Is Wikipedia Link Structure Different?

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

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

ABSTRACT

In this paper, we investigate the difference between Wikipedia and Web link structure with respect to their value as indicators of the relevance of a page for a given topic of request. Our experimental evidence is from two IR test-collections: the .GOV collection used at the TREC Web tracks and the Wikipedia XML Corpus used at INEX. We first perform a comparative analysis of Wikipedia and .GOV link structure and then investigate the value of link evidence for improving search on Wikipedia and on the .GOV domain. Our main findings are: First, Wikipedia link structure is similar to the Web, but more densely linked. Second, Wikipedia's outlinks behave similar to inlinks and both are good indicators of relevance, whereas on the Web the inlinks are more important. Third, when incorporating link evidence in the retrieval model, for Wikipedia the global link evidence fails and we have to take the local context into account.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*relevance feedback, retrieval models*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms

Measurement, Performance, Experimentation

Keywords

Web information retrieval, link evidence, Wikipedia

1. INTRODUCTION

The principal difference between Web retrieval and general information retrieval, is the abundant link structure of the Web which can been exploited to improve information retrieval in algorithms like PageRank [25] and HITS [16]. Similar to the earlier use of citations in bibliometrics, a link from a page A to a page B, can be considered as a vote by the author of page A for page B as being authoritative [16]. Wikipedia's links are a special case of the general hyperlinks that connect the World Wide Web. Internal links in Wikipedia are typically based on words naturally occurring in a page and link to another "relevant" Wikipedia page. As it is put in `http://en.wikipedia.org/wiki/Wikipedia:Only_make_links_that_are_relevant_to_the_context`:

> Only make links that are relevant to the context. It is counterproductive to hyperlink all possible words. ... A high density of links can draw attention away from the high-value links that readers would benefit from following. Redundant links clutter up the page and make future maintenance harder. A link is analogous to a cross-reference in a print medium. Imagine if every second word in an encyclopedia article were followed by "(see:)". Hence, the links should not be so numerous as to make the article harder to read.

Our conjecture is that the links in Wikipedia are different from links between arbitrary Web documents. Whereas in Web documents, an author can arbitrarily link his page to any other page, whether there is a topical relation or not, in Wikipedia, links tend to be relevant to the local context. A link from page A to page B shows that page B is semantically related to (part of) the content of page A. It is tempting to speculate about differences between the internal link structure of Wikipedia, and the link structure of the Web at large, and how these may affect the value of link based methods. First, as suggested by the quote above, links seem to signal a semantic relation between pages rather than serve pure navigational purposes, and may therefore provide a strong source of evidence for the relevance of a given page. Second, due to the shared authorship and encyclopedic organization of Wikipedia, we may expect a far more complete link graph where all (or a large fraction of all) relevant links are present, leading to higher link density and connectedness of the link graph, and promoting the effectiveness of the link evidence. Third, due to the encyclopedic organization the Wikipedia has relatively little redundant information. In addition, the huge Wikipedia is dwarfed by the size of the Web at large. This may have a number of consequences such as bounding the number of directly related incoming and outgoing links, as well as causing a quick loss of topical focus when traversing the link graph. Fourth, given that we search within a single domain, the authoritativeness of individual pages is essentially the same, and the value of link evidence is primarily to signal topical relevance. On the highly heterogeneous Web, link evidence may be used to signal other aspects of

WSDM'09, February 9–12, 2009, Barcelona, Spain

relevance, such as the general importance or authoritativeness of a site compared to other sites, or to indicate the best entry-page or entry-pages of the site.

Our main research question in this paper is to find out if, and how, the link structure of Wikipedia differs from the Web at large with respect to its value for promoting retrieval effectiveness. That is, we work in an information retrieval context where a user has a particular search request, and link evidence may help promote the quality of the search results on top of an effective text retrieval algorithm. To investigate our main research question, we need sets of search requests and associated relevance judgments. We use two IR test collections consisting of documents plus search requests and associated relevance judgments. For Wikipedia, we use the INEX 2006 and 2007 collections, together consisting of 217 ad hoc topics and an XML version of Wikipedia containing over 650,000 articles. For the Web, we use the TREC 2004 Web track collection, consisting of 225 topics and the 1.2 million documents .GOV collection. The Web track topics are a mix of 75 Named Page finding, 75 home page finding and 75 Topic Distillation topics. Although this collection provides us with the necessary topics and relevance judgments, and is reasonably comparable in size and number of topics to the Wikipedia collection, it is a relatively small crawl of a specific domain. We make no particular claims on the representativeness of this data set for the current Web, which is infinitely large and highly heterogeneous, but expect it to be a close enough approximation for our purposes [27].

Our main research question breaks down in two parts. We start by investigating the Wikipedia link structure with an extensive comparative analysis of the two IR test collections, Wikipedia and .GOV. Specifically, we want to know:

- What is the degree distribution of Wikipedia and the .GOV collections?
- Are there differences between distributions of incoming and outgoing links?
- And, in particular, how does the link topology relate to the relevance of retrieval results?

The second part of our main research question is about the effectiveness of link-based evidence. At TREC, we have seen that link degree is not effective for general ad hoc retrieval [10, 17]. However, for web-centric retrieval tasks such as topic distillation and known-item search, link indegree have been proved to improve retrieval performance [18, 30]. Link indegree can be considered on a global level, i.e. indegree over the whole collection (similar to PageRank), or on a local level, i.e. indegree within the subset of articles retrieved as results for a given topic (similar to HITS). And there are various ways in which link evidence can be implemented in the retrieval model. We continue our investigation by doing a range of experiments on the effectiveness of link based evidence. More specifically, we want to know:

- How can global or local link evidence be incorporated in our information retrieval models?
- What is the impact of link evidence on .GOV and Wikipedia retrieval? And, in particular, does it lead to improvement of retrieval effectiveness?

To answer our second set of questions, we work in the language modelling framework and build on and extend earlier approaches [13, 18, 24]. We define a range of operationalizations to incorporate link evidence into the retrieval model and conduct retrieval experiments with them on the TREC 2004 Web track topics and on the combined INEX 2006 and 2007 Ad Hoc track topics.

