Wikipedia Pages as Entry Points for Book Search

Marijn Koolen
ILLC
University of Amsterdam, The Netherlands
m.h.a.koolen@uva.nl

Gabriella Kazai
Microsoft Research
Cambridge, UK
gabkaz@microsoft.com

Nick Craswell
Microsoft Research
Cambridge, UK
nickcr@microsoft.com

ABSTRACT

A lot of the world's knowledge is stored in books, which, as a result of recent mass-digitisation efforts, are increasingly available online. Search engines, such as Google Books, provide mechanisms for searchers to enter this vast knowledge space using queries as entry points. In this paper, we view Wikipedia as a summary of this world knowledge and aim to use this resource to guide users to relevant books. Thus, we investigate possible ways of using Wikipedia as an intermediary between the user's query and a collection of books being searched. We experiment with traditional query expansion techniques, exploiting Wikipedia articles as rich sources of information that can augment the user's query. We then propose a novel approach based on link distance in an extended Wikipedia graph: we associate books with Wikipedia pages that cite these books and use the link distance between these nodes and the pages that match the user query as an estimation of a book's relevance to the query. Our results show that a) classical query expansion using terms extracted from query pages leads to increased precision, and b) link distance between query and book pages in Wikipedia provides a good indicator of relevance that can boost the retrieval score of relevant books in the result ranking of a book search engine.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval—*Relevance feedback, Search process*; H.3.7 [**Information Search and Retrieval**]: Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Domain specific, Wikipedia, query expansion, link graph

1. INTRODUCTION

Libraries are traditionally regarded as the gateways to mankind's knowledge that has been accumulated throughout the centuries. The bulk of this knowledge is stored in the form of texts, printed in books. With the mass-digitization efforts of recent years, this knowledge is becoming increasingly available online, as collections of digitised books in digital libraries and on the Web. For example, the Million Book project[1] led by Carnegie Mellon, has scanned over 1.5 million books. Even more ambitiously, Google[2] has created a consortium of libraries, including the Harvard University Library system, and aims to digitise every book in their collections. The Open Content Alliance[3], a library initiative with over 80 contributing libraries, is offering a more transparent framework for mass digitisation and provides access to its digitised books over the Internet Archive.

A popular form of access to collections of digitised books is by online search services, such as Google Books and Amazon. These search engines provide mechanisms for searchers to enter this vast knowledge space using queries as the entry points. In this case, as it is traditional in IR, the user's query is matched against a representation (e.g., index) of a collection of books. The index may be built based on book content (e.g., full text), metadata (e.g., publication information, reviews, etc.), or a combination of both. As a result of the matching process, the books estimated relevant to the query are then returned to the user. A common trait of such traditional IR approaches is that the user's query is directly matched against the collection's representation (where features extracted from external resources may form part of the representation).

In this paper, we explore alternative retrieval approaches, which incorporate an intermediary between the user's query and the target collection being searched. The intermediary resource that we introduce into the retrieval framework is Wikipedia[4], an online collaborative encyclopedia. We build on the search scenario of a user looking for books on a given topic, where the unit of retrieval is the whole book. However, instead of matching the user's query directly against the collection of books, we aim to discover relevant books in the target collection by exploiting the content and link structure of the intermediary resource of Wikipedia.

Our goal is to incorporate additional sources of evidence extracted from Wikipedia, providing richer context to the user's information need, with the aim to improve the retrieval effectiveness of a book search engine. Our approach is motivated by the observation that encyclopedias in general can be regarded as summaries of all branches of knowledge[5], and thus can be viewed as entry points into the world of knowledge that is stored in books. We view Wikipedia as such a summary of human knowledge and each Wikipedia article as a window onto the knowledge space that is focused on a

[1] http://www.ulib.org/
[2] http://books.google.com/
[3] www.opencontentalliance.org/
[4] http://www.wikipedia.org
[5] See, for example, the Wikipedia article on encyclopedias.

WSDM '09, Barcelona, Spain

given topic of interest or branch of knowledge.

This view offers the potential to exploit the content of a Wikipedia article as a rich source of information for augmenting a user's query. Our first approach explores this prospect: Using established methods for query expansion, we build a richer description of the user's information need from Wikipedia articles that match the user's original query. We refer to these Wikipedia pages as *query pages*. Query pages are identified by using exact matching of the query string to the title of Wikipedia pages. We then use the expanded query to search our index of the book corpus.

Using the INEX 2007 Book Track test collection [16], we experiment with traditional query expansion techniques, but extracting terms from the intermediary resource instead of the target book corpus. The goal of these experiments is to answer the following research question:

- Can we automatically extract useful terms from related Wikipedia pages to improve the retrieval effectiveness of a book search system?

In addition to using Wikipedia for query expansion in book retrieval, our second approach aims to exploit the link structure of Wikipedia in order to connect a user's query directly with relevant books. This novel method is based on retrieving books that are cited by Wikipedia articles related to the user's query. This idea is motivated by the observation that Wikipedia pages often contain references to other information sources such as web pages, journal articles, as well as to books on the topic of the article. We make the intuitive assumption that the books cited by a Wikipedia page are related to the topic of the article and are thus relevant to a user's query on that topic. We associate Wikipedia pages that cite books with the cited books themselves as retrieval units. We refer to the citing Wikipedia pages as *book pages*.

Using the link structure of Wikipedia, we can obtain a full chain that connects a user's query with query pages and then a set of relevant books on book pages through related Wikipedia pages. The chain can also be extended (branched) to include books cited by inter-related articles. Assuming that articles close to each other in the link graph are topically related, we exploit this topical clustering effect to find references to further books. The research question we aim to explore is as follows:

- Is the link distance between *query pages* and *book pages* related to relevance and can we use this to the improve retrieval effectiveness of a book search system?

The main contribution of this paper is the introduction of Wikipedia within a book search scenario as an intermediary between the users' queries and the collection of books being searched. We exploit both the content and the link structure of Wikipedia with the aim to improve the retrieval effectiveness of a book search system.

The paper is structured as follows. In Section 2, we discuss previous research on book retrieval, query expansion, link analysis and the use of Wikipedia for IR. Section 3 analyses the coverage of user queries and books by Wikipedia pages, a necessary condition of the proposed approaches. Section 4 describes how we model Wikipedia and incorporate it as an intermediary resource within a retrieval framework. Experiments and results are described in Section 5. Finally, conclusions and future work are presented in Section 6.

2. RELATED WORK

In this paper we report on experiments that aim to improve the effectiveness of book retrieval approaches by exploiting Wikipedia as an intermediary source by applying query expansion and by traversing the link graph. To cover all these aspects, we divide our review of the related work into three sections: 1) full-text book retrieval, 2) query expansion using Wikipedia, and 3) exploiting Wikipedia's link graph.

