The Importance of Link Evidence in Wikipedia

Jaap Kamps [a,b] Marijn Koolen [a]

[a] *Archives and Information Studies, University of Amsterdam, The Netherlands*
[b] *ISLA, University of Amsterdam, The Netherlands*

Abstract

The full paper appeared as: Jaap Kamps and Marijn Koolen. The importance of link evidence in Wikipedia. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Rutven, and Ryen W. White, editors, *Advances in Information Retrieval: 30th European Conference on IR Research (ECIR 2008)*, volume 4956 of *Lecture Notes in Computer Science*, pages 270–282. Springer Verlag, Heidelberg, 2008.

1 Introduction

Wikipedia is one of the most popular information sources on the Web. The free encyclopedia is densely linked. The link structure in Wikipedia differs from the Web at large: internal links in Wikipedia are typically based on words naturally occurring in a page, and link to another semantically related entry. Our main aim is to find out if Wikipedia's link structure can be exploited to improve ad hoc information retrieval. We first analyse the relation between Wikipedia links and the relevance of pages. We then experiment with use of link evidence in the focused retrieval of Wikipedia content, based on the test collection of INEX 2006.

2 Link Evidence in Wikipedia

2.1 Analysis

We have conducted an extensive analysis of Wikipedia link structure. Figure 1 shows the global link indegree

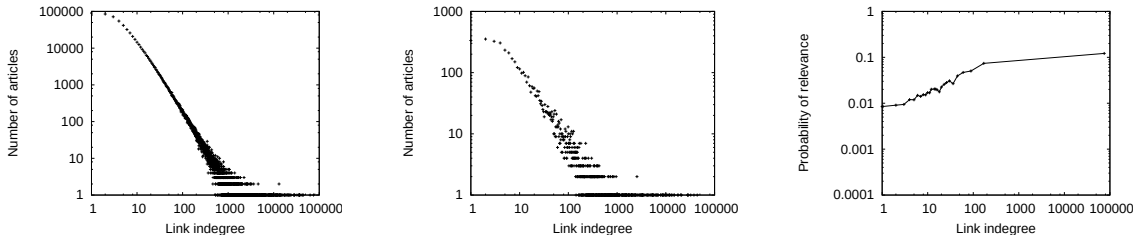

Figure 1: Wikipedia link degree distribution off all pages (left), of relevant pages (middle) and prior probability of relevance (right)

distribution of all pages (left), and of all pages relevant for an INEX 2006 topic (middle), allowing us to calculate the prior probability of a page being relevant given its indegree (right). We clearly see an increasing curve, suggesting that link evidence can be used as (possibly weak) indicator of relevance. For retrieval we have to combine content-based retrieval scores with a score based on the link topology. The crucial issue when incorporating link evidence is to retain the focus on the topic at hand, and avoid the retrieval of important but off-topic pages.

2.2 Priors

On top of a standard language model for information retrieval, we implemented a range of link evidence priors making the link evidence increasingly sensitive to the local context. First, we use a *standard indegree*

prior by multiplying the retrieval score with $1+$ the indegree:

$$P_{\text{standard}}(d) \propto 1 + indegree(d).$$

Here, the indegree score for a page may be based on either the *global* link graph or the *local* link graph restricted to pages with a content-based retrieval score. Second, we use a *log indegree prior* using the logarithm of the indegrees:

$$P_{\log}(d) \propto 1 + log(1 + indegree(d)).$$

The logged indegree values will reduce the impact of the indegrees and hence may act as a safe-guard against the infiltration of loosely related pages with very high (global) indegrees. Again, the indegree score may be based on *global* or *local* evidence.

We experiment with weighting the local indegree (the number of links from pages in the relevant set) by its global indegree (the number of links from arbitrary pages). That is, our third prior is a combination of the local *and* global link evidence computed as:

$$P_{\text{LocGlob}}(d) \propto 1 + \frac{indegree_{\mathsf{local}}(d)}{1 + indegree_{\mathsf{global}}(d)}.$$

This is similar to the well-known *tf.idf* weighting scheme used to determine term importance.

2.3 Results

The scores for three INEX 2006 ad hoc retrieval tasks are in Table 1. We see that the global link evidence leads to a loss of performance, it tends to favor important but off-topic pages. The more conservative local link evidence fares much better, with the less aggressive logged version leading to improvement on two of the tasks and a small loss on the third. The even more careful combined local/global indegree prior is effective on all tasks.

Table 1: Results of link evidence on three INEX 2006 ad hoc retrieval tasks. Best scores are in bold-face. Significance levels are 0.05 (*), 0.01 (**), and 0.001 (***).

Run ID	Thorough MAep,off		Focused nxCG@10,off		Relevant in Context MAgP	
Baseline	0.0353		0.3364		0.1545	
Global Indegree	0.0267	-24.40***	0.1979	-41.16***	0.1073	-30.57***
Log Global Indegree	0.0335	-4.99	0.3066	-8.87**	0.1352	-12.50***
Local Indegree	0.0405	+14.75*	0.3218	-4.34	0.1467	-5.02*
Log Local Indegree	0.0418	+18.46***	0.3460	+2.85	0.1515	-1.96
Local/Global Indegree	**0.0463**	+31.08***	**0.3629**	+7.88**	**0.1576**	+1.99*

3 Findings

Our main findings are: First, our analysis of the link structure reveals that the Wikipedia link structure is a (possibly weak) indicator of relevance. Second, our experiments on INEX ad hoc retrieval tasks reveal that if the link evidence is made sensitive to the local context we see a significant improvement of retrieval effectiveness. Hence, in contrast with earlier TREC experiments using crawled Web data, we have shown that Wikipedia's link structure can help improve the effectiveness of ad hoc retrieval.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO, grants # 639.072.601, 612.066.513, and 640.001.501), and the E.U.'s 6th FP for RTD - project MultiMATCH contract IST-033104.

References

[1] Jaap Kamps and Marijn Koolen. The importance of link evidence in Wikipedia. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Rutven, and Ryen W. White, editors, *Advances in Information Retrieval: 30th European Conference on IR Research (ECIR 2008)*, volume 4956 of *Lecture Notes in Computer Science*, pages 270–282. Springer Verlag, Heidelberg, 2008.