The rest of the paper is structured as follows. Next, in Section 2 we discuss earlier work on link structure and the use of link evidence in information retrieval. In Section 3, we perform a comparative analysis of the link structure of Wikipedia and .GOV and the relation between the link topology and the relevance of retrieval results. We continue in Section 4 by discussing our retrieval model and principal ways of incorporating link evidence into the model. Then, in Section 5, we perform a range of retrieval experiments, investigating the impact of link evidence on retrieval effectiveness. Finally, we end in Section 6 by summarizing our findings, and discussing their impact.

2. RELATED RESEARCH

There are three broad strands of related research, which we will discuss in turn. First, there is related research in studying web structure and its potential role on improving access to information. Various researchers have investigated the structure of the web [9, 20], its growth [2], or the emergence of 'cyber' communities [19]. The standard model of Internet is the so-called Bowtie model of [4], based on the link structure of 200 million pages and 1.5 billion links in an Alta-Vista crawl. Broder et al. [4] find a Strongly Connected Component (SCC) of 56M pages (28%), a set IN containing pages with a path to all SCC and a set OUT containing pages with a path from all SCC. The Weakly Connected Component (SCC+IN+OUT) contains 186 million pages (91%). The link structure of the web invites social network analysis [29], in particular notions of authority or importance [15, 26]. Particularly intriguing is the question whether such a link-based notion of importance can help improve search results. This question has been addressed by using either the global link structure, PageRank [25], or the local link structure, HITS [16]. Amento et al. [1] showed that link based approaches were effective in picking out high-quality results for a set of 5 queries, with indegree performing at least as well as PageRank and HITS authorities. They also found that ranking results by the total number of pages on their containing site performs nearly as well.

Second, there is currently emerging research in the nature of Wikipedia. Bellomi and Bonato [3] analyse PageRank and HITS on the Wikipedia link graph and provides lists of most authoritative pages, countries and cities, historical events, people and common nouns. Voss [28] analyses a range of characteristics of Wikipedia, and provides an analysis of Wikipedia link distributions. Buriol et al. [5] analyse the Wikipedia link graph over time, and amongst other things observe that the link density of Wikipedia is increasing over time, and that a far greater fraction of pages belongs to the strongly connected component than in earlier studies of Web crawls.

Third, there is related research in studying web retrieval within the narrower context of experimental IR test collections. Retrieval using Web data has been studied at TREC since TREC-8 in 1999. Despite high expectations, TREC experiments failed to establish the effectiveness of link evidence for general ad hoc retrieval [e.g., 10, 17]. Hawking and Craswell [11, p.215] explain why Web search is different

from traditional TREC ad hoc search: "Web searchers typically prefer the entry page of a well-known topical site to an isolated piece of text, no matter how relevant. For example, the NASA home page would be considered a more valuable answer to the query 'space exploration' than newswire articles about Jupiter probes or NASA funding cuts." These observations led to the definition of a range of Web-centric tasks, like known-item (home page, named-page) search and topic distillation.

Craswell et al. [6] investigated incoming anchor texts as a document representation and showed its effectiveness for home page or entry-page finding. Kraaij et al. [18] investigated the importance of query independent evidence for home page finding and show that document length is not helping, but the number of incoming links and especially the URL-form is promoting retrieval effectiveness. The importance of various document representations, such as incoming anchor texts and title-fields was further established by Ogilvie and Callan [24] for more general known-item search (home page finding and named-page finding). Craswell et al. [7] studies query independent evidence for a mixed query set of topic distillation, home page finding and named-page finding topics, and find that, in order of impact, PageRank, indegree, URL length and click-distance improve the effectiveness over the mixed query set. Kamps [13] conducts similar experiments and shows that indegree and URL evidence promotes topic distillation and home page finding, but gives mixed results on named-page finding. URL evidence plays no role in Wikipedia searching, so we will focus in this paper on link evidence.

There has been an important attempt to bridge the gap between the scale of scientific IR test collections and the web at large. Najork et al. [23] studies the effectiveness of link-based evidence on 464 million web pages, 28,043 queries and evaluate on the top 10 results. They find that combining link-based features with the content based scores lead to substantial improvements, with features based on incoming links (PageRank, indegree, and HITS authorities) superior to features based on outgoing links (outdegree and HITS hubs).

3. COMPARATIVE ANALYSIS OF LINK STRUCTURE

In this section, we look in close detail at the link structures of the Wikipedia and Web collections. We base our analysis on two IR test collections, consisting of a collection of documents, a large set of search requests and relevance judgments. For the Web, we take the .GOV collection used at the TREC Web tracks (2002-2004) which is based on a crawl of the `.gov` domain in early 2002. For Wikipedia, we take the Wikipedia XML Corpus used at INEX (2006-2007) which is based on an XML'ified version of the English Wikipedia in early 2006 [8]. This collection is based on the regular Wikipedia dumps provided by the Wikimedia foundation, and includes all pages including stubs. However, the pages do not include the side-bar with navigational links present in the online rendition of the pages. Therefore, in the online version of Wikipedia there will be more links on a page than we report here.

3.1 Web and Wikipedia Graph

The .GOV collection contains 1,247,753 documents and 11,110,989 unique links between these pages (we ignore links which point to, or from, pages outside the collection). The Wikipedia collection contains 659,304 documents and a total of 13,602,613 unique links between these pages. We have also looked at how many of these links are reciprocal, i.e., a link from page A to B in combination with a link from page B to A. There are 1,269,988 (11.4%) reciprocal links in the .GOV collection, and 1,182,558 (8,7%) reciprocal links in the Wikipedia collection. The higher fraction of reciprocal links in the .GOV collection is likely due to the presence of navigational links within web-sites. Table 1 gives some statistics on the incoming (indegree) and outgoing (outdegree) links and document lengths of both collections. We calculate length in characters. The pages in .GOV have mean length 6,345 characters (median 1,892) and the pages in Wikipedia are shorter with a mean length of 2,473 characters (median 1,309).

