The meaning of structure: the value of link evidence for information retrieval
Koolen, M.H.A.

Citation for published version (APA):
Koolen, M. H. A. (2011). The meaning of structure: the value of link evidence for information retrieval
Amsterdam: IR Publications

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Now that we have analysed the nature of link evidence in Wikipedia, we increase our scope to the Web at large. In Chapter 4 we already looked at link evidence for Web-centric search tasks. For Web-centric tasks, global link evidence is very effective and needs no curbing, nor to be made sensitive to the topical context. In this chapter we look at ad hoc retrieval, and compare the impact of link evidence on the Web with our findings on Wikipedia for the same search task. In 2009, a new TREC Web Track was organised, using a large collection of Web pages called ClueWeb09 and a new set of topics and relevance judgements, which allow us to address the issue of Web ad hoc retrieval on a collection that is arguably more representative of the real Web than the wt2g and wt10g collections used for previous Web Ad Hoc tasks (see Table 2 on page 34). The links in Wikipedia are a special case of hyperlinks in the Web. Our findings about the value of link evidence might either be specific to the hyperlinks in Wikipedia or apply to hyperlinks in general. We used Wikipedia because its closed domain, dense link structure, categorical organisation and the availability of high-quality IR test collections allowed us a detailed analysis of the nature of link evidence. We found that local link evidence is more effective for ad hoc search than global link evidence and that for local link evidence, the direction of the links is not important. Do these findings also hold in general? In this chapter, we put our findings to the test on the larger and more general structure of links on the Web and address the question:

• To what extent does the value of link evidence in Wikipedia hold for link evidence in general?

In the course of writing this thesis, a new Web information retrieval test collection, called ClueWeb09, has been constructed based on a recent crawl of the Web with ad hoc search topics and relevance judgements. This new collection is meant to be a good representation of the first tier of the highest-quality pages in commercial search engine indexes and allow a better evaluation of Web search. There are two versions of the collection. ClueWeb09 Category A consists of 1 billion Web pages in various languages, the English pages making up roughly half of the collection. ClueWeb09 Category B contains a subset of 50
million pages in English, which were the first 50 million pages crawled. In this chapter, we use the Category B collection.

The new Web ad hoc retrieval test collection allows us to study the impact of link evidence on a more recent collection that should better resemble the collection of pages in the first tier of commercial search engine indexes. In the previous chapters we have established the value of link evidence for Wikipedia ad hoc search. We conduct experiments using link evidence for Web ad hoc search and compare the impact to the following findings in Wikipedia:

- **Global link evidence needs toning down, local does not**: The global degrees lead to heavy infiltration and can only improve mean average precision when we tone down their impact by using the log global degrees. The local degrees are more sensitive to the topical context and become less effective when toned down.

- **Local link evidence is more effective than global link evidence**: In Wikipedia, query-independent link evidence is less effective than query-dependent link evidence.

- **Local link evidence works in both directions, global link evidence does not**: In Chapter 4 we saw that log global in-degrees lead to a small but significant improvement in MAP while log global out-degrees have almost no impact on MAP. The local in- and out-degrees lead to very similar, significant improvements in MAP.

We then compare the new collection with older TREC Web collections in terms of the density and degree distribution of the link structure. At TREC 1999–2001, participants identified a number of aspects of hyperlinks that could affect the value of link evidence for retrieval, such as the number and density of links, and the relative importance of intra-server and inter-server links. Intra-server links are links between pages on the same server or Web site, while inter-server links are links from a page on one server to a page on a different server. To support meaningful experiments with hyperlinks and algorithms such as HITS and PageRank, test collections should have a large enough number of inter-server links (Bailey et al., 2003). As mentioned in Section 4.3.4.1, site-internal links are considered less useful as indicators of authority (Kleinberg, 1999) while Davison (2000) showed that linked pages on the same server are more similar to each other than linked pages on different servers. In this chapter, we address a number of more specific research questions:
• How has the Web changed in the decade between the TREC Web tracks of 1999–2001 and 2009?

• How is the effectiveness of link evidence affected by the density of the link graph?

• What is the relative impact of inter-server and intra-server links on the effectiveness of link evidence?

One important difference between the new collection and earlier TREC Web collections is that the new collection contains the full English Wikipedia, in which we now know that link evidence is effective. Differences observed between ClueWeb09 and earlier TREC Web collections might be due to the presence of Wikipedia. Therefore, another question we will address is:

• What is the impact of Wikipedia on the effectiveness of link evidence for Web ad hoc search?

This chapter is organised as follows. In Section 7.1 we describe the new Web collection and look at the relation between link degrees and the relevance of retrieval results. Then, in Section 7.2 we describe our experiments with using link evidence for the TREC 2009 Ad Hoc task, and in Section 7.3 we seek to understand the nature of link evidence in the new collection and address the research questions above in turn. We draw conclusions in Section 7.4.

7.1 THE CLUWEB09 COLLECTION

The ClueWeb09 collection (CMU-LTI, 2009), consists of over a billion Web pages in several languages and contains some 25 terabytes of data. Because this amount of data might be too much to process for some research groups to participate in the TREC Web Track, a smaller, 10% subset of the collection is provided as an alternative. This subset, the ClueWeb09 category B collection consists of the first 45 million English Web pages of the crawl, and around 5 million pages representing the full English Wikipedia, which was crawled separately. In this chapter we use the category B collection, referred to as ClueWeb09 B.

The ClueWeb09 B collection was used for the TREC 2009 Web Track, which consisted of two tasks: the traditional Ad Hoc search task and the Diversity task. The ad hoc task is similar to the INEX Ad Hoc task on which the experiments in the previous chapters are based, and to the Ad Hoc task of the TREC 1999–2001 Web Tracks. The Diversity task
uses the same set of topics as the ad hoc task, but has a list of subtopics for each of those topics and challenges participants to develop retrieval techniques that return a list of diverse search results that cover relevant information for different aspects and interpretations of search queries.

### 7.1.1 Relevance judgements

Apart from the changes in the Web, any possible difference between the results in 1999–2001 and 2009 might be due to a difference in the setup. Although both have ad hoc search tasks, there might be differences in the numbers of pages judged, the specific judgement criteria (what counts towards the relevance of a page), and the types of queries and topics.

The 50 topics of the TREC 2009 Web Track were sampled from the query log of a real search engine (Clarke et al., 2009), with a preference for topics of medium popularity. Highly popular queries were assumed to be navigational and therefore less challenging, and rare queries were considered inappropriate because they may contain personally identifiable information.

For each of these topics, the relevant documents are identified by pooling the retrieved results of systems participating in the track and judging their relevance. Because of the large size of the ClueWeb09 collection, relevance judgements are necessarily incomplete. Two recently developed pooling strategies have been used to make evaluation over very large collections possible.

