Exploratory Search in Wikipedia

Sisay Fissaha
ISLA, Informatics Institute
University of Amsterdam
sfissaha@science.uva.nl

Maarten de Rijke
ISLA, Informatics Institute
University of Amsterdam
mdr@science.uva.nl

ABSTRACT

We motivate the need for studying the search, discovery and retrieval requirements of Wikipedia users. Based on a sample from an experimental Wikipedia search engine, we hypothesize that the fraction of Wikipedia searches that are exploratory in nature is at least the same as that of general web searches. We also describe a questionnaire for eliciting search, discovery and retrieval requirements from Wikipedia users.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms

Design, Experimentation, Human Factors

Keywords

Wikipedia, interfaces, exploratory search

1. INTRODUCTION

In this paper, we describe work in progress that is aimed at eliciting the user requirements of Wikipedia users, with a special focus on so-called undirected (or exploratory) queries. Understanding what users search for, and how their information needs can best be met, is of increasing interest, both for the scientific community and for society at large, especially where it concerns valuable, and increasingly popular resources such as Wikipedia. Increasingly, the IR research community seeks to understand *why* users conduct searches since it is believed that knowing the user intentions or goals may help tailor output of search engines to the needs of a particular user. As a result, understanding what users search for and why users conduct searches, which we refer to as *user requirements*, has become an active area of research.

Broder [2] and Rose and Levinson [7] identified the following broad classes of goals of web users: navigational and informational. Informational goals are further classified as directed, undirected, locate, advice, and list request. Directed requests may be open or close ended. Users issuing directional queries are typically looking for a focused answers whereas those issuing an undirected informational query about X want to learn something/everything about X. As the name implies, locate, advice and list requests are specialized requests that attempt to locate, get advice or list of items. Rose and Levinson [7] showed that over 60% of web search queries are informational in nature, and so-called undirected (or exploratory) queries form a significant subclass of these.

SIGIR '06 Seattle, Washington USA

Do the above findings carry over from the general web to Wikipedia? For instance, are there navigational goals in Wikipedia searches? Is the distribution of directed and undirected queries different in the context of Wikipedia? We hypothesize that users of encyclopedias—electronic or otherwise—have relatively uniform information needs which may largely be characterized as *informational.* Is this correct?

Studies aimed at answering such questions range from developing generalized models, which identify the different factors that affect the information search behaviour, to devising specific strategies for collecting empirical data required for testing the different hypotheses [6, 4]. Most studies adopt different strategies such as user studies or query log analysis to identify typical user search requirements and behaviours. Recently, Pharo and Järvelin [6] indicated that such strategies are limited in their ability to provide the necessary data for studying the different factors and their relationships, and proposed an extensive data collection and analysis methods. It is, however, a very time consuming and expensive approach and, hence, it may not be feasible to conduct relatively large surveys. As a result, we adopt commonly used methods of data collections for the current study.

Specifically, we use a two-fold approach towards answering the above questions. First, we create pilot applications that implement novel ways of accessing the information provided by Wikipedia; see, e.g., [3] (link suggestions), [8] (focused access at the sub-document level), [11] (exploratory question answering using Wikipedia), and try to mine useful information from the query logs. Second, we are in the process of setting up an online questionnaire aimed at eliciting the requirements of Wikipedia users.

The remainder of the paper is organized as follows. In Section 2 we provide background on Wikipedia and on accessing Wikipedia. Then, in Section 3 we examine a sample from a Wikipedia search engine query log. After that we describe the design of our user survey, and we conclude in Section 5.

2. SEARCH, DISCOVERY AND RETRIEVAL IN WIKIPEDIA

2.1 About Wikipedia

Wikipedia posseses a number of special and useful characteristics which call for possibly different access methods and which make investigation of the information access problem especially challenging and interesting. Among these are:

- Wikipedia is the result of a collaborative content development effort regulated by group concensus, without strict guidelines and control. Therefore, individuals may want to have a flexible browsing and search facility which will enable them to have a global view of Wikipedia while editing locally.
- Wikipedia is an encyclopedia, hence contains different types of information which may call for different modes of access. Access to geographic information, for example, can be enhanced with a map-based explorative search interface.
- Wikipedia's content consists of both structured and unstructured textual data, and it also other data types such as images, video. This in turn provides a useful experimental setup to apply the methods developed for different data types.
- Wikipedia's content can be edited by anyone. The same user may be a reader with a specific information need, or an author who wants to create an entry. The only source of information for anyone who wants to create a page is the general guidelines provided in Wikipedia website. Authors are not normally trained. It should be possible for individuals to learn more about Wikipedia in the process of using it or contributing to it which in turn may call for a more explorative search interface.