Table 1: Statistics of the .GOV and Wikipedia collections

		min	max	mean	median	stdev
GOV	Indegree	0	44,228	8.90	1	126.00
	Outdegree	0	653	8.90	4	16.61
	Length	2	102,069	6,345	1,892	13,377
Wiki	Indegree	0	74,937	20.63	4	282.94
	Outdegree	0	5,098	20.63	12	36.70
	Length	16	281,150	2,473	1,309	4,238

In .GOV, the average number of in- and outlinks per document is 8.90, in Wikipedia 20.63. Recall that, here, we are only using within-collection links, so every outgoing link is also an incoming link. The median number of incoming links is 1 in .GOV and 4 in Wikipedia and the median number of outgoing links is 4 in .GOV and 12 in Wikipedia. Also the maximal outdegree in Wikipedia (5,098) is much higher than in the .GOV collection (653). Again, we make no particular claims on the .GOV collection being a good representative of the Web at large. On the one hand, the indegrees should increase if we would consider a larger set of pages (since we cannot detect incoming links from pages outside the collection) leading us to underestimate the indegrees. On the other hand, the limited crawl will likely have favored pages with larger numbers of incoming links (e.g., how to crawl pages with no incoming links?) leading us to overestimate the mean indegree. To put these numbers in perspective, Najork et al. [23] use a Web crawl of 464 million pages and 18 billion hyperlinks, and find mean indegree of 6.10 and a mean outdegree (not limited to pages in the crawl itself) of 38.11.

The Wikipedia collection is thus more densely linked. This is surprising in the sense that the .GOV domain is much older, and link density tends to increase over time [21, 22]. There are at least two effects which help explain why the Wikipedia link graph is more "complete" than the .GOV link graph. First, due to the strongly structured nature of Wikipedia and the existence of author guidelines, it is much clearer for Wikipedia authors where to link to and when. Second, due to peer editing and automatic link detection, "missed" links will be added in a matter of time. With the high link densities, we see a single giant component, i.e., a large set of connected pages. The giant strongly connected component (SCC) of the .GOV collection contains 912,794

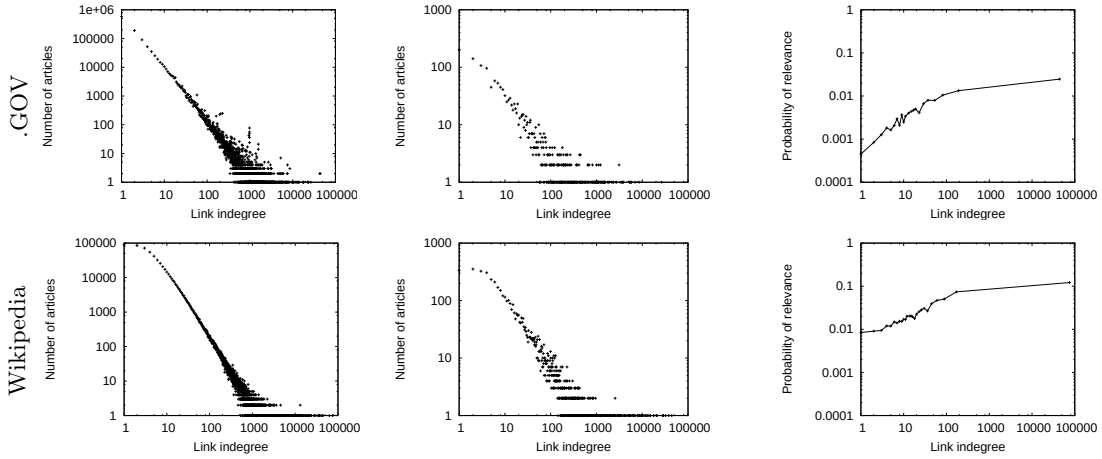

Figure 1: Link indegree distribution of all pages (left), of relevant pages (middle) and prior probability of relevance (right) for .GOV (top) and Wikipedia (bottom).

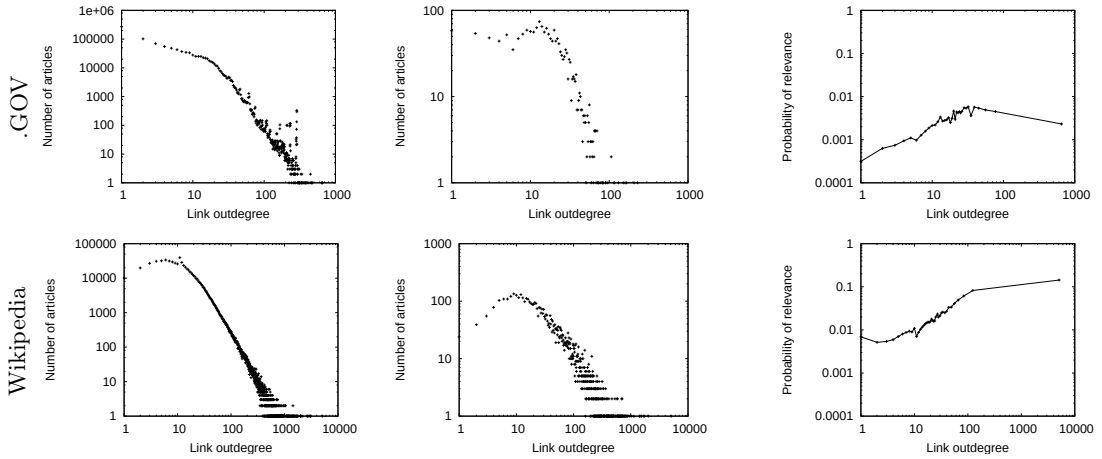

Figure 2: Link outdegree distribution of all pages (left), of relevant pages (middle) and prior probability of relevance (right) for .GOV (top) and Wikipedia (bottom).

