What's in a Link?
From Document Importance to Topical Relevance

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

[1] Archives and Information Studies, University of Amsterdam,The Netherlands
[2] ISLA, University of Amsterdam,The Netherlands

Abstract. Web information retrieval is best known for its use of the Web's link structure as a source of evidence. Global link evidence is by nature query-independent, and is therefore no direct indicator of the topical relevance of a document for a given search request. As a result, link information is usually considered to be useful to identify the 'importance' of documents. Local link evidence, in contrast, is query-dependent and could in principle be related to the topical relevance. We analyse the link evidence in Wikipedia using a large set of ad hoc retrieval topics and relevance judgements to investigate the relation between link evidence and topical relevance.

1 Introduction

Web information retrieval is best known for its use of the Web's link structure as a source of evidence. PageRank [11] is a query-independent algorithm that measures document importance on a global level and is not concerned with a topical relation to the query at hand. The alternative is to analyse the link structure of local sets of documents—e.g., the initial text-based results—to identify topically authoritative pages for broad topics [1, 7]. What is the value of links in topic relevance tasks? This question was addressed by constructing an IR test collection during the 1999 Small Web Task at TREC [13], where participants tried to answer the question "whether hyperlink information could be used to improve ad hoc retrieval effectiveness" [4]. The results from the experiments failed to demonstrate the value of link information for ad hoc retrieval.

Arguably, the notion of what is relevant for typical Web searches is different from the traditional IR interpretation of a document containing text relevant to a precisely defined information need. New Web-oriented tasks were designed to better reflect Web search behaviour. In these Web tasks, the goal was to identify entry pages to particular sites (in the case of home page finding and topic distillation) or another important document (in the case of named page finding). These tasks also dictated a different notion of relevance [12]. The experiments showed that, although links were not effective for singling out the documents with topically relevant textual content, they are useful for locating the documents that are important for these Web-oriented tasks. This leads to our main research question:

- To what extent is link evidence related to the importance of documents, and to the topical relevance of documents?

Here, global and local link evidence seem to play different roles. Links are also directed, and link evidence is typically used for the documents they point *to*, i.e., inlinks. Thinking of incoming links as some sort of vote, inlinks are attractive to measure document importance. However, insofar as a link is evidence that the two documents it connects are topically related, the direction of the link seems not to matter. Topical relatedness works both ways.

In this light, Wikipedia is an interesting data source to investigate the value of links. It is one of the most popular web sites and, being an encyclopedia, it contains entries on single topics, that are densely linked to related content. It is also a natural source for informational search, where it makes sense to study topical relevance aspects of links. Moreover, an extensive IR test collection based on Wikipedia is available thanks to the INEX Ad hoc Tracks of 2006 to 2007. Clearly, the Wikipedia differs considerably from the Web at large, and even the links in Wikipedia are different. We make no particular claims on the representativeness of the Wikipedia for the general Web. Still, the same link-related phenomena (global and local, incoming and outgoing links) are present, and looking at the Wikipedia allows us to study them in great detail. The INEX Ad hoc test collection allows us to study the impact of query-dependent and query-independent link evidence with respect to the topical relevance of retrieval results. In fact, because the INEX test collections are constructed to study the effectiveness of focused retrieval, we have exact information on where and how much relevant text is in each article.

We will first analyse the effectiveness of link evidence for ranking retrieval results, addressing the questions:

- What are the characteristics of Wikipedia link structure?
- How do global and local link evidence impact retrieval effectiveness?
- How do incoming, outgoing and undirected links impact retrieval effectiveness?

Then, we look at how related the different types of link evidence are:

- How do incoming, outgoing and undirected link evidence correlate?
- How is link evidence related to the amount of relevant text in articles?

The rest of this paper is structured as follows. In Section 2 we discuss related work. After discussing the experimental data in Section 3, we compare the different types of link evidence in a retrieval setting in Section 4. Then, in Section 5 we analyse the relation between the different degrees structures and the amount of relevant text in articles. We draw conclusions in Section 6.

2 Related Work

In the TREC Web Tracks of 1999 to 2004, participants were unable to show the effectiveness of link evidence for general ad hoc retrieval [3]. However, it was argued that traditional ad hoc retrieval is very different from how people search on the Web. To study the value of link information, tasks closer to real Web search are required [4]. With tasks adjusted to Web search scenarios, link information proved highly beneficial [8, 10]. This difference in effectiveness of link evidence for these tasks indicates that link evidence does not reflect topical relevance.