1. With the Minimal Test Collections (mtc) pooling strategy (Carterette et al., 2006), documents are selected for judgement that are most likely to determine the difference between two participating systems. The relevance judgements are meant to determine of any pair of participating systems which one is better than the other. Relevance judgements for documents ranked similarly by both systems are less useful, as they will not help identify which system is best. Therefore, documents are typically picked that are ranked very differently by the two systems. This pooling strategy is very sensitive to the particular systems contributing to the judgement phase and might be inappropriate for non-contributing systems that produce different rankings.

2. The second pooling strategy (Aslam et al., 2006) is based on the assumption that most relevant documents will be found in the top of the ranking, and samples documents from different parts (strata)
of the ranking using different sampling rates for each part. In the top 100 results the sampling rate is higher than in the results at ranks 100–1000. The sampling rate is then used to determine how many documents a pooled document represents. A document with a low sampling probability—meaning it was found far down the results list—represents more documents than a document with a high sampling probability.

Both pooling strategies use probabilities to determine how many documents a judged document represents, and both strategies skip many top ranked documents. Because we want to evaluate runs that did not contribute to the judgement pool, many of the top ranked results may not be judged. This might be problematic for our analysis of the impact of link evidence on reranking the top 100 results. The unknown relevance status of highly ranked documents might stop us from properly measuring qualitative differences between two results lists. To alleviate this possible problem, we can combine the Ad Hoc judgements of category B with the TREC 2009 Diversity relevance judgements. The Diversity judgements are based on the same topics and collection (Clarke et al., 2009), but on a different pooling strategy, namely, the traditional method of pooling the top \( n \) results of all officially submitted runs. In this case, the top 20 results of all runs were pooled and judged for relevance of a number of subtopics. Although the list of subtopics is not exhaustive, such that a document might be relevant for the overall topic but not for any of the subtopics, the extra judgements give us better information about the relevance of the top ranked results of our runs.

We evaluated all results with both the Ad Hoc relevance judgements and the combined relevance judgements, and found no qualitative differences. Systems doing better for one set of judgements tend to do better for the other set of judgements as well. This shows that the unjudged results introduce no bias in the evaluation. We use the category B documents and judgements in the experiments described below.

7.1.2 Degree distribution

We extracted over 1.5 billion links pointing to pages within the collection. Quite a large number of pages have multiple links to the same target URL (repeated links). If we collapse those repeated links and ignore self-referencing links (a link from a page to itself), we end up with 1.16 billion links between just over 50 million pages (see Table 31), which leads to a mean in-degree of 23.12. The median in-degree is 2,
the median out-degree is 9. There are 35.7 million pages (71%) with at least one incoming link and 43.7 million pages (87%) with at least one outgoing link.

We only look at collection-internal links, that is, links from pages in the collection to other pages in the collection. Because the collection contains only a small part of the entire Web, many of the outgoing links of ClueWeb09 pages point to pages outside the collection. Therefore, the out-degree is lower than the actual out-degree of those pages. The same holds for the in-degree, as pages outside the collection may have links to pages in the collection as well. Wikipedia is a notable exception to this. Although it does have links to pages outside Wikipedia, the vast majority of its links are internal links to other Wikipedia pages. The English Wikipedia was crawled separately and is included in its entirety. The out-degrees of the Wikipedia pages will be much closer to their actual out-degrees.

The in- and out-degree distributions are shown in Figure 26. The straight vertical line at the high end of the in-degrees shows that a small
number of pages have the same high in-degree. These are navigational and administrative pages in Wikipedia such as the Wikipedia main page, "Contact us", "Special:Random", "General_disclaimer" and portal pages. Every single Wikipedia page has a link to these pages, explaining why they have roughly the same in-degree.

Are the link degrees also related to the relevance of ad hoc retrieval results? The prior probability of relevance over in- and out-degrees is shown in Figure 27. Similar to the plots in Chapter 4, the in-degrees show a clear relation with relevance. The out-degrees also show a relation with relevance, although the probability does not increase all the way with out-degree. The maximum out-degree is 1,929, but the probability of relevance peaks at 500. Does this mean global link evidence can be used to improve ad hoc retrieval effectiveness in ClueWeb09 B? And is in-degree more effective than out-degree?

7.2 CLUEWEB EXPERIMENTS

We used Indri (Indri, 2009) for indexing. Stopwords are removed and all other terms are stemmed with the Krovetz stemmer. The main index is a standard full-text index.

For the full-text runs we again use a language model approach and linear smoothing. However, ad hoc search in large collections requires little smoothing (Kamps, 2006a). Therefore, we adjust the weight of the
smoothing parameter to $\lambda = 0.85$. That is, in the language modelling formula, the document model $P(q|d)$ has weight $\lambda = 0.85$ and the background model $P(q|D)$ has weight $1 - \lambda = 0.15$. We experimented with different smoothing values and found that on the ClueWeb09 B collection, the specific value for $\lambda$ has little impact on the results. There are many documents matching all the words in the short Web queries used for the TREC 2009 Web Track, so the smoothing value has little effect in the top ranked documents. In Chapter 4 we saw that document length is not an effective document prior for Web-centric search tasks, but for ad hoc search, longer documents have a higher prior probability of relevance. The ClueWeb09 collection contains only Web data, so we first compare baseline runs with and without a document length prior. Recall that when document length is used as a prior, documents are scored using the document length as a prior probability $p_{\text{length}}(d) = \frac{|d|}{\sum_{d' \in D} |d'|}$, where $d$ and $d'$ are documents in collection $D$.

There are two sets of effectiveness measures used for the TREC 2009 Web Track: the expected precision measures based on the MTC pooling strategy and the statAP (Yilmaz and Aslam, 2006) measures based on the stratified sampling strategy (see Section 7.1.1). We use statAP, as it is more robust than the MTC-based measures when evaluating systems that did not officially participate in the Web Track and therefore did not contribute to the pool. We test for significant changes with respect to the full-text baseline using a one-tailed bootstrap test with 100,000 resamples.

The first two lines in Table 32 show the baseline results with ($\beta = 1$) and without ($\beta = 0$) length prior. As with ad hoc search on other collections, document length is related to relevance (Singhal et al., 1996). We conduct our link-based experiments on the $\beta = 1$ baseline. Recall that the union of in- and out-degree is the degree derived from an undirected version of the link graph. That is, bidirectional links count as one link. One notable difference with earlier Web tracks is that the full-text baseline on ClueWeb09 has a low early precision. Is this caused by spam or by the much larger size of the Web, with many low quality pages? Several Web Track participants found that spam filters and page quality scores improved performance (Cormack et al., 2010, Hauff and Hiemstra, 2010, Lin et al., 2010, Rajput et al., 2010), suggesting that standard text retrieval approaches indeed suffer from spam and low-quality pages in ClueWeb09.