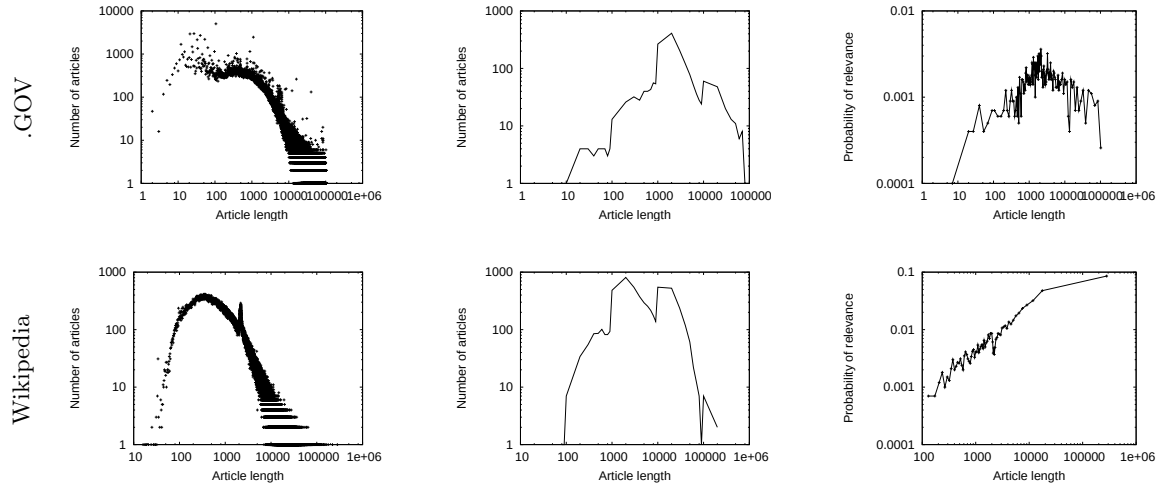

Figure 3: Article length distribution (left), relevant article length distribution (middle) and prior probability of relevance (right) for .GOV (top) and Wikipedia (bottom).

(or 73.16%) documents and the giant weakly connected component (WCC) contains 1,209,324 (or 96.92%). The giant SCC of the Wikipedia collection contains 605,952 (or 91.91%) and the giant WCC contains 657,601 (or 99.74%). The WCC and especially the SCC of the Wikipedia collection contain a much larger part of the entire collection than the SCC and WCC of the .GOV collection. For both the .GOV and Wikipedia collection, the percentage of pages in the SCC is considerably larger than in the large crawl of [4]. Soboroff [27] finds that the .GOV collection structurally resembles much larger web crawls but is very closely connected due to either starting the crawl from a small number of seeds or to the `.gov` domain being much more densely linked than the Web in general. The Wikipedia collection is a complete dump and the high link density, as also observed in [5], cannot be a crawling artifact.

3.2 Degree Distributions

We now look at the indegree (the number of incoming links) and the outdegree (the number of outgoing links). More precisely, we count unique pages that link to a given page, or are linked to by that page. If we look at the distribution of these degrees over the entire collections, in the left of Figure 1 for indegree and Figure 2 for outdegree, we see a power-law for all degree distributions. In .GOV, the power-law distribution is especially clear for the indegrees and less clear for the outdegrees. In Wikipedia, we see much smoother distributions and surprisingly little difference between the incoming and outgoing links. This suggests that outlinks in Wikipedia behave very much like inlinks. This is consistent with a semantic nature of links in Wikipedia: if a link from A to B means that B is relevant (in some sense) to A, then it is also likely A is relevant (in some sense) to B.

3.3 Relevant Link Distribution

Recall that we base our analysis on IR test collections and hence also have available sets of search requests and associated sets of relevant pages. How are the link degrees of these relevant pages distributed? For .GOV, we use the TREC 2004 Web track data consisting of 225 retrieval topics and in total 1,763 relevant pages for .GOV. For Wikipedia, we use the combined INEX 2006 and 2007 Ad Hoc track data, consisting of 217 topics and in total 11,896 pages with relevance. The INEX topics have relevance assessments on the passage level, in this paper we only consider full Wikipedia article retrieval and regard an article as relevant for a topic if and only if some part of the article is judged relevant.

Since we are interested in generic query-independent features, like the numbers of incoming or outgoing links, we will simply accumulate all relevant pages over topics. We will be stretching this argument to include features like the local indegree amongst top ranked documents. Although these degrees are not query-independent, and have to be calculated specifically for a given topic at query time, they are similar in character to global query-independent features and will play a similar role in the retrieval models discussed in Section 4.

The middle parts of Figures 1 (indegree) and 2 (outdegree) show the degree distributions over the relevant pages (for any topic) of the collections. The indegree distributions still show a power-law, but the outdegrees—especially in the .GOV collection—adhere less to a power-law distribution. It is not clear whether this is a true deviation from the standard power-law distribution (since relevant documents on average have a higher outdegree than non-relevant documents) or whether it is a consequence of the limited number of relevant documents. The higher number of relevant pages available in the Wikipedia collection also explains why these curves are smoother than those of the .GOV documents. What is interesting to see is that, for the Wikipedia collection, the outdegree distribution shows a clear upward trend over relevant documents with the lowest degrees (between 1 and 10). Does this mean that documents with a higher degree are more frequently relevant than documents with a lower degree?

3.4 Prior Probability of Relevance

We will now analyse the prior probability of relevance of a page with a particular degree. If the degrees of relevant documents deviate from the degrees of non-relevant documents in the collection, they may possibly be used as indicators of relevance. We have calculated the prior probability of relevance as follows. The documents are sorted into bins of equal size with ascending degree. Each bin contains 10,000 documents and the prior probability of relevance for these documents is computed by dividing the number of relevant documents in a bin by the total number of documents in that bin. That is, the 10,000 documents with the lowest degrees go into the first bin, the next 10,000 in the second bin, etc. Since there are many documents with an in- or outdegree of 0 or 1, they fill up several bins. In calculating the degree priors of these bins, we merge bins with the same maximal degree and assign each of them the same prior.

The probabilities are shown on the right of Figures 1 (indegree) and 2 (outdegree). In the .GOV collection, the probability of a document being relevant increases with indegree. For outdegree, the probability of relevance peaks somewhere between 10 and 100 and then drops as the outdegree further increases.

In the Wikipedia collection both in- and outdegree seem to be good indicators of relevance: a higher degree corresponds to a higher probability of relevance. Recall from above that the fraction of reciprocal links in Wikipedia is actually lower than that of .GOV; it is not a result of pages linking back-and-forth. This, again, signals the difference in the link structure of Wikipedia and the Web at large. For the semantic links of Wikipedia, the difference between incoming and outgoing links seems to disappear and both can be used as indicators of relevance.