Table 1. Link statistics of the Wikipedia collections. Local statistics are macro averages over 221 topics

	Global					Local				
Degree	**min**	**max**	**mean**	**median**	**stdev**	**min**	**max**	**mean**	**median**	**stdev**
Indegree	0	74,937	20.63	4	282.94	0	48.83	3.17	1.14	6.65
Outdegree	0	5,098	20.63	12	36.70	0.04	21.01	3.17	2.34	3.37
Union	0	75,072	37.65	16	287.87	0.04	51.11	5.14	3.14	7.19
Intersection	0	1,488	3.62	2	9.10	0	14.68	1.20	0.44	2.15

Najork et al. [9] compared HITS authorities and hubs with several link-based ranking algorithms – PageRank, Indegree and Outdegree – and found that the choice of algorithm makes little difference on the effectiveness of link evidence. What does have a big impact is the *direction* in which the evidence is used. Although adding evidence based on outgoing links to a content-based retrieval baseline does lead to improvements, it is much less effective than evidence based on incoming links.

Kamps and Koolen [5] showed that in Wikipedia, indegree is related to topical relevance and found that incoming link evidence can be effective for ad hoc retrieval. Later, they found that, unlike in the Web, incoming and outgoing link evidence is equally effective for document retrieval on Wikipedia [6].

3 Wikipedia Link Structure

For the analysis, we use the INEX Wikipedia collection [2], containing 659,304 documents, and a set of 221 topics with relevance judgements from the INEX 2006–2007 Ad hoc Tracks. The union of the in- and outdegree is the undirected degree, or the total number of pages that a page is connected to. The intersection of in- and outdegree is the set of bidirectional links, where pages A and B link to each other. The graph contains 12.4M undirected links and 1.2M bidirectional links (9.5%). We also look at local link evidence—considering only links between the top 100 ranked pages for a given query.

Degree statistics are shown in Table 1. Looking at the global link structure, the maximum indegree is much higher than the maximum outdegree. The maximum and spread of the undirected degree are very similar to those of the indegree, but the median is more similar to that of the outdegree. The bidirectional degree is much lower because only a small proportion (9.5%) of the links are bidirectional. When we look at the local degrees, we see a similar pattern. The indegrees have a bigger spread than the outdegrees, with the undirected degrees having a maximum and spread close to those of the indegrees and a median closer but above that of the outdegrees.

The number of local links is of course smaller than in the whole link graph, but the link density is higher. Globally, a document is connected to 0.0057% of the collection on average, whereas in the local set, it is connected to 5.14%. We also look at the proportion of bidirectional links by looking at the fraction of intersection within union. This proportion is much higher in the local set (23.4%) than in the global set (9.5%). This can be explained, at least in part, by the higher link density in the local set. The nature of Wikipedia links may also play a role: the Wikipedia guidelines on linking [14] state that a link to another document should only be made when it is relevant to the context. Thus, in a set of documents related to the same query, many documents will be related to each other and therefore cross-linked.

Table 2. Impact of link evidence on the INEX 2006 and 2007 Adhoc Track topics. Significance levels are 0.05 ($^{\circ}$), 0.01 ($^{\odot}$) and 0.001 ($^{\bullet}$), bootstrap, one-tailed.

		Global				Local			
	Run	**MAP**	**P@5**	**P@10**	**P@30**	**MAP**	**P@5**	**P@10**	**P@30**
Link only	*Indegree*	14.41	29.68	28.14	24.77	21.20	47.06	41.49	32.17
	Outdegree	13.56	25.70	25.29	24.34	21.46	44.07	41.09	32.96
	Union	14.05	27.96	27.33	24.18	22.26	47.15	42.31	33.92
	Intersection	14.36	30.95	27.56	24.66	20.45	44.43	39.28	30.71
Content only	*baseline*	30.65	55.57	48.91	35.87	30.65	55.57	48.91	35.87
Content+Link	*Indegree*	26.66	50.50	41.90	31.79	31.71$^{\bullet}$	59.00$^{\bullet}$	50.27	36.80°
	Outdegree	27.73	52.13	43.98	32.38	31.83$^{\bullet}$	56.47	49.82	37.12$^{\odot}$
	Union	27.51	50.86	43.89	32.08	32.09$^{\bullet}$	57.83°	50.50$^{\odot}$	37.53$^{\bullet}$
	Intersection	28.41	53.12	45.61	32.87	31.75$^{\bullet}$	57.83°	50.18	37.10$^{\odot}$

