OUC's participation in the 2009 INEX Book Track

Michael Preminger[1], Ragnar Nordlie[1], and Nils Pharo[1]

Oslo University College

Abstract. In this article we describe the Oslo University College's participation in the INEX 2009 Book track. This year's tasks have been featuring complex topics, containing aspects. These lend themselves to use in both the book retrieval and the focused retrieval tasks. The OUC has submitted retrieval results for both tasks, focusing on using the Wikipedia texts for query expansion, as well as utilizing chapter division information in (a number of) the books.

1 Introduction

In recent years large organizations like national libraries, as well as multinational organizations like Microsoft and Google have been investing labor, time and money in digitizing books. Beyond the preservation aspects of such digitization endeavors, they call on finding ways to exploit the newly available materials, and an important aspect of exploitation is book and passage retrieval.

The INEX Book Track, which has now been running for three years, is an effort aiming to develop methods for retrieval in digitized books. One important aspect here is to test the limits of traditional methods of retrieval, designed for retrieval within "documents" (such as news-wire), when applied to digitized books. One wishes to compare these methods to book-specific retrieval methods.

One of the aims of the 2009 Book Track experiments[1] was to explore the potential of query expansion using Wikipedia texts to improve retrieval performance. Another aim, which this paper only treats superficially, is to compare book specific retrieval to generic retrieval for both (whole) book retrieval and focused retrieval. In the short time Wikipedia has existed, its use as a source of knowledge and reference has increased tremendously even if its credibility as a trustworthy resource is sometimes put to doubt [2]. This combination of features would make it interesting to use digitized books as a resource with which one can verify or support information found in the Wikipedia.

The most interesting part of this year's topics, which also constitutes the essence of this years task, is, no doubt, the Wikipedia text that is supplied with each aspect. The first thing coming to mind is, of course, using the Wikipedia texts for query expansion, which could intuitively provide a precision enhancing device. The Wikipedia texts are quite long, and the chances of zero hits using

the entire text as a query are quite significant (particularly if using logical AND to combine the terms). Query expansion needs thus be approached with caution.

Whereas a query text (even a test query) is said to originally be formulated by the user, a Wikipedia article does not origin with the user, so that there may be elements in the article that the user would not have endorsed, and thus are unintentional. Used uncritically in a query, those parts may reduce experienced retrieval performance.

A measure to counter this effect would be either using only parts of the Wikipedia text that (chances are that) the user would knowingly endorse, or to use the topic title to process the Wikipedia text, creating a version of the latter that is closer to the user's original intention, while still benefitting from the useful expansion potential the text entails.

In investigations involving book retrieval, [3] have experimented with different strategies of retrieval based on query length and document length. Their conclusion has been that basing one's book retrieval on collating results obtained from searching in *book pages* as basic retrieval units (shorter documents), performed better than using *the book as a whole* as a basic retrieval unit. Moreover, manually adding terms to the query improved page level (shorter document) retrieval, but did not seem to improve retrieval of longer documents (whole books).

At the OUC, we wished to pursue this observation, and pose the following questions:

- Can the Wikipedia page partly play the same role which manually added terms would play in a batch retrieval situation (laboratory setting)?
- In case it does, would it also benefit users in real life situations?
- What kind of usage of the Wikipedia text would, on average, provide better retrieval?

2 A brief Analysis of a Wikipedia topic text

A Wikipedia text may vary in length. Even if we assume the text is very central to – and a good representative of – the user's information need, we hypothesize that using the entire text uncritically as a query text or expansion device, would be hazardous.

Being a collaborative resource, the Wikipedia community has a number of rules that contributors are asked to adhere to "using common sense" [4]. This means that Wikipedia has a realtively free style. An examination of a number of articles treating various subjects indicates that they resemble each other in structure, starting off with a general introduction, followed by a table of contents into the details of the discussed subject. Given that it is the general topic of the article which is of interest to the user (or the writer of the article for that matter), This makes the beginning of an article quite important as a source for formulating the search.

However, a glance at the Wikipedia texts supplied with [5] (both on topic and aspect level) leaves the impression that using even the initial sentences or sections uncritically may result in poor retrieval, or no retrieval at all.