We first look at the impact of the global degrees as true document priors over all retrieved results (rows 3–5). This leads to heavy infilt-
### 7.2 ClueWeb Experiments

<table>
<thead>
<tr>
<th>Reranked ID</th>
<th>baseline $\beta = 0$</th>
<th>baseline $\beta = 1$</th>
<th>All results</th>
<th>Top 100</th>
<th>Log Global in</th>
<th>Log Global out</th>
<th>Log Global union</th>
<th>Local in</th>
<th>Local out</th>
<th>Local union</th>
<th>Log Local in</th>
<th>Log Local out</th>
<th>Log Local union</th>
</tr>
</thead>
<tbody>
<tr>
<td>statMAP</td>
<td>0.0991</td>
<td>0.1442</td>
<td>0.0182</td>
<td>0.1417</td>
<td>0.1449</td>
<td>0.1563</td>
<td>0.1552</td>
<td>0.1514</td>
<td>0.1562</td>
<td>0.1575</td>
<td>0.1497</td>
<td>0.1561</td>
<td>0.1565</td>
</tr>
<tr>
<td>MFC(30)</td>
<td>0.2208</td>
<td>0.3079</td>
<td>0.2632</td>
<td>0.4488</td>
<td>0.3344</td>
<td>0.4689</td>
<td>0.3786</td>
<td>0.3704</td>
<td>0.4114</td>
<td>0.4210</td>
<td>0.3836</td>
<td>0.4049</td>
<td>0.4202</td>
</tr>
<tr>
<td>MAP</td>
<td>0.0932</td>
<td>0.1516</td>
<td>0.1090</td>
<td>0.1347</td>
<td>0.1423</td>
<td>0.1545</td>
<td>0.1541</td>
<td>0.1520</td>
<td>0.1628</td>
<td>0.1623</td>
<td>0.1542</td>
<td>0.1660</td>
<td>0.1658</td>
</tr>
<tr>
<td>MRR</td>
<td>0.2982</td>
<td>0.3061</td>
<td>0.1688</td>
<td>0.5008</td>
<td>0.5179</td>
<td>0.5040</td>
<td>0.5217</td>
<td>0.6616</td>
<td>0.5550</td>
<td>0.6395</td>
<td>0.5329</td>
<td>0.4724</td>
<td>0.5183</td>
</tr>
<tr>
<td>P@10</td>
<td>0.2180</td>
<td>0.2780</td>
<td>0.0640</td>
<td>0.2280</td>
<td>0.2380</td>
<td>0.3660</td>
<td>0.3440</td>
<td>0.2940</td>
<td>0.3880</td>
<td>0.3540</td>
<td>0.2980</td>
<td>0.3660</td>
<td>0.3620</td>
</tr>
</tbody>
</table>

Table 32: Results for the 2009 Ad Hoc Task. Significance tests are with respect to the full text baseline, confidence levels are 0.95 ($^*$), 0.99 ($^\circ$) and 0.999 ($^{\circ\circ}$).
ration and is disastrous for early and overall precision, with statMAP going from 0.1442 to 0.0182, 0.489 and 0.0381 using the in-degrees, out-degrees and their union, respectively. If we limit the prior to the top 100 retrieved results (rows 6–8), we see improvements in early precision with in- and out-degrees and their union (MRR goes up from 0.3061 to 0.5008, 0.4347 and 0.5010 respectively), and improvements in overall precision with the out- and union degrees (statMAP goes up to 0.1512 and 0.1511 respectively). This is in direct contrast to the impact of global link evidence on Web-centric search tasks (see Chapter 4, Section 4.3.4), where global link evidence is best used as a true prior probability over all results. Ad hoc search requires more careful use of link evidence. The global in-degree prior over the top 100 results gives an improvement in MRR and MPRC(30), but is not effective for average precision: MAP and statMAP go down. The low early precision of the baseline result and the improvements of the global degrees suggest that the text-based ranking has many low-quality documents at the top and the role of link evidence is to identify and push up the important pages. This fits with the findings of the TREC 2009 Web Track participants. The global out-degree priors (row 7) are more effective than the in-degree priors (row 6), and show improvement on all reported measures. The log global degree priors are even more effective (rows 9–11), and, apart from statMAP, the log global out-degrees are the most effective. Clearly, the number of outgoing links is a good signal for the relevance of a page. But why are pages with many outgoing links more often relevant? In Chapter 4 we saw that in Wikipedia the maximum out-degree is higher than in the .gov collection (see Table 8 on page 75). Perhaps this is also the case in ClueWeb09 B, and the global out-degree favours Wikipedia pages, which might be higher quality documents, and therefore more often judged relevant. We discuss this in Section 7.3.4.

If we look at the local degrees (rows 12–14), we see that especially the in-degrees are effective for improving MRR. The local out-degrees are more effective for all other measures and thus for later and overall precision, consistent with our findings in Chapter 4. We note that this collection includes the full English Wikipedia, which might partly explain this consistency. The log local degrees (bottom 3 rows) only slightly improve traditional MAP but not the other measures. The local degrees need no toning down.

How do these results compare to the impact of link evidence in the INEX 2006 Wikipedia collection?

Global link evidence needs toning down, local does not: The normal, non-log global degree priors hurt performance only when used over all retrieved
results but lead to improvements when used over the top 100 results. In Wikipedia, the non-log global degree priors hurt performance both when applied to all results and when applied to only the top 100 results. Global link evidence is more effective on the Web than on Wikipedia. As in Wikipedia, using the log of the degrees improves the effectiveness of the global degrees but not of the local degrees. In Wikipedia, the normal local degrees were clearly better than the log local degrees. Here the difference is smaller.

*Local link evidence is more effective than global link evidence:* Local link evidence is more effective than global link evidence for statmap, but not for mp C(30). The difference between the impact of local and global link evidence is also smaller. Link evidence for document importance seems more useful in the Web than in Wikipedia, which is possibly due to the large variation in document quality in the Web.

*Local link evidence works in both directions, global link evidence does not:* In Wikipedia, local in-degree has the same impact as local out-degree. Our hypothesis is that local links provide evidence for topical relevance, which is a symmetric relation, so should have the same impact in both link directions. In contrast to Wikipedia, there is a large difference between the impact of local incoming and outgoing link evidence in the Web. In contrast to the findings in Chapter 4, the local outgoing link degrees are more effective than the incoming link degrees in the ClueWeb09 B collection. This suggests that in the ClueWeb09 B collection, local link evidence signals more than just topical relevance. Perhaps the local out-degree is so effective because it is related to the global out-degree and also signals document importance. Again, this might be because the out-degree favours Wikipedia pages.