2.1 Books as a New Vertical in Search

As a result of mass-digitisation efforts and evaluation initiatives, such as the Book Track at INEX[6], there is growing interest in full-text book retrieval challenges and opportunities (see e.g., [15]). Wu et al. in [28], for example, investigate the relative effectiveness of different types of book specific data, such as table of contents, back of book indexes, footnotes, and bibliography, using a multi-field inverted index. They show that table of contents and back of book indexes are prominent features for estimating relevance in book retrieval. Similarly, the results in [21] indicate that certain parts of books, particularly titles and headers (akin to titles and anchor texts in web search), are more valuable than other book parts for indexing. The table of contents and back of book indexes were also studied as searching and browsing tools in e-books by Abdullah and Gibb in [1]. Their study showed that the back of book index is a significantly more efficient user tool for finding information.

Whereas all these studies focus on the content of the book itself, in this paper we explore the use of Wikipedia as an external source of information for book retrieval.

2.2 Query Expansion

Traditionally, query expansion (through relevance feedback) is used to modify a user's initial query by, e.g., adding extra terms drawn from documents in the searched collection [11]. The goal is to arrive at an improved description of the user's information need, which then leads to the retrieval of additional relevant documents. The underlying principle is that potentially relevant documents that do not match any query terms may be found by a more descriptive query that contains additional, closely associated terms drawn from documents retrieved using the original query [26]. Typically, the documents used in the relevance feedback are from the target collection being searched.

The work of He et al. in [13] is different from this in that they compare relevance feedback from documents in the *target* collection, i.e., the collection from which the final results are returned to the user, with relevance feedback from documents in an external collection, i.e., a collection of documents only used for term selection, and find that both techniques improve on the baseline. However, they use two subsets of a larger newswire corpus as the target and external collections. In our case, the *target* and *external* collections are very different, both in nature and content. Our *target* collection (i.e., the INEX book corpus) contains very long texts written in most cases by a single *expert* author, whereas the *external* collection (i.e., Wikipedia) contains relatively short texts written and edited by many *expert* and *non-expert* authors.

The use of Wikipedia as an external resource for query expansion has been explored by several studies which have shown that this strategy can improve retrieval effectiveness. Li et al., in [20], for example, exploit Wikipedia as an external corpus to expand poor performing queries in the TREC Robust track. They take the top 100 Wikipedia articles retrieved for each query, re-rank them based on the number of articles sharing the same category, and expand the query with 40 terms from the top 20 documents. Somewhat similarly, in [2], Arguello et al. use Wikipedia as an external source for traditional pseudo relevance feedback. Using the original query, the top R Wikipedia articles are considered relevant and the top W articles are used to score anchor text for the links pointing to

[6]INEX: INitiative for the Evaluation of XML retrieval: `http://www.inex.otago.ac.nz/`

a document in the R set. Collins-Thompson & Callan [6] build a network of terms associated with the user's query terms by retrieving an initial set of Wikipedia articles and using an average mutual information measure to select highly associated terms. They then employ a random walk on these networks to "obtain probability estimates that a potential expansion term reflects aspects of the original query". Our approach differs from the above three methods in that we use only a single Wikipedia page (the query page) for query expansion, assuming that an encyclopedic article whose title matches the user's query exactly gives a more specific description of the topic and more precise book retrieval.

2.3 Wikipedia and Link Analysis

Wikipedia has become a popular subject of study in recent years. Its size and growth have been studied, e.g., in [27], as well as its topical coverage, e.g., in [10]. Others looked at Wikipedia's link structure as a complex social network [3, 5, 29]. Link analysis in general has been extensively studied in information retrieval (IR), e.g., [12, 18, 24]. It has been shown to be particularly effective in Web retrieval, using query dependent [17] and query independent evidence [23]. Analysis of the link graph of Wikipedia has been used more recently by Kamps & Koolen [14] to improve XML element retrieval performance on ad hoc topics for searching over the Wikipedia test corpus used at INEX, which is marked up in XML. They show that by zooming in on the local context of retrieved XML elements, i.e. the links between the top retrieved results, the number of incoming links can be used as an indicator of relevance to re-rank the result list.

Whereas most of these studies focus on the number of incoming links as relevance indicator, we use an approach more similar to the random walk model of Craswell and Szummer [8]. They use a random walk model to produce a ranking of documents for a given query using a graph of queries, documents, and user clicks. The queries and documents form the nodes in a bipartite graph, with the clicks representing the edges connecting the queries and documents. Using a fixed length random walk, documents that have more distinct paths to a certain query are assumed to be more related to the query, and thus receive higher scores than documents with less number of distinct paths. While Craswell and Szummer use the queries and documents as nodes, we map the queries and books to Wikipedia pages and use the Wikipedia link graph to compute closeness scores between query and book nodes. Provided this closeness score is related to topical relevance, it can be used to re-rank retrieval results or extend the results list with books not found by a more traditional retrieval approach.

3. WIKIPEDIA COVERAGE

In this paper we build on the idea that Wikipedia articles may be useful sources of information to intermediate between a user's query and a collection of books being searched. This, however, relies on implicit assumptions regarding the coverage of Wikipedia, both with respect to the topics of the user queries and the topics covered by a given collection of books. On the one hand, a necessary condition is that a user's query can be matched to relevant content in Wikipedia. On the other hand, it assumes that Wikipedia provides adequate coverage of the portion of the world's knowledge that is stored in the collection of books being searched by the user. To investigate these assumptions, in this section, we look at the relationship between user queries and Wikipedia pages and between Wikipedia articles and collections of books.

3.1 Wikipedia's Coverage of Search Topics

Similarly to printed encyclopedias, Wikipedia can be considered as a summarisation of our ever growing world knowledge [27]. Unlike printed versions, however, this online encyclopedia is collectively edited by its users. While this has led to questions about the trustworthiness of the resulting articles, it has made Wikipedia become the world's largest encyclopedia, with the English version consisting of over 2.4 million articles on all branches of knowledge. In comparison, the largest commercially available paper encyclopedia, Encyclopædia Britannica, has only 85,000 articles.

Given that the construction of Wikipedia is a collective effort, where the users themselves provide the content, it is reasonable to assume that the resulting Wikipedia articles describe topics of interest to the general public. At the same time, it is also plausible that the titles chosen for these articles reflect the type of queries that people may use when searching for information on a given topic.

To validate these intuitions and investigate the level of coverage between the topics that users search on and the topics they write about on Wikipedia, we matched a large sample of queries taken from a web search engine log (covering 10 months) against the titles of the English Wikipedia pages in the Wikipedia dump of 12 March 2008. We converted both queries and Wikipedia page titles to lower case and counted only exact string matches. Queries with a frequency of less than 4 were excluded from the logs to remove possibly unintended queries, i.e., misspellings and typing errors.

