) it made a distinction between the user, his/her need, the information source or service, and use of the information; and (ii) it regarded the interaction between these concepts as an iterative process [362]. Here, the concept *information need* is defined as the desire to locate and obtain information to satisfy a conscious or unconscious need [317]. In a revision of his information behavior model, Wilson [364] added a number of contextual factors that influence the user in engaging in or stopping the process of satisfying his/her information need. For example, stress factors that prevent a user from engaging in information seeking, task, or role related factors that intervene with the search process, and reward factors that motivate continuation of search. The models identified a void in studies of information behavior, i.e., the journey of the user towards resolving his/her information need. The sense-making theory characterized this as the sense of uncertainty and incompleteness that users experience in the gap between a situation in which an information need arises and its resolution [117]. In the same publication a methodology for the investigation of information behavior was introduced, based on asking detailed interview questions about a user's subsequent activities conducted to resolve an information need.

These general models of information behavior gave rise to more specific studies on specific types of information behavior. Figure 2.2 shows Wilson's nested model of studies focusing on particular sub-sets of information behavior, i.e., information seeking behavior and information search behavior, which are discussed below.

### 2.2.1 Information Seeking Behavior

Models that focus on the process that spans from a user engaging in information related activities in response to a need until its resolution, are known as models of *information seeking behavior* [363] and are more prevalent than the information behavior models.

A series of studies into the information seeking behavior of users in a variety of disciplines, e.g., social scientists, engineers, and physicists, revealed a set of typical behaviors that these users engage in to resolve an information need [128, 130]. Specifically, these search behaviors are: (i) initiating search, e.g., asking a colleague (*starting*); (ii) following footnotes and citations (*chaining*); (iii) semi directed searching (*browsing*); (iv) filtering sources based on known characteristics (*differentiating*); (v) keeping up to date with a topic (*monitoring*); (vi) selecting relevant material (*extracting*); (vii) determining the accuracy of information (*verifying*); (viii) a final search, making sure nothing was missed (*ending*). Starting and ending hints at an order for these behaviors, but no particular order was suggested.

In later views these behaviors were seen as part of a larger behavioral process in which a user explores an unknown information space and applies various search behaviors influenced by the information encountered. For example, the "berry picking" model in which a user's query evolves and different behaviors alternate, e.g., chaining and browsing [41]. Russell et al. [291] associated costs to activities as finding and accessing information during a exploration of an information space. This notion was elaborated on in the hunter-gatherer culture inspired theory of information foraging, that explains changes in search, gathering, and consumption behavior through people's motivation to engage in the behavior that maximizes their rate of gaining valuable information in a specific environment [272]. O'Day and Jeffries [258] described the process of a user executring incremental steps in order to find an answer to an information need as *orienteering*. Instead of trying to jump to the answer directly with a long precise query, i.e., teleporting [318], users start with a general query and use information from their current search to determine their next steps.

Through a series of studies Kuhlthau [197] arrived at a model of information seeking behavior as an overarching process and named it the *information search process* (ISP). The model connected several behaviors to stages in the ISP of library users: initiation, recognition of a need, selection and identification of a topic, exploration of relevant information, formulation of a focused topic, collection of relevant information, and presentation of search results. To corroborate and refine this work, Vakkari [338] integrated the findings of a longitudinal study into the ISP model. In this study students' information seeking behaviors were monitored during the proposal writing phase for their theses. Fine-grained activities were observed in the search behavior of students and identified as *search tactics*. In particular, students changed the information they sought for, their criteria to determine relevance, and their search terms at various stages of the process. Search tactics were introduced earlier by Bates [42] in the context of bibliographic and reference searches and are defined as the low level actions that constitute search, e.g., adding search terms to a keyword search or inspecting results.

The above discussion illustrates that there are views of information seeking behaviors at different levels of granularity. Several propositions for the number of levels and their scope have been made in the literature. Bates [43] identified four levels: (i) a *move* is the smallest unit consisting of an identifiable thought or action in information searching; (ii) one level up a *tactic* consists of a move or a number of moves applied to advance the search process; (iii) a *stratagem* combines multiple tactics that exploit a specific system or source with a specific mode of searching; and (iv) a *strategy* represents a plan build up from moves, tactics, and/or stratagems to complete a search process. Similarly, Marchionini [230] proposed four levels of behavior with on the lowest level *moves*, followed by *tactics*, i.e., choices for a particular sequence of moves. The other two levels, however, do not align exactly with those of Bates as the concept of *strategy* is defined as a set of ordered tactics to solve a search process and *patterns* are combinations of strategies and tactics favoured over time and across search processes.

In this thesis we view the *information seeking process* as consisting of all the thoughts, actions, and feelings a user goes through from the conceptualization of an information need until its resolution. The information seeking process consists of search processes with a certain system, source, and goal. The particular search move and tactics exhibited during the search process are determined by the strategy for the search process.

In Chapter 3 we identify how the behavior of a group of humanities reseachers during a research project aligns with these models of information behavior.

### 2.2.2 Information Search Behavior

As technological developments continued, the focus of models of information seeking behavior shifted focus towards the information retrieval system as an important factor that influences the strategies applied in the information seeking process. Studies in information search behavior focus on models that capture the interaction between the user, a system and how this interaction influences user behavior.

An early model taking the system perspective describes information needs as anomalous states of knowledge which the searcher attempts to resolve through various interaction strategies with a retrieval system [46, 47]. Belkin et al. [48] later proposed the episode model in which the interaction of the user with an information retrieval system is guided through scripts. A script specifies the interactions between a user and a system necessary to complete a certain search strategy and can be combined in sequences of search strategies (episodes). Other work reflects the same focus on interaction, e.g., Spink [310] described search strategies as sequences of cycles in which search tactics, interpretation, and user feedback alternate.

Saracevic [298] summarizes the dimensions involved in the information seeking process as: (i) informational resources; (ii) computational resources; (iii) interface; (iv) query characteristics; (v) user knowledge; (vi) situation; and (vii) environment. Marchionini [230] describes a similar set of dimensions but adds task as a factor influencing behavior in the information seeking process.

Ingwersen and Järvelin [175] provide a comprehensive model of the information seeking process. Their model describes several dimensions each containing additional

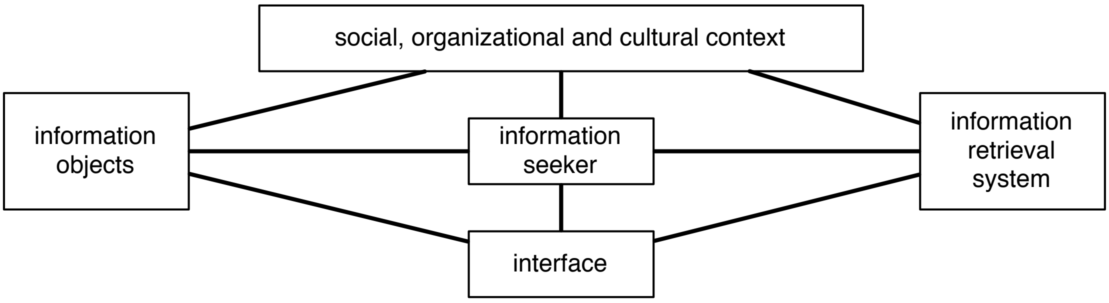

social, organizational and cultural context
information objects
information seeker
information retrieval system
interface

**Figure 2.3:** Simplified version of Ingwersen and Järvelin [175]'s model of the dimensions involved in the information seeking process. Lines indicate possible paths of interaction between dimensions.

factors that influence the information seeking process. A simplified version of the model showing the dimensions and their interactions is presented in Figure 2.3. The dimensions are: (i) the information retrieval system; (ii) the interface; (iii) the information objects; (iv) the information seeker; and (v) the social, organizational, and cultural context. The lines represent interactions between dimensions, e.g., an information seeker is influenced by the type of interface he/she is presented with. Ingwersen and Järvelin [175] further argue that an information seeking process does not stand on itself and needs to be considered in the context of other information seeking processes as part of a larger *work task*, which in itself is part of a social, organizational, and cultural context. Byström and Hansen [81] share this view and provide a conceptual model of the work task. In their model the work task consists of multiple information seeking tasks, each further consisting of multiple information retrieval tasks.

These conceptual models provide the 'big picture' of the dimensions involved in the information seeking process. However, to develop hypotheses and design experiments these dimensions need to be instantiated with details to arrive at a scenario involving a specific type of information seeker (a humanities researcher), system (a specific tool/interface), and environment (a particular type of information object or habit the system supports). In §2.3 we discuss previous work on the information needs, research habits, and tools of humanities researchers. This is complemented by our investigation in Chapter 3 on a specific group of researchers, i.e., media studies researchers. Together, these provides the necessary details to instantiate the scenario of the humanities researcher as information seeker used in the experiments in Chapter 4 and 5. In §2.4 we discuss specific experiments investigating scenarios closely related to ours, e.g., the use of the same type of interface within a different context.

## 2.3 Humanities Researchers' Needs, Habits, and Tools

The theoretical models of information behavior in the previous section generalize over studies and observations of users in different environments. They provide a theoretical framework to organize existing work, but lack the details necessary to design effective tools [60, 324, 336]. Although models of information seeking behavior are more detailed in describing tactics and search moves, specific instantiations of these depend on the specific group of users under study, their context such as the task, and the environment in which the information seeking processes take place. The empirical line of research in information behavior investigates the specific information needs and habits of various groups of users. The long tradition of this work and continued study of specific groups is further motivated by the observation that the needs and behaviors of researchers continually change as new technologies are introduced [60, 241, 336].

The work in this thesis is focused on the development of tools and algorithms to support humanities researchers. Therefore, in our discussion of information needs, habits, and tools, we limit ourselves to studies involving the humanities and leave out work on the information needs and habits of other groups, e.g., social scientists or engineers. Below we organize our discussion in three parts and start with investigations of the information needs of humanities researchers, followed by studies of the humanities' research cycle, and end with a review of tools developed to support humanities research practices.

### 2.3.1 Information Needs in the Humanities

Although there is no consensus on which disciplines constitute the humanities, generally included are fields of philosophy, religion, languages and literature, linguistics, music, art, history and media studies [311, 359].

Work investigating the information needs of humanities researchers has primarily focused on the study of information needs through observations of library use. In an overview of research on information needs and uses in the humanities from 1970 to 1982, Stone [311] observed that humanities researchers tend not to use databases or electronic tools, that they work alone, primarily use books and journals, and to a lesser extent primary source material, such as photos and diaries.

In a 1988 assessment of information needs in the humanities, a broadening of interest to include popular culture as well as high culture across all disciplines is observed. This had an effect on the types of materials studied by researchers. In history, comic books and radio shows became objects of study; art scholars started to consider the culture surrounding the production of art; literary scholars interested in the historical context of literary works began seeking out court records, medical texts, and contemporary accounts of witchcraft and exorcism; and in music, composers became an object of study [151]. This change in interest towards primary source materials, e.g., films, photographs, videotapes, prints, and works of art, suggested the need for more sophisticated descriptions of and access to these materials.

In a study following 11 scholars between the 1980s and 1990s it was found that in accessing primary source materials, such as stored in archives, scholars relied heavily on archivists for support in their search for material [360]. In contrast, researchers in search of bibliographic material are self reliant. This illustrates the difficulty humanities scholars have in discovering primary source material in archives as compared to bibliographic material.

In an overview of studies published from 1983 through 1992 some additional observations are made about the information needs and habits of humanities scholars [351]. Citation analyses show that both primary source material and bibliographic (secondary) materials are frequently used. The primary source materials are the central focus, while secondary material provides facts or opinions. To identify relevant material for a new research topic, browsing was found to be a popular strategy, e.g., going through references, browsing library shelves, and scanning periodicals. Feelings about the use of electronic tools were mixed, some researchers confirmed the need, but expressed difficulties in keeping up to date with the technology, while others preferred to rely on their knowledge of the field and using references in books and journals.

As electronic library database systems became available, the use of these tools by researchers became the subject of study. Bates [44] observed the search behavior of a small group of scholars with the bibliographic and full text search system DIALOG. The scholars in the study did not make frequent use of the system and continued to use books and articles to locate related literature. There were some positive reactions to the system regarding its support for exploring and discovering unexpected material.

The rapid technological changes of the last 15 years have motivated numerous studies on how these changes have impacted the information seeking habits of scholars. A summary of studies investigating electronic text usage up to 2002 through citation analyses found that use of electronic texts among humanities researchers increased [245]. Researchers noted as advantages the accessibility and additional features, e.g., hyperlinks. Disadvantages included inadequate indexing and lack of standardization. In a comparison of the information needs and habits of historians observed in 1981 and 2004 it was found that characteristics such as informal means of discovery through book reviews and browsing did not change. An increased use of electronic resources to access primary and secondary sources was observed [109]. In a 2007 case study of the information use of Jewish Studies scholars it was found that they had a positive attitudes towards technology and use of electronic channels [39]. Similar results were reported in studies of literature researchers [129] and students in the humanities [38]. These findings suggest a trend in the humanities towards a positive attitude regarding information technology.

The above discussion illustrates how the information needs of humanities researchers have changed over time. In Chapter 3 we investigate the current information needs and information gathering challenges faced by media studies researchers.

### 2.3.2 Models of the Humanities' Research Cycle

Complementary to studies of the need for and use of material, models of the research cycle provide insight in why and how certain materials are used, the contexts in which information needs arise, and the various other processes surrounding researchers' information seeking behavior.

Work in the 1970s identified three stages in the research cycle of economists: (i) the problem stage, in which an idea is developed and hypotheses are formulated; (ii) the methodology stage, in which first the technique for collecting data is determined and then the data is gathered; and (iii) the presentation stage, in which the data is analyzed, interpreted, and disseminated [353]. During these stages several research processes were observed: (i) perception of idea; (ii) definition of problem; (iii) development of methodology; (iv) provision of data; (v) suggestion of information source; (vi) analytical assistance; and (vii) practical assistance. Although the processes imply a certain order, they are recurrent and each process is observed during all three research stages. Subjects in the study became more purposeful in their search in the second stage. This observation is consistent with earlier findings that researchers become better at formulating their information need and determining relevance at later stages of their research [159]. A similar division of the research cycle in stages was proposed in the field of psychology, i.e., idea generation, development of the problem, and presentation of results [142].

In the 1980s Stone [311] suggested five steps that scholars go through during research: (i) thinking and talking to people about the topic; (ii) reading what has already been written on it; (iii) studying original sources of information and making observations and notes; (iv) drafting a document on what is found, and (v) revising the draft into a final document.

Regarding historians Uva [337] identified five stages in the research cycle: problem selection; detailed planning of data collection; data collection; analysis and interpretation; writing-rewriting. In the data collection stage historians expressed the importance of obtaining primary source materials. The need for archival search to locate primary source materials was recognized in several earlier studies [89, 151, 351]. Contrary to the attention that users of library systems received from research on information behavior, few studies examined users' information seeking behavior in the context of archives and electronic archival search tools. Duff and Johnson [123] identified four types of information-seeking activities exhibited by historians, including (i) orienting oneself to archives, finding aids, sources, or a collection; (ii) seeking known material; (iii) building contextual knowledge; and (iv) identifying relevant material. During the research cycle these activities of revisiting material and (re-)examining finding aids lead to refinement of the questions as historians build their contextual knowledge and increase their understanding of the research topic. This contextual knowledge is imperative to historians, without it interpretation of the material under study is pointless. Use of archival research tools was found to be limited. The most popular aids to locate relevant material were published finding aids, citations in published resources, and colleagues.

Cole [98] described a particular (recommended) way in which history students should access archival material. A student would start by reading secondary literature, becoming familiar with the topic of interest, and noting down all proper names. Only then is it recommended to start the search for these names in finding aids and the student is warned not to expect relevant material to be accessible by subject.

Case [89] studied the motivations behind the needs for primary and bibliographic material. He observed that the steps in the research cycle of historians are not sequential. Rather, historians embed themselves in material related to a general topic and are guided through their knowledge, intuitions, and interests that continue to shape and focus their questions. This process of embedding oneself in material is a purposeful process in which historians meticulously review and structure material. Historians organize primary sources according to topic, period, or person, and move back and forth between writing and inspecting material in order to arrive at a holistic view of the topic.

Chu [91] investigated another humanities discipline, i.e., literary criticism, and introduced a model of the research cycle with six stages: *idea*, *preparation*, *elaboration*, *analysis and writing*, *dissemination*, and *further writing and dissemination*. He described several variants of the model with three, four, and five stages to accommodate the various behaviors exhibited by individual critics. A model of the research cycle of music scholars identified similar stages [73]. The model contained an additional stage to categorize activities related to preparation and organization of controlled experiments and interviews.

**Table 2.1:** Overview of the models proposed for the research stages of various disciplines.

| | economists<br>White [353] | humanities<br>Stone [311] | historians<br>Uva [337] | literary critics<br>Chu [91] | music scholars<br>Brown [73] | media studies<br>Lunn [222] |
|---|---|---|---|---|---|---|
| 1 | problem | thinking; talking | problem selection | idea generation | idea generation | overview of broadcasts |
| 2 | methodology | reading literature | planning data collection | preparation | background work | selection for analysis |
| 3 | presentation | studying original material; notetaking | data collection | elaboration | preparing; organizing | identify exemplars |
| 4 | | drafting | analysis; interpretation | analysis; writing | analyzing | verification of facts |
| 5 | | revising | writing; rewriting | dissemination | writing; revision | |
| 6 | | | | writing; dissemination | dissemination | |

Lunn [222] studied the information needs of users of an audiovisual archive. He identified four phases in the information needs of one group of users, i.e., media studies researchers: (i) getting an overview of broadcasts; (ii) selection of specific broadcasts for analysis, (iii) identification of borderline exemplars, (iv) verification of facts. However, the focus of his work are the information needs and not the relation between these phases and the overall research cycle.

Table 2.1 provides an overview of the various stages in the research process of humanities researchers. Although some of these stages appear to be a single activity, each stage is associated with several activities, e.g., the *idea generation* stage of Brown [73] consists of activities such as studying previous work, reading (music) literature, and discussions with colleagues. The specific activities that occur during these stages depend on the discipline, e.g., listening to music is particular to the research cycle of music scholars.

The above discussion shows that models of the research process of several humanities disciplines share similar stages. In Chapter 3 we investigate how the research cycle of media studies researchers aligns with existing models of the humanities research cycle.

### 2.3.3 Development and Adoption of Tools in the Humanities

The discipline, the research topic, and characteristics of the individual humanities researcher all play a roll in determining the sequences of stages observed during research. Depending on the stage certain research activities are more or less prominent. Unsworth [335] proposed a list of research processes or so called *scholarly primitives* that humanities researchers engage in: discovering, annotating, comparing, referring, sampling, illustrating and representing. In 2003 Unsworth [336] observed that 10 years of tool development for the humanities had yet to result in portable, reliable, and extensible tools. He advocated building a modular and extensible toolkit that supports the scholarly primitives of humanities researchers. Toms and O'Brien [324] also noted a lack of support offered by tools that were being developed for humanities researchers and listed a number of requirements for such tools: (i) scanning and browsing to enable exploration of a text or set of texts; (ii) providing access and overviews of resources via links and portals; (iii) supporting downloading, storing, and organizing texts; and (iv) enabling note taking, text analysis, and communication with colleagues.

Palmer et al. [263] provided an updated overview of the various types of activities humanities researchers engage in during the research cycle. She defines six categories of scholarly primitives: searching, collecting, reading, collaborating, writing, and cross-cutting primitives. Primitives associated with searching correspond to search strategies [41] such as *directed searching*, *chaining*, *browsing*, *probing*, and *accessing*. In Chapter 4 and 5 we investigate whether interfaces designed for exploration can provide support for primitives such as directed searching, browsing, and probing.

Other primitives are not search activities but are considered to be on the same level as search tactics and several of these primitives can take part in a scholars' information seeking behavior. Search leads to discovery of relevant material and *information gathering*. The collected material becomes part of humanities researchers' personal collections. Material is gathered through, for example, visiting a library, archive, or downloading. To facilitate access to and re-finding of material humanities researchers organize material into a structure that suits their research purpose. Depending on the type of material *reading* or *viewing* activities are part of the research cycle. Humanities researchers scan material, assess its value, and reread or review material in depth. These activities are also described as differentiating, comparing, and sifting. Although *collaboration* is observed less in the humanities than in other disciplines, this activity still occurs. Primitives involved in collaboration are coordinating, networking, and consulting. Humanities researchers' *writing* activities are aimed at composing their thoughts regarding the material and topic of study. The goal of assembling is to obtain an overview of the concepts and relationships that exist within the domain of a research topic. In Chapter 8 and 9 we look into methods that support this type of activity.

*Cross-cutting primitives* are scholarly primitives that may occur in any of the above categories. Monitoring is the activity of maintaining awareness of current developments in the field. Note-taking occurs in all stages as it helps researchers to order their thoughts. Translating occurs in cross-disciplinary studies when researchers need to learn the concepts and research practices of another discipline. Finally, data practices are the activities related to generating, managing, and sharing of data studied or produced by humanities scholars.

Rather than supporting existing practices, however, most tools demonstrate novel ways to apply technology to transform practices in the humanities. An exception is Pliny, which is a tool developed to support existing practices of humanities researchers, i.e., interpretation of texts [62]. It supports actions associated with a model consisting of three research phases: (i) annotations and note taking during reading; (ii) developing interpretation by organizing texts and annotations; and (iii) publishing through exporting representations of annotations and organized material.

MONK[1] is a data analysis laboratory that provides visualizations and data mining tools. It supports the analysis of datasets that have been collected and transformed in the appropriate format. Similarly, TaPoR[2] is a portal for a variety of indexing, searching, and

[1] http://monkproject.org/

[2] http://portal.tapor.ca/

text-analysis tools. Another example of this type of visualization laboratory is Manyeyes, a web service that allows humanities researchers to upload data sets and select various visualization options [343]. Turknett et al. [333] are exploring new visualization possibilities with large displays. They provide a high level programming interface that allows humanities researchers to visualize data on multi panel screens.

The Google ngram viewer[3] is a tool that enables comparison of trends in term occurrences on a timeline in a large book collection supporting cultural analytics [243]. A limitation of this tool is is the lack of support for exploration or inspection of the actual books. In the DARIAH project a general framework to support virtual research environments is under development [59]. In this framework scholarly practices are supported through collaborative text editors, text analysis, and communication modules. To support a certain research cycle an environment is constructed by combining various modules [58].

Amin et al. [8] proposed a number of tools to support the information seeking tasks of experts in the cultural heritage domain. One tool focused on supporting source selection by displaying credibility ratings for each source [10]. Other tools supported comparison of data and autocompletion of search terms based on thesaurus information [9, 11].

Despite efforts towards the development of effective tools, use remains sporadic. Borgman [60] made a call to action for the humanities. She invited more behavioral research in the humanities as the discipline is evolving rapidly. Regarding the infrastructure she notes that it has been designed for the sensor based data common to data intensive disciplines such as astronomy. Now, it is time for the humanities research community to articulate its requirements.

Others seek the reason for the lack of adoption of tools in the theory of satisficing, i.e., people will make a trade-off between effort and reward and settle for something that is good enough [319]. Information management students were observed to whenever possible select search strategies and information sources that they are comfortable with and only engage in more complex search strategies if it was required by the assignment. Wiberley and Jones [358] make a similar observation in a 10 year study of a group of humanities researches. They point out that researchers stick to existing practices unless a gain in time is demonstrated. If this gain in time is marginal or perceived not to outweigh the start-up time to learn the new technology, the researcher is not inclined to adopt the tool.

From the above discussion we learn that the success or failure of tools to support the humanities depends on how well they support the scholarly primitives. Further, the adoption of tools depends on whether tools are perceived to save humanities researchers time and are easy to use. In Chapter 4 we investigate whether the preference of media studies students for a particular tool changes during a research project. As this has implications on its perceived usefulness. In Chapter 5 we address the challenge of supporting a particular scholarly primitive, i.e., comparing, with a novel type of interface.

[3] http://books.google.com/ngrams

## 2.4 Interactive Information Retrieval

The conceptual models of information seeking from §2.2 provide an overview of the various dimensions involved in the information seeking process. The studies from §2.3 provide insights in the practices of humanities researchers and the context in which they work. From these discussions a picture emerges of humanities researchers who place different demands on search systems at different times depending on the stage of the research cycle. The development of successful search systems to support humanities researchers, therefore, depends on whether these systems provide the right support at the right time under the appropriate circumstances. In our discussion below we review experiments that investigate specific factors related to the interaction of the user with a particular search system, in a specific environment, and how these affect each other during the information seeking process.

Kelly [190] characterizes interactive information retrieval studies by placing them on a continuum bounded on one side by studies focused on system factors, such as information objects and retrieval functions, and on the other side by studies focused on human and search environment factors. On the one hand, focusing on the system side allows experiments to be carefully controlled, but leads to rigid assumptions about the human and search environment. On the other hand, the number of factors involved when considering the user and his/her search environment prohibit controlled experimentation. Typical interactive information retrieval experiments balance between the two extremes by investigating the support that a particular system or interface provides for a particular user involved in a particular task. We focus here on studies that investigate one of the following dimensions: (i) the influence of the task on a user's interaction with a search system; (ii) the differences in behavior of different types of information seeker with the same system; and (iii) the benefits of one type of search interface over another. These studies inform the development of our experiments in Chapter 4 and 5.

### 2.4.1 The Task

One way of defining tasks is by their goal. With the rise of the Web, an increasing number of studies has focused on web search task goals. Broder [64] introduced a taxonomy of web search goals and identified three types of need: (i) *navigational*, i.e., reach a particular site; (ii) *informational*, i.e., obtain information assumed to be available somewhere on the Web; and (iii) *transactional*, i.e., perform some web based activity. Kellar et al. [188] classified web behaviors as tasks and defined five types: fact finding, e.g., looking up a phone number; information gathering, e.g., searching for summer school courses; (iii) browsing, e.g., looking for something to read; (iv) transactions, e.g., banking; and (v) other, i.e., meta behavior such as viewing web pages during development. Several other classifications of web search task types [90, 252, 289, 292] and web search goals [180, 288] have been proposed.

In general, depending on the context and the goal, a task may be characterized as simple, e.g., looking up an information object of which the distinguishing characteristic are already known, i.e., title, location, or author. Other tasks are more complex, e.g., when information seekers are unable to specify their information need before hand and iteratively search for and process information about a topic before resolving the task [356].

The complexity of a task tends to be attributed to the level of structure (familiar patterns), uncertainty, and cognitive effort involved. Byström and Järvelin [82] identified five levels of increasing task complexity: (i) *automatic information processing tasks*, which have the lowest complexity as they are a-priori completely determinable; (ii) *normal information processing tasks*, which are almost completely determinable but require some judgement or reasoning; (iii) *normal decision tasks*, which are structured tasks according to a familiar pattern but require major reasoning or judgement steps, e.g., grading a student term paper; (iv) *known, genuine decision tasks*, which have a clear goal but the procedures to arrive at the goal have not emerged, e.g., deciding on a location for building a factory; and (v) *genuine decision tasks*, which are unexpected, new, and unstructured, e.g., a research project.

Li and Belkin [210] provide a faceted classification scheme for tasks. According to this scheme a research task such as engaged in by humanities researchers is classified as: (i) having originated and being conducted by the task doer; (ii) a long term task consisting of several stages; (iii) delivering an intellectual product; (iv) consisting of multiple and repeated processes; (v) and a concrete goal. Attributes of this type of task are that it is complex and has low dependence on collaboration. The influence of specific instantiations of particular facets of tasks have been investigated and found to influence the information seeking process. Aula et al. [21] found that when the difficulty of a task increases users tend to formulate more queries, use more advanced operators, and spend more time on search result pages. Wu et al. [368] had similar findings and related the difficulty of the task to its cognitive complexity. In Chapter 5 we simulate the task of exploring a new research topic by asking a particular group of humanities researchers to develop a research question based on the information found using an experimental search system. As we will see, users spent a great amount of time using the system using various interaction styles. This may not be a bad thing as the goal is to learn about and explore information available about a topic to arrive at a research question.

A task, such as writing a term paper, may also consist of multiple sessions. In such a task an information seeker engages in multiple information seeking tasks spread over several time periods, and possibly with various systems [81]. Liu and Belkin [215] mimicked a multi-session search task by asking subjects to complete a journalists' assignments, i.e., write a feature story on hybrid cars for a newspaper consisting of three sections. Two task types were used: a dependent condition where the information from one sub-task, i.e., writing one section, was related to information in the following sub-tasks, while in the independent condition the topics covered in the sub-tasks were independent. The relative complexity of the tasks was kept constant. During this type of multi-session tasks users' behavior changes. For example, depending on the stage of a task, the time taken to inspect a relevant document changes, e.g., in earlier stages useful documents are inspected longer than in later stages [215]. Others found that during the stages of a task users' judgement of document usefulness changes as well as their query modification patterns [197, 338]. In Chapter 4 we investigate how these changes in behavior during a multi-session search task affect user preferences for particular types of display.

Tasks do not exist independently of the environment and may originate from the environment or the environment may shape an information seekers' personal interests and ambitions to undertake a certain task. Marchionini [230] identifies two of these factors to be the domain and the setting. The domain determines the extent to which information in the environment is available and potentially relevant to the information seeker. The domain will differ per discipline, for example, for a historian information will often be limited to the primary source materials in archives, while in the data intensive disciplines, e.g., astronomy, information is available in the form of measurements. The setting determines the social limitations and pressures explicitly or implicitly put on the information seeker. For example, researchers tend to conduct more extensive and prolonged searches for information than the average web searcher. They are concerned with the novelty, and completeness of the information they find as it influences the quality of their work and the chance it gets published.

This has implications for the design of experiments to investigate the usefulness of interactive information retrieval systems for humanities researchers. Whether a system is used in a naturalistic setting, e.g., during a research project, or in a laboratory setting influences the validity of the study. We take this in consideration in the design of our experiments in Chapter 4 and 5, where we endeavor to assign subjects realistic tasks in natural settings.

### 2.4.2 Information Seekers

To facilitate experimentation with various system and interface features, studies adopt a static model of a prototypical user. Users of information systems, however, all differ in their capability, knowledge, and preferences, making a particular variant of a system more suitable for some than for others. For example, Duggan and Payne [124] found that users' knowledge about a topic influences their ability to search for answers to questions about that topic on the web. Users with greater knowledge of a topic were better able to find answers to questions on that topic, they spent less time inspecting documents, used shorter queries, and gave up a line of search faster. Similar effects of domain expertise on search effectiveness were found in a large scale log analysis [357]. Not just domain expertise but also expertise and experience in performing searches have been found to influence search performance [170]. Tabatabai and Shore [315] discovered that especially the use of cognitive and meta cognitive strategies, such as using clear criteria to evaluate sites, reflecting on strategies, and monitoring progress, determined search effectiveness. In the design of our experiments in Chapter 4 and 5 we take these characteristics into account and ask subjects about these factors.

Other characteristics are more difficult to elicite from users. For example, in an eye tracking study of users of a standard web search interface Aula et al. [20] found that searchers differ in the way they inspect result lists and identified two types of searchers: economic and exhaustive. Economic searchers spent less time to decide on the next action such as query reformulation or following a link after result presentation. As a result only some of the top most results are inspected, while exhaustive searchers inspect more results and scroll down the result page before their next action. In a study of intelligence analysts persistence was found to be a deciding factor that determined search performance. Analysts that read more documents and spent more time on the task out performed those who did not [266]. We discuss inherent differences in search strategies and persistence between users as possible confounding variables in Chapter 4.

User factors are further shaped by task context factors and may change over time. Kelly [189] identified five measures for the effect of task context on users: (i) task endurance, i.e., expected time to complete a task; (ii) persistence, i.e., time a user expects to remain interested in a task; (iii) progress of the task; (iv) frequency, i.e., number of times subjects expect to conduct information seeking activities related to a task; and (v) familiarity, i.e., current state of knowledge of a topic. In a naturalistic study following seven users during a 14-week period the self reported ratings of users for these measures were found to change from week to week, e.g., for familiarity as users became more familiar with a topic. Further, the frequency with which information seeking activities were conducted tended to cluster together in intense episodes. Liu et al. [216] investigated to what extent users' knowledge of a topic increased during a multi-session search task. Users' topic knowledge was generally found to increase although for some participants a ceiling effect was observed. In Chapter 4 we investigate the effects of subjects that work with the same topic during all sessions of a multi-session search task compared to subjects that change topic in each session.

### 2.4.3 The Search Interface

The familiar design of web search interfaces allows the user to input a keyword-based query via a single search box and presents results as a ranked list of snippets. This type of interface is optimized for looking up facts. A user's information seeking process, however, does not necessarily stop after submitting a query and inspecting some results. Depending on the characteristics and stage of a task a user engages in various other types of search. In this section we discuss various interface features designed to support alternative search tasks. Specifically, we discuss interfaces that support aggregated search, which provides the necessary background for our investigation of aggregated search display preferences in Chapter 4. This is followed by a discussion of interfaces that support exploratory search, which informs the design of our subjunctive exploratory search interface investigated in Chapter 5.

#### Aggregated Search Systems

Search engines use crawlers to discover and download documents on the web. Documents that are missed as they are not linked, behind a login, or contain dynamic content, are absent from search results [278]. The number of documents in this so called *hidden web* is estimated to be many times larger than the number of visible pages [52]. Instead of indexing the hidden web, federated search systems pass a users' query to the search service of each source individually. Techniques are focused on selecting the most promising results from each source and merging results in a single result list. This type of search is often seen in the context of digital libraries to enable searching across multiple catalogues [306].

Aggregated search originated as a new direction of web search as aggregated search systems aim to integrate results from various heterogeneous sources (called verticals), e.g., images, videos, and news collections, next to general web pages. It can be seen as an instance of federated search as studies investigate which vertical to present given a query and where to present it in the result list. Two general presentation styles for aggregated search results exist: tabbed and blended [203]. Tabbed interfaces provide access to each source in separate tabs, while blended interfaces combine multiple sources into a single result page, see Figure 1.1. Within the blended display style a distinction is made between presenting results from each vertical per group or by interleaving results. Seo et al. [301] investigated the use of click through data (clicks on vertical results) to optimize vertical presentation. The top 5 results for a query are retrieved from each vertical. The top 5 results are ranked and presented per group on the result page. It was noted that the limited number of results returned by each vertical introduced a bias in the click counts towards only the top ranked results for each vertical. Arguello et al. [14] studied approaches that learn how to mix results from different verticals in a single result page. The best approaches learned to interpret the relation between features and relevance for each vertical individually. For example, temporal features are effective for a news vertical but not for a question answering vertical.

Other work focuses on evaluating the quality of aggregated search result pages. A common approach is to obtain pair-wise judgements of blocks of results by human annotators and then evaluate pages based on whether the preferred blocks are ranked closer to the top of the page [15, 379]. Another possibility is to simulate result pages and to evaluate based on the number of relevant results visible on the screen of a device [220]. Santos et al. [296] noted the need for result diversification across verticals and extended existing diversification metrics to account for aggregated search.

Next to techniques for selecting verticals and merging results, however, increased attention is paid to the effects of presentation style on vertical use [203]. Previous studies in aggregated search have investigated whether users prefer tabbed or blended displays for single-session search tasks of varying complexity. In a study with sixteen participants Sushmita et al. [313] found that in complex tasks blended displays in which vertical results are presented per group are preferred over a tabbed display. Additionally, subjects preferred multimedia verticals over textual verticals. In a follow-up study a blended display with grouped results was compared with a blended display with an interleaved result presentation. Results showed that using vertical specific wording in tasks influenced click through rates on the mentioned vertical. With respect to complexity a later study confirmed that more verticals are clicked when task complexity increases, but that users do not necessarily prefer a blended or tabbed display [16]. As a possible explanation the search experience of the subjects is suggested. As we will see in Chapter 4 we find that the users actually switch between display types during a multi-session search task and investigate factors that influence this behavior.

### Exploratory Search Systems

When a user's information need is vague or a user does not know how to accurately describe his/her information need, the simplicity of the standard interface becomes a limitation. Instead, a user requires support for exploration to better understand his/her problem before engaging in more focused search activities [231, 356]. White and Roth [354] proposed a set of eight features that increase the support systems provide users in resolving exploratory information needs: (i) support querying and rapid query refinement; (ii) offer facets and metadata-based result filtering; (iii) leverage search context, e.g., query expansion; (iv) offer visualizations to support insight and decision making; (v) support learning and understanding; (vi) facilitate collaboration; (vii) offer search histories, workspaces, and progress updates; and (viii) support task management, e.g., store previous search tasks. We discuss a number of interfaces developed to support exploratory search below, see Hearst [164], Shiri [305], White and Roth [354] and Wilson and White [361] for more complete reviews of various types of exploratory search systems.

Relevance feedback techniques support querying and query refinement by using terms from initially retrieved documents to enhance a users query. Explicit relevance feedback techniques expect users to indicate the relevant terms themselves [294], while implicit relevance feedback techniques automatically select terms from documents examined in a previous search [191].

Result filtering based on facets provides users with the ability to quickly zoom in on a particular topic. Facets must be a set of meaningful labels that reflect the concepts relevant in a domain. One way of selecting which labels to use is to base facets on metadata. Flamenco is an example of a tool that provides hierarchical faceted browsing of museum collections [372]. Lin et al. [214] extracted named entities from a document collection to enable faceted filtering over entity types. Capra and Marchionini [86] provide an option to preview the result of clicking a facet by means of a mouse hover. Showing facets or automatically generated term suggestions from a set of result documents is considered to be helpful to give insight in the topics discussed in a set of documents as well as for naviational pruposes. Presentation as an orderly list, however, is preferred over a “cloud” visualization [166].

Golovchinsky and Pickens [148] investigated various visualizations to give the user insight in his/her search context. For example, showing the rank at which a document was returned for previous queries or when entering a query showing the overlap between the documents returned for the current query and previous queries.

Visualizations should allow users to quickly gain insight in the characteristics of the collection they are searching in and the results they obtain for a query. FeatureLens is an example of a dashboard like system that allows users to explore and analyze patterns in a collection of documents. It provides charts showing query term distribution across documents in the collection and the distribution per document as well as provides filters based on frequent patterns [122]. Another example of such an analysis tool is TAKMI, which additionally shows the distribution of concepts in result sets over time [253]. A problem with this type of analysis, however, is that knowledge is required about how the collection is created or obtained in order to interpret patterns, e.g., a peak in the occurrence of a topic may indicate increased attention to that topic or an artifact of the collection.

Other visualizations serve to provide insight in the occurrence of query terms in result documents. Hearst [163] uses a tilebar visualization to indicate for each document the presence or absence of query terms, their proximity, and the length of the document. The length of a bar indicates the document length, while grayscale scores indicate the occurrence of query terms. In a comparison of various visualizations techniques for query terms in retrieval results Reiterer et al. [279] found that users preferred tilebars over histogram visualizations and histograms over a scatterplot visualization. Visualizations are also helpful in navigating within documents. For example, Byrd [80] proposed a visualization that highlights the position of query terms in documents the position of query terms in a document.

Another way to support insight generation is to enable users in comparing alternatives. Subjunctive interfaces have been suggested for this purpose as this type of interface allows a user to perform multiple actions in parallel and compare the results, i.e., editing a document or searching a database. Typically multiple versions of a standard interface, e.g., a standard document editor, are presented side-by-side to create a subjunctive interface [223]. In a web search context Villa et al. [344] proposed a browser that presents multiple traditional web search interface displays side-by-side to allow users to explore more aspects of a topic than a single view variant.

The experimental interfaces to support humanities researchers described in Chapter 5 incorporate these elements of exploratory search systems, e.g., faceted filtering, and visualizations of result set statistics. In the aggregated search interfaces in Chapter 4 we also provide metadata-based filtering to supplement keyword search for various verticals.

To evaluate the benefits of these interfaces for the humanities we study their use by a particular group of humanities researchers, i.e., media studies researchers. In the next chapter we introduce the field of media studies and investigate their research practices in detail.

# 3

# Media Studies in the Data-Driven Age: How Research Questions Evolve

Having outlined the models of information behavior and the research practices in the humanities in general (Chapter 2), we now turn to a particular group of humanities researchers, i.e., media studies researchers. On the one hand, the focus on media studies researchers provides a relative homogeneous group of participants in our studies that share a common research background. This facilitates our experiments with interactive information retrieval systems in Chapter 4 and 5, as a more diverse group of humanities researchers would require larger samples of participants to account for the differences in background. On the other hand, media studies is a diverse field in the objects studied, technologies used, and methodologies applied. This makes that findings for this group are potentially relevant for other humanities disciplines as well.

To gain a better understanding of the research practices of media studies researchers and how these practices may be supported by interactive information retrieval systems we obtained interviews about the research projects of twenty-seven media studies researchers. This data is analyzed in light of our first research question as put forward in Chapter 1:

**RQ 1.** What does the research process of media studies researchers look like?

**a.** Can we identify sequences of activities in research projects of media studies researchers and how does the resulting model compare to other models of the humanities research cycle?

**b.** Do research questions of media studies researchers change during research projects, and can we identify factors that influence this change?

**c.** Which information needs and information gathering challenges do media studies researchers face during research projects?

The remainder of this chapter is organized as follows. Section 3.1 provides a brief introduction to media studies; Section 3.2 describes methods used for interviewing and analysis. Section 3.3 presents the results of our analysis, followed by a discussion in Section 3.4. We conclude in Section 3.5.

## 3.1 Introduction to Media Studies

The field of media studies, a field that is part of, and often synonymous to, cultural studies and communication studies, is difficult to delineate and has its roots in areas such as literature, sociology, psychology, economics, history, and journalism. The media are the research subject of the field of media studies and in this field media aspects such as production, reception, and/or content are studied. The media are present in many forms, through various channels, and have given rise to subareas such as film studies, technology studies, advertising and marketing, as well as more practical subjects such as video production, radio production, printing and journalism [255]. Various research practices that have emerged in the field of media studies to deal with the various media types, publishing technologies, access tools, and representation formats.

In media studies, the media are studied according to one or more of the following three aspects: *production*, *texts*, and *reception*. Production concerns the production of media, the industry, or the institutional context, e.g., a study of journalists or the Hollywood film industry [107]. Text concerns the content of media and includes oral, print, still, moving image, and computer-generated communications [255]. Reception concerns the "effect" of media on audience beliefs, attitudes, and behavior on the one hand, and the use and interpretation of media by audiences on the other [255]. However, the rise of digital communication channels has diluted the analytical boundaries between the production, text, and reception aspects, as audiences have become both the producers and receivers of media [320].

Among the methodological tools available to media studies researchers for the study of production, text, and reception, are content analysis, participatory observation, focused interviewing, and surveys. The first, content analysis, can also be referred to as "document analysis": the analysis of text materials to identify statements in the proper context for analysis [5]. Altheide and Schneider [6] identify three classes of documents. First are primary documents, which are the objects of study, this includes newspapers, magazines, TV newscasts, diaries, or archeological artifacts. Next are secondary documents, which are records about primary documents and other objects of research. This includes field notes, published reports about primary documents, and other accounts. Media studies researchers find context information particularly valuable for data selection and data interpretation [50]. Third are the catchalls, the auxiliary documents, which can supplement a research project or some other practical undertaking but are neither the main focus nor the primary source of data for understanding.

The practice of document analysis is not exclusive to media studies, but is broadly practiced throughout the humanities. Typical steps in the document analysis process are: developing an original idea about a topic, in step one, to gathering some ethnographic materials about a relevant or related setting, context, or culture in step two. The third step entails actually examining a few relevant documents with this awareness in mind and then, following step four to twelve, drafting a protocol for data collection, coding and analysis, and drafting the report [6]. Media studies researchers, and humanities researchers alike, often use a grounded theory approach [312]: through the analysis of data, they discover theory. The research practice then is a continuous going back and forth between analysis and theory. It contradicts the more conservative research paradigm, in which first a theoretical framework is designed and only then the analysis is started. This process can be very time consuming [264].

## 3.2 Method

To gain a better understanding of the current research practices in media studies, we conducted interviews with twenty-seven media studies researchers. In this section we first describe the methodology that we used to conduct our interviews with media studies researchers. We then describe the characteristics of the media studies researchers in our sample. Finally, we discuss the method used for analysis of the interviews.

### 3.2.1 Interviews

Interviews were conducted using the same methodological approach as described by Chu [91] and Brown [73] in their investigation of the research cycle of literary critics and music scholars. The methodology is a combination of the structured personal account [74] and the "time line interview" [118]. In the account interview participants are asked to describe a previously experienced event from a personal point of view. The time-line method is aimed at reconstructing each step taken in a specific situation with a focus on information gaps experienced and how these were resolved.

**Table 3.1:** Topic list used during open questions part of the interview.

| | |
|---|---|
| Q1 | Do you remember how the research project started? |
| Q2 | What were your research questions? |
| Q3 | Did you often search in media archives? |
| Q4 | Which archives did you use? |
| Q5 | What did you expect to find in these archives? |
| Q6 | What problems did you encounter? |
| Q7 | Next to media items, what other information did you search for? |
| Q8 | What additional information did you need that you did not manage to obtain? |
| Q9 | What tools did you use, e.g., search engines, websites, or analytical software? |
| Q10 | Did your research questions change during the research project? |
| Q11 | If you would divide your research project into stages which would you identify? |

The style of the interview was semi-structured and consisted of three parts: (i) identification of a recent research project; (ii) open questions about research activities and research questions during the project, see Table 3.1; and (iii) an interactive part in which participants wrote down the research activities on index cards and ordered them chronologically. Interviews lasted between thirty and forty-five minutes, were tape-recorded and later transcribed. The chronologically ordered cards representing the research cycle were numbered and photographed before being collected at the end of the interview. Interviews were conducted in Dutch or English depending on the nationality of the interviewee. Quotes extracted from Dutch interviews have been translated to English via Google Translate[1] and where necessary corrected to arrive at a proper translation. Note

[1] `http://google.translate.com`

that when using quotes, square brackets [. . . ] indicate modifications to the original quote to improve understanding or to protect the anonymity of the participant.

The interviews were conducted by two interviewers: a media studies researcher and a computer scientist. The first seven interviews were conducted jointly. Later interviews were conducted separately as the interviewers gained a shared understanding of the domain and interview style. A limitation of the interview method is that it is an account of how the researcher remembers a research project. It does not necessarily accurately describe how the project was carried out as parts may have been omitted or receive extra attention depending on the impact events made on the individual.

### 3.2.2 Sample

The participants were recruited based on availability. The investigators contacted colleagues to obtain an initial set of interview participants. Some of the participants suggested additional candidates who were contacted and invited to participate according to the snowball-method [181]. Finally, the investigators recruited several participants by contacting researchers at the media studies department of an institution during research visits. As the resulting selection is not a proper probability sample, it is difficult to generalize findings to the entire population. The aim of this study, however, is not to define the research practices of the entire population of media studies researchers. Rather, it explores their current information needs and habits.

Table 3.2 shows the demographics recorded for our participants. We briefly discuss each of the columns below. Some of the information has been anonymized to protect the identity of the interview participants. For identification purposes participants (P) are assigned a number, i.e., P01 to P27. Similarly, for institutions (I) we use I1 to I7. In total, researchers from seven different universities participated, four located in the Netherlands, the others in Israel, Denmark, and Belgium. Seven researchers are from institute I1, three from I2, nine from I3, one from I4, one from I5, five from I6, and one from I7.

Regarding the research discipline, more than half of the participants (16 out of 27) answered that they are part of the humanities, ten participants identified themselves as part of the social sciences, and one participant mentioned science and technology studies as discipline. Institutes generally have a media studies (med std) department (15 out of 27) or a communication science (com scs) department (7 out of 27), in some cases it is part of another department, i.e., political sciences (pol scs), or cultural studies (cul std).

Academic positions of participants include senior researchers, i.e., four full professors, two associate professors (associate pr), and nine assistant professors (assistant pr), as well as some junior researchers, i.e., four post docs, seven phd students (phd st), and one master student (mst st).

Regarding the preferred media studied by the participants, television (tv) is the dominant medium mentioned nineteen times, followed by newspapers mentioned eight times, and both radio and new media mentioned five times. Other media studied are film, documentaries, games, and music. The study of texts is practiced by most researchers (23 out of 27). Additionally, some study production and/or reception aspects of media. In terms of research methods, both qualitative (qual), quantitative (quan) and combinations of these techniques are uses.

**Table 3.2:** Participants' characteristics, see § 3.2.2 for abbreviations.

| P | I | Discipline | Department | Position | Media focus | Research focus | Method |
|---|---|---|---|---|---|---|---|
| 01 | I1 | soc scs | com scs | phd st | tv | reception | qual |
| 02 | I1 | soc scs | com scs | phd st | newspaper | production; text | both |
| 03 | I1 | soc scs | com scs | post doc | tv | production; text | qual |
| 04 | I1 | soc scs | com scs | assistant pr | tv | production; text; reception | qual |
| 05 | I1 | soc scs | pol scs | assistant pr | tv; newspaper | text | both |
| 06 | I1 | soc scs | com scs | assistant pr | games; music | reception | both |
| 07 | I1 | soc scs | com scs | professor | tv; newspaper; radio | production; text; reception | both |
| 08 | I2 | hum | cult std | phd st | new media | text | both |
| 09 | I2 | hum | cult std | post doc | radio | text | qual |
| 10 | I2 | hum | cult std | assistant pr | radio | text | both |
| 11 | I3 | hum | cult std | mst st | tv; new media | production; text; reception | qual |
| 12 | I3 | hum | med std | phd st | tv; film | text | qual |
| 13 | I3 | hum | med std | phd st | documentary | production | both |
| 14 | I3 | hum | med std | post doc | tv | text | qual |
| 15 | I3 | soc scs | med std | post doc | tv | reception | qual |
| 16 | I3 | hum | med std | assistant pr | tv; newspaper | text; reception | qual |
| 17 | I3 | hum | med std | assistant pr | tv; newspaper; radio | production; text | qual |
| 18 | I3 | hum | med std | associate pr | tv | text | qual |
| 19 | I3 | hum | med std | professor | tv | production; text | qual |
| 20 | I4 | hum | med std | assistant pr | tv | production; text | qual |
| 21 | I5 | hum | med std | professor | tv | production; text; reception | qual |
| 22 | I6 | soc scs | com scs | phd st | tv; newspaper | text | quan |
| 23 | I6 | hum | med std | phd st | tv | production; text; reception | qual |
| 24 | I6 | hum | med std | assistant pr | newspaper; new media | production; text; reception | qual |
| 25 | I6 | hum | med std | assistant pr | tv; new media | text | qual |
| 26 | I6 | sc & tch | med std | professor | new media | text | both |
| 27 | I7 | soc scs | med std | associate pr | radio; tv; newspaper | production; text; reception | qual |

### 3.2.3 Analysis

The analysis started with reviewing the research activities written on index cards. Cards with activities that exactly match were used to align the research activity sequences of each of the participants. Next, we adopted an open coding strategy [312] and grouped cards with similar descriptions in the interview either mentioned during the description of the stages (Q11) or during the creation of the index cards. Activities that did not match were placed in a separate category. During a number of iterations categories were renamed and merged to arrive at the set of codes listed in Table 3.3.

To test the reliability of the coding scheme the descriptions of the activities were coded

**Table 3.3:** Overview of the codes used to annotate the activities in the research cycle mentioned during the interviews.

| Code (abbreviation) | Description |
|---|---|
| initial idea (ii) | An idea, observation, or proposal that starts a project. |
| background study (bg) | Identify literature and background material for a topic. |
| initial research questions (ir) | Identify research question or instrument, e.g., sampling. |
| initial data gathering (ig) | Initial search, exploration, or collection of data. |
| revised research questions (rr) | Revision of research questions and instruments. |
| targeted data gathering (tg) | Collect, search, or select data following guidelines. |
| analysis (an) | Inspect, read, code, compare, or organize data. |
| write (wr) | Write, select examples, drawing of conclusions. |
| report (rp) | Integrate findings into articles, chapter, or presentation. |

by an additional investigator. Digital excerpts were created from the index cards and the coding scheme was explained. Excerpts were presented one by one, in random order, and without context from the interview. The results are presented in Table 3.4. Cohen's kappa is a measure for inter-annotator agreement and generally the following rule of thumb is used for interpretation: $<.20$ is poor agreement, .21–.4 is fair agreement, .41–.6 is moderate agreement, .61–.8 is good agreement, and .81–1.0 is very good agreement [204]. The pooled kappa [115] of the agreement between the two annotators over all individual codes is .53 indicating moderate agreement.

Closer inspection of the agreement on individual codes reveals that the *initial data collection* and *targeted data collection* codes have low agreement. The main difficulty turned out to be in determining whether a data collection activity was initial or targeted without the context of the interview or other index cards. To resolve this issue, the two annotators discussed all excerpts where a disagreement existed using the interview and index cards as context to arrive at a 100% agreement on the code assignment.

The variety and richness of the issues that participants touch upon in the remaining interview questions (Q1–Q10) are not easily captured by a specific set of codes. Where appropriate we counted how many times items were mentioned in the answers or provide broad observations and illustrate each with quotes.

**Table 3.4:** Cohen's kappa coefficient indicating inter-annotator agreement for individual codes and pooled kappa.

| Code | $\kappa$ | Code | $\kappa$ |
|---|---|---|---|
| initial idea | 0.49 | background study | 0.75 |
| initial research questions | 0.54 | initial data collection | 0.17 |
| revised research questions | 0.46 | targeted data collection | 0.34 |
| analysis | 0.64 | write | 0.56 |
| report | 0.75 | pooled kappa | 0.53 |

## 3.3 Results

In this section we first discuss answers to interview questions related to research activities and their sequences in the research cycle, i.e., Q11 and the third part of the interview. Then, we discuss questions related to changes in the research questions (Q2 and Q10). Finally, we discuss questions related to information needs and data gathering challenges (Q3–Q9). Although the following sections address different questions and illustrate different points, some overlap exists within participants' answers to the interview questions.

### 3.3.1 Research Cycle

Table 3.5 shows the sequences of the research activities described on index cards and mentioned during the interviews for each of the participants. Columns represent participants, rows are labeled with codes described in Table 3.3. Each block represents an iteration of the research cycle and each row represents an activity. A "*" indicates that a certain activity was identified for a participant. A new iteration (block) is started whenever an earlier activity is repeated. The sequences have been ordered with on the left the research cycles with the most iterations and on the right those with the smallest number of iterations.

#### Initial Exploration of a Topic

In the first iteration, initial idea, background study, developing the initial research questions, and initial information gathering are the top four most frequently identified activities. Nine participants explicitly indicated that a research project starts with an initial idea or observation that sets the direction for the research topic. For example, participant P21 stated: "The first phase is the conceptual phase, thinking about what is your question. Why is that interesting. And the development of a concept, or a research plan how you could study that." According to P11: "The first decision making phase was to pick this phenomenon to write a paper about and not something else. Then identifying what literature relates to it."

Others indicated that research starts with studying the literature (P12, P25, P06, P18, P02, P27, P14) or an initial look at the data (P03, P17, P20, P19, P10, P04). Participants P17 and P20 indicated that initial information gathering comes first: "the first phase was to see what was there" (P17) and "start with the written archives to get the more general picture" (P20). Alternatively, participants P01 and P02 stated that literature is the starting point: "Exactly according to the scientific research process. So first the question, then the literature review ..." (P01), and "I spent most of the first year reading and defining the theme" (P02).

In general, the comments regarding the activities in the first iteration of the research cycle suggest the need for exploration to gain an overview of the data, topic, and literature. Participant P22 remarked: "first you must have a subject and know that it is interesting and has not been done before. But I never start by thoroughly figuring out a theoretical framework, which is actually the official procedure, [...] pretty quickly I go and see if the material is available." Another stated: "Always first explorative. A little

**Table 3.5:** Overview of the research activities, for abbreviations see Table 3.3.

| | 05 | 12 | 23 | 26 | 09 | 07 | 16 | 21 | 15 | 03 | 11 | 24 | 25 | 06 | 18 | 22 | 01 | 20 | 19 | 27 | 17 | 14 | 08 | 10 | 13 | 02 | 04 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ii | * | | | | * | * | * | * | | | * | | | | | * | | | | | | | * | | * | | | 9 |
| bg | * | * | | | | * | | | | | * | | * | * | * | | | | | * | | * | * | | * | * | | **12** |
| ir | * | * | * | * | * | | * | * | * | | | | | | * | | * | | | | | | | | * | * | | **12** |
| ig | | | * | | | | | | | * | | | | | | * | | * | * | | * | * | | * | | | * | 9 |
| rr | | * | | * | | * | | | | | | | | | | * | | | | | | | | | | * | | 5 |
| tg | | | | | | | | | | | | * | | | | | | * | | * | | | * | | | * | | 5 |
| an | | | | | * | | | | | | | * | | * | | | | | | * | * | * | | * | * | * | * | **10** |
| wr | | | | | | | | | | | | | | | * | | | | | | | | | | * | | * | 3 |
| rp | | | | | | | | | | | | | | | * | | | | | | | | | | * | * | | 3 |
| ii | | | | | | | | | | | | | * | | | | | | | | | | | | | | | 1 |
| bg | * | * | * | | | | * | * | * | | * | | * | | | * | * | | * | * | | * | * | | | | | **14** |
| ir | | | | | | | | | | * | * | | | | | | | | | | | | | | | | | 2 |
| ig | | | | * | | * | | | | | | | * | | | | * | | | | | | | | | | | 4 |
| rr | * | | | | | * | | | | | | * | | | | | | * | * | | | | | | | | | 5 |
| tg | | | | | * | * | * | * | * | * | | | | * | * | | * | * | * | | * | | | | | | | **12** |
| an | | | | * | * | * | * | * | * | * | | | * | | * | * | * | * | * | | * | | | * | | | | **15** |
| wr | | | | | | | | | | | | * | | | * | | | * | * | * | * | * | * | | | | | 8 |
| rp | | * | * | * | | | | * | * | | | | * | | * | * | * | | | | | | | | | | | 9 |
| bg | | | | | | * | | | * | | * | | * | | | | | | | | | | | | | | | 4 |
| ig | * | * | | | | | | | | | * | | | | | | | | | | | | | | | | | 3 |
| rr | | | | | | | * | | | * | | | | | | | | | | | | | | | | | | 2 |
| tg | | | * | | | | | | | * | | * | * | * | | | | | | | | | | | | | | 5 |
| an | | * | | | * | * | | | | * | * | * | | | | | | | | | | | | | | | | 6 |
| wr | | | | * | | | * | * | | * | * | | | | | | | | | | | | | | | | | 5 |
| rp | | * | | * | | | * | * | * | | | | | | | | | | | | | | | | | | | 5 |
| bg | * | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| rr | * | | * | | | | | | | | | | | | | | | | | | | | | | | | | 2 |
| tg | * | * | | | | * | | | | | | | | | | | | | | | | | | | | | | 3 |
| an | * | * | | | * | | | | | | | | | | | | | | | | | | | | | | | 3 |
| wr | | * | | * | * | | | | | | | | | | | | | | | | | | | | | | | 3 |
| rp | | | | * | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| bg | * | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |
| tg | | * | * | | | | | | | | | | | | | | | | | | | | | | | | | 2 |
| wr | * | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 |

bit of browsing, everywhere. Then, scoping not really. Because you will keep your eyes open for things you may discover. To not limit yourself in the beginning" (P19).

Additionally, some interviewees noted that the goal is to arrive at an initial research question through these interactions with data and literature. One researcher (P16) noted: "for me it starts with developing the research questions and data collection. This happens in parallel, so the question changes by the material you see" and another (P07) noted: "the data influences the research question, because the data is not available or because you start to see, oh this is so naïve."

