Searching Wikipedia: learning the why, the how, and the role played by emotion

Hanna Knäusl
Department of Information Science
University of Regensburg
93040 Regensburg
hanna.knaeusl@sprachlit.uni-regensburg.de

ABSTRACT

Searching Wikipedia has been the focus of study for an increasing number of information retrieval publications. In recent years different IR tasks have used Wikipedia as a basis for evaluating algorithms and interfaces for various types of search tasks, including Question Answering, Exploratory Search, Entity Search and Structured Document retrieval. Despite being associated with these well-defined task types, little is known about why people *actually* search wikipedia, what they try to find, how and why they try to find it or the criteria they use to define success. We argue that the way wikipedia content is generated influences the way it is used, including search behaviour. We are particularly interested in learning about affective aspects of search, which have been suggested to be an important motivating factor in wikipedia search behaviour, particularly in leisure scenarios. In this position paper we motivate the investigation of wikipedia search behaviour in the wild and present our ideas on the best way to study this behaviour.

1. INTRODUCTION AND MOTIVATION

Wikipedia[1] is a free online encyclopedia, which due to its open source design and community-based editing policy has become one of the largest reference works of all time. The large volume of information, the breadth of topics covered and open-access nature of the collection has made Wikipedia a natural target of study within the Information Retrieval research community. Wikipedia is now used as the document collection for several retrieval evaluation efforts at CLEF [4] and INEX [3] and has formed the basis of evaluations in several IR domains including:

- Question answering, e.g. [4], which attempts to provide answers to questions such as "How fast can a Cheetah run?", sometimes supplementing answers with additional relevant snippets that might be helpful to the user.
- Entity search, e.g. [2], which assumes the user has an information need that could be solved by with a list of entities that satisfy some properties. A query might, for example, indicate the type of entities to be retrieved (e.g., "castle") and distinctive features (e.g., "German", "medieval").
- Structured retrieval e.g. [3], which aims to retrieve relevant parts of documents in a collection in response to given information need.
- Exploratory search e.g. [5], whereby the user has a poorly defined information need, little knowledge of the topic of interest or is unfamiliar with the search space.

Each of these examples are associated with well-defined tasks or situations. However, it is unclear how reflective these tasks are of real-life wikipedia search behaviour. Are these the most appropriate tasks to be investigating? Are we evaluating these tasks appropriately? Are there more pressing aspects that we, as a research community, should be investigating?

As a starting point to answering these questions, in the following section, we briefly review research that informs on wikipedia search behaviour in naturalistic situations.

2. SEARCHING WIKIPEDIA

The main source of knowledge of wikipedia search behaviour comes from transaction log analyses. Sakai and Nogami [6], for example, logged user interaction with a wikipedia search interface, designed to encourage exploration and development of information needs. They discovered that information needs tend to progress and develop in small steps, usually within query type. For example, users tended to browse pages from person to person or from place to place etc. The implicit structure of wikipedia most likely encourages this behavior

Fissaha and de Rijke [1] also used log analyses to learn about wikipedia searches, distinguishing between "directed" and "undirected" searches by analysing the phrasing of queries. They [also] discovered that a large percentage of searches were undirected and exploratory in nature.

Log-based investigations such as these have the advantage of collecting large quantities of data from naturalistic situations. However, they are limited in that they say nothing about the intention of the user, his experience, or the outcome of the search. For example, the work of Wilson and Elsweiler [7] asserts that many searches will not be motivated by information needs per se, but purely by the user

[1] http://www.wikipedia.org

having an interest in a topic. In their work, they found example search tasks that were motivated by the desire to achieving a particular mood, emotional or physical state or by the presence or need of someone else in the social context. In such cases, the support the user would need from the system and the criteria that should be used to evaluate system performance would be very different to those currently featured in information retrieval research.

We believe that the way wikipedia is constructed, i.e., collaboratively by a subset of the users, the large collection size and broad topic range, linked structure, as well as multimedia prominence of multimedia content will mean that wikipedia will be used for leisure-time tasks. People are motivated to create / edit wikipedia pages as it mirrors their interests. This may not always be positive.