Table 1 provides details on the frequency distribution of the queries in the log data. From the sample of 5.76 billion queries, 2.19 billion (38%) are directly related to Wikipedia articles based on an exact match between the query string and the title of the Wikipedia page. Looking at the number of distinct queries, we see that only 1.5 million queries match Wikipedia pages on their titles out of a total of 114 million distinct queries. This indicates that the queries that match Wikipedia titles are the more frequently issued search queries. This is clearly visible when we look at the frequency of queries in the log. The average frequency of all distinct queries is 50.5, whereas the queries that match a Wikipedia title have an average frequency of 1,390. This is also visible in Figure 1, which plots the frequency distributions of a) all queries in the log, b) the queries that match Wikipedia page titles and c) both the previous distributions. In this last plot, it is clear that, with increasing query frequency, the two distributions show increasing overlap.

These findings suggest that Wikipedia topics provide good coverage over a high ratio of user queries. However, since the analysis here is over web query log data, our results to book search as a specialised domain may not be directly applicable. For instance, one could imagine that the pattern of user queries may vary in the book domain from the web domain in general. Currently, there is no published evidence to support or reject such a hypothesis. However, given the size of the query log studied here, our findings are expected to provide a reliable indicator of topical coverage for users searching over collections of books (which represent subsets of the available media on the Web). A further evidence supporting this is that from the 250 book retrieval queries that form part of the INEX 2007 Book Track test collection (see Section 5.1), 176 have matching Wikipedia page titles (i.e., 70.4%).

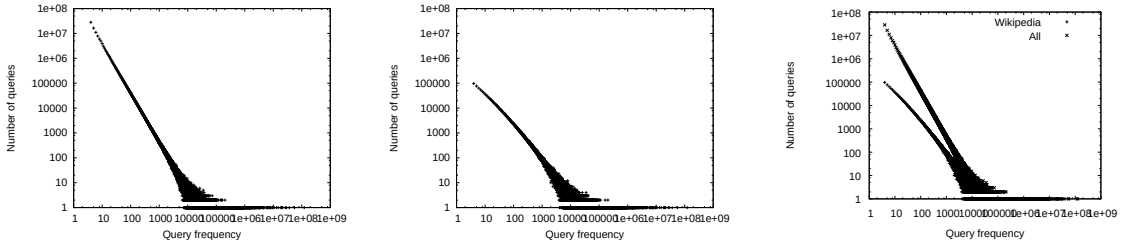
3.2 Wikipedia's Coverage of Topics in Books

In this section, we are interested in answering the question of 'How well does Wikipedia cover the topics found in books'? One way to answer this is to look at the books that are referenced in Wikipedia pages. When editing Wikipedia pages, editors are encouraged to add citations[7]. Cited sources can be, amongst others, web pages, journal or newspaper articles, videos, or books. We rea-

[7]See: `http://en.wikipedia.org/wiki/Wikipedia:Citing_sources`

Table 1: Web queries and Wikipedia titles overlap

			Query Frequency				
Queries	**Total**	**Distinct**	**Min**	**Max**	**Median**	**Mean**	**Stdev**
All queries	5,760,459,257	114,070,521	4	121,778,003	7	50.5	14,776
Wikipedia	2,194,678,715	1,578,602	4	121,778,003	32	1,390.0	125,157



(a) All queries (b) Queries matching Wikipedia titles (c) All queries and Wikipedia matches

Figure 1: Query frequency distributions.

sonably assume that such sources are directly or indirectly related to the citing Wikipedia page and thus to the topic of the page.

For example, consider the Wikipedia page on *Abraham Lincoln*. The page is extensive and contains references to many books (over 50) about various aspects of Abraham Lincoln's life. These cited sources are likely to be relevant to a query on Abraham Lincoln. A question however concerns the coverage of the cited sources within a given collection of books that is indexed and searched by a system. Furthermore, not all topics will have such an extensive page with so many references, and for many topics, their Wikipedia page may not even include any citation at all.

To investigate this, we parsed the 6.5 million pages in our Wikipedia dump and extracted 240,099 book citations from 96,256 Wikipedia pages (2.49 citations per page) based on the Wikipedia cite-book template (introduced in 2005). Although the Wikipedia dump includes many templates, images, redirects and meta-pages, the percentage of pages containing citations to books is still very small: According to the Wikipedia website at the time of the dump, the English Wikipedia contained roughly 2.3 million pages. This means that for 96% of the topics found on Wikipedia, no related books are cited. This may then limit the effect of using direct linkage to books for improving retrieval effectiveness.

A related study in [22] compared the use of citations to scientific journals in Wikipedia pages against journal citation statistics from *Journal Citation Reports* (JCR) with the aim to measure the quality of Wikipedia as an information organiser for science. Although citations in Wikipedia were not found to reflect the JCR impact factors, high correlation with the total number of citations to each journal was reported. Since the introduction of the citation template in Wikipedia in February 2005, the number of citations to journals and books has grown rapidly, thereby increasing the confidence in Wikipedia as a well-organised pointer to further information.

An alternative approach to the above is to compare the topics covered by both Wikipedia and a given book collection. This question has been investigated extensively in [10]. Halavais & Lackaff completed a detailed survey and analysis of how different subject areas, given by the Library of Congress classification (LCC) system, are covered by Wikipedia articles, compared to the topical distribution of *printed* books listed in the Bowker's Books In Print[8] catalogue. They found high coverage (counted by number of articles on each topic) between books and Wikipedia for areas like *Music* and *Fine arts* and reasoned that this was due to Wikipedia's coverage being driven by popular interests. High coverage was also found for other areas, like *History* and *Geography*, which is due to whole collections of census data being automatically imported into Wikipedia. On the other hand, the social sciences, and more expert areas like *Law* and *Medicine* were found to be underrepresented in Wikipedia compared to printed books.

If we look at the number of articles about each topic, some areas seem less covered by Wikipedia than by printed books. However, Halavais & Lackaff also point out that if we look at the length of Wikipedia pages, we find that these "underrepresented" areas (law, medicine and social sciences) contain the longest articles. With the enormous amount of entries and the fact that both the number of topics in Wikipedia and the citations to external sources are growing rapidly, Wikipedia promises to be a good intermediary resource to enhance traditional IR approaches.

3.3 Wikipedia Link Structure

In order to study the relation between the relevance of a book to a query and the link distance from the query entry page to the Wikipedia page that cites a book, we need to look at the connectedness of the Wikipedia link graph. If the link graph consists of many unconnected sub graphs, this will impact upon the measure of link distance between two pages.

Following Broder et al. [4], we study the connectedness of the Wikipedia link graph, see Figure 2 and Table 2. The total graph contains 6,552,490 pages (including templates, image descriptions and meta-pages) and 84,970,770 links. The giant Strong Connected Component (SCC), defined as the set of nodes that can be reached from any other node by following links, contains 3,448,104 pages. The IN set, containing pages with a path *to* the SCC, has just under 2.3 million nodes. The OUT set, containing pages with a path *from* the SCC, has only about 400,000 nodes. The giant Weak Connected Component, combining SCC, IN and OUT components, consists of 94% of all Wikipedia pages, showing that the vast majority of pages are connected. What is interesting to see, is that the IN component is far larger than the OUT component, whereas in the Web, these sets are much more balanced. A possible explanation may be found by looking at the distribution of the number of links pointing to and from Wikipedia pages.

