Using Similarity Links as Shortcuts to Relevant Web Pages

Mark D. Smucker and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
{smucker, allan}@cs.umass.edu

ABSTRACT

Successful navigation from a relevant web page to other relevant pages depends on the page linking to other relevant pages. We measured the distance to travel from relevant page to relevant page and found a bimodal distribution of distances peaking at 4 and 15 hops. In an attempt to make it easier to navigate among relevant pages, we added content similarity links to pages. With these additional links, significantly more relevant documents were close to each other. A browser plug-in or other tool that provides links to pages similar to a given page should increase the ability of web users to find relevant pages via navigation.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Experimentation

Keywords: Web, shortest paths, cluster hypothesis, relevance feedback, content and link similarity, web graph

1. INTRODUCTION

When a user finds a web page that the user considers relevant and wants to find other relevant web pages, a seemingly reasonable approach the user can take is to follow the page's links to other pages. The degree to which this approach is reasonable depends on the extent to which the well-known cluster hypothesis is true on the web when the similarity measure is the distance to navigate from one document to another using hyperlinks.

We first investigated the extent to which the cluster hypothesis is true on the web graph, and then we attempted to improve the navigability of the web with the automatic addition of links to similar web documents. If our automatically created links make relevant documents easier to find, such a capability could be packaged into some sort of browser plug-in that allows the user to request a list of web pages similar to the current web page. We found that:

- Relevant documents are either within distance 5 of another relevant page or are as likely to be reached at greater distances as non-relevant documents.
- The automatic addition of content similarity hyperlinks can significantly increase the number of relevant documents reachable from a given relevant document.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

2. EXPERIMENTS

We ran our experiments using the wt10g TREC web collection. Soboroff [5] has shown the wt10g collection to have structural characteristics similar to the web. We used the TREC 2001 web ad-hoc topics numbered 501-550.

We compared the web graph with two augmented versions of the graph. For each topic, we augmented the graph by adding 10 out-links to each relevant document. In the first case, we added links to the 10 most content similar documents. This case corresponds to our envisioned browser plug-in that provides a list of the 10 web pages most similar to the current page. In the second case, we added 10 random links. This case allows us to make sure that the mere addition of links does not make relevant documents closer to each other on the web graph.

For each topic, we measured the shortest path distance from each relevant document to all other documents. Traversing a link or one "hop" is a distance of 1.

We did not add any out-links to the non-relevant documents. We think that searchers would only utilize a feature providing these similarity links when they are looking for pages similar to a current relevant web page. In effect, we say the links do not exist on non-relevant pages because we do not believe that users will utilize the feature on non-relevant pages. In addition, it is unlikely that non-relevant documents can be used to find relevant documents based on content similarity.

Because we measure shortest paths, the distance from a relevant document to another relevant document is an upper bound on path length. Augmenting non-relevant documents with additional links could only have shortened the path lengths. On the other hand, if we had augmented all documents with additional out-links, the non-relevant documents would be closer than we report.

We computed overall averages by first averaging the measurements for a topic's relevant documents and then averaging all the topics.

We constructed the web graph using the wt10g out_links file. To compute the document to document content similarity, we created a maximum likelihood estimated model of each document. We truncated each model to consist of only the document's 50 most probable terms. Using this model, we measure the similarity of the other documents using the KL-divergence. We used Dirichlet prior smoothing and set its parameter to 1500. We stemmed using the Krovetz stemmer and used an in-house list of 418 stop words. We used the Lemur toolkit for our experiments.