Why do the in- and out-degrees behave the same way in Wikipedia but differently in the ClueWeb09 collection? And why is link evidence effective for ad hoc search in ClueWeb09 where it is not in the wt10g collection? Perhaps link evidence is more effective than in previous trec experiments because this collection contains the full Wikipedia, where we have seen that link evidence is effective for ad hoc search. Wikipedia pages are edited by many contributors, so the quality might be higher than that of many Web pages. Or perhaps the higher link density makes link evidence more effective. In the next section, we look at several factors of the ClueWeb09 link structure that play a role in the nature of link evidence.
7.3 Why Link Evidence Works in ClueWeb09

Why is link evidence effective for ad hoc search in the ClueWeb09 collection, but not in older collections (see Table 2 in Section 2.3.3)? And why is global out-degree more effective than global in-degree and the local degrees, while our findings in the previous chapters would suggest otherwise? Several aspects have been mentioned as possible factors determining the effectiveness of link evidence.

Differences in the collection: The new collection is bigger, more recent, crawled using a different strategy. All these aspects could affect the value of link evidence in a collection.

The impact of link density: Gurrin and Smeaton (2004) have mentioned collection size and (inter-server) link density. Fisher and Everson (2003) also mention the link density of collections as a crucial factor to make links useful. In the previous chapter we have seen that link evidence can have a positive impact even with relatively small set of links.

The impact of inter-server links: Bailey et al. (2003) describe the construction of the wt10g test collection with the aim of having a larger inter-server link density. What is the relative impact of inter- and intra-server links?

The impact of Wikipedia: Another notable difference between the wt10g and ClueWeb09 collections is the presence of the full English Wikipedia in ClueWeb09. In the previous chapters we have seen that link evidence is effective in Wikipedia, so the positive impact of link evidence on the ClueWeb09 collection might be due to the presence of Wikipedia.

We will discuss each of these aspects in detail.

7.3.1 Differences in the collection

We compare the trec 2009 Web Track with the earlier Web Tracks of 1999–2001 in terms of the collection, the relevance judgements and the link graph. The new ClueWeb09 collection is different from earlier trec Web collections in several ways:

Collection size: The ClueWeb09 B collection (50 million pages) is much larger than the wt10g (1.7 million pages) and .gov (1.2 million pages) collections and twice the size of the .gov2 collection (25 million pages).

Size of Web: The Web has grown from an estimated 320 million pages in late 1997 (Lawrence and Giles, 1999) to tens or hundreds of billions

*Page quality:* The ClueWeb09 collection was planned to reflect the first tier of highest quality Web pages in commercial search engine indexes (Callan et al., 2008), and was crawled in early 2009 (CMU-LTI, 2009) using a seed set of the highest PageRank pages from an earlier crawl (Fetterly et al., 2009b) and a crawling policy that schedules the most important pages to be crawled first (Abiteboul et al., 2003, Fetterly et al., 2009a). The value of link evidence might change with the quality of the pages. The INEX 2006 Wikipedia collection contains the full English Wikipedia, which arguably has little spam, and high-quality pages and links edited by many contributors.

*Average link degree:* The OPCR crawling policy (Abiteboul et al., 2003) determines page importance by counting incoming links. As a consequence, such a crawl leads to a densely interlinked collection of Web pages. As we can see in Table 1 on page 33, the average incoming and outgoing link degree of pages in the ClueWeb09 B collection is higher than in the earlier Web collections. With more links there is more evidence.

*Age:* The TREC Web Tracks of 1999-2000 used the wt2g, wt10g and vlC2 collections which were based on a truncated crawl of the Web of February 1997 (Hawking and Craswell, 2005), only four years after CERN declared the World Wide Web was freely available to anyone (W3C, 2010). The pages in the ClueWeb09 collection were crawled in early 2009, 12 years later. The World Wide Web has changed a lot since then. It has grown immensely in the meantime, as have commercial interests. The link graph is more complete and in some sense more stable.

*Access:* Users access many Web sites via search engines nowadays whereas in the early days access through hyperlinks was more common. Web site authors strive to get their Web pages as high in the search results ranking as possible to draw Web traffic. A lot of effort is put into search engine optimisation (SEO) so that site entry pages are placed high in search results rankings for particular queries, with particular attention to site-internal links structure and anchor text. Companies have analysed what makes sites and pages end up high in search results list. One particular strategy is to have pages in the site link to each other with high quality descriptive anchor text. Whereas in 1997 many authors used terms such as “here”, “next” and “home” as anchor text
to allow users to easily navigate within the site, focus has shifted to using anchor text terms that describe the content of a page so that a user typing the same terms as a query can find that particular page.

*Spam:* Because of the growing commercial interest and much larger number of users, Web spam has become an ever growing problem, with many different forms. The taxonomy of Web spam (Gyöngyi and Garcia-Molina, 2005) identifies two broad classes: text spam and link spam. Much like filtering out low-quality pages, text-based spam can be combated with page importance measures such as PageRank, HITS and Online Page Importance Computation (opic, Abiteboul et al. (2003)).

*Graph evolution:* In Section 4.2.2 we looked at the phases of development of the inex Wikipedia and .gov collection, and observed that both collections are in the final phase, where almost all the pages are connected to the giant component. The numbers in Table 33 show that the .gov, inex Wikipedia and ClueWeb09 B collections have reached the final phase of the evolution of link graphs. The wt10g and .gov2 collections—which were used for the earlier trec Ad Hoc Tracks—have not reached this phase yet. In the previous chapter we saw that the links in Wikipedia are still effective when we randomly remove most of the links, which suggests that even in a relatively undeveloped link graph, link evidence is still effective. But perhaps randomly filtered fully evolved link graphs are different from link graphs in earlier evolutionary phases. Perhaps ad hoc search requires a fully developed link graph that is crawled to focus on the most important pages, for links to be effective.

Differences might also be found in the types of queries used and the relevance criteria given to relevance judges. Above we noted that the queries for the trec 2009 Web track are mid-frequency queries from a search engine query log. If these queries are more general than those of earlier tracks, the number of relevant pages for these queries might also be higher.

We compare the relevance judgements of the trec Web Tracks of 2000–2001 and 2009 in Table 34. The 2000–2001 Ad Hoc topics have around 70,000 judgements (1400 per topic) and around 3000 relevant pages (60 per topic). In contrast, the 2009 Ad Hoc topics have only 13,118 judgements (262 per topic) but 4002 relevant pages (80 per topic). This could mean that the 2000–2001 topics are more specific. On the other hand, theory suggests that precision at a fixed cut-off (and therefore the number of relevant pages pooled from the participating

---

1 Because of the sampled pooling strategy used for the TREC 2009 Web Track, the 4002 judged relevant pages represent 25,036 estimated relevant pages in the entire collection.
systems) increases with collection size because the number of easy-to-find relevant documents increases (Hawking and Robertson, 2003). This might be the reason that document importance is more effective here than in Wikipedia. As Kleinberg (1999) argued, the problem for broad search topics is not finding the relevant pages—which are abundant—but identifying the authoritative pages.