4 Link Evidence

In this section, we investigate the impact of link evidence on the effectiveness of ad hoc retrieval. After that, we use standard IR effectiveness measures to evaluate a baseline run using a language modelling framework and runs derived from the baseline but re-ranked 1) using only link evidence and 2) using a combination of content and link evidence.

4.1 Using Only Link Evidence

We show the results in Table 2 and will first discuss the impact of using only link evidence (and not content) for ranking. When re-ranking on global link evidence only, indegree leads to higher early precision than outdegree, which is consistent with the hypothesis that global link structure signals 'important' pages. With the union of the degrees, precision is lower than with indegree alone, but higher than with outdegrees. The intersection of the degrees leads to a higher early precision than the indegree. The intersection creates a symmetric link graph that is still query-independent and seems more effective than indegree alone. Compared to the content-only run, though, the global link degrees are nowhere near as effective.

Although still well below the content-only run, local link degrees give much higher scores than global degrees, indicating that by biasing the link evidence by considering only links between documents related to the query, link information becomes more 'semantic'. The indegrees lead to higher early precision than the outdegrees but, overall, the outdegrees lead to a better ranking. The undirected or union degrees give even higher scores, showing that both individual degrees contribute complementary evidence on the relevance of documents. The scores of the intersection of the degrees are somewhat lower than those of the other degrees, which is probably due to the fact that the number of bidirectional links in the local set is relatively low.

4.2 Combining Link and Content Evidence

We now look at re-ranking using the combination of content and link evidence, which we do by multiplying the retrieval score by a link degree score:

$$P_{\text{Degree}}(d) \propto 1 + \mathsf{Degree}(d) \tag{1}$$

The bottom left part of Table 2 shows the combination of the baseline with the global link evidence. It is clear that the global link evidence universally hurts the baseline performance. The impact of the outdegree is smaller than that of the indegree, which makes sense given the bigger spread of the indegrees (see Table 1). The impact of the union of the degrees is closer to that of the outdegree, whereas the small number of bidirectional links in the local set keep the negative impact of the global link evidence small. Both in isolation and in combination with content evidence, global link degrees fall short of the performance of the content-only baseline.

Using local evidence (bottom-right part of Table 2), both indegree and outdegree can significantly improve the baseline run. Although the indegree gives bigger improvements in early precision and the outdegree gives bigger improvements further down the ranking, at P@30, their overall improvements are very similar. Links in Wikipedia can be used effectively in both directions as evidence to re-rank retrieval results. We then expect that ignoring the direction of links and counting the number of connections to other documents in the local set will lead to even better performance. The scores indeed show further, albeit small, improvements. Precision at rank 5 is higher than for the outdegrees, and later and overall precisions are higher than with both in- and outdegrees. When we use only the smaller set of bidirectional links, the results are still surprisingly good. With less than a quarter of the total links, the intersection of the degrees gives the same performance boost as the in- and outdegrees individually.

To summarise, global link evidence may be an indicator of document importance, but fails to help locate topically relevant documents to a specific information need. Local link evidence fares much better. In isolation, it gives much better performance than global link evidence, although it cannot compete with content-based evidence. In combination with this content-only baseline, it does lead to improvements. In fact, this combination is effective, whether we use only incoming links, outgoing links or their union or intersection. This result supports our intuition that for topical relatedness, the direction of links is of no importance. But in- and outdegrees affect the ranking differently. In what way do incoming and outgoing link evidence differ from each other? We address this question in the next section.

5 Relation between Degrees

In this section we analyse the extent to which degrees are correlated to each other. The main difference found so far is between global and local link evidence. The differences between incoming, outgoing, undirected and bidirectional link degrees are relatively small. A simple explanation would be that all these degrees are strongly correlated. Incoming and outgoing link evidence are necessarily related in some way: a link between two documents is incoming link evidence for one document and not the other, and vice versa for outgoing link evidence.

