Work in Progress: Navigating Document Networks

Mark D. Smucker

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
smucker@cs.umass.edu

ABSTRACT

While much research effort has been expended on innovative user interfaces for information retrieval (IR), deployed IR user interfaces have adopted few innovations. Rather than design another novel user interface tool that users never adopt, we decided that our first step would be to better understand the nature of an adopted tool. In that vein, we are in the process of studying the potential and the actual performance of find-similar, which is a widely adopted tool that allows a user to request documents similar to a given document. Find-similar is a compelling IR interface tool for the very reason that users appear to have adopted it and that it has the potential to provide to users the power long known to be available via relevance feedback. Our hope is that by better understanding find-similar, we'll be able to take that understanding and apply it to other user interface tools that will both be powerful and be adopted by users.

1. INTRODUCTION

Find-similar allows a user to request a list of documents similar to a given document. As such, find-similar provides a way for users to navigate from one document to another and browse by document similarity. This feature is typically instantiated as a button or link next to each result in the list of search results. For example, the Excite search engine labeled its find-similar link "More Like This: Click here for a list of documents like this one."

Find-similar can be an important and valuable tool for improving IR systems. Spink et al. [6, 7] analyzed samples of Excite's query logs and reported that between 5 and 9.7 percent of the queries came from the use of the "more like this" find-similar feature. Lin et al. [2] have reported that for the US National Library of Medicine's search engine, PubMed, 18.5% of non-trivial search sessions involve clicks on articles suggested by PubMed's find-similar, which PubMed refers to as *related articles* [3].

While relevance feedback is well known to be a powerful technique for improving retrieval performance, it has seen little adoption by popular search systems. We've shown that find-similar has the potential to match the performance of relevance feedback [4]. Earlier work by Wilbur and Coffee [8] found that certain browsing patterns could improve over the original query's ranking.

HCIR'07 Workshop on Human-Computer Interaction and Information Retrieval, October 23, 2007, Cambridge, Massachusetts.

2. DOCUMENT NETWORKS

Find-similar can be studied and understood in graph theoretic terms. Each document or web page is a node in a graph. When find-similar is applied to a document, find-similar provides the user with links to the similar documents and these links are effectively added to the document. For a web page, these automatically created links join the already existing links on the page. If the added links are good, users should be able to use the links to navigate to other relevant documents. These links make documents in the graph closer to each other, which is good, but these links also increase the amount of time that a user needs to spend examining the page, which is bad.

In a broad sense, find-similar aims to add links to documents such that the time for a user to get from relevant document to relevant document is minimized. These added links can represent many different types of similarity. The most studied form of similarity is content-based, which typically involves comparison of the terms in each document. The web's hyperlinks are another form of similarity; a similarity defined by the authors of the web pages. Find-similar could add links to documents from a similar time period or documents written by the same author, for example.

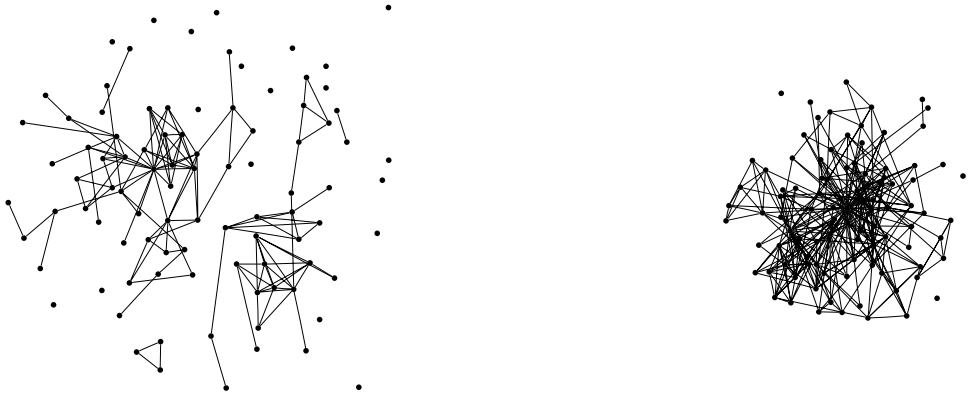
We've proposed a pair of metrics that can be used to measure the navigability of documents [5] and preliminary experiments show that existing web hyperlinks are ill-suited for navigation from relevant document to relevant document compared to links produced by a content similarity measure.