Figure 2 shows the distributions of the number of articles over the number of incoming links (a) and the number of outgoing links (b). We observe power-law distributions in both cases, which is

[8] http://www.booksinprint.com

Table 2: Statistics on the Wikipedia link structure

Description	# nodes
Total links	84,970,770
Total nodes	6,552,490
nodes in SCC	3,448,014
nodes in IN	2,292,363
nodes in OUT	401,954
Unconnected	410,159

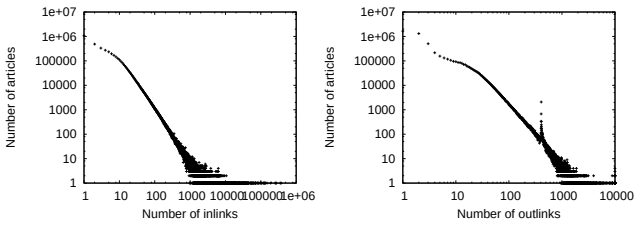


Figure 2: Link degree distributions.

typical of scale-free networks like Wikipedia [5] and the Web [9], that is, most pages have only a few incoming links and only a few pages have hundreds of thousands of incoming links, and similarly for the outgoing links.

The link graph of Wikipedia is very dense, with many pages containing thousands of links to other pages and several pages with more than 100,000 incoming links. This makes the graph very compact, requiring just a few steps to go from one page on a given topic to another on an entirely different topic. The pages with the most incoming links are often the so-called *year* pages, describing events that took place in a certain year. On many Wikipedia pages dates are often automatically linked to such a *year* page, which would explain why the IN component is so much larger than the OUT component. This makes the graph well connected, but a consequence is that these date pages often connect many topics with each other that are otherwise unrelated. This may then adversely impact on a possible correlation between the topical similarity of two pages and their link distance, which may then also affect retrieval approaches that incorporate link distance as a feature.

4. WIKIPEDIA AS INTERMEDIARY FOR BOOK SEARCH

As mentioned earlier, we view Wikipedia as a window onto the world's knowledge stored in books and thus as a useful resource to mediate between a user's search request and relevant books in a book corpus. Our aim is to investigate different ways of exploiting this relationship to improve retrieval performance. This section describes our two main strategies for using Wikipedia as an intermediary for book search. We build on the assumption of sufficient coverage between Wikipedia articles and the users' topics of interest, and between Wikipedia and knowledge stored in books (see previous section).

4.1 Query Expansion

We start by connecting a user query to related Wikipedia articles by matching exactly the query string and the article's title. We consider these *query pages* as entry points to the user's topic of interest, providing a rich description of the topic, complete with category tags, and a network of related Wikipedia pages. In order to make use of a Wikipedia article on the topic of the user's query, we aim to extract useful terms from these *query pages* to expand the query, i.e., terms that describe the topic complementarily to the terms in the original query. As mentioned in Section 2.2, several studies have shown the effectiveness of using Wikipedia for query expansion for a range of different tasks and collections. All these approaches ([2, 20, 6]) use the top n retrieved Wikipedia pages to extract terms or phrases for expansion. Our approach differs in that we exploit the overlap between search queries and titles of Wikipedia articles to select one entry page for query expansion, to keep more focus on the topic of the query. Although using multiple articles might give a larger vocabulary of related terms to choose from, often exploited to improve recall, our choice of using a single article specifically about the topic of the query is hoped to lead to related terms closer to the original query terms and thus result in increased precision. In addition, we reason that Wikipedia articles are often edited by multiple authors, who together have a larger vocabulary than each author individually.

In order to extract useful terms for query expansion from the associated query page in Wikipedia, we use the well known $tf.idf$ formula with the aim to select the N terms that best discriminate a given topic page from the rest of the Wikipedia articles.

The $tf.idf$ score of a term t is calculated as:

$$tf.idf(t) = \frac{tf_d(t)}{|d|} * log(\frac{D}{df(t)}) \quad (1)$$

where $tf_d(t)$ is the frequency of term t in document d, $|d|$ is the length of document d, D is the total number of documents in the collection, and $df(t)$ is the number of documents containing term t. An advantage of using query pages as entry points for query expansion is that we can pre-compute the best terms for every Wikipedia page, thus making query expansion very fast at query time. As often done in query expansion, we use term weights to place more emphasis on the original query terms in the expanded query. We employ a simple term weighting method, whereby the original query terms are weighted N times more than the N added terms. That is, if 5 terms are added, the original query terms each receive a weight of 5 while the added terms each get a weight of 1.

We note that we have also experimented with a normalised weighting scheme to compensate for queries of different lengths, but this led to disappointing results. The weighting scheme built on the $tf.idf$ scores to differentiate between the added terms, and normalised these to sum to 1. The original query term weights were also normalised by dividing by the number of terms.

4.2 Modeling Topical Closeness

In this section, we introduce the concept of topical closeness and look at how it is related to the notion of relevance. Our goal is to employ topical closeness as an indicator of a cited book's relevance.

We measure topical closeness as the link distance between two points of the Wikipedia link graph. The two points we are interested in are given as the query page whose title matches a user's query and the book page that cites a book in the collection being searched.

Assuming that a Wikipedia page matching a query string is a description of the topic of the query, we follow a user aiming to find references to relevant books. The user starts at the *query page* and traverses the link graph by clicking on links to other Wikipedia pages. We hypothesise that books cited by pages that are close to the *query page* are more closely related to the topic than books cited by pages at a greater distance in the link graph. This is supported by the finding that closely neighbouring pages, i.e., pages that link to or are linked from a given page, are often related to each other [14]. Traversing the link graph further, the topics of the pages become more diverse and likely less related to the topic at hand. In other words, with increasing distance from the *query page*, we expect to find less citations to relevant books. We thus aim to explore the use

of this distance as a notion of "closeness" to reflect the relevance of cited books to the topic of the *query page*.

As we mentioned already, the query pages are identified through exact string matching between the user's query string and the title of Wikipedia articles. What is now needed is a way in which book pages can be identified. In the following sections, we propose three different methods that exploit different properties of Wikipedia and the increasing use of citations in Wikipedia articles.

4.2.1 Linking by Citations

To connect citations on Wikipedia pages with books in our target collection, we first extract all book citations from Wikipedia and match them against the books in the INEX 2007 Book Track collection. A citation is considered a match with a book in our collection if both the book title and the authors match. Since author names usually vary in orthographic representation (e.g. *W./William H./Harrison Ainsworth*), we use n-gram matching to find similar names. This resulted in 2,494 matches between 1,382 books in the collection and 696 pages in Wikipedia (a book may be cited by multiple pages).