In the beginning of a Wikipedia article, there often occurs a "to be" (is, are, were a.s.o) or a "to have" ("has", "had" a.s.o.) inflection. The occurrence is not necessarily in the first sentence[1], but relatively early in the document. The idea is to use this occurrence as an entry point to the important part of the text[2].

Our hypothesis is that on both sides of such an occurrence, one generally finds words that could be used to compose meaningful queries. There are also grounds to assume that the user, who hunts for book references to this Wikipedia article the way he or she conceives its contents, has read this part of the text and approves of it as representative before proceeding to the rest of the article. Hence this part may arguably be a good representative of the text as seen by the user.

3 Extracting important terms from a Wikipedia article

A way of testing the idea mentioned above is, for each applicable Wikipedia text, to locate the first salient occurrence of a "to be" or "to have" inflection (see also footnotes 1 and 2, and define a window of both preceding and succeeding text (omitting stop-words). The length of the window may vary (See Figure 1). the content of this window is either used as the query or is added to an existing query as an expansion text.

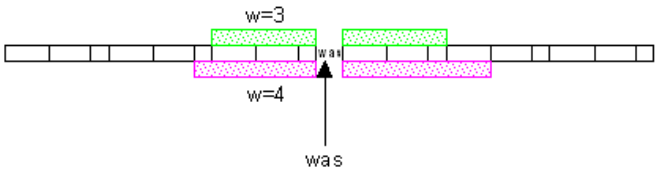


Fig. 1. Using windows of varying widths around an entry word in a Wikipedia text

[1] it is sometimes omitted in favor of phrases like "refers to", or another similar combination. This observation may be important to follow up in future research.

[2] On a too early occurrence, the second or third occurrence might be considered as an alternative entry.

3.1 Word occurrences in windows

Using this year's topics' Wikipedia texts as a sample, we have tried to analyze the occurrences of words in the samples, based on some parameters. The most important parameter was the length of the window. Based on the window length, it was interesting to categorize the distribution of expansion terms into nominals, verbs, adjectives etc.

Experiments performed by [6] indicate that nouns have an advantage when using only certain parts of speech in retrieval, in the sense that using only nouns in queries entails very little loss in precision recall. This would call for experimenting with Wikipedia text extracting only nouns. In this paper we are not pursuing this, merely observing the number of nouns in the queries we are composing. The purpose is to see how the presence of particular types of words in the query contribute to the retrieval quality, and what role Wikipedia can play here.

4 Book retrieval

4.1 Generic retrieval

Books seen as traditional documents are, on average, very long. This means that word may co-occur in line with their co-occurrence in a query, without necessarily being very relevant to the query. For this reason we have experimented with two types of queries. The one that looks at the book as a single large chunk of text, and the other that looks at each page as a chunk of text and combines the performances of the query against all pages.

We choose to regard this as a type of "generic retrieval" for the sake of the current discussion, although it does incorporate page division. In practice, page is not a part of the semantic structure, as page boundaries are mostly a function of physical attributes (font size, page size, illustration size etc.). Moreover we feel that the former type of retrieval may provide results that are unrealistically good.

4.2 Book-specific retrieval

One of the research objectives of this year's book track is to compare the performance of generic retrieval methods with more book-specific retrieval methods.

There are probably a large number of possibilities of utilizing book structure in the collection. We have chosen to identify chapter titles with the help of the TOC entries of the book[3]. In addition to the indication of being TOC sections (or lines or words), the marker elements also have references to the page where

[3] Results may suffer due to the fact that only some 36000 of the 50000 books indeed feature these markup attributes "refid" and "link" of the page and marker element respecively

the referred chapter begins. The chapter names are available and can be used to boost the chances of certain pages (title pages or inner pages of the chapter) to be retrieved as response to queries that include these words.

We create an index for which we identify words in chapter titles, section titles and the like, so we can enhance their significance at query time, and then try to run the expanded queries against this index as well. In practice we identify words constituting chapter or section titles in the TOC section of the book, find the physical location of their respective chapter and prepend them, specially tagged, to the chapter. This tagging facilitates different weighting of these words related to the rest of the text. Different weighting strategies can then be tested.