Three participants' research cycles end without repeating any research activities (P02, P04, and P13). These researchers may have left out some of details of their research activities. As P04 noted: "the thing is with qualitative research that these phases are not so easily broken up into parts."

### Targeted Information Gathering and Analysis

In the second iteration, we observe that again the background study activity is frequently identified (by 14 participants). Seven researchers explicitly mentioned studying background material a second time during their research. As interviewee P27 noted: "Literature starts before everything and it comes again at time of writing."

The emphasis in this iteration, however, is on targeted information gathering and analysis. Researchers engage in targeted information gathering activities based on experiences from earlier explorations: "If you have consulted more sources, then you always get more focus. You get more of a story line. Otherwise, it is anecdotal: this program says that, and that program says this. No, it should mean something together. You only realize that in second instance. Only when you have seen material, you get an idea of what goes together, like this may well be related to that. And that is what you will investigate further" (P19).

This iteration ends the research cycle for ten additional researchers. Five described a sequence of targeted data gathering, analysis, and writing and/or reporting in the (P01, P17, P18, P19, P20), while five others mentioned either a targeted or initial information gathering activity in the first iteration of the research cycle before finishing in the second iteration through activities of analysis, writing and/or reporting (P10, P08, P14, P22, P27). For example, according to interviewee P01: "Yes, the [selection method] was already decided upon before the students started, the method actually came from the theory [. . . elaborating on theory] Then there is data collection, analysis, and writing" and P17 noted "The first phase was to see what was there. The second phase I only dealt with the programme guides. So I have sampled from a few weeks of four or five years and then analyzed."

Fourteen researchers engaged in one or more additional iterations of the research cycle. Eight researchers (P03, P05, P06, P07, P12, P23, P24, P25) mentioned the need for additional data collection and provided various reasons for repeating this step: getting a representative sample (P03: "so that [first analysis] was followed by collecting new data, using the methods identified in the literature and guided by insights from the earlier analysis"), being overwhelmed by the amount of information (P23: "So here I had the most stress, I got lost in it. Then I made my research question more specific. And defined case studies. So making choices in systematically searching the archive"), and lack of suitable material (P07: "and then you return again to the data and sometimes the literature, while part of the data has already been collected, because you feel that something is there but it does not come out"). The other six participants instead focused on studying additional literature, analyzing data, and writing.

After the third iteration eight researchers did not engage in additional activities and work with the material they have: "by organizing [the material] you create the story, I chose to use a chronological ordering, if I had organized my archive differently I would have written a different story. I could have organized it in supporters and opponents"

(P16). In this case a researcher chose a certain view on the data and organized it accordingly. Although multiple lines of inquiry were possible only one is explored. Another noted: “Otherwise, you can not [find] those sources when you are in an archive and your time is precious, you can not sit there for five months. No you go there for two weeks and then again two weeks and then you should know exactly what you are looking for. You also have to interpret the material on the spot and be able to say this is important and that is less important” (P21). In this case, time was a limiting factor in the research cycle and prevented the collection of new material. To limit the amount of time it takes to collect data some researchers reuse data, for example, P02 noted: “However, interestingly, if you continue to use the same data than you do not lose a lot of time with collecting.” Others ask for help in getting access to data as P21 described: “Because the archive service did not seem to co-operate. So then you ask around to journalists who have easier access to that archive. [. . . ] That is all a big hassle. That is just not fun. So that is why this project is also nice to take. I have spent a lot of time on it and achieved few results.” These quotes illustrate one possible reason for researchers not to continue with additional information gathering and analysis activities in the fourth and fifth iteration, i.e., the time and effort involved in data collection and analysis.

### Writing and Reporting

When one or more iterations of data collection and analysis activities have finished researchers engaged in writing and reporting activities. During writing the results of the analysis are interpreted, suitable examples are found, and conclusions are drawn. This is a creative process that requires integrating original data, the results of analysis, literature and background material. For example, a researcher (P05) noted: “Then write out the data analysis. Then the conclusions. But again there are a lot of things in between . . . Yes, here you go again back to the literature, certainly after the data analysis and during that analysis. Actually those [cards] should also be put somewhat on top of each other because these things often overlap.” Another described writing as: “And then the third phase is analysis and linking, you will link all kinds of information together and conduct your analysis on that. Until you come to some observations. That connecting of all kinds of material is important and then you describe your vision on it” (P19).

Seven researchers mentioned that reporting occurred during and not at the end of the project. Giving presentations about preliminary work is a common practice in the humanities to obtain feedback from peers [73, 91], for example as P18 mentioned: “Explore themes, formulate research questions, then what in my case structured my thinking, is writing a paper and presenting. That way you find out if it is worth it and that way you can also find people who share that idea. Is it worth thinking about that theme. Then the collection of resources . . . ”

Some of the research cycles do not end with writing or reporting activities (P06, P07, P10, P12, P23, P24, P25). Although care was taken to pick a project that was finished some researchers did not have such a project, or described a project that was part of a larger project where the results served as input to another investigation.

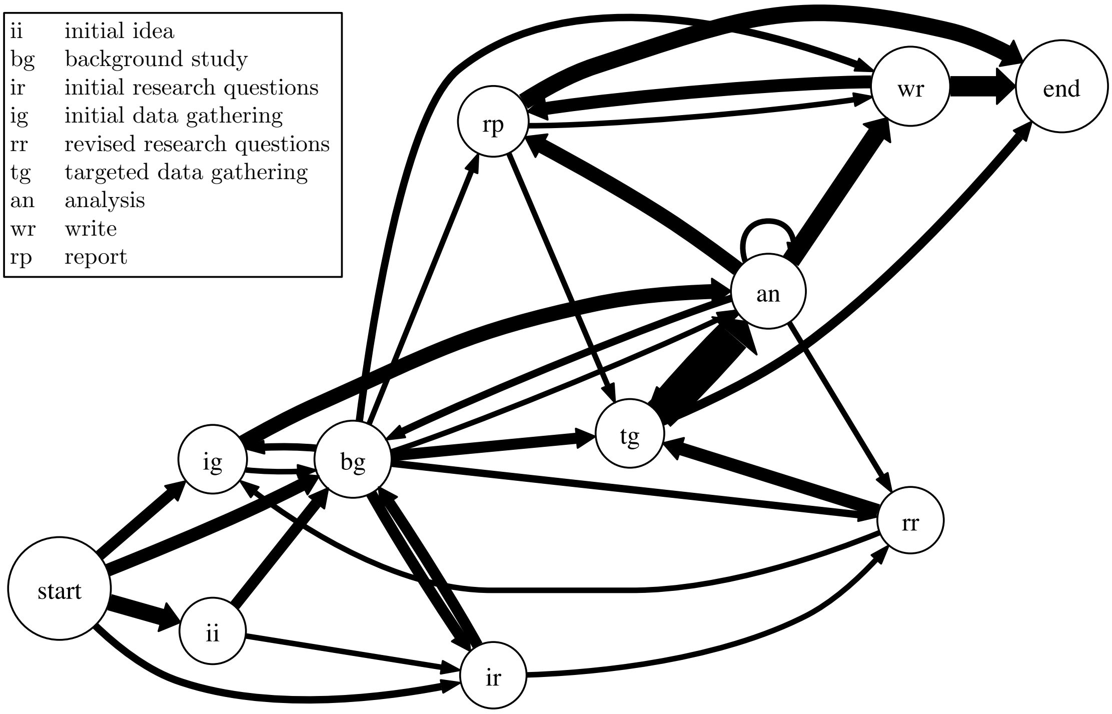

ii initial idea
bg background study
ir initial research questions
ig initial data gathering
rr revised research questions
tg targeted data gathering
an analysis
wr write
rp report
rp
wr
end
an
tg
ig
bg
rr
start
ii
ir

**Figure 3.1:** Transitions between activities in the media studies research cycle. Edge thickness indicates frequency. Transitions occurring once or twice have been removed for clarity.

### Summary of the Research Cycle

Figure 3.1 shows the one step transitions between activities in the research cycle of media studies researchers. We observe that at the start of a research project media studies researchers transition between studying background material, developing initial research questions, and initial information gathering. We identify this as the *exploration phase* in which the initial idea becomes more focused as media studies researchers become more familiar with the topic and the material. The goal of this phase is to get an overview of the topic and to formulate an initial research question.

In the next phase a more focused data collection starts. Initially gathered material is supplemented with material to place it into context or a theoretically motivated data selection is made. We define this as the *contextualization phase*. Three paths lead from the exploration phase to activities in the contextualization phase. The first path leads from initial information gathering, studying background material, or initial research questions to revised research questions. The second path leads directly to analysis activities and the third path leads from studying background activities to targeted information gathering. In most cases targeted information gathering leads (18/20) to analysis of the material.

After the analysis we observe some additional transitions to information gathering (4/32) and studying background material (4/32) activities. However, at some point the data is fixed and the next phase starts. At this point a relevant sample of the data has been collected and this data is interpreted in the context of focused research questions. This phase consists of interpreting and writing. The media studies researcher builds up a case to support his/her research questions by organizing the data and selecting appropriate qualitative evidence. We refer to this phase as the *presentation phase*.

### 3.3.2 Changes in Participants' Research Questions

In this section we discuss how the research questions of media studies researchers change during the research process. Table 3.6 shows for each participant (P) whether: (i) their research question changed (Ch) indicated by yes (Y) or no (N); (ii) how it changed (How), i.e., became more specific (S), additional question added (A), or perspective changed (CP); (iii) during which activity it changed (When), i.e., during analysis (ana) or an information gathering activity (gat); and (iv) why it changed (Why). Five participants indicated that they did not change their research question during their research project. Participant P04, for instance, started with a specific question and did not change it because he was able to collect all television series he wanted. P06 and P17 said they did not change their research questions, but added that they "slightly refined" (P06) or "focused

**Table 3.6:** Changes in research questions, for abbreviations see §3.3.2.

| P | Ch | How | When | Why |
|---|---|---|---|---|
| 01 | N | | | |
| 04 | N | | | |
| 06 | N | | | RQ in essence the same, slightly refined |
| 12 | N | | | |
| 17 | N | | | a bit more focused |
| 02 | Y | S | ana | RQ became more specific, since s/he discovered themes in the material |
| 07 | Y | S | gat | RQ changed due to availability and type of material |
| 09 | Y | S | gat | became more realistic: it was too much, narrowed down the time period |
| 10 | Y | S | ana; gat | RQ became more specific since s/he wanted a better focus and realized not all the material available |
| 15 | Y | S | gat | limited data collection for pragmatic reasons |
| 16 | Y | S | ana | watched material and got a more specific idea |
| 19 | Y | S | ana | RQ became more specific, since s/he had seen more documents |
| 22 | Y | S | gat | RQ changed since s/he did not have access to all material |
| 23 | Y | S | ana | RQ became multilayered, two RQs derived from the initial one |
| 03 | Y | A | ana | additional RQ about production and start new data collection |
| 05 | Y | A | ana | found that s/he could not properly answer the RQ and had to search again for specific material and enlarge his corpus of programmes. |
| 08 | Y | A | ana | studied additional literature and added a theoretical RQ |
| 13 | Y | A | ana | discovered that s/he should analyse a larger variety of cases |
| 18 | Y | A | ana | discovered shows lack of popularity in some countries, investigated why |
| 25 | Y | A | ana | found another aspect s/he could focus on |
| 27 | Y | A | ana | initial research turned out to be too limited |
| 11 | Y | C | ana | noticed that technology is an important factor |
| 14 | Y | C | gat | RQ changed since s/he could not access to the programmes s/he wanted |
| 20 | Y | C | ana | RQ was based on an assumption and should be rethought completely |
| 24 | Y | C | ana | noticed that some people tweet in multiple languages, investigated why |
| 21 | Y | S; C | ana | discovered during analysis that his initial RQ could not be answered |
| 26 | Y | S; A | ana | exploratory research, some RQs are dropped others made more focused |

a little bit maybe" (P17). P17 explains that "the programme guides and manuscripts provided everything [I] wanted."

The other twenty-two participants did change their research question. Our interview data shows that the research question changed in three ways: (1) by moving from broad to specific, (2) by adding other research questions, and (3) by changing the perspective of the research question. For twenty participants the research question only changed in one of these three ways. For two participants, P21 and P26, the research question changed in two ways.

### More Specific Research Questions

The research question turned more specific (i.e., got more focus), during the projects of eleven participants. In all cases the research question was refined during analysis of the material and/or during the data collection activity. Seven participants got a more focused research question during the analysis of material. P19 explains it as follows: "You are going to focus, to sharpen your research question if you know more about the programmes and more about the context. The more material you have analysed, the more focus you get." A participant (P02) who made his research question more specific during analysis of material, also referred to the literature study activity. He focused on four themes in the newspapers, since he "found that there were re-occurring themes in the literature. And in the journalistic debates [in the newspapers]." He adds: "In fact, the research questions got more focused along the way, but of course it is also an interactive process [between analysis and literature study]. I started out very broad." During data collection, the research question of six participants got more specific. P22, for instance, had to focus his research question as it turned out material was not available: "I had to change the research questions, [. . . ] because I could not obtain material of two [type A] broadcasters, and therefore could not compare [type A] and [type B] broadcasters."

### Additional Research Questions

In eight cases participants decided to add additional research questions. During analysis some participants discovered additional aspects of their research topic and added a research question to account for this. For example, P25 noted: "It is related to my original research question, but it is in a different direction, because I can see while analyzing material, ah, there is another aspect," while another participant answered: "Yes, in the beginning I was interested in how [object of study] imagines its audience. And towards the way, I found that there were things that were interesting that are not related to this. For instance, location" (P27).

Other participants added additional questions as their original research question turned out to be too limited and did not cover the trends discovered in the data sample. As P18 mentioned: "Then you discover that it is a format that is also produced elsewhere and whether it is popular over there, or not. It appears that there is a big difference between northern and southern countries" and P05 stated: "at one time I decided to add a qualitative part because I found that on the basis of the broadcasts that I had analyzed that I did not have all . . . That I was not yet able to fully answer the research question." Some participants mentioned that the patterns they expected to find in the material were not present: “I expected [medium A] to be more involved with [medium B]. Because [medium A] was the mass medium, the leading medium in the [...]s. I thought, in [medium A], they probably intensively discussed [medium B], but no” (P21).

### Research Questions with Changed Perspective

Five participants changed the perspective of their research question. According to P11: “when I was doing [analysis] it became clear that the availability of the shows depends on the technology. The paper ended up talking more about the technology. Because of the way [medium] criticism has moved in the way it works” and P20: “it was when I started having interviews that I realized that it is not so black and white and that is when the direction of my research changed completely.” Similarly, P14 had to change the perspective of her research question from “[perspective A]” to “[perspective B]” because she “did not get access to the archives. [...] So I investigated what I could get access to”.

In summary, the analysis of Q2 and Q10 suggests that research questions often change during the media studies research cycle, i.e., questions become more specific, additional questions are added, or the perspective of questions changes. Changes are related to activities of information gathering and analysis. During these activities participants learn about new aspects related to their research topic and gain insight in the availability and trends in the material covering their topic. Responses indicate that several iterations of information gathering and analysis activities alternate before the final research questions take shape.

## 3.3.3 Information Needs and Challenges

In this section we describe which information needs and information gathering challenges media studies researchers encountered during their research projects. First, we focus on the primary source materials that were the focus of the research projects. Then, we discuss what kind of additional information media studies researchers gathered. Finally, we elaborate on the challenges of gathering the initial and additional material.

### Information Needs

In the exploration phase, media studies researchers started with broad collections, in order to select interesting cases. Eleven participants went to physical archives. Nine participants collected material by buying newspapers/magazines/DVDs/games, or gathering online material such as websites and blogs. Two participants taped television programs when they were broadcast. Five interviewees combined archival material with online material and/or by taping recordings.

A large proportion of the participants (13) started by collecting audiovisual material: television programs, commercials, and documentaries. Others started with audio material, i.e., radio broadcasts (4). Six participants collected print material such as program guides, institutional material, and newspapers. Three out of six did this in conjunction with collection of television broadcasts. The other three had print material as their main collection. Five participants focused on new media collections such as Twitter feeds, websites, blogs, and games. Two out of five collected new media in conjunction with television broadcasts. Three participants also indicated that they conducted interviews as part of the initial data collection (i.e., next to collection of audiovisual, printed and/or new media material).

In contrast to data collected in the exploration phase, data collected in the contextualization phase is more specific, i.e., ratings/data on popularity (2), critical reviews (6), debates (3), blogs (4), online fora (2), letters of viewers (1), and biographies (4). The first five types of material are all collected to add a reception study on how television programs, radio broadcasts or films are received in the press and by the audience. The last category, biographies, is collected to dig into the background of producers, journalists, cast members and people mentioned in newspapers and news broadcasts. Lastly, interviews with producers are also often mentioned: nine participants indicated that they conducted interviews.[2] P03 explains why interviews are an important source of information: "For the production context in general, one depends of information on the internet, or the website of the production company or broadcaster. In general it remains very superficial. Therefore, we had to do interviews."

Interestingly, the information gathered in the contextualization phase is literally referred to as "contextual information" by eight participants. P21 describes contextual information as "essential" and explains it as "metadata in the language of archivists." Participants often named the collections they used in an attempt to find information ("program schedules in program guides," "interviews in newspapers," "reviews in newspapers," "biographies on Wikipedia"). On the one hand, they find it online, i.e., on Wikipedia, websites of broadcasters, and online newspapers. On the other hand, they look in paper archives for newspapers, magazines and program guides. P13, for instance, studied international newspapers for reflections on the audiovisual material she studied: "Yes, I also used newspaper articles, namely [newspaper 1] and [newspaper 2], [...] to search for a reflection on what happened with [main media subject]."

Newspapers are not only valuable in that they provide reviews, reflections and production information, but also to contextualize per se. Five participants used newspapers to get a better understanding of the political and social context of the media they studied. For instance, P24 mentioned that he "looked for international newspaper articles about [media subject] to contextualize it." (Historical) books are also mentioned in five interviews as useful contextual information. P14 explains the use value as follows: "journalists and directors do not come up themselves with ideas. They often get inspired by what happens in society. This is something you can find in historical books."

### Challenges

Participants reported on a variety of challenges that they encountered during the initial and targeted data collection activities. Participants did not mention a specific activity during this interview question (Q6), hence we discuss the general challenges in information gathering here. Participants mention up to three problems each. The problems can be divided in five categories: (i) availability of material; (ii) archival search system;

[2] Note that researchers collected multiple types of materials and the number of types exceeds the number of participants.

(iii) archival cataloguing; (iv) technological challenges outside the archives; and (v) institutional challenges.

A lack of availability of research material was mentioned eighteen times. In eleven cases (out of eighteen) the lack of availability referred to material that was not preserved. Six participants indicated that the television programmes, radio broadcasts and commercials that they needed were not preserved. P04 explains how it affected his research: "I got access to a lot of material that was not publicly available, [. . . ] but still there was some material that was not preserved and, therefore, not available for my research." Five participants said that it was difficult to find production information as this material is not preserved by archives. P01 considers "especially financial and budgetary information of television programs is difficult to obtain, but nevertheless important for research." She also suggests that "producers are not very likely to share this sensitive information with you." In four cases (out of eighteen), it was regarded as a problem that material was not digitized. P21, for instance, accounted for a situation in [country]: "Only a little is digitized. You often have to pay for viewing material. So it costs a lot. That makes it very difficult to study audiovisual material [in that country]." In the other three cases (out of eighteen), the needed material was available but in bad condition. For instance, P10 really had to rely on the recordings as he had to hear what is being said: "Quality various quite a lot [. . . ] Especially, those really bad recordings of the 19[. . . ]s." P20 gives as example that "it often happened that the film broke during my viewing [of audiovisual material]."

A second category of problems is related to the search system of the archives, a problem that was identified six times. Three participants mentioned that they were not allowed to use the search system themselves, and had to work with an archivist. For instance, P20 mentioned that "initially I did not have access to search myself, so I would tell them a big keyword or a specific title of a program, a long running program for instance and then they would give me pages with the reference number of the programs without description and then I had to do my research based on that." Others noted that the search system was "not good" (P04, P13) or "non-existing" (P17).

Third, a bad archival cataloguing system was mentioned as a problem by three researchers. According to participant P16, "the programs were not well described in the archive and difficult to retrieve." The same P16 also said that "there were missing metadata fields of radio programs." P24 "missed information on the page numbers of newspapers." Fourth, the participants who did not obtain their research material (solely) from an archive, also reported technological problems, such as that "it was difficult to scrape everything" (P25) or "difficult to tape multiple programs at the same time" (P03, P05).

Lastly, institutional challenges, related to the institute of the archive or the legislation of copyright owners and broadcasting companies were mentioned nine times. Five participants could not get permission to access the archive. One participant got no permission from webplatforms to access the data. Three participants said that it was too expensive to obtain material from the archives and, therefore, they all three decided to collect television programs themselves by recording them. Two other researchers had to deal with geographically dispersed archives. P17, for instance, said that he had to go both to the central archive and to all regional archives in every state in order to obtain all material, "which cost time, effort and money." According to two researchers, archives were also slow in releasing material. P09, for instance, says that "it took seven months to obtain all the material."

In summary, we found that next to traditional objects of study such as monographs, media studies researchers are turning towards new sources provided by the Web, e.g., fora, for primary source material as well as contextual material. Additionally, we found that gathering primary and contextual material requires search across various sources, i.e., newspaper services, archives, and the Web, and that each provides its own challenges in terms of accessibility and discoverability of the material.

## 3.4 Discussion

In this section we discuss our answers to our first research question from Chapter 1, where we asked:

**RQ 1.** What does the research process of media studies researchers look like?

**a.** Can we identify sequences of activities in research projects of media studies researchers and how does the resulting model compare to other models of the humanities research cycle?

**b.** Do research questions of media studies researchers change during research projects, and can we identify factors that influence this change?

**c.** Which information needs and information gathering challenges do media studies researchers face during research projects?

To address **a** we identified several activities within the media studies research cycle, as detailed in §3.3.1 and found that it is an iterative process, where activities such as literature study, data collection, and refinement of the research question alternate. A model with three phases emerged, i.e., the exploration phase, the contextualization phase, and the presentation phase, as during a research project media studies researchers transition from one set of activities to the next.

The model shares similarities with models of the research cycle of other humanities researchers (cf., §2.3.2). Most closely related to our work are models of literary critics and music scholars. The preparation stage of literary critics corresponds to our exploration phase. The elaboration, and the analysis and writing stages correspond to our contextualization phase. While the dissemination, and further dissemination and writing stages correspond to our presentation phase [91]. The additional *preparation and organization* stage in the model of music scholars' research cycle is characterized by information gathering activities such as interviews and participant observation [73]. We observed similar activities in the exploration phase of our model. These models describe similar stages and activities as our model of the media studies research cycle, however, they do not describe how these stages influence the research questions during the research cycle. The value of our model is that it makes the sequences of activities and the gradual refinement of the research questions in the media studies research cycle explicit.

The phases identified in our model of the media studies research cycle are consistent with the stages of models of the information seeking process (cf., §2.2.1). Kuhlthau

[197]'s ISP model describes the following six stages: initiation; recognition of a need; selection and identification of a topic; exploration of relevant information; formulation of a focused topic; collection of relevant information; and presentation of search results. Here, the first four stages fit our exploration stage, the next two fit our contextualization phase, and the final stage fits our presentation phase. Vakkari [338] used three stages in the *task performance model*, i.e., pre-focus, formulation, post-focus. An important difference is that these models focus on a single information seeking process which corresponds to a single activity in our model, e.g., a specific instance of the targeted information gathering activity. Moreover, our model considers these activities in the context of other activities, e.g., information gathering, analysis, or background study.

Strategies such as browsing and differentiating [41, 128] are observed during the activities of the media studies' research cycle. Whether researchers utilize these strategies depends on the stage and phase of the research cycle they are in.

Models of information behavior proposed by Wilson [364], Byström and Hansen [81], or Ingwersen and Järvelin [175], provide a general framework to describe users' information behavior in context (cf., §2.2). We have not considered all possible variables indicated by these models as these are too broad. Instead, we focused on information seeking behavior (information gathering and use) and the change of the information need in the research cycle.

With the second part of our research question (**b**) we investigated whether research questions of media studies researchers change during research projects, and whether we can identify factors that influence this change. We found that during the media studies research cycle the research questions become more focused as media studies researchers become increasingly familiar with the material and the topic under study. Additionally, we found that two of the main activities in the media studies research cycle are responsible for changes in the research questions of media studies researchers: information gathering and analysis. During information gathering media studies researchers discovered the extent to which accounts in primary source materials cover their research topic and whether this material was available. While during analysis media studies researchers gained insight in trends in the data and the existence of alternative views on their research topic.

A consequence of changing or adding an additional research question is that additional activities of data collection and analysis are required. Figure 3.1 shows this in the connections between the analysis, targeted information gathering, and revised research question activities. Not all participants explicitly mentioned the iterations between these three steps in response to the interview questions about the activities during the research cycle, e.g., mentioning only going back and forth between information gathering and analysis or between several analysis activities. Consequently, the edges in Figure 3.1 are not all equally strong. The analysis of Q10, however, indicates a strong connection between information gathering and analysis activities and changes in the research question.

Finally, with respect to **c** we found that media studies researchers have various information needs and that the specificity with which media studies researchers are able to characterize their information need and the type of material sought for depends on the phase in the research cycle. In the exploration phase media studies researchers engage in an activity of a broad gathering of a single type of primary material, e.g., television. As the media studies researcher becomes familiar with the material available on a topic the work becomes more focused. In the contextualization phase, theoretically informed selection criteria are used to select or gather the media type that is the focus of the research. In contrast to the exploration stage, however, additional types of primary and secondary materials are consulted, e.g., newspapers, to provide context for the interpretation of the primary source material focused on in the research.

Earlier work identified how the information needs of humanities researchers become more focused as they move through stages of the humanities research cycle and the need for context information arises (cf., §2.3.2). Duff and Johnson [123] discovered that historians need to orientate themselves on the archive before starting the search for relevant material. The process of (re-)examining finding aids leads to refinement of the questions and builds up contextual knowledge that increase historians' understanding of the research topic. Chu [91] identified how literary critics go through stages of preparation, in which the context of the work is identified, and elaboration where the exact area of interest is determined. Brown [73] noted as well that music scholars do background work to establish the viability of the research idea before information gathering is narrowed to a certain topic and organized in a certain way. We have found a similar pattern in the research cycle of media studies researchers. The differences, however, are in the type and number of sources used by media studies researchers. Next to traditional sources of contextual material such as books and monographs, media studies researchers now also turn to the Web and use websites, fora, and blogs. Archives are picking up on this and the need for tools that add contextual material of various kinds to for example audiovisual material [19]. Newspapers are popular context documents as well and consulted for various reasons, e.g., television schedules, interviews, or reviews. The accessibility to newspaper archives via services such as Lexisnexis,[3] may have increased their use.

We further identified four types of challenge that media studies researchers face in their data collection activities with current technologies. One of these challenges is the availability of information, e.g., production information. We found that conducting interviews is an important way of acquiring this type of contextual information that has not been preserved or that is difficult to obtain. Another observation was that although participants mentioned the need for analogue material and digitization thereof, few actually visited archives. Participants mentioned that material was not available because it was not digitized: "availability" is often taken to mean "digital availability." The activity of visiting physical archives is considered time-consuming and often inefficient.

Other challenges derive from archival search systems and cataloguing practices. Participants mentioned they were not always allowed to operate search systems themselves and had to work through an intermediary or if they had access systems turned out to be difficult to operate. Additionally, material was not described or made accessible as researchers expected due to the limited capacity of archives and libraries to extensively catalogue material. This problem may be expected to become worse with the advent of digitally born material. Even if material is accessible, copyright issues and the cost of acquiring material are posing challenges to media studies researchers. Although tools that support browsing and filtering are becoming available in libraries and archives [305], these challenges underline the importance of support for gaining an overview of the available material on a topic and determining suitable selection criteria.

[3] http://academic.lexisnexis.com

## 3.5 Conclusion

In this chapter we have investigated the research cycle of media studies researchers, a group of humanities researchers that deals with various information types and technologies and on which few studies in information behavior have focused. In particular, we have developed a model of the media studies research cycle that captures how research activities relate to changes in the research questions of media studies researchers. We found that information gathering and analysis activities are especially influential on research outcomes and result in additional questions or a changed perspective in the research questions of media studies researchers. Reasons identified are that media studies researchers learn about the availability of material, discover trends in the material or gain alternative views on a topic. To address this issue, in Chapter 5, we develop an interactive information retrieval system aimed at supporting media studies researchers in gaining insight in trends in the data and comparing alternative hypotheses. We evaluate the support provided by this system and investigate how it effects the research questions of media studies researchers.

Additionally, we found that media studies researchers turn to new information sources for contextualization on the Web, e.g., blogs, fora, as well as a diverse set of sources accessible though individual search services, e.g., newspapers and archives. In the second part of the thesis we investigate algorithms to support various types of contextualization. That is, discovering records related to the same event or related events across archives in Chapter 7 and finding a group of entities based on a common relation in Chapter 8 and 9.

Finally, we found that challenges of accessibility and discoverability of primary source materials remain a concern as collections are distributed across various repositories, cataloging capacities are limited, and IR tools are difficult to use. In the next chapter we address the issue of which source to consult and investigate how media studies researchers may benefit from various aggregated display styles in an interactive information retrieval system.

# 4

# Aggregated Search Interface Preferences in Multi-Session Search Tasks

In the previous chapter we found that one of the challenges encountered by media studies researchers is the search for material across archives. Primary source materials as well as contextual materials are held at various institutions and made discoverable through individual search engines. Aggregated search interfaces provide one solution to this problem by presenting information from multiple sources in a single display.

Two general types of display exist: tabbed, with access to each source in a separate tab, and blended, which combines multiple sources into a single result page, see Figure 1.1 in Chapter 1. Findings, however, regarding users preference and search effectiveness with respect to these two display variants are mixed. Moreover, subjects were found to prefer recent and multimedia content over textual content while performing single session search tasks (cf., §2.4). In contrast, humanities researchers engage in multiple search episodes (cf., §2.4) and prefer older material with textual content as well as multimedia material (cf., §2.3).

The general question we seek to answer in this chapter is whether media studies researchers engaging in multiple search sessions during a research project benefit from alternative display methods in an aggregated search system. The research behavior of humanities researchers with a novel tool is difficult to study especially over a longer period of time, due to time constraints and difficulties in adopting a new tool (cf., §2.3). Therefore, we invite media studies students to participate in our study, which provides us with a larger pool of potential participants that are trained in media studies research. In this setting we seek answers to our second research question, which we recall from Chapter 1:

**RQ 2.** How do media studies students use alternative display methods in an aggregated search system during a research project?

**a.** Do media studies students switch between tabbed and blended display types during a multi-session search task and what is the motivation to switch between display types?

**b.** Do changes in media studies students' information need across sub-tasks influence preference for a particular display type?

**c.** What other factors are related to changes in display preference during a multi-session search task?

Few studies have focused on the design of controlled experiments to investigate the use of interactive information retrieval systems during multi-session search tasks [190]. As the work is exploratory, in this study, we use multiple approaches and conduct two types of studies, i.e., a longitudinal study and a laboratory study. In our *longitudinal* study we follow 25 students during a four week research project. We provide students with an interface that allows switching between a tabbed display, blended display, and blended display with a find-similar functionality. This allows for the study of display use and switching in a naturalistic setting. Through questionnaires and focus group discussions we elicit the motivation for using a particular display.

In a *laboratory* study we present 44 students with a multi-session search task consisting of three complex sub-tasks. Each sub-task is carried out with a different display: the first task with the tabbed display; the second task with the blended display; and the third task with the blended display with a find-similar feature. We zoom in on the influence of changes in information need associated with recurring search sessions by manipulating whether a subject is assigned three sub-tasks about the same topic or three sub-tasks about different topics. This allows us to investigate the factors associated with changes in information need and whether these influence preference for a tabbed or blended display in different stages of a multi-session search task.

The remainder of this chapter is organized as follows. In Section 4.1 we describe our three variants of the aggregated interface that we will contrast. In Section 4.2 and 4.3 we describe the experimental setup and results of the longitudinal and laboratory study, respectively. In Section 4.4 we discuss the results of both studies in light of our research questions and we conclude in Section 4.5.

## 4.1 Aggregated Search Displays

An aggregated search interface provides access to multiple, often heterogeneous collections. In building an aggregated search interface there are three aspects to consider: (i) how to retrieve relevant information from each vertical; (ii) which verticals to show; and (iii) where to place the verticals on the screen. Below we first give details of the back-end of our aggregated search interface and then describe three types of aggregated search interface displays: a tabbed display, a blended display and a blended display with find-similar functionality. The source code for the interface is available, see Appendix A.

### 4.1.1 Data and Retrieval Back-end

The theme of the course selected for our longitudinal study is television history and the research projects carried out by the students are centered around television personalities between 1900 and 2010. To provide students with relevant material we obtained six collections from several archives and libraries: (i) a television program collection (metadata records for .5M programs); (ii) a photo collection related to television personalities (20K photos); (iii) a Wiki dedicated to television programs and presenters (20K pages); (iv) scanned television guides (25K pages); (v) scanned newspapers starting from 1900 till 1995 (6M articles); (vi) digital newspapers starting from 1995 till 2010 (1M articles). Each collection was indexed using Lucene SOLR 4.0 and the retrieval model used was BM25.

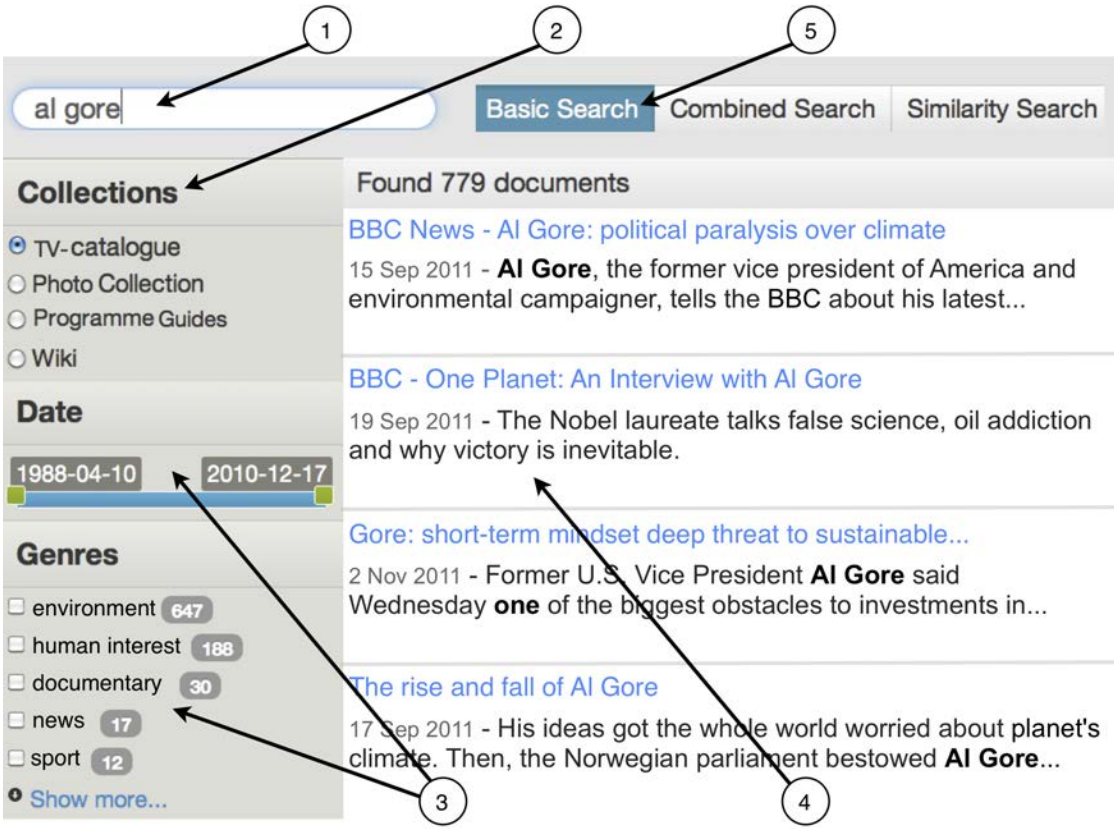
A single retrieval model may not be equally effective for each collection due to differences in term statistics and the presence of specific metadata fields in different collections [306]. To overcome this issue we provide faceted search and query preview capabilities in the aggregated search displays. These enable users to explore and learn about the characteristics of individual collections [354]. We did not optimize a retrieval model for each collection as we did not possess the relevance judgements or click logs required to optimize each vertical (cf., §2.4.3). The available facets depend on the collection as documents in some collections have rich metadata while others do not. The interaction model behind the facet values operates as follows: values within a single facet are combined using an OR operator, while values across facets are combined using an AND operator [332].

### 4.1.2 Tabbed Display

The tabbed display mimics the functionality and layout of a typical web search engine and is the default display presented to the user when opening the search interface. The tabbed display is shown in Figure 4.1; we use numbers 1, ..., 5 to reference specific components in the display. It consists of: (1) a search box; (2) a collection selection menu; (3) values for several facets; (4) a result list; and (5) an option to select the tabbed or blended interface.

A search is initiated by submitting a query via the search box (1). The television program collection is selected by default (2). In response to a query ten document snippets are shown for the selected collection on the result page (4). At the top of the result list the number of documents found is displayed and at the bottom of the page a pagination button enables moving to the next and further result pages. To further refine the results the top five values for several facets are available (3). Pressing on the *show more* button extends the list of facets up to the top 100. By selecting a different collection (2) results for this collection are displayed for the current query. Each tab displays the same number of results (ten) as a ranked list, five of which are generally visible above the fold depending on the size of the display.

A document can be viewed by clicking on a result snippet. The search result page is then covered by an overlay and the content and metadata of a document are displayed. There are two special cases where the displayed information depends on the type of document, i.e., image and Wiki. Records from the photo or tv-guide collections consist of several images; in the overlay the photo and its metadata are shown with the additional option of viewing the next or previous photo without leaving the overlay. Wiki page content is not shown in the overlay, only the metadata and a link that opens the page in a new tab. To save a particular document a bookmark button is available, when hovering over a search result and in the document view overlay. Clicking the bookmark button saves the title of the document to a bookmark list available as a drop down list at the top of the screen. Selecting a title from the drop down list triggers an overlay with the document.



Found 779 documents

BBC News - Al Gore: political paralysis over climate
15 Sep 2011 - **Al Gore**, the former vice president of America and environmental campaigner, tells the BBC about his latest...

BBC - One Planet: An Interview with Al Gore
19 Sep 2011 - The Nobel laureate talks false science, oil addiction and why victory is inevitable.

Gore: short-term mindset deep threat to sustainable...
2 Nov 2011 - Former U.S. Vice President **Al Gore** said Wednesday **one** of the biggest obstacles to investments in...

The rise and fall of Al Gore
17 Sep 2011 - His ideas got the whole world worried about planet's climate. Then, the Norwegian parliament bestowed **Al Gore**...

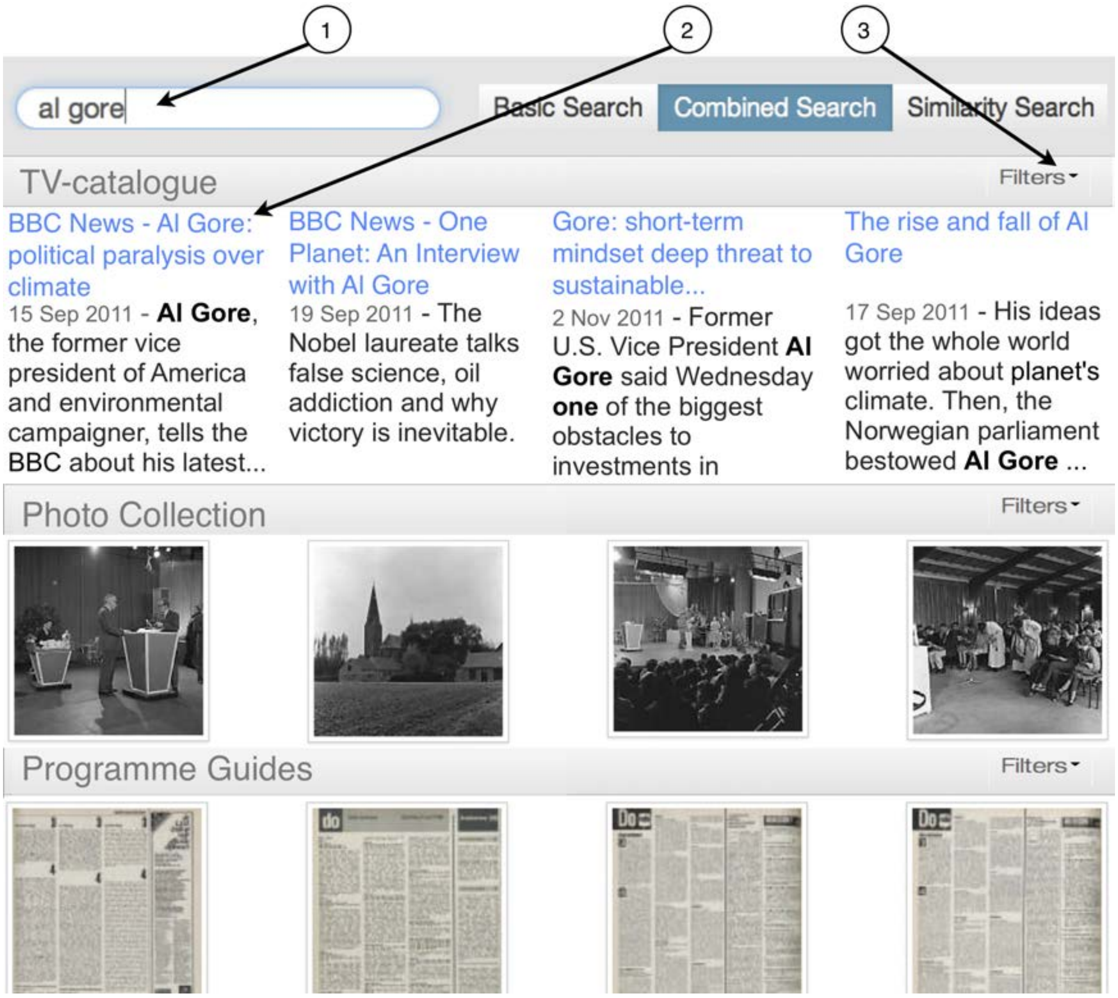
**Figure 4.1:** Tabbed display.

### 4.1.3 Blended Display

The blended display is based on the layout of interfaces typically used within digital libraries [305]. The blended display is shown in Figure 4.2; we use numbers 1, . . . , 3 to reference specific components in the display. It consists of: (1) a search box; and six horizontally orientated rectangular panes one for each collection, i.e., grouped display (cf., §2.4.3). The vertical order in which the six collections are displayed is fixed and is the same as the order of the collection selection menu in the tabbed display, e.g., the top most pane shows results for the television program collection. Within a collection pane four results are shown ordered from left (most relevant) to right (less relevant) (2). To free up screen space for displaying results horizontally the facets for each collection are hidden. A button is available to toggle the display of facet values (3). The facets, bookmark options, and document views are all the same as in the tabbed interface.

The blended display presents six by four results of which eight are generally visible above the fold depending on display size. More results are available as each source provides an independent pagination button. By scrolling down more results are available than in the tabbed display, however, only four results are available from each source. This is a trade-off between the two display types. The tabbed display supports searching through a single source, while the blended display supports searching through multiple sources.

In our blended display we use a fixed order and fixed number of results. This is in



BBC News - Al Gore: political paralysis over climate
15 Sep 2011 - **Al Gore**, the former vice president of America and environmental campaigner, tells the BBC about his latest...

BBC News - One Planet: An Interview with Al Gore
19 Sep 2011 - The Nobel laureate talks false science, oil addiction and why victory is inevitable.

Gore: short-term mindset deep threat to sustainable...
2 Nov 2011 - Former U.S. Vice President **Al Gore** said Wednesday **one** of the biggest obstacles to investments in

The rise and fall of Al Gore
17 Sep 2011 - His ideas got the whole world worried about planet's climate. Then, the Norwegian parliament bestowed **Al Gore** ...

**Figure 4.2:** Blended display.

contrast to a trend seen in current web search engines, which is to display a variable number of results from a variable number of collections in a vertically oriented ranked list in a query dependent manner (cf., §2.4.3). The latter display method requires careful tuning, either on large numbers of users or on log data, both of which are unavailable for complex or multi-session search tasks.

### 4.1.4 Similarity Search

An additional feature was added to the blended display and provided as an additional screen, i.e., similarity search. By clicking on a document the user submits the current query and the first 100 words of the clicked document as an OR query [67, 308]. This type of explicit feedback is often implemented in a digital library context to discover related material across multiple sources [240]. However, the additional effort of using this feature and lack of understanding of how it works limits the use of find-similar like features [45, 355]. We add this feature to the blended display in order to explore its use in multi-session search tasks.

By clicking the find-similar button a query is issued and the document used for the query is placed in the query history pane appearing as a new row above the first row of search results. There are four slots, one of which is filled each time the find-similar button is clicked ordered from newest (left) to oldest (right). When all slots are occupied the oldest document is removed and the others shifted to the right to free up space for the new document. Whenever a new query is issued the query history is emptied by removing the top row. Note that without clicking the find-similar button the similarity search display is exactly the same as the blended display.

## 4.2 A Longitudinal Study

The goal of our first, longitudinal study was to investigate in which way humanities researchers change between display type in an aggregated search system for a given work task in a natural setting and how this swapping behavior can be motivated, if it happens at all. Hence, the type of multi-session work task we make use of is a scholarly task, i.e., the task of writing a paper, which covers multiple search tasks to be carried out to collect and investigate the material that finally contributes to the paper [209]. Below we first describe the details of our experimental setting followed by an analysis of the data and discussion of the findings.

### 4.2.1 Study Setting

Through inquiries among staff at the humanities department of our university we searched for a course suitable for our study, one in which students complete the cycle of a research project including the development of a research question, searching for and interpreting materials, and reporting about their findings. The course that offered this structure and on which we settled is entitled: “Reception of media in historical perspective.” The lecturer is experienced in teaching the course and used the same course schedule and assignments as in previous years. The course consists of two parts, we only focus on the first five weeks in which students conclude the following research project: “reconstruct the historical context of the 1950s (start of television) or 1920s (start of film) in order to explain the emancipatory role of a famous female television/film personality. Write a photo essay in which you incorporate primary and secondary sources that place the photos in context.” The research project is split into four assignments, one due each week: (i) “familiarize yourself with the television/film personality of your choice and compose a list of five additional television/film personalities that fit your theme;” (ii) “start collecting images centered around your theme and collect material that motivates choosing these images;” (iii) “select ten images and add keyword descriptions to create a coherent story;” (iv) “prepare a presentation explaining the theme of your project, using the collected material.” These assignments provide structure for the students as well as a natural separation of the research project into four parts.

#### Subjects

In total twenty-five students participated in the course and all were at the postgraduate level in the area of media studies. The sample consists of twelve men and thirteen women, aged in terms of median ($MD$) and interquartile range ($IQR$) around twenty-three years ($MD = 23$, $IQR = 22$–$24$). We asked subjects background questions using a five point Likert-type scale, where a one indicates no agreement and a five indicates extreme agreement. Subjects reported high levels of experience in general computer use ($MD = 4$, $IQR = 4$–$5$) and using online search tools ($MD = 4$, $IQR = 4$–$5$),

### Procedure

At the end of the first lecture the experimenters presented the aggregated search system, explained how the three displays work and described the available data sources. After the presentation subjects were invited to sign up for the experiment, fill in a consent form and create a login. Subjects who signed up were not required to use the interface but were encouraged to only use the system as a supplement to the sources normally used in the class. The incentive for using the system was the availability of unique sources otherwise unavailable. The experiment lasted four weeks.

### Data Collection

In conducting a naturalistic study there is a tension between collecting as much data as possible and affecting the environment through this data collection process. After negotiating with the lecturer we settled on three points of interaction with the subjects: at the end of the first lecture, once during the project and at the end of the final lecture (of the first part of the course). Participation of the students in all parts of the study was optional and no requests were made to use any specific features or to complete specific tasks with the aggregated search system, e.g., bookmarking. It was not deemed appropriate to collect the project grades of the students for use in the experimental analysis.

Two methods were used to collect qualitative data: open question surveys, and focus group discussions. In preparation of the focus group, subjects were asked to complete two online questionnaires, one before the second class and one before the fourth class. Questions focused on the motivation to use a particular type of display. At the end of the second lecture and the final lecture a fifteen minute focus group discussion was conducted. The discussion focused on the motivation for using a particular type of display and switching between displays. The discussion was tape recorded and transcribed for later analysis. In addition, we collected quantitative data by logging all actions with the aggregated search system.

## 4.2.2 Analysis

To investigate what type of display subjects use and whether they switch between display types in a multi-session search task, we first analyze the log data and then place our findings in the context of the qualitative data collected.

### Log Analysis

We first investigate whether there is a change in the amount that each of the three displays was used during the project. We considered three alternative indicators for the amount of use of a display. First, the amount of time that a subject spent logged into the system may

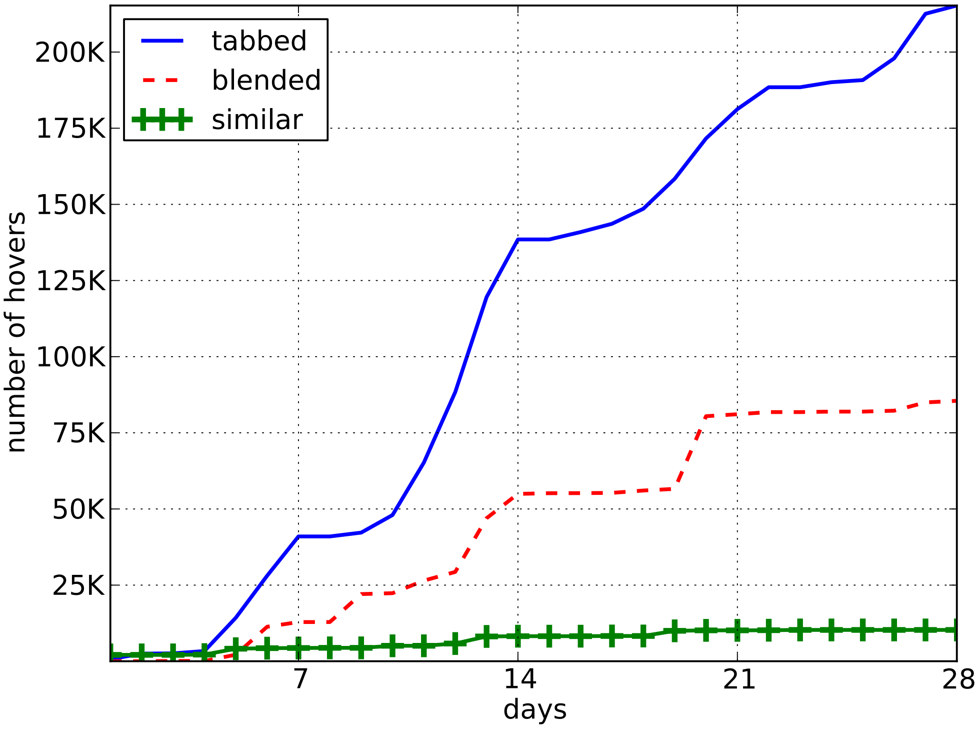


**Figure 4.3:** Cumulative number of mouse hovers recorded per day for the tabbed display (solid line), blended display (dashed line) and similarity display (cross marked).

overestimate the time a subject actually spent using the system. Similarly, clicks may underestimate the use of the system as inspection of result pages does not necessarily lead to clicks. We finally settled on the number of mouse hovers within a particular display as an indication for the amount of use of a display. Note, however, that mouse movement styles depend on the individual participant, the interface, and the task [78, 287]. Therefore, when comparing between subjects and interfaces hovers only provide an indication whether a display was used. Comparison of relative amounts is only meaningful within subject and display type. A mouse hover is recorded when the cursor remains in the same position for 40ms and is only recorded again if the position changes [172]. Figure 4.3 shows the cumulative number of hovers with each of the three displays per day. We observe a sharp increase in the number of hovers recorded for each display before the 7th, 14th, 21th, and 28th day of the project, followed by a plateau of inactivity. These days coincide with the lecture and assignment deadline for each week and provides a natural separation of the course project into four stages. We further observe that most hovers are recorded for the tabbed display (solid) throughout the twenty-eight days of the project. The blended display (dashed) receives less hovers, and the similarity display (crossed) receives the least hovers. In the second week more hovers are recorded for the tabbed and blended display compared to the first and last week. This may be due to the assignment of that week or the focus group after the second lecture. We return to this issue in the discussion of the qualitative analysis.

Next, we investigate whether individual subjects differ in the of use of each display. In the log data we find that not all subjects used all of the displays provided by the aggregated search interface. For all twenty-five subjects, hovers are recorded for the tabbed display, for eight-teen subjects hovers are recorded for the blended display, and for eleven subjects hovers are recorded for the similarity display. Figure 4.4 shows the total number of hovers recorded per user during the project on the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar) ordered by the total number of hovers. We observe that for twenty-three out of twenty-five subjects most hovers are recorded for the tabbed display, while for two subjects most hovers are recorded for the blended display. There is a notable difference in the number of hovers recorded for eight subjects with the blended display as at least 25% or more of their total hovers occur with the blended display. This may indicate a difference in use or search style between users with the displays. We explore this further in the qualitative analysis.

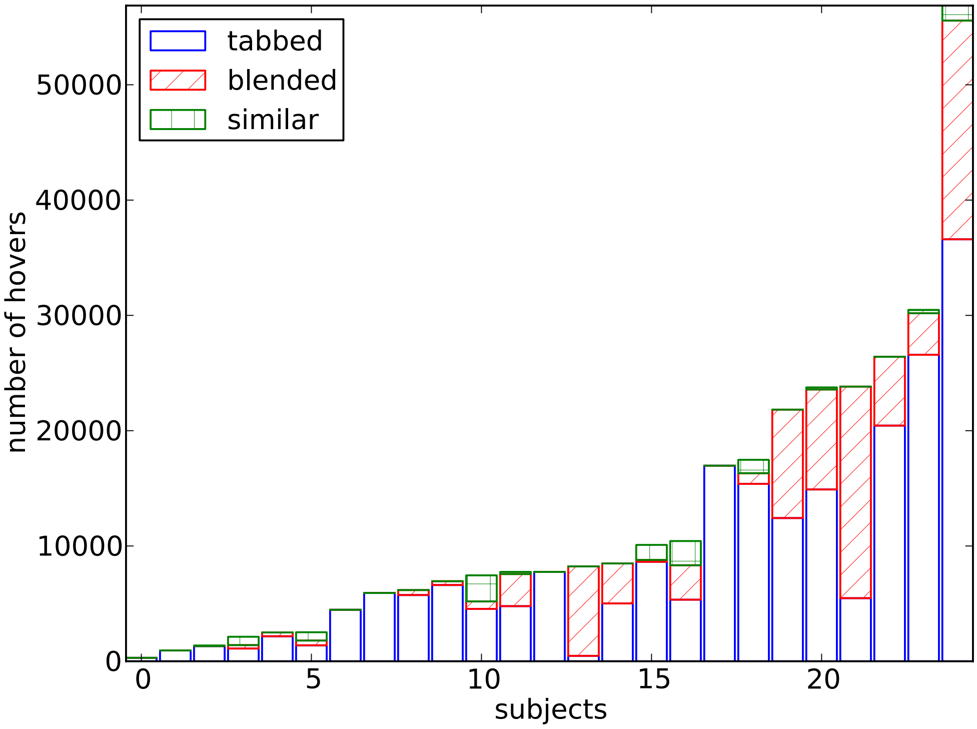


**Figure 4.4:** Total number of hovers received per user during the project on the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar).

Finally, we investigate whether individual subjects differ in the amount of use of each display per week. Figure 4.5 shows the number of hovers recorded for each subject per week of the project with the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar). We observe that during the first week (bottom row in Figure 4.5) out of the twenty-five subjects eleven make no or limited use of the displays. Of the remaining subjects most have at least 75% of their hovers in the first week recorded with the tabbed display (12/25). For the last two remaining subjects (2/25) one predominantly uses the blended display, the other the find-similar display. In the second week (second row Figure 4.5, use of all displays increases. Out of twenty-five subjects ten have around 100% of their hovers recorded for the tabbed display. The blended displays are used more than in the first week as for eight subjects about 50% or more of their hovers are recorded with the blended display and two with the find-similar

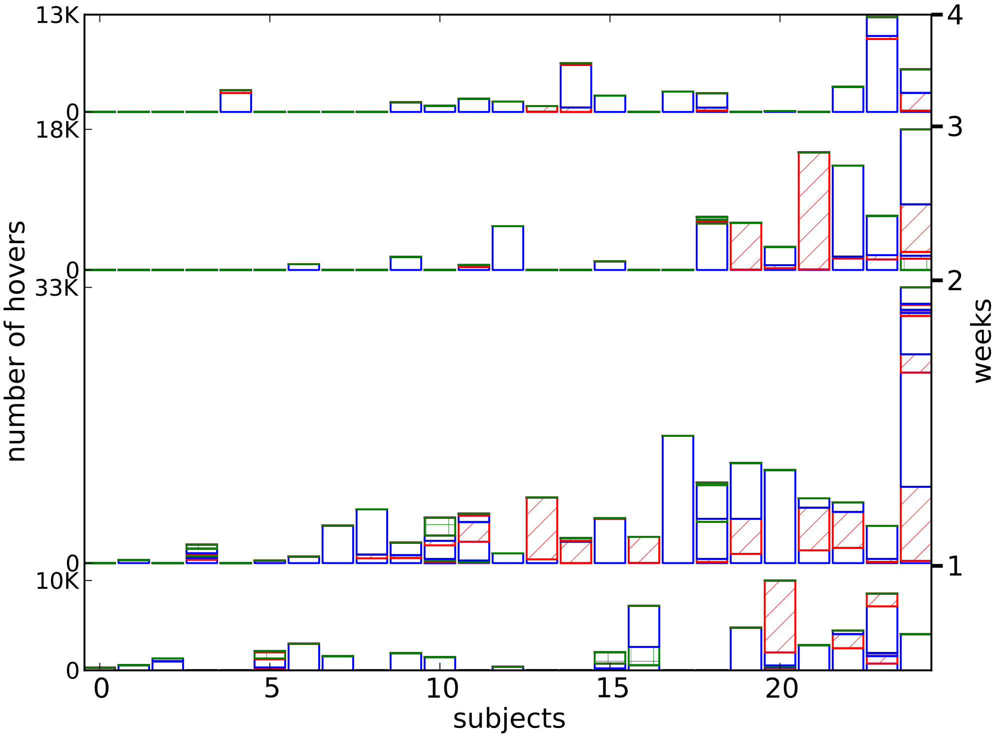


**Figure 4.5:** Number of hovers per user for each of the four stages of the project with the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar).

display. The remaining five subjects have no or hardly any hovers recorded for any of the displays. In the third week no or a limited number of hovers are recorded for most subjects (13/25) with the aggregated search system. The remaining subjects (12/25), however, continue to use the system. In the fourth week some subjects start to use the various displays again while they did not in the third week. For others use increases or decreases compared to the third week. Of the active subjects (13/25) most subjects use the tabbed display.

