The Impact of Semantically Related Links on Retrieval

Marijn Koolen   Jaap Kamps
University of Amsterdam, The Netherlands
{m.h.a.koolen, kamps}@uva.nl

ABSTRACT

Link-based ranking algorithms are based on the often implicit assumption that linked documents are semantically related to each other, and that link information is therefore useful for retrieval. The main aim of this paper is to test this underlying assumption on why link evidence works. We observe that links between semantically related pages are more effective for ad hoc retrieval than links between unrelated ones. These findings shed further light on our understanding of the nature of link evidence. This paper is a compressed version of [3].

1. INTRODUCTION

Link-based ranking algorithms such as relevance propagation [8], and SALSA [5], use the assumption that linked documents have related content. So far, this assumption has remained implicit, because it is hard to measure the semantic relatedness of linked documents independent from the feedback of a retrieval system given a search query. Kurland and Lee [4] showed that generating links based on document similarity can help improve ad hoc retrieval effectiveness. However, these results do not show such links are effective *because* they connect semantically related documents.

Wikipedia allows us to measure the semantic relatedness of documents independently, and, with the INEX Wikipedia Ad Hoc test-collections since 2006, also allows us to study its impact on the effectiveness of link evidence for retrieval. Kamps and Koolen [2] found that Wikipedia links behave very much like links on the larger Web. Wikipedia being a part of the Web, we expect our findings to be generally applicable.

We distinguish between algorithms that use global, query-independent evidence, such as PageRank [7], and local, query-dependent algorithms such as SALSA, which use the links between a subset of documents retrieved for a given topic. In this paper, local link evidence refers to the links between the top 100 results. Najork [6] found that query-dependent link evidence is more effective on a large Web corpus than query-independent link evidence. The query-dependent set of links is a subset of the global link structure. Evidently, some, but not all, links are useful for retrieval. This confronts us with the question:

- Are links between semantically related documents more effective for ad hoc retrieval than links between unrelated ones?

We use the category hierarchy of Wikipedia to measure the semantic distance between two documents. By filtering links based on their semantic distance we study the impact of the semantic nature of links on the effectiveness of link evidence. Because filtering also changes the quantity of links, we compare semantic filtering against a random filter.

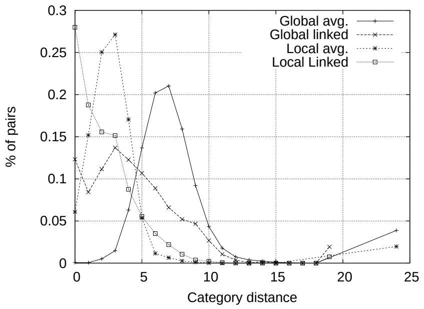

Figure 1: Distribution of category distances between documents.

2. LINKS AND CATEGORIES

We use the INEX 2006 Wikipedia collection [1], consisting of over 650,000 documents, and the 221 topics and associated relevance judgements created by participants of the INEX 2006–2007 Ad Hoc Tracks. The Wikipedia category structure allows us to determine how semantically related two documents are.

We opt for the path-based measure using the category hierarchy because it is simple and has proven to be reasonably effective in semantic relatedness evaluations [9]. The category distance between two documents d_a and d_b is the minimum number of edges between any pair of the categories of d_a and d_b. The distribution of distances between documents is shown in Figure 1. We randomly sampled one million pairs of documents and found the average distance is 6.60. Over all linked document pairs the average distance is 4.04. There is a relation between links and semantic relatedness.

Among the documents in the top 100 results[1] of the 221 queries, the average distance is 2.56, showing documents in the top results are more semantically related to each other than in the overall collection. The average distance of linked document pairs in the top 100 results is 2.22. Local link evidence provides a stronger signal that two documents are semantically related than the text evidence.

3. SEMANTIC RELATEDNESS AND EFFECTIVENESS OF LINKS

We use the category structure to filter links and thereby control the semantic nature of link evidence. Our baseline run is a standard language model run with linear smoothing ($\lambda = 0.15$) and a document length prior $P_{length}(d) = |d| / \sum_{d' \in D} |d'|$, where d and d'

Copyright is held by the author/owner(s).
DIR '12, 23-24 February, 2012, Ghent.
ACM X-XXXXX-XX-X/XX/XX.

[1] The baseline retrieval system is described in the next section.