2.2 Access to Wikipedia

Traditionally, access to (paper-based) reference works such as encyclopedias has relied on alphabetic listings of the titles of the entries, on cross-references, and on multiple indexes. Many of these strategies seem to have been carried over to their online counterparts—Wikipedia is no exception. Today, Wikipedia has become one of the primary reference sites; its main site (`http://wikipedia.org`) consistently ranks amongst the top 50 sites in terms of traffic [1]. Wikipedia's increasing popularity has gone hand in hand with its growth in size, which has been steady and exponential, going from 0 articles in 2001 to over 1,000,000 (for the English part alone) during the first half of 2006 [10]. This growth in size and popularity call for effective support methods; as the distinction between reader and author in the Wikipedia context is being blurred [5], such support methods are needed for both readers who want to locate information in Wikipedia and for authors who want to contribute to the growing number of stubs and articles.

Currently, Wikipedia provides a keyword-word based search facility which allows users to enter a set of keywords and get a ranked list of Wikipedia pages. There are also other search engines that provide efficient and focused access to the Wikipedia content though the basic search facilities remain more or less the same. In addition, Wikipedia has dense networks of hyperlinks which eases browsing through the content. Furthermore, each page is assigned to categories or lists that groups pages into some kind of semantic classes. These features allow for extra browsing and navigation facilities.

Though the facilities sketched above—and other ones not listed, such as the Wikipedia categories—ease the burden of searching and browsing through the Wikipedia content, we believe that the nature of Wikipedia and its development process may require more advanced search facilities that go beyond what is currently available. For example, one peculiar property of Wikipedia is that the distinction between readers and authors is blurred—for both it may be useful to be able to generate templates with slots for prototypical facts about entities falling within a particular category, and for authors/editors it may be useful to be able to automatically check for structural consistency such as link structure.

Results of a recent Wikipedia-based study [8] call for a proper investigation of the requirements of the new generation of users in order to better meet their information needs. Sigurbjörnsson et al. [8] conducted an experiment on the advantage of providing focused access (i.e., direct access to the sections: "go and read here") to the content of Wikipedia over a full-document retrieval baseline. The result showed that "focused access allows users to solve their search task quicker, at least when the information need is specific." But what if the information is not specific, and users need to *explore* Wikipedia's content, because they need to "find out" about a topic? In the following section we look at a sample from a Wikipedia search engine log file—the sample suggests that many users have such undirected information needs.

3. A WIKIPEDIA SEARCH ENGINE QUERY LOG

We analysed a random sample of 200 queries that are taken from an experimental Wikipedia search engine [9] that is publicly available. As we only have access to the query log—and not to the users submitting the queries—it is difficult to carry out a detailed classification of the user intention or goals. Hence, we could not use the classification used in [7]. For our classification, we adopted three classes: directed informational goal ("I want to learn something specific about my topic"; D), undirected or exploratory informational goal ("tell me about my topic"; X), and unknown (ones that we were unable to classify; UN).

For directed goals, we checked for the presence of the following properties in the queries:

- is the query in the form of a factoid question
- does the question have the following form: capital Iceland, population Netherland, inventor computer, Woody Allen married etc.—in short, queries that can typi-

cally be answered in terms of a specific named entity or clause.

For undirected informational goals, we checked the following properties in the queries:

- is the query a well-formed phrase and does it represent a well-defined concept or entity?
- does the query match the title of a Wikipedia page?
- does the query represent a person or name of a place or location?

All queries not covered by the above conditions were classified as unknown. The results of classifying a random sample of 200 queries from the log are given in Table 1. A significant portion of the queries are undirectional or exploratory queries. Of the 145 undirectional queries, 47 (32%) have a Wikipedia entry. This preliminary result shows that undirected search queries seem to be the dominant type of queries, as with general web queries. Since the above analysis is very limited and far from accurate, we plan to carry out a survey, to verify results obtained from the search engine query—setting up this survey is the topic of the next section.

