Time-Aware Exploratory Search: Exploring Word Meaning through Time

Daan Odijk
ISLA, University of Amsterdam
d.odijk@uva.nl

Giuseppe Santucci
Sapienza, University of Rome
santucci@dis.uniroma1.it

Maarten de Rijke
ISLA, University of Amsterdam
derijke@uva.nl

Marco Angelini
Sapienza, University of Rome
angelini@dis.uniroma1.it

Guido Lorenzo Granato
Sapienza, University of Rome
granato@dis.uniroma1.it

ABSTRACT

With more longitudinal archives becoming digitized and publicly available, new uses emerge. Collections that span centuries call for a time-aware exploration approach, a coordinated environment supporting understanding the development of word usage and meaning through time, with the means to leverage this for exploration. We present ongoing work on a coordinated time-aware exploratory search approach and present a case study on a prototype system. With this approach, a user is able to gain a deeper understanding of the relevant parts of the collection.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.5.2 [**User Interfaces**]: Evaluation/methodology

Keywords

Exploratory search, time-aware, temporal IR, statistical semantics

1. INTRODUCTION

With more longitudinal archives becoming digitized and publicly available, new uses emerge for these collections. Language and culture are inherently dynamic phenomena. Scholars in the humanities—linguists, literary scholars, historians—try to understand processes of change and variation in these phenomena and to uncover their internal and external causes. The implicit assumption is that the meaning of a word can be inferred from observing its usage, i.e., its distribution in text, and changes to its usage. Longitudinal collections of documents that span multiple decades or even centuries call for a time-aware exploration approach.

In [13] this problem was addressed using an exploratory search interface, featuring visualizing of the temporal distribution of documents and word clouds for getting a quick insight through summarization. This approach was effective for searching specific word usage change and indicated that there is valuable insight to be gained from observing changes in word usage and meaning over time.

One problem, with the interface in [13] is that the user had to go out and look for the usage of a specific word at a specific point in time. What is needed in support of time-aware exploratory search is a coordinated environment in which a user is supported in understanding the development of word usage and meaning through time and is able to leverage this understanding in finding interesting periods and interesting bits of information. In particular, based on the experience in [13] we propose to bring together Information Retrieval (IR) and Visual Analytics (VA) to support the exploration of longitudinal document collections. Specifically, what we propose in this paper is an approach to provide a user with information on volume and correlation of words and documents across time.

The aim of this paper is to describe our ongoing work on this coordinated time-aware approach to exploratory search. With this we aim to further facilitate insight into temporal aspects of document collections. Exposing and exploiting temporal aspects of a collection has been identified as one of the key open challenges in recent IR research, specifically in exploratory search systems [2].

In Section 2 we discuss related work in both IR and VA. In Section 3 we detail our approach to time-aware exploratory search. Section 4 is devoted to a worked example. In Section 5 we conclude, step back and discuss our findings so far.

2. BACKGROUND

Exploratory search is a form of information retrieval where users start off without a clear information need. They do not know beforehand what they are looking for, nor where to find it [11]. Exploratory search systems therefore try to interactively and iteratively guide the user to interesting parts of the collection. Exploratory search interfaces often try to provide a quick overview, while allowing users to quickly zoom into details. Providing this quick overview can be done by visualizing the information that was retrieved [5, 6, 13].

In [13] an exploratory search system was used for the specific task of selecting a subset of documents for historical research. For this task, the most common approach was to manually sample documents. The proposed method of semantic document selection allows a historian to interactively select a subset. Here, time was an important dimension, but the focus was only on volume of documents. In contrast, in this paper we propose an combined approach that also considers word usage and correlations for words and documents.

Alonso et al. [2] review current research trends in temporal IR and identify open problems and applications. Exploratory search is one of these applications, with exposing and presenting temporal information as open problems. In time-aware exploratory search, as opposed to time-unaware exploratory search, there is a clear emphasis on the evolution, development and changes of documents, topics and word usage. In collections spanning long periods, time can be an important retrieval cue and insight into word usage is an important aspect to understanding the collection that is explored.

Understanding word meaning by studying how words are being used and how their usage changes has a tradition that goes back at least half a century. The field of statistical semantics focusses on the statistical patterns of words and how they are used [16]. The under-

TAIA'12, August 16, 2012, Portland.
ACM xxx-x-xxxxx-xxx-x/xx/xx.

lying assumption is that "a word is characterized by the company it keeps" [8]. Recent innovations have allowed statistical semantics methods to be applied to larger and larger datasets. Michel et al. [12] have used these methods on millions of digitized books, ∼4% of all books ever published, to observe cultural trends. Odijk et al. [13] describe initial steps towards a similar approach to providing access to 100 million news articles spanning multiple centuries.