For example, Wilson and Elsweiler [7] describe one study participant reporting frustration that he has again wasted a lot of time aimlessly browsing ebay. This negative outcome - realised through a negative emotion - would not be considered in any current IR methodology.

In the following section we outline our thoughts on what we believe to be a more suitable study design to learn about wikipedia search tasks. We would like to use the workshop as a platform for discussion to improve on this design.

3. LEARNING ABOUT BEHAVIOUR WITH A LOG / DIARY HYBRID

We need to design a study that helps us learn about the the user's motivation for searching, his behaviour in response to this motivation, his satisfaction with the experience as well as his emotional response to the experience.

To investigate these aspects we propose combining the log based approaches scholars have used previously with user diaries. Diary Studies offer the ability to capture factual data, in a natural setting, without the distracting influence of an observer. They also offer the chance to question the user regarding his motivation to search, as well as the search process and feelings and emotions experienced during the search process.

Diary studies also have limitations. These include difficulties in maintaining participant dedication levels throughout the period of study and getting the participants to remember that situations of interest should be recorded. These negative aspects can be offset, however, through careful study design. For example, since Wikipedia is digital and accessed within a web browser, it makes sense to use a digital diary that can also be filled out in a web-browser session, perhaps as a pop up. We plan to build an extension to the Firefox web-browser that detects when a wikipedia page is accessed and if a certain time threshold has elapsed since the last diary entry, the user will be asked to record details about his information need and the motivating situation surround the search. The extension will also record interactions with wikipedia (e.g. pages viewed, search queries submitted etc.), allowing analyses similar to those published previously to be complemented by the diary study data.

To limit the irritation that filling out such a form would cause and to minimise distraction to the search process we plan only to ask two short questions at that time point. The user will be asked to give a brief description of what they are looking for and why. This will be enough information to remind them of the situation at a later time point when we ask more detailed questions regarding the experience, success of the task, how the feelings realized and the factors that influenced these. This data will be elicited through a mixture of fixed and free-form questions.

We plan to triangulate the data collected from the various aspects of our study to create a rich understanding of user needs and behaviour. For example, we plan to look at the content of visited pages; the topic and the kind of media used etc. and look to see how this relates to how participants describe their experiences. We want to see, what affects user behaviour, e.g. does the link structure or the way information is presented, certain content influence behaviour or emotions experienced. The different sources of data we will collect will help us to learn about these complicated behavioural aspects.

4. CONCLUSIONS

So what will we learn from the study and why is it important? The most important point is to find out what makes the users happy; what do they need, how do they behave to achieve these needs and emotional aspects are involved when Wikipedia is searched? An understanding of these issues will inform us on the kind of functionality a wikipedia search tool should offer. Do users want to browse to related topics? Do they like a wide range of possible interesting information or just quirky look up pieces of information as and when they are needed? The proposed study would offer the chance to answer these questions by providing naturalistic data, as well as additional comments from the participants of interest.

5. REFERENCES

[1] S. F. Adafre and M. de Rijke. Exploratory search in wikipedia. In *Proceedings SIGIR 2006 workshop on Evaluating Exploratory Search Systems*, 2006.

[2] G. Demartini, C. Firan, T. Iofciu, R. Krestel, and W. Nejdl. Why finding entities in wikipedia is difficult, sometimes. *Information Retrieval*, 13:534–567, 2010. 10.1007/s10791-010-9135-7.

[3] INEX. Initiative for the evaluation of xml retrieval, 2006. url: http://inex.is.informatik.uni - duisburg.de/2006/.

[4] V. Jijkoun and M. de Rijke. Overview of WiQA 2006. In A. Nardi, C. Peters, and J. Vicedo, editors, *Working Notes CLEF 2006*, September 2006.

[5] B. Kules and R. Capra. Designing exploratory search tasks for user studies of information seeking support systems. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '09, pages 419–420, New York, NY, USA, 2009. ACM.

[6] T. Sakai and K. Nogami. Serendipitous search via wikipedia: a query log analysis. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 780–781, New York, NY, USA, 2009. ACM.

[7] M. L. Wilson and D. Elsweiler. Casual-leisure searching: the exploratory search scenarios that break our current models. In *4th International Workshop on Human-Computer Interaction and Information Retrieval*, Aug 2010. New Brunswick, NJ.