3.5 Length

So far we have considered only link evidence. Another type of evidence is document length. Figure 3 (left) shows the length distribution for .GOV and Wikipedia. We see no power-law distribution for .GOV and more of a log-normal distribution for Wikipedia. The distribution of relevant pages is too sparse to give an interpretable plot, and we crudely bin it by rounding length to a single significant digit. The resulting plot is shown in Figure 3 (middle). The prior probability of relevance is calculated as the degree distributions above, and shown in Figure 3 (right). For .GOV, there is no evidence for the value of document length as indicator of relevance. For Wikipedia, in contrast, the distribution suggests that document length can be used as indicator of relevance.

Table 2: Correlation between length and degrees for Web and Wikipedia collections

	Web			Wiki		
Variables	In	Out	Length	In	Out	Length
In	1.00	0.10	-0.01	1.00	0.19	0.16
Out	–	1.00	-0.07	–	1.00	0.65
Length	–	–	1.00	–	–	1.00

Document length might actually be correlated to the link degree. Naively, we would expect that a document with many links going out is longer than a document with very few links going out. How is the length of documents related to the link degree? Moreover, we have seen above similar behavior for Wikipedia indegree and outdegree: how do these correlate? Table 2 gives the correlation between the indegrees, the outdegrees and the document length for both collections. For .GOV, we see a low correlation between indegree and outdegree and no correlation between length and the degrees. For Wikipedia, we see a low correlation between indegree and outdegree and between length and indegree. However, there is a strong correlation between outdegree and document length in the Wikipedia collection. This makes sense, since pages containing more textual content will naturally give rise to more links inside Wikipedia.

4. INCORPORATING LINK EVIDENCE

In this section, we will discuss how link evidence can be incorporated in a state-of-the-art retrieval model.

4.1 Retrieval Model

We use a language model where the score for a document d given a query q is calculated as:

$$P(d|q) = P(d) \cdot P(q|d) \tag{1}$$

where $P(q|d)$ can be viewed as a query generation process—what is the chance that the query is derived from this document—and $P(d)$ a document prior that provides an elegant way to incorporate link evidence and other query independent evidence [12, 18].[1]

We estimate $P(q|d)$ using Jelinek-Mercer smoothing against the whole collection, i.e., for a collection D, document d and query q:

$$P(q|d) = \prod_{t \in q} \left((1-\lambda) \cdot P(t|D) + \lambda \cdot P(t|d) \right), \tag{2}$$

where

$$P(t|d) = \frac{\text{freq}(t,d)}{|d|} \tag{3}$$

$$P(t|D) = \frac{\text{freq}(t,D)}{\sum_{d' \in D} |d'|} \tag{4}$$

For the Web collection, we use a mixture language model over three document representations: document text, incoming anchor texts and title field, i.e., for a collection D, document d and query q:

$$P(q|d) = \prod_{t \in q} ((1-\lambda_1-\lambda_2-\lambda_3) \cdot P(t|D) + \lambda_1 \cdot P_{\text{doc}}(t|d) + \lambda_2 \cdot P_{\text{anchor}}(t|d) + \lambda_3 \cdot P_{\text{title}}(t|d)),$$

[1]Note that we are ranking here documents given a query and can use this simplified version instead of the Bayesian $P(d|q) = (P(d) \cdot P(q|d))/P(q)$.

Table 3: Titles with the highest indegrees in the Wikipedia collection for INEX topic 339, "*Toy Story*"

Title	Global	Title	Local
Test cricket	1,405	Toy Story	33
Nobel Prize in Physics	557	Toy Story 2	22
Sequel	529	Pixar	20
1999 in film	427	Buzz Lightyear	8
Jet Engine	341	Cars (film)	6
Pacifism	339	Toy Story 3	6
Unix-like	339	John Ratzenberger	5
Portrait	339	John Lasseter	5
Psychedelic music	339	Sheriff Woody	5
Toy	331	1755 Lisbon earthquake	3

where each of the document language models is estimated as above [13].

4.2 Baseline

For the mixture model run, all three models are weighted the same with $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$. For the Ad Hoc Wikipedia task we use a standard language model run, with the default smoothing parameter setting $\lambda = 0.15$. We performed the experiments discussed below against two baselines with or without a length prior and found the same qualitative patterns. Below, we will show only the experiment against the highest scoring baseline, which is without length prior for the Web and with length prior for the Wikipedia (as was already suggested by Figure 3).

4.3 Link Evidence as Document Priors

We use global and local link evidence, which we will illustrate by discussing in detail two topics. Wikipedia topic 339 has title *Toy Story* and is about the computer animated movie from 1995. .GOV topic 119 has as title *Groundhog day Punxsutawney* and is about a celebration day in Punxsutawney, where groundhog Phil makes a weather forecast for the whole year.

We first look at global degrees, i.e. the total number of incoming links, or outgoing links for a page. To illustrate the effect of global degrees, we took the top 1,000 articles from the baseline run described above and list the 10 articles with the highest global indegree in Table 3 (left hand side) for Wikipedia and in Table 4 (left hand side) for .GOV. What we see is that some pages with little bearing on the topic at hand, for example the page on *Toys* in the case of Wikipedia, have a very high global indegree, but most pages have no relation to the topic at all. The same holds for the .GOV pages. The only slightly related page is *National Weather Service Forecast Office - Memphis, TN*. All the other pages seem to be entirely unrelated to the topic, yet could infiltrate the top ranks when too much weight is put on the degrees.

Hence, we also look at local degrees, i.e. the number of incoming or outgoing links between the top 100 pages according to the query-based retrieval score. To illustrate the effect of local degrees, we also list the 10 articles with the highest local indegree in Table 3 (right hand side) for Wikipedia and in Table 4 (right hand side) for .GOV. We see that the local degrees keep better focus on the topic of request, although local links are also sparser and just a few local links are all it takes to infiltrate the lower ranks.