In time-aware exploratory search, Information Visualization (IV) is an important aspect and statistical semantics methods lend themselves well for this. Shneiderman [14] identified temporal data as a basic data types most relevant for IV. Visualizing time-oriented data has been extensively studied [1]. Our aim is to combine IV and exploratory search facilities so as to arrive at a visual exploratory search system. To this end, we build on VA [9], an emerging multidisciplinary area taking into account analytical reasoning and IV techniques, combining the strengths of human and electronic data processing. Visualization becomes the medium of a semi-automated analytical process, where humans and machines cooperate for the most effective results. Decisions on the direction analysis takes to accomplish a task are left to the user. Combining these strengths, VA can attack problems whose size and complexity make them too difficult to solve otherwise. Although IV techniques have been extensively explored [1, 7, 14], combining them with automated data analysis for specific application domains is still a challenge [10].

The challenges in time-aware exploratory search call for an approach that combines IR, VA and IV techniques. In this approach, the analysis is mostly done by a search engine, with the user steering the system into finding documents deemed worthy of further inspection. To this end the user is supported by visualizations on changing word usage and meaning over time. This provides a user with insight on the development of both documents and words through time.

3. TIME-AWARE EXPLORATORY SEARCH

To support time-aware exploratory search, we propose an interactive search system with an interface that combines faceted search techniques and interactive visualizations in which time is a first-class citizen. These visualizations use methods from statistical semantics and are presented alongside the search results and facets.[1]

A sketch of the exploratory search interface is shown in Figure 1. A user formulates a search query, that can be as simple as a single word, as complex as the search engine allows it to be or even encompass the entire collection. On issuing a query, the search results form a document set, for which multiple coordinated views are presented. Each view shows a different perspective on the document set, providing different insight and different entries into the data.

We present three types of views onto the document set. Probably the most familiar view is a ranked list of documents, represented by snippets and some metadata. This metadata is also presented in the second type of view, as a faceted view on the document set and summarizing the values for the range of values for that facet. Time is represented as a separate facet, to allow for ways of presentation and means of interaction that take advantage of its inherent semantic structure. Finally, the visualizations of the document set form the last type of view. Before going into the details of the visualizations, we will first describe the interaction between the interface components.

Coordinating the multiple views ensures that the user is able to maintain a sense of context and allows for insight to arise from combining multiple dimension. The views operate on different primitives, of which time, words and documents are the most important. We identify two distinct actions that can be performed on these primitives: select and highlight. Selecting means we want to filter the selection of documents to only show those selected. Highlighting

[1]The interactive system is implemented in a web-based interface, using d3 [4] for visualizations and ElasticSearch (built on top of Apache Lucene).

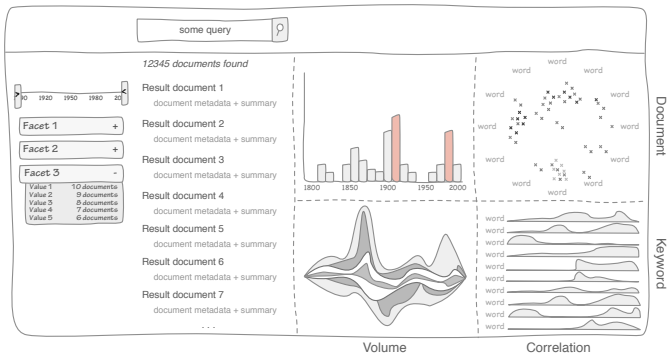


Figure 1: Sketch of the exploratory search interface, showing the faceted search part (left) and four visualization (right).

will focus on a document subset and give less focus or hide the rest. Coordinating these actions means that they are reflected in all views.

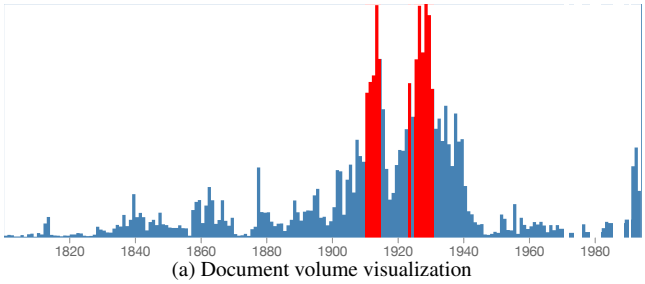
The interface we propose features four visualizations that provide insight into the result set for the query. These visualizations can be grouped along two dimensions. They are either document-centric or keyword-centric and they present information on either volume or correlation. Time is the central concept in these visualizations, represented as an axis in all, except for the document correlation visualization, where interaction allows for inspection of temporal patterns. The temporal axis is represented with the same scale in all visualizations. In the keyword-centered and document correlation visualization, a set of keywords is used. These can be chosen by the user or can be based on the most frequently occurring words in the result set. Each keyword is color-coded consistently in all visualizations it occurs, to establish the relations between all visualizations.