We perform a Chi square test to investigate whether these differences in hovers per week are due to chance and find that there is significant association between project week and hovers with each display ($\chi^2(df = 6, N = 311061) = 14458.8, p < 0.001$). Table 4.1 shows a cross-tabulation of display type and project week. We find that in the second and third week relatively more attention is spent to the blended display (50% and 30%) than the tabbed display (45% and 20%), which receives hovers more throughout the project. The standardized residuals of the hover values that most contribute to the significant effect are highlighted in boldface. Especially the number of hovers of the similarity display in the first week is surprising, as well as the increased hovers with the blended display in the third and diminished hovers in the fourth week. Hovers with the tabbed display dominate the fourth week.

The absolute number of hovers provides an indication of the amount of use. Whether subjects switch between displays, however briefly, and in what sequence remains unclear. Figure 4.6 shows the same data as Figure 4.5—the number of hovers recorded for each subject per week of the project with the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar)—on a log scale. We observe that six out

**Table 4.1:** Cross-tabulation of display type and project week, in percentages of both the week marginals and the interface marginals. The number of hovers are distributed as: tabbed $N = 215284$, blended $N = 85488$, and similarity $N = 10289$. Standardized residuals are show in brackets, the largest are highlighted in boldface

| | Tabbed | Blended | Similarity | |
|---|---|---|---|---|
| Week 1 | 70.4%<br>19.0% (3.5) | 22.1%<br>15.0% (-24.7) | 7.5%<br>42.4% (**55.5**) | 100% |
| Week 2 | 67.7%<br>45.3% (-6.7) | 29.5%<br>49.7% (14.8) | 2.7%<br>38.3% (-11.8) | 100% |
| Week 3 | 60.5%<br>19.9% (-27.9) | 36.9%<br>30.5% (**47.8**) | 2.6%<br>17.9% (-10.2) | 100% |
| Week 4 | 89.1%<br>15.8% (**46.7**) | 10.5%<br>4.7% (**-63.2**) | 0.4%<br>1.4% (-31.7) | 100% |
| | 100% | 100% | 100% | |

of twenty-five subjects only use the tabbed display and four out of twenty-five subjects only switch once to another display during the four weeks of the study. The remaining fifteen subjects switch two or more times between displays and switching occurs in all four weeks. The tabbed display is the default whenever subjects login to the system and switching between displays takes a conscious effort. It is to be expected that in the first week subjects explore the system and try the various options. That switching between displays occurs in any of the weeks and for some subjects several times within a week indicates that subjects have a function for each display during the project.

### Qualitative Analysis

The above analysis shows that a majority of the subjects switches display during the weeks of the research project. In the first week subjects predominantly use the tabbed display, while a minority prefers to use another display. In the second and third week of the project usage of the blended display increases over that of the first week. The number of subjects using the system, however, decreases. Below we summarize the qualitative data gathered from two surveys and two focus group discussions and discuss subjects' motivation to switch between displays.

Before the second lecture fifteen subjects completed a survey with open questions about the use of the three displays during the first week. On the question why subjects did or did not use the tabbed display all subjects indicated to have used the system: eight subjects indicated to use the display to explore the collections and to get an impression of the available material; five subjects used the display to try out the system; one subject used the display but provided no motivation; and one subject liked the ability to specify in which collection to search. On the question why subjects did or did not use the blended display ten subject indicated not to use the blended display: five subjects indicated not to have decided on a specific topic for their project yet and that they preferred the tabbed

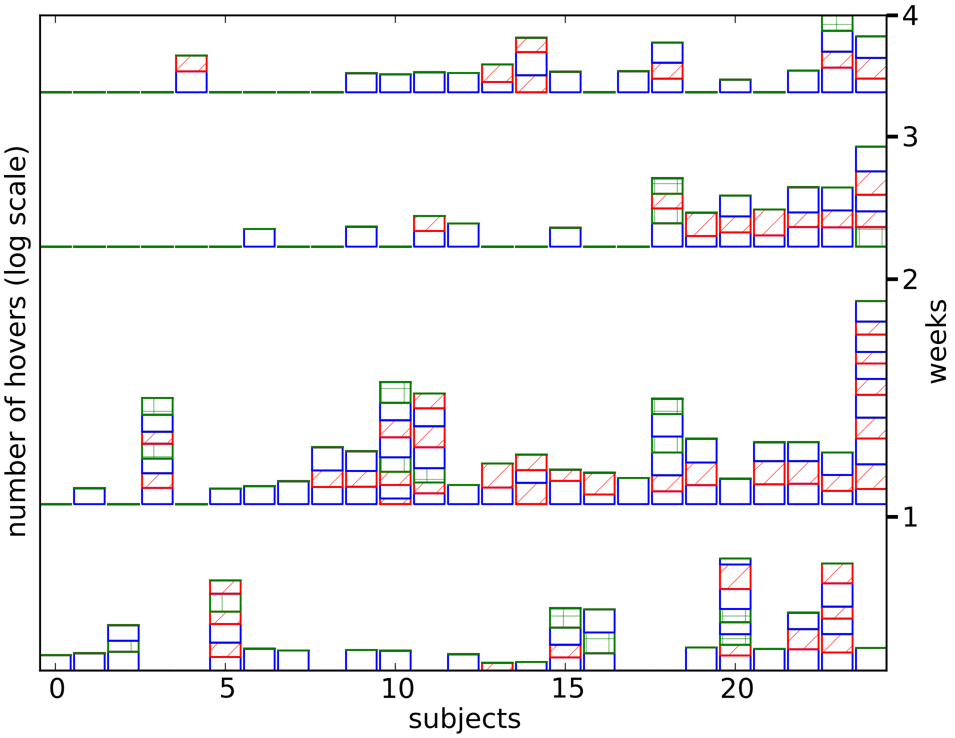


**Figure 4.6:** Number of hovers log scaled per user for each of the four stages with the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar).

display to explore available material; and five subjects did not provide a motivation. Of the remaining five subjects that did use the blended display: three subjects indicated to like the blended display as it provided them with a better overview than the tabbed display; one subject found the blended display confusing; and one subject was trying out the display. On the question why subjects did or did not use the similarity display eight subjects indicated not to use this display: four subjects indicated not to have decided on a specific topic; and four provided no motivation. Of the remaining seven subjects that did use the similarity display: three subjects expected to find related television personalities; two subjects found that the similarity display did not provide relevant information; and two subjects were trying out the display.

In a fifteen minute focus group discussion with twenty-five participants conducted at the end of the second lecture we focused on the motivation for using or not using the system with a particular display. The consensus among participants is that the information provided by most of the sources is too specific for the current stage of the project: "in this phase I want general information, now when I search for photos of her I would get very specific information." Due to the stage of the project the tabbed display, with which a single collection can be explored, was preferred: "I found the Wiki interesting, because of our assignment this week" and "next week I will search differently, as this week was for exploration." In the similarity and blended display all collections are presented at the same time, which subjects found less useful: "you get results with which you can not get an overview of the topic." Subjects also noted to prefer other resources, e.g., Wikipedia, and web search engines in order to get an overview of the possible topics for their project.

In summary, the above qualitative data suggests several reasons why subjects preferred the tabbed display during the first week and why use of the system was lower compared to the second week. First, the search behavior in this stage was of an exploratory nature as subjects tried to get an overview of available information and to decide on a topic for their project. Second, the information in most sources other than the Wiki was too specific. For example, a photo of a presenter and a guest during a specific broadcast of a television show provides little background information necessary to get oriented on a topic, for this purpose Wikipedia and similar sources are more useful. Finally, subjects preferred to use other resources such as Wikipedia and web search engines.

Before the fifth lecture fifteen students again completed a survey with open questions about the use of the three displays during the second third and fourth week. Regarding the use of the tabbed display six subjects remarked that they liked to use it to find specific information in a specific collection, while three subjects preferred it to get an overview, and one provided no motivation. Subjects who did not use the tabbed interface favored the use of external sources (three subjects) or preferred the overview of multiple collections (two subjects). Regarding the blended display seven subjects remarked that it was useful to find additional information: “to find next to images, more specific information about a person or event” or “get an as complete as possible overview of what is available about a specific program.” To search for images two subjects preferred the blended display and one subject preferred the overview without searching for anything specific. The remaining four subjects did not find the blended display useful as it was confusing or did not provide relevant material. The similarity display was used by two subjects that mentioned being curious about the material that would be found. The remaining thirteen subjects did not use the similarity display because they did not get around to using it or because results were irrelevant.

We conducted a second fifteen minute focus group discussion with twenty-five students at the end of the fourth week. We focused again on the motivation for using a particular display and centered the discussion around three points: (i) did becoming more familiar with the search system affect display choice; (ii) did increased knowledge of the research topic affect display choice; and (iii) did the assignment each week affect display choice. Subjects did not experience difficulties in using the displays: “it is a reasonably intuitive system, I understand it.” They did not associate any changes in display use with an increase in familiarity with the search system. Both increasing knowledge of the research topic and type of assignment are factors that subjects associate with changes in display use. Two groups emerged in this discussion. One group that changed from using the tabbed display to using the blended display: “I started to use the combined display more, because I know what I need, and then I want to see everything that is available about her.” The other group changed from using the blended display to using the tabbed display: “I felt the same, that my search started to become more focused, but then I preferred using the tabbed display because in the first week you do not know how to use photos and program guides, and now I wanted to know what was said in those sources about a specific person.” Some subjects remarked that they did not use the system in later weeks of the project as it did not contain any relevant material for the theme they had chosen. There were two options for a project theme, i.e., female role models in either television or film, see Section 4.1. The material in the repositories is predominantly focused on television and does not contain as much relevant material for the second theme.

In summary, the second survey and focus group suggests that after the first week subjects' use of the system increased as their search topic became more concrete and the various sources became more relevant. We find several motivations for the switching behavior between displays throughout the project. The use of the tabbed display is motivated in the first week by an exploratory information need, where the Wiki was the only suitable source. In later weeks searching a single source is motivated by a more specific information need, e.g., photos for the essay. The use of the blended display is initially motivated by a need to explore the content of the various sources. Later the blended display provides a way to explore multiple sources simultaneously for material about a specific entity or topic. We observed several switching patterns in Figure 4.6 that can be explained using the above motivations. Whether subjects engaged in a particular pattern seemed to depend on the individual subject. Some subjects indicated to have searched with the blended display in the first week and preferred to continue searching with the tabbed display in the second week. In this case the blended display is only used for an initial exploration of the sources. Other subjects repeatedly switched between the tabbed and blended display during the project. In these cases the blended display is likely used to search for additional information about a specific entity across sources.

Finally, the use of the find similar display was sporadic. Subjects mention several reasons, i.e., not getting around to using this display or not yet having decided on a specific topic yet. Some subjects note that the results deemed similar by the system did not match their expectation of similar results.

### Limitations

The setup of the longitudinal study involved that subjects were asked several times, i.e., focus group and surveys, about their use of and switching between displays. This may have influenced the behavior of the subjects. To avoid influencing subjects two measures were taken. First, subjects do not get a reward and have no obligation to use the tool. Their priority is to finish the assignment for each week and pass the course. To this end subjects choose to use the display which they feel will be most effective in solving their task [350]. As one subject remarked: "I have to search longer before I find something [with the aggregated search system] and I notice that I want to be efficient and then I quickly switch to a web search engine. I want to use the [aggregated search system] because I know you [the experimenters] need that, but at some point I think [insert expletive] I need to find my material and then I choose for something that quickly gives me results..." Although use of the system diminished in the third and fourth week, a number of subjects used the system and switched displays throughout the project. This suggests that the tabbed and blended displays are both found useful by some of the subjects and that each display supports a recurring need. Second, subjects were encouraged to use the system and explore all of its functionality with equal emphasis. The find similar and blended displays were both mentioned in the first focus group as being used less than the tabbed display in order to elicit a discussion about why this is. If the focus group discussion had any influence on the use of the find-similar display than this had a diminishing effect on its use.

Another limitation of the study is the relative poor quality of the result rankings. We did not evaluate or optimize the ranking for each of the verticals and some subjects commented that it was difficult to find material. Although this may have been a reason for the switching behavior, this is not reflected in the comments from the focus group or surveys.

## 4.3 A Laboratory Study

In our first, longitudinal study we observed that the majority of subjects switched between displays and that use of the blended display was motivated by a need to explore the content of the collections available in the aggregated search system. Initially subjects desired an overview of the sort of information generally available in the sources, later the blended display was used to obtain an overview whenever a new specific information need arose, i.e., what is available in each source about a specific entity. We are interested in whether user preferences change in later stages of the multi-session search task depending on whether there are changes in the users' information need. We recreated this process in a laboratory study by assigning subjects three tasks to complete: the first task with a tabbed display, the second with a blended display, and the third with the similarity display. We compare two conditions: (i) one in which the three tasks are about the same topic allowing subjects to become increasingly familiar with the topic; and (ii) one in which tasks are about different topics simulating changes in information need. In this setting we investigate whether subjects preferences for displays deviate as some become more familiar with a topic, while others encounter new topics.

### 4.3.1 Study setting

The study uses a mixed methods design with three search tasks as within subject factor and the (in)dependency of the tasks as between subject factor. This design is similar to other work on multi-session search tasks [215]. The following three sub-tasks emulate a multi-session search task: (i) imagine that you work at the editorial office of a current affairs program and are asked to collect information about [celebrity]. Collect at least five items deemed to be relevant for this collection, (ii) search for events that were key in the career of [celebrity]. Collect at least five items for this collection, for example articles and photographs, about these events, (iii) key in the career of [celebrity] was [event]. Collect at least five items about the run-up to the program, the program itself, and the aftermath. In this experiment [celebrity] stands for one of three television personalities and [event] represents an event related to the corresponding celebrity. The tasks are modeled after the assignments subjects received in the longitudinal study, i.e., starting out broad and gradually becoming more specific. Randomizing this order would obfuscate the simulation of a multi-session search task.

A subject completes three sub-tasks that are either dependent or parallel [215]. In a dependent series of sub-tasks all tasks are performed with respect to the same celebrity, while in a parallel series of sub-tasks each task has a different celebrity as topic simulating changes in a users information need. In the parallel condition the celebrities are randomized across tasks and in the dependent condition celebrities are randomized across subjects. In this manner each celebrity and task combination is seen an equal number of times in each condition.

All subjects are presented with the same display for each task. To complete the first task subjects are provided with the tabbed display, for the second task with the blended display, and for the third task with the similarity display. We used a fixed order for the displays as we are interested in the preference of the blended display in later stages of the multi-session search task. In this way subjects are able to become familiar with the repositories using the tabbed display before encountering the blended display.

### Subjects

A group of 44 undergraduate students participated in the study. For two of the subjects a technical failure prevented recording the pre-experiment questionnaire data. All analyses reported in the results section are excluding these two subjects and are conducted on the remaining 42 subjects. The students (30 female, 12 male) were around nine-teen years of age ($MD$ = 19.0, $IQR$ = 19–21.75). Random assignment to conditions resulted in two groups. The dependent task group contained twenty-three students (6 male and 17 female), where the parallel condition group contained nine-teen subjects (6 male and 13 female). We asked subjects of both groups to answer background questions using a 5 point Likert-type scale, where a one indicates no agreement and a five indicates extreme agreement. Subjects generally reported high levels of experience in general computer use ($MD$ = 4, $IQR$ = 4–5).

We additionally measured topic knowledge (1 item scale) and search experience (7 item scale) to make sure that these characteristics were balanced over the two groups. We found no significant difference between the two groups in terms of topic knowledge $U(N = 42) =$ 191.5, $p =$ 0.242), general computer use $U(N = 42) =$ 211.5, $p =$ 0.430), or search experience $U(N = 42) =$ 212.0, $p =$ 0.440), using Mann-WhitneyU hypothesis tests.

### Procedure

The study was conducted on two separate occasions in a computer lab equipped with thirty computers. Subjects could sign up for one of two time slots to participate in the study. Subjects of the first slot were asked not to communicate about the experiment until the second run had been finished. Each occasion of the study started with an introduction to the research project and a viewing of a five minute tutorial video explaining the use of the search displays. After viewing the tutorial, subjects were invited to fill in a consent form and create a login. Subjects were not allowed to talk, use mobile phones or open any other browsers. Four experimenters invigilated each of the experiments.

After completing a background questionnaire eliciting demographics subjects were randomly assigned to either the dependent or parallel condition and presented with a pre-experiment questionnaire to collect information about search experience and prior knowledge about the task topic(s). Next, subjects started on the multi-session search task. For every task the subjects were given ten minutes after which they were redirected to a post-task questionnaire asking about the topic difficulty, the perceived usability of the display, and the search effectiveness of the display. After the final search task a post-experiment questionnaire was presented that asked subjects to order the displays by preference, and to express any remarks they might have about the experiment. In total an experiment lasted one and a half hours.

**Table 4.2:** Questions used in the search effectiveness scale.

| | |
|---|---|
| 1 | The display has supported me in solving the task. |
| 2 | The display provided enough information for every collection to solve the task. |
| 3 | The display provided surprising results relevant to the task. |
| 4 | The display supported me in finding relevant results. |

### Data Collection

We use two Likert-scales to measure subjects' preference for a particular type of display. For each item in the scales subjects indicated their agreement with the statement on a scale from one to five, where a one indicates no agreement and a five indicates extreme agreement. To arrive at a single score for each Likert-scale the mean of the responses was taken. To measure perceived usability we use an adaptation of the perceived usability sub-scale of the O'Brien Engagement Scale [257]. We modified the scale to apply to an aggregated search setting by substituting the word "shopping" and "website" by "searching" and "display" in the items. Additionally, we rephrased some of the items to a positive wording to arrive at an alternating ten item scale.

To measure search effectiveness we used the items in Table 4.2. A final 3 items were devoted to subjects' perception of the task: (i) "the task was difficult to complete;" (ii) "there was one collection most useful in solving the search task;" and (iii) "to get an overview of the results in different collections is important in solving the task." These were not combined and are analyzed separately. All actions with the aggregated search interface were logged.

## 4.3.2 Results

We first investigate whether subjects' preferences differ within conditions in terms of relations between sub-task and the dependent variables usability, task difficulty, and search effectiveness. We use non-parametric tests as data is collected on ordinal scales, and not all variables meet the assumption of normality. Table 4.3 shows the median ($M$) and inter quartile range ($IQR$) for each of the dependent variables split over each topic and the two conditions, i.e., dependent and parallel. We observe that there is a significant interaction between sub-task and perceived usability of the display in the within subject dependent condition (Kruskal-Wallis $H(2, N = 42) = 17.7$, $p < 0.001$, top part of Table 4.3). Post hoc comparisons show a significant difference between task 1-task 2 (Mann-Whitney $U(N = 42) = 155$, $p = 0.008$) and between task 1-task 3 (Mann-Whitney $U(N = 42) = 81.5$, $p < 0.001$) at the Bonferroni corrected significance level of $\alpha = .05/3$. These results show that when the topic of the search task remains the same, subjects find the tabbed display easiest to use, and that the blended displays provided in later stages are considered more difficult to use. In the parallel condition we do not find a significant interaction between sub-tasks and usability. The median and inter quartile range values are stable for the first two topics and decreases for the third topic. With this result we are unable to determine whether users find the blended display easier to use to explore a new topic.

**Table 4.3:** Interaction effects of the perceived usability, search effectiveness, topic difficulty, one collection is important and an overview is important responses in terms of median (interquartile range) across conditions (dependent and parallel) and across topics. The Kruskal Wallis hypothesis test is used to test for significant within subject effects, the Mann-WhitneyU hypothesis test is used to test for significance between subjects.

| | Task 1 | Task 2 | Task 3 | |
|---|---|---|---|---|
| Condition | | Perceived usability | | K-W (H, p) |
| Dependent | 3.9 (3.8-4.1) | 3.6 (3.3-4.1) | 3 (2.7-4) | **(17.7, $<$ 0.001)** |
| Parallel | 3.7 (2.9-4) | 3.7 (3.1-4.1) | 3.4 (2.9-3.7) | (2.90, 0.24) |
| M-W (U, p) | **(108, $<$ 0.001)** | (197, 0.30) | (190, 0.24) | |
| | | Task difficulty | | |
| Dependent | 3 (2-3) | 2 (2-3) | 3 (2-4) | (5.95, 0.051) |
| Parallel | 3 (3-4) | 3 (2-3) | 2 (2-3) | **(6.45, 0.040)** |
| M-W (U, p) | **(134, 0.017)** | **(151, 0.045)** | (155, 0.054) | |
| | | Search effectiveness | | |
| Dependent | 3.3 (3.3-3.8) | 3.3 (2.8-3.9) | 3.3 (2.4-3.9) | (2.68, 0.26) |
| Parallel | 3.3 (3.3-3.8) | 3.3 (3.0-3.6) | 3.3 (2.6-3.6) | (3.21, 0.20) |
| M-W (U, p) | (194, 0.27) | (201, 0.33) | (194, 0.27) | |
| | | Overview important | | |
| Dependent | 4 (3-4) | 3 (3-4) | 3 (3-4) | (4.02, 0.13) |
| Parallel | 4 (3-4) | 4 (3-4) | 3 (3-4) | (2.75, 0.25) |
| M-W (U, p) | (214, 0.46) | **(149, 0.040)** | (183, 0.19) | |
| | | One collection important | | |
| Dependent | 4 (3.5-4) | 3 (2-3.5) | 3 (2-3.5) | **(16.4, $<$ 0.001)** |
| Parallel | 4 (3-4) | 3 (2-4) | 3 (2-4) | (2.70, 0.26) |
| M-W (U, p) | **(148, 0.038)** | (211, 0.43) | (218, 0.50) | |

We further observe that there is a significant interaction between sub-task and task difficulty in the within subject parallel condition (Kruskal-Wallis $H(2, N = 42) = 6.45$, $p = 0.04$, second row of Table 4.3). Post hoc comparisons show a significant difference between task 1-task 3 (Mann-Whitney $U(N = 42) = 94.5, p = 0.006$). In the dependent condition we do not find a significant interaction, although in terms of the median values the first and third sub-task are considered more difficult than the second sub-task. This suggests that in the parallel condition subjects find that the tasks become easier as they search material for new topics with the blended displays.

We find no relation between sub-tasks and the search effectiveness that subjects experience with the displays.

Two possible factors that moderate the preference of subjects for a certain display are: (i) whether for a particular task a subject considers a single collection as the most important source to search for material; and (ii) whether a subject considers it important to get an overview of the material available in different collections for a certain task. We do not find any interaction between the sub-tasks and the importance of getting an

**Table 4.4:** Cross tabulation of search moves in stage 1 for the dependent and parallel conditions. Search moves are given in percentages with the standardized residuals in brackets.

| Move | $(N = 926)$ Dependent | res. | $(N = 728)$ Parallel | res. |
|---|---|---|---|---|
| paginate | 6.48 % | (-1.38) | 9.34 % | (1.55) |
| bookmark | 11.77 % | (-1.37) | 15.52 % | (1.55) |
| filter | 17.93 % | (1.21) | 14.29 % | (-1.36) |
| change tab | 21.38 % | (1.14) | 17.58 % | (-1.29) |
| queries | 11.23 % | (-0.55) | 12.64 % | (0.62) |
| view docs | 22.14 % | (0.28) | 21.15 % | (-0.32) |
| delete bookmark | 0.32 % | (-0.46) | 0.55 % | (0.52) |
| unique queries | 8.75 % | (-0.08) | 8.93 % | (0.09) |

overview. We do find a significant interaction between sub-tasks and the fact that a single collection is considered to be the most important in solving a task (Kruskal-Wallis $H(2, N = 42) = 16.4$, $p < 0.001$). Post hoc comparisons show a significant difference between task 1-task 2 (Mann-Whitney $U(N = 42) = 109.5$, $p < 0.001$) and between task 1-task 3 (Mann-Whitney $U(N = 42) = 103.5$, $p < 0.001$). Subjects in the dependent conditions initially consider a single collection important to solve the search task, while in later tasks more collections become important.

Next, we investigate between subject effects in terms of relations between changes in information need and the dependent and moderating variables. We find that in the first task there is a significant difference between the perceived usability reported by subjects from the dependent and parallel condition (Mann-Whitney $U(N = 42) = 108$, $p < 0.001$). We further observe that in the first and second task there is a significant difference between the task difficulty reported by subjects from the dependent and parallel condition (Mann-Whitney $U(N = 42) = 134$, $p = 0.017$), i.e., in the dependent condition the first and second task are considered easier.

Regarding the moderating variables, we find that in the second task subjects in the parallel condition consider getting an overview more important than subjects in the dependent condition (Mann-Whitney $U(N = 42) = 149$, $p = 0.40$). With respect to the importance of a single collection to solve the first search task, subjects in the dependent condition consider this more important than subjects in the parallel condition.

During the first stage subjects are assigned the same task and possess the same level of prior knowledge. Subjects were randomly assigned to the dependent and parallel conditions and no significant differences were found in terms of subjects' background, i.e., prior knowledge, search experience, and computer use. Finding significant effects in the first stage indicates the presence of additional factors that possibly interact with the dependent variables. Table 4.3 shows that subjects in the dependent condition consider the first task easier, a single collection more important, and find the tabbed display easier to use than subjects in the parallel condition.

We also investigate to what extent search strategy is involved in the effects between the two conditions. Table 4.4 shows a cross tabulation of the search moves with the tabbed display recorded within the dependent and parallel condition. We find that there is a significant relation between the moves made in a display and the conditions ($\chi^2(df = 7, N = 1654) = 16.2, p = 0.022$). From the standardized residuals we find that especially the following search actions contribute to the significant effect: paginating, bookmarking, using filters and changing collection (tab). Subjects in the dependent condition tend to switch between collections and use the facet filters, while subjects in the parallel condition tend to paginate and bookmark more frequently. This suggests that subjects in the dependent condition initially explore more of the collections while subjects in the parallel condition dig deeper into material of a single source.

## 4.4 Discussion

In this section we discuss our findings in light of our second research question from Chapter 1, where we asked:

**RQ 2.** How do media studies students use alternative display methods in an aggregated search system during a research project?

**a.** Do media studies students switch between tabbed and blended display types during a multi-session search task and what is the motivation to switch between display types?

**b.** Do changes in media studies students' information need across sub-tasks influence preference for a particular display type?

**c.** What other factors are related to changes in display preference during a multi-session search task?

Regarding **a**, we find that the majority of subjects switch between displays during the project. The main motivation to use the tabbed display is to zoom in on a single source as other sources are not considered relevant at that stage of the project, e.g., Wiki documents are relevant at the start of the project to gather general background information, while photos of specific events become relevant at a later point. The use of the blended display is initially motivated by a need to explore the content of various sources. Later the blended display provides a way to explore multiple sources simultaneously for material related to a specific information need.

With respect to **b**, we find that when subjects are completing search tasks about the same topic there is a negative influence on the usability of a blended display when switching from a tabbed to a blended display. In the longitudinal study we found that some subjects turn to the blended display when investigating a new aspect of a topic, but prefer the tabbed display to zoom in when their research topic has become concrete. These findings suggest that subjects are less likely to switch to a blended display when subjects are engaged in a sequence of search tasks related to the same topic.

Finally, in answer to **c**, we find that there are several other factors that influence the preferences of subjects for the tabbed display, i.e., whether subjects find the first search task easy, whether subjects use a particular search strategy, and whether they find that one of the collections is most important in solving the search task. These factors limit the generalizability of our findings as the observed interactions between changes in information need and usability may be specific to our sample with this particular configuration of factors. This is to be expected in an exploratory study, however, our findings suggest an array of factors that should be considered when investigating user preferences for aggregated search display types in multi-session search tasks.

When comparing the search strategies observed in the laboratory study to those from our longitudinal study we find parallels with the two types of search behavior observed there, i.e., some subjects preferred to get an overview of the various verticals first, while others prefer a particular vertical at first before exploring other verticals. Rodden et al. [287] found differences in users' strategies to examine search result pages, i.e., economic and exhaustive searchers. The behaviors observed here may be variants of these, i.e., broad and narrow searchers when inspecting aggregated search result pages.

Other studies in aggregated search have investigated whether users prefer tabbed or blended displays for single-session search tasks of varying complexity. Sushmita et al. [313] found that for complex tasks users prefer blended displays. Arguello et al. [16] found that more verticals are clicked with a blended display, but that users do not necessarily prefer a blended or tabbed display. As a possible explanation the search experience of the subjects is suggested. Our findings suggest that users differ in their initial search strategy and that during a multi-session search task depending on strategy and changes in information need users change display type.

## 4.5 Conclusion

Aggregated search interfaces are a promising way to provide humanities researchers with an overview of results from various sources. In this chapter we investigated the use and preferences of media studies students for a tabbed and blended display within the context of a particular multi-session search tasks, i.e., a research project. In our first, longitudinal study we observed that the use of the tabbed display is predominantly motivated by a need to zoom in on specific sources. The majority of subjects, however, switched between the tabbed and blended displays. Use of the blended display was motivated by a need to explore the content of the sources available in the aggregated search system. Initially, some subjects desired an overview of the sort of information generally available in the sources, later the blended display was used to obtain an overview whenever a new specific information need arose.

In a laboratory study a multi-session search task was recreated, composed of three sub-tasks. The first sub-task was completed with a tabbed display, the remaining subtasks with blended displays. The conditions were manipulated by either providing three sub-tasks about the same topic or about three different topics. We found that a certain combination of factors, i.e., subjects' search strategy and changes in formation need across sub-tasks, negatively influences perceived usability of the blended display. Two types of searchers emerged, i.e., one in which users explore a single source before consulting other sources and one in which users gain an overview of several sources before focusing on a particular source. The combined results from both studies suggest that subjects change display preference during a specific type of multi-session search task, i.e., seeking archival materials across multiple heterogeneous sources during a research project.

These findings have implications for archives and digital libraries that aim to improves accessibility of their collections through aggregated search interfaces. In particular, an institutions' landing page that presents a single type of aggregated search display by default does not provide an optimal experience for every user. Moreover, during longer search sessions, enabling users to easily switch between different aggregated search display styles may improve user experience.

This chapter focused on investigating the utility of three aggregated search displays for supporting users to search across multiple heterogeneous sources of records from various archives and libraries. One such display, i.e., find similar, was not adopted by subjects as it failed to provide relevant material and the interpretation of what constitutes a "similar" result was unclear. In Chapter 7 we investigate an approach to automatically link records from two different archives, i.e., a newspaper archive and a multi-media archive, based on events. Such a method has the potential to support the type of search for records across institutions by providing high quality links to related material. In the next chapter we move away from the challenge of supporting search across several collections and instead focus on developing a tool to support exploration of a single collection.

# 5 A Subjunctive Exploratory Search Interface to Support Media Researchers

In Chapter 3 we identified two key activities in the media studies research cycle that often lead to changes in the research questions of media studies researchers: information gathering and analysis. During information gathering media studies researchers discover to what extent accounts in primary source materials cover their research topic and whether this material is available. During analysis media studies researchers gain insight in trends in the data and the existence of alternative views on their research topic.

The process of methodically collecting and analyzing material is a time intensive process. Changing a research question comes at a cost as new data has to be collected, organized and analyzed. In order to mitigate the possibility of having to change a research question, media studies researchers familiarize themselves with a topic and the available material during the initial phase of the research cycle. In this chapter we focus on developing and assessing a search interface to support media studies researchers in refining their research question in the *exploration* phase of the research cycle (cf., §3.3). Our requirements for such an interface are that it supports (i) exploring multiple views on a topic and (ii) discovering patterns in the data.

A number of exploratory search tools exist, that support various ways to explore collections, e.g., through filtering by facets, or relevance feedback, see §2.4.3 for more details. As work on exploratory search systems focuses on supporting exploration in a general setting, few have considered how exploratory search systems can support humanities researchers in discovering alternative views and trends in the data during exploration. To support users in complex search tasks (see §2.4.1), investigating multiple aspects of a topic in a subjunctive interface was shown to reduce task complexity and task completion time [223, 344] (see §2.4.3). In the humanities, standard practices to discover patterns in data are organizing and comparing [235, 262, 335] (see §2.3.3).

We propose to extend the traditional exploratory search system design in two ways. First, we incorporate two side-by-side versions of an exploratory search interface in a single interface. Second, we add visualizations in which the characteristics of the result sets are shown and can be compared. An interface that incorporates multiple instances of the same search tool in a single interface is referred to as a subjunctive interface [223]. According to this convention we propose a subjunctive exploratory search interface and investigate whether the ability to make comparisons supports media studies researchers in refining their research question. Specifically we address our third research question from Chapter 1, which asked:

**RQ 3.** Does the ability to make comparisons support media studies researchers in exploring a collection of television broadcast metadata descriptions?

**a.** Does a subjunctive exploratory search interface better support media studies researchers in a complex exploratory search task than a standard exploratory search interface?

**b.** Does the subjunctive exploratory search interface better support media studies researchers in refining a research question than a standard exploratory search interface?

**c.** Does the increase in complexity caused by the inclusion of additional features affect the usability of the subjunctive interface as compared to a standard exploratory search interface?

The remainder of this chapter is organized as follows: in Section 5.1 we describe the subjunctive exploratory search interface; in Section 5.2 we detail the experimental design; in Section 5.3 we present the results of assessing the subjunctive interface; we provide a discussion in Section 5.4; and we conclude in Section 5.5.

## 5.1 A Subjunctive Interface

The model of the media studies research cycle in Chapter 3 provides insight in the requirements for a successful search interface for media studies researchers. Below we describe the development process of our subjunctive exploratory search interface, simply referred to as *subjunctive interface* in the remainder of the chapter, followed by a detailed description of the interface features.

### 5.1.1 Development Procedure

In developing the subjunctive interface care has been taken to follow user centered design principles [63]. Here, we motivate the initial design of the subjunctive interface, describe the data used in our prototype, and findings from two rounds of usability testing.

#### Initial Design

We established two requirements for an interface for media studies researchers: (i) to provide users with support for exploration, i.e., support in formulating queries, query refinement and exploring various aspects of a topic; and (ii) to provide support for discovering patterns in the data, i.e., to compare alternatives and to observe trends in the data. A large body of work exists on interfaces for supporting exploratory search (see §2.4.3). Such interfaces provide support for the first requirement through visualizations, filters and facets. We start our interface development with the design of a prototypical exploratory search interface, such as the one by Capra and Marchionini [86].

Not as well supported by this type of interface is the ability to compare alternatives. Subjunctive interfaces have been suggested for this purpose as this type of interface allows a user to perform multiple actions in parallel and compare the results, i.e., editing a document or searching a database. Typically multiple versions of a standard interface, e.g., a standard document editor, are presented side-by-side to create a subjunctive interface [223]. In web search, a multi-view interface has been proposed that supports multiple views of a traditional web search interface and allows users to explore more aspects of a topic than a single view variant [344]. Another requirement not as well supported is the discovery of trends in the data. Visualization and data mining laboratories, as those used in data intensive disciplines, e.g., astronomy, are better suited for this purpose as these offer various visualization techniques, i.e., curves, scatter plots and renderings, to analyse large numerical datasets [168].

Given our requirements, we adapted the standard exploratory search interface design in two ways: (i) we extended the design to a subjunctive exploratory search interface by incorporating two side-by-side versions of a standard exploratory search interface; and (ii) we added a timeline visualization and a term statistics visualization in which the characteristics of the result sets obtained with each side of the interface are shown and can be compared.

### Data Set

Television studies (a sub-discipline of media studies) concerns the study of production and/or reception of television (cf., §3.1). From an audiovisual archive we obtained a catalogue of about 1.5M television program descriptions to serve as the data set to be accessed through our interface. We use descriptions as the actual programs are often not directly accessible due to copyright legislation [222]. The program descriptions are created by archivists, and primarily consist of metadata fields describing the program. For example, keywords, summary, and fields with program production information, e.g., broadcast date and program creator. The back-end of the interface consists of a Lucene SOLR index, where stopwords have been removed and stemming has been applied. For retrieval the Lucene implementation of the Vector Space Model is used.[1]

### Usability Testing

In a first round of usability testing we presented a prototype of the subjunctive interface to two groups of media studies researchers, consisting of 12 and 16 subjects. A presentation with a walk-through of the interface was followed by a group interview. The three main findings are: (i) the importance of production information such as program broadcast date and program maker, next to the content of programs; (ii) program genre information is an essential subject in media studies; and (iii) television production/reception is often studied over time. We also received feature requests, i.e., the ability to exclude certain terms, to view the query history, and to load alternative archives such as news archives and television magazine collections.

After a new round of development we performed a usability study of the subjunctive interface. The subjects consisted of 30 first year information science students that participated as part of a class project. The main concerns of the subjects were with the cosmetics of the interface, the response time, and the size of the result snippets. After in-

[1] `http://lucene.apache.org/solr/`

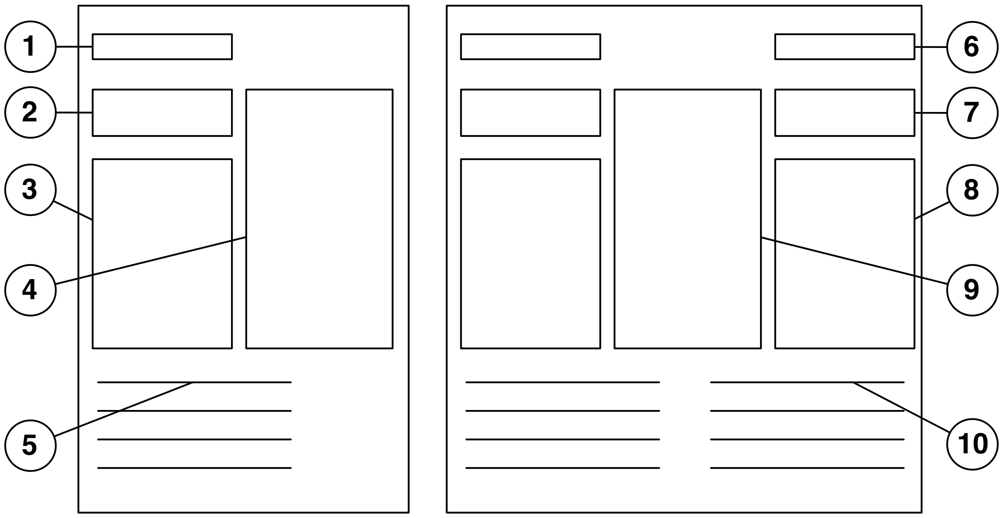


**Figure 5.1:** Schematic view of the baseline interface (left) and the subjunctive interface (right). Numbers are used for reference.

corporating this feedback we performed a series of small pilot studies with media studies researchers to test the final design and to remove any further usability issues.

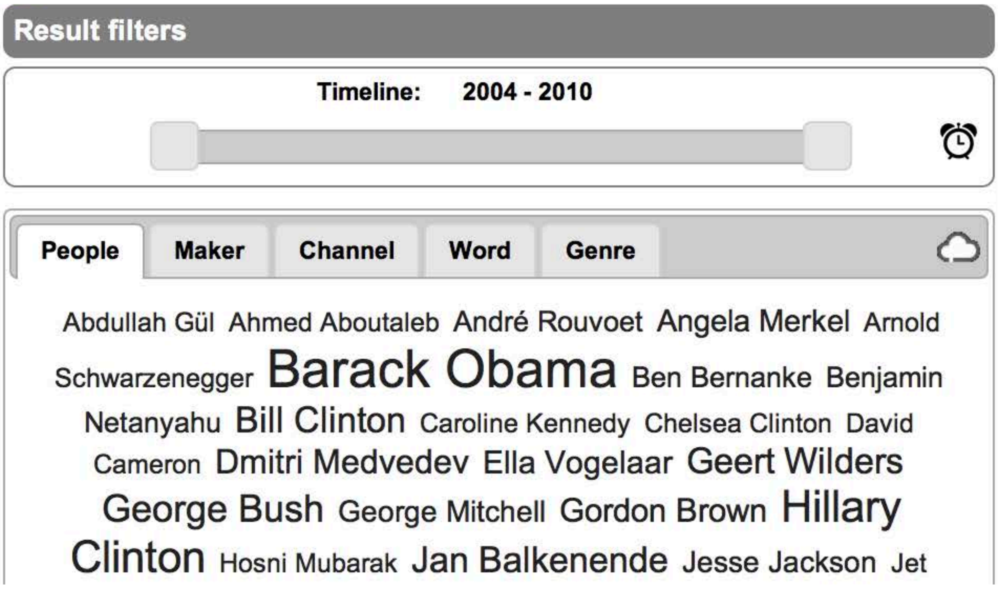
### 5.1.2 Subjunctive Interface Description

We start by describing an exploratory search interface that will serve as the basis for the subjunctive interface. In the remainder of the chapter we refer to this interface as the *baseline* as it is used for comparison in our evaluation of the subjunctive interface described in Section 5.2. Figure 5.1 shows a schematic view of two interfaces: on the left-hand side the baseline and on the right-hand side the subjunctive interface. We use the numbers (1, . . . , 10) in Figure 5.1 to reference specific components in the interfaces. The source code for the interface is available, see Appendix A.

#### Baseline

The baseline interface consists of a search box (1), two filters: a timeline (2) and a term-cloud filter (3), a timeline and term statistics chart (4), and a result list (5). The baseline interface provides traditional search functionality in that typing a query in the search box (1) results in a ranked list of document snippets (5). Each result snippet describes a program with a title, broadcast date, and a maximum of forty words from the summary of the program. Next to each snippet, a bookmark button is available. Bookmarking a program adds it to the query history, available as drop-down list, showing for each query the programs bookmarked in its result set. When clicking on a snippet an *overlay* with the complete program description appears. In the result set twenty-five program snippets are shown per page and the result set is limited to a maximum of five-hundred programs to keep the interface responsive.

The filters (2, 3) enable a user to rapidly refine the result set returned for a query [354]. Each subsequent filter that is applied operates on the remaining program descriptions.



Result filters
Timeline: 2004 - 2010
People
Maker
Channel
Word
Genre
Abdullah Gül Ahmed Aboutaleb André Rouvoet Angela Merkel Arnold Schwarzenegger Barack Obama Ben Bernanke Benjamin Netanyahu Bill Clinton Caroline Kennedy Chelsea Clinton David Cameron Dmitri Medvedev Ella Vogelaar Geert Wilders George Bush George Mitchell Gordon Brown Hillary Clinton Hosni Mubarak Jan Balkenende Jesse Jackson Jet

**Figure 5.2:** Screenshot of the timeline filter (top) and the term-cloud filter with the *people* facet selected (bottom).

Filters are reset by issuing a new query or pressing a "clear filter" button. The timeline filter removes programs with a broadcast date outside of the selected range, see top Figure 5.2.

The term-cloud filter enables a type of faceted search over the result set, see bottom Figure 5.2. Next to query refinement, faceted search also provides support for gaining insight in a topic [198, 354]. We decided on five facets based on the focus group interviews: *people* mentioned in a program, *makers* of a program, *channel* a program is

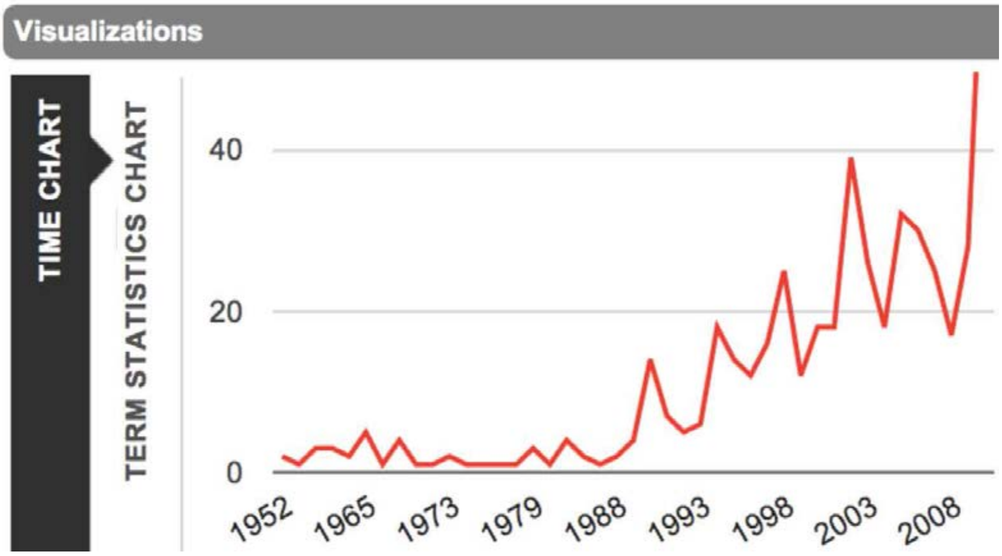


**Figure 5.3:** Timeline chart: y-axis showing the number of programs broadcast per year; x-axis showing the years.

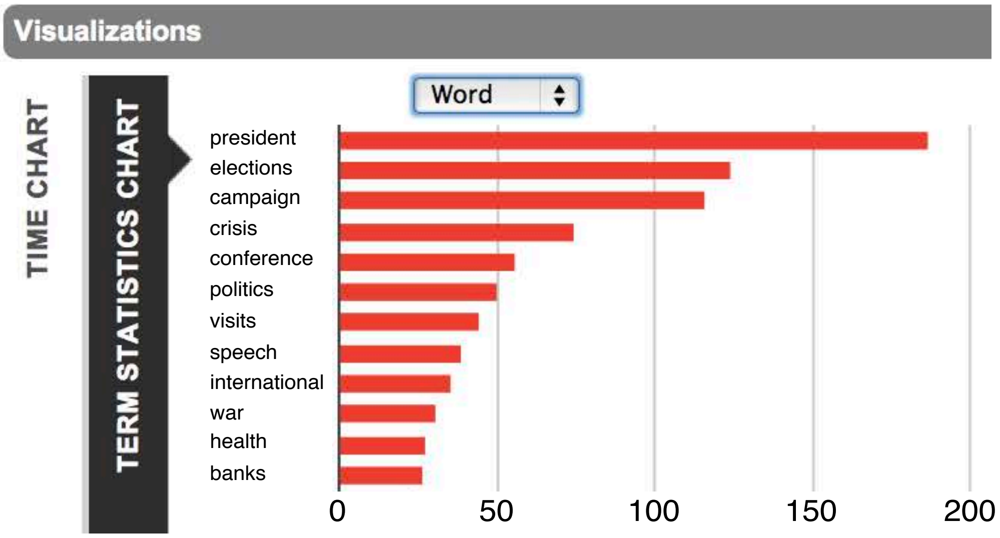


**Figure 5.4:** Term statistics chart: y-axis showing the terms with the highest frequency in the program descriptions of the current result set, x-axis showing the number of program descriptions that contain the term.

broadcasted on, *words* are keywords characterizing the program, and *genre* of a program. Each cloud provides two modes of filtering: *retain* and *remove*. To *retain* a user clicks a term and only program descriptions that contain the term are kept. To *remove*, a user clicks a term and holds the mouse button, drags it slightly and releases the mouse button causing program descriptions that contain the term to be removed from the selection. Repeating an action deactivates a filter and "un-hides" documents affected by this filter.

The final parts of the baseline interface are the timeline and term statistics charts (4). These visualizations offer support for discovering trends in the data. They are not shown simultaneously, but are accessed through a *slide deck*. The timeline chart, see Figure 5.3, is shown by default. By clicking on the *term statistics slide* an animation shows the term statistics chart "sliding" over and covering the timeline chart. The timeline chart is subsequently accessible through the *timeline slide*. An example of the term statistics chart is shown in Figure 5.4. A drop-down menu allows the user to select one of the facets (people, maker, channel, words, and genre) to inspect the terms that occur most frequently in the program descriptions.

### Subjunctive Exploratory Search Interface

The left-hand side of the subjunctive interface consists of the same features as the baseline interface. On the right side the subjunctive interface further consists of an additional search box (6), timeline and term-cloud filter (7, 8), and result list (10); see Figure 5.1. The two search boxes with their respective filters and result lists are independent and in essence provide the user with a second exploratory search interface. The visualizations in the subjunctive interface (9) differ from those in the baseline (4). The timeline chart shows two curves, one for each result set, see Figure 5.5. The curves are color coded black and red. Similarly, the search boxes are colored black and red on the left and right-hand side of the interface, respectively, to indicate their correspondence to the user.

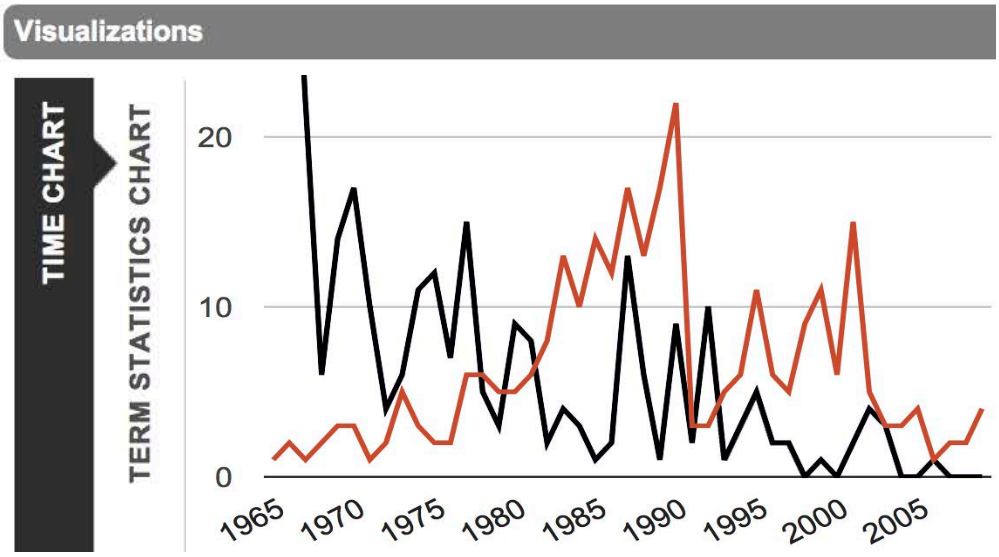


**Figure 5.5:** Subjunctive interface timeline chart: black corresponds to the result set for the query "protests", the lighter shade (red) to the query "riots". The axes are defined in Figure 5.3.

Analogously, the term statistics chart shows two bars per term, one for the frequency of the term in the left result set and one for its frequency in the right result set. The terms are required to occur in both result sets and are ordered by the total frequency in both sets, see Figure 5.6.

Note that although care has been taken in the design of the interface, we do not claim that this design is optimal. One suggestion for improvement is a tabbed view allowing a user

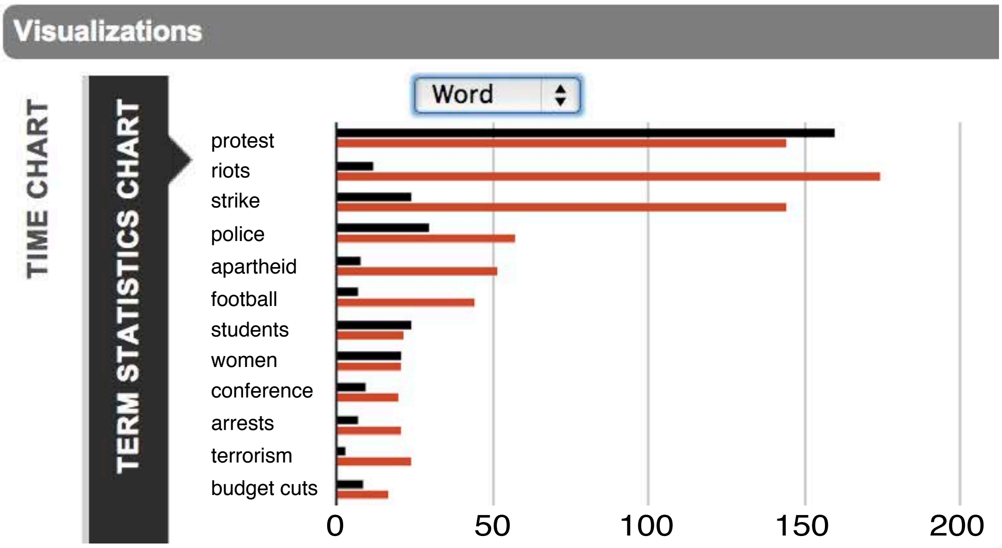


**Figure 5.6:** Subjunctive interface term statistics chart: y-axis showing terms with the combined highest frequency in the two result sets, x-axis showing the number of program descriptions that contain the term. Bar colors correspond to those in Figure 5.5.

to operate any number of instances of the interface and thus make any number of comparisons. Another issue arises from the term-cloud tabs, although offering faceted search the facets are hidden and only one facet is available at a time. Finally, we opted for cloud visualizations which have been found to be inferior to alphabetical listings [166]. The current design however satisfies the essential requirements for a subjunctive exploratory search interface and is suitable to answer our research questions.

## 5.2 User Study

To assess the support provided by the subjunctive interface we conduct a user study with media studies researchers. Next, we describe the experimental design and our evaluation methodology.

### 5.2.1 Experimental Design

The experiment was set up as a remote user study with a between subjects design [190]. We decided on a remote user study to be able to reach a wider audience of media studies researchers. A disadvantage is that there is less control over the setting of the study.

#### Study Procedure

Subjects were recruited by spreading a URL among researchers and students at six media studies institutes. The URL directed subjects to a webpage explaining the experiment. Then subjects were presented with a consent form and a background questionnaire. Next, a three minute tutorial video of the interface was shown followed by a 3 minute practice session.

After practicing, subjects were given the following complex exploratory search task: “As preparation for writing a research paper on the topic of migrants you investigate an audiovisual repository. You are interested in how migrants are represented on television. The goal of exploring the repository is to help you establish the initial research question for your paper.” Subjects were instructed to bookmark programs deemed relevant to formulate their research question and given 30 minutes to search for relevant programs with one of the two interfaces. After 30 minutes, or when subjects pressed a done button, a form was presented in which subjects were asked to submit a research question. While formulating their research question the subjects had access to the program descriptions bookmarked earlier during search. The final step consisted of a usability questionnaire. A session took about 45 minutes per subject; as a reward, subjects received a 10 Euro gift certificate.

#### Subjects

The interface is developed to support media studies researchers, we therefore targeted subjects that had at least completed a Bachelor’s degree in media studies. Out of 61 subjects, 38 fully completed the experiment. Two subjects that did complete the experiment were excluded from the experiment: one subject that spent a total of 26 seconds interacting with the interface and one subject that had not yet completed a Bachelor's degree. This left us with 36 subjects, 17 for the baseline and 19 for the subjunctive interface.

In terms of research experience, subjects are from a wide range of academic positions in media studies: 13 Master level students, 9 PhD students, 3 post doctoral researchers, 6 assistant professors, 1 full professor, and 4 research support staff. The research experience of subjects, in terms of the median (*MD*) and interquartile range (*IQR*), varies as subjects are a mix of researchers and students (*MD* = 3, *IQR* = 0–6.5). We asked subjects background questions using a 5 point Likert-type scale, where a one indicates no agreement and a five indicates extreme agreement. Subjects generally reported high levels of experience in general computer use (*MD* = 4, *IQR* = 4–5) and using online search tools (*MD* = 4, *IQR* = 4–5). Subjects had little previous experience with the topic of the search task, e.g., media and migration (*MD* = 2, *IQR* = 1–3). We found no significant differences between the groups in terms of these statistics using Wilcoxon rank-sum tests.

## 5.2.2 Evaluation Methodology

We assess the support the subjunctive interface provides for media researchers in terms of three aspects: (i) support in exploration of different views of a topic; (ii) support in refining a research question; and (iii) general usability.

### Exploration of Different Views

It is difficult to obtain a fixed set of relevance judgements for a complex exploratory search task, e.g., gathering documents that serve as a basis to formulate a research question, as relevance is hard to determine in such a broad task [190]. Instead of using precision and recall, we evaluate support for exploration in terms of user interaction derived from server side log files. We hypothesize that an interface that provides better support for exploration will enable subjects to generate more query formulations and that subjects will bookmark more diverse documents.

### Research Question Refinement

To evaluate subjects' research questions we asked three media studies researchers, experts in the field of media and migration, to act as assessors: an associate professor (judge$_1$), a full professor (judge$_2$), and a post-doctoral researcher (judge$_3$). Research questions were judged on five criteria: (i) general quality (g); (ii) extent to which a scope is defined, i.e., limiting the question to a certain person or time (s); (iii) clarity of formulation (f); (iv) embedding (e), i.e., the degree to which the research question relates to literature; and (v) originality (o). In the media studies research cycle one of the factors influencing the refinement of the research question are changes in data selection criteria. We therefore hypothesize that research questions formulated by subjects with the subjunctive interface will be judged higher on the scope criterion.

**Table 5.1:** Medians and interquartile ranges for user interactions with of the baseline and subjunctive interface.

| Interface feature | Baseline | | Subjunctive | |
|---|---|---|---|---|
| query formulations | 3 | (2–6) | **5** | **(3.3–7.8)** |
| bookmarks | 9 | (5.8–22.3) | 9 | (2.8–12.8) |
| document views | 3 | (1–7) | 2 | (0.3–5.3) |
| timeline filter | 3.5 | (2.5–9.5) | 3 | (0–8.5) |
| term-cloud filter | 28 | (13.5–41) | 16 | (9–36) |
| filter/analysis time (sec) | 303 | (204.8–589.3) | 384 | (222.5–544.3) |
| inspect result time (sec) | 253 | (51.3–438.8) | 202 | (94.8–385.3) |
| total time (sec) | 532 | (308–995) | 575 | (386.3–947.3) |

### Usability

Exploratory search systems are more complex than standard web search interfaces [354]. We introduce a subjunctive version of an exploratory search interface that essentially doubles the number of features in the interface. We assess the subjunctive interface in terms of usability and use the following criteria: (i) usefulness, (ii) intuitiveness; (iii) ease of use; and (iv) interestingness, based on [221].

All judgements regarding the research questions and questions in the exit-questionnaire are given on a five point Likert-type scale, where the level of agreement is indicated in the range from one (not at all) to five (extremely). When we report results, a Wilcoxon rank-sum test is used to determine significant differences between groups at the $\alpha < .05$ level. In tables significant differences are always in comparison to the baseline and indicated in bold face.

## 5.3 Results

### 5.3.1 Exploratory Search Support

We address our third research question (**RQ3**) as raised in Chapter 1 in three parts. In the first part we asked: (**a**) does a subjunctive exploratory search interface better support media studies researchers in a complex exploratory search task than a standard exploratory search interface? To address **a** we evaluate the performance of the baseline (bl) and subjunctive (sj) interface on a complex exploratory search task in terms of user interaction statistics and in terms of search patterns.

### User Interaction Statistics

Our first hypothesis states that with the subjunctive interface subjects will formulate more queries and bookmark more diverse documents. We first compare subjects' interactions with the baseline and subjunctive interface, followed by an analysis of the diversity of the bookmarked documents.

Table 5.1 shows the medians and interquartile ranges, for interactions of subjects with features of the two interfaces. We find that the number of query formulations is higher for subjects using the subjunctive interface (bl *MD* = 3, sj *MD* = 5). The difference is significant as indicated by a Wilcoxon ranksum test (*W* = 255.5, *p* = .0395) indicating that the subjunctive interface provides more support for generating new query formulations. We observe that subjects bookmark a similar number of documents with the interfaces (bl *MD* = 9, sj *MD* = 9). The high-end of the interquartile range for the baseline is higher but the difference is not significant. That subjects do not bookmark more documents may be due to the task description, i.e., the goal is to formulate a research question and not to bookmark as many relevant documents as possible.