For practical reasons, we implement the global document priors to the top 1,000 retrieved pages and the local document priors to the top 100 pages based on the content scores.

Table 4: Titles with the highest indegrees in the .GOV collection for TREC topic 119, "*Groundhog day Punxsutawney*"

Title	Global	Title	Local
Site Map	3,119	Bureau of Labor Statistics Home Page	61
Online Library - HUD	2,119	NTP Meetings & Events	58
Bureau of Labor Statistics Home Page	1,119	Recalls and other Press Releases	5
AMS - Search	730	What's New	3
The United States Mint	722	NCDC: Climate of 2001 - Climate Perspectives Reports	3
NHGRI: In The News	518	Youth Opportunity Movement Highlights	3
Metadata Records By Catalog Title	448	California Department of Motor Vehicles home page	2
FCC Universal Licensing System	348	Washington State Senate Democratic Caucus Home Page	2
The Embassy of the U.S.A., Ottawa - Canada - United States Relations	256	Conferences and Events in California related to the Career Development and Curriculum Leadership Unit	2
National Weather Service Forecast Office - Memphis, TN	249	Hewitt celebrates Groundhog Day with 'shadow' from Columbia High School	2

That is, we compute the new score by multiplying the content score with the link degree prior.

First, we use a *standard degree prior* by multiplying the retrieval score with 1+ the degree:

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

Here, the degree score for a page may be based on either *local* or *global*, and either *indegree* or *outdegree* (leading to four logical cases). We will, for convenience, refer to the link evidence as prior, even though we do not actually transform it into a probability distribution. Note that we can turn any prior into a probability distribution by multiplying it with a constant factor $\frac{1}{\Sigma_{d \in D} prior(d)}$, leading to the same ranking.

Second, we use a *log degree prior* using the logarithm of the degrees:

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

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

In earlier work on XML element retrieval [14], we found that weighting the local degree (the number of links to or from pages in the relevant set) by the global degree (the number of links to or from arbitrary pages) keeps more focus on the topic by removing infiltrations in the local set from pages with very high global degree. This is similar to the well-known *tf.idf* weighting scheme used to determine term importance. Our third prior is a combination of the global *and* local link evidence computed as:

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

For the degree score we may take either *indegree* or *outdegree*. For the combined LocGlob prior, the logged version uses only the log of the global degree.

5. EXPERIMENTS

In this section we experiment with using both indegree and outdegree evidence, either with the standard or log indegree priors. First we will discuss the experiments with the link evidence on the Web collection, then on the Wikipedia collection. The Web collection uses the TREC 2004 Web track data: a mixed query set of 225 topics in equal fractions of topic distillation, home page finding and named-page finding topics. Here the known-item search topics tend to have a single relevant document (possibly more due to duplicates in the collection), and the distillation topics tend to have a larger set of key results.[2] The Wikipedia collection uses the combined INEX 2006 and 2007 Ad hoc track data: a set of 217 ad-hoc retrieval topics, and much larger sets of topically relevant documents. The tables show Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) scores. In our discussion we will mainly focus on MAP.

5.1 Length Prior

We choose our baseline runs based on experiments with the document length priors. In line with the plots showing the probability of relevance over document length, the best run for the Web track collection uses no length prior (MAP drops from 0.3970 to 0.3419 when using the length prior, MRR drops from 0.4662 to 0.3868), whereas the best run for the Wikipedia collection uses a document length prior (MAP goes up from 0.2544 to 0.3069, MRR goes up from 0.6923 to 0.8102). We choose these best runs as baseline runs for the experiments with the link priors. Note that the higher MAP for the Web data can be attributed to differences in the tasks, with a large fraction of known-item search topics used on the Web collection.

5.2 Web

Table 5 shows the results for the link prior runs on the Web track collection. We tested all the runs for the significance of the increase or decrease in performance over the baseline using the bootstrap test, one-tailed, using 100,000 resamples. We report three significance levels, p<.05 (°), p<0.01 (°) and p <0.001 (•).

If we look at the global degree prior, columns 2 (MAP) and 5 (MRR), we see significant improvements for both in- and outdegree. The indegree prior leads consistently to a greater improvement than the outdegree, which is in line with the probability of relevance plots in Figures 1 and 2. For the logged version we see a similar pattern. Indegree works better than outdegree, but the scores are lower than with the standard priors.

The results for the local priors are given in columns 3 and 6. Again, the indegree is more effective than the outdegree. If we compare the local with the global priors, we see that the

[2]TREC experiments on Web data failed to establish the effectiveness of link evidence for general ad hoc retrieval [11]. Since we want to compare the impact of link evidence in Wikipedia with that on Web data, we focus on a web-centric retrieval task where link evidence is known to be effective.

Table 5: Results of the different link priors over in- and outdegree on the 225 topics of the Web track collection

	MAP			MRR		
Run id	**Glob**	**Loc**	**Loc/Glob**	**Glob**	**Loc**	**Loc/Glob**
baseline		0.3970			0.4662	
in	0.4738$^{\bullet}$	**0.4799**$^{\bullet}$	0.3966$^{\bullet}$	**0.5885**$^{\bullet}$	0.5655$^{\bullet}$	0.4646$^{\bullet}$
out	0.4299°	0.4497$^{\bullet}$	0.4042$^{\bullet}$	0.5046$^{\bullet}$	0.5199$^{\bullet}$	0.4744$^{\bullet}$
log.in	0.4449$^{\bullet}$	0.4410$^{\bullet}$	0.4395$^{\bullet}$	0.5209$^{\bullet}$	0.5148$^{\bullet}$	0.5101$^{\bullet}$
log.out	0.4082°	0.4181$^{\bullet}$	0.4289$^{\bullet}$	0.4789$^{\bullet}$	0.4879$^{\bullet}$	0.4913$^{\bullet}$

Table 6: Results of the different link priors over in- and outdegree on the 217 topics of the Wikipedia collection