Time slices can be selected or highlighted in all visualizations where it is represented, i.e. document volume and both keyword-centric visualizations. Selecting a time slice filters the document set to documents from a specific time. Highlighting a time slice, will highlight this time slice and the documents from that time, in all other views. This gives a context for comparing different perspective on the documents from a specific time. Similarly, keywords can be selected or highlighted in the three graphs where they are represented, i.e., the keyword-centric and the document correlation visualizations. Selecting a keyword will filter out all documents from the document set that do not contain that specific word.

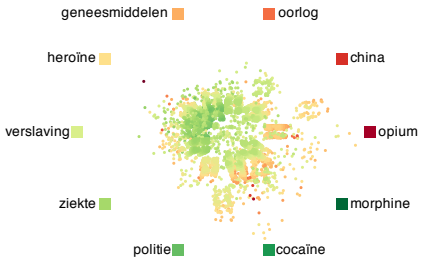
Selecting or highlighting time slices or keywords will implicitly select a subset of documents. In the document correlation view, individual documents are represented and a subset can be highlighted by brushing, i.e., drawing one or more areas around them. Coordination ensures that interactions with one view are propagated to all other views in order to maintain consistency among the views and to give the user a sense of context.

Document Volume To present information on the volume of documents, we use a histogram, showing the temporal distribution of the search results (see Figure 2a). In the histogram, bars get a different color if the volume of documents in a period is substantially higher than average, drawing the user's attention to possibly interesting periods. This is an interactive graph, allowing for semantic zooming (when zooming to a shorter time scale, granularity of axis switches and the data will show more detail, e.g. from year to month, allowing for detailed inspection).

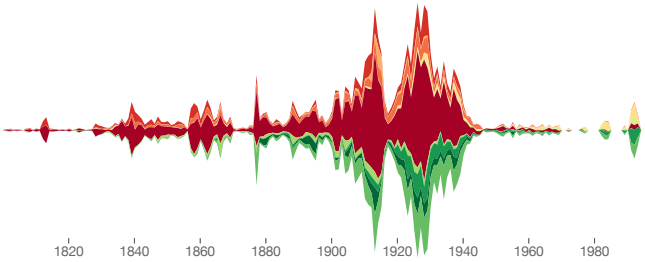
Keyword Volume To visualize changes in word usage over time, we visualize the volume of the keywords as a stream (see Figure 2c). This visualization shows patterns that are similar to the document volume visualization, as the underlying patterns for the increase in volume are the same. The insight that can be gained from the keyword volume visualization is clearly different. The volume of



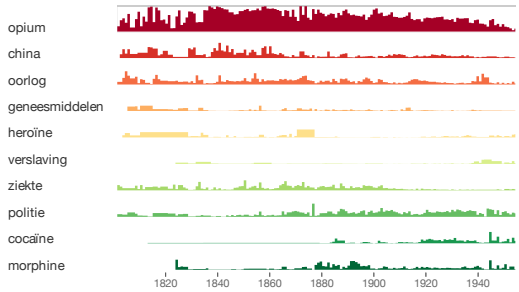
(a) Document volume visualization



(b) Document correlation visualization



(c) Keyword volume visualization



(d) Keyword correlation visualization

Figure 2: Screenshots of the exploratory search interface, showing the graphs for a query on drugs related terms. The keywords shown in the interface are topic-specific and can be translated as: *opium, China, war, medicines, heroine, addiction, disease, police, cocaine, morphine*.

word usage can be measured as the number of occurrences or documents containing the keyword. Furthermore, the volume of word usage can be measured in the entire collection or in the document selection that results from the query. The best choice depends on the application. If a user is specifically looking for changing word usage, the number of occurrences in the entire collection might be the most informative choice. In our exploratory search scenario, we are trying to guide the user to an interesting part of the collection. With a clear focus on the selection of documents in this scenario, we choose to show the volume of documents containing the keyword in the search results. In this way, the information that is presented also remains consistent with the other visualization.

Document Correlation For visualizing the correlation between documents, we use a visualization based on radial distance, namely RadViz [3] (see Figure 2b). We organize a number of words on a circle, with equal distances between them. A document is represented as a point inside the circle, with each word pulling the document close to it. The force each word puts on a document is based on the relatedness between the document and the word. The relatedness is taken to be the relevance assigned by the search engine to the document using the word as a query.