In very large collections with many relevant documents, the local graph might become sparse if too small a number of pages is chosen for the local set. With billions of Web pages, a larger local set might be required to derive meaningful evidence for topical relevance.

### 7.3.2 The impact of link density

What is the impact of link density? One of the hypotheses at TREC was that link evidence was not effective for ad hoc search because the

<table>
<thead>
<tr>
<th>Edition</th>
<th>Topics</th>
<th>Coll. size</th>
<th># Judgements</th>
<th># Relevant</th>
<th># Relevant # Judged</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC 9 (2000)</td>
<td>50</td>
<td>1.7M</td>
<td>70,070</td>
<td>2617</td>
<td>0.0373</td>
</tr>
<tr>
<td>TREC 2001</td>
<td>50</td>
<td>1.7M</td>
<td>70,400</td>
<td>3363</td>
<td>0.0478</td>
</tr>
<tr>
<td>TREC 2009</td>
<td>50</td>
<td>50M</td>
<td>13,118</td>
<td>4002</td>
<td>0.3051</td>
</tr>
</tbody>
</table>

Table 34: Comparison of the topics and relevance judgements of TREC systems 2000–2001 and 2010.
number of (inter-server) links was too low. If there are no links at all, there is no impact of link evidence. With a complete graph (all pages link to all other pages), all pages have the same amount of evidence and link evidence has no impact. For link evidence to be effective, the link density must be somewhere in between. But is the optimal density closer to zero links or to the maximum number of links? As in the previous chapter, we randomly filter out links, this time to study the impact of link density.

Is the difference in effectiveness due to the density of the link graph? In fact, the link density of the ClueWeb09 B collection is lower than that of the wt10g collection. Link density is the fraction of all possible links that is actually present in the graph (Wasserman and Faust, 1994). With a collection of 1.7 million documents and 8 million links (see Table 1), the link density of the wt10g graph (treating all links as undirected) is $5.6 \times 10^{-6}$. The link density of the ClueWeb09 B collection with 50 million pages and 1.16 billion links, is $9.2 \times 10^{-7}$. However, the number of links per page, or average link degree, of the ClueWeb09 B collection is higher than that of the wt10g collection. Pages in the ClueWeb09 B collection have an average in-degree of 23.12, whereas pages in the wt10g collection have an average in-degree of 4.77.

The impact of filtering links on the effectiveness of link evidence is shown in Figure 28. Note that below 21% of the links, the average degree of the ClueWeb09 B collection is lower than that of the wt10g collection (The average degree in the wt10g collection, 4.77, is 21% of the average degree of the ClueWeb09 collection, 23.12). So, if we remove 80% of the links in the ClueWeb09 B collection, the average degree is roughly the same as that of the wt10g. Does link evidence become ineffective in ClueWeb09 if we randomly remove 80% of the links?
We show the impact of random filtering on the most effective global and local degrees, namely the log global out-degrees and the local undirected degrees. The impact on the other degrees is similar. The impact of global link evidence on mrr remains stable as we randomly remove links, up to the last 5% of the links. Link density has almost no impact on the effectiveness of global link evidence for mrr. The impact on the local degree is bigger, as the mrr slowly drops as we remove links and has the same impact as the global out-degrees after 95% of the links are removed. Link evidence is effective for mrr even at the lowest density.

For P@10, we see that the log global out-degrees remain effective when filtering links, and the impact is stable with up to 60% of the links removed. Beyond 60%, the improvement slowly drops but performance remains well above that of the baseline. The local union of in- and out-degrees is clearly affected by filtering. The improvement over the baseline steadily drops to zero as we filter out more links. The local graph is sparse to start with, so at 5% there are very few links left, changing very little in the ranking.

For statmap we see a similar pattern as for P@10, although the local union of in- and out-degrees is more effective than the log global out-degrees without filtering. However, filtering has a bigger impact on local link evidence than on global link evidence, and as a consequence, local link evidence becomes less effective than global link evidence when more than 30% of the links are removed. The global link graph is very rich and many links can be removed before the structure falls apart and becomes meaningless. The local link graph is sparse and, with most links removed, carries almost no information to aid the ranking of retrieval results. However, even a very sparse local link graph can be enough to improve overall precision.

In sum, link density plays almost no part in the impact of global link evidence on the entire collection. Even a small set of links is enough to distinguish the important pages from the rest. In contrast, link density plays an important part in the impact of local link evidence. Local link evidence becomes less effective on all three measures as we randomly remove links. The impact is bigger for measures that take into account a larger part of the ranking, like statmap. If the global link graph is sparse, there is not enough local link information to have any impact beyond the first ranked documents. However, even the smallest samples contain enough links to improve retrieval performance with local link evidence.
Can link density explain why link information was not useful for Web ad hoc search in the \textit{wt10g} collection? With 5\% of the links the average degree is just above one, whereas in the \textit{wt10g} collection where links were not effective, the average degree is almost 5. Link density alone cannot fully explain why links are effective in the ClueWeb09 B collection, but might be part of the explanation.

7.3.3 \textit{The impact of inter-server links}

Another hypothesis brought forward at \textsc{trec} 8 for the disappointing results of using hyperlinks for retrieval is that the \textit{wt2g} collection has very few inter-server or site-external links. Links across sites are considered to be more meaningful than links between pages on the same site. The reasoning behind this is twofold. First, similar to the distinction between incoming and outgoing links, it is harder for a Web author to increase the number of links to her own page(s) from other sites than from pages under her control. Second, links within a site often serve a purely navigational purpose, such as links to the entry page and links in a navigation bar that allow users to quickly jump to the part of the site they are interested in. Perhaps the inter-server link density is high enough in the ClueWeb09 collection and is the determining factor in the effectiveness of link evidence. On the other hand, the internal links in Wikipedia are also intra-server or site-internal links, and have been proven meaningful in the previous chapters. Assuming that a single Web site is created and modified by a single author or group of cooperating authors, these authors have full control over the link structure on a global level, just as in Wikipedia. In this section, we want to find out:

- What is the relative impact of inter-server and intra-server links on the effectiveness of link evidence?