Query Types	Frequency	
Directed	8	(4%)
Undirected	145	(72.5%)
Unknown	47	(23.5%)

Table 1: Classification results of the queries

4. SOLICITING REQUIREMENTS

We plan to extend our experimental Wikipedia search engine with a questionnaire. The questions are organized into four categories.

4.1 General

The first set of questions are used to identifying the type of user (reader or author or both), the frequency of access to Wikipedia, and purpose of use.

- How often do you use Wikipedia?
- Do you edit Wikipedia articles?
- What do you use Wikipedia for?

4.2 Type of Information

The next set of questions attempt to identify what sort of information people are looking for. This may roughly map to the user goals or intentions enumerated in the introduction, such as directed, undirected, etc.

- Did you have a specific question in mind?
- Would you prefer to express your queries in terms of natural language?
- Given the three types of information needs illustrated by the following questions, which one illustrates your information needs best?
  - I want to know the capital city of Kenya. I want to know who invented the telephone.
  - I want to know how to make pasta. I want to know why bears hibernate.
  - I want to know something about lung cancer. I want to know who Micheal Jackson is.
  - I want the list of countries in Latin America. I want the names of some programming languages.
  - If none of the above, could you formulate a question expressing your information need?

4.3 Author-related Questions

Unlike the previous set of questions, which are targeted more to the readers of Wikipedia, the following set of questions is geared more to the requirements of authors of Wikipedia. Though the distinctions between the two may be blurred or non-existent from the user's perspective, they may still pose different requirements when viewed from the system development perspective.

- I want to be able to automatically create hyperlinks.
- I want to be able to check the consistency of the hyperlink structure.
- I want to validate my sentences against snippets extracted from the web.
- I want to get automatic update support for a page's content.
- Could you specify other information needs that you may think will fall under this category?

4.4 "Professional" users

So far the focus has been on the readers or authors of Wikipedia. As the size and popularity of Wikipedia increases, the intended use of it also varies a lot. Recently, its content has also become target of scientific enquiries. Though it might be very hard to characterize the type of users under this category, it might still be useful to include some possible questions in the survey.

- I want to retrieve sentences with a particular named-entities or describing a particular events.
- I want statistical summaries of the corpus. What sort of statistical summary do you need?
- I want to visualize Wikipedia content. What should the visualization should include?
- Do you have any particular requirements?

5. CONCLUSIONS

In this paper we motivated the need for studying the requirements of Wikipedia users. We hypothesized that the fraction of Wikipedia searches that are exploratory in nature is at least the same as that of general web searches. We described a questionnaire for eliciting search, discovery and retrieval requirements from Wikipedia users. We expect to be able to report on initial results from our questionnaire at the time of the EESS workshop.

6. ACKNOWLEDGMENTS

Sisay Fissaha Adafre was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 220-80-001. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.-065.-120, 612-13-001, 612.000.106, 612.066.-302, 612.069.006, 640.001.501, and 640.002.501.

7. REFERENCES

[1] Alexa, 2005. Traffic ranking for Wikipedia. URL: `http://www.alexa.com/data/details/?url=wikipedia.org`, accessed October 2005.

[2] A. Broder. A taxonomy of web search. In *SIGIR Forum*, pages 3–10, 2002.

[3] S. Fissaha Adafre and M. de Rijke. Discovering missing links in Wikipedia. In *Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005)*, 2005.

[4] P. Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1):3–50, 1996.

[5] N. Miller. Wikipedia and the disappearing "Author". *ETC: A Review of General Semantics*, 62(1):37–40, 2005.

[6] N. Pharo and K. Järvelin. The SST method: a tool for analysing web information search processes. *Information Processing & Management*, 40(4):633–654, 2004.

[7] D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th intern. conf. on World Wide Web*, pages 13–19, New York, NY, USA, 2004. ACM Press.

[8] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Focused access to wikipedia. In F. de Jong and W. Kraaij, editors, *6th Dutch-Belgian Information Retrieval Workshop (DIR 2006)*, pages 73–80, 2006.

[9] Wikiii, 2006. Wikiii: A focused search engine for Wikipedia. URL: `http://berk.science.uva.nl:8080/wikiii/`, accessed May 2006.

[10] Wikipedia, 2005. Size of Wikipedia. URL: `http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia`, accessed October 2005.

[11] WiQA, 2006. Question Answering Using Wikipedia. URL: `http://ilps.science.uva.nl/WiQA/`, accessed May 2006.