How Different are Wikipedia and Web Link Structure?

Jaap Kamps[1,2] Marijn Koolen[1]

[1] Archives and Information Studies, University of Amsterdam
[2] ISLA, Informatics Institute, University of Amsterdam
{kamps,m.h.a.koolen}@uva.nl

ABSTRACT

In this paper, we investigate the difference between Wikipedia and Web link structure with respect to their value as indicators of the relevance of a page for a given topic of request. Our main findings are: First, Wikipedia link structure is similar to the Web, but more densely linked. Second, Wikipedia's outlinks behave similar to inlinks and both are good indicators of relevance, whereas on the Web the inlinks are more important. Third, when incorporating link evidence in the retrieval model, for Wikipedia the global link evidence fails and we have to take the local context into account.

1. INTRODUCTION

The principal difference between Web retrieval and general information retrieval, is the abundant link structure of the Web which can been exploited to improve information retrieval in algorithms [4, 7]. Similar to the earlier use of citations in bibliometrics, a link can be considered as a "vote" for a page being authoritative. Wikipedia's links are a special case of the general hyperlinks that connect the World Wide Web. Internal links in Wikipedia are typically based on words naturally occurring in a page and link to another "relevant" Wikipedia page. Our conjecture is that the links in Wikipedia are different from links between arbitrary Web documents.

Our main research question is to find out if, and how, the link structure of Wikipedia differs from the Web at large with respect to its value for promoting retrieval effectiveness. To investigate this, we use two IR test collections consisting of documents plus search requests and associated relevance judgments. For Wikipedia, we use the INEX 2006 and 2007 Ad hoc collections, together consisting of 217 ad hoc topics and an XML version of Wikipedia containing over 650,000 articles [1]. and for the Web we use the TREC 2004 Web Track collection, consisting of 225 topics and the 1.2 million documents .GOV collection. We make no particular claims on the representativeness of this data set for the current Web, which is infinitely large and highly heterogeneous, but expect it to be a close enough approximation for our purposes [8].

Our main research question breaks down in two parts. We start by investigating the Wikipedia link structure with a comparative analysis of the two IR test collections, Wikipedia and .GOV. The second part of our main research question is about the effectiveness of link-based evidence. At TREC, we have seen that link degree is not effective for general ad hoc retrieval [2]. However, for web-centric retrieval tasks like entry page finding, link indegree proved highly beneficial [5]. What is the impact of link evidence on Web-centric retrieval on .GOV and ad hoc retrieval on Wikipedia?

Table 1: Statistics of the .GOV and Wikipedia collections

		min	max	mean	median	stdev
GOV	Indegree	0	44,228	8.90	1	126.00
	Outdegree	0	653	8.90	4	16.61
Wiki	Indegree	0	74,937	20.63	4	282.94
	Outdegree	0	5,098	20.63	12	36.70

2. COMPARATIVE ANALYSIS

In this section, we look in detail at the link structures of the Wikipedia and Web collections. The .GOV collection contains 1,247,753 documents and 11,110,989 unique links between these pages (we ignore links which point to, or from, pages outside the collection). The Wikipedia collection contains 659,304 documents and a total of 13,602,613 unique links between these pages. We have also looked at how many of these links are reciprocal: there are 1,269,988 (11.4%) reciprocal links in the .GOV collection, and 1,182,558 (8,7%) reciprocal links in the Wikipedia collection. The higher fraction of reciprocal links in the .GOV collection is likely due to the presesence of nagivational links within web-sites. Statistics of the degree distributions is given in Table 1. The Wikipedia collection has fewer documents and a larger number of links and is thus more densely linked. This is surprising in the sense that the .GOV domain is much older, and link density tends to increase over time [6]. There are two effects which help explain why the Wikipedia link graph is more "complete" than the .GOV link graph. First, due to the structured nature of Wikipedia, it is much clearer for Wikipedia authors where to link to. Second, due to peer editing and automatic link detection, "missed" links will be added over time.

We analyse the prior probability of relevance (PoR) of a page with a particular link degree. We use IR test-collections with search topics and associated relevance judgments. For the 225 topics of the .GOV collection we have 1,763 relevant documents, for the 217 topics of the Wikipedia collection we have 11,896 relevant documents. If the degrees of relevant documents deviate from the degrees of non-relevant documents, they may possibly be used as indicators of relevance. We calculate the PoR as follows. We sort all documents on ascending degree into bins of 10,000 documents. The PoR for the documents in each bin is the ratio of relevant documents in that bin. If link degree is related to relevance, we expect the PoR to go up with increasing degree. Figure 1 shows the results. In the .GOV collection, the probability of a document being relevant increases with indegree. For outdegree, the probability of relevance initially rises but then drops as the outdegree further in-

*This is an extended abstract of: J. Kamps and M. Koolen. Is Wikipedia link structure different? In *Proceedings WSDM 2009*, pages 232–241. ACM, 2009.

DIR-2010 January 25, 2010, Nijmegen, the Netherlands.