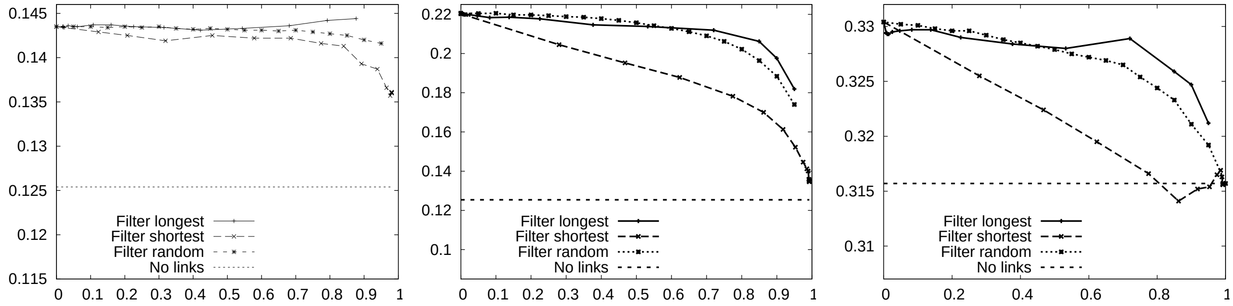

Figure 2: The impact of filtering links on MAP when ranking the top 100 results by global in-degree (left), local union degree (middle) and local union degree and content score. The x-axis shows the percentage of links removed.

are documents in collection D. The length prior promotes longer documents and improves MAP from 0.2561 to 0.3157.

We use link evidence to re-rank the top 100 results. For global link evidence we take the in-degree and for local link evidence we take the union of the in- and out-degrees. We found that these are equally or more effective than other types of link evidence [2]. We filter links based on the path length measure described above. In the first filtering step the SD (shortest distance) filter removes the links at distance 0, in the second step the links at distance 1, etc. The LD filter first removes the longest distance links. For comparison, we also look at the impact of randomly filtering links.

The impact of filtering on the effectiveness of the global in-degrees is shown in Figure 2. The x-axis shows the percentage of links filtered. The left figure shows the impact on global in-degree. The Random and LD filters have little impact on the in-degrees, but removing the shortest distance links hurts performance when more than 80% of links are removed. We compare this with a random ordering (averaged over 20 iterations) of the top 100 results. Even a small number of links is better than random ordering. Filtering has little impact on global link evidence, probably because the link graph is very rich and the high-degree pages are very robust against filtering. Its effectiveness seems unrelated to semantic relatedness.

Local links (middle) are far more effective than global links. But here, random filtering has a bigger impact. If we remove the shortest distance links first, performance drops faster than with random filtering, while if we remove the longest distance links first, performance remains stable longer. The shortest semantic distance links are the more effective links.

We also combine text and link evidence by multiplying the language model probabilities with local link degrees to obtain a new ranking (Figure 2, right). The baseline scores are the straight dotted lines. The local union degrees improve upon the baseline performance. With random filtering, MAP gradually drops as we remove more links. If we remove the SD links first, the improvement drops faster and the score even falls below that of the baseline. With the LD filter, MAP stays higher than with random filtering. Again, the shortest distance links are the most effective. Effectiveness of local link evidence is directly related to the semantically relatedness of the linked documents.

Note that filtering does not improve the effectiveness of local link evidence, which may be because the local link graph is already filtered on the search topic, which is a semantic filter in itself.

4. CONCLUSIONS

We looked at whether links between semantically related pages are more effective for retrieval than links between unrelated ones.

When the aim of link evidence is to identify important documents, links between semantically related documents are not more effective than links between unrelated ones. When we make link evidence sensitive to the context of the search topic, the role of link evidence shifts to identifying topically relevant documents, and here links between semantically related documents are indeed more effective than links between unrelated ones.

More generally, our findings confirm the assumption that (query-dependent) link information is effective for retrieval because it signals the semantic relatedness of linked documents. This adds to our understanding of why link evidence works, which can help in developing better link-based ranking methods.

Acknowledgments

This work was generously supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.513, 639.-072.601, and 640.001.501).

References

[1] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40(1):64–69, June 2006.

[2] J. Kamps and M. Koolen. Is Wikipedia link structure different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM 2009)*, pages 232–241. ACM Press, New York NY, USA, 2009.

[3] M. Koolen and J. Kamps. Are semantically related links effective for retrieval? In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, and V. Murdoch, editors, *Advances in Information Retrieval: 33rd European Conference on IR Research (ECIR 2011)*, volume 6611 of *LNCS*, pages 92–103. Springer, 2011.

[4] O. Kurland and L. Lee. Respect my authority!: Hits without hyperlinks, utilizing cluster-based language models. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. Järvelin, editors, *SIGIR*, pages 83–90. ACM, 2006.

[5] R. Lempel and S. Moran. Salsa: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, 2001.

[6] M. Najork. Comparing the effectiveness of hits and salsa. In M. J. Silva, A. H. F. Laender, R. A. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falcão, editors, *CIKM*, pages 157–164. ACM, 2007.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[8] A. Shakery and C. Zhai. A probabilistic relevance propagation model for hypertext retrieval. In P. S. Yu, V. J. Tsotras, E. A. Fox, and B. Liu, editors, *CIKM*, pages 550–558. ACM, 2006.

[9] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, July 2006.