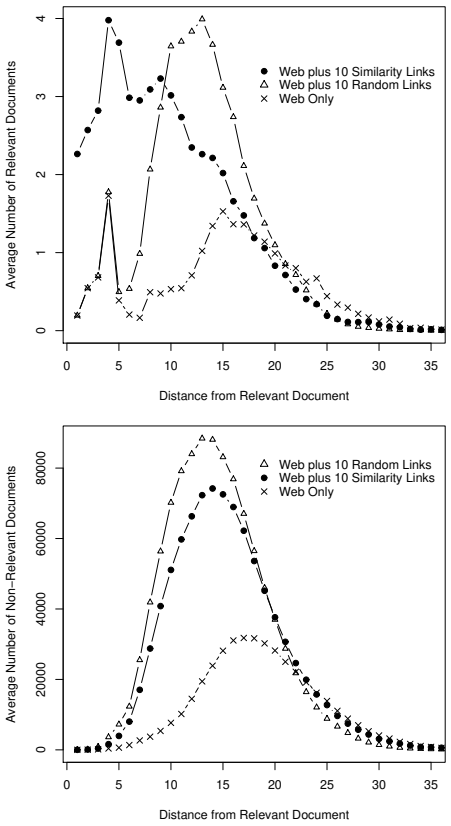

Figure 1: The distance of relevant and non-relevant documents from relevant documents.

3. RESULTS AND DISCUSSION

Figure 1 shows the distribution of relevant and non-relevant documents as a function of their web graph distance from relevant documents. The "Web Only" distribution of relevant documents shows a clear bimodal shape peaking at distances of 4 and 15. The distribution of non-relevant documents is unimodal peaking around 17. The relevant documents reached at distances greater than 6 are most likely reached simply as a result of the interconnectedness of the web and are no easier to reach than non-relevant documents. The peak of relevant documents at distances less than 6 demonstrate that the cluster hypothesis is true on the web graph at least for some relevant documents.

The addition of 10 content similar links brings a significant number of relevant documents closer to the average relevant document. A similar peaking at a distance of 4 shows that the relevant documents reached using the similarity links are likely independent of the ones being found using the existing links. With existing web links, on average 0.19 relevant documents are within distance 1 from a relevant document. With the addition of 10 content similarity links, an average of 2.26 relevant documents are distance 1 from a relevant document. The performance boost continues for the larger distances. At distance 4, the web-only distribution has an average 1.73 relevant documents and the web plus similarity links has 3.98.

The addition of 10 random links results in no significant gains at distance 5 or less. The additional links do bring documents closer but do not help at short distances. Thus, the gains obtained by similarity linking appear to result from relevant documents being more similar to each other than simply from increasing the graph connectivity.

4. RELATED WORK

Allan [1] studied the automatic construction of hypertext with a focus on the typing of links while here we make no attempt to address the issue of helping a user choose among links. We have previously studied using a single document as query to find similar pages [4], but here our emphasis has been on understanding the distribution of relevant and non-relevant documents on the web graph with and without the addition of similarity links. Chakrabarti et al. [2] studied the rate at which topics dissipate on the web graph, while we looked at the related – but different – notion of distance between relevant pages. Our results are in line with Menczer's prediction that one would be able to infer relevance of a page at a maximum of around 4 to 5 hops [3]. Vassilvitskii and Brill [6] used distance on the web graph to perform a reranking of search results given that relevant documents link to other relevant documents. For each top ranked search result, they performed a limited breadth first search and found that searching to a distance of 4 resulted in the best performance. Our results explain their finding by showing that relevant documents are found within a distance of 5 or are as likely to be found as non-relevant documents.

5. CONCLUSION

We found a bimodal distribution for the distance of relevant documents to each other on the web graph. Relevant documents are within a distance of 5 of each other or as likely to be reached as non-relevant documents. The automatic addition of 10 content similarity links brings significantly more relevant documents close to each other. Augmenting hypertext with similarity links should aid the searcher attempting to navigate from a relevant document to other relevant documents.

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] J. Allan. Building hypertext using information retrieval. *IPM*, 33(2):145–159, 1997.

[2] S. Chakrabarti, M. M. Joshi, K. Punera, and D. M. Pennock. The structure of broad topics on the web. In *WWW '02*. ACM Press, 2002.

[3] F. Menczer. Lexical and semantic clustering by web links. *JASIST*, 55(14):1261–1269, 2004.

[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] I. Soboroff. Do TREC web collections look like the web? *SIGIR Forum*, 36(2):23–31, 2002.

[6] S. Vassilvitskii and E. Brill. Using web-graph distance for relevance feedback in web search. In *SIGIR '06*, pages 147–153. ACM Press, 2006.