	MAP			MRR		
Run id	**Glob**	**Loc**	**Loc/Glob**	**Glob**	**Loc**	**Loc/Glob**
baseline		0.3090			0.8121	
in	0.3018^{-}	0.3190°	0.3140°	0.8139^{-}	0.8236^{-}	0.8161^{-}
out	0.3016^{-}	0.3199$^{\bullet}$	0.3123°	0.8262^{-}	0.8266^{-}	0.8119^{-}
log.in	0.2865^{-}	0.3176$^{\bullet}$	**0.3234**$^{\bullet}$	**0.8322**$^{\circ}$	0.8289°	0.8263^{-}
log.out	0.2890^{-}	0.3156$^{\bullet}$	0.3203$^{\bullet}$	0.8291°	0.8225°	0.8211^{-}

Table 7: Results of HITS runs on the Web track collection (using all links or using only transverse links) and on the Wikipedia collection

	Web				Wikipedia	
	All links		Transverse		All links	
Run id	**MAP**	**MRR**	**MAP**	**MRR**	**MAP**	**MRR**
baseline	0.3970	0.4662	0.3970	0.4662	0.3090	0.8121
Authority 100	0.0391$^{\bullet}$	0.0638$^{\bullet}$	0.1014$^{\bullet}$	0.1801$^{\bullet}$	0.0106$^{\bullet}$	0.1063$^{\bullet}$
Authority 200	0.0348$^{\bullet}$	0.0571$^{\bullet}$	0.0601$^{\bullet}$	0.1262$^{\bullet}$	0.0061$^{\bullet}$	0.0803$^{\bullet}$
Authority 100 prior	0.3960^{-}	0.4657^{-}	**0.4056**$^{\bullet}$	**0.4758**$^{\circ}$	0.2766$^{\bullet}$	0.7982^{-}
Authority 200 prior	0.3958^{-}	0.4656^{-}	0.3990°	0.4690°	0.2740$^{\bullet}$	0.7987^{-}
Hub 100	0.0163$^{\bullet}$	0.0303$^{\bullet}$	0.0219$^{\bullet}$	0.0495$^{\bullet}$	0.0135$^{\bullet}$	0.1161$^{\bullet}$
Hub 200	0.0127$^{\bullet}$	0.0207$^{\bullet}$	0.0165$^{\bullet}$	0.0337$^{\bullet}$	0.0054$^{\bullet}$	0.0585$^{\bullet}$
Hub 100 prior	0.3967^{-}	0.4663^{-}	0.3994^{-}	0.4687^{-}	0.2773$^{\bullet}$	0.7990^{-}
Hub 200 prior	0.3963^{-}	0.4662^{-}	0.4009^{-}	0.4706^{-}	0.2755$^{\bullet}$	0.7955^{-}

local degree priors tend to lead to a higher MAP, while the global degree priors tend to lead to a higher MRR. In other words, the global priors are more effective for improving early precision, while the local priors are more effective for overall precision.

The combined local/global priors are much less effective. Only the combined outdegree priors lead to a very small improvement for both MAP and MRR, while indegree priors lead to small decreases in performance. The logged version fares somewhat better, with significant improvements in all cases.

In sum, the outdegree is clearly the least effective of the two degrees, whether used as a standard or log prior and whether evaluated by MAP or MRR. For global link and local link evidence, the standard prior works better than the logged prior. The global priors are most effective for early precision, while the local priors are most effective for MAP.

5.3 Wikipedia

Now, we move to the Wikipedia collection and experiment with the degree priors on the ad hoc topic set. Table 6 shows the results for the baseline and reranked runs on the Wikipedia collection. The second (MAP) and fifth (MRR) columns show the results for the global prior. We see that the global prior has a small positive effect on MRR, but a negative effect on MAP. For the standard prior, none of the differences are significant. The positive effect on MRR and the negative effect on MAP are stronger for the logged prior and the improvements for MRR are significant. The global prior can be used to improve early precision, but is not effective for pushing up the lower scoring relevant documents. There seems to be little difference between using either in- or outdegree priors, as the scores are very close to each other. The local prior results are in columns three and six. Like the global prior, the local prior improves early precision, but also boosts the MAP scores. The standard prior is more effective for MAP, while the logged prior is more effective for MRR. While the standard local degree priors generally lead to a higher score than the log local degree priors, the log prior improvements are more significant. The local outdegrees priors seem to perform at least as well as the local indegree priors.The standard combined local/global prior (columns 4 and 7) shows significant improvements for MAP, but the improvements are less than those of the standard local prior. For MRR the combined priors have very little effect. When we use the logged version, the MAP scores improve further, leading to the highest MAP scores overall.

To sum up, when using link degrees for reranking results in the Wikipedia collection, both incoming and outgoing links can be used as evidence but—compared to the Web track collection—we have to be more careful when using it. The global degrees alone seem to be ineffective for improving ad hoc retrieval results, leading only to improvements in early precision. The more informed local degree priors fare much better, with a significant improvement in MAP for ad hoc retrieval. The even more careful, weighted local/global prior can further improve MAP when using the log of the global

degree.

5.4 HITS

We also look at the effectiveness of HITS [16]. Given that the local degrees tend to be more effective on Wikipedia than the global degrees, we expect HITS to be more effective than PageRank [25]. Specifically, we look at the HITS authority and hub scores in isolation and using them as a prior:

$$P_{\text{HITS}}(d) \propto 1 + \mathsf{HITS}(d).$$

We use an initial base set of either 100 or 200 top ranked pages and expand it with pages linking to a page in the base set (maximally 50 pages per page in the base set) and with all pages linked to from a page in the base set. In Wikipedia most links are to semantically related pages, but on the Web links may exist for a variety of reasons. We can try to distinguish between intrinsic web links (for example, navigational links within a site) and transverse web links (for example, a link to related content on a different site). We first identified the site of a page as its base URL, with the removal of any prefix starting with www and excluded links between pages within the same domain. We further reduced the set by removing links between base URLs when either is a substring of the other. For example, a link between `www.nih.gov` and `www.nlm.nih.gov` regarded as intrinsic, while a link between `www.nlm.nih.gov` and `www.nichd.nih.gov` is regarded as transverse. The resulting set of transverse links contains 1,693,477 links (or 15% of all links). For the Web collection, we use either the full link graph, or the transverse links. We used the transverse link graph for the degree priors above as well, but while they improve performance, using the full link graph is much more effective.