One of the reasons for the low number of matches is that many of the books in the collection are books out of print, published before 1930, while most of the books cited by Wikipedia pages are published after 1970. We use two alternative mechanisms to match more books with Wikipedia pages.

4.2.2 Linking by LCC Labels

We use the Library of Congress Classification (LCC) labels assigned to the books in the INEX collection as links to Wikipedia. For this, we exploit the fact that there are Wikipedia pages dedicated to many of the topics classified by the LCC system (i.e., the class *PF – West Germanic Languages* links directly to the page *West Germanic Languages*). We can associate the books that have LCC labels with the Wikipedia page on that topic. This leads to 512 Wikipedia pages matching 20,682 books in the INEX collection (the other books in the collection have no or erroneous LCC labels). Since these matches are based on more general topics, with on average over 40 books associated with each of the 512 Wikipedia pages, we expect this to be a much more noisy approach than using the explicit citations described above.

Figure 3 shows the distribution of topics as defined by the LCC labels covered by the books in the INEX collection and the subset of these cited by Wikipedia pages. As it can be seen, most of the (available) LCC labels associated with books in the INEX collection are *P – Literature*, *B – Philosophy* and *E – History*. The distribution of topics covered by the cited books in Wikipedia is similar to the distribution of topics covered by the entire INEX book collection, although there are, for example, no matches in some categories (e.g., in category A, which covers general books like dictionaries, encyclopedias and periodicals among others, category R – Medicine, or category U – Military Science).

4.2.3 Linking by Document Similarity

We employ a document similarity measure to associate a book from the INEX collection with the Wikipedia page that is most similar to it in content. However, computing scores of document similarity between a whole book and every single Wikipedia page would be an expensive step. Instead, we use a shortcut by representing books using only their top 50 terms, i.e., with the highest $tf.idf$ scores, and we match these as query terms against the full-text index of all Wikipedia articles. To keep memory requirements limited, we base the $tf.idf$ scores on sets of 500 books at a time. This way, we obtain ranked lists of Wikipedia pages for each book in our collection. We then associate each book with its top ranked

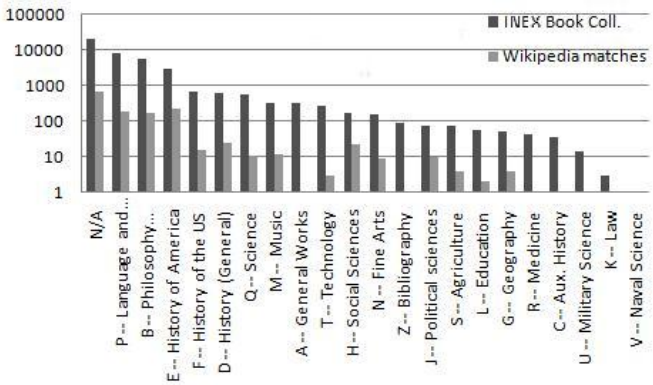


Figure 3: Distribution of topics (as LCC labels) covered by books in the INEX 2007 Book Track collection and books cited by Wikipedia pages.

Wikipedia page. Each such Wikipedia page, thus, represents a book page that is linked to the book with similar content.

4.2.4 Calculating Closeness Scores

We apply a random walk model [8] to compute closeness scores between a *query page* and the *book pages*, using transition probabilities based on the number of links. The probability of going from node j at step s from the *query* node to node k is computed as:

$$P_{s+1|s}(k|j) = P_{s|s-1}(j) * \frac{l_{jk}}{l_j} \quad (2)$$

where l_{jk} is the number of links from node j to node k, l_j is the total number of links from node j and $P_{s|s-1}(j)$ is the probability of being at node j after step s.

Experimentally, we find that with 4 steps, 95% of the books are reached via at least one path. Furthermore, given the size and density of the link graph, computing scores for long random walks is an expensive operation. Since almost all of the *book pages* can be reached within 4 steps, we restrict the maximum path length to 4 in all our experiments.

5. EXPERIMENTS AND RESULTS

To evaluate our proposed approaches to incorporate Wikipedia as an intermediary resource within a retrieval framework, we conducted a series of experiments on the INEX 2007 Book Track data. We report on these experiments and on our findings in this section.

5.1 Experimental Setup

For all our experiments, we used the test collection provided by the INEX 2007 Book Track. This corpus consists of over 42,019 out-of-copyright books, totalling around 210GB [16]. 39,176 books in the collection have associated MARC (MAchine Readable Cataloguing) records containing publication and classification information, where LCC labels are available for 20,692 books.

The test collection contains 250 user queries, extracted from the query log of a commercial book search engine. Out of these 250, 176 queries directly matched titles of Wikipedia pages. We thus use this subset for all our experiments.

The relevance assessments for the 250 queries of the test collection were collected at the book level from paid human judges. Assessment were made along a four point scale: *Excellent*, *Good*, *Fair* and *Non-relevant*. We have transformed these graded judgements into binary judgements to be able to use standard measures

Table 3: Results for query expansion using the top N $tf.idf$ terms. Significance levels are 0.05 (*), 0.01 (**), and 0.001 (***)

Run id	# jugded rel.	# jugded non-rel.	MAP	bpref	P@10
baseline	1666	808	0.3771	0.6131	0.3040
$N = 5$	1666	808	0.3725	**0.6205**	0.3080
$N = 10$	1671	808	**0.3874****	0.6168	0.3119*
$N = 20$	1667	807	0.3837**	0.6149	**0.3136*****
$N = 50$	1666	806	0.3780	0.6136	0.3074*
$N = 100$	1666	807	0.3780**	0.6133	0.3063***

like Mean Average Precision (MAP), Precision at rank 10 (P@10) and binary preference (bpref). An explanation of these measures can be found in [25]. All labels *Excellent* and *Good* are assigned the relevance score of 1, while all labels *Fair* are assigned the relevance score of 0.

5.2 Query Expansion

For the query expansion experiments, we used Indri[9] for indexing and retrieval, with no stopwords removal, content words stemmed using the Porter stemmer and default values for smoothing. All XML structure within the books was ignored.

We compare the Indri baseline run, based on the 176 queries in their original form, with 5 runs obtained using the expanded queries. We experimented with adding 5, 10, 20, 50 and 100 terms to the original query, with results shown in Table 3. Differences with respect to the baseline were tested for significance using the bootstrap test (one-tailed, significance levels are 0.05 (*), 0.01 (**), and 0.001 (***)). What is interesting to see is that the expanded queries lead to improvements in P@10, but the number of relevant books found in total is very similar across all runs, including the baseline. Only for $N = 10$ is the number of relevant books retrieved in the top 1000 results slightly higher than the baseline. However, finding only an additional 5 relevant books over 176 topics shows that for most topics no new relevant books are found (even though the total number of relevant books for the 176 topics is 1,859). Thus, query expansion leads to higher precision but not better recall. This is likely an effect of limiting the source document where additional terms are drawn from to the query page in Wikipedia. Adding 10 or more terms leads to improvements in MAP as well, but with larger N, i.e. 50 or more, the improvements decrease. This is most likely due to the weighting scheme we applied. With large N, the relatively very low weights for the added terms curb the impact these terms might have. Thus, although adding 10 terms leads to the best MAP score and adding 20 terms leads to the best P@10 score, a more balanced weighting scheme might show that adding more than 20 terms may be more effective.