The remaining interaction statistics demonstrate no apparent differences. We note that subjects spend a similar amount of time searching (bl *M* = 532, sj *M* = 575), but that users of the subjunctive interface spend more time operating the filters and/or analysing the visualizations (bl = 303, sj = 384). While in the baseline interface more time is spend inspecting results (bl = 254, sj = 202), i.e., reading result snippets and viewing documents. Although the difference is not significant it is to be expected that subjects spend more time analysing and filtering with the subjunctive interface as more information is presented. That subjects spend less time, or a similar amount of time inspecting results is surprising as the subjunctive interface presents twice as many results. We look further into this when analysing the interaction patterns.

Next we investigate whether there are differences in the diversity of bookmarked documents. We use cosine similarity as a distance measure and calculate the average pairwise cosine similarity of the documents bookmarked ($D_s$) by a subject ($s$):

$$avg\_sim(s) = \frac{1}{|P_s|} \sum_{(d,d') \in P_s} sim(d, d'),$$

here $P_s$ is the set of pairs of documents bookmarked by a subject: $P_s = \{(d, d') : d, d' \in D_s, d \neq d'\}$ and $sim(d, d')$ is defined as:

$$sim(d, d') = \frac{\sum_{i=1}^{n} d_i \cdot d'_i}{\sqrt{\sum_{i=1}^{n} (d_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (d'_i)^2}}.$$

The average similarities for subjects using the baseline ($avg\_sim$ *MD* = .62, *IQR* = .56–.69) are higher than the similarities of subjects using the subjunctive interface ($avg\_sim$ *MD* = .52, *IQR* = .43–.63 ) and this difference is significant (*W* = 195 *p* = .0496). That documents bookmarked with the subjunctive interface are less similar than those bookmarked with the baseline indicates that with the subjunctive interface a more diverse set of documents are explored.

### User Interaction Patterns

We first describe the process of creating the interaction patterns based on maximal repeating patterns [307] and then analyze the patterns generated with the two interfaces.

An interaction pattern consists of all of a subject's search actions during a search session. We identify the following action types: submitting queries (Q), using filters (F), inspecting results (I), bookmarking (B), viewing program descriptions (D), paginating to new result pages (P), and closing the interface (S). Repeated actions are aggregated

**Table 5.2:** Users' most frequent maximal repeated patterns with the baseline (bl) and subjunctive (sj) interface. Here Q is submitting queries, F is using filters, I is inspecting results, B is bookmarking, D is viewing documents, P is paginating to reach new result pages, and S is closing the interface.

| Q starts pattern | | | | F starts pattern | | | | B starts pattern | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # | bl | # | sj | # | bl | # | sj | # | bl | # | sj |
| 9 | QFIF | 12 | QFQ | 12 | FIQ | 18 | FIFIFIF | 8 | BPBP | 6 | BIQ |
| 9 | QIQI | 12 | QIFI | 14 | FIFIF | 19 | FQ | 10 | BPB | 7 | BS |
| 12 | QIQ | 13 | QIF | 15 | FQ | 23 | FIB | 11 | BDB | 8 | BIB |
| 13 | QIFI | 18 | QIQI | 15 | FID | 29 | FIFIFI | 13 | BD | 11 | BDB |
| 15 | QIF | 22 | QIQ | 25 | FIB | 34 | FFIF | 14 | BFI | 11 | BFI |
| 27 | QFI | 23 | QFI | 30 | FIFI | 55 | FIFI | 15 | BP | 14 | BF |
| 32 | QF | 37 | QF | 37 | FIF | 62 | FIF | 15 | BI | 18 | BD |
| 37 | QI | 57 | QI | 97 | FI | 116 | FI | 16 | BF | 22 | BI |

into a single action type, e.g., queries submitted in the left or right search box of the subjunctive interface are considered as a single query (Q) action. The purpose of the resulting interaction pattern is to reveal transitions between interaction types. Per subject group all occurrences of possible sub-patterns of at least two subsequent actions are counted to find the maximal repeated patterns (MRP) for each interface. For example, the sequence of actions: QFIFIQFI, contains the following MRP: QFI and FI, as these are the longest sequences that are repeated.

Table 5.2 shows the top 8 MRPs that start with a query action (Q), a filter action (F), or a bookmark action (B). In both interfaces the most frequent transition after submitting a query is to inspect the results (bl QI = 37, sj QI = 57). After this initial behavior, subjects using the subjunctive interface more often reformulate their query (bl QIQ = 12, sj QIQ = 22), while users of the baseline prefer filtering (bl QIF = 15, sj QIF = 13). This is consistent with the earlier finding, see Table 5.1, that the subjunctive interface provides more support for formulating new queries.

When starting with a filtering action and then inspecting the results (FI) subjects using the subjunctive interface tend to transition more from filtering actions to bookmarking or viewing documents (bl FIFI = 30, sj FIFI = 55). With the subjunctive interface subjects spend more time refining and inspecting result snippets (sj FIFIFI = 29); the number of times a filter and inspection sequence leads to a bookmark is comparable (bl FIB = 25, sj FIB = 23). The extra time spent refining and inspecting can be explained by the presence of the second result set in the subjunctive interface as subjects have a larger set of program descriptions at their disposal.

In the baseline interface a bookmark action is often followed by moving to the next result page (bl BP = 15) and (bl BPBP = 8), while in the subjunctive interface more often program descriptions are viewed (sj BD = 18). This suggests that subjects using the baseline interface are unable to formulate new queries or use filters to refine the result set and resort to browsing more result pages in the result set. This is consistent with the observation in the interaction statistics, see Table 5.1, that with the baseline more time is

**Table 5.3:** Medians and interquartile ranges for judgements of the research questions on five criteria: general quality, scope, formulation, embedding, and originality, by three media researchers for the baseline (bl) and subjunctive (sj) interface.

| | $\text{Judge}_1$ | | $\text{Judge}_2$ | | $\text{Judge}_3$ | |
|---|---|---|---|---|---|---|
| | bl | sj | bl | sj | bl | sj |
| general quality | 3 (2–4) | 4 (3–4) | 4 (2–4) | 3 (2.3–4) | 4 (3.8–4) | 4 (3.3–5) |
| scope | 3 (2–4) | 4 (3–4) | 2 (2–4) | 2 (2–3) | 4 (3–4) | 4 (2.3–4) |
| formulation | 3 (3–4) | 4 (2.3–4) | 3 (2–4) | 3 (2.3–4) | 4 (4–4.3) | 4 (4–4.8) |
| embedding | 3 (3–4) | 3 (2.3–4) | 2 (2–3) | 3 (2–3.8) | 3 (1.8–4) | 3 (2–4) |
| originality | 4 (3–4) | 4 (3–4) | 4 (3–4) | 4 (3–4) | 3 (2–4) | 4 (2–4) |

spent inspecting results; this is similar to the behavior observed in web search when users face a difficult search task [21]. Users of the subjunctive interface on the other hand, tend to bookmark documents on the first result page (sj BIB = 8 ) and (sj BDB = 11).

We have determined that there are differences in the interaction patterns of subjects using the baseline and subjunctive interface. Interaction patterns show that with the subjunctive interface subjects alternate more between formulating queries and inspecting results than subjects using the baseline interface. We also find that users of the baseline more often resort to an exhaustive search of the results suggesting that they are unable to refine their information need.

## 5.3.2 Research Question Formulation Support

In the second part of our third research question we asked: (**b**) does the subjunctive exploratory search interface better support media studies researchers in refining a research question than a standard exploratory search interface? To address **b** we evaluate how the exploration provided by the two interfaces affects the research questions formulated by the subjects. We perform two types of evaluation: (i) a quantitative evaluation where we use explicit judgements of the research questions; and (ii) a qualitative analysis where we divide research questions into phrases and classify these into several types to compare the composition of the research questions.

### Research Question Formulation Performance

Our second hypothesis states that research questions formulated by subjects with the subjunctive interface will be judged higher on the scope criterion. The top of Table 5.3 shows the medians and interquartile ranges for the judgements of the research questions on the five criteria, i.e, general quality, scope, formulation, embedding, and originality, by three media researchers. Overall agreement of the assessors on the criteria is low as indicated by Fleiss' Kappa ($\kappa < 0.2$). Agreement is stronger on the scope criterion ($\kappa = 0.236$).

Subjects' research questions are judged to be good in terms of quality, formulation, and originality, for all three assessors (*MD* $\geq$ 3). The level of embedding is lower (*MD* $\leq$ 3), as researchers were unable to consult any literature during the experiment. The

**Table 5.4:** Number of occurrences (unique occurrences) of types of phrases that determine the scope of a research question.

| Interface | Genre | Group | Theme | Period | Production |
|---|---|---|---|---|---|
| baseline | 11 (6) | 14 (8) | 8 (7) | 10 (10) | 2 (2) |
| subjunctive | 10 (7) | 17 (8) | 12 (11) | 9 (9) | 3 (3) |

judgements for the scope of the research questions, are mixed. We observe, however, no apparent differences between the baseline and subjunctive interface for the judgement criteria.

That we observe no differences may be due to the open ended nature of the task of formulating a research question and the difficulty of judging research questions without any further context.

### Research Question Composition

The research questions formulated by the media studies researchers provide a rich source of qualitative data. To investigate if there is a difference in the views and topics of the research questions, we perform a qualitative analysis of the phrases that define the scope in the research questions of the 36 subjects. We identify five types of phrases: defining (i) a program genre, e.g., fiction; (ii) a group to study, e.g., Muslims; (iii) a focus on a theme or person, e.g., Geert Wilders; (iv) a time period; or (v) a part of the program production process, e.g., ethics in the production of game shows.

Table 5.4 shows the number of occurrences of types of phrases that determine the scope of a research question. A similar number of phrases of type genre, period, and production is found in research questions generated with the baseline and subjunctive interface. There is a difference in the number of phrases of type group (bl = 14, sj = 17), but not in terms of unique phrases. Here, the phrase “migrants” is used to specify the population of study, most likely influenced by the use of the term in the search task. We observe that in research questions generated with the subjunctive interface more often phrases that describe a specific theme occur as compared to the baseline (bl = 8, sj = 12) and in most cases these themes are unique (bl = 7, sj = 11). This suggests that the subjunctive interface provides subjects with more support to explore different themes surrounding a topic and that they use this information to scope their research question. There is little influence on other types that determine the scope of a research question, e.g., a time period or television genre, as both interfaces provide users with the ability to spot trends through the timeline and term statistics chart.

Next we provide an example-based comparison to further illustrate the effect of the interfaces on the scope of the research questions. Table 5.5 shows the top 3 research questions for the baseline and subjunctive interface at the top and bottom respectively. Questions are ranked in terms of the sum of the three assessors’ judgements on the scope criterion. We observe that when subjects use the baseline, questions are based on observations of trends in the timeline, i.e., the increase and decrease in the occurrence of a query term in programs. This leads to research questions that focus on a single topic during a certain period, i.e., the representation of the Islam (bl $\text{rq}_1$) and the representation of a political figure (bl $\text{rq}_2$). The ability to compare changes in frequency of query terms on a timeline in the subjunctive interface, however, inspires subjects to consider more views. This leads to research questions that include interactions between multiple aspects of a topic, e.g., representation of Muslims and the influence of terrorism (sj $\text{rq}_1$) and difference in representation of refugees' children in fictional programs compared to documentaries. In the last research questions (bl and sj $\text{rq}_3$) the effect of the subjunctive interface is most obvious. Although both questions follow the change in use of terminology over time, in the subjunctive interface a contrast is made between two terms.

**Table 5.5:** Top 3 research questions for the baseline (top) and subjunctive interface (bottom), ranked in terms of the scope criterion. Alterations to the research questions are indicated by [..] and serve to protect the anonymity of subjects or to improve clarity.

| | |
|---|---|
| bl $\text{rq}_1$ | How is the Islam represented in factual television genres during the period 2000–2010? |
| bl $\text{rq}_2$ | How is [political figure] represented by the public broadcasting corporation during the elections for the house of representatives in 2010? |
| bl $\text{rq}_3$ | Investigation of the evolution of the term integration in news and human interest programs broadcasted between 1992 and 2012. |
| sj $\text{rq}_1$ | How are Muslim immigrants represented on television and what is the role of terrorism in this representation. Case study of several episodes of [$\text{program}_A$] and [$\text{program}_B$] about Muslims in [country]. |
| sj $\text{rq}_2$ | How are the experiences of refugees' children represented on television in fiction and documentaries from 1990 to 2010? Do we find any differences or changes and can these be explained? |
| sj $\text{rq}_3$ | In 1987 we observe a diminishing in the use of the term migrant worker and a rise in the use of the term immigrant. Is it possible to identify a cause in the broadcasting schedule of that time? Are there specific programs that started this development? |

### 5.3.3 Usability

To answer the third part of our third research question, i.e., (**c**) does the increase in complexity caused by the inclusion of additional features affect the usability of the subjunctive interface as compared to a standard exploratory search interface, we look at the usability of the interfaces.

Table 5.6 shows the medians and interquartile ranges of subjects' judgements of the usability of the two interfaces. Subjects indicate that both are intuitive (bl = 4, sj = 4) and that they are not difficult to use (bl = 2, sj = 2). Of the subjects, 75% do not find the subjunctive interface difficult to use (difficult $\leq$ 3). This suggests that subjunctivity can be added to exploratory interfaces with little cost to the difficulty and intuitiveness of the system. We further asked subjects if the interfaces were interesting to use and useful for media research. Subjects indicate that both interfaces are interesting (bl = 4, sj = 4) and useful, in case of the subjunctive interface subjects indicate it to be extremely useful (bl = 4, sj = 5). Regarding the visualizations subjects indicate a preference for the timeline chart. We suspect that the information in the term statistics chart is more difficult to interpret and therefore used less.

**Table 5.6:** Medians and interquartile ranges for the usability of the baseline and subjunctive interface.

| Question | Baseline | Subjunctive | Question | Baseline | Subjunctive |
|---|---|---|---|---|---|
| intuitive | 4 (4–4) | 4 (3–4) | difficult | 2 (2–2) | 2 (2–3.8) |
| interesting | 4 (4–5) | 4 (4–5) | useful | 4 (4–5) | 5 (4–5) |

## 5.4 Discussion

In this section we discuss the answers we presented in this chapter to address our third research question, which we recall from Chapter 1:

**RQ 3.** Does the ability to make comparisons support media studies researchers in exploring a collection of television broadcast metadata descriptions?

**a.** Does a subjunctive exploratory search interface better support media studies researchers in a complex exploratory search task than a standard exploratory search interface?

**b.** Does the subjunctive exploratory search interface better support media studies researchers in refining a research question than a standard exploratory search interface?

**c.** Does the increase in complexity caused by the inclusion of additional features affect the usability of the subjunctive interface as compared to a standard exploratory search interface?

Regarding **a**, we find significant evidence that with the subjunctive interface, media studies researchers exhibit different search behavior than with the baseline interface on a complex exploratory search task. Subjects are able to formulate more queries and bookmark more diverse documents than with a traditional exploratory search interface. Inspection of the interaction patterns confirms these findings. Users of the subjunctive interface follow a pattern of reformulating a query and inspecting results followed by another query reformulation and result inspection, while users of the traditional exploratory search interface formulate fewer queries and look through more result pages.

With regard to **b**, we find that with both interfaces researchers are able to formulate high quality research questions. A qualitative analysis of the research questions shows that there is a subtle difference in the research questions that subjects formulate. With the subjunctive interface subjects use more diverse themes to scope their research question. There is no influence on other types of defining the scope, e.g., a time period, as both interfaces provide users with the ability to spot trends in visualizations. An example based comparison of the top 3 research questions in terms of scope illustrates the difference in the number of views on a topic incorporated in the research questions.

Turning to part **c**, we find that although the complexity of the subjunctive interface in terms of features almost doubled compared to the standard exploratory search interface, most users indicate that the subjunctive interface is intuitive and not difficult to use. Users indicate that the subjunctive interface is interesting and judge it to be extremely useful for media research.

The subjunctive interface was developed to support the exploration phase in the media studies research cycle by supporting multiple views on a topic and discovering trends in the data. The above findings demonstrate that a subjunctive exploratory search interface can indeed provide this type of support for media studies research. A limitation of this study is that the time and data restrictions in the experiment make it an abstraction from the real research cycle. A longitudinal study where the subjunctive interface is used by media studies researches over a longer period of time will have to be conducted. Over time media studies researchers will determine whether this type of functionality provides additional value for their research, as we observed with the use of the aggregated search interfaces by media studies students in Chapter 4.

## 5.5 Conclusion

In this chapter we have presented a subjunctive exploratory search interface to support media studies researchers. Based on the analysis of the media studies research cycle in Chapter 3 we have found that media studies researchers require support in discovering multiple views on a topic and discovering trends in data to refine their research question. We have developed a subjunctive exploratory search interface and performed a user study to assess its value for media studies researchers. We have found that with the subjunctive interface media studies researchers are able to formulate more queries and bookmark more diverse documents compared to a standard exploratory search interface. With respect to the effect of type of interface on the research questions we found that with both interfaces quality research questions are formulated and we observed few differences in terms of originality, theoretical embedding and formulation. In a qualitative analysis of the research questions formulated by media studies researchers we have found some evidence to suggest that the influence of the subjunctive interface is predominantly on the scope of the research question. Specifically, users of the subjunctive interface incorporate more views on a topic in their research question than users of the standard exploratory search interface. We have observed no advantage for other types of defining the scope as visualizations in both interfaces enable spotting trends in the data. In terms of usability, media studies researchers report that the subjunctive interface is intuitive and not difficult to use, suggesting that the additional complexity in terms of features in the subjunctive interface does not reduce its usability.

These findings suggest that providing access to archival collections through a subjunctive interface is a promising way to allow researchers to discover new material and to develop their research questions. Several projects are underway to investigate the utility of subjunctive interfaces further, which we discuss in Chapter 10.

In this and the previous chapter we investigated the benefits of interactive information retrieval systems, i.e., aggregated and exploratory search systems, to support media researchers in learning about a new research topic and gaining insight in the material available in various information sources. As discussed in Chapter 3 exploration is only part of the research process. Once appropriate information sources are selected and a particular focus for a research topic has been chosen, a process of exhaustively gathering material to place the research topic into context begins. For example, a researcher investigating how actresses in the 1920s served as role models for emancipation of women on television would require information such as the directors and producers they worked with, the organizations that funded their movies, and the societal events that steered public opinion. Current information retrieval systems provide limited support for answering this sort of questions. Web search systems return documents with high precision but not necessarily a list with relevant contextual material, and although interactive information retrieval systems support discovery of this sort of contextual material, they require considerable effort on the part of the user. In the second part of the thesis we investigate methods that automatically discover this type of information.

# Part II

# Contextualization through Concepts

# 6
# Background on Information Retrieval

In Chapter 2 we saw that in the late 1940s, driven by a need to manage the increasing amount of information, two lines of information behavior research emerged: the user centered line, i.e., information science, and the system centered line, i.e., information retrieval (IR). In the first part of the thesis we focused on supporting humanities researchers in exploring archival collections. Exploration is a process in which a user learns about a collection and gains new insights through interactions with a search system. We therefore focused on the user and investigated the utility of IR systems that allow richer means of interaction for exploration.

In this chapter we provide background material for the work in the second part of the thesis, which is focused on contextualization. Here, we focus on work from the area of IR that motivates our contextualization methods in later chapters based on related concept finding. We start with a brief overview of the key concepts in standard IR in Section 6.1. We then move beyond the traditional unit of retrieval in IR, i.e., documents, and discuss retrieval tasks that return information related to a particular concept, e.g., events, questions, or entities in Section 6.2. Finally, we discuss work on semantic search related to the use of structured data to support related concept finding in Section 6.3.

## 6.1 A Brief Overview of Information Retrieval

The field of IR concerns the development of efficient and effective systems to search and manage information. The classic model of an information retrieval system considers four factors: (i) the input, usually a query representing a user's information need; (ii) a repository containing information objects; (iii) a matching rule that determines which information objects are relevant to the query; and (iv) a result output, which is typically a ranked list of information objects, and its evaluation [22, 75]. Below we discuss each of these factors individually.

### 6.1.1 Information Objects

In traditional IR the information objects returned to a user in response to a query are commonly referred to as *documents*, e.g., web pages, scientific papers, or news articles. This is in contrast to, for example, database systems that return data in response to a query or question answering systems that select a particular piece of information from a document. More generally, information objects capture information in a particular format, e.g., a painting, a video, or a CV. These information objects, however, are not directly retrievable and require some form of representation. In IR a term-based representation is predominant, e.g., a web page may be represented by the terms (unigrams) or phrases that occur on the web page. Information objects that do not have an innate textual representation, such as paintings or videos, are often assigned certain categories, tags or descriptions that serve as metadata. These metadata annotations may then be used as a representation for a information object, e.g., a video [92]. A vocabulary determines the set of terms that may be used to represent information objects. The vocabulary may be controlled or uncontrolled. Examples of controlled vocabularies are subject heading lists and thesauri that define a particular set of keywords that may serve as representations of information objects. An uncontrolled vocabulary allows the use of natural language terms. Controlled vocabulary terms are generally manually assigned while an index created automatically from the terms within a document will result in an uncontrolled vocabulary.

The creator of a vocabulary needs to decide how he/she represents the information objects in a collection and which terms to use. Terms should be chosen in such a way that a user with a particular information need is likely to use the same terms to request a particular information object as the vocabulary creator used to represent it. On the one hand, an uncontrolled vocabulary provides the most flexibility in representing information objects but also introduces ambiguity [84, 314]. On the other hand, the terms from a controlled vocabulary have high precision, but may not match a user's expectation of how information objects in a collection are represented. A technique to improve precision is to use more specific terms to represent information objects, e.g., to use phrases or sequences of terms [131]. Other techniques focus on improving recall. For example, when information objects have sparse descriptions it has been found that document expansion can help improve retrieval performance [316]. Using morphological variants of terms [177] or applying stemming [156] are techniques also often applied to improve recall. In order to efficiently access information objects an inverted index is used. An index is a specific data structure that allows for fast lookup of documents that contain specific terms or phrases. It provides a mapping from the vocabulary of terms used to represent information objects to the actual information objects themselves [22].

Representations of information objects are rarely unstructured, e.g., web pages are build using HTML markup, books have chapters and sections, and in metadata descriptions a distinction is made between information in different parts of the document (fields). Indexing of documents that contain some kind of structure as separate fields allows users and retrieval methods to focus on the relevance of a particular part of a document. One difficulty with nested structures, e.g., such as present in XML documents, is that one has to choose between indexing individual elements or collapsing elements. By indexing individual elements context is lost, but collapsing multiple elements dilutes the influence of a particular element [22].

### 6.1.2 Information Needs

Analogous to information objects, information needs (cf. §2.2) require a (machine-interpretable) representation that is compatible with the representation of the information objects. Taylor [317] defined four levels of information need: (i) the actual, but unexpressed, need for information; (ii) the conscious within-brain description of the need; (iii) the formal statement of the question; and (iv) the question as presented to the information system. In information retrieval the focus is on the last level of information need, i.e., the information need as presented to a system.

There are various ways in which an information need can be expressed, but generally this happens in the form of a number of keywords, i.e., a query. Queries as formulated by users have been found to be short with the majority of queries consisting of one or two keywords [179]. This representation is not necessarily adequate to represent a user's information need. To obtain a more accurate representation of a user's information need various techniques to perform query expansion are used. In a relevance feedback system terms are selected from documents based on relevance judgements, e.g., as obtained through clicks in a previous search iteration, and used for expansion [284]. In the lack of user feedback, one might assume that the top ranked documents are relevant. This technique is known as blind relevance feedback. It uses terms from the top ranked documents for query expansion and was shown to be effective when the top ranked documents are relevant [1, 248].

Other techniques do not assume a relevance feedback scenario. Instead, they identify clusters of related terms based on term co-occurrence data and expand a query with the clusters that contain a query term [207, 247]. Another technique is thesaurus-based query expansion where terms with similar meanings are mapped to a thesaurus term [275].

Structured query languages such as common in database systems provide an alternative way to express queries, but are less prominent in an information retrieval setting as the semantics of the structure contained in documents is not as well defined. An early example of an SQL-like query language for the Web is W3QS that allowed users to specify the type of a document to search and constraints on hyperlinks [195]. Through use of information extraction techniques some structure can be imposed on web documents [83], however, these techniques suffer from a lack of coverage. Some collections contain documents with (semi-)structured data such as information objects with metadata held in archives and libraries. This structure allows users to query specific metadata fields and hence increase the control over the results that are returned [101]. One difficulty with structured query languages is that users have to be familiar with the underlying structure. A way to aid users in formulating structured queries is through using specialized interfaces with query assistance services [40]. Instead of providing a structured query, users might provide structural hints to supplement a keyword query. For example, by specifying a date range, a target category, or entities (see Chapter 8 and 9).

Zloof [384] suggested a query technique based on examples. Query by example, also referred to as find similar or related article finding, allows a user to find additional information objects by presenting a search system with an example of a relevant information object. For example, Smucker and Allan [308] found that in simulating browsing patterns using find-similar achieved improvements in precision over patterns that do not. Lin and Wilbur [213] also found improvements in terms of precision over a probabilistic retrieval model (BM25) on a task to find articles that discuss the same topic in a collection of medical literature.

### 6.1.3 Retrieval Models

Given a representation of a user's information need, i.e., a query, and representations of the information objects the retrieval model determines the relevance of the information objects to the query in order to obtain a ranking of the information objects. For text-based retrieval three types of classic models are distinguished: *the boolean model*, *the vector space model*, and *the probabilistic model*.

The classic boolean retrieval model assumes that the query and document representations are sets of terms. A query is specified as a boolean expression and matching is based on whether the documents contain the query terms. The advantage of the model is that the matching principle is transparent and intuitive [22]. Disadvantages, however, are that it is hard to differentiate between documents of different quality and to impose an ordering on the result set. In later versions of the boolean model term weighting was introduced to overcome some of these issues [349].

The vector space model was first used by Salton [293] in the SMART system, an information retrieval system which was used for the development and experimentation with various retrieval models. In the vector space model documents and queries are represented by vectors of weighted terms. The matching is based on the similarity between document and query vectors and these similarity values are used to obtain a ranking of results [295]. One effective term weighting scheme is TFIDF. It is a combination of the term frequency (TF) of a term in a document, which promotes frequent terms, and the inverse document frequency (IDF) of the term in the collection, which promotes terms that are rare. We discuss this model in more detail in Chapter 7, where we experiment with linking records from different archives based on whether the records discuss the same or related events.

The probabilistic model was first proposed by Robertson and Spärck-Jones [285] as an alternative retrieval approach based on a probabilistic framework. The central assumption in the probabilistic model is the probability ranking principle (PRP) that states: "given a user query $q$ and a document $d_j$ in the collection, the probabilistic model tries to estimate the probability that the user will find the document $d_j$ interesting (i.e., relevant). The model assumes that this probability of relevance depends on the query and the document representations only, i.e., on the information available to the system. Further, the model assumes that there is a subset of all documents which the user prefers as the answer set for the query $q$. Such an ideal answer set is labeled R and should maximize the overall probability of relevance to the user. Documents in the set R are predicted to be relevant to the query documents not in the set are predicted to be non-relevant" [283]. Two issues with the PRP are: (i) it does not consider information outside of the system; and (ii) it relies on relevance judgements to estimate the probability of relevance of a document to a query, which are not available for new queries. In the absence of relevance judgements the probabilistic model reduces to a similarity function based on IDF term weighting. The lack of a term frequency weighting component or document length normalization were addressed in an extension of the probabilistic model, i.e., BM25 [282].

There are several other families of probabilistic models in IR such as language models [273], divergence from randomness [7], and Bayesian networks [334]. Language models capture regularities in language usage by modeling the distribution of linguistic units, e.g., words. Language models were first used in speech recognition to predict the likelihood of the next unit in a sequence of previous units. In information retrieval language models are used to capture the language usage in documents and to predict the likelihood that a document would generate a query. We discuss language models and their estimation in more detail in Chapter 8 and 9.

The advantage of Bayesian networks is that other types of evidence can be taken into account such as term co-occurrence, user clicks and location, to decide on the relevance of a document, next to the current query. We use this formalism in Chapter 9 to develop models for related entity finding.

### 6.1.4 Evaluation

An important part of IR research revolves around evaluation of retrieval models and the various ways of representing information needs and information objects. IR systems are evaluated based on their ability to return relevant information objects in response to various information needs. There are various definitions of relevance in IR. Mizzaro [249] defined each relevance as a point in a four dimensional space, where the four dimensions are: (i) a manifestation of the information source, i.e., a document representation, an actual document, or a piece of information; (ii) a manifestation of the information need, i.e., the real information need, perceived information need, a formulation of a perceived information need into a request, and translation of a request into a query to an IR system; (iii) time, i.e., the change of relevance over time; and (iv) components, i.e., the relevance with respect to a certain task, topic, and context. For example, in Chapter 7, where we investigate a method to link archival records from a news archive to a television archive based on events, relevance has been judged based on representations of records and not on the actual content. In this case only the representation is used as the IR system does not have access to the content of the records, e.g., video data, and the representation is assumed to be adequate. In Chapter 8 and 9 relevance is judged based on topicality, i.e., the relevance of a representation of a document to a request. This is the type of relevance generally used to evaluate IR systems.

Topical relevance was at the heart of the Cranfield experiments that were conducted in the 1960s and resulted in the Cranfield paradigm, which prescribes a methodology for the evaluation of information retrieval systems [95]. It assumes that the same set of documents and queries can be used to evaluate different IR systems once relevance judgements for all documents in relation to these queries have been obtained. A collection of documents with associated information needs and relevance judgements is referred to as a reference collection. Early reference collections were rather small, e.g., the Cranfield collection (1.2K documents) and communications of ACM collection (3.2K documents). A problem with this methodology arises when document collections become larger and judging all documents for relevance to a set of queries becomes unfeasible.

This changed in 1992 when the Text Retrieval Conference (TREC) was started as part of the TIPSTER Text program, co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense. The goal of TREC is to provide a common evaluation platform for IR systems Harman [157]. To be able to create judgements for larger collections a method called pooling is used where only the top k results for various systems are collected and assessed for relevance. TREC follows a particular yearly cycle to create reference collections: (i) organizers provide a collection of documents and a list of questions (topics); (ii) participating teams run their IR systems on the topics and return results in the form of a list of ranked documents; (iii) human assessors judge the documents in the result list on relevance to the topics and the systems are evaluated based on these judgements; and (iv) NIST organizes the TREC workshop where participating teams exchange ideas about implementation and approaches. In later years the focus of TREC changed from the evaluation of document retrieval in general to the creation of reference collections for more specific types of retrieval tasks. Some of the tasks considered are legal document retrieval, chemical document retrieval, question answering and expert finding. Several other evaluation campaigns have since emerged to address different retrieval challenges such as the Initiative for the Evaluation of XML Retrieval (INEX), which is focused on XML retrieval, and Conference and Labs of the Evaluation Forum (CLEF) (formerly known as the Cross-Language Evaluation Forum), which was originally directed at multilingual aspects of retrieval.

## 6.2 Beyond Document Retrieval

Up to this point we have discussed work that considers individual documents as the unit of retrieval. Documents, however, are only the carrier of what users are generally after, i.e., information. For example, someone interested in learning about information retrieval may ask: “which books are a good starting points to learn about information retrieval?” An answer to this question would be a list of recommended book titles instead of a list of documents that may or may not contain references to these book titles. Alternatively, someone may not be interested in just a single document but a set of documents related to a particular topic. For example, an infrequent news reader that would like to catch up on the current state of the economy requires a selection of articles that are non-redundant. Such a scenario violates the assumption of the PRP that the relevance of a document is independent of other documents [150].

Since the 1990s an increasing number of Cranfield style evaluation campaigns have emerged within the IR and natural language processing (NLP) communities aimed at creating reference collections for tasks that go beyond document retrieval such as news summarization, question answering, and entity retrieval. This has resulted in a myriad of tasks each with variations in the setup of the task such as the input to the system, e.g., keywords or examples, and the type of corpus operated on, e.g., web pages or structured data.

Table 6.1 provides an overview of the major evaluation campaigns and some of the tasks they have put forward. We do not review the work aimed at addressing tasks at the MUC, the ACE program, the TAC, and the CoNLL as these forums focus on computationally intensive NLP techniques to extract information in structured form from relatively small collections of well written documents. We refer the reader to overview papers of the respective campaigns for further details [51, 111, 112, 121, 152, 154, 226, 234, 322, 323]. We also forgo a discussion on work on multilingual retrieval or retrieval in other languages than English as pursued at CLEF, the NTCIR, and the FIRE, see [17, 144, 185, 227, 270] for more on these campaigns.

Instead we focus on tasks from the TDT campaign, INEX, TREC, and SemSearch as far as they provide an IR setting, go beyond document retrieval, and/or operate on struc-

**Table 6.1:** Overview of major NLP and IR evaluation forums that include English collections.

| Forum | Summary |
|---|---|
| MUC [154] | Originally focused on IE technology to extract information from military messages. Later editions included tasks such as template filling and co-reference on a corpus of Wall Street Journal articles. Ran from 1987–1997 and paved the way for ACE. |
| ACE [121] | Ran from 1999–2008 and aimed at further improving IE technology. Objectives included the detection and characterization of entities, relations, and events. Focused on computationally intensive NLP techniques on smaller collections (500K words) of broadcast transcripts, newswire and newspaper data. Later merged with TAC. |
| TAC [51, 111, 112, 234] | Started in 2008 to encourage research on particular sub-problems in NLP, i.e., question answering, textual entailment, summarization, and knowledge base population. Data included blogs [226] and newswire articles [152]. |
| CoNLL [322, 323] | Investigates machine learning methods applied to NLP tasks. Since 1999, each year a shared task is organized to compare approaches to a certain NLP problem. |
| TDT [2] | A forum running from 1997–2004 for the development of topic detection and tracking techniques, e.g., story segmentation. first story detection, and link detection. |
| TREC [346] | Explores various sub-problems in IR organized in tracks. Some dedicated to various aspects of document retrieval, e.g., the Web Track. Others focus on special document collections, e.g., the Legal, Medical, and Microblog Tracks. Several tracks have investigated interaction, i.e., Interactive, Relevance Feedback, and Session Track. Finally, some tracks investigate tasks that go beyond document retrieval, e.g,. the Enterprise, Question Answering, and Entity Tracks. |
| CLEF [17, 144, 270] | Encourages research on retrieval systems for multilingual, multimodal, and structured data. Examples of tracks are: GeoCLEF, focused on geographic IR; WePS, focused on people search; and CHiC, focused on cultural heritage data. |
| INEX [143] | Explores evaluation of focused retrieval tasks that go beyond documents, e.g., in the Book and Entity tracks. Other tasks focus on structured data, e.g., in the Link the Wiki and Linked Data tracks. |
| SemSearch [327] | A workshop series aimed a evaluating IR approaches to retrieval tasks on Linked Data that organized two entity search tasks in 2010 and 2011. |
| QALD [219] | A series of challengese aimed to answer natural language questions (e.g., "Who is the mayor of Berlin?") using Linked Data sources, i.e., DBpedia and MusicBrainz. |
| NTCIR [185] | Started in 1997 and with goals similar to TREC but aimed at enhancing IR research for Asian languages. |
| FIRE [227] | Information retrieval evaluation forum like TREC but for Indian languages. |

tured data collections. One type of task, that we consider in Section 6.2.1, asks systems to recommend, link, or group documents related to the same concept, e.g., TDT Link Detection tasks. Systems that address this type of task could support users in gathering related material and is relevant to our work on linking archives in Chapter 7.

A second type of task, discussed in Section 6.2.2 and 6.2.3, asks systems to return a specific bit of information instead of documents such as the question answering, expert search, and entity search tasks at TREC and INEX. Such systems would support identifying the concepts and relationships that exist within the domain of a research topic and motivate our work on entity oriented search tasks in Chapter 8 and 9.

Finally, in Section 6.3 we discuss work on tasks in which systems make use of or link to structured data such as seen at the INEX Link the Wiki, SemSearch Adhoc Object Retrieval, and TREC Entity List Completion (ELC) tasks, which are specifically related to our work in Chapter 8. Table 6.2 provides an overview of the tasks we consider in Chapter 6, 7, 8, and 9.

**Table 6.2:** Overview of the tasks considered in Chapter 6, 7, 8, and 9. For each task we specify the input, output, data collection used, and forum that organized the task.

| Task | Input | Output | Collection | Forum |
|---|---|---|---|---|
| topic tracking | documents about a topic up to a certain point in time | subsequent documents judged as on topic or not | newswire; broadcast news | TDT [135] |
| link detection | pair of documents each about an event | whether the documents are about the same event | | |
| question answering | a natural language question | 5 answer strings each with an answer to the question | TREC disks 1–5 | TREC [348] |
| expert search | a description of an area of expertise | a list of experts in that area | W3C corpus | TREC [103] |
| related entity finding | source entity, the type of the target entities, and a relationship description | homepages of entities that satisfy the relationship and target type constraint | Clueweb09 | TREC [29] |
| entity search | keyword query representing an entity | URI representation of the entity | BTC2009 | Sem-Search [274] |
| list search | natural language description of a group of entities | group of entities that match the description | | |
| entity ranking | description of a group of entities and a set of Wikipedia categories | set of entities that satisfy the constraints | Wikipedia | INEX [114] |
| entity list completion | description of a group of entities and a set of examples entities | set of entities that satisfy the constraints | | |
| entity linking | a Wikipedia page | links from phrases in the text to Wikipedia pages | | INEX [171] |

### 6.2.1 Related Article Finding

Related article finding also referred to as *find similar*, *query by example*, and *more like this*, is the general task of returning additional documents related to an initial document about a certain topic. It shares similarities with relevance feedback. However, in related article finding a user query is not necessarily involved. Recommender systems also provide this type of functionality using approaches based on collaborative filtering, i.e., overlap between user interests, and information filtering, i.e., finding items with similar characteristics as previous items a user liked [280]. We focus on information filtering approaches within the context of news archives.

Topic Detection and Tracking was a multi-site research project to develop techniques for news summarization systems that ran from 1997 to 2004. A number of techniques were studied to enable finding topically related material in streams of data such as topic segmentation, detection, and tracking. Most relevant to the work in Chapter 7 are techniques for the topic tracking and the link detection tasks. In the topic tracking task, where given a document about a certain news topic, the goal is to find other documents that discuss the same seminal event or related events [2].

In the link detection task the goal is to detect whether a pair of documents discuss the same topic. Here, a *seminal event* is a high impact news event that generates follow-up events, and a *related event* is caused or predicts the seminal event but is not a seminal event by itself. Topic tracking may be done within a collection consisting of a single news source [138] or a collection consisting of various sources such as newswire text, broadcast news speech transcriptions, closed captioning, and machine translation results [276, 377]. Some approaches in topic detection and tracking focus on detecting novelty and redundancy using language models [377]. Others support new event detection by using an adaptation of the vector space model and assigning more importance to named entities [199].

With respect to event linking, expansion on both the initial document and candidate document side by selecting representative terms from the documents most relevant to those documents was found to be especially effective [205]. Further, it has been found that when faced with very short documents (i.e., documents with sparse text), *document expansion* can help improve retrieval performance [316]. Document expansion refers to combining text from related documents with the text of an original document. The focus of work on document expansion, however, has been in the context of traditional document search [120]. Document expansion is one of the techniques further investigated to enable cross-archive linking of records based on events in Chapter 7.

In a more general context of finding related material across news sources work has been done on linking news articles to blog posts. It was found that using the structure of news articles (title, lead, body, etc.) to model a query helps in identifying related blog posts [328]. An early paper on the topic of cross-media linking investigates generating connections between news photos, videos, and text on the basis of dates and named entities present in texts associated with the items [87]. Ma et al. [225] investigated cross-media news content retrieval to provide complementary news information. This was done on the basis of news articles and closed captions from news broadcasts, and focused on differences in topic structure in the captions to find complementary news articles for broadcasts. Also relevant is work on linking passages from the closed captioning of television news broadcasts to online news articles [167]. Here, the focus was on the time-based aspect of identifying articles about the news subject being discussed at any particular point in time. An interesting finding was that term selection was valuable in identifying the correct relevant articles. We investigate the effects of term selection as well as document expansion and utilizing document structure further in Chapter 7.

### 6.2.2 Focused Retrieval

Traditional information retrieval returns documents in response to an information need and leaves the task of locating relevant information within those documents to users [184]. The aim of focused retrieval is to return a particular piece of information to the user. We discuss three areas of work related to the entity oriented search tasks addressed in Chapter 8 and 9: (i) question answering, where entities may be returned as answers; (ii) expert search, which is a specific type of entity search; and (iii) entity retrieval.

#### Question Answering

The goal of question answering (QA) is to return answers, rather than documents containing answers, in response to a question, e.g., factoid questions and list questions [345]. Factoid questions require a single answer in the form of a snippet, e.g., "550 calories" in response to the query: "How many calories are there in a Big Mac?" To answer list questions, systems have to return instances of the class of entities that match the description in the question, e.g., "What are 12 types of clams." A prototypical pipeline for a QA system consists of four components: query analysis, document retrieval, document analysis and answer selection [250]. In the query analysis component the question is analyzed and classified into a certain class of questions, e.g., a "who" or "what" question. Next, the document retrieval component identifies documents that are relevant to the questions and the document analysis component selects candidate answers from these documents. The candidate answers are send to the answer selection component which selects the most likely answer from the candidates. An issue with most systems, however, is the use of knowledge intensive techniques for these components [93, 200], which makes them unsuited for efficient processing of large volumes of data. An alternative lightweight approach exploits the redundancy of the web to extract answer strings. The fact that answers occur multiple times in different forms allow for the creation of simple answer selection patterns based on the question [125].

#### Expert Search

The TREC 2005–2008 Enterprise track [31] focused on answering a more specific type of question, i.e., finding experts on a particular topic. Two important families of retrieval models for expert finding have emerged: *candidate-centric* models that first compile a textual representation of candidate experts by aggregating the documents associated with them and then rank these representations with respect to the topic for which experts are being sought; and *document-centric* models that start by ranking documents with respect to their relevance to the query and then rank candidate experts depending on the strength of their association with the top ranked documents [23, 25]. Many variations on these models have been examined, for a range of expertise retrieval tasks, exploring such features as proximity [26, 269], document priors [383], expert-document associations [24] external evidence [303], and co-occurrence [85]. While expert finding focused on a single entity type ("person") and a specific relation ("expert in"), the proposed methods are typically not limited to these relations. Therefore, most of the approaches devised for expert finding have also been applied to the more general task of entity search. A comprehensive overview of expertise retrieval is given in Balog et al. [37]

### 6.2.3 Entity Retrieval

The roots of entity retrieval go back to natural language processing, specifically to IE. One typical IE task is to find all entities for a certain class, for example “cities.” The general approach taken uses context based patterns to extract entities; e.g., “cities such as * and *,” either learned from examples [281] or created manually [165].

A range of commercial providers now support entity oriented search, dealing with a broad range of entity types: people, companies, services and locations. Examples include TextMap,[1] ZoomInfo,[2] Evri,[3] and the Yahoo! correlator demo.[4] They differ in their data sources, in the entity types they support, functionality, and user interface. Common to them, however, is their ability to rank entities with respect to a topic or to another entity. In web search engines it is also increasingly common to present users with information about a specific entity or a list of entities for specific queries, e.g., “Chinese restaurants in Amsterdam.” Little is known, however, about the algorithms underlying these applications.

Early work by Conrad and Utt [100] introduced techniques for extracting entities and identifying relationships between entities in large, free-text databases. The degree of association between entities is based on the number of co-occurrences within a fixed window size. A more general approach is also proposed, where all paragraphs containing a mention of an entity are collapsed into a single pseudo document. Raghavan et al. [277] re-state this approach in a language modeling framework and use the contextual language around entities to create a document-style representation, that is, entity language model, for each entity. This representation is then used for a variety of tasks: fact-based question answering, classification into predefined categories, and clustering and selecting keywords to describe the relationship between similar entities. Sayyadian et al. [299] introduce the problem of finding missing information about a real-world entity from text and structured data. Results show that entity retrieval over text documents can be significantly aided by the availability of structured data, e.g., Google’s Knowledge Graph (cf. §6.3.3).

#### Entity Retrieval at INEX

Since 2006 INEX features an entity track that consists of two tasks: entity ranking and list completion. In the entity ranking task the input question consists of a query (“Paul Austen novels”) and a Wikipedia category (“Novels”); the expected output is a ranked list of Wikipedia pages representing relevant entities. The list completion task is similar but takes as extra input a number of example entities [114, 116]. These tasks extended earlier work on entity retrieval in Wikipedia that only considers the surface forms of entities and does not attempt to use documents associated with those entities or category information [374]. Fissaha Adafre et al. [136] addressed an early version of the entity ranking and list completion tasks and explored different representations of list descriptions and example entities (using textual descriptions plus descriptions of related entities); other

[1] http://www.textmap.com/

[2] http://www.zoominfo.com/

[3] http://www.evri.com/

[4] http://sandbox.yahoo.com/correlator

early work on the topic is due to Vercoustre et al. [342]. Most approaches to the entity ranking task use the query to retrieve relevant entities (Wikipedia pages) and use the categories as a filter [32]. The Wikipedia category structure, however, is not a strict hierarchy making expansion of the input category necessary. Vercoustre et al. [342] expand with categories that match the terms in the query, while others [178, 330, 352] use the category structure to expand categories.

There is a wide range of approaches looking for evidence in other documents, that is, besides the Wikipedia page corresponding to the entity. Both Zhu et al. [382] and Jiang et al. [182] employ a co-occurrence model, which takes into account the co-occurrence of the entity and query terms (or example entities) in other documents, by borrowing methods from the expert finding domain ([381] and [25], respectively). Many entity ranking approaches utilize the link structure of Wikipedia, e.g., as link priors [186] or using random walks to model multi-step relevance propagation between linked entities [329]. Fissaha Adafre et al. [136] use a co-citation based approach; independently, Pehcevski et al. [267] expand upon a co-citation approach and exploit link co-occurrences to improve the effectiveness of entity ranking. Likewise, Kamps and Koolen [183] show that if link-based evidence is made sensitive to local contexts, retrieval effectiveness can be improved significantly.

In the list completion task several participants use the categories of example entities for constructing or expanding the set of target categories, using various expansion techniques [104, 178, 182, 341, 352, 382]; some use category information to expand the term-based model, see, e.g., [186, 352]. Balog et al. [35] introduce a probabilistic framework that can incorporate most of the approaches for both tasks and show the effectiveness of (blind) relevance feedback on the entity ranking and list completion tasks. A challenge in these tasks is to find parameters for weighting the query, category and example entity information. A commercial service, Google Sets, also tackles the list question task. Here, a user provides a number of examples of a class as input and the service retrieves entities that closely match the examples [149]. Ghahramani and Heller [145] developed an algorithm for completing a list based on examples using machine learning techniques.

Another method combines the two approaches using a linear combination, where the mixing parameter depends on the difficulty of a topic [340]. A model is trained to predict each topic's difficulty and the combination weight is set accordingly. We investigate a variant of the entity list completion task in Chapter 8 in the context of the Linking Open Data cloud. We also propose a query dependent method that combines text-based retrieval with additional structure, but differ from supervised machine learning based approaches [304] in that our method does not require any training data. Moreover, machine learning based approaches do not necessarily outperform unsupervised approaches in this setting [141].

### Entity Retrieval at TREC

A major development in evaluating entity oriented search was the introduction of the Entity track at TREC in 2009 with the aim of performing entity oriented search tasks on the web [29]. The first edition featured the Related Entity Finding (REF) task: given a source entity, a relation and a target type, identify home pages of target entities that enjoy the specified relation with the source entity and that satisfy the target type constraint [29]. For example, for a source entity ("Michael Schumacher"), a relation ("Michael's teammates while he was racing in Formula 1") and a target type ("people") return entities such as "Eddie Irvine" and "Felipe Massa."

A general pipeline for REF is to first collect documents or text snippets from the collection that are relevant to the REF query, next obtain entities by performing named entity recognition on the snippets, implement some sort of ranking step and finally find home pages. A popular method to rank entities is to use a language modeling approach, where the entity model is constructed from snippets containing the entity and the query is the relation [369, 371, 378]. Mccreadie et al. [232] present a successful adaptation of the voting model for people search to REF. Fang et al. [132] use a hierarchical relevance retrieval model that linearly combines relevance scores of the query given either the document, a passage or an entity. They further improve their model by exploiting list structures, training logistic regression models for type filtering and applying several heuristic filtering and pattern matching rules. Zhai et al. [376] propose a probabilistic framework to estimate the probability of an entity given a REF query, with two components: the probability of the relation given an entity and source entity, and the probability of an entity given the source entity and target type.

The dataset used for the REF task is the ClueWeb09 Category A web crawl [96], consisting of about 500 million documents, including English Wikipedia. The size of the corpus limits the complexity of document analysis and entity extraction techniques that can be used. As a result a number of approaches rely heavily on Wikipedia, i.e., as a repository of entity names, to perform entity type filtering based on categories, and to find home pages through external links [187, 232, 302]. We will investigate methods to address the challenge of performing entity search on a web corpus and the effectiveness of various heuristics commonly applied in Chapter 9.

## 6.3 Semantic Search

The information retrieval models discussed up to this point operate on strings and have a limited understanding of what concepts these strings refer to. For example, "Michael Jackson" could refer to a famous pop star or a computer scientist. Moreover when we as humans think of a person, we know that he/she has certain attributes, i.e., was born on a particular day and lives in a certain place, while this knowledge is generally unavailable to retrieval models. The aim of semantic search is to use Semantic Web technologies to improve traditional web searching [155, 228]. Below we first describe methods to map strings to concepts. We then discuss linked data and the Semantic Web initiative to create a web of data. Finally, we detail some of the approaches to entity search in the web of data.

### 6.3.1 From Strings to Concepts

Named Entity Normalization (NEN) is the task of finding an unambiguous referent for a string representing an entity. Entity normalization has a long tradition in the context of databases, where it is also known as record linkage [366]. The structured way in which entities are represented in databases allows entities to be normalized if they have more than a certain number of relations, or type of relations, in common with another entity in the database [49]. For example, when combining databases of customer data of two companies a person that is registered with both may be matched based on name and telephone number even when the address data does (fully) not match.

Pure string-distance based matching methods [97] are not enough to reliably normalize entities. Köpcke et al. [196] characterize eleven entity matching frameworks for structured and semi-structured data in terms of four criteria: (i) the entity data considered, i.e., XML or relational; (ii) the blocking method used, i.e., clustering candidate entities together in order to reduce the matching space; (iii) the matching function, i.e., based on attributes and or context; and (iv) whether the framework uses examples for training.

To be able to do entity normalization it is necessary to have a knowledge source with referents against which entities can be compared. In a web setting it is difficult to find a referents list that provides enough coverage of entities found on the web. Often, knowledge bases such as DBpedia[5] or Freebase[6] are used.

The INEX Link the Wiki Track introduced the task of finding phrases in Wikipedia pages that are likely to be anchor text and link those to their corresponding Wikipedia page [139]. Wikipedia has since been heavily used in entity normalization approaches. One successful approach to identify candidate phrases was based on how likely a phrase was to be used as an anchor in Wikipedia [244]. Milne and Witten [246] identified the utility of using unambiguous entities from the surrounding context of a surface form of an entity on a Wikpedia page as features to normalize entity surface forms. Cucerzan [106] showed that Wikipedia data can be used to derive context models to disambiguate entities in news articles. He et al. [161] found how previous Wikipedia based linking approaches failed when applied to linking medical terms from radiology reports to Wikipedia due to the frequent occurrences of modifiers and conjunction in noun phrases. They showed that identifying sub-sequences is the key to resolving this issue. Meij and de Rijke [236] addressed the challenge of linking entities from queries to DBpedia. This work was later extended to also identify and link entities in tweets [239].

### 6.3.2 The Web of Data

The Semantic Web initiative is a movement that aims to turn the unstructured web of documents into a web of data that is understandable by machines [13]. In order to realize the Web of Data web pages need to be associated with metadata. To this end the World Wide Web Consortium (W3C), an organization responsible for standards on the Web, has proposed standards for metadata annotation of web pages such as XML and the Resource Description Framework (RDF).

Providing an XML document that for each web page provides its information in structured form is one solution to this problem. For example, a HTML page with the time table of the university library opening hours could be accompanied by an XML document that offers this data in structured form. Another approach is to mark up HTML documents with microformats to identify strings as a specific type of object, e.g., a location or a date.

[5] `http://dbpedia.org/About`

[6] `http://www.freebase.com/`

Microformats are gaining popularity as the markup is directly applied to the web document and does not require an additional metadata document next to the web page [193].

A disadvantage of these markup approaches is that relations between objects can only be expressed through nesting of elements and that the interpretation associated with a particular way of nesting is up to the application. RDF has been suggested as a data model for describing relations between objects. RDF allows statements to be made about a domain in terms of triples. An RDF triple consists of a subject, a predicate, and an object. A subject is always a URI and represents a thing. Subject URIs serve as unique identifiers for entities. An object is either a URI referring to another thing or a string holding a literal value. Predicates are also always URIs and specify the relations between subjects and objects. The RDF Schema defines the classes of things and the predicates that exist in a certain domain as well as the relations between classes. It defines the semantics of being an object of a particular class in terms of relations to other classes in a domain, e.g., a person has a date of birth. The definition of classes and their organization in hierarchies in the RDF Schema allows inference mechanisms to derive new information. Although the expressiveness of RDF(Schema) is rather limited, it allows a reasoner to derive information such as that an object in a sub-class of class A will have the properties associated with objects of class A. We do not further consider this aspect of RDF(Schema) as reasoning over large collections of RDF data remains impractical and instead investigate methods that do not depend on reasoning.

Linked Open Data[7] is data that is published openly on the web following the linked data principles: (i) use URIs to identify "things"; (ii) make URIs dereferenceable; (iii) attach "useful" information to URIs; and (iv) link URIs to other URIs. Linked Data is typically represented using the RDF format. Many organizations and companies have published their data as linked data resulting in the so called Linking Open Data (LOD) cloud.[8] At the center of the LOD cloud is DBpedia as it provides information spanning multiple domains and connects data from different domains into a single data space [18, 54]. A problem with the LOD cloud is that data quality control is based on trust and left to the person publishing his/her data. As the number of contributers to the LOD cloud increases so does the number of RDF Schemas used to define the classes and properties of datasets. Appropriately linking a new data set to objects already present in the LOD cloud is becoming increasingly difficult [4]. This is also complicating the development of applications that use linked data as each schema introduces new requirements. In Chapter 8 we will investigate how to overcome this challenge in an entity search application.

### 6.3.3 Entity Search in the Web of Data

Using structured data to provide contextual information about entities is gaining popularity as evidenced by the recent introduction of Google's *knowledge graph* (KG). The KG is used in various applications, e.g., to present a *KG biography*[9] or a *KG image carousel*[10] with additional information related to a query. Given a query, e.g., Asta Nielsen, the KG

[7] `http://linkeddata.org`

[8] `http://lod-cloud.net/`

[9] `http://googleblog.blogspot.nl/2012/05/introducing-knowledge-graph-things-not.html`

[10] `https://plus.google.com/+google/posts/KpsbyvHUotN`

biography will display images of her, a short summary describing her as a 1910 Danish movie star, and some facts such as her date of birth. However, the concepts and facts that are presented are limited to the information available in sources such as Wikipedia and Freebase. The KG image carousel provides a set of images of related concepts given a particular concept, e.g., actors in a film or books by an author. The carousel only appears, however, for a specific set of relations such as books by author X or actors in movie Y and when a certain popularity threshold is met.[11] Furthermore, little is known about the methods behind these applications.

A traditional way of obtaining information about an object from collections of linked data is through structured query languages such as SPARQL that express queries through constraints on relations (*links*) between URIs. These languages, however, are difficult to use and require knowledge of the underlying ontologies. More recent user-oriented approaches address this issue by automatically mapping keyword queries to structured queries [326, 380]. A number of services provide keyword based interfaces to search in Linked Data for URIs of entities [55]. Other approaches use keyword queries against a free text index of Linked Data [274, 331].

Hybrid approaches to ranking entity URIs exploit the link structure and textual information contained in Linked Data. For example, one approach returns both URIs that contain query terms as well as URIs that link to those URIs [286]. Yet others propose a combination of structured and keyword-based retrieval methods [34, 113]. Common to the text-based and hybrid approaches mentioned here is their focus on retrieving URIs for entities given a name or a description.

A hybrid method able to retrieve entities that engage in a certain relation with another entity is proposed by Elbassuoni et al. [127]. This method uses a language modeling approach to construct exact, relaxed, and keyword augmented graph pattern queries. In order to estimate the language models, RDF triple occurrence counts and co-occurring keywords are extracted from a free text corpus.

### Adhoc Object Retrieval

As a first step towards evaluating these approach to search Linked Data the Semantic Search Workshop launched the “adhoc object retrieval task” [57, 274], which was focused on retrieving URIs for entities described by free text. Ad-hoc object retrieval differs from entity list completion in its focus on resolving entity names to URIs in the LOD cloud, instead of locating entities that stand in some specified relation. A popular approach to this task is to adapt standard retrieval models to operate on Linked Data. For example, by using a linear combination of the language model scores for different textual entity representations or applying a variant of the BM25F model that takes into account various statistics of the attributes in entity representations [56]. One promising approach is to combine both traditional retrieval techniques and structure-based queries in a hybrid system as demonstrated by Tonon et al. [325].

[11] http://searchengineland.com/googles-image-carousel-and-knowledge-graph-search-results-167007

### Entity List Completion

The TREC Entity track's variation on the ELC task extends the INEX ELC task in that entities are no longer Wikipedia pages but URIs in a sample of the LOD cloud. Approaches to this task where evaluated on a limited (8) number of topics. They include a text-based [133] method using a filtering approach based on WordNet and link-based methods [66, 110] using link overlap and set expansion techniques.

In the entity list completion task of the 2011 Semantic Search Challenge entities are represented by URIs as well, but no example entities are given and only a textual description of the common relation between the target entities is provided. Approaches to this task are predominantly text-based [36]. A notable exception is an approach that re-ranks an initially retrieved list of entities using spread-activation [94].

There are other unsupervised approaches to combining results, also known as late data fusion, that use different ways of weighting the scores from various result lists [137]. These, however, do not exploit features other than those available in the result lists, i.e., they do not consider example entities.

We investigate structure-based, text-based, and hybrid approaches to the entity list completion task in Chapter 8. Different from other work we do not consider a fixed set of examples: instead, we study the effect of varying the composition of the set of example entities on the retrieval performance of these methods.

# 7 Linking Archives Using Document Enrichment and Term Selection

Words are an example of the more general notion of signs, i.e., things we use to refer to our surroundings. As signs may originate within different situations their meaning varies [271]. Interpretation is the process of deriving the meaning of a sign and contextual information is the information necessary to establish the meaning of a sign at a particular time [339].

Users of archival records such as humanities researchers as well as the archivists responsible for maintaining collections of such records, have long recognized the importance of contextual information. Lytle [224] described the need for contextual information within archives as follows: "Items in isolation from an archival body lose part of their meaning; the reason for this is that the file, not information in the records alone, is related to activity." To this end archives often organize records according to the entity that created them, i.e., the provenance method [123, 224]. This type of organization, however, does not necessarily meet the needs of humanities researchers. For example, historians require information about how a record associated with a particular event relates to other events in the particular period under study to be able to determine its significance [123]. In Chapter 3 we observed a similar need for contextual information about events in the media studies research cycle, where researchers use newspapers to obtain reflections about events.