Different types of content similarity measures create different document networks. Using simulated browsing behaviors, we have found that a query-biased document-to-document similarity consistently outperforms a baseline "regular" similarity. Figure 1 shows an example of how query-biased similarity can result in a better clustering of relevant documents. We intend to study query-biased similarity using our navigability metrics and obtain a better understanding of the more navigable networks produced by query-biased similarity.

3. USER BEHAVIOR

While understanding of the find-similar document networks is important to understanding find-similar and maximizing its performance, we also want to better understand how people use find-similar.

In a study by Huggett and Lanir [1], users found more relevant documents using an interface that provided find-similar over an interface without find-similar. Huggett and Lanir's study used small newswire collections of 2000 doc-



(a) (b)

Figure 1: Simplified depictions of the relevant document networks for TREC topic 337, "viral hepatitis." The network on the left (a) uses regular similarity while the network on the right (b) uses query-biased similarity, which better clusters relevant documents. The documents are closer in figure (b) because they are higher ranked in each other's ranked lists. Links are drawn between two documents when one of the pair is close to the other. The actual relevant document network is a weighted, directed graph [5].

uments and limited test subjects to two minutes for each search. We would like to examine find-similar's usage on much larger TREC collections and on the web. As Huggett and Lanir did, we will also compare our find-similar implementations to IR systems allowing query reformulation and one of our measures of performance will be the rate at which relevant documents are found. We hypothesize that users adopt IR interface features that provide better rates of information discovery as opposed to tools that may improve ranking performance but overall slow the rate of finding relevant documents.

We also want to learn about how users navigate the document networks formed by find-similar and what forms of user interface support are needed to maximize performance. For example, how far away from the original query will users navigate? Do users apply find-similar to documents that are non-relevant but which they think might lead them to relevant documents? How is find-similar usage interleaved with query reformulation? We intend to answer these and other questions as part of a planned user study.

4. CONCLUSION

Find-similar provides a chance for us to study a user interface feature that has been adopted by search engines and shown to be frequently used by users. To date, we've shown that find-similar has the potential to match a traditionally styled multiple item relevance and that different forms of similarity offer different levels of inherent navigability. Our next steps include a closer examination of similarity functions such as query-biased similarity and actual user studies. As much as possible, we hope to learn what has allowed find-similar to become a useful tool for search when so many other user interface features have failed to succeed and be adopted outside of the laboratory.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] M. Huggett and J. Lanir. Static reformulation: a user study of static hypertext for query-based reformulation. In *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, pages 319–328. ACM Press, 2007.

[2] J. Lin, M. DiCuccio, V. Grigoryan, and W. J. Wilbur. Exploring the effectiveness of related article search in pubmed. Technical Report LAMP-TR-145/CS-TR-4877/UMIACS-TR-2007-36/HCIL-2007-10, College of Information Studies, University of Maryland, College Park, July 2007.

[3] Pubmed, `www.pubmed.gov`. "Related articles": `www.nlm.nih.gov/bsd/pubmed_tutorial/m5002.html`.

[4] M. D. Smucker and J. Allan. Find-similar: Similarity browsing as a search tool. In *SIGIR '06*, pages 461–468. ACM Press, 2006.

[5] M. D. Smucker and J. Allan. Measuring the navigability of document networks. In *SIGIR '07 Web Information-Seeking and Interaction Workshop*, 2007.

[6] A. Spink, B. J. Jansen, and H. C. Ozmultu. Use of query reformulation and relevance feedback by excite users. *Internet Research: Electronic Networking Applications and Policy*, 10(4):317–328, 2000.

[7] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *JASIST*, 52(3):226–234, 2001.

[8] W. J. Wilbur and L. Coffee. The effectiveness of document neighboring in search enhancement. *IPM*, 30(2):253–266, 1994.