Focused Browsing: Providing Topical Feedback for Link Selection in Hypertext Browsing

Gareth J. F. Jones and Quixiang Li

Centre for Digital Video Processing & School of Computing
Dublin City University, Dublin 9, Ireland
email: gareth.jones@computing.dcu.ie

Abstract. When making decisions about whether to navigate to a linked page, users of standard browsers of hypertextual documents returned by an information retrieval search engine are entirely reliant on the content of the anchortext associated with links and the surrounding text. This information is often insufficient for them to make reliable decisions about whether to open a linked page, and they can find themselves following many links to pages which are not helpful with subsequent return to the previous page. We describe a prototype *focusing browsing* application which provides feedback on the likely usefulness of each page linked from the current one, and a *term cloud* preview of the contents of each linked page. Results from an exploratory experiment suggest that users can find this useful in improving their search efficiency.

1 Introduction

Users approach search engines with a wide range of types of information need; some are very focused, while others are much vaguer. In the latter case the user will often need to browse through multiple documents to address their information need. When browsing between documents returned by a standard search engine users rely on the contents of the item, and in particular the link anchortexts, to determine whether to select a link and progress their search to the linked item. The available information can often be insufficient to enable them to make an informed decision, and they often end up needing to follow a link to determine whether it leads to useful information, and then returning to the previous page when they determine that it does not.

This paper introduces a new enhanced browsing framework of *focused browsing* where the user is provided with feedback on the potential utility of available links from the current page and a summary preview of the content of linked items in the form of a *term cloud*. The principle underlying focusing browsing is analogous to the established concept of *focused crawling* [3][1]. The goal of a focused crawler is to selectively seek out web pages relevant to pre-defined topics. Focused browsing aims to focus a user's browsing behaviour towards "on topic" pages that will be of interest to them.

The remainder of this paper describes details of our prototype browsing application and a preliminary pilot user study carried out with this application.

2 Focused Browsing

Conventional information retrieval (IR) systems for hypertext documents do not provide feedback to a user on the potential utility of links appearing in retrieved documents. This is in contrast to adaptive hypermedia (AH) systems designed for constrained domains where the contents typically comprise manually selected information fragments. These fragments can be composed into personalised presentations based on complex models of the user's goals, preferences and knowledge of the subject at hand. Various navigation support mechanisms are supported by these systems including link ranking, link annotation, and link disabling [4] [2].

A recent emergence in web technology is the facility to enable users to add tags to web content, potentially building up rich descriptions from a community of users. Browsing of these tags is supported using *tag clouds* [5] which provide a means for visualizing groups of tag words. Words to be included in the tag clouds are typically selected based on frequency of user annotation. The significance of individual words can be indicated by varying their font size based on their score. These words are then presented to the user in the form of a simple graphical cloud.

Our focused browsing application provides the user with interactive feedback to support their browsing. It uses IR methods to score document links. While the models used are less sophisticated than those typically used in AH systems, they do not have their domain specific limitations and can easily be applied to any linked documents. A related more complex approach was introduced in [6]. The concept of tag clouds is extended to provide term cloud document summaries.

2.1 Link Scoring

In order to assist the user in determining which link they should follow from the current page, each link is scored with respect to the search request. The intensity of the display of the link is then selected based on its score relative to others leading from the page.

In our current prototype links are scored using a standard $tf \times idf$ approach. After removal of stop words and Porter stemming, terms are weighted as follows. $w(i,j) = tf(i,j) \times \log N/n(i)$, where $tf(i,j)$ is the frequency of term i in the linked page j, N is the no of pages linked from the current page, and $n(i)$ is the no of them which contain i.

A matching score is then computed by summing matching term weights between the user's search query and each linked document. The scores for each linked page are then ranked and the links displayed with their intensity determined by their relative rank score. While the current prototype only uses the query to score the links, many other factors could be used. For example, a user profile based on current or ongoing browsing and search behaviour, or an absolute page significance factor such as PageRank [7] or one based on access behaviour [8].

2.2 Term Clouds

Varying the intensity of the links obviously does not indicate to the user what the linked pages contain. Indeed the score may actually be misleading if the user changes the focus of their information need. To further assist the user in determining whether to follow a link the focused browser makes available a *term cloud* of each linked page. Term clouds are related to the “fluid links” introduced in [10].