One way to support locating contextual material over (multiple) archives is to create links between individual records. On the web this is supported through hyperlinks between documents that allow users to move from one document to the next and to obtain background information about topic of interest. In an archival setting the creation of links has received little attention, likely due to the focus on annotation and preservation of provenance information, rather than on supporting browsing behavior.

We examine the linking problem in an archival setting, focusing on events. Here, we aim to connect a record from one archive to records in another archive that discuss the same or related events. Links to records describing the same event allow users to access different views of the same event, while links to records describing related events provide potential contextual information about the build-up and impact of events.

We focus on a specific instance of the task as introduced in Chapter 1: linking records from a newspaper archive with a rich textual representation to records from a multimedia archive that tend to have sparse annotations, see Figure 1.2. Our scenario is characterized by two somewhat complementary challenges. First, the targets of our linking task are multimedia records. Retrieval models, generally, do not have access to the actual media content of such records due to unresolved challenges in multimedia retrieval [208, 370]. Therefore, we investigate methods for document expansion to address the challenge of low recall introduced by the relatively sparse metadata representation of these records (cf. §6.2.1). Second, the source of a link in our task is a news article in a news archive. Such articles are textually rich and may contain interviews and debates next to accounts of an event. To alleviate the potential of irrelevant terms to give rise to a precision problem, we investigate methods for term selection (cf. §6.2.1). These two challenges motivate our fourth research question, which we recall from Chapter 1:

**RQ 4.** How can we automatically generate links from a record in a newspaper archive with a rich textual representation to records in a television archive that tend to have sparse textual representations?

**a.** Does expanding sparse record representations with text from other sources improve linking performance?

**b.** What effect does modeling reduced versions, e.g., by selecting informative terms of the original richly represented records from the source archive, have on linking performance?

We approach the linking task as a retrieval problem: given a source record, retrieve target records it should be linked to. To address our first research question we improve the representation of target records by enriching their sparse annotations with representations from the target archive, the source archive, and Wikipedia. To address our second research question we reduce the representation of the source record by selecting a subset of terms from the original representation, i.e., from the metadata as well as the content of articles.

The contributions of this chapter are three-fold: (i) we define and motivate a new task, i.e., linking archives based on events, and identify future directions; (ii) we report on a set of experiments investigating the effect of record representation along two dimensions; and (iii) we demonstrate the effectiveness of linking based on enrichment of sparsely annotated records with content from a different archive.

We describe our record enrichment and linking approaches in Section 7.1. In Section 7.2 we describe our data sets and experimental setup. We report the results of our experiments and provide a discussion in Section 7.3. We conclude in Section 7.4.

## 7.1 Approach

We formally define the variants of the linking archives tasks addressed in this chapter: *same event linking* and *related event linking*. Let $A_s = \{d_{s,1}, \ldots, d_{s,n}\}$ be a set of source archive records, and $A_g = \{d_{g,1}, \ldots, d_{g,m}\}$ be a set of target archive records. From now on we drop the $n$ and $m$ indices for clarity in our notation, but note that $d_s$ and $d_g$ denote individual records from the source and target archive respectively. Let $f_s(d_s, d_g)$ be a binary function that is true if $d_s$ and $d_g$ are about the same event. We define a similar function $f_r(d_s, d_g)$ that is true if $d_s$ and $d_g$ are about related events. The notions of same and related event will be defined in §7.2.

In the same event linking task we aim to identify all pairs $(d_s, d_g)$ for a given source record $d_s$ describing a certain event $e$, where each $d_g$ describes the *same* event $e$ as $d_g$. In the case of related event linking we aim to find all pairs $(d_s, d_g)$, where each target record $d_g$ describes an event $e'$ that is *related* to $e$, which is described by a given source record $d_s$.

A record is *textually rich* when its representation consists, on top of human-annotated metadata from a controlled vocabulary also textual content from an uncontrolled vocabulary (cf. §6.1.1). *Sparse* representations only contain human-annotated metadata. In the specific setting in which we are working, source records are news articles, hence textually rich, and target records are sparsely represented records in a video catalog (§7.2).

### Linking Model

In choosing a linking model we have two considerations. First, computing the functions $f_s(d_s, d_g)$ and $f_r(d_s, d_g)$ for all $(d_s, d_g)$ pairs is computationally expensive, and scales exponentially with the size of the data set. Second, if we base our binary linking decisions on a similarity measure then we require a threshold parameter that determines when records should be linked. Setting this threshold is a nontrivial as it varies per collection and type of linking task.

Instead of making binary decisions for pairs of elements, we model the task of linking archives as finding a ranked list of target records whose representation is most similar to the representation of the source record. For each source record we generate a ranked list of same/related events described by target records based on their similarity to the source record. This approach has a number of advantages. First, it is computationally more efficient. Second, a single model can be used for finding both same events and related events. Third, it doesn't involve a threshold parameter, instead, it presents a ranked list of link targets and allows the user to select which link to follow.

In our linking model we represent records from the source archive $(d_s)$ and target archive $(d_g)$ as vectors. Each dimension of our vectors relates to the weight $(w)$ of a term in the respective records, i.e., $\vec{d_s} = [w_{s,1}, \ldots, w_{s,n}]$ and $\vec{d_g} = [w_{g,1}, \ldots, w_{g,n}]$, where $n$ is the number of terms in the vocabulary. We use the vector space model [295] as our similarity function:

$$\mathrm{sim}(d_s, d_g) = \frac{\vec{d_s} \cdot \vec{d_g}}{|\vec{d_s}||\vec{d_g}|} = \frac{\sum_{i=1}^{n} w_{s,i} w_{g,i}}{\sqrt{\sum_{i=1}^{n} w_{s,i}^2}\sqrt{\sum_{i=1}^{n} w_{g,i}^2}}. \tag{7.1}$$

Here, the weight $w_d$ of each term $t$ in record $d$ is given by its TFIDF score:

$$\text{TFIDF}(t, d) = w_d = \frac{tf(t, d)}{|d|} \cdot \log\left(\frac{|D|}{\sum_{d \in D} \delta(t, d)}\right),$$

where $tf(t, d)$ gives the count of term $t$ in record $d$, $|d|$ is the length of a record, $|D|$ is the number of documents in the collection and $\delta(t, d)$ is a binary indicator function defined as:

$$\delta(t, d) \begin{cases} 0 & \text{if term } t \text{ occurs in } d \\ 1 & \text{otherwise.} \end{cases}$$

The term frequency (TF) component is given by the first factor, the inverse document frequency (IDF) is given by the second factor in Equation 7.1.

To obtain a ranked list of link targets we compute the similarity between a fixed source record $d_s$ and every potential target record in the target archive collection according to 7.1.

### Document Expansion

To address sparseness of the representation of a target record $d_g$ we use a set of additional records for expanding the representation of $d_g$. Below we consider multiple sources $A_x$ for the expansion records: the source archive $A_s$, the target archive $A_g$, and an external archive $A_e$. Let $d_{g_x}$ be an expanded representation for a target record. Given $d_g$, we obtain $d_{g_x}$ as follows. We compute the similarity between $d_g$ and each record $d_x$ in an expansion achieve $A_x$, using the same similarity function as defined in (7.1). Then, we obtain a set of expansion records $X_{d_g}$ by iteratively selecting records $d_x$ into $X_{d_g}$ such that

$$d_x = \arg\max_{d_x \in A_x} \mathrm{sim}(d_t, d_x), d_x \notin X_{d_t}.$$

We limit the size of $X_{d_g}$ with a threshold parameter on the number of expansion records and $\mathrm{sim}(.,.)$ is the same similarity function as in (7.1). We then obtain the expanded target record vector as $\vec{d}_{g_x} = [w_{g_x,1}, \ldots, w_{g_x,n}]$ where the weight for a term $t$ is defined as:

$$w_{gX} = \frac{tf(t, d_g) + \sum_{d_x \in X_{d_g}} tf(t, d_x)}{|d_g| + \sum_{d_x \in X_{d_g}} |d_x|} \cdot \log\left(\frac{|D|}{\sum_{d \in D} \max(\delta(t, d), \delta(t, X_d))}\right).$$

Here, $X_{d_g}$ is the set of expansion records for $d_g$. The total number of records $|D|$ remains the same, however, each record is associated with a set of expansion records. Therefore, to arrive at an inverse expanded record frequency (IDF) for a term $t$, we count an occurrence of $t$ when it occurs in the original records or any of its expansion records $X_d$ for each record in $D$.

### Selecting Representative Terms

Recall that our source records are textually rich. Although we apply TFIDF weighting in our similarity function, the high dimensionality of the documents may result in poor similarity values. To address the potential of topic drift that may result from textual richness, we investigate the effect of automatically selecting a reduced set of terms $R$ from the text associated with a source record $d_s$ (instead of using all terms) when ranking candidate target terms. For a source record $d_s$, we select a reduced set of terms $R$ by iteratively select the top $K$ terms from $d_s$ into $R$, according to their TFIDF scores:

$$t = \arg\max_{t \in V} \mathit{TFIDF}(t, d_s), t \notin R,$$

where $t$ is a term and $V$ is the set of terms in the vocabulary. Given the set of selected terms $R$ and a reduced record representation $\vec{d}_{s_R} = [w_{s_R,1}, \ldots, w_{s_R,n}]$, the weight of a term in the reduced record vectors is defined as:

$$w_{s_R} = \frac{tf(t,d_s) \cdot \delta(t,R)}{|\sum_{\{t' \in R : b(t',d_s)\}} tf(t',d_s)|} \cdot \log\left(\frac{|D|}{\sum_{d \in D} \delta(t,d) \cdot \delta(t,R)}\right).$$

Here, $\delta(t,R)$ results in 1 if term $t$ occurs in $R$ and 0 otherwise.

### Selecting representative entities

We experiment with the selection of representative terms by only considering named entities. Named entities are a special type of term found to be important in identifying related events [199]. To select entities we apply a named entity recognizer [134] based on conditional random fields to the content of all source archive records. We then select the top $K$ entities based on their TFIDF value as with the terms.

### Date Filter

Finally, we also examine the use of the date field present in the metadata of both source and target records. Not only is the date field one of the most consistently used fields in archival data, but it has also been shown that dates are useful when detecting same events [211]. We use a simple date filter that only allows a link from a source record $s$ to a target record $t$ if $t$'s date is within an $N$ day window around the date of $s$. That is, target archive records with a publication date within $s - \frac{N}{2}$ and $s + \frac{N}{2}$ days of $s$.

# 7.2 Experimental Setup

In this section we describe our experimental setup. We start by describing the collection used for evaluating the linking rich-to-sparse archive task, and follow with a description of our experiments with document expansion and term selection.

## 7.2.1 Evaluation Collection

Our evaluation collection consists of a source archive containing textually rich newspaper articles and a target archive of textually sparse television news broadcasts to which we want to link. We single out a set of source records as our test cases for linking, and for each test source record, we have a set of relevance judgments indicating which records in the target archive refer to (i) the same seminal event, and (ii) related events.

### Source Archive

Our source archive consists of 346,559 newspaper articles published by a Dutch newspaper, the NRC Handelsblad,[1] from 3 Jan. 2005 to 8 Jun. 2010. Each article consists of the article text (article title and body) and a series of metadata fields created by editors at the newspaper. These metadata fields comprise of the *persons*, *locations*, *organizations*, *events* and *keywords* that are the subject of an article. Rather than exploiting the specific

[1] http://www.nrc.nl/

structure of the metadata schema of NRC's archive, we combine all of the data from the metadata fields for an article together into a single representation; we refer to this aggregated set of fields as the *metadata* for a source archive record $s$. We refer to the article text as the *content* of $s$. On average, source record content has 409 terms and metadata has 8 terms, for a total of 417 terms per record.

### Target Archive

Our target archive in this chapter consists of 73,666 television news stories obtained from the Netherlands Institute for Sound and Vision, the Dutch national audiovisual broadcast archive.[2] We restrict the target archive to news stories broadcast during a period that encompassed the period of the source article collection, (1 Jan. 2005–20 Dec. 2010). We limited the target archive to news stories as other program categories, e.g., game shows and soap operas, are unlikely to yield suitable link targets for news articles. Each news story is manually described by professional archivists, with free-text *description* and *summary* fields and structured fields describing *persons*, *locations*, *keywords* and *other names* that are the subject of the news story. Once again, rather than considering the text of all these fields individually, we combine them to form the *metadata* representation for a given target record $t$. On average, target record metadata consists of 13 terms, illustrating the relative sparsity of text in the target archive as compared to the source archive.

### Archive Statistics

Table 7.1 provides an overview of the type of fields present in the records in the source and target archive. Relatively few of the records in the target archive have annotations other than the title and date field resulting in a sparse event representation for the target records. In contrast about a third of the source records have explicitly annotated entities and more than half have keywords assigned to them. On top of that each source record has a content field containing the article text and a title field, yielding rich event representations for source records.

**Table 7.1:** Statistics of the fields present in the items in the source and target archives.

| Field | Source | Target | Field | Source | Target |
|---|---|---|---|---|---|
| id | 346,559 | 73,666 | content | 346,559 | – |
| date | 346,559 | 73,666 | persons | 117,742 | 2,042 |
| title | 346,559 | 73,618 | locations | 123,412 | 4,736 |
| description | – | 8,632 | other names | 114,726 | 2,601 |
| summary | – | 26 | keywords | 269,691 | 4,770 |

[2] http://instituut.beeldengeluid.nl/

### Events

We use the definition of event used at the Topic Detection and Tracking (TDT) campaign which makes a distinction between a *seminal event*, i.e., a high impact news event—along with all its necessary preconditions and unavoidable consequences—that generates follow-up events, and *related events* that are caused by or predict the seminal event but are not seminal events by themselves.[3] Let us consider the following seminal event: *In February 1998, a low-flying U.S. Marine jet sliced through the cable supporting a funicular at a ski resort in Cavelese, Italy. The funicular then came crashing down, killing 20 people and injuring many more.* Within the TDT definition the funicular's fall to the ground and the subsequent deaths and injuries were all unavoidable consequences of the jet flying into the cable, and are thus considered part of the same seminal event.

Related events are events that are directly related to the seminal event, e.g., the rescue efforts, statements made by the US Marines about policies for training missions in civilian areas, and the criminal investigation that followed the cable car crash.

### Test Collection of Source Archive Records

In order to evaluate our linking approaches, we select a set of source records to use as test set in our linking task. We use two requirements for our selection: the selected record should contain a clear seminal event (to facilitate judgments of system-generated links) and there should be at least one record in the target archive that covers the same or a related event. To satisfy the first requirement, we randomly select news events from Wikipedia listings of important events per month[4] and manually search the source archive to identify a newspaper record describing the event. To satisfy the second requirement, we search in the target archive to make sure that there is at least one television broadcast that describes the same or a related event. If so, the record is selected as a test source record. In total we selected 50 test source records, describing a range of events such as *16 May 2007: Nicolas Sarkozy is sworn in as the new president of France*; *17 December 2009: Heavy snowfall in Belgium and the Netherlands disrupts trains and causes traffic jams*; and *7 July 2009: A memorial service is held in the Staples Center in Los Angeles for the deceased pop icon Michael Jackson.*

### Relevance Judgments

We create relevance judgments using the pooling method adopted by TREC [158], the de-facto standard for creating relevance judgments for test collections (cf. §6.1.4). We performed pooling on the basis of the sets of results produced by different retrieval systems. For each system and source record, the top 20 ranked documents were selected for inclusion in the pool. These results were then merged and duplicate documents were removed. The merged lists of results were then shown to human assessors, with results for each individual source record being judged by the same assessor to ensure consistency.

The assessors were instructed to make a distinction between target records that describe the same event as the source record and targets that describe related events. The assessors' instructions were based on instructions from the TDT assessor manual.[3]

[3] `http://projects.ldc.upenn.edu/TDT5/Annotation/TDT2004V1.2.pdf`

[4] See e.g., `http://nl.wikipedia.org/wiki/Januari_2009`

All judgements in the TDT tasks are binary, i.e., related to the source record event or not. An important difference between the TDT task and our work here is that we make a distinction between same events and related events. This allows us to study the impact of document expansion and term selection on linking performance for these two types of event. In our evaluation set we find that for each source record the average number of target records describing the same event is much lower than the number describing related events (2.4 vs. 11.8).

## 7.2.2 Experimental Setup

Below we detail our experimental setup used to investigate the utility of document expansion and term selection on linking performance. Experiments are performed on the two tasks: *same event linking*, i.e., linking to records that describe the same event, and *related event linking*, i.e., linking to records that describe a related event. In all experiments our baseline is to perform linking using the original representation of the source and target records without document expansion or term selection.

### Expanding Sparse Text Representations

We investigate the effect of increasing the number of documents used to expand target records on linking performance. An overview of the experiments is given in Table 7.2. We experiment with three sources of information for document expansion: the target archive itself, expanding target records with representations from other records in the archive; Wikipedia, an online encyclopedia; and the richly represented, news-focused records in the source archive.

**Table 7.2:** Description of the expansion models. In all cases the original sparse target metadata is concatenated with $n$ expansion documents to form the expanded record representation.

| Exp. model | $A_x$ | Description |
|---|---|---|
| baseline | – | no expansion |
| $n$ target docs | $A_t$ | add $n \in \{1, \dots, 10\}$ documents from target archive |
| $n$ Wikipedia docs | $A_e$ | add $n \in \{1, \dots, 10\}$ documents from Wikipedia |
| $n$ source docs | $A_s$ | add $n \in \{1, \dots, 10\}$ documents from source archive |

### Term Selection for Rich Text Representations

We investigate the effect of reducing the amount of text in a source record on linking performance. An overview of our term selection experiments is given in Table 7.3. First, we experiment with using the fields in the newspaper article metadata to reduce the source record representation, following the framework presented in [328]. We then experiment with using only the most representative terms and entities from the content of a source archive record. We also investigate using only the manual annotations, i.e., metadata, of a source record. Finally, we experiment with using the optimal combination of these options.

**Table 7.3:** Descriptions of the term selection models evaluated in the term selection experiments. In all experiments the number of terms in the *source* item representation (content and metadata) is reduced. Target items consist of their original metadata without expansion.

| TS model | Description |
|---|---|
| baseline | all text associated with $s$, including content text and metadata text |
| content | $s$ content text only |
| metadata | $s$ metadata text only |
| title | $s$ title |
| lead | first 2 sentences of content $s$ |
| $x\%$ terms | select top $x\%$ terms from *content* $s$, using TFIDF ($x \in \{10, 20, \ldots, 100\}$) |
| $y\%$ ne | select top $y\%$ entities in *content* $s$, using TFIDF ($y \in \{10, 20, \ldots, 100\}$) |
| combined | combine metadata $s$ with optimal $x\%$ *terms* and $y\%$ *ne* from content $s$ |

#### Evaluation Measures and Significance Testing

We use three evaluation metrics for evaluating linking performance. Mean Average Precision (MAP), the average of the Average Precision (AP) scores over all test records, evaluates the number of correct link targets in a list (of length 100 in our case), where correct targets higher in the list are assigned more importance. Precision at rank five (P@5) only considers link targets in the top five. A perfect score of 1.0 indicates that all five targets at the top are correct. When fewer correct targets exist the maximum score will be lower. Mean Reciprocal Rank (MRR) is the average of the Reciprocal Rank (RR) for each source record. The RR is the inverse of the first correct answer and indicates at which rank of the list of target records the first correct target is found. We use a standard paired t-test to determine significant differences between results. We use $^{\vartriangle}$ or $^{\triangledown}$ ($^{\blacktriangle}$, $^{\blacktriangledown}$) to indicate whether a score is significantly higher or lower than the baseline with a significance level of $\alpha < .05$ ($\alpha < .01$).

## 7.3 Results

### 7.3.1 Document Expansion

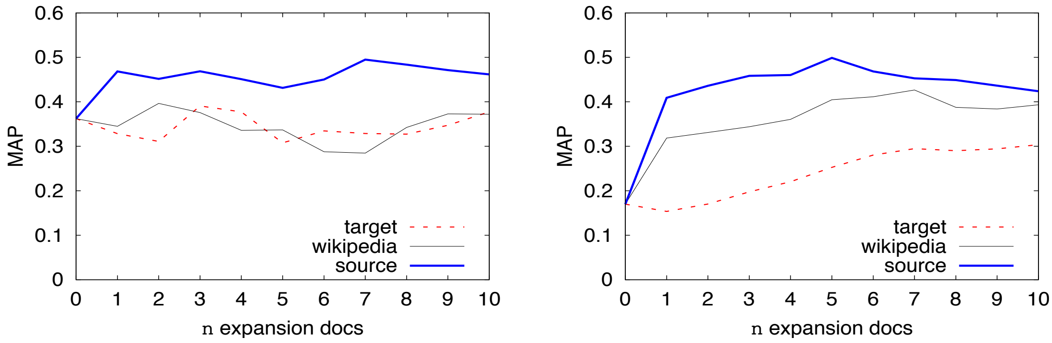
We first contrast the effect of using records from different archives for expansion on linking performance, i.e., records from the source archive, target archive, and Wikipedia. Figure 7.1a shows the MAP scores for *same event linking* using different archives for expansion. Expanding with documents from archives other than the source archive does not result in consistent improvements over the baseline even with the optimal number of expansion documents. Figure 7.1b shows that for *related event linking* expansion with documents from all three archives improves over the baseline. Again expanding with source archive documents achieves best performance. The performance scores for both event linking tasks, with the optimal number of expansion documents, are given in Table 7.4. The optimal number of documents to expand with from the source archive is seven for *same event linking* and five for *related event linking*; both yield a significant improvement over the baseline. We note that although optimized for MAP, the other

**Table 7.4:** Results of document expansion; significance is tested against the baseline.

| Same event | | | | |
|---|---|---|---|---|
| Exp. model | *detail* | *MAP* | *P5* | *MRR* |
| *baseline* | – | .3623 | .2000 | .4819 |
| $n$ *target docs* | $n = 3$ | .3907 | .2227 | .4654 |
| $n$ *wikipedia docs* | $n = 2$ | .3964 | .2136 | .4425 |
| $n$ *source docs* | $n = 7$ | .4949$^{\vartriangle}$ | .2818 | .5435 |
| **Related event** | | | | |
| Exp. model | *detail* | *MAP* | *P5* | *MRR* |
| *baseline* | – | .1699 | .2732 | .5082 |
| $n$ *target docs* | $n = 10$ | .3036$^{\blacktriangle}$ | .3854 | .5705 |
| $n$ *wikipedia docs* | $n = 7$ | .4266$^{\blacktriangle}$ | .4537$^{\blacktriangle}$ | .6988$^{\vartriangle}$ |
| $n$ *source docs* | $n = 5$ | .4988$^{\blacktriangle}$ | .4829$^{\blacktriangle}$ | .6864$^{\vartriangle}$ |

early precision metrics follow the same trend in that the optimal number of documents for MAP is also the optimal number for the other metrics. The P5 scores for *same event linking* do improve (by 40.9%), but remain relatively low; this is due to the small number of relevant target records per test record (on average 2.4).

Let us examine the source record that benefits most from document expansion in the *same event linking* task. The title of this source record is "Openness expenses Dutch Royal Family." The description of the target record is: "Prime Minister Balkenende promises the House of Representatives transparency in the expenses of the Royal Family." The underlying event of the source and target record is the same, i.e., a parliamentary discussion about transparency with respect to the expenses of the Dutch royal house. However, the viewpoint of the event is described from a different angle in each record: the source record focuses on a request for more transparency from the house of representatives, while the target record focuses on the prime minister promising this trans-



**(a)** Same event **(b)** Related event

**Figure 7.1:** Document expansion with $n$ documents, from the target archive, the Wikipedia encyclopedia, and the source archive. Here 0 indicates no expansion.

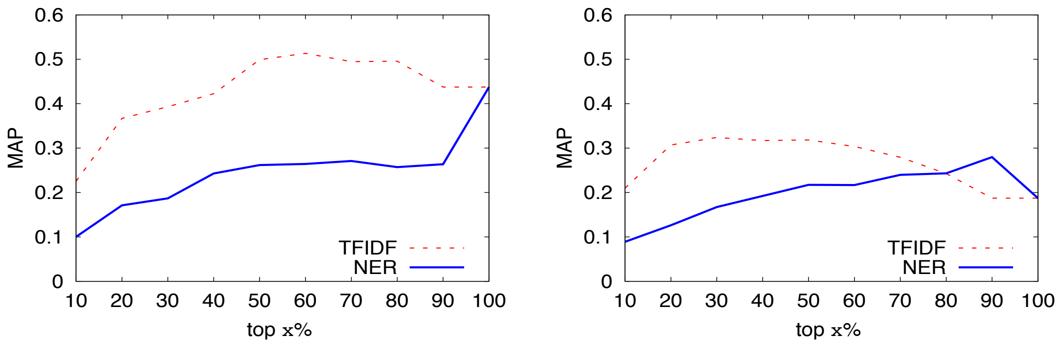
parency. Document expansion works as it adds text from multiple news articles about the parliamentary discussion on transparency to the target record, compensating for different views. Similarly, for *related event linking*, document expansion increases the number of viewpoints of a seminal event covered in a source record to improve linking performance.

### 7.3.2 Term Selection

On the *source* record side we experiment with different term selection techniques. In this section, we link to the original unexpanded records in the target archive. Table 7.5 shows that using only terms from a specific field, e.g., lead or title, improves over using the whole document in terms of absolute scores for both *same event linking* and *related event linking*, but not significantly so. We also select terms and named entities from the content of the *source* record based on their TFIDF score. Figure 7.2a shows the MAP score for *same event linking* while using only the top $x\%$ of the terms (dotted line) or named entities (solid line). We observe that removing any named entities decreases performance. For selecting terms there is an optimum when only 60% of the terms (ranked by TFIDF) are selected. Table 7.5 shows that *same event linking* with the optimum of 60% of the terms selected from the source record, a significant improvement over the baseline is achieved. When linking to related events, selecting terms from the source record does not lead to significant improvements over the baseline; this is not surprising as *related event linking* is more recall oriented and benefits from having a source record

**Table 7.5:** Results of the term selection experiments; significance tested against the baseline.

| | | Same event | | |
|---|---|---|---|---|
| TS model | *detail* | *MAP* | *P5* | *MRR* |
| *baseline* | – | .3623 | .2227 | .4820 |
| *content* | – | .3582 | .1955 | .4800 |
| *metadata* | – | .1636$^\blacktriangledown$ | .0636$^\blacktriangledown$ | .1863$^\blacktriangledown$ |
| *title* | – | .4157 | .2227 | .4597 |
| *lead* | – | .4428 | .2318 | .5386 |
| $x\%$ *terms* | $x = 60\%$ | .5133$^\vartriangle$ | .2682 | .6390 |
| $y\%$ *ne* | $y = 100\%$ | .4374 | .2091 | .5592 |
| *combined* | $x$=60%, $y$=100% | .4660 | .2409 | .5849 |
| | | Related event | | |
| TS model | *detail* | *MAP* | *P5* | *MRR* |
| *baseline* | – | .1699 | .2732 | .5083 |
| *content* | – | .1583 | .2634 | .4838 |
| *metadata* | – | .1768 | .2000 | .2887$^\blacktriangledown$ |
| *title* | – | .2264 | .2829 | .4300 |
| *lead* | – | .2681 | .3366 | .5294 |
| $x\%$ *terms* | $x = 30\%$ | .3229 | .3268 | .4799 |
| $y\%$ *ne* | $y = 90\%$ | .2796 | .2829 | .4724 |
| *combined* | $x$=30%, $y$=90% | .3387 | .3317 | .4459 |



**(a)** Same event

**(b)** Related event

**Figure 7.2:** Term selection with the top $x\%$ (ranked by TFIDF) of the original terms and named entities from the original source item retained.

description that covers all aspects of a seminal event. When less terms from the source record are selected, less aspects of an event are covered making linking to related events more difficult.

We take a closer look at the source record that benefits most from term selection on the *same event linking* task. The title of the source record is "Ukrainian president dissolves parliament." One of the target records is described by: "President Yushchenko of Ukraine dissolves parliament and issues new elections." The source record content consists of 342 words and mentions various aspects of the event, e.g., comments of the opposition leader and protests leading to the dissolution of parliament. Each aspect potentially matches with the description of a target record. As the political situation in the Ukraine was unstable for a number of years, many target records cover aspects of this topic. By only selecting a small number of terms specific to the seminal event, term selection prevents a drift in topic towards other aspects of the source record description.

In the case of *related event linking* the benefit of term selection is less in terms of linking performance than for *same event linking*. We perform the same type of analysis as for same event linking and find the source archive records on which query modeling performs worst. That is, given a source topic about: "Congress wants US out of Iraq" and the following events described by target records: "US senate wants to set a timeline for withdrawing from Iraq" and "President Bush in conflict with Senate US about leaving Iraq." By removing terms from the source record less aspects of the seminal event are covered, i.e., conflict with Bush or the timeline, making linking more difficult. In contrast, expansion supplies additional context to the target records increasing the number of aspects of an event that are covered and thus increasing the similarity between target and source record.

### 7.3.3 Further Improving Linking Performance

In order to see how far we can push linking performance we conduct two additional experiments: (i) by combining document expansion and term selection techniques; and (ii) by filtering target records with dates that do not fall within a certain time window

**Table 7.6:** Results of experiments of combining document expansion with term selection and experiments with date filtering. For details about the optimal settings see Table 7.4 and 7.5.

| | | Same event | Related event |
|---|---|---|---|
| Expansion | TS model | *MAP* | *MAP* |
| *baseline* | *baseline* | .2227 | .1699 |
| *optimal* | *baseline* | .4949 | .4988 |
| *baseline* | *optimal* | .5133 | .3387 |
| *optimal* | *optimal* | .4801 | .4641 |

around the date of a source archive record.

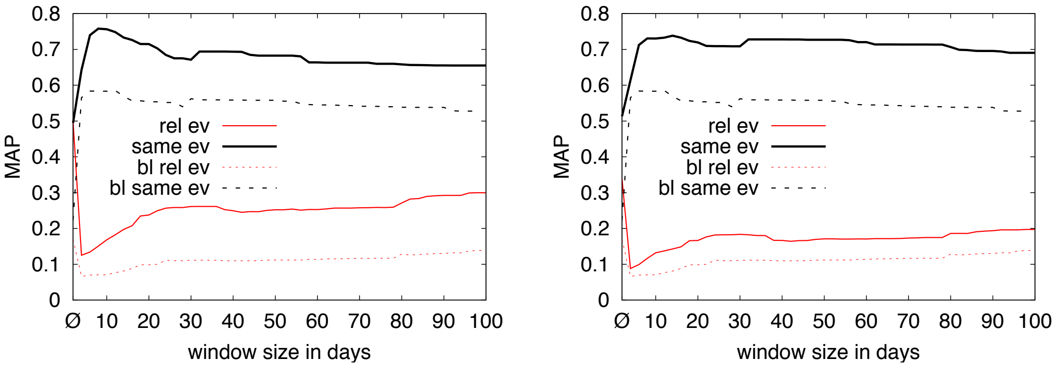
### Combining document expansion and term selection

In our first attempt to improve linking performance, we combine the best expansion and term selection models, i.e., the best term selection is used to find targets and the target records have been expanded with the optimal number of documents, see Table 7.6. The combination achieves a MAP of .4801 on the *same event linking* task, which does not improve over using document expansion (.4949) or term selection (.5133) by itself. We find similar results for *related event linking*. We find that for records where document expansion helps, term selection has relatively poor performance, and vice versa. This fits the intuition that term selection and expansion have opposite effects: one makes a record's event description more specific, while the other broadens the description. Depending on whether the source record is focused on a single topic or discusses several aspects only one of these effects may be desired.

Existing approaches to related article finding generally incorporate either document expansion or term selection (cf. §6.2.1). Our results suggest that each of these approaches are effective. To further improve linking performance, however, a straightforward combination of these techniques is not enough. Lavrenko et al. [205] introduced an effective approach to related article finding that first expands both source and target archive records before selecting informative terms from these expanded representations. Their setting as defined by the TDT task, however, is different from ours. First, we deal with rich textual descriptions in a source archive and sparse textual descriptions in a target archive. Second, relevance judgements in our setting make a distinction between same events and related events, while these are conflated in the TDT task. We did not observe any benefit in a combined approach where we applied term selection on the source record side and document expansion on the target record side. Further investigations are necessary to determine in which setting combining document expansion and term selection is viable.

### Applying a Window Based Date Filter

Our second experiment is with a date filter that restricts target records to those with publication dates within a time window, i.e., a certain number of days, around the date of the source record. Figure 7.3 shows the results on linking performance for the same event linking (*same ev*) and related event linking (*rel ev*) tasks using the baseline, document expansion and term selection methods when increasing the number of days included in the date filter window. The left most sides of Figure 7.3a and 7.3b show the results for our



**(a)** document expansion **(b)** term selection

**Figure 7.3:** Results for related event linking (*rel ev*) and same event linking (*same ev*) using the baseline, document expansion and term selection methods when increasing the window size in terms of days. Here Ø indicates that no window based filtering is applied.

linking methods without applying the date filter (indicated with Ø). We observe that the MAP score for the baseline method is optimal for same event linking with a window size of 10 days, i.e., 5 days before and 5 days after the source record publication date, and increases from .2227 to .5855. The performance of both the document expansion and term selection methods follow a similar pattern and improve from .4949 to .7485 and from .5133 to .7327 MAP for the best document expansion and term selection models, respectively. Although scores for all models go up, including the baseline, the same significant differences in performance remain between the baseline and the best models. Increasing the window size beyond two weeks results in a drop in performance for all methods. The effect of the date filter on performance in the *same event linking* task is unsurprising: this is a high precision-oriented task where news broadcast and newspaper articles about the same event are generally published around the same day. We note, however, that this effect is specific to linking news based on same events and that the optimal values for the filter window are specific for our evaluation set.

On the *related event linking* task using a date filter decreases performance for all methods. Unlike the same event linking task, related events may be distributed over a long period of time. For example, the achievements of a sports team during an international tournament or the apprehension of a longtime fugitive.

## 7.4 Conclusions

The meaning of an archival record is not solely determined by the record itself but arises through interpretation of the record within the context of related records. In this chapter we investigated a way of providing access to contextual information by automatically generating links between records across archives based on events. In particular, we defined the task of linking archives as follows: given a representation of a record from an archive (the source archive), connect it to the representation of a record in another archive (the target archive). In this setting we investigated two variants of the linking archives task, i.e., *same event linking* and *related event linking* and asked:

**RQ 4.** How can we automatically generate links from a record in a newspaper archive with a rich textual representation to records in a television archive that tend to have sparse textual representations?

**a.** Does expanding sparse record representations with text from other sources improve linking performance?

**b.** What effect does modeling reduced versions, e.g., by selecting informative terms of the original richly represented records from the source archive, have on linking performance?

In answer to (**a**), we find that expanding *target* records with documents from other sources improves performance for both *same event linking* and *related event linking*. Using expansion documents from the source archive, however, is most effective as the content has the same focus as the target archive.

In answer to (**b**), we find that reducing the number of terms in the *source* record representation is most effective for *same event linking*. Removing any number of entities, however, has an adverse effect. The reduced records are more robust to topic drift and form a better match for the short event descriptions in the target archive. *Related event linking* also improves by applying term selection but not as much as with target record expansion. Related events benefit more from rich descriptions (as obtained through expansion) that cover all aspects of an event. Further, we found no benefit in combining document expansion and term selection techniques.

These findings have implications for the design of applications to support humanities researchers in finding contextual material based on automatically generated links. As observed in studies of the research habits of historians, getting an overview of the various aspects related to a topic is an important part of their research cycle (cf. §2.3.2). We made a similar observation in our study of the research cycle of media studies researchers in Chapter 3. Our results here show that different methods are appropriate to link to records that cover different aspects of a topic, i.e., same events or related events. Applications to support contextualization based on linking should therefore not be limited to a single method and provide users with a range of methods to discover links to different aspects of a topic. An important observation from Chapter 4 was that the criteria on which such links are based should be made explicit. Otherwise, a mismatch between a user's expectation of what constitutes a link to related information and a system's criteria for creating a link, may lead to frustration.

In this chapter we focused on methods that generate links between events that have the potential to support humanities researchers in the discovery of contextual material in news and multimedia archives. A research topic, however, is not necessarily centered around events and may revolve around a person or an organization. In the next two chapters we investigate methods that enable discovery of entities that share a particular relation. Such methods could support humanities researchers in gathering contextual material based on relations between entities. In Chapter 8 we investigate methods that utilize structured data to support this type of search and in Chapter 9 we explore methods that operate on unstructured data.

# 8
# Example Based Entity Finding in the Web of Data

To be able to arrive at an understanding of a topic it is necessary to identify the concepts and relationships that exist within the domain of a topic. In the discussion of humanities research habits in Chapter 2 and in our investigation of the media studies research cycle in Chapter 3 we observed the practice of humanities researchers to embed themselves in primary and secondary source material in order to arrive at a holistic view of their research topic. Novak and Gowin [256] suggested that this process of learning about a topic resembles that of creating a map in which the concepts related to a particular topic are organized. In the case of humanities researchers such a concept map slowly evolves as they encounter new material and identify new concepts or relationships.

In the previous chapter we developed a method that enables gathering additional material based on links between archival records. This material provides contextual information for the interpretation of concepts and relations already identified within the domain of a topic as well as sources for discovering new concepts and relations. Another way to support contextualization would be to directly identify the concepts related to the ones already identified as relevant to the topic. Once all, or sufficiently many, concepts have been identified, contextual material may be systematically located, for example, by using these concepts as search terms.

The Linking Open Data (LOD) cloud is part of an interconnected Web of Data that contains information about relations between concepts. The Web of Data is formed by connections between a multitude of knowledge bases and information repositories [53]. This type of structured data has the potential to be helpful in identifying relations between concepts. There are two general types of approach to querying the Web of Data: structure-based and text-based approaches. Structure-based queries, e.g., using SPARQL,[1] enable the retrieval of concepts by specifying the relations these concepts should have to other concepts. Table 8.1 shows an example of a structure-based query and its results. This query finds scientists with their associated religion and field. Such a query might be submitted to a SPARQL endpoint interface[2] by a humanities researcher to obtain a list of concepts related to the topic of science and religion. Constructing a structure-based query, however, requires knowledge of a structured query language as well as the

[1] http://www.w3.org/TR/rdf-sparql-query

[2] http://dbpedia.org/snorql/

**Table 8.1:** An example of a SPARQL query and some of the concepts it returns. These objects satisfy the requirement that they have both a religion and field as property and are scientists. Variables are prefixed with a ? and we use `dbpprop:` and `dbpont:` as abbreviations to refer to the namespaces `http://dbpedia.org/property/` and `http://dbpedia.org/ontology` respectively.

```
SELECT ?scientist ?religion ?field
WHERE {      ?scientist dbpedia:property/religion ?religion .
             ?scientist ?p dbpedia:ontology/Scientist .
             ?scientist dbpedia:ontology/field ?field }
```

| scientist | religion | field |
|---|---|---|
| `dbpedia:Isaac_Newton` | Christian | Physics |
| `dbpedia:Johannes_Kepler` | Lutheranism | Mathematics |
| `dbpedia:Nicolaus_Copernicus` | Roman Catholic | Astronomy |

schemas underlying the Web of Data to be able to specify the relationships that relevant concepts should have to other concepts. In this case, it is required that a relevant concept should have the predicates `dbpprop:religion` and `dbpont:field` relating it to some other objects (indicated by the variables ?religion and ?field respectively) and a third unspecified predicate (?p) relating it to the particular object `dbpont:ontology/Scientist`.

From a user's point of view it is easier to specify keyword queries to retrieve information about concepts. These text-based approaches, however, make limited use of the potential of the available structure and instead focus on the sparse textual information associated with objects (cf. §6.3.3).

An alternative to using keyword queries is to allow humanities researchers to submit examples of the concepts related to their topic in order to find additional concepts. Providing examples is easier than specifying structure-based queries, while the structural information associated with the examples could be used to provide input for structure-based methods. Possible scenarios for users to obtain examples are to use keyword queries to retrieve examples from an initial result set or to use a schema browser allowing a user to wander from one entity to the next until one or more examples have been found [331].

In this chapter we look into the challenge of utilizing examples to find related concepts. The context in which we evaluate our methods is modeled after the entity list completion task as seen at various evaluation platforms (cf. §6.3.3). We define the task as follows: *given a query consisting of a relation and example entities, complete the list of examples by finding URIs of entities that all share the specified relationship with a particular concept.* Given this task we now take on our fifth research question as introduced in Chapter 1:

**RQ 5.** How can we exploit the structural information available in the Web of Data to find a set of entities, that all have the same relationship with a particular concept in common, based on a number of example entities?

**a.** Is a structure-based method that uses examples competitive when compared against a text-based approach?

**Table 8.2:** One of the ELC test topic descriptions.

| | | |
|---|---|---|
| query $Q$ | $R$ : | Apollo astronauts who walked on the Moon |
| | $X$ : | dbpedia:Buzz_Aldrin dbpedia:Neil_Armstrong |

**b.** Does the performance of text- and structure-based methods depend on the quality and the number of examples that are given?

**c.** Can a hybrid method automatically balance between the two approaches in a query-dependent manner?

The contributions of this chapter are three-fold: (i) a deeper understanding of the issues involved in using examples for entity search; (ii) an error analysis of text-based and structure-based methods; and (iii) the introduction of an alternative hybrid method that is more effective in combining text-based and structure-based approaches than current hybrid approaches, i.e., linear combinations of text-based and structure-based components that use a fixed mixture weight for all queries. We describe the details of our text-based, structure-based, and hybrid approaches in Section 8.1. We provide details of our experimental setup in Section 8.2. In Section 8.3 we present our results and provide a discussion. We conclude in Section 8.4.

## 8.1 Task and Approach

In the *Entity List Completion* (ELC) task a query (Q) consists of (i) a textual representation for the relation ($R$) and (ii) a URI based representation for the example entities ($X$); see Table 8.2 for an example topic. The goal is to complete the list of examples by finding URIs of entities that have the specified relation. The data we consider for this task consists of a sample of the LOD cloud. Linked Data is typically represented using the RDF format and defines relations between objects in the form of triples. We briefly recall the details from §6.3.2.

An RDF triple consists of a subject, a predicate, and an object. A subject is always a URI and represents a "thing" (in our case: an entity), such as `dbpedia:Isaac_Newton` which is the URI representation for Isaac Newton in DBpedia.[3] Subject URIs serve as unique identifiers for entities. An object is either a URI representing another "thing," e.g., `dbpont:Scientist`, or a string holding a literal value, e.g., the values for the field and religion variables in Table 8.1. Predicates are also always URIs and specify the relations between subjects and objects, e.g., `dbpprop:religion`.

### 8.1.1 Text-based Approach

There are two choices to be considered in designing a text-based approach to entity finding in Linked Data: (i) the representation of entities and (ii) the retrieval model. A popular approach to representing entities is to group all triples that have the same URI as subject together [36, 108, 218]. We follow [56, 254, 268] and use a fielded

[3] The `dbpedia:` prefix is an abbreviation for the namespace `http://dbpedia.org/resource/`.

**Table 8.3:** An example of the entity representation consisting of the aggregation of triples with the same subject, i.e., `dbpedia.org/resource/Isaac_Newton`. A distinction is made between three types of field: attributes, types, and links. The abbreviations `rdf:`, `cyc:`, `owl:` `rdfs:` and `dcterms:` are used for the namespaces `http://www.w3.org/1999/02/22-rdf-syntax-ns#`, `http://sw.cyc.com/concept/`, `http://www.w3.org/2002/07/owl#`, `http://www.w3.org/2000/01/rdf-schema#` and `http://purl.org/dc/terms/` respectively. For other namespaces see Table 8.1

| | predicate | object |
|---|---|---|
| attributes | `rdfs:label` | Isaac Newton |
| | `dbpprop:shortDescription` | English mathematician, physicist, and astronomer |
| | `rdfs:comment` | Sir Isaac Newton (25 December 1642 20 March 1727) was an English physicist, mathematician, astronomer, natural philosopher, alchemist, and theologian. His monograph Philosophiae Naturalis Principia Mathematica lays the foundations for most of classical mechanics. |
| types | `dcterms:subject` | `dbpedia:category:Christian_mystics` |
| | | `dbpedia:category:English_astronomers` |
| | | `dbpedia:category:Theoretical_physicists` |
| | `rdf:type` | `http://umbel.org/umbel/rc/Scientist` |
| | | `http://schema.org/Person` |
| links | `dbpprop:nationality` | `dbpedia:Kingdom_of_England` |
| | `dbpprop:religion` | `dbpedia:Christian` |
| | `dbpprop:fields` | `dbpedia:Mathematics` |
| | `dbpprop:birthPlace` | `dbpedia:Woolsthorpe-by-Colsterworth` |
| | `owl:sameAs` | `cyc:Mx4rwETmR5wpEbGdrcN5Y29ycA` |

representation where triples associated with an entity are grouped into a small set of predefined categories. We consider the following three categories: (i) *attributes*, i.e., triples that have a string as object; (ii) *types*, i.e., triples for which the predicate is one of a predefined set of common predicates to indicate type information (`rdfs:type`, `skos:subject`,[4] `dcterms:subject`); and (iii) *links*, i.e, triples that have another node as object and are not of the *types* category. The objects of the links and types categories are URIs. This results in an entity representation as shown in Table 8.3. To obtain a meaningful textual representation we expand these URIs with the text associated with an object through the `rdfs:label` predicate, which is widely used to provide a natural language description for Linked Data objects. For example, to convert the identifier of object `cyc:Mx4rwETmR5wpEbGdrcN5Y29ycA` to the more human readable string "Isaac Newton."

For the retrieval model, we adopt a language modeling approach because of its probabilistic foundations and effectiveness in entity-oriented search tasks [108, 127, 254]. Our goal is to obtain a ranking of document representations of entities ($e$), e.g., the rep-

[4] The `skos:` prefix abbreviates the namespace `http://www.w3.org/2004/02/skos/core#`.

resentation of Isaac Newton in Table 8.3, based on the probability of being relevant to the relation ($R$) as specified in a query (Q): $P(e|R)$. As the limited number of terms in the relation make it hard to estimate this probability directly we apply Bayes' rule to reformulate this to $P(R|e)P(e)/P(R)$. We drop the denominator $P(R)$ as it remains the same for all entities and does not influence the ranking. For the entity prior, $P(e)$, we assume a uniform distribution.

Following the language modeling framework, we model the entity document representation $e$ as a multinominal probability distribution ($\theta_e$) over the vocabulary of terms, i.e., the set of terms that occur in any of the entity representations in our corpus. This model captures the regularities in the language usage for each entity representation and allows us to predict how likely the entity model will produce the relation $R$. We refer to the probability that an entity model $\theta_e$ generate a relation $R$ as $P(R|\theta_e)$. If we assume that terms are generated independently we obtain $P(R|\theta_e)$ as the product over the terms in the relation:

$$P(R|\theta_e) = \prod_{t \in R} P(t|\theta_e).$$

What remains to be done is to estimate the probability of a term $t$ given the Dirichlet smoothed language model. We follow the standard language modeling approach [375] and estimate $P(t|\theta_e)$ as:

$$P(t|\theta_e) = \tfrac{tf(t,e)+\mu P(t|\theta_c)}{|e|+\mu},$$

where $tf(t, e)$ is the term frequency of $t$ in the representation document of $e$, $|e|$ is the number of terms in the entity representation, and $P(t|\theta_c)$ is the Dirichlet smoothed model of the entire collection of triples. The smoothing parameter $\mu$ is set to the average document length in the collection.

To obtain a ranking for different entity representations, we estimate $P(t|\theta_e^f)$ for each field type ($f$) in Table 8.3, where $\theta_e^f$ is a multinomial distribution estimated over the terms occurring in the triples of a particular field type $f$. Previous work on ad-hoc entity search has shown that a linear mixture of the representation language models is effective [254]. We follow this approach and re-estimate the probability of a term given the weighted representation language models as follows:

$$P(t|\theta_e^w) = \textstyle\sum_{f \in \{tp,lk,at\}} P(t|\theta_e^f)P(f),$$

where $P(f)$ is the weight given to a specific field type model, i.e., types ($tp$), links ($lk$), and attributes ($at$). The $w$ superscript indicates that this is a weighted language model. The probability of the weighted text-based model then becomes:

$$P(R|\theta_e) = \prod_{t \in R} P(t|\theta_e^w).$$

In summary, we have described two ways of estimating $P(R|\theta_e)$: using a language model that collapses terms from all fields into a single field (LM$_{all}$) and a weighted model (LM$_{weighted}$) that combines language models estimated for each field, i.e., LM$_{links}$, LM$_{types}$, and LM$_{attributes}$.

### 8.1.2 Using Examples with a Structure-based Approach

An alternative to the text-based approach is to represent an entity by the links it has to other entities. Taking an entity URI as starting point we consider all RDF triples that have that URI as subject (i.e., outlinks) or object (i.e., inlinks). From these triples we extract the predicate-object or predicate-subject pairs depending on whether our entity occurs as subject or object respectively. For example, from the triple (`dbpedia:Isaac_Newton`, `rdf:type`, `dbpont:scientist`) we extract the pair set {`rdf:type`, `dbpont:scientist`} for the entity `dbpedia:Isaac Newton`. Together, these pairs form the link-based representation of an entity ($e_l = \{pr_1, \ldots, pr_m\}$, where $pr_i$ is a pair set extracted from RDF triples containing entity $e$.).

Under this representation, entities consist of sets of pairs. The set of example entities becomes a set of sets of pairs ($X = \{x_1, \ldots, x_n\}$ and $x_i = \{pr_1, \ldots, pr_k\}$). The goal is to rank entities according to the probability of the entity's link-based representation $e_l$ given a set of example entities $X$: $P(e_l|X)$.

If we model $X$ again as a multinomial distribution over pairs analogously to the text-based method and compute $P(e_l|X)$ as the product of the pairs in $e_l$, then the structure-based method would prefer entities with smaller representations all else being equal. To account for this and to incorporate the intuition that predicate-object pairs observed with more examples are more important than others, we expand this term to incorporate the pairs $pr$ explicitly: $P(e_l|pr, X) \cdot P(pr|X)$.

This probability is difficult to estimate directly as it requires observations of pairs, sets of pairs associated with entities, and sets of examples. Therefore, we assume independence between the examples and the entity given the pairs which allows us to factorize this probability as follows: $P(e_l|pr)P(pr|X)$. Taking X to be a multinomial distribution over relations, $\theta_X$, and marginalizing over the relations observed with the examples we obtain:

$$P(e_l|\theta_X) = \sum_{pr \in \bigcup_{x \in X}} P(e_l|pr)P(pr|\theta_X),$$

where $\bigcup_{x \in X}$ is the union of the triples associated with each example. We estimate $P(tr|\theta_X)$ as follows:

$$P(pr|\theta_X) = \frac{\sum_{x \in X} n(pr, x)}{\sum_{pr' \in \bigcup_{x \in X}} \sum_{x \in X} n(pr', x)}.$$

Here, $n(pr, x)$ is 1 if $pr$ occurs in the representation of example $x$ and 0 otherwise. For $P(e_l|pr)$ we use a function which is 1 if $pr$ occurs in the context of $e_l$ and 0 otherwise.

### 8.1.3 Combining Approaches

Supervised merging and learning to rank methods that combine various ranked lists have gained in popularity [77, 217, 306]. One of the major factors contributing to their popularity is the increased availability of labeled data (relevance judgements) in the form of clicks. Such judgements, however, are not available for emerging tasks such as searching in the Web of Data due to the limited effectiveness of current search tools. To get past this cold start problem effective retrieval models are needed that do not require training data. Therefore, we experiment with unsupervised versions of such combination methods and focus on two prototypical approaches: (i) that use a linear combination of methods, which is effective when multiple methods return similar lists; and (ii) that select a method in a query dependent manner, which is effective when methods return dissimilar lists of results. Our linear combination method combines the normalized similarity scores of the text and structure-based method. In contrast, our query dependent selection method uses the example entities to choose between the text-based approach, the structure-based approach, or a combination of these two approaches.

In the linear combination approach we use the parameter $\lambda$ to control the weight assigned to the structure and text-based methods as follows:

$$P_{cmb}(e|Q) = \lambda \cdot P(e_l|\theta_X) + (1 - \lambda) \cdot P(R|\theta_e),$$

where $Q$ consists of the relation $R$ and the set of examples $X$.

Our second, alternative method is to predict the effectiveness of the text-based and structure-based techniques by capitalizing on the availability of explicit relevance feedback in the form of example entities. This *switch* method chooses between the text-based and structure-based method depending on which method is better able to retrieve the example entities. If both methods achieve similar performance, the linear combination method is used.

We formalize this method as follows: given two ranked lists, one produced by the text-based method for a query ($L_{P(R|\theta_e)}$) and one produced by using the examples with the structure-based ($L_{P(e|\theta_X)}$), we use the example entities as relevance judgements and calculate the average precision (AP) for each of the lists. Based on the difference between the AP scores, $\lambda$ is set to 0, to 1, or to the same value as in the linear combination method:

$$P_{switch}(e|Q) = \begin{cases} P(e|\theta_X) & \text{if overlap} < \gamma \\ & \text{and AP}(L_{P(e|\theta_X)}) > \text{AP}(L_{P(R|\theta_e)}) \\ P(R|\theta_e) & \text{if overlap} < \gamma \\ & \text{and AP}(L_{P(e|\theta_X)}) < \text{AP}(L_{P(R|\theta_e)}) \\ \lambda \cdot P(e|\theta_X) + & \\ (1-\lambda) \cdot P(R|\theta_e) & \text{otherwise,} \end{cases} \tag{8.1}$$

where overlap is defined as:

$$\text{overlap} = \frac{\min(\text{AP}(L_{P(R|\theta_e)}), \text{AP}(L_{P(e|\theta_X)}))}{\max(\text{AP}(L_{P(R|\theta_e)}), \text{AP}(L_{P(e|\theta_X)}))},$$

and $\gamma$ is a threshold parameter that determines how much the performance of the two methods is allowed to overlap, before one is chosen over the other. In case both methods have similar performance, a combination of both methods is used; otherwise, the best performing method is picked. Note that unlike previous work [35], we focus on using the structure of the examples, and not the associated text, e.g., through relevance feedback.

## 8.2 Experimental Setup

Before we discuss our results we briefly recall our research questions for this chapter and describe the experimental settings in which we evaluate the text-based, structure-based, and hybrid methods. In Chapter 1 we asked: (**a**) is a structure-based method that uses examples competitive when compared against a text-based approach; (**b**) does the performance of text- and structure-based methods depend on the quality and the number of examples that are given; and (**c**) can a hybrid method automatically balance between the two approaches in a query-dependent manner? As there is no reference corpus available that allows us to address these specific questions we adapted three test collections to our setting.

The dataset in our experiments is the Billion Triple Challenge 2009 (BTC2009) data set.[5] The first set of topics we use consists of 50 semantic search challenge list completion task topics (SemSearch'11). This task was conducted on the BTC2009 data set and the evaluation data (qrels) with relevant URIs for each topic have been made available, see Appendix A. In this specific setting no explicit examples are provided, only the desired relation that the target entities should satisfy is specified. The relevance judgements are graded on a relevance scale of 0 to 2. We consider URIs judged as either relevant (2) or somewhat relevant (1) the same in our experimental setting as 454 of the 650 judgements are considered somewhat relevant.

In addition, we convert the original INEX'07 and INEX'08 topics to conform to the semantic search setting. INEX topics contain a description similar to the semantic search topic relation (R), e.g., *I want a list of the state capitals of the United States of America.* The topic further contains example entities, e.g., *Lincoln, Nebraska.* In the original INEX entity list completion task the goal is to retrieve entities from Wikipedia. The evaluation data also consists only of titles of Wikipedia pages. We experimented with several approaches to create an initial mapping of Wikipedia entities (pages) to DBpedia URIs [161, 239, 268] and refined this mapping through manual inspection.[6] The examples provided with each topic were added to the evaluation data. This results in a set of 25 and 35 topics with 423 and 849 URIs judged as relevant, respectively. We use the official TREC evaluation measures: R-precision (Rprec), Mean Average Precision (MAP) and number of relevant URIs returned (rel_ret). Result lists are evaluated till rank 100.

In order to obtain example entities we randomly sample relevant entities for each topic from the evaluation data. In our experiments we select 10 random samples for each setting of our *number of examples* parameter as we increase the number of examples provided to the structure-based method. In order to make a fair comparison between methods we remove the sampled examples from the evaluation data. This procedure generates a different evaluation data set each time a different set of examples is selected.

## 8.3 Results

We first consider the results of our text-based approach. Table 8.4 shows the results of the language modeling (LM) approaches described in Section 8.1.1. Each model uses a different combination of field types, i.e., all terms associated with an entity collapsed into a single field (LM$_{all}$), only considering terms from the link (LM$_{links}$), type (LM$_{types}$),

[5] `http://km.aifb.kit.edu/projects/btc-2009/`

[6] See `http://ilps.science.uva.nl/ecir2013elc` for topics and ground truth.

**Table 8.4:** Results of text-based language modeling (LM) approaches estimated using fields within the entity representation: only attributes, only triples containing type information, only triples linking to other nodes, all triples, and a weighted combination of the representations.

| | SemSearch’11 | | | |
|---|---|---|---|---|
| | MAP | Rprec | rel_ret | rel |
| LM$_{attributes}$ | .0726 | .1096 | **193** | 650 |
| LM$_{links}$ | .0854 | .1028 | 169 | 650 |
| LM$_{types}$ | **.0891** | **.1176** | 144 | 650 |
| LM$_{all}$ | .1311 | .1488 | 247 | 650 |
| LM$_{weighted}$ | **.1632** | **.1935** | **270** | 650 |
| | INEX’07 | | | |
| LM$_{attributes}$ | .0497 | .0699 | 40 | 432 |
| LM$_{links}$ | **.0746** | .0673 | **76** | 432 |
| LM$_{types}$ | .0651 | **.0821** | 67 | 432 |
| LM$_{all}$ | .0713 | .0942 | 58 | 432 |
| LM$_{weighted}$ | **.1187** | **.1370** | **93** | 432 |
| | INEX’08 | | | |
| LM$_{attributes}$ | .0173 | .0330 | 82 | 849 |
| LM$_{links}$ | .0670 | .0816 | 186 | 849 |
| LM$_{types}$ | **.0816** | **.0922** | **197** | 849 |
| LM$_{all}$ | .0298 | .0537 | 152 | 849 |
| LM$_{weighted}$ | **.0898** | **.1073** | **217** | 849 |

or attribute (LM$_{attributes}$) fields, and a weighted model (LM$_{weighted}$) that combines the language models estimated for each field,

We find that of the representations that use a specific field associated with an entity, the type representation generally outperforms the other representations in terms of MAP and Rprec. This is in line with our expectations as at the INEX Entity Ranking track treating type information as a special field was a popular and effective approach [33, 340]. We observe that when using all triples as entity representation, precision and recall improve over using any subset of triples as representation for the SemSearch’11 data set and that results decrease for both INEX data sets.

We perform a grid search over the parameter space to obtain the optimal weight values to combine the language models for each individual field. The best performance is achieved with this weighted combination of the different representations. We use the optimal weights in the remainder of this chapter. These values are the same across the three data sets, i.e., to $0.4$ for the *attributes*, $0.2$ for the *links*, and $0.4$ for the *types* entity representation.