Looking at bpref, we see that adding terms improves performance, but smaller N gives better results. This might also be due to the weighting scheme. In general, adding terms improves early precision and bpref, but this particular weighting scheme makes the impact of query expansion smaller with larger N. However, the results also show that using a single *query page*, matched only on the title, provides a description of the topic that can be used to improve both early and overall precision with query expansion.

5.3 Topical Closeness

We now turn to the question of whether link distance between a query page and pages citing relevant books is related to relevance.

We obtained closeness scores from 176 query pages to 696 *book pages* citing books from the INEX collection based on the linking by citation method (see Section 4.2.1), and to 512 *book pages* identified based on the linking by LCC labels approach. For the document similarity method we experiment with associating books with the top N ranked Wikipedia pages, with N set to 1, 3 or 5. Since books can cover multiple topics, they can be associated with each of those topics in the Wikipedia collection. Thus, we want to know if multiple *book pages* per book could lead to better closeness scores.

The results presented in this section are all based on 4-step forward walks, i.e., following links from the linking page to the linked page. Using a backward walk, i.e., following links from the cited page to the citing page, leads to very similar results.

As discussed in Section 3.3, not all Wikipedia articles can be reached from a given page, e.g., a *query page*, by following the links, and as a consequence, not all cited books will receive a closeness score. For the citation pages (i.e., book pages based on linking by citation method), we obtained an average of 1,207 closeness scores per topic, which means that with our random walk we find 1,207 books per topic on average[10]. Given that 1,382 books in the INEX collection are cited on Wikipedia, on average 87% of them receive a closeness score for each topic. For the LCC pages (i.e., book pages based on the LCC method), we obtained an average of 18,796 closeness scores per topic (representing 91% of all books associated with LCC pages), and for document similarity based *book pages*, we obtained 24,036, 34,764, and 37,862 closeness scores (57%, 83% and 90% of all books in the INEX collection) per topic, respectively for 1, 3 and 5 *book pages* per topic.

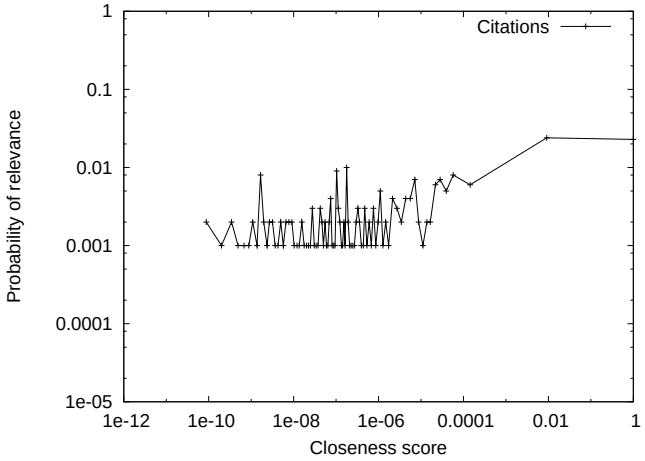
We investigate whether the obtained closeness scores are related to topical relevance. To this aim, we first turn these scores into probabilities of relevance. Recall that each score connects a query to a book in the INEX collection (via the matching query and book pages in Wikipedia), where some scores connect queries with books that are known to be relevant for that query. In other words, each query/book pair that has a closeness score has a relevance score as well. Judged pairs have an explicit relevance score, provided within the relevance assessment set of the INEX Book Track test collection. Un-judged pairs are assumed irrelevant, giving an implicit relevance score. Based on this, our hypothesis can be stated as: if the closeness score is positively related to relevance, then the relevant query/book pairs should on average have higher closeness scores than non-relevant pairs.

We first sort the query/book pairs (across all 176 topics) into bins of equal size with ascending closeness score. Each bin contains 100,000 pairs (10,000 for the citation scores since we have less data there) and the probability of relevance for these pairs is computed by dividing the number of relevant pairs in a bin by the total number of pairs in that bin. That is, the first 100,000 pairs with the lowest closeness scores go into the first bin, the next 100,000 in the second bin, etc. Figure 4 shows the probability of relevance over ascending closeness score (each bin is represented by the mean of the closeness scores in that bin) for closeness scores based on citations, LCC pages and document similarity, respectively. The closeness score for a book associated with multiple nodes in the graph is just the sum of the closeness scores of the *book pages*.

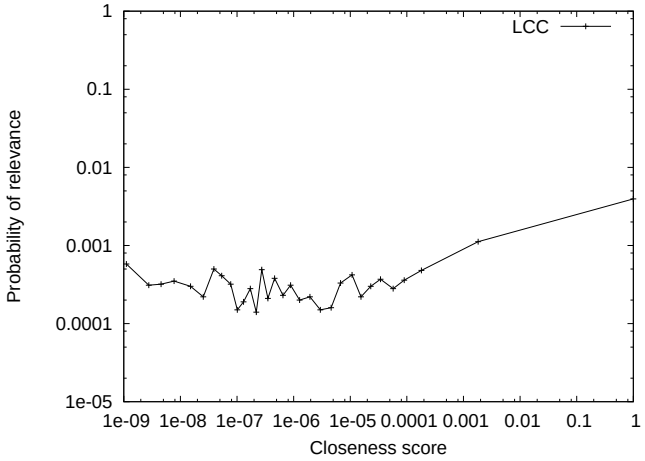
With citation and LCC based closeness scores up to 0.0001, there seems to be no relation with the probability of relevance. With higher scores, i.e. above 0.0001, we do see a relation. As the closeness score increases, so does the probability of relevance. With the citation based scores the relation seems to be weaker than with the LCC based scores. One reason might be that the citation method

[9] http://www.lemurproject.org/indri/

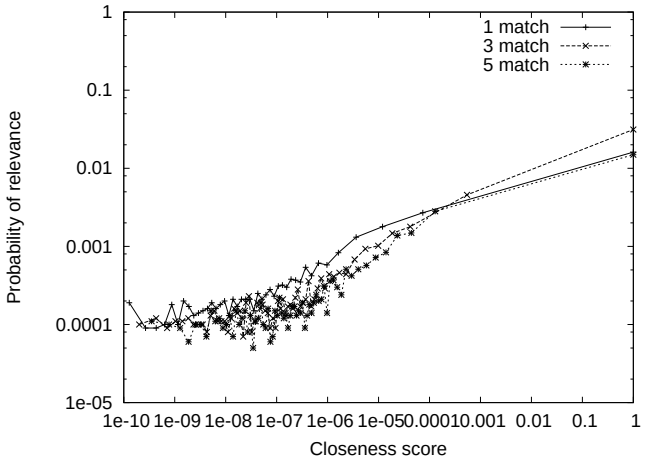
[10] Note that multiple books that are cited from the same book page receive the same closeness score, since the link distance from the query page to the book page is the same.