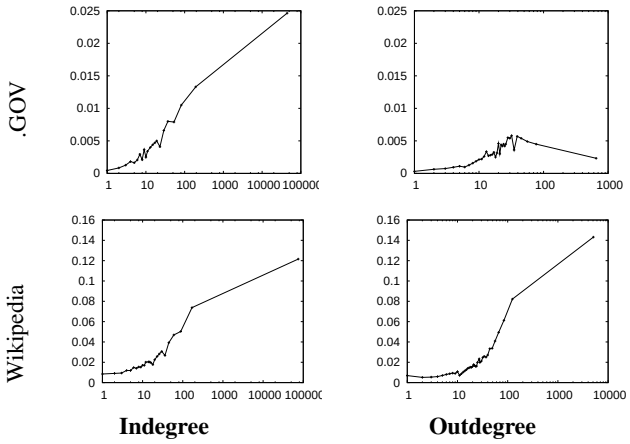

Figure 1: Prior probability of relevance of indegree (left) and outdegree (right) for .GOV (top) and Wikipedia (bottom)

creases. In the Wikipedia collection both in- and outdegree seem to be good indicators of relevance: a higher degree corresponds to a higher probability of relevance. This is not a result of pages linking back-and-forth, the fraction of reciprocal links in Wikipedia is actually lower than in .GOV. This suggests that outlinks in Wikipedia behave very much like inlinks. This is consistent with a semantic nature of links in Wikipedia: if a link from A to B means that B is relevant (in some sense) to A, then it is also likely A is relevant (in some sense) to B. This signals differences in the link structure of Wikipedia and the Web at large. For the semantic links of Wikipedia, the difference between incoming and outgoing links seems to disappear and both can be used as indicators of relevance.

3. EFFECTIVENESS OF LINK EVIDENCE

We work in the language modelling framework, incorporating link evidence into the retrieval model similar to Kraaij et al. [5]. We multiply the content-based retrieval score with the link degree and conduct experiments with them on the TREC 2004 Web track topics and on the combined INEX 2006 and 2007 Ad Hoc track topics. Link indegree can be considered on a global level, i.e. indegree over the whole collection (similar to PageRank), or on a local level, i.e. indegree within the subset of articles retrieved as results for a given topic (similar to HITS). For the local link degrees we use only the links between the top 100 ranked results.

Results for the .GOV collection are shown in Table 2. As we expected from the PoR plots, the indegrees are much more effective than the outdegrees, although the outdegrees are still effective. The local degrees are more effective for Mean Average Precision (MAP), but the global outdegrees are the most effective for Mean Reciprocal Rank (MRR). Taking the log of the priors to tone down their impact is less effective.

Results for the Wikipedia collection are shown in Table 3. Here, both the global in- and outdegrees improve MRR but hurt MAP, even when logged. For ad hoc retrieval, with many relevant documents, global link evidence leads to infiltration of important but off-topic pages that are ranked low on content score. Local link degrees lead to significant improvements, with little difference between the impact of in- and outdegrees.

4. DISCUSSION AND CONCLUSIONS

We investigated the difference between Wikipedia and Web link structure, based on evidence from two IR test-collections. Wikipedia is more densely linked than .GOV. We observe that Wikipedia

Table 2: Results of the different link priors over in- and outdegree on the 225 topics of the Web track collection

	MAP		MRR	
Run id	**Glob**	**Loc**	**Glob**	**Loc**
baseline	0.3970		0.4662	
in	0.4738$^{\bullet}$	**0.4799**$^{\bullet}$	**0.5885**$^{\bullet}$	0.5655$^{\bullet}$
out	0.4299$^{\circledcirc}$	0.4497$^{\bullet}$	0.5046$^{\bullet}$	0.5199$^{\bullet}$
log.in	0.4449$^{\bullet}$	0.4410$^{\bullet}$	0.5209$^{\bullet}$	0.5148$^{\bullet}$
log.out	0.4082$^{\circledcirc}$	0.4181$^{\bullet}$	0.4789$^{\bullet}$	0.4879$^{\bullet}$

Table 3: Results of the different link priors over in- and outdegree on the 217 topics of the Wikipedia collection

	MAP		MRR	
Run id	**Glob**	**Loc**	**Glob**	**Loc**
baseline	0.3090		0.8121	
in	0.3018^{-}	0.3190$^{\circledcirc}$	0.8139^{-}	0.8236^{-}
out	0.3016^{-}	**0.3199**$^{\bullet}$	0.8262^{-}	0.8266^{-}
log.in	0.2865^{-}	0.3176$^{\bullet}$	**0.8322**$^{\circ}$	0.8289°
log.out	0.2890^{-}	0.3156$^{\bullet}$	0.8291$^{\circledcirc}$	0.8225°

inlinks and outlinks are similar in character, leading to the conflation of the notions of authority and hub [4].

In our retrieval experiments, we wanted to know what the impact is of link evidence on retrieval. For the Web track collection, all global and local outdegree priors are less effective than the corresponding indegree priors, supporting the claim that document importance is a major aspect in Web retrieval. Global indegree is more effective for early precision, which is important for Web search.

For the Wikipedia collection, the outdegree priors behave very similar to the indegree priors. The brute force of the global degree priors is too much for the task of ad hoc retrieval. Even the more subtle log degree prior is not effective for MAP. The local degrees stay more on topic and can improve early and later precision, showing that link evidence has to be carefully weighted and made sensitive to the local context.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO), CATCH programme, under project number 640.001.501.

REFERENCES

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

[2] D. Hawking. Overview of the TREC-9 web track. In *TREC*, 2000.

[3] J. Kamps and M. Koolen. Is Wikipedia link structure different? In *Proceedings WSDM 2009*, pages 232–241. ACM, 2009.

[4] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[5] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings SIGIR 2002*, pages 27–34. ACM, 2002.

[6] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings KDD '05*, pages 177–187. ACM, 2005.

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