For the evaluation of the text-based method we use the verbatim evaluation data with all entities included. This allows us to compare our results to those obtained at the 2011 Semantic Search Challenge. We find that our implementation of the text-based approach is able to achieve these results, e.g., the highest pure text-based approach achieved a

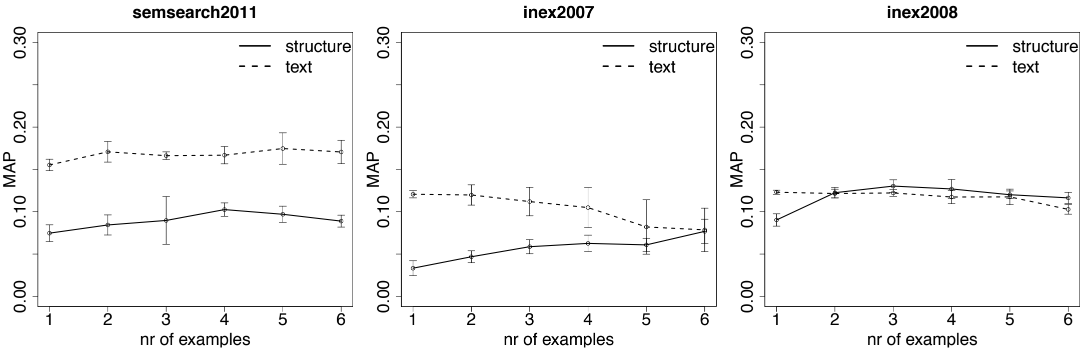


**Figure 8.1:** The average MAP and standard deviation achieved by the text-based method (dotted line) and the structure-based method (solid line).

MAP of 0.1625.[7] Higher performance is achieved by approaches that re-rank an initial ranked list based on the link structure between top ranked entities. We focus on a pure text-based approach as baseline in order to analyze the individual contributions of text- and structure-based methods.

### 8.3.1 Results Using Examples With A Structure-based Approach

We now consider whether the number of examples influences performance, how the structure-based method compares to the text-based method, and how performance varies with the quality of the examples. The solid line in Fig. 8.1 shows the mean and standard deviation of MAP achieved by the structure-based method over 10 samples for different numbers of examples for the INEX and SemSearch data sets. The dotted line shows the mean and standard deviation of MAP achieved by the text-based method. Note that the evaluation data changes with every sample and that the results here are not directly comparable to those in Table 8.4. We observe that on the INEX'07 and SemSearch'11 topics the text-based approach outperforms the structure-based approach, while on the INEX'08 data set comparable performance is achieved. On the INEX'07 data the performance of the text-based method decreases as the number of examples increases, but this phenomenon is not observed on the other topic sets. Performance of the structure-based method increases on all three topic sets when the number of examples is increased and levels off when more than 4 examples are provided. With more examples the structure-based method is better able to determine the importance of triples in the example set but as more examples are added this results in diminishing returns.

Regarding the standard deviation of MAP scores achieved by the structure-based method we observe no obvious pattern and performance of the structure-based method does not become more or less robust as more examples are added. The performance of the text-based method also varies, this as a consequence of sampling entities and removing them from the evaluation data. This variation in performance suggests that the text-based method is dependent on a particular set of entities being relevant.

Next we take a closer look at the per query performance of the text and structure-based methods. Fig. 8.2 shows the difference in Average Precision (AP) achieved by the

[7] `http://semsearch.yahoo.com/results.php#`

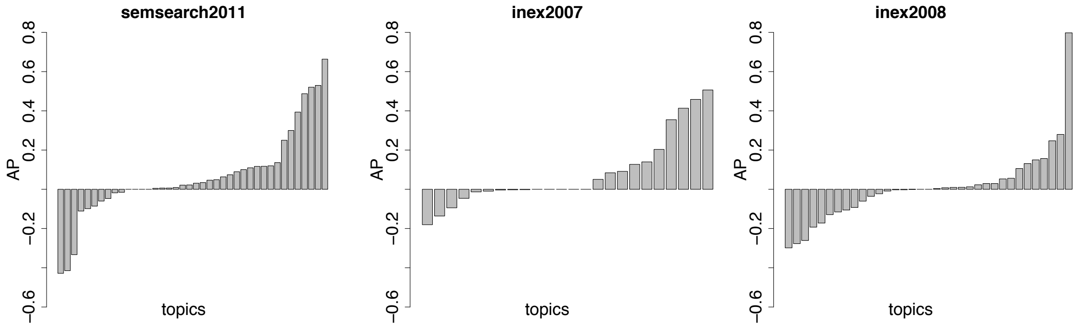


**Figure 8.2:** Barplot of the difference in AP achieved by each topic. A negative value indicates that the structure-based method achieves better AP for that topic than the text-based method. A positive value indicates that the text-based method performs better.

two methods per topic. A positive value indicates that the text-based method is more effective and a negative value indicates that the structure-based method achieves higher AP. The run on which these differences are based uses two examples and was further picked at random. We observe that the text-based method achieves a higher AP on more topics than the structure-based method on the INEX'07 and SemSearch'11 topics. On the INEX'08 topics there is no clear winner. We find that a considerable number of topics exists on which the structure-based method outperforms the text-based method. These results suggest that the text-based and structure-based methods work well on different queries and sets of example entities, motivating the use of a hybrid method.

## 8.3.2 Combined Approaches

A standard approach to combine structured information with a text-based approach is to use a linear combination ($P_{comb}(e|Q)$), where the contribution of each method is governed by a parameter ($\lambda$). To investigate the potential of this approach we perform a sweep, i.e., initialize $\lambda$ from 0 to 1 with steps of 0.1, and find the optimal setting of $\lambda$ over the number of examples: 0.1. For the switch method ($P_{switch}(e|Q)$) we likewise set $\gamma$ to the optimal value (0.0 for INEX'07, 0.1 for INEX'08, and 0.0 for SemSearch'11) and we use the same $\lambda$ as for the linear combination. When $\gamma$ is set to 0 the switch method decides to mix if there is any overlap in the first 100 results of the two methods and otherwise uses the method that was able to return the examples. Note that using optimal settings allows us to investigate how the performance of text- and structure-based methods relate under ideal conditions. We leave an investigation of parameter sensitivity as future work. Fig. 8.3 shows the average and standard deviation of the MAP achieved by the linear combination method (dashed black line) and the switch method (dotted black line). We observe that on all three topic sets the performance of the switch method increases when the number of examples provided increases. In contrast, the performance of the linear combination method decreases when more examples are provided. When providing 3 or more examples the switch method outperforms the linear combination on each data set. On the INEX'07 dataset using 3 or more examples results in significantly ($\alpha = .05$) better performance in terms of MAP than the text- and structure-based meth-

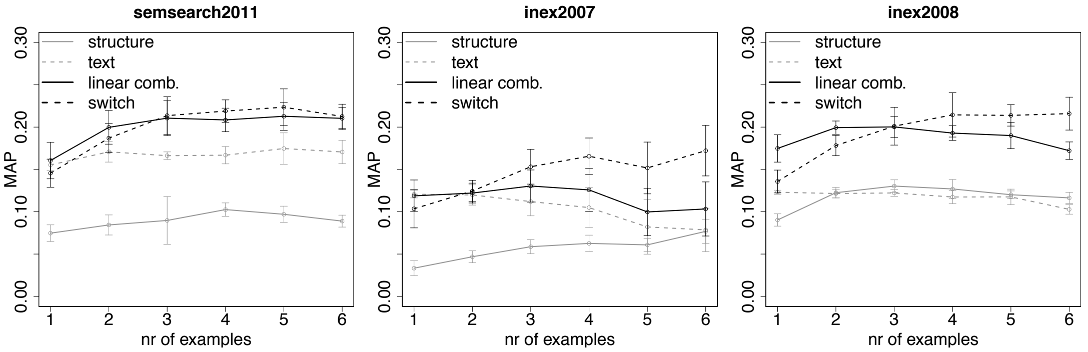


**Figure 8.3:** Average and standard deviation of the MAP achieved by the linear combination method (solid black line) and the switch method (dotted black line). The structure-based method (solid grey line) and text-based method (dotted grey line) are added for comparison.

ods. On the INEX'08 dataset the same holds when using 4 or more examples. On the SemSearch'11 dataset we find that both combination methods significantly outperform the individual methods when using more than 1 example.

These results confirm our earlier observation that the text and structure-based methods return different sets of entities and are effective for different topics. The switch method is able to use the examples to determine which of these two methods will be most effective. The linear combination method performs initially better but is not able to utilize the information provided by the structure-based method. This has implications for such methods in a scenario where users may provide any combination of example entities and are no longer interested in re-finding them.

We observe that the variance for the linear combination and switch method increases compared to the structure-based approach. The methods become more sensitive to the specific examples that are available. This adds another challenge to using examples for entity search, i.e., how to asses the quality of the examples provided to our methods.

## 8.4 Conclusion

A necessary part of gaining a full understanding of a research topic is to determine the concepts and relationships that exist within its domain. In this chapter we investigated the utility of methods based on Linked Data to support this process in the context of an entity list completion task. Specifically, we addressed our fifth research question as raised in Chapter 1:

**RQ 5.** How can we exploit the structural information available in the Web of Data to find a set of entities that all have the same relationship with a particular concept in common based on a number of example entities?

**a.** Is a structure-based method that uses examples competitive when compared against a text-based approach?

**b.** Does the performance of text- and structure-based methods depend on the quality and the number of examples that are given?

**c.** Can a hybrid method automatically balance between the two approaches in a query-dependent manner?

In answer to **a** and **b** we found that depending on the number and quality of the examples, a structure-based approach achieves comparable performance to a competitive text-based approach. Through a per topic analysis, however, we find that each method returns different sets of entities, thereby motivating the use of a hybrid approach. We have performed an analysis of the performance of two hybrid methods on repeated samples of example entities and relevance judgements. Results showed that a standard linear combination approach is suboptimal when the set of examples and entities considered relevant changes. This has consequences for the applicability of linear combination approaches in scenarios where a user provides examples, i.e., the particular set of entities the text-based method is effective in finding may overlap with the examples.

Regarding **c** we found that a hybrid method that uses example entities to determine whether to use a text-based, structure-based, or linear combination approach, outperforms a standard linear combination. We have also found that the variance in the performance achieved by both hybrid methods increases over the text-based and structure-based methods based on the specific set of examples provided. This suggests that a new direction in using examples for entity search lies in assessing the quality of examples provided.

These findings inform the design of applications for humanities researchers that support identifying related concepts. In §2.3.3 we discussed how start-up time in learning to use new technology and the time gain an application is perceived to provide are factors that determine whether an application is adopted by humanities researchers. Our results show that both text and examples are viable as input options to an application supporting related concept finding based on Linked Data. However, allowing users multiple input options has additional advantages in that such an application provides more flexibility and achieves potential gains in terms of performance when multiple options are used.

Although the hybrid methods are able to utilize Linked Data more successfully than either text-based or structure-based methods alone, performance remains low in absolute terms. Further investigations are necessary to determine whether this technology has reached the level of maturity necessary for it to be incorporated in real world applications. The dependence on Linked Data introduces other limitations as well. Therefore, in the next chapter we investigate methods that are able to identify related concepts based on natural language descriptions in a web corpus.

# 9 Ranking Related Entities: Components and Analyses

In the previous chapter we investigated methods to support the discovery of concepts and relationships within the domain of a topic based on structured data. Many of the concepts and relations that humanities researchers seek to identify, however, are difficult to obtain through these methods as only a limited number of relations are encoded in structured form. For example, in Chapter 3 one media studies researcher investigating the rise and fall of a certain political figure noted the potential of organizing material in terms of supporters and opponents, while another sought to identify media that were critically reflected upon in newspapers. In Chapter 4 media studies students were asked to reconstruct the historical context of the 1950s in order to explain the role of female television personalities in the emancipation of women. To support the identification of concepts based on these types of relationship a system would need to be able to answer questions such as: who are the opponents/supporters of political figure X, list the critics who reflected upon television show Y, and who have been colleagues of female television presenter X.

Over the past decade, increasing attention has been devoted to retrieval technology aimed at identifying concepts relevant to an information need not regularly captured by structured data. For example, the TREC Question Answering track focused on fact-based questions such as "Who invented the paper clip?" and the expert finding task, studied at the TREC Enterprise track, focused on identifying experts on a topic (cf. §6.2.2). To address the broader task of identifying relations between concepts the INEX Entity Ranking task and the TREC Entity track were introduced. The INEX Entity Ranking task, however, focused solely on Wikipedia, while the TREC Entity track sought to utilize a web corpus (cf. §6.2.3).

In this chapter we focus on the *related entity finding* (REF) task introduced at the TREC Entity track as it aims at making arbitrary relations between concepts searchable. The task is defined as follows: given a source entity, a relation and a target type, identify homepages of target entities that enjoy the specified relation with the source entity and that satisfy the target type constraint. E.g., for a source entity ("Bill Clinton"), a relation ("His political opponents during his presidential term") and a target type ("people") return entities such as "George H.W. Bush" and "Robert Dole."

Below, we first introduce the REF task in more detail and then recall the research questions from Chapter 1 that we seek to answer in this chapter.

## 9.1 Introduction

A system that aims to extract information from a large corpus of unstructured text as opposed to documents requires a dedicated pipeline of data cleaning and preprocessing components. Figure 9.1 shows an example of such an architecture for an idealized entity retrieval system. Computations take place at two levels: the entity repository is built offline, using tools and techniques for named entity recognition and normalization. Queries are processed online, through a retrieval pipeline. This pipeline resembles a question answering architecture [12], where first candidate answers are generated, followed by type filtering and the final ranking (scoring) steps. Candidate generation is a recall-oriented step, while the subsequent two blocks aim to improve precision. Our work sets out the challenge of adopting this general architecture to the REF task, and addresses the issue of balancing precision and recall when executing a search.

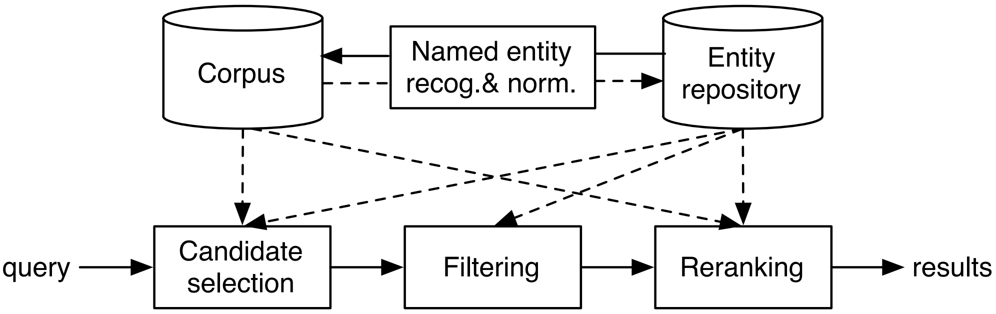


**Figure 9.1:** Components of an idealized entity finding system. Solid arrows indicate control flow, dashed arrows data flow.

When developing a system to perform a task such as REF, the most important evaluation is on the end-to-end task. The end-to-end focus, however, means that it is difficult to disentangle the performance contributions of individual components. Moreover, REF is a relatively new task and a canonical architecture has yet to emerge. In this chapter we go through a series of ablation studies and contrastive runs so as to obtain an understanding of each of the components that play a role and the impact they have on precision and recall.

Specifically, we address the REF task as defined at TREC 2009 and consider a particular instantiation of the idealized entity finding system, shown in Figure 9.2. Our focus is on retrieval and ranking rather than on named entity recognition and normalization; to simplify matters we use Wikipedia as a repository of (normalized) known entities. While the restriction to entities in Wikipedia is a limitation in terms of the number of entities considered, it provides us with high-quality data, including names, unique identifiers and type information, for millions of entities. Our framework is generic and conceptually independent of this particular usage of Wikipedia.

Given our focus on entities in Wikipedia, it is natural to address the REF task in two phases. In the first we build up our retrieval pipeline (the blocks at the bottom of Figure 9.2) working only with Wikipedia as a corpus and Wikipedia pages as representations

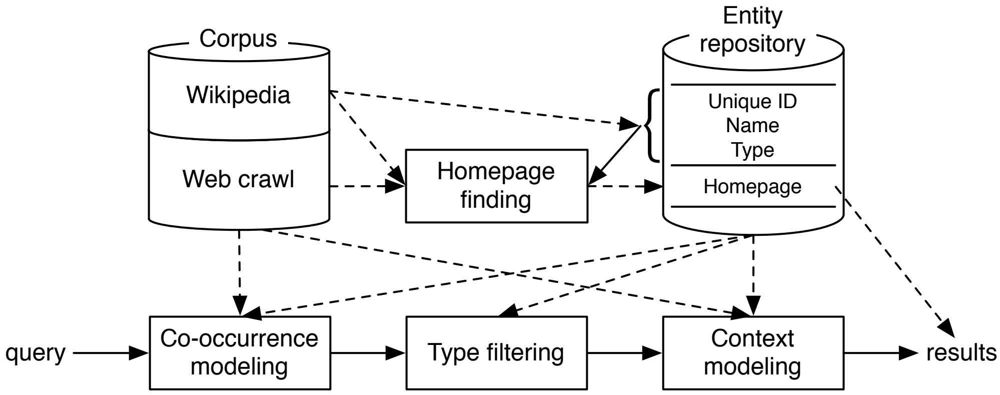


**Figure 9.2:** Components of our REF system.

of entities; in the second phase we go beyond Wikipedia by applying our methods to a web corpus and mapping entities to their homepages. Therefore, in the first phase we ask:

**RQ 6A.** How can we find related entities for which the specified relation with a given source entity holds and that satisfy the target type constraint?

**a.** How do different measures for computing co-occurrence affect the recall of a pure co-occurrence based related entity finding (REF) model?

**b.** Can a type filtering approach based on Wikipedia categories successfully be applied to REF to improve precision without hurting recall?

**c.** Can recall and precision be enhanced by combining the co-occurrence model with a context model, so as to ensure that source and target entities engage in the right relation?

To address the questions in phase one we investigate the use of a generative framework to combine three components of a REF system. The first component is a co-occurrence-based model that selects candidate entities. While, by itself, a co-occurrence-based model can be effective in identifying the potential set of related entities, it fails to rank them effectively. Our failure analysis reveals two types of error that affect precision: (1) entities of the wrong type pollute the ranking and (2) entities are retrieved that are associated with the source entity without engaging in the right relation with it. To address (1), we add a type filtering component based on category information in Wikipedia. To correct for (2), we complement the pipeline with contextual information, represented as statistical language models derived from documents in which the source and target entities co-occur. The addition of context proves beneficial for both recall and precision.

We then move away from the specific characteristics of a Wikipedia corpus and investigate the performance of our REF system on a web corpus. We conform to the official TREC definition of the REF task by adding the additional challenge of mapping entities to their homepages. Regarding the second phase we ask:

**RQ 6B.** How can we find related entities and their homepages in a web corpus?

**a.** Does the use of a larger corpus improve estimations of co-occurrence and context models?

**b.** Is the initial focus on Wikipedia a sensible approach; can it achieve comparable performance to other approaches?

**c.** Can our basic framework effectively incorporate additional heuristics in order to be competitive with other state-of-the-art approaches?

To address the questions in phase two, we add a homepage finding component that maps entities represented by Wikipedia pages to homepages. We show that our approach achieves competitive performance on the official task, especially in terms of recall. We then demonstrate the generalizability of our framework by expanding it with two heuristics: one aimed at improving type filtering, the other at co-occurrence. We find that these methods have a very positive impact on all measures.

The main contribution of this chapter is two-fold. First, we propose a transparent architecture for addressing the REF task. Second, we provide a detailed analysis of the effectiveness of its components and estimation methods, shedding light on the balance between precision and recall in the context of the REF task.

The remainder of this chapter is organized as follows. In Section 9.2 we detail our approach to the REF task. Our experimental setup is described in Section 9.3. In Section 9.4 we analyze the effectiveness of a pure co-occurrence model, a type filtering component, and adding contextual information. Improved estimations of co-occurrence and context models are considered in Section 9.5. We address the (sub)task of homepage finding (mapping entities to their homepages) in Section 9.6. In Section 9.7 we discuss TREC Entity results as well as additional heuristics that can be incorporated into our framework. We conclude in Section 9.8.

## 9.2 Approach

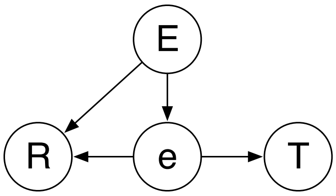
The goal of the REF task is to return a ranked list of relevant entities $e$ for a query, where a query consists of a source entity ($E$), target type ($T$) and a relation ($R$) [29]. We formalize REF as the task of estimating the probability $P(e|E, T, R)$. This probability is difficult to estimate, due to the lack of training material, which is exacerbated by the fact that relations do not come from a closed vocabulary. Also, the model should capture a particular relation conditioned on the two entities involved. To address these concerns we turn to a generative model. First, we apply Bayes' Theorem and rewrite $P(e|E, T, R)$ to:

$$P(e|E,T,R) \quad = \quad \frac{P(E,T,R|e) \cdot P(e)}{P(E,T,R)}. \tag{9.1}$$

Next, we drop the denominator as it does not influence the ranking of entities, and derive our final ranking formula as follows:

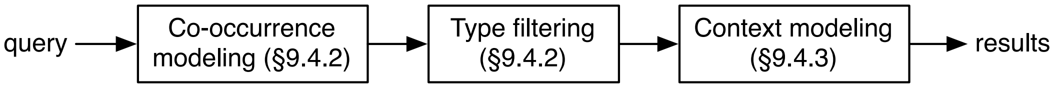
$$\begin{aligned}
& P(E,T,R|e) \cdot P(e) \\
& \quad \propto P(E,R|e) \cdot P(T|e) \cdot P(e) && (9.2) \\
& \quad = P(E,R,e) \cdot P(T|e) = P(R|E,e) \cdot P(E,e) \cdot P(T|e) \\
& \quad = P(R|E,e) \cdot P(e|E) \cdot P(E) \cdot P(T|e) && (9.3) \\
& \quad \overset{\text{rank}}{=} P(R|E,e) \cdot P(e|E) \cdot P(T|e) && (9.4)
\end{aligned}$$

In (9.2) we assume that type $T$ is independent of source entity $E$ and relation $R$. We rewrite $P(E, R|e)$ to $P(R|E, e)$ so that it expresses the probability that $R$ is generated by two (co-occurring) entities ($e$ and $E$). Finally, we rewrite $P(E, e)$ to $P(e|E) \cdot P(E)$ in (9.3) as the latter is more convenient for estimation. We drop $P(E)$ in (9.4) as it is assumed to be uniform, thus does not influence the ranking.



The generative model, shown above, functions as follows. The input entity $E$ is chosen with probability $P(E)$, which generates a target entity $e$ with probability $P(e|E)$. The input and target entities together generate a relation $R$ with probability $P(R|E, e)$. Finally, the target entity generates a type $T$ with probability $P(T|e)$.

Assuming that input entities are chosen from a uniform distribution, we are left with the following components: (i) pure co-occurrence model ($P(e|E)$), (ii) type filtering ($P(T|e)$) and (iii) contextual information ($P(R|E, e)$). We summarize the developments to come in the figure below; we analyze a pure co-occurrence model and its performance on the REF task in §9.4.1. We then add type filtering and contextual information to the pipeline; these are introduced and examined in §9.4.2 and §9.4.3, respectively. The components are combined using Equation 9.4.



## 9.3 Experimental Setup

In this section we provide details about our document collections, the tools used for named entity recognition and normanlization as well as the test topic set and evaluation measures.

### Document Collection

Our document collection is the ClueWeb09 Category B subset [96] (“CW-B” for short), with about 50 million documents, including English Wikipedia. We use the Wikipedia part of ClueWeb09 and filter out duplicate pages, page not found errors and non-English pages. This leaves us with about 5M documents, 2.6M of which correspond to unique entities. The total number of unique entity occurrences in Wikipedia documents (i.e., each unique entity occurring in a document counts only once, independent of the actual number of occurrences) is 373M.

### Entity Recognition and Normalization

While named entity recognition and normalization are not our focus, they are key preprocessing steps. We use Wikipedia as a repository of known (normalized) entities. We handle Named Entity Recognition (NER) in Wikipedia by considering only anchor texts as entity occurrences. We obtain an entity's name by removing the Wikipedia prefix from the anchor URL. For named entity normalization (NEN) we map URLs to a single entity variant. Here we make use of Wikipedia redirects that map common alternative spellings or references (e.g., "Schumacher," "Schumi" and "M. Schumacher") to the main variant of an entity ("Michael Schumacher"). Below, when using the full CW-B subset, we use the entity names as queries to an index of this collection. This effectively bypasses NER as the resulting document lists identify in which documents the entities occur. In this case, we do not perform NEN; while potentially introducing noise, we believe that the amount of data partly compensates for this.

### Test Topics

We base our test set on the TREC 2009 Entity topics. A topic consists of a source entity ($E$), a target entity type ($T$) and the desired relation ($R$) described in free text. Since we are restricting ourselves to entities in Wikipedia, we are not able to use all 20 TREC Entity topics, but only 15 of them. Specifically, we exclude three topics (#3, #8 and #13) without relevant results in Wikipedia and another two (#2 and #16) with source entities without a Wikipedia page. For the remaining topics we manually mapped the source entity to a Wikipedia page, this is the only manual intervention in the pipeline; the topics are listed in Table 9.1.

**Table 9.1:** Description of our 15 test topics. Target entity types are ORG=organization, PER=person and PROD=product.

| ID | Source entity ($E$) | Relation ($R$) | Type ($T$) |
|---|---|---|---|
| 1 | Blackberry | Carriers that Blackberry makes phones for. | ORG |
| 4 | Philadelphia, PA | Professional sports teams in Philadelphia. | ORG |
| 5 | Medimmune, Inc. | Products of Medimmune, Inc. | PROD |
| 6 | Nobel Prize | Organizations that award Nobel prizes. | ORG |
| 7 | Boeing 747 | Airlines that currently use Boeing 747 planes. | ORG |
| 9 | The Beaux Arts Trio | Members of The Beaux Arts Trio. | PER |
| 10 | Indiana University | Campuses of Indiana University. | ORG |
| 11 | Home Depot Foundation | Donors to the Home Depot Foundation. | ORG |
| 12 | Air Canada | Airlines that Air Canada has code share flights with. | ORG |
| 14 | Bouchercon 2007 | Authors awarded an Anthony Award at Bouchercon in 2007. | PER |
| 15 | SEC conference | Universities that are members of the SEC conference for football. | ORG |
| 17 | The Food Network | Chefs with a show on the Food Network. | PER |
| 18 | Jefferson Airplane | Members of the band Jefferson Airplane. | PER |
| 19 | John L. Hennessy | Companies that John Hennessy serves on the board of. | ORG |
| 20 | Isle of Islay | Scotch whisky distilleries on the island of Islay. | ORG |

We perform two types of evaluation. First, throughout §9.4 and §9.5 we focus on finding entities as represented by their Wikipedia page. We establish ground truth by extracting all primary Wikipedia pages from the TREC 2009 Entity qrels. We handle

Wikipedia redirects and duplicates in our evaluation; a Wikipedia page returned for any of the variants of a relevant entity is considered to be correct, but once found, other variants of that page are ignored. This setup constitutes a change to the original TREC REF task, arguably making it easier, therefore the reported numbers are not directly comparable with those of the TREC 2009 Entity track [29]. Our second type of evaluation, on the original TREC REF task, is performed in §9.7, where we compare our scores with those of TREC Entity participants; based on their original submissions, we also compute their Wikipedia-based evaluation scores. The evaluation script, judgements and additional resources, used in our experiments are available, see Appendix A.

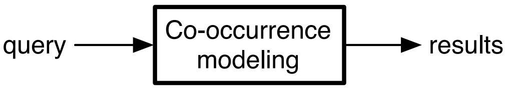
### Evaluation Metrics

We focus on two measures: precision and recall. Specifically, we use R-Precision (R-prec), where R is the number of relevant entities for a topic, and recall at rank N (R@N), where we take N to be 100, 2000 and "All" (i.e., considering all returned entities). In Table 9.12 we also report on the metrics used at the TREC 2009 Entity track: P@10, NDCG@R, and the number of primary and relevant entity homepages retrieved. We forego significance testing as we do not have the minimal number of topics (25) recommended [347].

# 9.4 An Entity Finding Pipeline

## 9.4.1 Co-Occurrence Modeling

The pure co-occurrence component is the first building block of our retrieval pipeline. It can produce a ranking of entities on its own.



Since we are planning on expanding this pipeline with additional components (that will build on the set of entities identified in this step), our main focus throughout this section will be on recall.

### Estimation

The pure co-occurrence component ($P(e|E)$) expresses the association between entities based on occurrences in documents, independent of context (i.e., the actual content of documents). To express the strength of co-occurrence between $e$ and $E$ we use a function $\text{cooc}(e, E)$ and estimate $P(e|E)$ as follows:

$$P(e|E) = \frac{\text{cooc}(e, E)}{\sum_{e'} \text{cooc}(e', E)}.$$

We consider four settings of $\text{cooc}(e, E)$: (i) as maximum likelihood estimate, (ii) $\chi^2$ hypothesis test, (iii) pointwise mutual information and (iv) log likelihood ratio. We briefly recall their details [229].

*(i) Maximum likelihood estimate (MLE)* uses the relative frequency of co-occurrences between $e$ and $E$ to determine the strength of their association:

$$\text{cooc}_{MLE}(e, E) = c(e, E)/c(E),$$

where $c(e, E)$ is the number of documents in which $e$ and $E$ co-occur and $c(E)$ is the number of documents in which $E$ occurs.

*(ii) The* $\chi^2$ hypothesis test determines if the co-occurrence of two entities is more likely than just by chance. A $\chi^2$ test is given by:

$$\text{cooc}_{\chi^2}(e, E) = \frac{N \cdot (\, c(e, E) \cdot c(\overline{e}, \overline{E}) - c(e, \overline{E}) \cdot c(\overline{e}, E)\,)^2}{c(e) \cdot c(E) \cdot (\, N - c(e)\,) \cdot (\, N - c(E)\,)},$$

where $N$ is the total number of documents, and $\overline{e}$, $\overline{E}$ indicate that $e$, $E$ do not occur, respectively (i.e., $c(\overline{e}, \overline{E})$ is the number of documents in which neither $e$ or $E$ occurs).

*(iii) Pointwise mutual information (PMI)* determines the amount of information we gain if we observe $e$ and $E$ together. It is useful to determine independence between entities, but of less value to determine how dependent two entities are. PMI is given by:

$$\text{cooc}_{PMI}(e, E) = \log c(e, E)/(c(e) \cdot c(E)).$$

*(iv) Log likelihood ratio* is another measure that determines dependence and is more reliable than PMI [126]. It is defined as:

$$\begin{aligned}\text{cooc}_{LLR}(e, E) &= 2(L(p_1, k_1, n_1) + L(p_2, k_2, n_2) \\ &\quad -L(p, k_1, n_1) - L(p, k_2, n_2)),\end{aligned}$$

where $k_1 = c(e, E)$, $k_2 = c(e, \overline{E})$, $n_1 = c(E)$, $n_2 = N - c(E)$, $p_1 = k_1/n_1$, $p_2 = k_2/n_2$ and $p = (k_1 + k_2)/(n_1 + n_2)$, while $L(p, k, n) = k \log p + (n - k) \log(1 - p)$.

## Results

Table 9.2 shows the results of the different estimation methods for the pure co-occurrence model. Out of the four methods, $\chi^2$ is a clear winner while PMI performs worst on all metrics. MLE and LLR deliver very similar scores; their recall is comparable to that of $\chi^2$, but they achieve much lower R-precision. All estimators return entities that co-occur at least once with the source entity, hence R@All is the same for all, just over 93%.

**Table 9.2:** Results of the pure co-occurrence models.

| Co-occ. | R-prec | R@100 | R@2000 | R@All |
|---|---|---|---|---|
| MLE | .0399 | .2957 | .7501 | .9311 |
| $\chi^2$ | **.1099** | **.3268** | **.8273** | .9311 |
| PMI | .0244 | .0981 | .4888 | .9311 |
| LLR | .0399 | .2957 | .7184 | .9311 |

### Analysis

The numbers presented in Table 9.2 demonstrate that simple co-occurrence statistics can achieve reasonable recall scores and can be used to obtain a candidate set of entities (e.g., top 2000) that can then be further examined by subsequent components in the pipeline. Fig. 9.3 (Top) shows the R@2000 scores of the methods per topic. For most topics at least one of the methods achieves high recall, with the exception of topic #4.

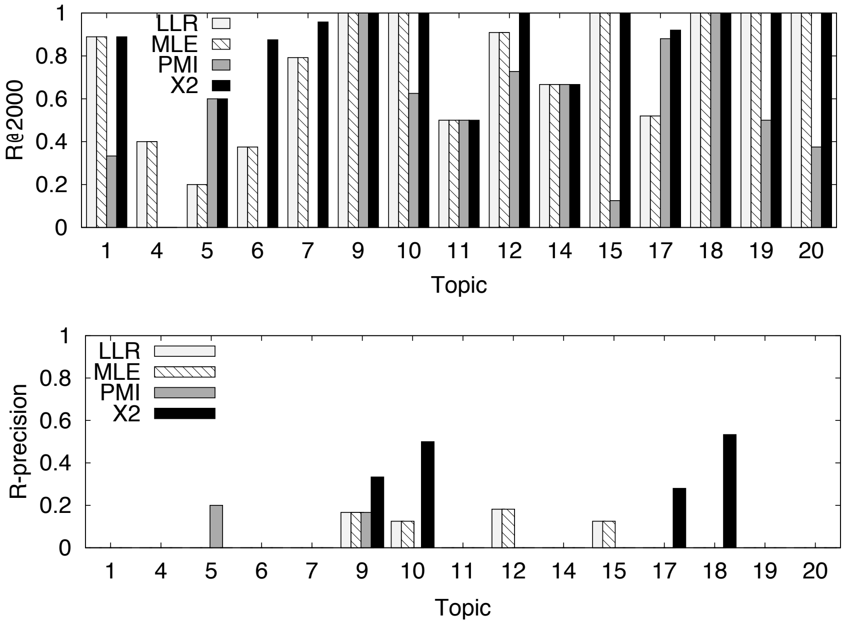


**Figure 9.3:** Per topic R@2000 (Top) and R-precision (Bottom) scores for the pure co-occurrence estimation methods.

Unlike recall, R-precision scores are very low, suggesting that pure co-occurrence is not enough to solve the REF task. Fig. 9.3 (Bottom) shows that all methods score zero on R-precision on all but 4 topics. To identify the types of errors made, we take topic #17 (cf. Table 9.1) as an example and list the top 10 entities produced by our co-occurrence methods in Table 9.3.[1] Clearly, $\chi^2$ finds relevant entities (bold) mixed with non-relevant entities that are not of the target type $T$ (normal font). The other methods suffer more heavily from this type of error and fail to return any relevant entities in the top 10. We also see another type of error: entities that are of the right type, but do not satisfy the target relation with the source entity. Note that one of the entities (indicated by †) is relevant, but not identified as such, as its Wikipedia page does not occur in the qrels.

Different co-occurrence methods display distinct characteristics in what they consider as strongly associated. MLE and LLR focus on popular entities; the top ranking entity, "Charitable Organization," occurs 5,271,075 times. The other extreme is demonstrated by PMI, which favors rare entities: the top ranking entity occurs 2 times and exclusively with the source entity. Finally, $\chi^2$ performs well when entities co-occur frequently with

[1] We use topic #17 as a running example throughout the chapter, to illustrate the impact of additional ranking and filtering components.

**Table 9.3:** Top 10 entities for topic #17. Relevant entities in bold, entities of the wrong type in roman, and entities of the right type but in the wrong relation in italics. MLE and LLR have the same top 10 ranking and are not displayed separately. † indicates a relevant entity for which no Wikipedia page is available in the qrels.

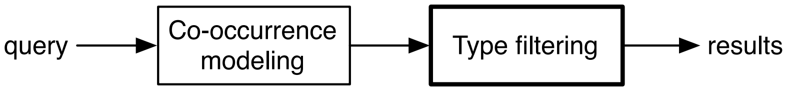
| | PMI | MLE/LLR | $\chi^2$ |
|---|---|---|---|
| 1 | Y'all (magazine) | Charitable organization | Iron Chef America |
| 2 | *Wayne Harley Brachman* | United States | **Paula Deen** |
| 3 | Veniero's | New York City | **Bobby Flay** |
| 4 | The Hungry Detective | 2007 | *Alton Brown*† |
| 5 | The FN Dish | 2006 | Fine Living |
| 6 | Super Suppers | NBC | **Rachael Ray** |
| 7 | Sugar Sugar, Inc | 2008 | Emeril Live |
| 8 | Stonewall Kitchen | Website | Unwrapped |
| 9 | *Raul Musibay* | Internet Movie Database | **Giada De Laurentiis** |
| 10 | Party Line with The Hearty Boys | American Broadcasting Company | Real Age |

the source entity and less with others; the top ranked entity occurs in 327 documents, in 187 cases together with the source entity, for the second best entity these numbers are 148 and 106, respectively.

As all methods and topics suffer from entities of the wrong type polluting the rankings, we address this issue next.

### 9.4.2 Type Filtering

To combat the problem that results produced by the pure co-occurrence model are polluted by entities of the wrong type, we add a type filtering component on top of the pure co-occurrence model; this is indicated by the thick box in the figure below.



The challenge will be to maintain the high recall levels attained by the pure co-occurrence model while improving precision. Recall from (9.4), that type filtering is formalized as $P(T|e)$. The entity type filtering component $P(T|e)$ expresses the probability that an entity $e$ is of the target type. Combined with the pure co-occurrence model, it yields the following model for ranking entities:

$$P(e|E,T) \propto P(e|E) \cdot P(T|e). \tag{9.5}$$

#### Estimation

In order to perform type filtering we exploit category information available in Wikipedia. We map each of the (input) entity types ($T \in \{PER, ORG, PROD\}$) to a set of Wikipedia categories ($\mathrm{cat}(T)$) and we create a similar mapping from entities to categories ($\mathrm{cat}(e)$). The former is created manually, while the latter is granted to us in the form of page-category assignments in Wikipedia (recall that Wikipedia pages correspond to entities). With these two mappings we estimate $P(T|e)$ as follows:

$$P(T|e) = \begin{cases} 1 & \text{if } \mathrm{cat}(e) \cap \mathrm{cat}^{L_n}(T) \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

Since the Wikipedia category structure is not a strict hierarchy and the category assignments are imperfect [114], we (optionally) expand the set of categories assigned to each target entity type $T$, hence write $\mathrm{cat}^{L_n}(T)$, where $L_n$ is the chosen level of expansion.

For the initial mapping of types to categories, $\mathrm{cat}^{L_1}(T)$, we manually assign a number of categories to each type as in [187]. To the person type we map categories that end with "birth," "death," start with "People" and the category "Living People." To the organization type we map categories that start with "Organizations" or "Companies" and to the product type we map categories starting with "Products" or ending with "introductions." Next, we use the Wikipedia category hierarchy to expand this set by adding all direct child categories of the categories in $L_1$, to obtain our first expansion set $L_2$. We continue expanding the categories this way, one level at a time until no new categories are added.

While this particular form of type filtering is specific and tailored to Wikipedia, it is reasonable to assume that a named entity recognizer would provide us with high-level type information; therefore, it is not a limitation of the generalizability of our framework.

### Results

By varying the expansion levels, we can optimize type filtering in two ways: for (R-)precision and for recall (R@2000). We first investigate the optimal levels of expansion for R-precision. Figure 9.4 (Left) shows that R-precision increases when moving from level 0 (no filtering, shown on the right end of the plot) to level 2 expansion, but drops as the level of category expansion is further increased. This is in line with our expectation that an increasing number of categories allow more entities of the wrong type; because of the imperfection of the Wikipedia category structure, expansion results in the addition of many irrelevant categories.

As to recall, Figure 9.4 (Right) shows R@2000 vs. level of expansion. R@2000 first increases and then decreases (PMI) or remains the same (MLE, LLR, and $\chi^2$) as cate-

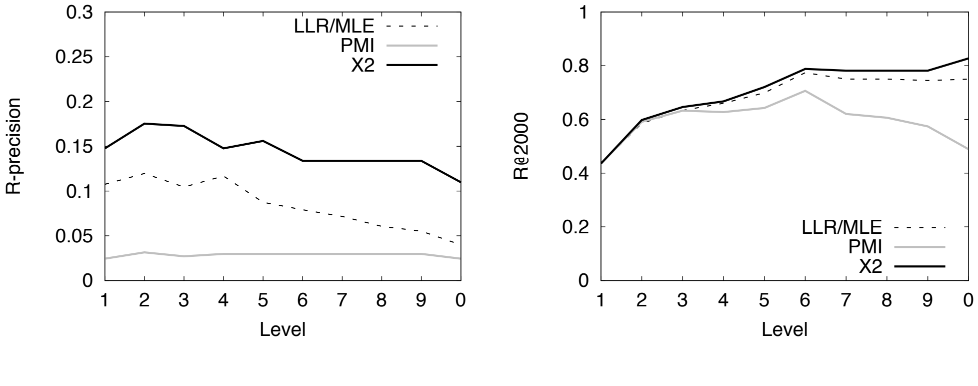


**Figure 9.4:** R-precision and R@2000 at increasing levels of category expansion.

**Table 9.4:** Results of type filtering with optimal level of filtering.

| Co-occ. | R-Prec | R@100 | R@2000 | R@All |
|---|---|---|---|---|
| *Optimized for Precision* | | | | |
| MLE | .1196 (+200%) | .3827 (+29%) | .5924 (-27%) | .5977 (-56%) |
| $\chi^2$ | **.1753** (+60%) | **.3976** (+22%) | **.5977** (-38%) | **.5977** (-56%) |
| PMI | .0316 (+30%) | .1920 (+96%) | .5910 (+21%) | .5977 (-56%) |
| LLR | .1196 (+200%) | .3827 (+29%) | .5857 (-23%) | .5977 (-56%) |
| *Optimized for Recall* | | | | |
| MLE | .0791 (+98%) | .3915 (+32%) | .7740 (+3%) | .8667 (-7%) |
| $\chi^2$ | **.1338** (+22%) | **.5012** (+53%) | **.7881** (-5%) | .8667 (-7%) |
| PMI | .0298 (+22%) | .1344 (+37%) | .7065 (+45%) | .8667 (-7%) |
| LLR | .0791 (+98%) | .3915 (+32%) | .7740 (+8%) | .8667 (-7%) |

gories are expanded. At level 6 or beyond, the number of non-relevant entities allowed into the ranking is large enough to push relevant entities out of the top 2000. Uniformly applying category expansion down to the same level for all types is not necessarily optimal; some relevant entities of type organization are removed at expansion levels smaller than 6, while those of type person are only filtered out at level 1.

Table 9.4 shows the results of applying type filtering to the pure co-occurrence model optimized for precision (top rows) and for recall (bottom rows); relative changes are given w.r.t. the results in Table 9.2. We see an increase in R-precision for all methods. The best results when optimized for R-precision are achieved with $\chi^2$, but we see large relative improvements for all methods in R-precision. Type filtering causes recall to drop sharply at low ranks; achieving max 60% R@2000 as opposed to 83% without filtering (cf. Table 9.2). The best R-precision scores averaged over all topics are achieved with type filtering at level 2 ($L_2$); this is the setting we will use when reporting scores optimized for R-precision.

As to the results optimized for recall, we find, again, that all methods improve both R-precision and R@100. The R@100 and R@2000 results suggest that $\chi^2$ ranks relevant entities closer to the top 100 than the other methods. We find 79% of the relevant entities in the top 2000 and in total only 7% of the relevant entities are lost by type filtering. We achieve the best recall scores with type filtering at level 6 ($L_6$); this is the value used for recall-optimized settings reported in the remainder of the chapter.

By varying the level of expansion we can effectively aim either for R-precision or for R@2000, without hurting the other. This decision is likely to be made depending on whether this is the last component of the pipeline or results will be passed along for downstream processing. Optimizing category expansion levels for precision and recall carry the risk of overfitting, especially on a small topic set. Our aim with this tuning, however, is not to squeeze out the last bit of performance, but to demonstrate that type filtering can effectively be used to balance precision and recall. Two reasons reduce the risk of overfitting: (i) the target types are of a high level causing the granularity of category expansion to be of a coarse nature and (ii) the level of expansion is the same for all types.

### Analysis

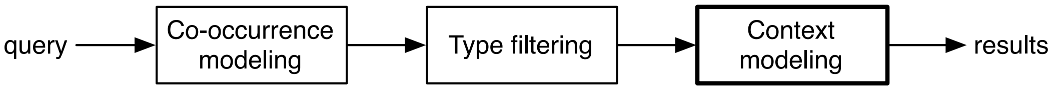
Table 9.5 shows the top 10 results for topic #17 after type filtering. We see that type filtering effectively removes entities of the wrong type from the ranking: all remaining entities are of type PER and no relevant entities were removed. Another type of error—entities of the right type but not engaged in the required relation $R$ to the source entity $E$ (“Chefs with a show on the Food Network”)—, is now more prominent (see, e.g., *Oprah Winfrey* and *George W. Bush*). In §9.4.3 we address this type of error by adding context to the co-occurrence model and only admitting co-occurrences in contexts that display evidence of the required relation.

**Table 9.5:** Top 10 entities for topic #17 with type filtering ($L_2$).

| | PMI | MLE/LLR | $\chi^2$ |
|---|---|---|---|
| 1 | Wayne Harley Brachman | Alton Brown$^\dagger$ | **Paula Deen** |
| 2 | Kerry Vincent | **Rachael Ray** | **Bobby Flay** |
| 3 | Jacqui Malouf | **Bobby Flay** | **Alton Brown** |
| 4 | Glenn Lindgren | Chef | **Rachael Ray** |
| 5 | Geof Manthorne | **Paula Deen** | **Giada De Laurentiis** |
| 6 | Anna Pump | **Mario Batali** | **Mario Batali** |
| 7 | **Alexandra Guarnaschelli** | Oprah Winfrey | **Guy Fieri** |
| 8 | Kenji Fukui | George W. Bush | **Michael Symon** |
| 9 | Warren Brown | **Giada De Laurentiis** | **Cat Cora** |
| 10 | Tatsuo Itoh | **Emeril Lagasse** | Charles Scripps |

## 9.4.3 Adding Context

To suppress entities that are of the right type $T$ but that do not engage in the required relation $R$, we add an additional component: modeling contextual information (the thick box below):



Recall from (9.4) that the context of a co-occurrence model is captured as $P(R|E, e)$. Putting things, this is how we rank (§9.2):

$$P(e|E, T, R) \propto P(R|E, e) \cdot P(e|E) \cdot P(T|e). \tag{9.6}$$

### Estimation

The $P(R|E, e)$ component is the probability that a relation is generated from (“observable in”) the context of a source and candidate entity pair. We represent the relation between a pair of entities by a co-occurrence language model ($\theta_{Ee}$), a distribution over terms taken from documents in which the source and candidate entities co-occur. By assuming independence between the terms in the relation $R$ we arrive at the following estimation:

$$P(R|E, e) = P(R|\theta_{Ee}) = \prod_{t \in R} P(t|\theta_{Ee})^{n(t,R)}, \tag{9.7}$$

where $n(t, R)$ is the number of times $t$ occurs in $R$. To estimate the co-occurrence language model $\theta_{Ee}$ we aggregate term probabilities from documents in which the two entities co-occur:

$$P(t|\theta_{Ee}) = \frac{1}{|D_{Ee}|} \sum_{d \in D_{Ee}} P(t|\theta_d), \tag{9.8}$$

where $D_{Ee}$ denotes the set of documents in which $E$ and $e$ co-occur and $|D_{Ee}|$ is the number of these documents. $P(t|\theta_d)$ is the probability of term $t$ within the language model of document $d$:

$$P(t|\theta_d) = \frac{n(t, d) + \mu \cdot P(t)}{\sum_t' n(t', d) + \mu}, \tag{9.9}$$

where $n(t, d)$ is the number of times $t$ appears in document $d$, $P(t)$ is the collection language model, and $\mu$ is the Dirichlet smoothing parameter, set to the average document length in the collection [201].

### Results

Table 9.6 shows the results of the context dependent model (including type filtering), optimized for precision (Top) and recall (Bottom); relative changes are w.r.t. the corresponding cells in Table 9.4. In both cases, R-precision and R@100 are substantially improved, while R@2000 and R@All remain the same or slightly improve. The best performing method across the board is MLE, but there is only a slight difference with the LLR and $\chi^2$ scores. PMI achieved the largest relative improvements, but it still lags behind the other three methods for both R-precision and R@100.

**Table 9.6:** Results of the context dependent model (including type filtering).

| Co-occ. | R-Prec | R@100 | R@2000 | R@All |
|---|---|---|---|---|
| *Optimized for Precision* | | | | |
| MLE | **.2099** (+76%) | .4929 (+29%) | .5950 (0%) | .5977 (0%) |
| $\chi^2$ | .2094 (+19%) | .4631 (+16%) | **.5977** (0%) | .5977 (0%) |
| PMI | .0678 (+115%) | .2715 (+41%) | .5889 (-1%) | .5977 (0%) |
| LLR | .2032 (+70%) | **.4955** (+29%) | .5950 (+2%) | .5977 (0%) |
| *Optimized for Recall* | | | | |
| MLE | **.1905** (+140%) | **.6221** (+60%) | .8344 (+8%) | .8667 (0%) |
| $\chi^2$ | .1798 (+34%) | .5708 (+14%) | **.8459** (+7%) | .8667 (0%) |
| PMI | .0678 (+127%) | .3313 (+147%) | .8315 (+18%) | .8667 (0%) |
| LLR | .1705 (+115%) | .5997 (+53%) | .8316 (+7%) | .8667 (0%) |

### Analysis

Looking at Table 9.7 we see that several entities have been replaced with others, “fresh” ones. Some that were in the “wrong” relation (i.e., *Oprah* and *Bush*, cf. Table 9.5) have been removed. For both MLE and LLR *Chef* and *Celebrity* are now returned at the top ranks; these entities are frequently observed together with relation terms, i.e., *chefs with a show on the Food Network* (and type filtering erroneously recognizes them as people). Some entities occur only in a handful of documents (<10), as a consequence of which very little evidence of the relation $R$ can be found in their contexts (examples from the qrels include *Alexandra Guarnaschelli*, *Aida Mollenkamp*, *Daisy Martinez*). We observe a larger performance gain for the MLE and LLR based models than for $\chi^2$. By introducing context, the result lists—consisting of frequent entities, favored by these models—are supplemented with entities that occur in suitable contexts. The entities found by the $\chi^2$ model show a large overlap with those identified on the basis of context, hence limiting the performance gain.

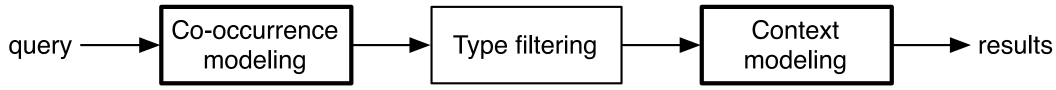
These observations point to two issues with using Wikipedia as a corpus: (1) estimates for the pure co-occurrence models are unreliable and (2) the corpus is too small for constructing accurate context models, i.e., there is simply not enough textual material for certain entities. In §9.5 we address these problems by considering a larger corpus to improve our estimations of the pure co-occurrence model and to gather contexts for more robust context models.

**Table 9.7:** Top 10 entities for topic #17 after adding context.

| R | PMI | MLE | LLR | $\chi^2$ |
|---|---|---|---|---|
| 1 | Gennaro Contaldo | Chef | Chef | **Bobby Flay** |
| 2 | Asako Kishi | Celebrity | Celebrity | **Anne Burrell** |
| 3 | Yutaka Ishinabe | B. Smith | B. Smith | **Robert Irvine** |
| 4 | Karine Bakhoum | **Bobby Flay** | **Bobby Flay** | **Tyler Florence** |
| 5 | Masahiko Kobe | **Mario Batali** | **Mario Batali** | **Aaron McCargo, Jr.** |
| 6 | Tamio Kageyama | **Tyler Florence** | Bravo | **Mario Batali** |
| 7 | Toshiro Kandagawa | Bravo | **Rachael Ray** | Sunny Anderson$^\dagger$ |
| 8 | Alpana Singh | **Rachael Ray** | **Tyler Florence** | **Guy Fieri** |
| 9 | Katie Lee Joel | **Robert Irvine** | **Paula Deen** | **Giada De Laurentiis** |
| 10 | Kazuko Hosoki | **Anne Burrell** | Alton Brown$^\dagger$ | Kevin Brauch |

## 9.5 Improved Estimations

We investigate how using a large corpus (CW-B, §9.3) for estimating our models can overcome the issue that for some entities their co-occurrences are limited to a small set of pages and that for some there is not enough context to be able to derive a robust language model. These changes affect two components of our pipeline:



### Estimation

Using a large corpus for REF presents two challenges: NER on the entire corpus is time consuming and the shear number of entities becomes prohibitively large for any but the simplest of methods. To deal with these issues, we limit ourselves to a "working entity set" consisting of the top 2000 entities produced by the context dependent co-occurrence model (estimated on Wikipedia). We chose the entities returned for PMI without filtering as this produced the highest R@2000 (i.e., 87%). For our pure co-occurrence model we need, for each source-candidate entity pair, the number of documents in which they occur separately and the number of documents in which they co-occur (§9.4.1). We estimate these numbers by submitting the top 2000 entities as queries to an indexed version of CW-B, which returns the document IDs. We do the same for the source entities and then compare the document ID lists to find documents with co-occurrences. In order to estimate the context dependent model we consider only documents containing the source entity. We then create the co-occurrence model for a source-candidate entity pair by using the candidate as a query, effectively collecting all documents in which they co-occur.

### Results

Table 9.8 shows the results for the co-occurrence models estimated using CW-B; relative changes in columns 2 and 3 are w.r.t. Table 9.4; those in columns 4 and 5 are w.r.t. Table 9.6. In the top left quadrant R-precision and R@100 of the pure co-occurrence model (optimized for precision) both improve over the same model using Wikipedia-based estimates for all methods: adding data solves the issue of sparse co-occurrences.

**Table 9.8:** Results for the context dependent model with filtering and estimations using the CW-B corpus.

| Co-occ. | Pure Co-Occurrence | | Context Dependent | |
|---|---|---|---|---|
| | R-Prec | R@100 | R-Prec | R@100 |
| *Optimized for Precision* | | | | |
| MLE | .1512 (+26%) | **.5423** (+42%) | .1898 (-11%) | **.5423** (+10%) |
| $\chi^2$ | **.2382** (+36%) | .4891 (+23%) | **.2623** (+25%) | .4747 (+3%) |
| PMI | .1363 (+331%) | .3545 (+85%) | .0649 (-8%) | .3137 (+16%) |
| LLR | .1540 (+29%) | .4947 (+29%) | .1767 (-15%) | .4873 (-2%) |
| *Optimized for Recall* | | | | |
| MLE | .0799 (+1%) | **.5821** (+49%) | .0966 (-97%) | **.6982** (+12%) |
| $\chi^2$ | **.2281** (+70%) | .5474 (+9%) | **.2399** (+33%) | .5418 (-5%) |
| PMI | .0966 (+224%) | .3748 (+179%) | .0577 (-18%) | .3308 (0%) |
| LLR | .0793 (0%) | .5655 (+44%) | .0988 (-73%) | .6469 (+8%) |

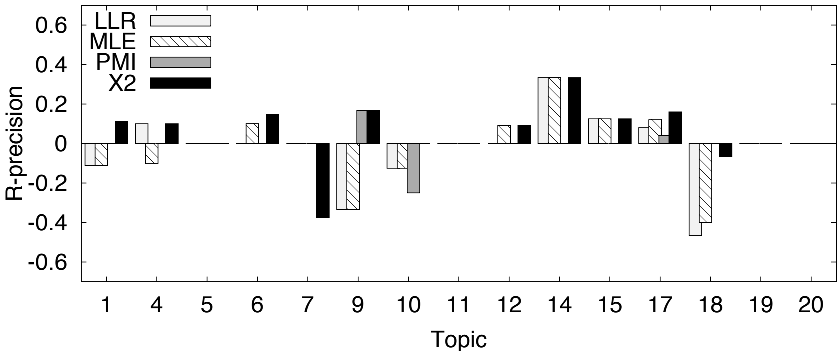


**Figure 9.5:** Differences in R-precision per topic; context dependent model using CW-B vs. Wikipedia. A negative score indicates greater precision for the Wikipedia-based model.

In the top right quadrant we see that the addition of context, using CW-B documents, further improves the $\chi^2$ results, similar to what we saw when adding context in §9.4.3. In this case however, R-precision is worse than that achieved by the Wikipedia-based model for MLE, PMI, and LLR. In contrast, $\chi^2$ shows a 25% improvement when adding CW-B documents. The models optimized for recall demonstrate a similar behavior; the pure co-occurrence model (bottom left) improves over the Wikipedia-based model, while the context dependent one does not, except for $\chi^2$. For the $\chi^2$ method, we seem to have reached a good balance between precision and recall, continuing to improve R-precision with improvements or little effect on R@100. For the other methods, the picture is more diverse, especially for recall-optimized type filtering.

## Analysis

Fig. 9.5 shows the difference per topic in R-precision of the context dependent model using either CW-B or Wikipedia; a negative score indicates higher R-precision for the Wikipedia-based model. Using Wikipedia documents greatly improves precision scores

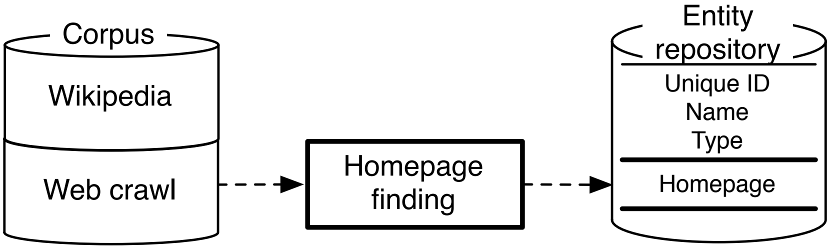
**Table 9.9:** Top 10 entities with improved estimations for topic #17; some names truncated for layout reasons.

| R | PMI | MLE | LLR | $\chi^2$ |
|---|---|---|---|---|
| 1 | Tamio Kageyama | Alton Brown$^\dagger$ | Alton Brown$^\dagger$ | **Bobby Flay** |
| 2 | Kazuko Hosoki | **Rachael Ray** | **Rachael Ray** | **Paula Deen** |
| 3 | Toshiro Kandagawa | **Bobby Flay** | **Bobby Flay** | Alton Brown$^\dagger$ |
| 4 | Ron Siegel | **Mario Batali** | **Paula Deen** | **Michael Symon** |
| 5 | Mayuko Takata | **Paula Deen** | **Mario Batali** | **Giada De Laure.** |
| 6 | Asako Kishi | Chef | Chef | **Rachael Ray** |
| 7 | David Evangelista | **Cat Cora** | **Cat Cora** | **Mario Batali** |
| 8 | Dave Spector | **Emeril Lagasse** | **Emeril Lagasse** | **Cat Cora** |
| 9 | Kazushige Nagash. | **Michael Symon** | **Giada De Laure.** | **Guy Fieri** |
| 10 | Chua Lam | **Giada De Laure.** | **Michael Symon** | Kenji Fuku |

for three of the topics for MLE and LLR. As we look into these topics we see that the Wikipedia page of each source entity contains a full list of all the relevant entities (e.g., "members of the Beaux Arts Trio" and "members of Jefferson Airplane"), making them relatively easy, with external evidence likely to generate noise. However, the CW-B based model improves R-precision scores on a number of topics, which suggests that we can effectively use a larger corpus to handle a more diverse set of topics. In our running example (cf. Table 9.9) we now achieve a near perfect ranking for $\chi^2$, MLE and LLR; PMI still finds only rare entities.