A) Closeness scores based on citations



B) Closeness scores based on Loc topics



C) Closeness scores based on document similarity

Figure 4: Probability of relevance over closeness scores for LCC matches and citation matches.

matches only a small number of books, with an even smaller number of them being relevant to any topic. Using the explicit citations works to some extent, but the few citations that match books in the INEX collection are not a good representation of the topics closely related with the *query page*. The more general LCC method matches much more books and also far more relevant books, giving a better representation of the topics around the *query page*.

If we look at the closeness scores based on document similarity, we again see no relation between closeness and relevance for scores up to 1e-5, whereas for scores above 1e-5, the probability of relevance increases with increasing closeness scores. What is interesting to see is that there is not much difference between the scores using 1, 3 and 5 *book pages*. However, compared to the other two methods, the scores based on document similarity seem to show a stronger relation with relevance (steeper curve). One reason why closeness based on document similarity shows a stronger relation with relevance, than closeness based on LCC, is that in the former the *book pages* are matched to books using the actual content of the book. The LCC topics are fairly broad, so all books on, e.g. American history, get associated with the same Wikipedia page. With document similarity, books with general topics can be matched to Wikipedia pages about the same broad topics, and books with specific topics can be associated with Wikipedia pages of similar specificity.

In general, it seems that the closeness score of 1e-5 represents an important point on the closeness scale. What is of interest then is the number of steps in the walk that this score relates to. If we take the document similarity based scores as an example, we find that books first encountered in the second step receive an average closeness score of 6.39e-05 at that step. At the third step, this is 1.22e-06. This means that on average, beyond the first two steps, any two books have a more or less equal chance of being relevant. In other words, within the first two steps, the books associated with nodes closer to the *query page*, are more likely to be relevant.

The observed relationship between closeness scores and relevance enables the effective use of closeness scores to re-rank our original retrieval results. Many papers describe ways to combine multiple sources of evidence for retrieval. Kraaij et al., in [19], experimented with estimating document priors for document length and link evidence in Entry Page search, using either general modelling assumptions or training data. They found a linear relation between the number of incoming links and the probability of relevance, which could be exploited to improve retrieval performance. Craswell et al. [7] guessed transformation functions from looking at distributions of log odds estimates for different features. Url length, link indegree and click distance (the minimum of clicks to the page from a root page) were modelled by sigmoid functions, leading to substantial improvements when combined with a BM25 baseline.

We choose a standard sigmoid function for the transformation:

$$sigmoid(b,q) = \frac{1}{1+e^{-cl(b,q)}} \quad (3)$$

where $cl(b,q)$ is the closeness score for book b and query q. The sigmoid function ensures that at the low end of the distribution, where there is no relation to relevance, the closeness scores are transformed to values very close to 0.5 (a closeness score of zero would be transformed to 0.5). Close to 1, the closeness scores rapidly increase to 0.73. Thus, only the books at the high end of the distribution receive a boost. We combine this with Indri's retrieval score by simple addition:

$$S(b,q) = Indri(b,q) + sigmoid(b,q) \quad (4)$$

Table 4 shows the results for our re-ranked runs. With the same

Table 4: Results for combination of Indri scores and sigmoid transformation of closeness scores. Significance levels are 0.05 (*), 0.01 (**), and 0.001 (***)

Run id	MAP	bpref	P@10
baseline	0.3771	0.6131	0.3040
Citation	0.3769	0.6150	0.3051
LCC	0.3445	0.6109	0.2756
Doc.Sim.1	0.3604	0.6010	0.2983
Doc.Sim.3	0.3790	0.6245*	**0.3091***
Doc.Sim.5	**0.3823***	**0.6251*****	0.3080*

baseline as in the query expansion experiments, we see that the citation based closeness scores lead to minor improvement on bpref and P@10, but the differences are so small that it seems the closeness score has almost no effect. This minor impact is to be expected, however: given the small number of cited books, only a few books receive a relative boost. The LCC based scores lead to a decrease in performance, possibly because these associations are based on much more general topic classifications that dilute the book set and thus boost possibly irrelevant books.

As the probability of relevance plots already suggested, the document similarity scores are better indicators of relevance, leading to small but significant improvements for books associated with multiple pages. Even though the closeness scores for books associated with a single Wikipedia page show some relation with relevance, its impact only hurts performance. The results show that multiple *book pages* per book better reflect the topical coverage of the books. Using either 3 or 5 Wikipedia pages per book, performance improves for MAP, bpref and P@10, with more pages leading to further improvement, except for P@10, where 3 pages work better than 5 pages. Intuitively, this makes sense. A book covering only a specific topic will be associated with multiple Wikipedia pages related to this specific topic and therefore these pages will be closely related with each other as well. If all these *book pages* are close in the link graph to the *query page*, the book will receive a very high closeness score (possibly above 1.0) and will subsequently receive a substantial boost in the final ranking.

To sum up, for all three methods, using citations, LCC labels or document similarity, the closeness scores between a *query page* and a *book page*, show a relation to relevance only up to a small number of steps in the link graph. The citations lead to *book pages* that are plausibly closely related to the topic of the book. However, the small number of *book pages* based on citations – due to the INEX book collection containing mainly out of print books – restricts the impact of this approach and thus its applicability to this test collection. The LCC *book pages* are on very general topics and, although they cover a much larger part of the INEX collection, are unable to distinguish between books within a fairly general classification and provide no help in identifying relevant books for more specific information needs. With document similarity, exploiting both the possibility to use multiple *book pages* to cover possibly multiple sub-topics in a book, and the possibility to associate *all* the books in the collection with a number of *book pages*, the closeness score can be used as query dependent evidence, complementary to evidence based on more traditional keyword based approaches.

6. CONCLUSIONS

In this paper we investigated ways to use Wikipedia as an intermediary resource for book search. Using the fact that many topics have a dedicated Wikipedia page, we looked at ways to use these rich sources of information to find additional terms to describe a user's topic of interest and use these for query expansion. Our first research question was:

- Can we automatically extract useful terms from Wikipedia entry pages to improve the retrieval of relevant books?

Using a single query page in Wikipedia, which is specifically about the user's search topic, and employing a term selection strategy based on $tf.idf$ weighting, our query expansion approach is able to keep focus on the topic, leading to significant improvements in early precision and MAP. Adding 10-20 terms from the *query page* of the topic leads to the best performance. Adding more terms still has a positive effect, but the improvements decrease, possibly due to the weighting scheme placing increasingly too much emphasis on the original query terms.