There are many more intra-server or site-internal links than inter-server or site-external links in ClueWeb09 B. There are 952 million intra-server links (88\% of the total) and 132 million inter-server links (12\%). If intra-server (site-internal) links are indeed less meaningful, the findings of the previous chapter suggest that the positive impact of link evidence mainly comes from inter-server (site-external) links. We test this by comparing in Table 35 the impact of using link evidence from only inter-server links versus using only intra-server links. We show only the log global degrees as they give the best improvements. The impact of inter-server versus intra-server links is largely the same.
Table 35: The impact of inter- and intra-server link evidence on retrieval effectiveness in ClueWeb09.

<table>
<thead>
<tr>
<th>Links</th>
<th>ID</th>
<th>statMAP</th>
<th>MFC(30)</th>
<th>MAP</th>
<th>MRR</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>0.1442</td>
<td>0.3079</td>
<td>0.1516</td>
<td>0.3061</td>
<td>0.2780</td>
</tr>
<tr>
<td>Intra</td>
<td>Log Global Out</td>
<td>0.1566</td>
<td>0.4818</td>
<td>0.1659</td>
<td>0.5114</td>
<td>0.3680</td>
</tr>
<tr>
<td></td>
<td>Log Global Union</td>
<td>0.1556</td>
<td>0.4701</td>
<td>0.1631</td>
<td>0.5129</td>
<td>0.3580</td>
</tr>
<tr>
<td>Inter</td>
<td>Log Global Out</td>
<td>0.1488</td>
<td>0.3336</td>
<td>0.1573</td>
<td>0.4333</td>
<td>0.3460</td>
</tr>
<tr>
<td></td>
<td>Log Global Union</td>
<td>0.1469</td>
<td>0.3201</td>
<td>0.1547</td>
<td>0.3942</td>
<td>0.3280</td>
</tr>
</tbody>
</table>

on global and local degrees. Although the inter-server links improve performance, the intra-server links lead to larger improvements. Of course, the intra-server links are larger in number, but these results show that the inter-server links are not the main contributors to the effectiveness of link evidence. In fact, the impact of the intra-server links is very similar to the impact of using both inter- and intra-server links.

The inter-server link structure is much more sparse and has less information to distinguish between pages. Site-entry pages are typically the pages with the most incoming inter-server links (Hawking et al., 2004), while most other pages within the same site have no incoming links from other sites. The inter-server links cover only a small part of the collection and mainly the entry pages. When searching information within a single (enterprise) Web site, Hawking et al. (2004) found that site-internal link evidence is effective for improving retrieval performance while site-external link evidence has almost no impact. Ad hoc search is closer to the enterprise search task of searching for information within a corporate Web site than to Web-centric tasks such as home page finding. In home page finding tasks, only the entry pages, which tend to have more incoming (inter-server) links than other pages, can be relevant. In ad hoc and enterprise search, any type of page can be relevant as long as it contains relevant information.

The bias towards intra-server links introduced by the fact that Web site authors can control both the incoming and outgoing links of a page has no negative impact on the value of intra-server links for ad hoc search. This bias is more troublesome for measuring authority and popularity.
7.3.4 The impact of Wikipedia

The English Wikipedia forms a substantial part of the ClueWeb09 B collection. With over 5.7 million pages, it takes up 11% of the collection, and could be the main reason for the effectiveness of link evidence.

To test this hypothesis, we indexed the ClueWeb09 B collection without Wikipedia, removed all links from and to Wikipedia from the link graph and re-ran our experiments. This version of Wikipedia is more recent than the version on which the inex 2006 collection is based, and contains many more pages—although the crawl contains some duplicate pages and redirects. The average degree has also increased. The 5.7 million pages have 446 million links (ignoring repeated links and self-referencing links). All Wikipedia pages have the same navigational bar with links to the main page and a number of other pages, which are not present in the inex 2006 Wikipedia collection. But even without those links, Wikipedia forms a very densely interlinked part of the Web. There are a further 21 million links from Wikipedia to other pages in ClueWeb09 B and 1.5 million links to Wikipedia from the rest of ClueWeb09 B, which we also exclude. The non-Wikipedia part of the Web has 615 million links between 45 million pages. What is the impact of link evidence on ad hoc search in the non-Wikipedia part of the Web?

The results are shown in Table 36. Link degree priors are applied to the top 100 results of the $\beta = 1$ baseline. Performance drops when we remove Wikipedia from the collection. Wikipedia is a high-quality Web site with good informational pages that are often ranked high and contain little to no spam. If we remove them, statMAP drops considerably, from 0.1442 to 0.1044.

However, link evidence still improves the non-Wikipedia baseline. Without Wikipedia, the global degrees (rows 3–5) are effective for early precision—MRR goes up from 0.2814 to 0.3566, 0.3596 and 0.3537 for, respectively, the global in-degrees, out-degrees and their union—but not for statMAP and MAP. Although $\text{mpc}(30)$ is improved, the traditional P@10 drops. The log global degrees (rows 6–8) are more effective than the normal global degrees. The local degrees (rows 9–11) improve on all reported measures. The log local degrees (rows 12–14) further improve normal MAP and P@10, but lead to a lower MRR than the normal local degrees. The presence of Wikipedia cannot explain the positive impact of link evidence on Web ad hoc search.

But these results reveal another interesting factor contributing to the impact of Wikipedia. With Wikipedia removed, local link evidence is
Table 36: Impact of link evidence on the non-Wikipedia part of ClueWeb09 B.

<table>
<thead>
<tr>
<th>ID</th>
<th>statmap</th>
<th>MFC(30)</th>
<th>MAP</th>
<th>MRR</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-wiki baseline $\beta = 0$</td>
<td>0.0880</td>
<td>0.2181</td>
<td>0.0802</td>
<td>0.2784</td>
<td>0.2160</td>
</tr>
<tr>
<td>Non-wiki baseline $\beta = 1$</td>
<td>0.1044</td>
<td>0.2528</td>
<td>0.1015</td>
<td>0.2814</td>
<td>0.2260</td>
</tr>
<tr>
<td>Non–Wiki Global In</td>
<td>0.1008</td>
<td>0.3692</td>
<td>0.0883</td>
<td>0.3566</td>
<td>0.1720</td>
</tr>
<tr>
<td>Non–Wiki Global Out</td>
<td>0.1030</td>
<td>0.3552</td>
<td>0.0877</td>
<td>0.3596</td>
<td>0.1680</td>
</tr>
<tr>
<td>Non–Wiki Global Union</td>
<td>0.1012</td>
<td>0.3594</td>
<td>0.0882</td>
<td>0.3537</td>
<td>0.1760</td>
</tr>
<tr>
<td>Non–Wiki Log Global In</td>
<td>0.1043</td>
<td>0.2699</td>
<td>0.0951</td>
<td>0.4207</td>
<td>0.1880</td>
</tr>
<tr>
<td>Non–Wiki Log Global Out</td>
<td>0.1080</td>
<td>0.3812</td>
<td>0.1019</td>
<td>0.3961</td>
<td>0.2540</td>
</tr>
<tr>
<td>Non–Wiki Log Global Union</td>
<td>0.1072</td>
<td>0.2980</td>
<td>0.1014</td>
<td>0.4174</td>
<td>0.2260</td>
</tr>
<tr>
<td>Non–Wiki Local In</td>
<td>0.1114</td>
<td>0.3131</td>
<td>0.1035</td>
<td>0.4705</td>
<td>0.2440</td>
</tr>
<tr>
<td>Non–Wiki Local Out</td>
<td>0.1115</td>
<td>0.3121</td>
<td>0.1053</td>
<td>0.4667</td>
<td>0.2700</td>
</tr>
<tr>
<td>Non–Wiki Local Union</td>
<td>0.1132</td>
<td>0.3312</td>
<td>0.1070</td>
<td>0.4827</td>
<td>0.2700</td>
</tr>
<tr>
<td>Non–Wiki Log Local In</td>
<td>0.1102</td>
<td>0.3185</td>
<td>0.1059</td>
<td>0.4291</td>
<td>0.2580</td>
</tr>
<tr>
<td>Non–Wiki Log Local Out</td>
<td>0.1101</td>
<td>0.3250</td>
<td>0.1073</td>
<td>0.3961</td>
<td>0.2760</td>
</tr>
<tr>
<td>Non–Wiki Log Local Union</td>
<td>0.1108</td>
<td>0.3270</td>
<td>0.1080</td>
<td>0.4253</td>
<td>0.2720</td>
</tr>
</tbody>
</table>