## 9.6 Homepage Finding

Up to this point in the retrieval pipeline entities have been identified by their Wikipedia page. However, according the TREC Entity track, an entity is uniquely identified by its homepage, therefore we now focus on the homepage finding component in our architecture:



### Approach

The 2009 Entity track allows up to three homepages and a Wikipedia page to be returned for each entity and judges pages as either primary,[2] relevant or non-relevant. Here, we define homepage finding as the task of returning the primary homepage for an entity. Our approach combines language modeling based homepage finding and link-based approaches (see below), as a linear mixture with equal weights on the components.

### Document-based Homepage Ranking

We address homepage finding as a document retrieval problem, and employ a standard language modeling approach with uniform priors [309]. The goal is to obtain a ranking of documents based on the probability that a document is a homepage given an entity's name ($P(d|e_n)$). We reformulate using Bayes' rule and rank homepages according to the query likelihood: here, we use the name of the entity $e_n$ as a query, $P(q = e_n|d)$. Successful approaches to named page and homepage finding use a combination of multiple document fields to represent documents [102, 259]. Following [259], we estimate

[2] A primary homepage is the main page about, and in control of, the entity, (e.g., `www.lufthansa.com`), whereas a relevant page merely mentions the entity (e.g., `www.staralliance.com/en/about/airlines/lufthansa/`).

$P(e_n|d)$ as a linear mixture of four components, constructed from the body ($P(e_n|f_b)$), title ($P(e_n|f_t)$), header ($P(e_n|f_h)$) and inlink ($P(e_n|f_i)$) fields:

$$P(e_n|d) = \lambda_b \cdot P(e_n|f_b) + \lambda_t \cdot P(e_n|f_t) + \lambda_h \cdot P(e_n|f_h) + \lambda_i \cdot P(e_n|f_i),$$

where each $\lambda$ is between 0 and 1 and $\sum_{\lambda \in \{\lambda_b, \lambda_t, \lambda_h, \lambda_i\}} \lambda = 1$. The specific setting of the $\lambda$ parameters of the model are estimated empirically, see below.

### Link-based Homepage Ranking

Since our REF system identifies entities by their Wikipedia pages, it is natural to use the information on those pages for homepage finding; external links often contain a link to the entity's homepage [187, 302]. We, again, view this as a ranking problem and estimate the probability that document $d$ is the homepage of entity $e$ given a link $e_{wl}$ on the entity's Wikipedia page: $P(d|e_{wl})$. We set this probability proportional to the position of the link among all external links on the Wikipedia page (pos($e_{wl}$)). Since we have to return "valid" homepages (i.e., that are present in CW-B), we perform an additional filtering step, and exlude URLs from our ranking that do not exist in CW-B.

We also employ a method based on DBpedia, which provides a list of entities with the URL of their homepage.[3] While these homepages may be more reliable than those found through the earlier external links strategy, the coverage of this method is limited. We set the probability of a homepage given a DBpedia URL, $P(d|e_{dl})$, to 1 if the URL exists in CW-B, and to 0 otherwise. To take advantage of the high quality, but sparse, data in DBpedia, while maintaining high coverage through external links in Wikipedia, we combine the external link and DBpedia strategies using a mixture model:

$$\lambda \cdot P(d|e_{wl})(1 - \lambda) \cdot P(d|e_{dl}),$$

for the sake of simplicity, we set equal weights to both components, i.e., $\lambda = 0.5$.

### Evaluation

Before we incorporate a document-based or link-based homepage ranking method as the homepage finding component into the end-to-end retrieval process, we evaluate their performance on a homepage finding task. For this purpose we create a test set of homepage finding topics from the TREC 2009 Entity qrels (Entity track-hp); we consider each entity with a primary homepage as a topic, and take the homepage as relevant document. The homepage finding topics and qrels are included in the resources made available, see Appendix A.

We first turn towards the estimation of the parameters of the document based homepage ranking method ($P(e_n|d)$), i.e., $\{\lambda_b, \lambda_t, \lambda_h, \lambda_i\}$. Parameter estimation is done in two ways. The first uses the TREC 2002 Web track data [102]. Our second parameter estimation method utilizes Wikipedia, using the page title as a name for the entity and considering external links with "official website" in their anchor text as homepages of that entity.

[3] Available at `http://wiki.dbpedia.org/Downloads33`.

The top two rows of Table 9.10 show the homepage finding performance of our document-based homepage ranking method with optimal settings on the web track and Wikipedia (Wikipedia-hp) data sets. The bottom three rows show the performance of the document-based homepage ranking method on the Entity track-hp evaluation set with optimal parameter settings for the Web track, Wikipedia-hp, and Entity track-hp sets respectively. Performance on the Web track's topics (MRR $0.6856$) is comparable to the best approaches at TREC 2002, however, this setting does not perform very well on CW-B (MRR $0.3569$). We find that the relative high weight assigned to the body parameter is recall oriented and causes many pages to be returned that push the primary homepage lower down the ranking. With the increase in size of the collection from the Web track (.GOV 18GB) to CW-B (1.5TB), the parameters should be tuned towards precision. The settings optimized for the Wikipedia topics, improve over the results obtained by optimizing for the Web track (MRR $0.47$). In this case less weight is given to the body of the document and the in-link field becomes more important. As a sanity check we also optimize the parameters on our test set of homepage finding topics, bottom row Table 9.10, and observe that the settings derived from the Wikipedia topics achieve close to optimal performance. We will use these settings (i.e., based on the Wikipedia topics) for homepage ranking in the remainder of this chapter.

**Table 9.10:** Homepage ranking results with optimal settings for the Web track, Wikipedia (Wikipedia-hp) and Entity track (Entity track-hp) based homepage finding topics.

| evaluation set | MRR | MAP | $\lambda_b$ | $\lambda_h$ | $\lambda_i$ | $\lambda_t$ |
|---|---|---|---|---|---|---|
| Web track 2002 | 0.6856 | 0.6749 | 0.7 | 0.1 | 0.1 | 0.1 |
| Wikipedia-hp | 0.1929 | 0.1929 | 0.0 | 0.0 | 0.9 | 0.1 |
| Entity track-hp | 0.3569 | 0.3051 | 0.7 | 0.1 | 0.1 | 0.1 |
| Entity track-hp | 0.4709 | 0.3938 | 0.0 | 0.0 | 0.9 | 0.1 |
| Entity track-hp | 0.4815 | 0.4062 | 0.0 | 0.0 | 0.4 | 0.6 |

Turning to the evaluation of the link-based homepage finding methods, we find that by itself, the DBpedia-based method results in very low scores (MRR $0.08$), due its low coverage, see Table 9.11. Using external links from Wikipedia pages of entities we achieve a more acceptable score (MRR $0.44$). The combined link-based method results only in minor improvements (MRR $0.45$); this is not surprising given that DBpedia is extracted from Wikipedia. The best performance overall is achieved when we use a linear mixture with equal weights of the language modeling-based approach and the combined link-based approach (MRR $0.62$); this is the method that we use in the homepage finding component in the remainder of this chapter.

## 9.7 Discussion

We now perform an end-to-end evaluation on the task specified at the TREC Entity track. According to the track's definition, up to 3 homepages and a Wikipedia page may be

**Table 9.11:** Results for link ranking with different data sources and the combination with the homepage ranking approach.

| | MRR | MAP |
|---|---|---|
| DBpedia | .0800 | .0658 |
| external links | .4467 | .3570 |
| link-based | .4517 | .3653 |
| link-based + document-based | .6248 | .5330 |

returned for each entity; each is judged on a 3-point scale (non-relevant, relevant or primary). We combine the pipeline developed in §9.4.1–§9.4.3 with the homepage finding component developed in §9.6. Table 9.12 presents the results. The *Baseline* row corresponds to our best performing run on CW-B (cf. §9.5). Note that we still only consider the 15 topics described in §9.3. Observe that our recall-oriented model (r1) outperforms other Entity track approaches in terms of the total number of primary pages found (#pri), while the precision-oriented model (p1) is in the top 6 in terms of precision ($P10$).

Next we take a look at how competitive our results are when we apply heuristic methods that were popular at the Entity track to our model. We experiment with two additional techniques.

### Improved Type Filtering

Serdyukov and de Vries [302] use the high quality type definitions provided by the DBpedia ontology[4] to perform type filtering. We follow this approach and map the ontology categories "Person" and "Organization" to their respective topic target types ($\mathit{PER}$ and $\mathit{ORG}$). We associate the class "Resource" with the product target type ($\mathit{PROD}$), as there is no specific product category in the ontology. In case an entity does not occur in the ontology, we fall back to our Wikipedia-based filtering (either precision- or recall-oriented), as described in §9.4.2. We incorporate this in the type filtering component of our model as follows:

$$P(T|e) = \begin{cases} 1 & \text{if } \mathrm{ont}(e) \neq \emptyset \text{ and } T \cap \mathrm{ont}(e) \neq \emptyset \\ 1 & \text{if } \mathrm{ont}(e) = \emptyset \text{ and } \mathrm{cat}(e) \cap \mathrm{cat}^{L_n}(T) \neq \emptyset \\ 0 & \text{otherwise,} \end{cases}$$

where $T \in \{\mathit{PER}, \mathit{ORG}, \mathit{PROD}\}$ and $\mathrm{ont}(e)$ returns the set of types for an entity in the DBpedia ontology.

The combination of our category-based filtering approach with DBpedia based filtering has a positive effect on both precision and recall, see Table 9.12 (p2 and r2). The two approaches complement each other as category-based filtering covers all entities, but is imprecise, while filtering based on the DBpedia ontology is precise, but only covers some of the entities in Wikipedia.

[4] Obtained from `http://wiki.dbpedia.org/Ontology`.

**Table 9.12:** Comparison of our best runs and TREC 2009 results. Wikipedia pages are counted as primary homepages.

| | TREC evaluation | | | | WP evaluation | |
|---|---|---|---|---|---|---|
| Method | P@10 | #pri | nDCG@R | #rel | R-prec | R@100 |
| *Optimized for Precision ($\chi^2$)* | | | | | | |
| (p1) Baseline (§9.5) | .2100 | 121 | .1198 | 54 | .2623 | .5423 |
| (p2) Improved typefiltering | .2350 | 157 | .1399 | 62 | .2959 | .6017 |
| (p3) Anchor based co-oc. | **.3000** | **174** | **.1562** | **76** | **.3473** | **.6667** |
| (p4) Adjusted judgements | .3900 | 186 | .1966 | 78 | .4759 | .6869 |
| *Optimized for Recall (MLE)* | | | | | | |
| (r1) Baseline (§9.5) | .0800 | 171 | .0880 | 105 | .0966 | .6982 |
| (r2) Improved typefiltering | .1000 | 177 | .1012 | 102 | .1408 | .7422 |
| (r3) Anchor based co-oc. | **.1950** | **187** | **.1444** | **143** | **.2730** | **.7496** |
| (r4) Adjusted judgements | .3450 | 214 | .2207 | 156 | .4134 | .8057 |
| *Best runs from the TREC 2009 Entity track* | | | | | | |
| KMR1PU [132] | **.4450** | 137 | **.2210** | 115 | **.5494** | **.5755** |
| ICTZHRun1 [376] | .3100 | 124 | .1525 | 69 | .3182 | .4638 |
| NiCTm2 [369] | .3050 | 124 | .1689 | 98 | .2721 | .3820 |
| tudpwkntop [302] | .2600 | **144** | .1506 | 128 | .2705 | .5721 |
| uogTrEpr [232] | .2350 | 135 | .1760 | **311** | .2945 | .4536 |

### Anchor-based Co-occurrence

Another approach employed at the Entity track [302, 369] is to only consider entities that link to, or are linked from, the Wikipedia page of the source entity. We view this as a special case of co-occurrence; its strength is proportional to the number of times source and target entities cross-link to each other on their corresponding Wikipedia pages. We estimate this anchor based co-occurrence as follows:

$$P_{anc}(e|E) = \lambda_a \cdot \frac{c(e, E_a)}{\sum_{e'} c(e', E_a)} + (1 - \lambda_a) \cdot \frac{c(E, e_a)}{\sum_{E'} c(E', e_a)},$$

where $c(e, E_a)$ is the number of times the candidate entity $e$ occurs in the anchor text on the Wikipedia page of the input entity $E$, and $c(E, e_a)$ is the other way around. We incorporate this into the pure co-occurrence component (§9.4.1) as a sum with equal weights.

With the addition of the anchor-based co-occurrence we further improve our precision and recall scores; see Table 9.12 (p3 and r3). Anchor-based co-occurrence works well in this setting as for most topics the relevant entities occur as anchor texts on the page of the source entity and vice versa (e.g., topics #9 and #20).

### Wikipedia-based Evaluation

Another way to compare our model and those of other TREC participants is to use the Wikipedia based evaluation employed throughout this chapter. From each participant's run we extract the Wikipedia fields and evaluate the number of primary Wikipedia pages for each topic in terms of R-precision and Recall@100. We observe that we outperform all but one of the other approaches in terms of R-precision (p3) and all approaches in terms of Recall@100 (r1, r2 and r3). The high precision achieved by the best performing team is due to their extensive use of heuristics, e.g., using a web search engine to collect relevant pages, crafting extraction patterns and exploiting lists and tables [132].

### Adjusted Judgements

Finally, the runs produced by our models are not official TREC runs and as such were not included in the assessment procedure; this might leave us with sparse judgments. Following standard TREC practice, non-judged documents are considered non-relevant—the resulting scores could therefore be an underestimation of our actual retrieval performance. To investigate how this affects our results we remove all entities for which there is no judgment available at all (neither primary, relevant or non-relevant, for neither the homepage or Wikipedia fields). We observe that only considering judged entities has a big affect on the precision and recall of our model (extended with anchor based co-occurrence and improved type filtering), see Table 9.12 (p4 and r4). In the precision oriented model 763 of the 6184 pages are judged (186 primary, 78 relevant). In the recall oriented model 1119 of the 6172 pages are judged (214 primary, 156 relevant). These numbers show that many of the returned entities have not been judged, impeding an assessment of the full potential of our models.

## 9.8 Conclusion

In this chapter we developed a system that supports identifying related concepts by specifying a particular relationship that should hold between them. We investigated the effectiveness of our system on the related entity finding (REF) task on a web corpus, where we focused on four core components: pure co-occurrence, type filtering, contextual information, and homepage finding. Initially we investigated the task on a smaller, less noisy corpus, using Wikipedia pages to uniquely identify entities. In this setting we addressed the first part of our sixth research question as formulated in Chapter 1:

**RQ 6A.** How can we find related entities for which the specified relation with a given source entity holds and that satisfy the target type constraint?

**a.** How do different measures for computing co-occurrence affect the recall of a pure co-occurrence based related entity finding (REF) model?

**b.** Can a type filtering approach based on Wikipedia categories successfully be applied to REF to improve precision without hurting recall?

**c.** Can recall and precision be enhanced by combining the co-occurrence model with a context model, so as to ensure that source and target entities engage in the right relation?

To address **a** we looked at four measures for computing co-occurrence and found that $\chi^2$ was most effective in identifying a potential set of related entities. An analysis showed that rankings of all methods were polluted by entities of the wrong type. In answer to **b** we found that even a basic category based type filtering approach is very effective and that the level of category expansion can be tuned towards precision or recall. Regarding **c** we found that adding context improves both recall and precision by ensuring that source and target entities engage in the right relation.

We then looked at the REF task in the setting of a web corpus which led us to the second part of our sixth research question:

**RQ 6B.** How can we find related entities and their homepages in a web corpus?

**a.** Does the use of a larger corpus improve estimations of co-occurrence and context models?

**b.** Is the initial focus on Wikipedia a sensible approach; can it achieve comparable performance to other approaches?

**c.** Can our basic framework effectively incorporate additional heuristics in order to be competitive with other state-of-the-art approaches?

In order to address **a** we re-estimated our co-occurrence and context models and found that using a larger corpus improves the estimations of both models. To conform to the official REF task we used a homepage finding component to map the Wikipedia entity representation to a homepage. In answer to **b** we found that our framework achieves decent precision and very high recall scores compared to other approaches on the official task. Finally, regarding **c** we found that our model can effectively incorporate additional heuristics that lead to state-of-the-art performance.

In this and the previous chapter we investigated methods that support identifying the concepts and relationships that exist within the domain of a topic. Applications incorporating these methods could support humanities researchers in discovering additional concepts related to their research topic by specifying concepts and relationships already identified. Our results show that both structured and unstructured data can be utilized to provide this type of functionality.

In the next chapter we review the answers we have provided in each of the chapters to our research questions from Chapter 1 and look at directions for future work.

# 10
# Conclusions

At the start of this thesis we observed how one by one research disciplines are becoming computationally more intensive as researchers seek to utilize the increasing amount of data available. In some disciplines, e.g., the natural sciences, switching to computationally intensive methods occurs more naturally than in others because of the focus on quantitative research methods and numerical data. In this thesis we focused on the development and evaluation of information retrieval tools and algorithms to support research in the humanities, where the research methodology is based on analytical, critical, and interpretative approaches applied to records of human culture. In particular, we aimed to address two challenges that humanities researchers face when dealing with large collections of digital records: *exploration* and *contextualization*. With respect to supporting exploration, we showed the utility of providing humanities researchers with information retrieval tools that enable richer means of interaction, i.e., that enable making comparisons between alternative search topics and that allow switching between alternative aggregated search displays. Regarding contextualization, we explored various algorithms that given a particular concept of interest, i.e., event or entity, provide suggestions for related concepts.

Below we first provide answers to the research questions posed in Chapter 1. Then in the final section, we look at efforts to incorporate the prototypes of information retrieval tools for exploration, developed in this thesis, in real world applications and future directions for the design and evaluation of tools that incorporate the algorithms developed for contextualization.

## 10.1 Answers to the Research Questions

In the first part of the thesis, where we focused on exploration, the central questions we addressed were: how do researchers in the humanities use existing information retrieval tools to search for research material in digital collections and can information retrieval tools with richer means of interaction be designed to provide better support for exploration? We started by reviewing theoretical models of information behavior, empirical models of the humanities research cycle, and information retrieval tools developed to support humanities researchers in Chapter 2. We have found that the success or failure of tools to support humanities researchers depends on how well they support existing research practices as well as that research practices are changing due to technological developments and an increasing amount of available digital material. In order to gain further insight in research practices in the humanities and how these practices may be supported we have investigated a particular discipline within the humanities, i.e., media studies, and asked:

**RQ 1.** What does the research process of media studies researchers look like?

We have developed a model of the media studies research cycle that captures how research activities relate to changes in the research questions of media studies researchers. A model with three phases emerged, i.e., the exploration phase, the contextualization phase, and the presentation phase, as during a research project media studies researchers transition from one set of activities to the next. The model shares similarities with models of the research cycle of other humanities researchers, which suggests that findings for this group are potentially relevant for other humanities disciplines. Although these models share similar stages and activities with our model of the media studies research cycle, an important difference is that they do not describe how these stages influence the research questions during the research cycle. The value of our model is that it makes the sequences of activities and the gradual refinement of the research questions in the media studies research cycle explicit. We have found that information gathering and analysis activities are especially influential on research outcomes and result in additional questions or a changed perspective in the research questions of media studies researchers. Reasons we have identified are that media studies researchers learn about the availability of material, discover trends in the material or gain alternative views on a topic.

The observation that media studies researchers make these discoveries, i.e., that certain material is unavailable or that there are particular trends in the data, during later stages of the research cycle gave rise to the two additional research questions that were addressed in the first part of the thesis. One challenge that media studies researchers face is to select suitable sources from which to gather material. We have investigated how media studies students are supported by two versions of aggregated search display style, i.e., tabbed and blended, during a research project. We asked:

**RQ 2.** How do master students conducting a media studies research project use alternative aggregated search result presentation methods?

We have followed a group of media studies master students during a four week research project and observed that the use of the tabbed display is predominantly motivated by a need to zoom in on specific sources. The majority of subjects, however, switched between the tabbed and blended displays. As motivation to use the blended display subjects noted a need to explore the content of the sources available to them. Subjects switched between displays at different times. Some subjects first preferred to zoom in on a specific source and later used the blended display to obtain an overview whenever a new specific information need arose. Other subjects initially desired an overview of the sort of information generally available in the sources, later the tabbed display was used to zoom in on a particular source.

To investigate the factors that underly these changes in display preference we have conducted a laboratory study. We recreated a multi-session search task composed of three sub-tasks. The first sub-task was completed with a tabbed display, the remaining sub-tasks with blended displays. One condition was manipulated by either providing three sub-tasks about the same topic or about three different topics. We have found that a certain combination of factors, i.e., subjects' search strategy and changes in formation need across sub-tasks, negatively influences perceived usability of the blended display. Two types of searchers emerged, i.e., one in which users explore a single source before consulting other sources and one in which users gain an overview of several sources before focusing on a particular source. The results from both studies have showed that subjects change display preference during a multi-session search task and suggest that among the possible factors influencing subjects' preferences are personal search strategies and changes in information need.

Another observation from the analysis of the media studies research cycle concerns the importance of discovering trends in the material and gaining alternative views on a topic. To support this type of research activities we extended a traditional exploratory search system design in two ways. First, we incorporated two side-by-side (subjunctive) versions of an exploratory search interface in a single display so multiple queries can be explored simultaneously. Second, we added visualizations that allow for contrasting and comparing of characteristics of the result sets. We then asked:

**RQ 3.** Does the ability to make comparisons support media studies researchers in exploring a collection of television broadcast metadata descriptions?

We have conducted an experiment in which a group of media studies researchers used either a standard version or a subjunctive version of an exploratory search interface to explore a collection of documents for the purpose of developing a research question. We have found that with the subjunctive interface media studies researchers are able to formulate more queries and bookmark more diverse documents compared to a standard exploratory search interface. With respect to the effect of the type of interface on the research questions we have found that with both interfaces quality research questions are formulated and we observed no differences in terms of originality, theoretical embedding, and formulation quality of the research questions. In a qualitative analysis of the research questions formulated by media studies researchers we have found some evidence to suggest that the influence of the subjunctive interface is predominantly on the scope of the research question. Specifically, users of the subjunctive interface incorporated more views on a topic in their research question than users of the standard exploratory search interface. In terms of usability, media studies researchers reported that the subjunctive interface is intuitive and not difficult to use, suggesting that the additional complexity in terms of features in the subjunctive interface does not reduce its usability.

Once a humanities researcher has established the focus of his/her research topic and gathered an initial sample of material through exploration, the next challenge is to gather the contextual information that is necessary for interpretation of the concepts and relationships within the domain of the research topic. In the second part of the thesis we focused on two approaches to support contextualization: (i) by generating links between material based on related concepts; and (ii) by identifying related concepts directly. We have reviewed existing methods to find related concepts in Chapter 6. To be able to compare and build upon these methods we moved away from the user centered approach of the first part of the thesis and adopted a system centered approach.

In this setting we have investigated a concept based contextualization method that automatically creates links between archival records from different institutions that cover related or similar events. Specifically, we have defined the task of linking archives, i.e., given a representation of a record and its associated metadata from one archive (the source archive), connect it to the representation of a record and its metadata in another archive (the target archive), where that record describes the same event or a related event. Given this task we asked:

**RQ 4.** How can we automatically generate links from a record in a newspaper archive with a rich textual representation to records in a television archive that tend to have sparse textual representations?

We have experimented with a retrieval approach to event linking and found that expanding target archive records with records from other archives improves performance for both *same event linking* and *related event linking*. Using records from the source archive for expansion, however, was most effective. Additionally, we found that reducing the number of terms in the source archive record representation is most effective for *same event linking*. The reduced record representations were more robust to topic drift and formed a better match for the short event descriptions in the target archive. *Related event linking* also improved but not as much as with target archive record expansion. The benefit for *related event linkin*, as obtained through expansion, stemmed from the rich descriptions that covered more aspects of an event.

We then shifted our attention from finding related material based on events to identifying related concepts directly based on relationships with concepts already identified as relevant to a particular research topic. To this end, we have defined the example-based entity finding task, i.e., find a group of target entities, that all have the same relationship with a particular concept in common, given a number of examples. We have looked into the benefits of using structured data in this entity-oriented search task and asked:

**RQ 5.** How can we exploit the structural information available in the Web of Data to find a set of entities, that all have the same relationship with a particular concept in common, based on a number of example entities?

We have found that depending on the number and quality of the examples, a structure-based approach achieves comparable performance to a competitive text-based approach. Through a per topic analysis, however, we found that each method returns different sets of entities, motivating the use of a hybrid approach. We have performed an analysis of the performance of two hybrid methods on repeated samples of example entities and relevance judgements. Results showed that a standard linear combination approach is sub-optimal when the set of examples and entities considered relevant changes. In contrast, a hybrid method that uses example entities to determine whether to use a text-based, structure-based, or linear combination approach, outperformed the linear combination method.

Although structured data is effective in improving performance on an entity-oriented search task, a limited number of relations are available in structured form. Therefore, we have investigated the potential of unstructured data to find contextual information and defined the related entity finding task, i.e., given a (source) entity, a free text description of a relation, and the type of the (target) entities, finds related entities or which the specified relationship with the source entity holds. We first turned to Wikipedia as a corpus which provides a middle ground between structured data and the noisy content on the web and asked:

**RQ 6A.** How can we find related entities for which the specified relation with a given source entity holds and that satisfy the target type constraint?

We have incrementally developed a pipeline architecture for a related entity finding system and first investigated a relation modeling component based on co-occurrence. We compared four measures for computing co-occurrence to identify a potential set of related entities and found that $\chi^2$ performed best. An analysis showed that rankings of all methods were polluted by entities of the wrong type. We added a filtering component and found that even a basic category-based type filtering approach is very effective and that the level of category expansion can be tuned towards precision or recall. Furthermore, we have found that adding a relation modeling component based on context improves both recall and precision by ensuring that source and target entities engage in the right relation.

We then went beyond the relatively clean setting provided by Wikipedia and have investigated the behavior of the pipeline developed above on a large web corpus. This setting added the challenge of finding homepages for the surface forms of related entities and so we asked:

**RQ 6B.** How can we find related entities and their homepages in a web corpus?

We have found that our related entity finding system is effective in addressing the related entity finding task on a web corpus and that using a larger corpus improves the estimations of both co-occurrence and context based relation modeling components. We added a homepage finding component to our related entity finding system and found that our pipeline achieves decent precision and very high recall scores compared to other approaches on a community-based benchmark task. Finally, we have found that our models can effectively incorporate additional heuristics that lead to state-of-the-art performance.

## 10.2 Future Work

The two main challenges that humanities researchers face when using large collections of digital records in their research are *exploration* and *contextualization*. In the first part of this thesis we have provided insights in how information retrieval tools can be improved to better support humanitites researcher in exploring collections through richer interactions. In the second part we have developed algorithms that support identifying contextual information through relations between concepts. Here, we discuss these findings in a broader perspective and highlight possible areas of application for the information retrieval tools and algorithms developed in this thesis within the humanities.

**Beyond media studies.** In Chapter 3 we saw that the research cycle of media studies researchers shares similarities with those of researchers in other humanities disciplines. Historians, for example, share the need for exploration of collections of primary and secondary source materials across various institutions [321]. This suggests that the findings in Chapter 4 and Chapter 5 may be applicable to a wider range of humanities researchers.

The switching behavior observed in Chapter 4 suggests that alternative display options are necessary as researchers require a different presentation of results depending on their personal search strategy, task, and the stage of the research project in which they are engaged. These results may be used to inform the design of new archival information retrieval tools. Another future direction is to use this type of information to predict whether a particular type of display method should be used for result presentation in response to a query.

The ability to explore alternative topics and compare queries provided by the subjunctive interface introduced in Chapter 5 may be adjusted to support researchers in other humanities disciplines as well. The specific type of comparison or visualization that is potentially helpful to humanities researchers depends on the specific activities and data studied within each discipline. For example, maps are often used in historical research as they provide a view of the world at a certain time. One extension of the subjunctive interface would be to enable comparison of the regions associated with certain search results. Other disciplines require more in depth comparisons of a small set of records. For example, researchers in comparative literature study word usage and patterns across time periods, borders, and languages [206]. Another direction is to detect the sentiment associated with search results in order to compare opinions expressed in records about a topic. Such technology could support researchers in memory studies to detect changes in cultural memory [373].

**Uptake of tools.** Several projects and follow up work is underway based on the work in this thesis. We discuss three: AVResearcher, Quamerdes, and TROVe.

AVResearcher [174] is an implementation of the prototype developed in Chapter 5 by the Dutch National Institute of Sound and Vision. It covers the same data, i.e., television program metadata descriptions, expanded with subtitles and tweets. It is planned to be released as an advanced search option next to the existing search interface of the archive. The release of this tool in a real world setting provides a unique opportunity to log the behavior of users with a subjunctive exploratory search interface in a naturalistic setting. The analysis of these logs could yield new insights in how such an interface is used for exploration and how its usability could be improved.

QuaMeRDES (Quantitative Content Analysis of Media Researchers' Data[1]) is a project that aims to develop a tool that enables quantitative content analysis of television and newspaper records across various sources. It extends the work from Chapter 5 by incorporating additional sources and shifting the emphasis from exploration to analysis. Users are able to define *codes* by specifying a set of keywords and several types of visualizations are available to compare occurrence counts of these codes, e.g., over time or per genre. In this thesis we did not investigate tools to support analysis as several tools exist that support this practice (cf. §2.3.3). The tight coupling between exploration, information gathering and analysis in the research cycle of media studies researchers we observed in Chapter 3, however, suggests the need for a tool that integrates support for exploration, developing a sampling strategy, and analysis.

The goal of TROVe (Transmedia Observatory[2]) is to develop a tool that enables analysis of how news is spread through various types of media and entities. Instead of making

[1] `http://www.clarin.nl/node/1404`

[2] `http://www.clariah.nl/trove/samenvatting`

comparisons within a single archive as we investigated in Chapter 5, it compares coverage of events across archives. TROVe allows media researchers to answer questions such as who are the key players in a debate, how an event is covered by various media, or the extent to which a media channel reaches its audience for a certain event.

The adoption and continued development of these tools is a sign of the need that exists among humanities researchers to gain more insight in the digital collections that are now available. Tools such as the ones described here are generally developed to operate on a single (or a fixed set of) collection(s) and to provide a particular type of functionality, e.g., visualization. One of the reasons is that in developing such tools for each collection considerable preprocessing efforts are required to allow a tool to operate on the content, as well as decisions about which fields to show, which to aggregate, and how to visualize them. At the moment these decisions are made by the tool builders. A direction for future work is to develop flexible and intuitive interfaces that enable adaptation of a tool to requirements of a specific collection by humanities researchers.

**Applications for algorithms.** Examples of applications that show contextual information are the *more on this story* feature on the BBC website that provides links to related news events with each article[3] or the biography information provided by Google's *knowledge graph* next to search results (cf. §6.3.3). These applications, however, depend on manual editing, are limited to popular concepts, and restricted to a particular set of relations.

We have investigated algorithms that return related concepts for any type of relation and regardless of popularity. We have found that methods are able to find relations between popular concepts and relations captured in structured data with high confidence. Accurately discovering relations between infrequent concepts, however, remains challenging. The next step is to extend these methods to enable concept based contextualization of the relative sparse occurring concepts in archival and library collections. One extension to the event-based linking method developed in Chapter 7 is to investigate the performance of linking to additional (news) archives. A natural extension of the work in Chapter 8 and 9 is to combine the methods developed to operate on structured data with the methods that use a web corpus to find related concepts. This could improve performance on both the example based entity finding and related entity finding tasks by combining evidence from multiple sources as well as support cases where relations are not available in structured form.

One step further, when performance reaches a certain level of quality, is to incorporate these methods in tools and evaluate their usability and effectiveness in supporting humanities researchers. Examples of applications are browsing concepts in archival collections based on a particular relation, e.g., correspondence, or visualizing a timeline of related and similar events covered in records from various archives given a record describing a particular event.

Finally, although some parts have focused on media studies, the tools and methods described in this thesis aim to support research practices of humanities researchers in general, i.e., to contribute to the toolkit to support scholarly primitives (cf. §2.3.3) as envisaged by Unsworth [336]. For example, the work in Chapters 4 and 5 supports

[3] `http://www.bbc.co.uk/news/world-middle-east-12313405`

primitives such as comparing, browsing, and probing, while the work in Chapters 7, 8, 9 has the potential to support primitives such as assembling and organizing. However, the introduction of these new tools as well as new types of data that will become available in the future are likely to again change the way humanities researchers work and the questions they seek to answer. Insights from the work in this thesis may be used to inform the design and evaluation of future tools to continually support new needs and developments in the humanities.

# A
# Interfaces and Resources

In order to carry out our user experiments in Chapter 4 and 5 we have developed prototypes of two interfaces. We release the code for these interfaces as open source packages and provide a short description of the content of each package in Section A.1 and A.2 for the interfaces introduced in Chapter 4 and 5 respectively. In our investigation of methods to support contextualization in Chapters 8 and 9 we have created several sets of relevance judgements. We make these relevance judgements available and provide a description of these resources in Section A.3

## A.1 Aggregated Search Interfaces

The source code is available on `http://ilps.science.uva.nl/resource/comerda`. It provides the three display types introduced in Chapter 4. A number of configuration options are available, e.g., which display type to show and whether to provide faceted search. The interface expects a multi-core SOLR index as back end, where each core represents a collection. Further details are available in the documentation.

## A.2 A Subjunctive Exploratory Search Interface

The source code is available on `http://ilps.science.uva.nl/resource/merdes`. This is a reïmplementation of the original subjunctive interface used in our experiments in Chapter 5 as the original system required a relatively high bandwidth to be responsive. It requires an ElasticSearch index as back end and only search within a single collection is supported.

## A.3 Relevance Judgements

Below we provide a brief overview of the various relevance judgements we created for the data sets used in our experiments in Chapter 8 (§A.3.1) and 9 (§A.3.2).

## A.3.1 Example Based Entity Finding Resources

The topics and judgements are made available on http://ilps.science.uva.nl/ecir2013elc. It includes judgements for three sets of topics: INEX 2007, INEX 2008, and SemSearch 2011. This package contains the following resources:

### Topics

The folder topics contains the following files with the topics we derived from the INEX 2007, INEX 2008, and SemSearch 2011 campaigns:

- inex2007.topics
- inex2008.topics
- semsearch2011.topics

Topics follow the following format:

```
topicID relation,
```

for example:

```
72 venice movies fully partly shot venice
```

Both punctuation and stopwords have been removed. The narrative of the INEX 2007 and 2008 topics is used as the relation, the SemSearch 2011 topics already have relations.

### Judgements

For each topic results were obtained from the BTC2009 corpus and judged for relevance. The folder qrels contains the files with judgements for the corresponding topics:

- inex2007.qrels
- inex2008.qrels
- semsearch2011.qrels

The original 2007 and inex 2008 qrels refer to Wikipedia URLs, these have been mapped to the DBpedia URIs in the BTC2009 corpus. Matching was done via several methods. We started by adding the http://dbpedia.org/resource/ prefix to the entity string after the last "/" in the Wikipedia URL and checking for the existence of the URI in the BTC2009 corpus. As these datasets are several years old, some Wikipedia URLs no longer exist. By using the entity string as query and retrieving Wikipedia URLs from a newer version of Wikipedia, we found additional DBpedia URIs. Finally, using a lucene index of the BTC2009 corpus, we used the entity string as query and manually inspect the top 20 results. The SemSearch 2011 qrels are obtained from the BTC2009 corpus and required no further preprocessing.

## A.3.2 Related Entity Finding Resources

The runs, judgements, and evaluation scripts are made available on http://ilps.science.uva.nl/resources/cikm2010-entity. This package contains the following resources:

### Judgements

The folder `qrels` containing judged documents (qrels) used by the various evaluation scripts to evaluate the runs produced by our methods and other TREC 2009 Entity Track participants.

**`09.entity.qrels.txt`** Used by `eval-entity.pl` for official Entity 2009 TREC related entity finding evaluation

**`09.wpvariants.qrels`** Used by `eval-cikm2010-entity.py` for Wikipedia-based related entity finding evaluation. The original Entity Track qrels did not contain all variants of the relevant Wikipedia pages (pages with a different cluewebID but that redirect to the relevant Wikipedia page and thus have the same content). These pages "variants" have been added to the `09.wpvariants.qrels` file.

**`2009.hpfinding.qrels`** Contains the ClueWeb09 ID (cwID) of homepages for each topic in TREC qrel format. Extracted from the 2009 Entity Track qrels. Only primary homepages were extracted. Used by `trec-eval` for the evaluation of homepage finding.

### Runs

The folder `runs/15top-trecstyle-runs` contains runs of top performing participants at the TREC 2009 Entity Track in the standard TREC format, that serves as input to the official TREC evaluation script `entity-eval.pl`.

- `ICTZHRun1`
- `KMR1PU`
- `NiCTm2`
- `tudpwkntop`
- `uogTrEpr`

The folder `runs/15top-wikipediastyle-runs` contains the runs of other top performing participants at TREC 2009 in Wikipedia evaluation format, taken by Wikipedia evaluation script `eval-cikm2010-entity.py`.

- `ICTZHRun1.wp`
- `KMR1PU.wp`
- `NiCTm2.wp`
- `tudpwkntop.wp`
- `uogTrEpr.wp`

The folder `runs/cikm-trecstyle-runs` contains our runs as described in Chapter 9 in TREC 2009 entity format taken by the official evaluation script `entity-eval.pl`.

- `res_adjusted_judgements_mle` (see Table 9.12: r4)
- `res_adjusted_judgements_x2` (see Table 9.12: p4)
- `res_anchorbased_cooccurrence_mle` (see Table 9.12: r3)
- `res_anchorbased_cooccurrence_x2` (see Table 9.12: p3)
- `res_improved_typefiltering_mle` (see Table 9.12: r2)
- `res_improved_typefiltering_x2` (see Table 9.12: p2)

Finally, the folder `runs/cikm-wikipediastyle-runs` contains our runs as described in Chapter 10 in format taken by the Wikipedia evaluation script `eval-cikm-2010-entity.py`.

- `WP-res_adjusted_judgements_mle` (see Table 9.12: r4)
- `WP-res_adjusted_judgements_x2` (see Table 9.12: p4)
- `WP-res_anchorbased_coocurrenc_mle` (see Table 9.12: r3)
- `WP-res_anchorbased_coocurrence_x2` (see Table 9.12: p3)
- `WP-res_improved_typefiltering_mle` (see Table 9.12: r2)
- `WP-res_improved_typefiltering_x2` (see Table 9.12: p2)

### Scripts

The folder `scripts` contains the scripts used to generate the evaluation scores as well as some convenience scripts.

**`eval-entity.pl`** The official 2009 TREC entity evaluation script.

**`eval-cikm2010-entity.py`** Evaluation script for Wikipedia-based evaluation.

**`test.run`** Used to test the Wikipedia-based evaluation script. contains the correct documents for each topic so should return prefect recall and R-precision. Note, this is not the case as topic 6 contains the same entity twice, while considering it as a different answer (it has a different cluster ID).

**`run-entity-trecstyle-eval.sh`** Do the official TREC Entity 2009 evaluation on all supplied runs.

**`run-entity-wikipediastyle-eval.sh`** Do the Wikipedia-based evaluation on all supplied runs.

### Topics

The folder `topics` contains the homepage finding topics and qrels used for the homepage finding experiments in Section 9.6.

**`2009.hpfinding.topics`** Contains the names of the entities for which to find homepages. Names are extracted from the 2009 Entity track qrels. If a Wikipedia page occurred in the same cluster as a primary homepage, we used the Wikipedia title as query. Otherwise we used the `NAME` field. If neither was available a name was found manually by inspecting the homepage.

**`2009.hpfinding.topics.trectext`** For convenience also the topics in trectext. All punctuation is replaced by space, trailing spaces have been removed.

### Data

The folder `data` contains the file `source_ent_wikipedia_mapping.txt`, which is a manual mapping of the 2009 Entity Track source entities to their respective Wikipedia pages.

# Bibliography

[1] J. Allan. Relevance Feedback with Too Much Data. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 337–343. ACM, 1995. (Cited on page 101.)

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop*, pages 194–218, 1998. (Cited on pages 105 and 107.)

[3] T. J. Allen. *Information needs and uses*. Information Today, 1969. (Cited on page 18.)

[4] J. Alonso-Jimene, J. Borrego-Diaz, A. M. Chavez-Gonzalez, and F. J. Martin-Mateos. Foundational Challenges in Automated Semantic Web Data and Ontology Cleaning. *Intelligent Systems, IEEE*, 21(1): 42–52, 2006. (Cited on page 113.)

[5] D. L. Altheide. The Elusive Mass Media. *International Journal of Politics, Culture, and Society*, 2(3): 414–419, 1989. (Cited on page 38.)

[6] D. L. Altheide and C. J. Schneider. *Qualitative Media Analysis*, volume 38. SAGE Publications, Incorporated, 2012. (Cited on pages 2 and 38.)

[7] G. Amati and C. J. Van Rijsbergen. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems*, 20(4):357–389, 2002. (Cited on page 102.)

[8] A. Amin, J. van Ossenbruggen, L. Hardman, and A. van Nispen. Understanding Cultural Heritage Experts' Information Seeking Needs. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL'08)*, pages 39–47. ACM, 2008. (Cited on page 28.)

[9] A. Amin, M. Hildebrand, J. Van Ossenbruggen, V. Evers, and L. Hardman. Organizing suggestions in autocompletion interfaces. In *Advances in Information Retrieval*, pages 521–529. Springer, 2009. (Cited on page 28.)

[10] A. Amin, J. Zhang, H. Cramer, L. Hardman, and V. Evers. The effects of source credibility ratings in a cultural heritage information aggregator. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 35–42. ACM, 2009. (Cited on page 28.)

[11] A. Amin, M. Hildebrand, J. van Ossenbruggen, and L. Hardman. Designing a thesaurus-based comparison search interface for linked cultural heritage sources. In *Proceedings of the 15th international conference on Intelligent user interfaces (IUI'10)*, pages 249–258. ACM, 2010. (Cited on page 28.)

[12] A. Andrenucci and E. Sneiders. Automated Question Answering: Review of the Main Approaches. In *Proceedings of the Third International Conference on Information Technology and Applications*, pages 514–519. ACM, 2005. (Cited on page 148.)

[13] G. Antoniou. *A semantic web primer*. the MIT Press, 2004. (Cited on page 112.)

[14] J. Arguello, F. Diaz, and J. Callan. Learning to Aggregate Vertical Results into Web Search Results. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*, pages 201–210. ACM, 2011. (Cited on page 33.)

[15] J. Arguello, F. Diaz, J. Callan, and B. Carterette. A Methodology for Evaluating Aggregated Search Results. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval*, pages 141–152. Springer, 2011. (Cited on page 33.)

[16] J. Arguello, W. Wu, D. Kelly, and A. Edwards. Task Complexity, Vertical Display and User Interaction in Aggregated Search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*, pages 435–444. ACM, 2012. (Cited on pages 33 and 77.)

[17] J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In *Proceedings of the Multilingual and Multimodal Information Access Evaluation, International Conference of the Cross-Language Evaluation Forum (CLEF'10)*. Springer, 2010. (Cited on pages 104 and 105.)

[18] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007. (Cited on page 113.)

[19] G. Auffret and B. Bachimont. Audiovisual cultural heritage: from TV and radio archiving to hypermedia publishing. In *Research and Advanced Technology for Digital Libraries*, pages 58–75. Springer, 1999. (Cited on page 55.)

[20] A. Aula, P. Majaranta, and K.-J. Räihä. Eye-tracking reveals the personal styles for search result evaluation. In *Human-Computer Interaction-INTERACT 2005*, pages 1058–1061. Springer, 2005. (Cited on page 31.)

[21] A. Aula, R. Khan, and Z. Guan. How Does Search Behavior Change as Search Becomes More Difficult.

In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*, pages 35–44. ACM, 2010. (Cited on pages 30 and 91.)

[22] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*, volume 463. ACM press New York, 1999. (Cited on pages 99, 100, and 102.)

[23] K. Balog. *People Search in the Enterprise*. PhD thesis, University of Amsterdam, June 2008. (Cited on page 108.)

[24] K. Balog and M. de Rijke. Associating People and Documents. In *Proceedings of the 30th European Conference on Information Retrieval (ECIR'08)*, pages 296–308. Springer, 2008. (Cited on page 108.)

[25] K. Balog, L. Azzopardi, and M. de Rijke. Formal Models for Expert Finding in Enterprise Corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'06)*, pages 43–50. ACM, 2006. (Cited on pages 108 and 110.)

[26] K. Balog, L. Azzopardi, and M. de Rijke. A Language Modeling Framework for Expert Finding. *Information Processing & Management*, 45(1):1–19, 2009. (Cited on page 108.)

[27] K. Balog, M. Bron, J. He, K. Hofmann, E. Meij, M. de Rijke, E. Tsagkias, and W. Weerkamp. The University of Amsterdam at TREC 2009: Blog, Web, Entity, and Relevance Feedback. In *Proceedings of the Eightteenth Text REtrieval Conferenc (TREC'09)*. NIST, 2009. (Cited on page 13.)

[28] K. Balog, M. de Rijke, and M. Bron. Related Entity Finding Based on Co-Occurrence. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC'09)*. NIST, 2009. (Cited on page 13.)

[29] K. Balog, A. P. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld. Overview of the TREC 2009 Entity Track. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC'09)*. NIST, 2009. (Cited on pages 106, 110, 111, 150, and 153.)

[30] K. Balog, J. He, M. Bron, M. de Rijke, and W. Weerkamp. *The University of Amsterdam (ISLA) at INEX 2009*. Springer Berlin Heidelberg, 2009. (Cited on page 13.)

[31] K. Balog, I. Soboroff, P. Thomas, N. Craswell, A. P. de Vries, and P. Bailey. Overview of the TREC 2008 Enterprise Track. In *Proceedings of the Seventeenth Text Retrieval Conference Proceedings (TREC'08)*. NIST, 2009. (Cited on page 108.)

[32] K. Balog, M. Bron, and M. de Rijke. Category-based Query Modeling for Entity Search. In *Advances in Information Retrieval*, pages 319–331. Springer Berlin Heidelberg, 2010. (Cited on pages 13 and 110.)

[33] K. Balog, M. Bron, M. de Rijke, and W. Weerkamp. Combining Term-Based and Category-Based Representations for Entity Search. In *Proceedings of the 8th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'09)*, pages 265–272. Springer, 2010. (Cited on pages 13 and 141.)

[34] K. Balog, E. Meij, and M. de Rijke. Entity search: building bridges between two worlds. In *Semantic Search Workshop'10*, pages 1–5, 2010. (Cited on page 114.)

[35] K. Balog, M. Bron, and M. de Rijke. Query Modeling for Entity Search Based on Terms, Categories, and Examples. *ACM Transactions on Information Systems*, 29(4):22, 2011. (Cited on pages 13, 110, and 139.)

[36] K. Balog, M. Ciglan, R. Neumayer, W. Wei, and K. Nörväag. NTNU at SemSearch 2011. In *Proceedings of the 4rd International Semantic Search Workshop (SemSearch'11)*. ACM, 2011. (Cited on pages 115 and 135.)

[37] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise Retrieval. *Foundations and Trends in Information Retrieval*, 6(2-3):127–256, 2012. (Cited on page 108.)

[38] A. Barrett. The Information-Seeking Habits of Graduate Student Researchers in the Humanities. *The Journal of Academic Librarianship*, 31(4):324–331, 2005. (Cited on page 24.)

[39] S. Baruchson-Arbib and J. Bronstein. Humanists as Information Users in the Digital Age The Case of Jewish Studies Scholars in Israel. *Journal of the American Society for Information Science and Technology*, 58(14):2269–2279, 2007. (Cited on page 24.)

[40] H. Bast, F. Bäurle, B. Buchhold, and E. Haussmann. A Case for Semantic Full-Text Search. In *Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search (JIWES'12)*, pages 4:1–4:3. ACM, 2012. (Cited on page 101.)

[41] M. Bates. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, 13(5):407–424, 1989. (Cited on pages 20, 27, and 54.)

[42] M. J. Bates. Information Search Tactics. *Journal of the American Society for Information Science*, 30 (4):205–214, 1979. (Cited on page 20.)

[43] M. J. Bates. Where Should the Person Stop and the Information Search Interface Start? *Information Processing & Management*, 26(5):575–591, 1990. (Cited on page 21.)

[44] M. J. Bates. The Getty End-User Online Searching Project in the Humanities: Report No. 6: Overview and Conclusions. *College & Research Libraries*, 57(6):514–23, 1996. (Cited on pages 2 and 24.)

[45] N. J. Belkin. Helping People Find what They Don't Know. *Communications of the ACM*, 43(8):58–61, 2000. (Cited on page 61.)

[46] N. J. Belkin, R. N. Oddy, and H. M. Brooks. ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation*, 38(2):61–71, 1982. (Cited on page 21.)

[47] N. J. Belkin, R. N. Oddy, and H. M. Brooks. ASK for Information Retrieval: Part II. Results of a Design Study. *Journal of Documentation*, 38(3):145–164, 1982. (Cited on page 21.)

[48] N. J. Belkin, C. Cool, A. Stein, and U. Thiel. Cases, Scripts, and Information-Seeking Strategies: On the Design of Interactive Information Retrieval Systems. *Expert Systems with applications*, 9(3):379–395, 1995. (Cited on page 21.)

[49] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom. Swoosh: A Generic Approach to Entity Resolution. Technical Report 2005-5, Stanford InfoLab, 2005. (Cited on page 112.)

[50] J. Bennett and N. Strange. *Television as digital media*. Duke University Press Books, 2011. (Cited on page 38.)

[51] L. Bentivogli, I. Dagan, H. T. Dang, D. Giampiccolo, and B. Magnini. The Fifth Pascal Recognizing Textual Entailment Challenge. In *Proceedings of the Second Text Analysis Conference (TAC'09)*, pages 14–24. NIST, 2009. (Cited on pages 104 and 105.)

[52] M. K. Bergman. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, 7(1):07–01, 2001. (Cited on page 32.)

[53] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284(5):28–37, 2001. (Cited on pages 4, 6, and 133.)

[54] C. Bizer. The Emerging Web of Linked Data. *Intelligent Systems, IEEE*, 24(5):87–92, 2009. (Cited on page 113.)

[55] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009. (Cited on pages 4, 6, and 114.)

[56] R. Blanco, H. Halpin, D. Herzig, P. Mika, J. Pound, and H. Thompson. Entity Search Evaluation over Structured Web Data. In *Proceedings of the First International Workshop on Entity-Oriented Search (EOS'11)*. ACM, 2011. (Cited on pages 114 and 135.)

[57] R. Blanco, P. Mika, and S. Vigna. Effective and Efficient Entity Search in RDF Data. In *Proceedings of the 11th International Semantic Web Conference (ISWC'11)*, pages 83–97. Springer, 2011. (Cited on page 114.)

[58] T. Blanke, L. Candela, M. Hedges, M. Priddy, and F. Simeoni. Deploying General-Purpose Virtual Research Environments for Humanities Research. *Philosophical Transactions of the Royal Society A*, 368(1925):3813–3828, 2010. (Cited on page 28.)

[59] T. Blanke, M. Bryant, M. Hedges, A. Aschenbrenner, and M. Priddy. Preparing DARIAH. In *Proceedings of the 7th IEEE International Conference on eScience (eScience'11)*, pages 158–165. IEEE, 2011. (Cited on page 28.)

[60] C. Borgman. The Digital Future is Now: A Call to Action for the Humanities. *Digital Humanities Quarterly*, 3(4), 2009. (Cited on pages 1, 2, 22, 23, and 28.)

[61] H. Borko. Information Science: what is It? *American Documentation*, 19(1):3–5, 1968. (Cited on page 18.)

[62] J. Bradley. Thinking About Interpretation: Pliny and Scholarship in the Humanities. *Literary and Linguistic Computing*, 23(3):263–279, 2008. (Cited on page 27.)

[63] T. Brinck, D. Gergle, and S. Wood. *Usability for the Web: designing Web sites that work*. Morgan Kaufmann, 2002. (Cited on page 80.)

[64] A. Broder. A Taxonomy of Web Search. *ACM SIGIR Forum*, 36(2):3–10, 2002. (Cited on page 29.)

[65] M. Bron, K. Balog, and M. de Rijke. Ranking Related Entities: Components and Analyses. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*, pages 1079–1088. ACM, 2010. (Cited on page 12.)

[66] M. Bron, J. He, K. Hofmann, E. Meij, M. de Rijke, M. Tsagkias, and W. Weerkamp. The University of Amsterdam at TREC 2010: Session, Entity and Relevance Feedback. In *Proceedings of the Nineteenth Text REtrieval Conference (TREC'10)*. NIST, 2011. (Cited on pages 13 and 115.)

[67] M. Bron, B. Huurnink, and M. de Rijke. Linking archives using document enrichment and term selection. In *Research and Advanced Technology for Digital Libraries*, pages 360–371. Springer Berlin Heidelberg, 2011. (Cited on pages 12 and 61.)

[68] M. Bron, J. van Gorp, F. Nack, and M. de Rijke. Exploratory Search in an Audio-Visual Archive: Evaluating a Professional Search Tool for Non-Professional Users. In *Proceedings of the 1st European Workshop on Human-Computer Interaction and Information Retrieval (EuroHCIR'11)*, pages 3–6.

CEUR, 2011. (Cited on page 12.)

[69] M. Bron, F. Nack, M. de Rijke, and J. van Gorp. Ingredients for a User Interface to Support Media Studies Researchers in Data Collection. In *Proceedings of the 2nd European Workshop on Human-Computer Interaction and Information Retrieval (EuroHCIR'12)*, pages 33–36. CEUR, 2012. (Cited on page 12.)

[70] M. Bron, J. van Gorp, F. Nack, M. de Rijke, A. Vishneuski, and S. de Leeuw. A Subjunctive Exploratory Search Interface to Support Media Studies Researchers. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*, pages 425–434. ACM, 2012. (Cited on page 12.)

[71] M. Bron, K. Balog, and M. de Rijke. Example based entity search in the web of data. In *Advances in Information Retrieval*, pages 392–403. Springer Berlin Heidelberg, 2013. (Cited on page 12.)

[72] M. Bron, J. van Gorp, F. Nack, L. B. Baltussen, and M. de Rijke. Aggregated Search Interface Preferences in Multi-Session Search Tasks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*, pages 123–132. ACM, 2013. (Cited on page 12.)

[73] C. D. Brown. Straddling the Humanities and Social Sciences: The Research Process of Music Scholars. *Library & Information Science Research*, 24(1):73–94, 2002. (Cited on pages 25, 26, 39, 46, 53, and 55.)

[74] J. Brown and J. Sime. A methodology for accounts. In M. Brenner, editor, *Social method and social life*, pages 159–188. Academic Press, London, 1981. (Cited on page 39.)

[75] M. Buckland and C. Plaunt. On the Construction of Selection Systems. *Library Hi Tech*, 12(4):15–28, 1994. (Cited on page 99.)

[76] M. K. Buckland. Information as Thing. *Journal of the American Society for Information Science*, 42 (5):351–360, 1991. (Cited on page 19.)

[77] C. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical report, Microsoft Research, 2010. (Cited on page 138.)

[78] G. Buscher, R. W. White, S. Dumais, and J. Huang. Large-Scale Analysis of Individual and Task Differences in Search Result Page Examination Strategies. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining (WSDM'12)*, pages 373–382. ACM, 2012. (Cited on page 64.)

[79] V. Bush. As We May Think. *Atlantic Monthly*, 176(1):101–108, 1945. (Cited on page 18.)

[80] D. Byrd. A Scrollbar-Based Visualization for Document Navigation. In *Proceedings of the fourth ACM conference on Digital libraries (DL'99)*, pages 122–129. ACM, 1999. (Cited on page 34.)

[81] K. Byström and P. Hansen. Conceptual Framework for Tasks in Information Studies. *Journal of the American Society for Information Science and Technology*, 56(10):1050–1061, 2005. (Cited on pages 22, 30, and 54.)

[82] K. Byström and K. Järvelin. Task Complexity Affects Information Seeking and Use. *Information Processing & Management*, 31(2):191–213, 1995. (Cited on page 30.)

[83] M. J. Cafarella, C. Re, D. Suciu, O. Etzioni, and M. Banko. Structured Querying of Web Text. In *Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR)*. CIDR, 2007. (Cited on page 101.)

[84] M. L. Calkins. Free Text or Controlled Vocabulary? A Case History Step-By-Step Analysis Plus Other Aspects of Search Strategy. *Database: The Magazine of Database Reference and Review*, 3(2):53–67, 1980. (Cited on page 100.)

[85] Y. Cao, J. Liu, S. Bao, and H. Li. Research on Expert Search at Enterprise Track of TREC 2005. In *Proceedings of the The Fourteenth Text REtrieval Conference (TREC'05)*. NIST, 2005. (Cited on page 108.)

[86] R. Capra and G. Marchionini. The Relation Browser Tool for Faceted Exploratory Search. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'08)*, page 420. ACM, 2008. (Cited on pages 34 and 80.)

[87] C. Carrick and C. Watters. Automatic Association of News Items. *Information Processing & Management*, 33(5):615–632, 1997. (Cited on page 107.)

[88] D. Case. *Looking for information*. Emerald Group Publishing, 2012. (Cited on page 18.)

[89] D. O. Case. The Collection and Use of Information by Some American Historians: a Study of Motives and Methods. *Library Quarterly*, 61(1):61–82, 1991. (Cited on pages 1, 2, and 25.)

[90] C. W. Choo, B. Detlor, and D. Turnbull. Information Seeking on the Web: An Integrated Model of Browsing and Searching. *First Monday*, 5(2), 2000. (Cited on page 29.)

[91] C. Chu. Literary Critics at Work and their Information Needs: A Research-Phases Model. *Library & Information Science Research*, 21(2):247–273, 1999. (Cited on pages 25, 26, 39, 46, 53, and 55.)

[92] H. Chu. *Information representation and retrieval in the digital age*. Information Today, Inc., 2003. (Cited on page 100.)

[93] J. Chu-Carroll, K. Czuba, J. Prager, A. Ittycheriah, and S. Blair-Goldensohn. IBM's PIQUANT II in TREC 2004. In *Proceedings of the Thirteenth Text Retrieval Conference (TREC'04)*. NIST, 2004. (Cited on page 108.)

[94] M. Ciglan, K. Nörväag, and L. Hluchý. The SemSets Model for Ad-Hoc Semantic List Search. In *Proceedings of the 2012 International Conference on the World Wide Web (WWW'12)*, pages 131–140. ACM, 2012. (Cited on page 115.)

[95] C. Cleverdon. The Cranfield Tests on Index Language Devices. *Aslib Proceedings*, 19(6):173–194, 1967. (Cited on page 103.)

[96] Clueweb09. The ClueWeb09 Dataset, 2009. URL: `http://boston.lti.cs.cmu.edu/Data/clueweb09/`. (Cited on pages 111 and 151.)

[97] W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proceedings of IJCAI-03 Workshop on Information Integration*, pages 73–78, 2003. (Cited on page 112.)

[98] C. Cole. Inducing Expertise in History Doctoral Students via Information Retrieval Design. *Library Quarterly*, 70(1):86–109, 2000. (Cited on pages 6 and 25.)