5.1 Correlation of Degrees

We computed rank correlations (Kendall's Tau) between the four degree types over the entire collection, and within the local set of retrieved results for all 221 topics (Table 3). Over the entire collection, the in- and outdegree are moderately correlated, and

Table 3. Rank correlations between global, top 100 and top 10 local degrees

	Global				Top 100				Top 10			
Degree	*In*	*Out*	*Union*	*Inter*	*In*	*Out*	*Union*	*Inter*	*In*	*Out*	*Union*	*Inter*
In	–	0.41	0.59	0.66	–	0.49	0.71	0.77	–	0.30	0.77	0.43
Out		–	0.83	0.46		–	0.82	0.58	0.13	–	0.47	0.24
Union			–	0.50			–	0.59	0.63	0.32	–	0.37
Inter				–				–	0.48	0.49	0.45	–

the undirected degree is very strongly correlated with outdegree and less strongly with the indegree. This means that, on a global level, the outdegree is the dominant factor in the undirected degree. The same holds for the bidirectional degree. The intersection correlates most strongly with the indegree. Over the local top 100 link graphs, using the average of the correlations of the 221 topics, the correlation between in- and outdegree is stronger, and thereby, their correlation with the union is more similar. An explanation might be the higher percentage of bidirectional links in the local sets. The overall correlations give a broad idea of the relationship between degrees. Given that most documents have a low in- and outdegree, the correlation is dominated by these low degrees while we are mostly interested in the other end with the highest degrees.

In Table 3 we also show the correlations between degrees over the top 10 results. That is, we take the top 10 results ranked by the column (say, indegree) and compare their ordering with how they are ranked by the row (say, outdegree). Note that over the top 10, the correlation is not symmetrical: the top 10 documents by indegree can be different from the top 10 documents by outdegree. Over the top 10, the rank correlation between indegree and outdegree is lower than over the top 100. The top 10 ranking by indegree corresponds better to their ranking by outdegree than the top 10 ranking by outdegree corresponds to their ranking by indegree. The average overlap between the two sets of top 10 documents is 4.7, thus each has 5.3 documents in the top 10 that are not in the top 10 of the other. Over the top 100, the outdegree correlates stronger with the undirected degree than the indegree, but over the top 10, it is the other way around. This is reflected in the precision scores in Table 2. The undirected degree has an early precision very similar to that of the indegree, while further down the results list, its precision is closer to that of the outdegree. The correlations with the intersection are much lower over the top than over the top 100, probably because of the lower degrees.

5.2 Correlation of Degree and Relevant Text Size

In Section 3, we saw that all four types of local link evidence show some relation to topical relevance, as evidenced by their positive effect on performance when combined with the content score. But not all documents are equally relevant. Some documents might be mostly off-topic and only mention the topic in a few sentences, while others might be fully on-topic and cover the topic exhaustively. For the INEX Ad hoc Track, assessors are asked to highlight in yellow all and only relevant text within each pooled document. This allows us to study the relation between link evidence and the amount of relevant text. We assume that documents that have more relevant text discuss the topic more exhaustively and are therefore more important to the topic.

Figure 1 shows the average amount of relevant text over the first 10 retrieved relevant documents when ranked by degree. The left-hand-side shows the content-only and

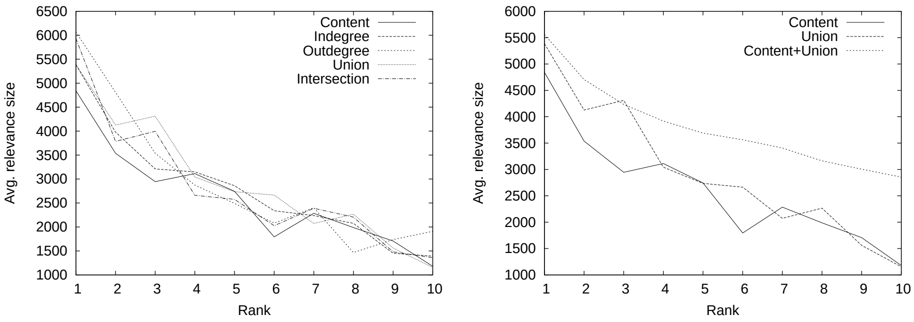

Fig. 1. The average amount of relevant text at ranks 1 to 10 for the retrieved relevant documents ranked by content or link degree