more effective than global link evidence, and the difference between incoming and outgoing link evidence has almost disappeared. In-degree is better for very early precision while out-degree is better for later and overall precision, just as in the INEX 2006 Wikipedia collection. The union of the two degrees is even more effective. We compare this again with our findings in previous chapters.

Global link evidence needs toning down, local does not: The normal, non-log global degrees hurt MAP when used over the top 100 results, just as in Wikipedia, and using the log of the degrees improves the effectiveness of the global degrees. The local degrees become slightly more effective for MAP but not for STATMAP. The log helps improve precision at a fixed rank cut-off but not MRR. In Wikipedia, the normal local degrees were clearly better than the log local degrees. In the Web, curbing the degrees is beneficial for performance.

Local link evidence is more effective than global link evidence: Without Wikipedia, local link evidence is more effective for Web ad hoc search than global link evidence, but global link evidence is still effective when applied with care. Document importance is useful for Web ad hoc search, but query-dependent evidence is more useful. Is local link evidence
more effective because it is a toned down version of the global degrees or because it is more related to topical relevance?

Local link evidence works in both directions, global link evidence does not: Local in-degrees are better for MRR while local out-degrees are better for P@10 and MPRC(30). For MAP the differences are small. The union of the in- and out-degrees further improves performance. This is almost exactly the same as for the INEX Wikipedia collection. For overall precision, the impact of link evidence is symmetric, suggesting that local links signal topical relevance.

Without Wikipedia, link evidence in the Web behaves similar to link evidence in Wikipedia. Only when we combine Wikipedia with the rest of the Web do the global degrees become effective and the out-degrees more so than the in-degrees. What is the impact of Wikipedia on the global in- and out-degrees?

In the $\beta = 1$ baseline, there are 15 Wikipedia pages in the top 100 on average. The baseline run has 1.2 Wikipedia pages in the top 10. The local degrees and the global in-degrees push up Wikipedia pages, all with roughly 2.2 Wikipedia results in the top 10. The global out-degrees and union degrees have 4.2 Wikipedia pages in the top 10. The global out-degrees favour Wikipedia pages more than other degrees do. For the non-Wikipedia results in the top 100, the median global in-degree is 3 and the median global out-degree is 9. Among the Wikipedia results in the top 100, the median global in-degree is 0 and the median global out-degree is 147. The global out-degree works as a Wikipedia filter much more than the other degrees.

This can explain why the global out-degrees are effective. Wikipedia is densely linked and most pages have a large amount of outgoing links. The global out-degree seems to push up Wikipedia pages in the ranking. Wikipedia pages are often considered high-quality pages with informational text (a pre-requisite for relevance in the ad hoc methodology) and are on average longer than the non-Wikipedia pages—Wikipedia pages are on average 7944 characters long while non-Wikipedia pages are 4635 characters long. The impact of the length prior shows that document length is related to relevance. This would mean that Wikipedia pages have a higher probability of being relevant than non-Wikipedia pages, and performance is improved by favouring Wikipedia pages high in the ranking. The same observation was made by He et al. (2010), who reranked the search results by pushing all Wikipedia results to the top of the ranking.

We ran the same queries on a Wikipedia-only index of the ClueWeb09 B collection (see Table 37) and found that on the Wikipedia-only index,
early precision is higher than on the whole ClueWeb09 B index, even though Wikipedia is a 10% subset. This is in direct contrast with the observation by Hawking and Robertson (2003) that precision at a fixed rank cut-off tends to increase with collection size. It is easier to find relevant pages in Wikipedia than in the rest of the ClueWeb09 B collection, suggesting that Wikipedia pages are of higher quality than non-Wikipedia pages.

The Wikipedia pages form 11% of the ClueWeb09 B collection. The judged Wikipedia pages form 18% of the judged pages. This means Wikipedia pages have a higher probability of being retrieved in the top results. The relevant Wikipedia pages form 21% of the relevant pages, meaning Wikipedia pages have a higher probability of being relevant.

That the log global degrees still lead to improvements indicates that in the Web, part of the role of link evidence is to identify important, high-quality documents. The global out-degrees lead to a larger improvement than the global in-degrees, which might be a document length effect.

It seems that the special nature of Wikipedia creates a bias in the ClueWeb09 B collection which muddles the analysis of the impact of link evidence. Below, we look at the impact of link density and inter-server links in ClueWeb09 B with and without Wikipedia.
Table 38: The impact of inter- and intra-server link evidence on retrieval effectiveness in the non-Wikipedia part of ClueWeb09 B.

### 7.3.4.1 Inter-server and intra-server links

We have seen that the Wikipedia part of ClueWeb09 B accounts for 446 million of the intra-server links and 22 million of the inter-server links (20.5 million links from Wikipedia to external pages and 1.5 million links from external pages to Wikipedia). With 5.7 million pages, it is by far the largest Web site in the collection. The next biggest Web site (in number of pages) has only 34,684 pages. The number of pages on the same site determines the maximum intra-server link degree. In a site with \( n \) pages, the maximum intra-server in- and out-degree is \( n - 1 \). Wikipedia pages can have much larger degrees than non-Wikipedia pages. Note that in ClueWeb09, Wikipedia has been crawled separately, which has an important impact on the composition of the collection. The crawling policy for the rest of the collection was to crawl new domains first and limit the number of pages per domain and the depth at which pages were crawled within each domain. Wikipedia is the only exception. In a normal crawl using the same policy, Wikipedia would form a much smaller part of the crawl and have a much sparser link graph. The change in the impact of link evidence on the ClueWeb09 B collection when we include Wikipedia might thus be an artefact of the way the collection is constructed. Because Wikipedia accounts for such a large part of the intra-server links and is a single giant Web site that accounts for 24% of the relevant documents, we look at the impact of inter- and intra-server links without Wikipedia in Table 38.