Term clouds present a simple summary of the linked page in the form of a cloud of words or phrases taken from the document. Selection of terms is similar to the principles used to form query-biased summaries (QBS) in [9]. Term clouds are more compact than QBSs and are intended to be more efficient in terms of cognitive user interaction and the display space needed, but use of QBSs as an alternative document visualization method will be explored in future work.

The components used to select words for inclusion in term clouds are as follows:

Term Weighting The first word significance score $ws1(i,j)$ is calculated using the $tf \times idf$ function from Section 2.1. The $ws1(i,j)$ is calculated for each non stop term appearing in each linked document. This is then multiplied by a scaling factor $ws1$.

Title Term Weighting Words contained in the title of a document are often important to its topic. Words appearing in the title of each document are given a boost of a fixed scalar value $ws2$.

Location Weighting Words appearing at the beginning of a document are more likely than those later in it to be significant to its topic. Words appearing in the first sentence are given a fixed scalar boost of $ws3$.

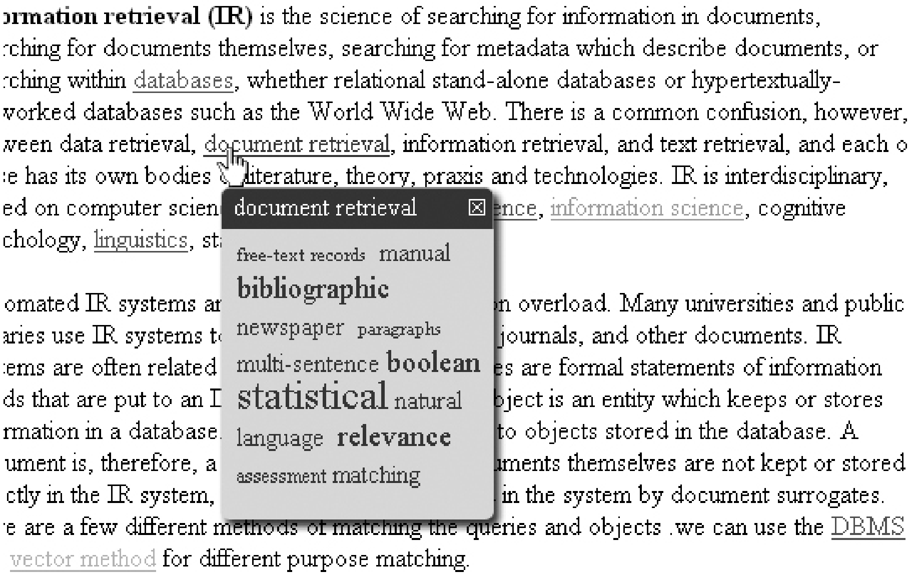
Relationship to Current Document Assuming that the user’s browsing is focused on the topic of the current page, we give an additional scalar boost $ws4$ to terms which appear in the current document.

Combining the Term Scores The final combined score for each term is a simple sum of the four components. The highest scoring ones are then selected for inclusion in the term cloud. The values of $ws1$, $ws2$, $ws3$ and $ws4$ were selected empirically based on experimentation with the small test evaluation site used in our pilot experimental study, with font size varied according to the combined term score. Values used in the current prototype are 0.5, 0.2, 0.1, 0.1 respectively.

3 Prototype Application

Figure 1 shows a sample interface from our prototype focused browsing application. The hyperlinks are highlighted at different levels indicating the estimated utility of each link to the user as described in section 2.1. In the example the user has moused over the link to “document retrieval”, in response to this the term cloud shown is produced.

Implementing a practical focusing browsing application requires a tighter integration between the search engine and browser than is generally the case. Computation of link scores and term clouds requires access to term data most easily available from a search engine, thus documents must be displayed from the search engine cache or if they are downloaded from their original sources, they must be automatically rewritten prior to display to include variations in link display and include links to term clouds.



rmation retrieval (IR) is the science of searching for information in documents,
rching for documents themselves, searching for metadata which describe documents, or
rching within databases, whether relational stand-alone databases or hypertextually-
worked databases such as the World Wide Web. There is a common confusion, however,
ween data retrieval, document retrieval, information retrieval, and text retrieval, and each o
e has its own bodies literature, theory, praxis and technologies. IR is interdisciplinary,
ed on computer scien ence, information science, cognitive
chology, linguistics, st

omated IR systems ar n overload. Many universities and public
aries use IR systems t journals, and other documents. IR
ems are often related s are formal statements of information
ds that are put to an I bject is an entity which keeps or stores
rmation in a database. to objects stored in the database. A
ument is, therefore, a ments themselves are not kept or stored
ctly in the IR system, in the system by document surrogates.
e are a few different methods of matching the queries and objects .we can use the DBMS
vector method for different purpose matching.

document retrieval ☒

free-text records manual bibliographic newspaper paragraphs multi-sentence boolean statistical natural language relevance assessment matching

Fig. 1. Prototype Focused Browsing Application.