As we have opposite forces that might attract a document, we need a way to differentiate between documents that are not attracted and documents that are attracted strongly, but equally strong from opposite sides. We use color coding with green representing documents that suffer a low force, via yellow to red for high force.

This visualization is focused on relations among documents and words. The purpose of this visualization is to characterize the documents set using a subset of keywords and to identify clusters of similar documents. Advantages of this visualization are the scalability in respect of the number of keywords, the absence of cluttering of the screen and the ability to add effective interaction techniques.

In the document correlation visualization, there is no axis representing time. If no time slice is highlighted, the correlation for all documents is presented. By highlighting a time slice in another view, a user can inspect the document correlation in a specific period.

Keyword Correlation The correlation between the query and specific keywords is shown in a series of simple visualizations, organized as small multiples [15], allowing for easy comparison (see Figure 2d). A series of bars shows for each word and through time the correlation with the query, encoded as the height of the bar. A full bar means every document in the search results for that period contains that specific word and no bar means there are no documents containing that word in the period.

4. A WORKED EXAMPLE

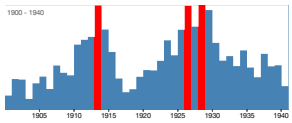
A full evaluation of our time-aware exploratory search approach is part of ongoing work. In this section we report on one of several case studies with individual historians. Our subject is a senior historian, studying the public opinion on drugs as represented in newspaper, with a specific interest in the early 20th century. We use a collection of ∼30 million newspaper articles from the Koninklijke Bibliotheek, the National library of the Netherlands.

In this example, we use a query created by our subject using a topic lexicon with drug related terms, in a process called semantic document selection [13]. This results in a document set of 32,921 documents, spanning almost two centuries (1800–1994). We have used this document set to create the screenshots shown in Figure 2.

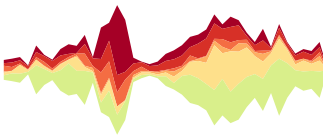
Looking at the temporal distribution of documents for the document set (Figure 2a), there are clear peaks in document volume. The historian selects the time period 1900–1940, resulting in the document volume shown in Figure 3a. With his domain knowledge he is able to connect these peaks to key events, such as the Opium Treaty from Shanghai (1912), the introduction of Dutch Opium laws (1920) and the tightening thereof (1928).

To gain a better understanding of the aspects associated with drugs, the historian looks at the terms associated with drugs over time, by examining the keyword volume. A comparison of the volume of two sets of words is shown in Figure 3b. In this visualization he can notice a shift from health issues (*medications, science, pharmacy*) to crime related issues (*addiction, trafficking, arrested*), after the Dutch Opium laws came into effect. Inspecting the keyword correlation for specific words, while highlighting key events, he can conclude that before the Dutch Opium laws came into effect the term *chloroform* was dominantly used; afterwards, the terms *opium*, *heroïne*, and *cocaïne* (shown in Figure 3c) are more prominent.

Based on his knowledge that the Netherlands had a monopoly on opium in their colony, the Dutch East Indies (now Indonesia), the historian select all documents that were published in the capital, Batavia (now Jakarta). Inspecting the document correlation, a distinction is apparent for the *medicine* versus *China* and *Chinese*.

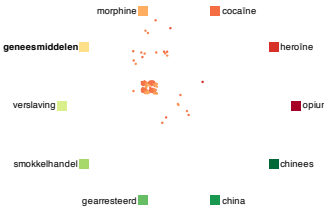


(a) Document volume for the drugs-related query, with peaks indicated at the years 1913, 1926 and 1928.

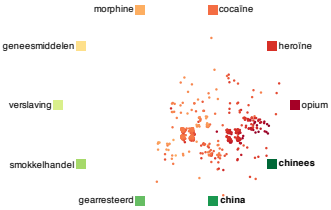


(b) Keyword volume for the translated keywords: *medicine, science, pharmacy, addiction, trafficking, arrested* (top to bottom).

(c) Keyword correlation for the drug *cocaine* and the query on drugs.



(d) Document correlation, with the word *medicine* highlighted.



(e) Document correlation, with words *China*, *Chinese* highlighted.

Figure 3: Screenshots showing specific visualizations for a query on drugs in the timespan 1900–1940.

For *medicine* (Figure 3d), the bulk of the documents are associated with morphine and almost none with the other keywords. The keywords *China* and *Chinese* (Figure 3e) are associated with *opium* and with crime related words, such as *arrested*. This is in line with the general image in the colony at that time: opium was mostly associated with Chinese drug users and crime syndicates. On top of this, our historian is able to select a location of publication, a subset of newspapers that represent a certain religion or the articles that appeared on the front page of a newspaper.