Next, we associated the books in the INEX Book Track collection with Wikipedia pages that either cite these books, cover the topic of the Library of Congress classification for these books or are the most similar in content. It is generally assumed that pages close to each other in the link graph are topically related. With both queries and books connected to Wikipedia pages, we wanted to know:

- Is the link distance between query pages and book pages related to relevance?

We modelled the closeness of a book to a query by computing transition probabilities between pages using a 4-step forward random walk model. We found that at the higher end of the closeness score distribution, roughly corresponding to articles at a distance of up to 2 steps from the *query page*, there is a clear relation between transition probabilities and the probability of relevance.

Due to the low number of books in the INEX collection that are cited in Wikipedia, using their closeness scores to re-rank results had little overall effect on retrieval effectiveness. The LCC pages cover many more books in the INEX collection, but at a much more general level, resulting in large groups of books, with varying subtopics, receiving the same closeness score, subsequently leading to a decrease in retrieval performance. Associating books with Wikipedia pages based on document similarity leads to the best overall results. Although picking a *single* Wikipedia page was found to be ineffective, using multiple pages leads to small but significant improvements in early precision and MAP.

6.1 Future Research

In our future work, we will look at several aspects of the closeness scores. We aim to study the impact of Wikipedia's link density local to a *query page*. With only a few neighbouring pages, the closeness scores will remain relatively high, which could have an impact on the relation between the closeness score and relevance, suggesting the need for a different transformation function. As the experiments in this paper show that retrieval performance improves when more *book pages* are used, we also want to investigate how the number of *book pages* affects the quality of the closeness scores. The optimal number of *book pages* may be book dependent, with books on specific topics benefiting from a smaller number of *book pages*. We will also look at other ways of measuring document similarity between books and Wikipedia pages. Since directly comparing the whole book content with every page is prohibitively expensive, we could instead look at term selection methods other than $tf.idf$, use collocations, that is, sequences of terms that co-occur more than would be expected by chance, instead of single terms, or apply latent semantic analysis.

With the experiments presented here we explored only a few possible ways to use Wikipedia as an intermediary resource, showing that Wikipedia pages can be used effectively as entry points to provide a richer context for retrieval. With the increasing number of citations of books, and the dense link graph connecting related pages

on a growing number of topics, Wikipedia is likely to become an increasingly valuable resource to mediate between users' information needs and knowledge stored in books.

We also note that during our analysis of a large Web log, we found substantial overlap between typical queries issued by Web users and the titles of Wikipedia pages. This suggests that the use of Wikipedia as an intermediary can be applied to areas in IR beyond book search. Many Wikipedia pages, for example, cite journal and newspaper articles, web pages and point to multimedia objects.

7. REFERENCES

[1] N. Abdullah and F. Gibb. Using a Task-Based Approach in Evaluating the Usability of BoBIs in an e-Book Environment. In *Proceedings of the 30th European Conference on Information Retrieval, Glasgow*, volume Lecture Notes in omputer Science, Vol. 4956, pages 246–257. Springer-Verlag, 2008.

[2] J. Arguello, J. L. Elsas, J. Callan, and J. G. Carbonell. Document representation and query expansion models for blog recommendation. In *Proceedings of the Second International Conference on Weblogs and Social Media (ICWSM 2008) 2008*, 2008.

[3] F. Bellomi and R. Bonato. Network Analysis for Wikipedia. *Proceedings of Wikimania*, 2005.

[4] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

[5] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: the case of Wikipedia. Physical Review E, Feb 2006.

[6] K. Collins-Thompson and J. Callan. Query expansion using random walk models. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 704–711, New York, NY, USA, 2005. ACM.

[7] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 416–423, New York, NY, USA, 2005. ACM.

[8] N. Craswell and M. Szummer. Random walks on the click graph. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 239–246, 2007.

[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262. ACM Press, New York NY, USA, 1999.

[10] A. Halavais and D. Lackaff. An Analysis of Topical Coverage of Wikipedia. *Journal of Computer-Mediated Communication*, 13(2):429–440, 2008.

[11] D. Harman. Relevance feedback revisited. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1–10, New York, NY, USA, 1992. ACM.

[12] D. Hawking. Overview of the TREC-9 Web Track. In *TREC*, 2000.

[13] D. He and Y. Peng. Comparing two blind relevance feedback techniques. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 649–650, New York, NY, USA, 2006. ACM.

[14] J. Kamps and M. Koolen. The Importance of Link Evidence in Wikipedia. In *Proceedings of the 30th European Conference on Information Retrieval, Glasgow*, volume 4956 of *Lecture Notes in Computer Science*, pages 270–282. Springer Verlag, Heidelberg, 2008.

[15] P. Kantor, G. Kazai, N. Milic-Frayling, and R. Wilkinson, editors. *BooksOnline '08: Proceeding of the 2008 ACM workshop on Research advances in large digital book repositories*, New York, NY, USA, 2008. ACM.

[16] G. Kazai and A. Doucet. Overview of the INEX 2007 Book Search track. *SIGIR Forum*, 42(1):2–15, 2008.

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

[18] W. Kraaij and T. Westerveld. How Different are Web Documents? In *Proceedings of the ninth Text Retrieval Conference, TREC-9*. NIST Special Publication, May 2001.

[19] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 27–34, New York, NY, USA, 2002. ACM.

[20] Y. Li, W. P. R. Luk, K. S. E. Ho, and F. L. K. Chung. Improving weak ad-hoc queries using Wikipedia as external corpus. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 797–798, New York, NY, USA, 2007. ACM.

[21] W. Magdy and K. Darwish. Book search: indexing the valuable parts. In *[15]*, pages 53–56, New York, NY, USA, 2008. ACM.

[22] F. Å. Nielsen. Scientific citations in Wikipedia. *First Monday*, 12(8), 2007.

[23] 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.

[24] A. Singhal and M. Kaszkiel. A case study in web search using TREC algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 708–716, New York, NY, USA, 2001. ACM.

[25] C. Tre. *Common evaluation measures*. The Twelfth Text REtrieval Conference (TREC 2003), 2003.

[26] C. J. Van Rijsbergen. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979.

[27] J. Voss. Measuring Wikipedia. In *Proceedings International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden, 2005.

[28] H. Wu, G. Kazai, and M. Taylor. Book search experiments: Investigating ir methods for the indexing and retrieval of books. In *Proceedings of the 30th European Conference on Information Retrieval, Glasgow*, volume 4956 of *Lecture Notes in Computer Science*, pages 234–245. Springer Verlag, Heidelberg, 2008.

[29] V. Zlatic, M. Bozicevic, H. Stefancic, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. Physical Review E, Jul 2006.