4 Pilot Study

To obtain some initial user feedback on our prototype application a small test website was built on the subject of “information retrieval.” This was intended for students wishing to explore topics in IR by browsing an online learning resource.

5 test users each entered 5 search queries on the topic of IR giving a total of 25 queries for this pilot study. 3 queries were judged to be off topic and thus not covered by the test web site; the analysis is thus based on the remaining 22 queries.

Question 1: do the different colour links help you finding target information, how good it is? A Poor, B good C excellent.

Poor: 7 queries, Good: 13 queries, Excellent: 2 queries.

Question 2: Do you find that the summary (keywords) of the documents accurately reflect the content of the documents. A very poor, B poor, C good, D very good, E excellent

Poor: 1 user, Good: 3 users, Very good: 1 user,

These results are reasonably encouraging. However, the test website used for the experiments was very limited in size and scope, meaning that the users had only very limited opportunity to actually browse, and the results of this very simple evaluation are obviously not significant. In order to explore the utility of focused browsing properly, we plan to incorporate focused browsing into a larger domain specific search tool in the near future.

5 Conclusions and Further Development

Focusing browsing provides feedback to users engaged in the exploration of hypertext document collections, including the web. The aim being to improve the efficiency with which they can access the information they require. We have developed a prototype application providing feedback using highlighting of available links and term cloud summaries. Potential extension of the application would be to extend it to make use of user tags [5] with the term clouds to form term/tag clouds. In further work we will be incorporating focused browsing into a larger experimental digital library and conducting a more formal experimental evaluation.

Acknowledgement

The authors are grateful to Hyowon Lee for technical assistance with graphics and for the very helpful comments of the reviewers. Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD — project MultiMATCH contract IST-033104. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

References

[1] de Assis, G. T., Laender, A. H. F., Gonçalves, M., A., and da Silva, A., S.: Exploiting Genre in Focused Crawling. In Proceedings of the 14th International Symposium on String Processing and Information Retrieval (SPIRE 2007), Santiago, Chile, pages 62-73, 2007.

[2] Brusilovsky, P. Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 11(1-2):87-110, 2001.

[3] Chakrabarti, S., van den Berg, M., and Dom, B.: Focused crawling: a new approach to topic-specific Web resource discovery. *Journal of Computer Networks*, 31(11-16):1623-1640, 1999.

[4] De Bra, P., Brusilovsky, P., and Houben, G.-T. Adaptive Hypermedia: From Systems to Framework. *ACM Computing Surveys*, 31(4), 1999.

[5] Hassan-Montero, Y., and Herrero-Solana, V.: Improving Tag-Clouds as Visual Information Retrieval Interfaces. In Proceedings of the International Conference on Multidisciplinary Information Sciences and Technologies (InSciT2006), Mèrida, Spain, 2006.

[6] Olston, C. and Chi, E. H.: ScentTrails: Integrating Browsing and Searching on the Web. *ACM Transactions on Computer-Human Interaction*, 10(3):177-197, 2003.

[7] Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Computer Science, Stanford University, 1998.

[8] Soules, C., and Ganger, G.: Connections: Using Context to Enhance File Search. In *Proceedings of the 20th ACM Symposium on Operating Systems Principles (SOSP'05)*, pages 119-132, Brighton, U.K., 2005.

[9] Tombros, A., and Sanderson, M.: Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of the Twenty First ACM Annual Conference on Research and Development in Information Retrieval (SIGIR 98)*, Melbourne, Australia, pages 2-10, 1998

[10] Zellweger, P. T., Chang, B.-W., and Mackinlay, J. D.: Fluid Links for Informed and Incremental Link Transitions. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*, Pittsburgh, U.S.A., 1998.