5. CONCLUSION AND DISCUSSION

The examples described in Section 4 show that using our coordinated time-aware exploratory search approach, a historian is able to infer high level knowledge from the patterns in volume and correlation of both words and documents. With this understanding of the development of word usage and meaning throughout time, the user is able to find interesting periods and bits of information, further supported by rich metadata facets. By coordinating the multiple views, a user maintains a sense of context and can find information using a multitude of search strategies. With this coordinated time-aware exploratory search approach, a user is able to gain a deeper understanding of the relevant parts of the collection.

The case study described is part of ongoing work to evaluate our time-aware exploratory search approach, with a full evaluation still work in progress. We foresee several approaches for this evaluation. For example, comparing this interface to another interface that uses only a search box and a result list, is not the right approach, as these interfaces have a very different purpose. In this evaluation, efficiency is not very important. In fact, users will probably not find things quicker, as the interface allows for casual browsing and exploring of a collection and is aimed at getting a deeper understanding of the collection. Evaluation should therefore focus on the quality of the insight that can be gained from using it.

For evaluating the effectiveness of our approach, we can assess the quality of the insight gained from exploring the collection. For example, Bron et al. [6] compared the quality of research questions obtained using two variants of a exploratory search tool. For document selection [13], the quality of the document selection can be evaluated. We can measure the quality of this selection in terms of diversity, both temporal and non-temporal. One way of measuring this non-temporal diversity is to look at the framing of an issue, i.e., the ways in which issues are presented. A more diverse selection of documents on a issue will have a broader range of perspectives on it. Another important aspect in time-aware exploratory search is to identify turning points in time, e.g. the moments words change meaning or association. An evaluation of our approach could evaluate how well different variants of the system are able to identify these turning points.

The case study uses an unique diachronic collection of digitized newspapers with proper document dating and rich metadata. Yet, there is nothing in our approach that is collection-specific. We use two dimensions present in nearly every collections: time and words. One can easily find examples in smaller timespans, e.g., investigating changing reputation of a company on Twitter, or finding when new words appear in a language, by analyzing books and news.

Acknowledgments This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE) and 288024 (LiMoSINe), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, 727.011.-005, 612.001.116, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP and BILAND projects funded by the CLARIN-nl program, the Dutch national program COMMIT, and by the ESF Research Network Program ELIAS.

REFERENCES

[1] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visualizing time-oriented data. *Computers & Graphics*, 31(3): 401–409, 2007.
[2] O. Alonso, J. Strötgen, R. Baeza-Yates, and M. Gertz. Temporal information retrieval: Challenges and opportunities. In *TWAW Workshop, WWW*, pages 1–8. Citeseer, 2011.
[3] M. Ankerst, D. Keim, and H. Kriegel. *Circle segments: A technique for visually exploring large multidimensional data sets*. Bibliothek der Universität Konstanz, 1996.
[4] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12): 2301–2309, 2011.
[5] M. Bron, J. van Gorp, F. Nack, and M. de Rijke. Exploratory search in an audio-visual archive. In *EuroHCIR '11*, 2011.
[6] 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 *SIGIR '12*. ACM, 2012.
[7] S. K. Card and J. Mackinlay. The structure of the information visualization design space. In *Proceedings of InfoVis '97*, pages 92–99. IEEE Computer Society, 1997.
[8] J. Firth. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pages 1–32. Oxford: Philological Society, 1957.
[9] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Visual analytics: Definition, process, and challenges. In *Information Visualization*, pages 154–175. Springer-Verlag, Berlin, Heidelberg, 2008.
[10] D. Keim, J. Kohlhammer, G. Santucci, F. Mansmann, F. Wanner, and M. Schäfer. Visual analytics challenges. In *Proceedings of the eChallenges 2009*, 2009.
[11] G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.
[12] J. Michel, Y. Shen, A. Aiden, A. Veres, M. Gray, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, and Others. Quantitative analysis of culture using millions of digitized books. *Science*, 331 (6014):176, 2011.
[13] D. Odijk, O. de Rooij, M.-H. Peetz, T. Pieters, M. de Rijke, and S. Snelders. Semantic Document Selection. In *TPDL 2012: Theory and Practice of Digital Libraries*. Springer, 2012.
[14] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343. IEEE Computer Society, 1996.
[15] E. Tufte. *The visual display of quantitative information*, volume 7. Graphics Press Cheshire, CT, 1983.
[16] W. Weaver. Translation. *Machine Translation of Languages*, 14: 15–23, 1955.