We show only the impact on the local and log local union degrees, which give the best performance. The impact is similar on all other degrees. Apart from the P@10 measure, intra-server link evidence outperforms inter-server link evidence on all measures. The presence of Wikipedia does not tip the balance in favour of the intra-server links. Even without Wikipedia, intra-server links are more effective than inter-
server links. The explanation given earlier is still valid: inter-server links cover a smaller number of pages than intra-server links and mainly point to site-entry pages. The bias of control over incoming links has no impact on the value of link evidence for topical relevance.

7.3.4.2 Link density

We repeat the link filtering method on the non-Wikipedia part of the collection and see the impact of filtering on effectiveness in Figure 29. The most striking difference from the impact on the whole collection is that, on the non-Wikipedia part, filtering has a larger impact on the effectiveness of the global degrees. Whereas on the whole collection the global degrees were barely affected by filtering, in the non-Wikipedia part the global degrees gradually lose their impact as we filter more links.

The small impact of filtering on the full collection can be explained by the fact that we used a random filter to reduce link density. With random filtering, all pages are affected in the same way. If page A has a higher link degree than page B in the full link graph, then A will also have a higher link degree on average in a randomly sampled link graph. In other words, the Wikipedia pages are promoted in the full collection regardless of the amount of links filtered. Highly connected pages are more robust against filtering links than other pages. The local graph is more fragile, so filtering has a larger impact on the local degrees.

7.4 CONCLUSIONS

In this chapter we looked at the impact of link evidence on ad hoc retrieval in the new ClueWeb09 B collection.
• What is the impact of link evidence on the ranking of Web ad hoc retrieval results?

Link evidence can significantly improve ad hoc retrieval effectiveness in the Web, when restricted to the top 100 results. This is in direct contrast to the findings of the TREC Web Tracks of 1999–2001. The main differences between the 1999–2001 evaluation on the Wt10g collection and the evaluation in this chapter on the ClueWeb09 B collection is that TREC 2009 topics are more general, the collection is much larger and more densely linked and contains the full English Wikipedia.

• How is the effectiveness of link evidence affected by the density of the link graph?

We randomly filtered links from the graph to study the impact of link density. The effectiveness of the global degrees is hardly affected by randomly filtering links while the effectiveness of local link evidence gradually decreases. The global graph is richer and therefore more robust against filtering. The top of the degree-based ranking changes little by random filtering. The local graph is much sparser and the ranking is more sensitive to the presence of specific links. However, even with a small amount of links, both local and global link evidence can improve a standard full-text retrieval baseline.

• What is the relative impact of inter-server and intra-server links on the effectiveness of link evidence?

Intra-server links make up a large part of the link graph. The inter-server link graph is much more sparse and covers only a small number of pages. If inter-server links in ClueWeb09 B mainly point to entry pages, as they do in the dataset used by Hawking et al. (2004), it would make them less suitable for ad hoc search, where any type of page can be relevant. Intra-server links cover a much larger part of the collection, including many pages deep within sites. The impact of link evidence on ad hoc search mainly comes from intra-server links.

The main difference between inter-server and intra-server links is that Web authors have more control over the intra-server links. That is, assuming that a single Web site is authored by a single author or a cooperating group of authors, authors control both the incoming and outgoing links within their own Web site, just as in Wikipedia. Intra-server links are more similar to Wikipedia links and show the same impact on link evidence. There is little qualitative difference between incoming and outgoing site-internal links, making in-degrees
and out-degrees equally effective. The bias of control over the incoming intra-server links does not affect the relation between link evidence and topical relevance.

- What is the impact of Wikipedia on the effectiveness of link evidence for Web ad hoc search?

In Wikipedia the link structure is much denser than in the rest of the Web, and Wikipedia pages tend to have higher global out-degrees than other Web pages. Although Wikipedia might be different in nature from the rest of the Web, the higher density is also partly due to the crawling policy restricting the number of pages to be crawled from a single Web site—except for Wikipedia—and perhaps also to the fact that only a limited part of the Web has been crawled. There might be links to the pages in the collection from Web pages that have not been crawled. As a consequence, the global out-degree pushes Wikipedia up the ranking. Because Wikipedia pages have a higher probability of relevance, the global out-degree is effective for improving Web ad hoc search.

Without Wikipedia, link evidence in the Web behaves similar to link evidence in Wikipedia, lending support to the findings of previous chapters. Local links are more effective than global links, and incoming and outgoing link evidence have a similar impact on overall performance. This suggests that in the Web, local link degrees provide evidence for the topical relevance of search results. The fact that global degrees are still effective when toned down by taking the log of the degrees shows that, with the presence of many low-quality pages, document importance is useful for Web ad hoc search.

With Wikipedia, the impact of link evidence changes radically. Global out-degrees are very effective because they favour longer articles and especially Wikipedia articles, which have a higher prior probability of relevance. The Wikipedia link graph is very different from most other site-internal link graphs. Wikipedia is a large domain with millions of relatively long articles that are densely interlinked, whereas the next biggest Web site in ClueWeb09 B has only 25 thousand pages. The impact of Wikipedia might be smaller in a larger crawl of the Web, where there is no restriction on the number of pages per Web site.

- To what extent does the value of link evidence in Wikipedia hold for link evidence in general?

The presence of Wikipedia affects the value of link evidence. Without it, our hypotheses mostly hold. The fact that local link evidence is more
effective than global link evidence supports our hypothesis that local link evidence signals topical relevance and that topical relevance is useful for ad hoc search. That incoming and outgoing links have a similar impact on overall performance supports our hypothesis that the evidence for topical relevance is symmetric. The positive impact of global link evidence shows that document importance is also useful for ad hoc, at least on the Web.

However, Wikipedia is part of the World Wide Web, and general aspects of hyperlinks should hold in both the entire Web and in Wikipedia. Any aspect of hyperlinks where Wikipedia differs from the whole Web cannot be a general aspect. How should we interpret the impact of Wikipedia on the value of hyperlinks for retrieval? Perhaps the impact of Wikipedia is not so much that it changes the nature of links, but the nature of informational search on the Web. Wikipedia forms a special part of the Web that is important for informational search.