The results for the HITS runs are shown in Table 7. What we see is the following: On the Web collection, HITS hubs or authorities alone perform poorly when compared to the baseline, although authorities do much better than hubs. The transverse links are more effective than the full link graph, which fails to improve upon the baseline. However, the improvements with the transverse links are still much lower than with the earlier discussed link degree priors. When using HITS the transverse link graph is more useful. We also did the experiments in Section 5.2 for the transverse link graph (not reported here) and found out that the full link graph is more effective than the transverse link graph for the indegree priors.

On the Wikipedia collection we also see that the HITS scores alone perform well below the baseline. Interestingly, the hubs (based on outgoing links) are performing better than the authorities (based on incoming links). In Wikipedia, the nature of the incoming and outgoing links is similar making the difference between hubs and authorities disappear. The results also show that using more results in the base set quickly leads to topic drift and lower performance.

6. DISCUSSION AND CONCLUSIONS

In this paper, we investigated the difference between Wikipedia and Web link structure. We first performed a comparative analysis of Wikipedia and .GOV link structure and then investigated the value of link evidence for improving search on Wikipedia and on .GOV. Our experimental evidence is from two IR test-collections consisting of documents, a large set of search requests and associated relevance judgments. The first is the .GOV collection used at the TREC Web tracks consisting of a crawl of the `.gov` domain. The second is the Wikipedia XML Corpus used at INEX consisting of a XML'ified dump of the English Wikipedia.

In our comparative analysis of Wikipedia and Web link structure we hoped to find out:

- What is the degree distribution of Wikipedia and the Web at large?
- Are there differences between distributions of incoming and outgoing links?
- How does the link topology relate to the relevance of retrieval results?

Analysis of the link structures of the .GOV and Wikipedia collections shows that the Wikipedia pages are more densely interlinked and that their outdegree distribution is much more similar to the indegree distribution than the .GOV pages. The .GOV collection has a giant strong connected component larger than general web crawls, but the giant strong connected component of Wikipedia covers an even larger part of the collection.

For the .GOV collection, the global indegree is a good indicator of relevance. Pages with many incoming links have a higher probability of being relevant than pages with few incoming links. The prior probability of relevance of the number of outgoing links is first increasing but then drops again, making the outdegree a less clear indicator of relevance than the indegree. For the Wikipedia collection, both indegree and outdegree are good indicators of relevance. More generally, we observe that Wikipedia inlinks and outlinks are similar in character, leading to the conflation of the notions of authority and hub [16].

In our retrieval experiments, we hoped to find an answer to the following questions:

- How can global or local link evidence be incorporated in our information retrieval models?
- What is the impact of link evidence on Web and Wikipedia retrieval? And, in particular, does it lead to improvement of retrieval effectiveness?

The language modelling framework allows for easy incorporation of document priors and we have experimented with different priors using the link degrees as evidence.

For the Web track collection, all global and local outdegree priors are less effective than the corresponding indegree priors. The Web task requires high precision, the global degree prior is effective to achieve this and no curbing is necessary. The indegree leads to greater improvements than the outdegree, supporting the claim that document importance is a major aspect in Web retrieval. The local degrees are still very effective for improving early precision, but are even more effective for general precision. The combined local/global evidence is less effective.

For the Wikipedia collection, the outdegree priors behave very similar to the indegree priors. The brute force of the global degree priors is too much for the task of ad hoc retrieval. Even the more subtle log degree prior is not effective for MAP. As in the Web collection, adding global link evidence can improve early precision, but hurts performance at lower ranks. The local degrees stay more on topic and

can improve early and later precision. The even more subtle, log version of the combined local/global indegree prior is most effective, showing that link evidence has to be carefully weighted and made sensitive to the local context.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] B. Amento, L. Terveen, and W. Hill. Does 'authority' mean quality? predicting expert quality ratings of web documents. In *SIGIR 2000*, pages 296–303. ACM Press, 2000.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] F. Bellomi and R. Bonato. Network analysis for wikipedia. In *Proceedings of Wikimania*, 2005.

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *WWW9*, pages 309–320. Elsevier Science, Amsterdam, 2000.

[5] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal analysis of the wikigraph. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 45–51. IEEE Computer Society, Washington, DC, USA, 2006.

[6] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *SIGIR 2001*, pages 250–257. ACM Press, 2001.

[7] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *SIGIR '05*, pages 416–423. ACM, New York, NY, USA, 2005.

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

[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM '99*, pages 251–262. ACM Press, 1999.

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

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

[12] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, Center for Telematics and Information Technology, University of Twente, 2001.

[13] J. Kamps. Web-centric language models. In *CIKM'05*, pages 307–308. ACM Press, 2005.

[14] 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 *Lecture Notes in Computer Science*, pages 270–282. Springer Verlag, Heidelberg, 2008.

[15] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

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

[17] W. Kraaij and T. Westerveld. How different are web documents? In *TREC-9*. NIST Special Publication, May 2001.

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

[19] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *WWW8*, pages 403–415. Elsevier Science, Amsterdam, 1999.

[20] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400:107–109, 1999.

[21] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM Press, New York, NY, USA, 2005.

[22] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1): 2, 2007.

[23] M. A. Najork, H. Zaragoza, and M. J. Taylor. HITS on the Web: How does it compare? In *SIGIR '07*, pages 471–478. ACM, New York, NY, USA, 2007.

[24] P. Ogilvie and J. Callan. Combining document representations for known-item search. In *SIGIR 2003*, pages 143–150. ACM Press, 2003.

[25] 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.

[26] J. R. Seeley. The net of reciprocal influence. *Canadian Journal of Psychology*, 3:234–240, 1949.

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

[28] J. Voss. Measuring wikipedia. In *ISSI 2005*, 2005.

[29] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*, volume 8 of *Structural Analysis in the Social Sciences*. Cambridge University Press, Cambridge MA, 1994.

[30] T. Westerveld, D. Hiemstra, and W. Kraaij. Retrieving web pages using content, links, URL's and anchors. In *The Tenth Text Retrieval Conference, TREC-2001*, pages 52–61, May 2002.