[99] E. Collins and J. Michael. How Do Researchers in the Humanities Use Information Resources? *Library Quarterly*, 21(2), 2012. (Cited on pages 1 and 2.)

[100] J. G. Conrad and M. H. Utt. A System for Discovering Relationships by Feature Extraction from Text Databases. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'94)*, pages 260–270. Springer Verlag, 1994. (Cited on page 109.)

[101] S. B. Cousins, A. Paepcke, T. Winograd, E. A. Bier, and K. Pier. The Digital Library Integrated Task Environment (DLITE). In *Proceedings of the second ACM international conference on Digital libraries (DL'97)*, pages 142–151. ACM, 1997. (Cited on page 101.)

[102] N. Craswell and D. Hawking. Overview of the TREC-2002 Web Track. In *Proceedings of the The Eleventh Text REtrieval Conference (TREC'02)*, pages 86–95. NIST, 2002. (Cited on pages 164 and 165.)

[103] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC 2005 Enterprise Track. In *Proceedings of the The Fourteenth Text REtrieval Conference (TREC'05)*. NIST, 2005. (Cited on page 106.)

[104] N. Craswell, G. Demartini, J. Gaugaz, and T. Iofciu. L3S at INEX2008: Retrieving Entities Using Structured Information. In *Proceedings of the 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'08)*, pages 253–263. Springer, 2009. (Cited on page 110.)

[105] S. Crawford. Information Needs and Uses. *ASIST*, 13(1):61, 1978. (Cited on pages 18 and 19.)

[106] S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*, pages 708–716. ACL, 2007. (Cited on page 112.)

[107] J. D'Acci. Cultural studies, television studies, and the crisis in the humanities. In J. T. Caldwell, C. Brunsdon, L. Spigel, and J. Olsson, editors, *Television after TV: Essays on a Medium in Transition*, pages 418–446. Duke University Press, NC, 2004. (Cited on page 38.)

[108] J. Dalton and S. Huston. Semantic entity retrieval using web queries over structured RDF data. In *Semantic Search Workshop '10*, 2010. (Cited on pages 135 and 136.)

[109] M. S. Dalton and L. Charnigo. Historians and their Information Sources. *College & Research Libraries*, 65(5):400–425, 2004. (Cited on page 24.)

[110] B. Dalvi, J. Callan, and W. Cohen. Entity List Completion Using Set Expansion Techniques. In *Proceedings of the Nineteenth Text REtrieval Conference (TREC'10)*. NIST, 2011. (Cited on page 115.)

[111] H. T. Dang and K. Owczarzak. Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In *Proceedings of the First Text Analysis Conference (TAC'08)*. NIST, 2008. (Cited on pages 104 and 105.)

[112] H. T. Dang and K. Owczarzak. Overview of the TAC 2008 Update Summarization Task. In *Proceedings of First Text Analysis Conference (TAC'08)*, pages 1–16. NIST, 2008. (Cited on pages 104 and 105.)

[113] J. Davies and R. Weeks. QuizRDF: Search Technology for the Semantic Web. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04)*, pages 1—8. IEEE, 2004. (Cited on page 114.)

[114] A. de Vries, A.-M. Vercoustre, J. A. Thom, N. Craswell, and M. Lalmas. Overview of the INEX 2007 Entity Ranking Track. In *Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'07)*, pages 245–251. Springer, 2008. (Cited on pages 106, 109,

and 157.)

[115] H. De Vries, M. N. Elliott, D. E. Kanouse, and S. S. Teleki. Using Pooled Kappa to Summarize Interrater Agreement Across Many Items. *Field Methods*, 20(3):272–282, 2008. (Cited on page 42.)

[116] G. Demartini, A. de Vries, T. Iofciu, and J. Zhu. Overview of the INEX 2008 Entity Ranking Track. In *Proceedings of the 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'08)*, pages 243–252. Springer, 2008. (Cited on page 109.)

[117] B. Dervin. An overview of sense-making research: Concepts, methods, and results to date. `http://faculty.washington.edu/wpratt/MEBI598/Methods/AnOverviewofSense-MakingResearch1983a.htm`, 1983. Online; accessed August 31 2013. (Cited on page 19.)

[118] B. Dervin and K. Clark. ASQ: Asking Significant Questions. Alternative Tools for Information Need and Accountability Assessments by Libraries. Technical report, California State Library, 1987. (Cited on page 39.)

[119] B. Dervin and M. Nilan. Information Needs and Uses. *ASIST*, 21(1):3–33, 1986. (Cited on page 18.)

[120] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *SIGIF '06*, pages 154–161. ACM, 2006. (Cited on page 107.)

[121] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. LREC, 2004. (Cited on pages 104 and 105.)

[122] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering Interesting Usage Patterns in Text Collections: Integrating Text Mining with Visualization. In *Proceedings of the sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM'07)*, pages 213–222. ACM, 2007. (Cited on page 34.)

[123] W. M. Duff and C. A. Johnson. Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives. *Library Quarterly*, 72(4):472–496, 2002. (Cited on pages 2, 25, 55, and 117.)

[124] G. B. Duggan and S. J. Payne. Knowledge in the Head and on the Web: Using Topic Expertise to Aid Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*, pages 39–48. ACM, 2008. (Cited on page 31.)

[125] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. Web question answering: Is more always better? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298. ACM, 2002. (Cited on page 108.)

[126] T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, 1993. (Cited on page 154.)

[127] S. Elbassuoni, M. Ramanath, R. Schenkel, M. Sydow, and G. Weikum. Language Model-Based Ranking for Queries on RDG-Graphs. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM'11)*, pages 977–986. ACM, 2009. (Cited on pages 114 and 136.)

[128] D. Ellis and M. Haugan. Modelling the Information Seeking Patterns of Engineers and Research Scientists in an Industrial Environment. *Journal of Documentation*, 53(4):384–403, 1997. (Cited on pages 20 and 54.)

[129] D. Ellis and H. Oldman. The English Literature Researcher in the Age of the Internet. *Journal of Information Science*, 31(1):29–36, 2005. (Cited on page 24.)

[130] D. Ellis, D. Cox, and K. Hall. A Comparison of the Information Seeking Patterns of Researchers in the Physical and Social Sciences. *Journal of Documentation*, 49(4):356–369, 1993. (Cited on page 20.)

[131] D. A. Evans and C. Zhai. Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL'96)*, pages 17–24. ACM, 1996. (Cited on page 100.)

[132] Y. Fang, L. Si, Z. Yu, Y. Xian, and Y. Xu. Entity Retrieval with Hierarchical Relevance Model. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC'09)*. NIST, 2009. (Cited on pages 111, 168, and 169.)

[133] Y. Fang, L. Si, N. Somasundaram, S. Al-Ansari, Z. Yu, and Y. Xian. Purdue at TREC 2010 Entity Track: a Probabilistic Framework for Matching Types between Candidate and Target Entities. In *Proceedings of the Nineteenth Text REtrieval Conference (TREC'10)*. NIST, 2011. (Cited on page 115.)

[134] J. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05*, pages 363–370. ACL, 2005. (Cited on page 121.)

[135] J. G. Fiscus and G. R. Doddington. Topic detection and tracking evaluation overview. In *Topic detection and tracking*, pages 17–31. Kluwer, 2002. (Cited on page 106.)

[136] S. Fissaha Adafre, M. de Rijke, and E. Tjong Kim Sang. Entity Retrieval. In *Recent Advances in Natural Language Processing (RANLP 2007)*. RANLP, 2007. (Cited on pages 109 and 110.)

[137] E. Fox and J. Shaw. Combination of Multiple Searches. In *Proceedings of the Third Text REtrieval Conference (TREC'94)*, pages 243–243. NIST, 1994. (Cited on page 115.)

[138] M. Franz, T. Ward, J. Mccarley, and W. Zhu. Unsupervised and supervised clustering for topic tracking. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, pages 310–317. ACM, 2001. (Cited on page 107.)

[139] N. Fuhr, J. Kamps, M. Lalmas, and A. Trotman, editors. *Focused access to XML documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, volume 4862 of *Lecture Notes in Computer Science*, 2008. Springer Verlag, Heidelberg. (Cited on page 112.)

[140] J. Gantz and D. Reinsel. Extracting Value from Chaos. Technical report, IDC, 2011. (Cited on page 1.)

[141] J. Gao, Q. Wu, C. Burges, K. Svore, Y. Su, N. Khan, S. Shah, and H. Zhou. Model Adaptation via Model Interpolation and Boosting for Web Search Ranking. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, pages 505–513. ACL, 2009. (Cited on page 110.)

[142] W. D. Garvey and B. C. Griffith. Communication and Information Processing within Scientific Disciplines: Empirical Findings for Psychology. *Information Storage and Retrieval*, 8(3):123–136, 1972. (Cited on page 25.)

[143] S. Geva, J. Kamps, R. Schenkel, and A. Trotman. *INEX 2010 Workshop Pre-proceedings*. Citeseer, 2010. (Cited on page 105.)

[144] F. Gey, R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras. GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview. In *Accessing Multilingual Information Repositories*, pages 908–919. Springer, 2006. (Cited on pages 104 and 105.)

[145] Z. Ghahramani and K. Heller. Bayesian Sets. In *Advances in Neural Information Processing Systems 18*, pages 435–442, 2006. (Cited on page 110.)

[146] W. Goffman. Information Science: Discipline or Disappearance. *Aslib Proceedings*, 22(12):589–596, 1970. (Cited on page 18.)

[147] M. K. Gold. *Debates in the Digital Humanities*. University of Minnesota Press, 2012. (Cited on page 1.)

[148] G. Golovchinsky and J. Pickens. Interactive Information Seeking via Selective Application of Contextual Knowledge. In *Proceedings of the Third Symposium on Information Interaction in Context (IIiX'10)*, pages 145–154. ACM, 2010. (Cited on pages 4 and 34.)

[149] Googlesets, 2009. `http://labs.google.com/sets`, accessed Jan. 2009. (Cited on page 110.)

[150] M. D. Gordon and P. Lenk. When is the Probability Ranking Principle Suboptimal? *Journal of the American Society for Information Science*, 43(1):1–14, 1992. (Cited on page 104.)

[151] C. C. Gould. *Information needs in the humanities: An assessment*. RLG, 1988. (Cited on pages 1, 23, and 25.)

[152] D. Graff. *The AQUAINT corpus of English news text*. Linguistic Data Consortium, 2002. (Cited on pages 104 and 105.)

[153] D. Graus, T. Kenter, M. Bron, E. Meij, and M. de Rijke. Context-Based Entity Linking - University of Amsterdam at TAC 2012. In *Proceedings of the Fifth Text Analysis Conference (TAC'12)*. NIST, 2012. (Cited on page 13.)

[154] R. Grishman and B. Sundheim. Message Understanding Conference-6: A Brief History. In *16th International Conference on Computational Linguistics, Proceedings of the Conference (COLING'96)*, pages 466–471. ACM, 1996. (Cited on pages 104 and 105.)

[155] R. Guha, R. Mccool, and E. Miller. Semantic search. In *WWW '03*, pages 700–709, 2003. (Cited on page 111.)

[156] D. Harman. How Effective is Suffixing? *Journal of the American Society for Information Science*, 42 (1):7–15, 1991. (Cited on page 100.)

[157] D. Harman. Overview of the First TREC Conference (TREC-1). In *Proceedings of the First Text REtrieval Conference (TREC-1)*, pages 1–20. NIST, 1992. (Cited on page 103.)

[158] D. K. Harman. The TREC Test Collections. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and evaluation in information retrieval*. MIT, 2005. (Cited on page 123.)

[159] G. Harmon. Information Need Transformation During Inquiry: a Reinterpretation of User Relevance. In *Proceedings of the American Society for Information Science (ASIS'70)*, pages 41–3. ASIS, 1970. (Cited on page 25.)

[160] J. He, M. Bron, and M. de Rijke. A query performance analysis for result diversification. In *Advances in Information Retrieval Theory*, pages 351–355. Springer Berlin Heidelberg, 2011. (Cited on page 13.)

[161] J. He, M. de Rijke, M. Sevenster, R. Van Ommering, and Y. Qian. Generating Links to Background Knowledge: A Case Study Using Narrative Radiology Reports. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*, pages 1867–1876. ACM,

2011. (Cited on pages 112 and 140.)

[162] J. He, M. Bron, and A. de Vries. Characterizing Stages of a Multi-Session Complex Search Task through Direct and Indirect Query Modifications. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*, pages 897–900. ACM, 2013. (Cited on page 12.)

[163] M. Hearst. TileBars: Visualization of Term Distribution Information in Full Text Information Access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'95)*, pages 59–66. ACM, 1995. (Cited on page 34.)

[164] M. Hearst. *Search user interfaces*. Cambridge University Press, 2009. (Cited on pages 4 and 34.)

[165] M. A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th conference on Computational linguistics (COLING'92)*, pages 539–545. ACM, 1992. (Cited on page 109.)

[166] M. A. Hearst and D. Rosner. Tag Clouds: Data Analysis Tool or Social Signaller? In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS'08)*, pages 160–169. IEEE, 2008. (Cited on pages 34 and 86.)

[167] M. Henzinger, B.-W. Chang, B. Milch, and S. Brin. Query-Free News Search. *World Wide Web*, 8(2): 101–126, 2005. (Cited on pages 4 and 107.)

[168] A. J. Hey and A. Trefethen. The Data Deluge: An e-Science Perspective. In F. Berman, G. Fox, and T. Hey, editors, *Grid Computing: Making the Global Infrastructure a Reality*. Wiley and Sons, 2003. (Cited on pages 1 and 81.)

[169] K. Hofmann, B. Huurnink, M. Bron, and M. de Rijke. Comparing Click-Through Data to Purchase Decisions for Retrieval Evaluation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*, pages 761–762. ACM, 2010. (Cited on page 13.)

[170] C. Hölscher and G. Strube. Web Search Behavior of Internet Experts and Newbies. *Computer Networks*, 33(1):337–346, 2000. (Cited on page 31.)

[171] D. W. C. Huang, Y. Xu, A. Trotman, and S. Geva. Overview of INEX 2007 link the wiki track. In *Focused Access to XML Documents*, pages 373–387. Springer, 2008. (Cited on page 106.)

[172] J. Huang, R. White, and S. Dumais. No Clicks, No Problem: Using Cursor Movements to Understand and Imporve Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*, pages 1225–1234. ACM, 2011. (Cited on page 64.)

[173] B. Huurnink, K. Hofmann, M. de Rijke, and M. Bron. Validating query simulators: An experiment using commercial searches and purchases. In *Multilingual and Multimodal Information Access Evaluation*, pages 40–51. Springer Berlin Heidelberg, 2010. (Cited on page 13.)

[174] B. Huurnink, A. Bronner, M. Bron, J. van Gorp, B. de Goede, and J. Wees. AVResearcher: Exploring Audiovisual Metadata. In *Proceedings of the 13th Dutch-Belgian Information Retrieval Workshop (DIR'13)*, pages 64–65. CEUR, 2013. (Cited on pages 12 and 176.)

[175] P. Ingwersen and K. Järvelin. *The turn: Integration of information seeking and retrieval in context*, volume 18. Springer, 2005. (Cited on pages 21, 22, and 54.)

[176] P. E. R. Ingwersen. *Information Retrieval Interaction*. Taylor Graham, 1992. (Cited on page 19.)

[177] C. Jacquemin, J. L. Klavans, and E. Tzoukermann. Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. In *Proceedings of the eighth conference of European chapter of the Association for Computational Linguistics (EACL'97)*, pages 24–31. ACl, 1997. (Cited on page 100.)

[178] J. Jämsen, T. Näppilä, and P. Arvola. Entity Ranking Based on Category Expansion. In *Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'07)*, pages 264–278. Springer, 2008. (Cited on page 110.)

[179] B. J. Jansen, A. Spink, and T. Saracevic. Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web. *Information Processing & Management*, 36(2):207–227, 2000. (Cited on page 101.)

[180] B. J. Jansen, D. L. Booth, and A. Spink. Determining the Informational, Navigational, and Transactional Intent of Web Queries. *Information Processing & Management*, 44(3):1251–1266, 2008. (Cited on page 29.)

[181] K. B. Jensen. *A handbook of media and communication research: qualitative and quantitative methodologies*. Routledge, 2002. (Cited on page 40.)

[182] J. Jiang, W. Liu, X. Rong, and Y. Gao. Adapting Language Modeling Methods for Expert Search to Rank Wikipedia Entities. In *Proceedings of the 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'08)*, pages 264–272. Springer, 2009. (Cited on page 110.)

[183] J. Kamps and M. Koolen. The Importance of Link Evidence in Wikipedia. In *Proceedings of the 30th*

*European Conference on Information Retrieval (ECIR'08)*, pages 270–282. Springer, 2008. (Cited on page 110.)

[184] J. Kamps, S. Geva, and A. Trotman. Report on the SIGIR 2008 Workshop on Focused Retrieval. *ACM SIGIR Forum*, 42(2):59–65, 2008. (Cited on page 108.)

[185] N. Kando. Overview of the Fifth NTCIR Workshop. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*. NII, 2005. (Cited on pages 104 and 105.)

[186] R. Kaptein and J. Kamps. Finding Entities in Wikipedia Using Links and Categories. In *Proceedings of the 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'08)*, pages 273–279. Springer, 2009. (Cited on page 110.)

[187] R. Kaptein, M. Koolen, and J. Kamps. Result Diversity and Entity Ranking Experiments. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC'09)*. NIST, 2009. (Cited on pages 111, 157, and 165.)

[188] M. Kellar, C. Watters, and M. Shepherd. A Goal-Based Classification of Web Information Tasks. *Journal of the American Society for Information Science and Technology*, 43(1):1–22, 2006. (Cited on page 29.)

[189] D. Kelly. Measuring Online Information Seeking Context, Part 2: Findings and Discussion. *Journal of the American Society for Information Science and Technology*, 57(14):1862–1874, 2006. (Cited on page 31.)

[190] D. Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval*, 3(1-2):1–224, 2009. (Cited on pages 29, 58, 86, and 87.)

[191] D. Kelly and J. Teevan. Implicit Feedback for Inferring User Preference: a Bibliography. *ACM SIGIR Forum*, 37(2):18–28, 2003. (Cited on page 34.)

[192] D. Kennedy and M. Bishop. Google Earth and the Archaeology of Saudi Arabia A Case Study from the Jeddah Area. *Journal of Archaeological Science*, 38(6):1284–1293, 2011. (Cited on page 1.)

[193] R. Khare. Microformats: the Next (Small) Thing on the Semantic Web? *IEEE Internet Computing*, 10 (1):68–75, 2006. (Cited on page 113.)

[194] S. Klarman, R. Hoekstra, and M. Bron. Versions and Applicability of Concept Definitions in Legal Ontologies. In *OWLED '08*, 2008. (Cited on page 13.)

[195] D. Konopnicki and O. Shmueli. W3QS A Query System for the World-Wide Web. In *Proceedings of 21th International Conference on Very Large Data Bases (VLDB'95)*, pages 54–65. Morgan Kaufmann, 1995. (Cited on page 101.)

[196] H. Köpcke, A. Thor, and E. Rahm. Evaluation of Entity Resolution Approaches on Real-World Match Problems. *Proceedings of the VLDB Endowment*, 3(1):484–493, 2010. (Cited on page 112.)

[197] C. Kuhlthau. Inside the Search Process: Information Seeking from the User's Perspective. *Journal of the American Society for Information Science*, 42(5):361–371, 1991. (Cited on pages 20, 30, and 54.)

[198] B. Kules, R. Capra, M. Banta, and T. Sierra. What Do Exploratory Searchers Look at in a Faceted Search Interface? In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'09)*, pages 313–322. ACM, 2009. (Cited on page 83.)

[199] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*, pages 297–304. ACM, 2004. (Cited on pages 107 and 121.)

[200] C. Kwok, O. Etzioni, and D. Weld. Scaling Question Answering to the Web. *ACM Transactions on Information Systems*, 19(3):242–262, 2001. (Cited on page 108.)

[201] J. Lafferty and C. Zhai. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'01)*, pages 111–119. ACM, 2001. (Cited on page 160.)

[202] R. Lagerweij, M. Bron, C. Monz, and M. de Rijke. University of Amsterdam at TAC 2011: English Slot Filling Task. In *Proceedings of the Fourth Text Analysis Conference (TAC'11)*. NIST, 2011. (Cited on page 13.)

[203] M. Lalmas. Aggregated search. In *Advanced Topics in Information Retrieval*, pages 109–123. Springer, 2011. (Cited on pages 32 and 33.)

[204] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. (Cited on page 42.)

[205] V. Lavrenko, J. Allan, E. Deguzman, D. Laflamme, V. Pollard, and S. Thomas. Relevance Models for Topic Detection and Tracking. In *Proceedings of the second international conference on Human Language Technology Research*, pages 115–121, 2002. (Cited on pages 107 and 129.)

[206] G. Lernout. Comparative Literature in the Low Countries. *Comparative Critical Studies*, 3(1):37–46, 2006. (Cited on page 176.)

[207] M. E. Lesk. Word-Word Associations in Document Retrieval Systems. *American Documentation*, 20 (1):27–38, 1969. (Cited on page 101.)

[208] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-Based Multimedia Information Retrieval: State of the Art and Challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):1–19, 2006. (Cited on page 118.)

[209] Y. Li. Exploring the Relationships Between Work Task and Search Task in Information Search. *Journal of the American Society for Information Science and Technology*, 60(2):275–291, 2008. (Cited on page 62.)

[210] Y. Li and N. Belkin. A Faceted Approach to Conceptualizing Tasks in Information Seeking. *Information Processing & Management*, 44(6):1822–1837, 2008. (Cited on page 30.)

[211] Z. Li, B. Wang, M. Li, and W. Ma. A probabilistic model for retrospective news event detection. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*, pages 106–113. ACM, 2005. (Cited on page 121.)

[212] P. Lieberman. *The biology and evolution of language*. Harvard University Press, 1984. (Cited on page 18.)

[213] J. Lin and W. J. Wilbur. PubMed Related Articles: a Probabilistic Topic-Based Model for Content Similarity. *BMC Bioinformatics*, 8(1):423, 2007. (Cited on page 101.)

[214] Y. Lin, J. Ahn, P. Brusilovsky, D. He, and W. Real. Imagesieve: Exploratory Search of Museum Archives with Named Entity-Based Faceted Browsing. In *Proceedings of the 73rd Annual Meeting of the Association for Information Science and Technology (ASIST'10)*, pages 1–10. ASIST, 2010. (Cited on page 34.)

[215] J. Liu and N. Belkin. Personalizing Information Retrieval for Multi-Session Tasks: The Roles of Task Stage and Task Type. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*, pages 26–33. ACM, 2010. (Cited on pages 30 and 71.)

[216] J. Liu, N. J. Belkin, X. Zhang, and X. Yuan. Examining Users' Knowledge Change in the Task Completion Process. *Information Processing & Management*, 49(5):1058–1074, 2012. (Cited on page 32.)

[217] T.-Y. Liu. *Learning to rank for information retrieval*. springer, 2011. (Cited on page 138.)

[218] X. Liu and H. Fang. A Study of Entity Search in Semantic Search Workshop. In *Proceedings of the 3rd International Semantic Search Workshop (SemSearch'10)*. ACM, 2010. (Cited on page 135.)

[219] V. Lopez, C. Unger, P. Cimiano, and E. Motta. Evaluating Question Answering over Linked Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(1):1, 2013. (Cited on page 105.)

[220] W. Lu, Q. Wang, and B. Larsen. Simulating Aggregated Interfaces. In *Workshop on Aggregated Search*, pages 24–28. ACM, 2012. (Cited on page 33.)

[221] A. Lund. Measuring Usability with the USE Questionnaire. *Usability Interface*, 8(2):8, 2001. (Cited on page 88.)

[222] B. Lunn. User Needs in Television Archive Access: Acquiring Knowledge Necessary for System Design. *Journal of Digital Information*, 10(6), 2009. (Cited on pages 26 and 81.)

[223] A. Lunzer and K. Hornbäek. Subjunctive Interfaces: Extending Applications to Support Parallel Setup, Viewing and Control of Alternative Scenarios. *ACM Transactions on Computer-Human Interaction*, 14 (4):17, 2008. (Cited on pages 4, 35, 79, and 81.)

[224] R. H. Lytle. Intellectual Access to Archives. *American Archivist*, 43(2):191–207, 1980. (Cited on page 117.)

[225] Q. Ma, A. Nadamoto, and K. Tanaka. Complementary Information Retrieval for Cross-Media News Content. *Information Systems*, 31(7):659–678, 2006. (Cited on page 107.)

[226] C. Macdonald and I. Ounis. The TREC Blogs06 collection: Creating and analysing a blog test collection. Technical report, University of Glasgow, 2006. (Cited on pages 104 and 105.)

[227] P. Majumder, M. Mitra, D. Pal, A. Bandyopadhyay, S. Maiti, S. Pal, D. Modak, and S. Sanyal. The FIRE 2008 Evaluation Exercise. *ACM Transactions on Asian Language Information Processing*, 9(3): 10, 2010. (Cited on pages 104 and 105.)

[228] C. Mangold. A Survey and Classification of Semantic Search Approaches. *International Journal of Metadata, Semantics and Ontologies*, 2(1):23–34, 2007. (Cited on page 111.)

[229] C. D. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999. (Cited on page 153.)

[230] G. Marchionini. *Information seeking in electronic environments*. Cambridge University Press, New York, NY, USA, 1995. (Cited on pages 21 and 30.)

[231] G. Marchionini. Exploratory Search: from Finding to Understanding. *Communications of the ACM*, 49 (4):41–46, 2006. (Cited on page 33.)

[232] M. Mccreadie, C. Macdonald, I. Ounis, J.Peng, and R. L. T. Santos. University of Glasgow at TREC

2009. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC'09).* NIST, 2009. (Cited on pages 111 and 168.)

[233] K. Mckeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *HLT '02*, pages 280–285, 2002. (Cited on page 4.)

[234] P. Mcnamee and H. T. Dang. Overview of the TAC 2009 Knowledge Base Population Track. In *Proceedings of the Second Text Analysis Conference (TAC'09).* NIST, 2009. (Cited on pages 104 and 105.)

[235] L. Meho and H. Tibbo. Modeling the Information-Seeking Behavior of Social Scientists: Ellis's Study Revisited. *Journal of the American Society for Information Science and Technology*, 54(6):570–587, 2003. (Cited on page 79.)

[236] E. Meij and M. de Rijke. Thesaurus-Based Feedback to Support Mixed Search and Browsing Environments. In *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'07)*, pages 247–258. Springer, 2007. (Cited on page 112.)

[237] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. de Rijke. Learning semantic query suggestions. In *The Semantic Web-ISWC 2009*, pages 424–440. Springer Berlin Heidelberg, 2009. (Cited on page 13.)

[238] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. de Rijke. Mapping Queries to the Linking Open Data Cloud: A Case Study Using DBpedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):418–433, 2011. (Cited on page 13.)

[239] E. Meij, W. Weerkamp, and M. de Rijke. Adding Semantics to Microblog Posts. In *Proceedings of the fifth ACM international conference on Web search and data mining (WSDM'12)*, pages 563–572. ACM, 2012. (Cited on pages 112 and 140.)

[240] M. Melucci and R. Baeza-Yates. *Advanced topics in information retrieval*, volume 33. Springer, 2011. (Cited on pages 3 and 61.)

[241] H. Menzel. The Information Needs of Current Scientific Research. *Library Quarterly*, 34(1):4–19, 1964. (Cited on pages 1 and 23.)

[242] H. Menzel. Information Needs and Uses in Science and Technology. *Annual Review of Information Science and Technology*, 1(1):41–69, 1966. (Cited on page 18.)

[243] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, and Others. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182, 2011. (Cited on pages 1, 2, and 28.)

[244] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07*, volume 7, pages 233–242, 2007. (Cited on page 112.)

[245] S. Miller. Information-seeking behaviour of academic scientists in the electronic age. `http://www.researchknowledge.ca/initiatives/evaluation/LitReview-SusanMiller.pdf`, 2002. Online. (Cited on page 24.)

[246] D. N. Milne and I. H. Witten. Learning to Link with Wikipedia. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*, pages 509–518. ACM, 2008. (Cited on page 112.)

[247] J. Minker, G. A. Wilson, and B. H. Zimmerman. An Evaluation of Query Expansion by the Addition of Clustered Terms for a Document Retrieval System. *Information Storage and Retrieval*, 8(6):329–348, 1972. (Cited on page 101.)

[248] M. Mitra, A. Singhal, and C. Buckley. Improving Automatic Query Expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'98)*, pages 206–214. ACM, 1998. (Cited on page 101.)

[249] S. Mizzaro. How Many Relevances in Information Retrieval? *Interacting with Computers*, 10(3): 303–320, 1998. (Cited on page 103.)

[250] C. Monz. *From Document Retrieval to Question Answering*. PhD thesis, University of Amsterdam, 2003. (Cited on page 108.)

[251] C. N. Mooers. Information Retrieval Viewed as Temporal Signaling. In *Proceedings of the International Congress of Mathematicians*, pages 572–573. AMS, 1950. (Cited on page 18.)

[252] J. B. Morrison, P. Pirolli, and S. K. Card. A Taxonomic Analysis of what World Wide Web Activities Significantly Impact People's Decisions and Actions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'01)*, pages 163–164. ACM, 2001. (Cited on page 29.)

[253] T. Nasukawa and T. Nagano. Text Analysis and Knowledge Mining System. *IBM Systems Journal*, 40 (4):967–984, 2001. (Cited on page 34.)

[254] R. Neumayer, K. Balog, and K. Nörväag. On the Modeling of Entities for Ad-Hoc Entity Search in the Web of Data. In *Proceedings of the 34th European Conference on IR Research (ECIR'12)*, pages 133–145. Springer, 2012. (Cited on pages 135, 136, and 137.)

[255] C. Newbold, O. Boydt-Barrett, and H. Van Den Bulck. *The media book*. London, 2002. (Cited on page 38.)

[256] J. D. Novak and D. B. Gowin. *Learning how to learn*. Cambridge University Press, 1984. (Cited on page 133.)

[257] H. O'Brien and E. Toms. The Development and Evaluation of a Survey to Measure User Engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, 2009. (Cited on page 73.)

[258] V. L. O'Day and R. Jeffries. Orienteering in an Information Landscape: how Information Seekers Get from here to there. In *Proceedings of the ACM CHI 93 Human Factors in Computing Systems Conference (CHI'93)*, pages 438–445. ACM, 1993. (Cited on page 20.)

[259] P. Ogilvie and J. P. Callan. Combining Document Representations for Known-Item Search. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'03)*, pages 143–150. ACM, 2003. (Cited on page 164.)

[260] J. C. Olney. Library Cataloging and Classification. Technical report, DTIC Document, 1963. (Cited on page 18.)

[261] W. Paisley and H. Menzel. Information Uses and Needs. *Annual Review of Information Science and Technology*, 38(1), 1968. (Cited on page 18.)

[262] C. Palmer. Scholarly Work and the Shaping of Digital Access. *Journal of the American Society for Information Science and Technology*, 56(11):1140–1153, 2005. (Cited on page 79.)

[263] C. L. Palmer, L. C. Teffeau, and C. M. Pirmann. Scholarly information practices in the online environment. `www.oclc.org/programs/publications/reports/2009-02.pdf`, 2009. Online; accessed 28-August-2013. (Cited on page 27.)

[264] N. R. Pandit. The Creation of Theory: A Recent Application of the Grounded Theory Method. *The Qualitative Report*, 2(4):1–14, 1996. (Cited on page 39.)

[265] E. A. Parsons. *The Alexandrian Library, glory of the Hellenic world: its rise, antiquities, and destructions*. Elsevier Press, 1952. (Cited on page 18.)

[266] E. S. Patterson, E. M. Roth, and D. D. Woods. Predicting Vulnerabilities in Computer-Supported Inferential Analysis under Data Overload. *Cognition, Technology & Work*, 3(4):224–237, 2001. (Cited on page 31.)

[267] J. Pehcevski, A.-M. Vercoustre, and J. A. Thom. Exploiting Locality of Wikipedia Links in Entity Ranking. In *Proceedings of the 30th European Conference on Information Retrieval (ECIR'08)*, pages 258–269. Springer, 2008. (Cited on page 110.)

[268] J. Pérez-Agüera, J. Arroyo, J. Greenberg, J. Iglesias, and V. Fresno. Using BM25F for Semantic Search. In *Proceedings of the 3rd International Semantic Search Workshop (SemSearch'10)*, page 2. ACM, 2010. (Cited on pages 135 and 140.)

[269] D. Petkova and W. B. Croft. Proximity-Based Document Representation for Named Entity Retrieval. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM'07)*, pages 731–740. ACM, 2007. (Cited on page 108.)

[270] V. Petras, N. Ferro, M. Gäde, A. Isaac, M. Kleineberg, I. Masiero, M. Nicchio, and J. Stiller. Cultural Heritage in CLEF (CHiC) Overview 2012. In *Proceedings of the Multilingual and Multimodal Information Access Evaluation, International Conference of the Cross-Language Evaluation Forum (CLEF'12)*. Springer, 2012. (Cited on pages 104 and 105.)

[271] C. S. Pierce. Logic as semiotic: The theory of signs. In *The philosophical writings of Pierce*, pages 98–119. Dover, 1955. (Cited on page 117.)

[272] P. Pirolli and S. K. Card. Information Foraging. *Psychological Review*, 106(4):643–675, 1999. (Cited on page 20.)

[273] J. M. Ponte and W. B. Croft. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'98)*, pages 275–281. ACM, 1998. (Cited on page 102.)

[274] J. Pound, P. Mika, and H. Zaragoza. Ad-hoc Object Ranking in the Web of Data. In *WWW '10*, 2010. (Cited on pages 106 and 114.)

[275] Y. Qiu and H.-P. Frei. Concept Based Query Expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 160–169. ACM, 1993. (Cited on page 101.)

[276] D. Radev, J. Otterbacher, A. Winkel, and S. Blair-Goldensohn. NewsInEssence: Summarizing Online News Topics. *Communications of the ACM*, 48(10):95–98, 2005. (Cited on pages 4 and 107.)

[277] H. Raghavan, J. Allan, and A. Mccallum. An Exploration of Entity Models, Collective Classification and Relation Description. In *Proceedings of the ACM SIGKDD Workshop on Link Analysis and Group*

*Detection (LinkKDD2004)*, page 33, 2004. (Cited on page 109.)

[278] S. Raghavan and H. Garcia-Molina. Crawling the hidden web. Technical report, Stanford, 2000. (Cited on page 32.)

[279] H. Reiterer, G. Tullius, and T. M. Mann. Insyder: a Content-Based Visual-Information-Seeking System for the Web. *International Journal on Digital Libraries*, 5(1):25–41, 2005. (Cited on page 34.)

[280] P. Resnick and H. R. Varian. Recommender Systems. *Communications of the ACM*, 40(3):56–58, 1997. (Cited on page 107.)

[281] E. Riloff. Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence ( AAAI'96)*, pages 1044–1049. AAAI, 1996. (Cited on page 109.)

[282] S. Robertson and H. Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009. (Cited on page 102.)

[283] S. E. Robertson. The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304, 1977. (Cited on page 102.)

[284] S. E. Robertson. On Term Selection for Query Expansion. *Journal of Documentation*, 46(4):359–364, 1990. (Cited on page 101.)

[285] S. E. Robertson and K. Spärck-Jones. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976. (Cited on page 102.)

[286] C. Rocha, D. Schwabe, and M. Aragao. A hybrid approach for searching in the semantic web. In *WWW '04*, pages 374–383, 2004. (Cited on page 114.)

[287] K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-Mouse Coordination Patterns on Web Search Results Pages. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems*, pages 2997–3002. ACM, 2008. (Cited on pages 64 and 77.)

[288] D. E. Rose and D. Levinson. Understanding User Goals in Web Search. In *Proceedings of the 13th international conference on World Wide Web (WWW'04)*, pages 13–19. ACM, 2004. (Cited on page 29.)

[289] H. D. Rozanski, G. Bollman, and M. Lipman. Seize the occasion! The seven-segment system for online marketing. `http://www.strategy-business.com/article/19940?gko=d29c9`, 2001. Online; accessed 28-august-2013. (Cited on page 29.)

[290] W. Rundell. *In Pursuit of American History: Research and Training in the United States*. University of Oklahoma Press, 1970. (Cited on page 1.)

[291] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. The Cost Structure of Sensemaking. In *Proceedings of the ACM CHI 93 Human Factors in Computing Systems Conference (CHI'93)*, pages 269–276. ACM, 1993. (Cited on page 20.)

[292] D. M. Russell, D. Tang, M. Kellar, and R. Jeffries. Task Behaviors During Web Search: The Difficulty of Assigning Labels. In *Proceedings of the 42nd Annual Hawaii International Conference on System Sciences*, pages 1–5. IEEE, 2009. (Cited on page 29.)

[293] G. Salton. *The SMART retrieval system-experiments in automatic document processing*. Prentice-Hall, Inc., 1971. (Cited on page 102.)

[294] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. In K. Sparck Jones and P. Willett, editors, *Readings in information retrieval*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. (Cited on page 34.)

[295] G. Salton, A. Wong, and C. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1975. (Cited on pages 102 and 119.)

[296] R. L. Santos, C. Macdonald, and I. Ounis. Aggregated search result diversification. In *Advances in Information Retrieval Theory*, pages 250–261. Springer, 2011. (Cited on page 33.)

[297] T. Saracevic. Information Science: Origin, Evolution and Relations. In *Proceedings of the International Conference held for the celebration of 20th Anniversary of the Department of Information Studies (CoLIS1)*, pages 5–27. Taylor, 1992. (Cited on page 18.)

[298] T. Saracevic. Modeling Interaction in Information Retrieval (IR): A Review and Proposal. In *Proceedings of the 59th Annual Meeting of the American Society for Information Science (ASIS'96)*, pages 3–9. ASIS, 1996. (Cited on page 21.)

[299] M. Sayyadian, A. Shakery, A. Doan, and C. Zhai. Toward Entity Retrieval over Structured and Text Data. In *Proceedings of the ACM SIGIR 2004 Workshop on the Integration of Information Retrieval and Databases (WIRD'04)*, pages 47–54. ACM, 2004. (Cited on page 109.)

[300] S. Schreibman, R. Siemens, and J. Unsworth. *A companion to digital humanities*, volume 26. Wiley-Blackwell, 2008. (Cited on page 1.)

[301] J. Seo, W. Croft, K. Kim, and J. Lee. Smoothing Click Counts for Aggregated Vertical Search. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information*

*Retrieval*, pages 387–398. Springer, 2011. (Cited on page 33.)

[302] P. Serdyukov and A. de Vries. Delft University at the TREC 2009 Entity Track: Ranking Wikipedia Entities. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC'09)*. NIST, 2009. (Cited on pages 111, 165, 167, and 168.)

[303] P. Serdyukov and D. Hiemstra. Being Omnipresent to be Almighty: The Importance of the Global Web Evidence for Organizational Expert Finding. In *Proceedings of the SIGIR 2008 Workshop on Future Challenges in Expertise Retrieval (fCHER)*, pages 17–24. ACM, 2008. (Cited on page 108.)

[304] D. Sheldon, M. Shokouhi, M. Szummer, and N. Craswell. LambdaMerge: Merging the Results of Query Reformulations. In *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM'11)*, pages 795–804. ACM, 2011. (Cited on page 110.)

[305] A. Shiri. Metadata-Enhanced Visual Interfaces to Digital Libraries. *Journal of Information Science*, 34 (6):763–775, 2008. (Cited on pages 4, 34, 55, and 60.)

[306] M. Shokouhi and L. Si. Federated Search. *Foundations and Trends in Information Retrieval*, 5(1): 1–102, 2011. (Cited on pages 32, 59, and 138.)

[307] A. Siochi and R. Ehrich. Computer Analysis of User Interfaces Based on Repetition in Transcripts of User Sessions. *ACM Transactions on Information Systems*, 9(4):309–335, 1991. (Cited on page 89.)

[308] M. D. Smucker and J. Allan. Find-Similar: Similarity Browsing as a Search Tool. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*, pages 461–468. ACM, 2006. (Cited on pages 61 and 101.)

[309] F. Song and W. B. Croft. A General Language Model for Information Retrieval. In *Proceedings of the eighth international conference on Information and knowledge management (CIKM'99)*, pages 316–321. ACM, 1999. (Cited on page 164.)

[310] A. Spink. Study of Interactive Feedback During Mediated Information Retrieval. *Journal of the American Society for Information Science*, 48(5):382–394, 1997. (Cited on page 21.)

[311] S. Stone. Humanities Scholars Information Needs and Uses. *Journal of Documentation*, 38(4):292–313, 1982. (Cited on pages 23, 25, and 26.)

[312] A. Strauss and J. Corbin. *Basics of qualitative research: Grounded theory procedures and techniques.* Sage Publications, Inc, 1990. (Cited on pages 38 and 41.)

[313] S. Sushmita, H. Joho, and M. Lalmas. A Task-Based Evaluation of an Aggregated Search Interface. In *Proceedings of the 16th International Symposium on String Processing and Information Retrieval (SPIRE'09)*, pages 322–333. Springer, 2009. (Cited on pages 33 and 77.)

[314] E. Svenonius. Unanswered Questions in the Design of Controlled Vocabularies. *Journal of the American Society for Information Science*, 37(5):331–340, 1986. (Cited on page 100.)

[315] D. Tabatabai and B. M. Shore. How Experts and Novices Search the Web. *Library & Information Science Research*, 27(2):222–248, 2005. (Cited on page 31.)

[316] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *HLT-NAACL '06*, pages 407–414, 2006. (Cited on pages 100 and 107.)

[317] R. S. Taylor. The Process of Asking Questions. *American Documentation*, 13(4):391–396, 1962. (Cited on pages 19 and 101.)

[318] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The Proceedings of the SIGCHI Conference on Human Factors in Computing Systemsq Search Engine is not Enough: a Study of Orienteering Behavior in Directed Search. In *CHI '04*, pages 415–422. ACM, 2004. (Cited on page 20.)

[319] R. Tennant. The Convenience Catastrophe. *Library Journal*, 126(20):39–40, 2001. (Cited on page 28.)

[320] N. Thumim. *Self-representation and digital culture*. Palgrave Macmillan, 2012. (Cited on page 38.)

[321] H. R. Tibbo. How Historians Locate Primary Resource Materials: Educating and Serving the Next Generation of Scholars. In *Proceedings of the 11th ACRL Conference*. ACRL, 2003. (Cited on page 175.)

[322] E. F. Tjong Kim Sang and S. Buchholz. Introduction to the CoNLL-2000 Shared Task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 127–132. ACL, 2000. (Cited on pages 104 and 105.)

[323] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, pages 142–147. ACL, 2003. (Cited on pages 104 and 105.)

[324] E. G. Toms and H. L. O'Brien. Understanding the Information and Communication Technology Needs of the E-Humanist. *Journal of Documentation*, 64(1):102–130, 2008. (Cited on pages 1, 22, and 26.)

[325] A. Tonon, G. Demartini, and P. Cudré-Mauroux. Combining Inverted Indices and Structured Search for Ad-Hoc Object Retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*, pages 125–134. ACM, 2012. (Cited on page 114.)

[326] T. Tran, H. Wang, S. Rudolph, and P. Cimiano. Top-k exploration of query candidates for efficient keyword search on graph-shaped (RDF) data. In *ICDE'09*, pages 405–416, 2009. (Cited on page 114.)

[327] T. Tran, P. Mika, H. Wang, and M. Grobelnik. SemSearch'11: the 4th Semantic Search Workshop. In *Proceedings of the 20th international conference companion on World wide web*, pages 315–316. ACM, 2011. (Cited on page 105.)

[328] M. Tsagkias, M. de Rijke, and W. Weerkamp. Linking Online News and Social Media. In *WSDM'11*, pages 565–574. ACM, 2011. (Cited on pages 107 and 124.)

[329] T. Tsikrika, P. Serdyukov, H. Rode, T. Westerveld, R. Aly, D. Hiemstra, and A. P. de Vries. Structured Document Retrieval, Multimedia Retrieval, and Entity Ranking Using PF/Tijah. In *Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'07)*, pages 306–320. Springer, 2008. (Cited on page 110.)

[330] T. Tsikrika Et Al. Structured Document Retrieval, Multimedia Retrieval, and Entity Ranking Using PF/Tijah. In *Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'07)*, pages 306–320. Springer, 2008. (Cited on page 110.)

[331] G. Tummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker. Sig.ma: Live Views on the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4): 355–364, 2010. (Cited on pages 114 and 134.)

[332] D. Tunkelang. Faceted Search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1 (1):1–80, 2009. (Cited on page 59.)

[333] R. Turknett, B. Westing, and S. Moore. 1000 Words: Advanced Visualization for the Humanities. In *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, page 66, 2013. (Cited on page 28.)

[334] H. Turtle and W. B. Croft. Evaluation of an Inference Network-Based Retrieval Model. *ACM Transactions on Information Systems*, 9(3):187–222, 1991. (Cited on page 102.)

[335] J. Unsworth. Scholarly Primitives: what Methods Do Humanities Researchers Have in Common, and how Might our Tools Reflect This. *Symposium on Humanities Computing: Formal Methods, Experimental Practice*, -(-), 2000. (Cited on pages 26 and 79.)

[336] J. Unsworth. Tool-time, or 'haven't we been here already?' Ten years in humanities computing. `http://people.lis.illinois.edu/~unsworth/carnegie-ninch.03.html`, 2003. Online; accessed 28-august-2013. (Cited on pages 2, 22, 23, 26, and 177.)

[337] P. A. Uva. Information-Gathering Habits of Academic Historians: Report of the Pilot Study. Technical report, SUNY Upstate Medical Center, 1977. (Cited on pages 25 and 26.)

[338] P. Vakkari. A Theory of the Task-Based Information Retrieval Process: a Summary and Generalisation of a Longitudinal Study. *Journal of Documentation*, 57(1):44–60, 2001. (Cited on pages 20, 30, and 54.)

[339] B. Van Oers. From Context to Contextualizing. *Learning and Instruction*, 8(6):473–488, 1998. (Cited on page 117.)

[340] A. Vercoustre, J. Pehcevski, and V. Naumovski. Topic Difficulty Prediction in Entity Ranking. In *Proceedings of the 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'08)*, pages 280–291. Springer, 2009. (Cited on pages 110 and 141.)

[341] A.-M. Vercoustre, J. Pehcevski, and J. A. Thom. Using Wikipedia Categories and Links in Entity Ranking. In *Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'07)*, pages 321–335. Springer, 2008. (Cited on page 110.)

[342] A.-M. Vercoustre, J. A. Thom, and J. Pehcevski. Entity Ranking in Wikipedia. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC'08)*, pages 1101–1106, 2008. (Cited on page 110.)

[343] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. Mckeon. Manyeyes: a Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007. (Cited on page 28.)

[344] R. Villa, I. Cantador, H. Joho, and J. Jose. An Aspectual Interface for Supporting Complex Search Tasks. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*, pages 379–386. ACM, 2009. (Cited on pages 35, 79, and 81.)

[345] E. Voorhees. Overview of the TREC 2004 Question Answering Track. In *Proceedings of the Thirteenth Text Retrieval Conference (TREC'04)*. NIST, 2005. (Cited on page 108.)

[346] E. Voorhees and D. K. Harman. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005. (Cited on page 105.)

[347] E. M. Voorhees and C. Buckley. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'02)*, pages 316–323. ACM, 2002. (Cited on page 153.)

[348] E. M. Voorhees and D. Tice. Overview of the TREC-9 Question Answering Track. In *Proceedings of*

*the Nineth Text REtrieval Conference (TREC'00)*. NIST, 2000. (Cited on page 106.)

[349] W. Waller and D. H. Kraft. A Mathematical Model of a Weighted Boolean Retrieval System. *Information Processing & Management*, 15(5):235–245, 1979. (Cited on page 102.)

[350] C. Warwick, J. Rimmer, A. Blandford, J. Gow, and G. Buchanan. Cognitive Economy and Satisficing in Information Seeking: A Longitudinal Study of Undergraduate Information Behavior. *Journal of the American Society for Information Science and Technology*, 60(12):2402–2415, 2009. (Cited on page 70.)

[351] R. Watson-Boone. The Information Needs and Habits of Humanities Scholars. *RQ*, 34(2):203–216, 1994. (Cited on pages 2, 23, and 25.)

[352] W. Weerkamp, J. He, K. Balog, and E. Meij. A Generative Language Modeling Approach for Ranking Entities. In *Proceedings of the 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'08)*, pages 292–299. Springer, 2009. (Cited on page 110.)

[353] M. D. White. The Communications Behavior of Academic Economists in Research Phases. *Library Quarterly*, 45(4):337–354, 1975. (Cited on pages 24 and 26.)

[354] R. White and R. Roth. Exploratory Search: Beyond the Query-Response Paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009. (Cited on pages 33, 34, 59, 82, 83, and 88.)

[355] R. W. White, I. Ruthven, and J. M. Jose. The Use of Implicit Evidence for Relevance Feedback in Web Retrieval. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval*, pages 93–109. Springer, 2002. (Cited on page 61.)

[356] R. W. White, B. Kules, and S. M. Drucker. Supporting Exploratory Search, Introduction, Special Issue, Communications of the ACM. *Communications of the ACM*, 49(4):36–39, 2006. (Cited on pages 29 and 33.)

[357] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the Influence of Domain Expertise on Web Search Behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM'09)*, pages 132–141. ACM, 2009. (Cited on page 31.)

[358] S. E. Wiberley and W. G. Jones. Time and Technology: A Decade-Long Look at Humanists' Use of Electronic Information Technology. *College & Research Libraries*, 61(5):421–431, 2000. (Cited on page 28.)

[359] S. E. Wiberley Jr. *Humanities Literatures and Their Users*. Taylor and Francis, 2009. (Cited on page 23.)

[360] S. E. Wiberley Jr and W. G. Jones. Patterns of Information Seeking in the Humanities. *College & Research Libraries*, 50(6):638–45, 1989. (Cited on page 23.)

[361] M. Wilson and R. White. Evaluating Advanced Search Interfaces Using Established Information-Seeking Models. *Journal of the American Society for Information Science and Technology*, 60(7): 1407–1422, 2009. (Cited on pages 4 and 34.)

[362] T. D. Wilson. On User Studies and Information Needs. *Journal of Documentation*, 37(1):3–15, 1981. (Cited on page 19.)

[363] T. D. Wilson. Information needs and uses: fifty years of progress. In B. Vickery, editor, *Fifty years of information progress: a Journal of Documentation review*, pages 15–51. Aslib, 1994. (Cited on page 20.)

[364] T. D. Wilson. Models in Information Behaviour Research. *Journal of Documentation*, 55(3):249–270, 1999. (Cited on pages 19 and 54.)

[365] T. D. Wilson. Human Information Behavior. *Informing Science*, 3(2):49–56, 2000. (Cited on page 18.)

[366] W. E. Winkler. The State of Record Linkage and Current Research Problems. In *Statistical Research Division, US Census Bureau*, 1999. (Cited on page 111.)

[367] C. Woods. *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond*. Oriental Institute Museum Publications, 2010. (Cited on page 18.)

[368] W.-C. Wu, D. Kelly, A. Edwards, and J. Arguello. Grannies, Tanning Beds, Tattoos and NASCAR: Evaluation of Search Tasks with Varying Levels of Cognitive Complexity. In *Proceedings of the 4th Information Interaction in Context Symposium (IIiX'12)*, pages 254–257. ACM, 2012. (Cited on page 30.)

[369] Y. Wu and H. Kashioka. NiCT at TREC 2009: Employing Three Models for Entity Ranking Track. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC'09)*. NIST, 2009. (Cited on pages 111 and 168.)

[370] R. Yan, B. Huet, and R. Sukthankar. Large-Scale Multimedia Retrieval and Mining. *IEEE Multimedia*, 18(1):2–4, 2011. (Cited on page 118.)

[371] Q. Yang, P. Jiang, C. Zhang, and Z. Niu. Experiments on Related Entity Finding Track at TREC 2009. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC'09)*. NIST, 2009. (Cited on page 111.)

[372] K. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted Metadata for Image Search and Browsing.

In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'03)*, pages 401–408. ACM, 2003. (Cited on page 34.)

[373] C.-M. A. Yeung and A. Jatowt. Studying how the Past is Remembered: Towards Computational History through Large Scale Text Mining. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*, pages 1231–1240. ACM, 2011. (Cited on page 176.)

[374] H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. Ranking Very Many Typed Entities on Wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM'07)*, pages 1015–1018. ACM, 2007. (Cited on page 109.)

[375] C. Zhai. Statistical Language Models for Information Retrieval a Critical Review. *Foundations and Trends in Information Retrieval*, 2(3):137–213, 2008. (Cited on page 137.)

[376] H. Zhai, X. Cheng, J. Guo, H. Xu, and Y. Liu. A Novel Framework for Related Entities Finding: ICTNet at TREC 2009. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC'09)*. NIST, 2009. (Cited on pages 111 and 168.)

[377] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*, pages 81–88. ACM, 2002. (Cited on page 107.)

[378] W. Zheng, S. Gottipati, J. Jiang, and H. Fang. UDEL/SMU at TREC 2009 Entity Track. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC'09)*. NIST, 2009. (Cited on page 111.)

[379] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating Aggregated Search Pages. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*, pages 115–124. ACM, 2012. (Cited on page 33.)

[380] Q. Zhou, C. Wang, M. Xiong, H. Wang, and Y. Yu. Spark: Adapting keyword query to semantic search. In *ISWC/ASWC '07*, pages 694–707, 2007. (Cited on page 114.)

[381] J. Zhu, D. Song, S. M. Rüger, M. Eisenstadt, and E. Motta. The Open University at TREC 2006 Enterprise Track Expert Search Task. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC'06)*. NIST, 2006. (Cited on page 110.)

[382] J. Zhu, D. Song, and S. Rüger. Integrating Document Features for Entity Ranking. In *Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'07)*, pages 336–347. Springer, 2008. (Cited on page 110.)

[383] J. Zhu, X. Huang, D. Song, and S. RÜger. Integrating Multiple Document Features in Language Models for Expert Finding. *Knowledge and Information Systems*, 23(1):29–54, 2009. (Cited on page 108.)

[384] M. M. Zloof. Query by Example. In *Proceedings of the May 19-22, 1975, national computer conference and exposition*, pages 431–438, 1975. (Cited on page 101.)

# Summary

Today's technology enables the continuous production, recording, and storage of all types of digital information. The abundance of information that is therefore available provides researchers with the opportunity to ask new questions but also requires new research methods and appropriate tools. Fields of study differ in the way they have adapted to deal with this flood of data. The natural sciences such as earth sciences and computational biology have readily adopted computationally intensive methods, as they study signals recorded by radars, sensors, or produced by simulations. This type of data lends itself well for analyses through data mining and visualization techniques. In contrast, the traditional objects of study in the humanities have always been analogue records such as books, letters, and photographs. These objects are studied using analytical, critical, and interpretative approaches instead of computational ones. As the introduction of new technology and information sources is changing the way humanities researchers work and the questions they seek to answer, a new challenge arises for the development of tools and algorithms that support new practices as well as traditional ones using new types of information.

Particular challenges for humanities researchers raised by the abundance of available material are to gain insight in which materials to consider for a study through exploration and once chosen to obtain a holistic view of the research topic through contextualization. This thesis investigates two dimensions along which tools to support humanities researchers in dealing with the flood of information may be improved: by providing richer means of interaction with information systems and developing algorithms that allow discovery of information through relations between concepts. One of the findings, along the interaction dimension, is that for a particular group of humanities researchers the ability to make comparisons between alternative search results leads to further exploration of the material available. Additional results, along the concepts dimension, include algorithms that support identifying relations between concepts based on structured and unstructured data. Tools incorporating these algorithms allow humanities researchers to identify additional concepts related to the ones already identified as relevant to their research topic.

The results in this thesis show how both richer interactions and more effective related concept finding algorithms may be used to improve tools to support the research practices of humanities researchers. The insights from the work in this thesis may be used to inform the design and evaluation of future tools to continually support new needs and developments in the humanities.

# Samenvatting

De huidige technologie maakt de continue productie, opname, en opslag van allerlei soorten data mogelijk. De overvloed aan informatie die hierdoor beschikbaar is, verschaft onderzoekers de mogelijkheid om nieuwe onderzoeksvragen te stellen maar maakt ook nieuwe onderzoeksmethoden en digitaal gereedschap noodzakelijk. Onderzoeksgebieden verschillen in hoe ze zich hebben aangepast om om te kunnen gaan met deze vloedgolf aan informatie. De natuurwetenschappen, zoals aardwetenschappen en computationele biologie, hebben computationeel intensieve methoden snel overgenomen, aangezien zij signalen bestuderen zoals opgenomen door radars, sensoren, of voortgekomen uit simulaties. Dit soort data leent zich goed voor analyses door middel van data mining en visualisatie technieken.

In contrast hiermee zijn de traditionele objecten bestudeerd door de geesteswetenschappen altijd analoge documenten geweest zoals boeken, brieven en foto's. Deze objecten worden bestudeerd door middel van analytische, kritische en interpretatieve methoden in plaats van computationele methoden. Aangezien de introductie van nieuwe technologie en informatiebronnen de manier waarop onderzoekers in de geesteswetenschappen werken en de vragen die zij stellen verandert, ontstaat een nieuwe uitdaging in het ontwikkelen van digitaal gereedschap en algoritmen die zowel nieuwe als ook de bestaande onderzoeksgebruiken toegepast op nieuwe soorten informatie ondersteunen.

De voornaamste uitdagingen voor geesteswetenschappers veroorzaakt door de overvloed aan beschikbaar materiaal zijn het verkrijgen van inzicht in welk materiaal te gebruiken voor een onderzoek door middel van exploratie en wanneer eenmaal een onderwerp is gekozen om een algeheel overzicht te krijgen van het onderzoeksonderwerp door middel van contextualisatie. Dit proefschrift onderzoekt twee dimensies langs welke digitaal gereedschap voor het ondersteunen van onderzoekers in de geesteswetenschappen in het omgaan met de overvloed aan materiaal verbeterd kunnen worden: door het verschaffen van rijkere manieren van interactie met informatie systemen en het ontwikkelen van algoritmen die het mogelijk maken informatie te ontdekken door middel van relaties tussen concepten. Eén van de bevindingen, met betrekking tot interactie, is dat voor een bepaalde groep onderzoekers in de geesteswetenschappen de mogelijkheid tot het maken van vergelijkingen tussen alternatieve zoekakties leidt tot verdere exploratie van het beschikbare materiaal. Additionele resultaten, met betrekking tot concepten, zijn ondermeer algoritmen die het vinden van relaties tussen concepten ondersteunen op basis van gestructureerde en ongestructureerde data. Digitaal gereedschap op basis van deze algoritmen stelt geesteswetenschappers in staat om additionele concepten te identificeren, gerelateerd aan de concepten die al als relevant aan het onderzoeksonderwerp bekend zijn.

De resultaten in dit proefschrift laten zien hoe rijkere interactie en meer effectieve algoritmen voor het vinden van gerelateerde concepten gebruikt kunnen worden voor het verbeteren van het digitale gereedschap dat de onderzoeksgebruiken van geesteswetenschappers ondersteunt. De inzichten van het werk in dit proefschrift leveren verder een leidraad voor het ontwikkelen en evalueren van nieuw digitaal gereedschap dat ondersteuning biedt aan nieuwe gebruiken en ontwikkelingen binnen de geesteswetenschappen.