link-only evidence. The right-hand-side shows the content and undirected link evidence and their combination. We see that the amount of relevant text decreases over rank for all types of evidence. The content-only evidence has the lowest amount of relevant text at rank 1, and the outdegree the highest. In the set of retrieved relevant documents, link evidence seems to be a good indicator of the amount of relevant text in a document. This makes sense if local link evidence is related to topical relevance. More links means more evidence of topical relevance, and thus more relevant to the topic. In the right-hand figure, it is clear that the undirected degree ranking has more relevant text at most ranks, especially at the first 3 ranks. Thus, although the content-only score is a better indicator of the relevance of a document—it has much higher precision and MAP scores in Table 2—the undirected link evidence seems a better indicator of the amount of relevant text in documents. The combination of both types of evidence has a large impact on the amount of relevant text found at all first 10 ranks. This indicates that the relevant articles are ranked more favourably, an important aspect that remains unnoticed by the standard evaluation measures.

6 Discussion and Conclusions

In this paper we investigated the relation between link evidence and topical relevance in Wikipedia. Our main aim was to find out to what extent link evidence is related to document importance and to topical relevance.

The local link structure is more dense than the global link structure and has a larger proportion of bidirectional links, making link evidence more symmetrical. Evidence based on incoming links gives better early precision, while outgoing link evidence gives better precision further down the ranking. Taking the union of these two degrees leads to further improvements, showing that in- and outdegrees contribute different information. At the local level, the different degrees all derived from the same link graph exhibit reasonably high correlations, indicating that they promote many of the same documents. However, this correlation is lower in the top of the in- and outdegree rankings. Given the substantial difference between documents in the top 10, in- and outdegree seem to promote different documents. The degrees can also help the internal ranking of the relevant documents by inducing a more favourable ranking in terms of the amount of relevant text in articles. One could think of notions of relevance for Web retrieval extending the traditional topical relevance, for example, by requiring pages to be both 'relevant' in the traditional sense, as well as 'important' or 'authoritative'. Such a view would impose an additional criterion on the topically relevant pages, which is supported by our analysis of the amount of relevant text in documents.

This paper is only a first step in understanding the value of link information. Wikipedia is different from the Web at large, including its links: the Web is much more heterogeneous and noisy, and the creation of Web links is not steered by clear guidelines, nor done for a single purpose. Nevertheless, the distinction between global and local evidence holds for the Web as well. But our analysis of links in Wikipedia has shown that the link structure contains valuable cues about topical relevance.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.513, 639.072.601, and 640.001.501).

References

[1] S. J. Carrière and R. Kazman. Webquery: Searching and visualizing the web through connectivity. *Computer Networks*, 29(8-13):1257–1267, 1997.

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

[3] D. Hawking. Overview of the TREC-9 web track. In *The Ninth Text REtrieval Conference (TREC-9)*, pages 87–102. NIST Special Publication 500-249, 2001.

[4] D. Hawking and N. Craswell. Very large scale retrieval and web search. In *TREC: Experiment and Evaluation in Information Retrieval*, chapter 9. MIT Press, 2005.

[5] J. Kamps and M. Koolen. The importance of link evidence in Wikipedia. In *Advances in Information Retrieval: 30th European Conference on IR Research (ECIR 2008)*, volume 4956 of *LNCS*, pages 270–282. Springer Verlag, Heidelberg, 2008.

[6] 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)*. ACM Press, New York NY, USA, 2009.

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

[8] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, pages 27–34. ACM Press, New York NY, USA, 2002.

[9] M. Najork, H. Zaragoza, and M. Taylor. Hits on the web: How does it compare? In *SIGIR '07*, 2007.

[10] P. Ogilvie and J. Callan. Combining document representations for known-item search. In *Proceedings of the 26th Annual International ACM SIGIR Conference*, pages 143–150. ACM Press, New York NY, USA, 2003.

[11] 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.

[12] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26:321–343, 1975.

[13] TREC. Text-REtrieval Conference, 2009. `http://trec.nist.gov/`.

[14] Wikipedia. Linking, 2009. URL `http://en.wikipedia.org/wiki/Wikipedia:Linking`.