For book retrieval it is not important where in the book the title words increase in weight, as they will increase the chances of retrieving this book as a response to a query featuring these words. For focused retrieval we have the limitation that we do not have an explicit chapter partition of the book, only a page partition. One choice of handling this is to identify all pages belonging to a partition, and adding the title words (with enhanced weights) to the text of each page. Within the title page (first page of the chapter) the same title words can be given a different relative weight than for the pages inside the chapter.

5 Focused retrieval

In our experiments, focused retrieval follows along similar lines as book retrieval, just that here the purpose is to retrieve book pages rather than books. We have submitted a number of runs participating in the effort to improve the quality of the test collection. As page-relevance assessments for focused runs were not available by the submission deadline of this paper we choose to defer further analysis regarding this task to a later stage of the research.

6 Runs

We have been running comparable experiments for book and focused retrieval, using the same Indri (`http://www.lemurproject.org`) index. We were using Indri'support for retrieval by extents. We added chapter titles (enclosed in elements we named *titleinfront* and *titleinside* respectively) in all the pages that constituted a chapter title page or a chapter content page.

In both book retrieval and focused retrieval we have experimented with generic as well as book specific retrieval as described above, using the code pattern as described in table 1.

The main partition follows the line of book vs. focused retrieval, so that parallel generic and book specific runs can be compared. The first row features runs with queries involving topic titles only. The second row has runs where the topic is composed of terms from the topic title and the Wikipedia text, and the third row represents queries composed of the Wikipedia texts only. The hash

Table 1. Code pattern of run submissions to the book track. Rows represent the composition of the query. For columns, *book specific* refers to retrieval where chapter title and chapter front page are weighted higher than the rest of the text. No weighting for *generic* retrieval. Grey cells represent the runs analyzed in this paper.

	book retrieval		focused retrieval	
	generic	book specific	generic	book specific
title only	book_to_g	book_to_b	focused_to_g	focused_to_b
title and wiki	book_tw_g#	book_tw_b#	focused_tw_g#	focused_tw_b#
wiki only	book_wo_g#	book_wo_b#	focused_wo_g#	focused_wo_b#

character is a place holder for the size of the window as described in Section 3, where applicable. In this submission we have experimented with window sizes of 3 and 5 for each run type (as represented by a table cell). This gives a total of 20 runs. The choice of 3 and 5 is somewhat arbitrary as hypothetically salient choices. The choice of 10 was later added for the sake of control. More extensive experimentation with different combinations will be beneficial.

7 Results

7.1 Word distribution in queries

In table 2 we are listing the proportion of nouns[4] in the queries.

Table 2. Average percentage of nouns in the topic queries

	to_g	tw_g3	tw_g5	tw_g10
nouns	44	99	145	227
total	55	148	213	376
%nouns	80,00	66,89	68,08	60,37

7.2 Retrieval performance

Below we show retrieval performance results. The results are shown for some of the runs we submitted, and are based on partial assessments. Therefore conclusions are drawn with caution. The results are shown separately for generic

[4] Based on the definition in [6], "A noun is any person, place or thing", both person name, place name and general nouns are included when nouns are counted.

retrieval (Figure 2) and for book-specific - structure supported retrieval (Figure 3).

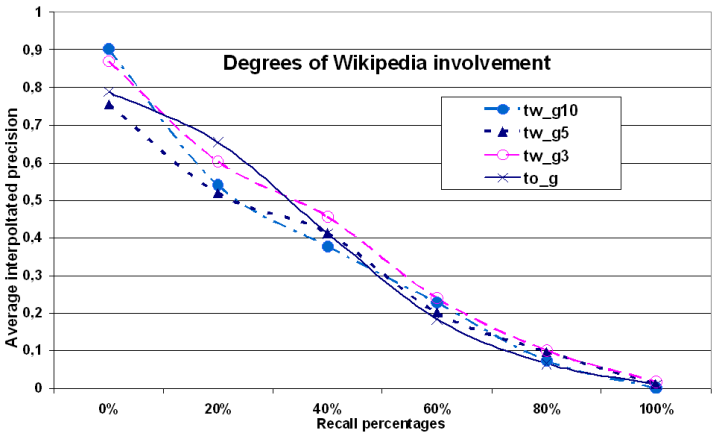
Generic retrieval In sub-figures 2(a) and 2(b) we are showing precision-recall curves of some of the runs, based on (a) each book as a long chunk of text - and (b) combination of single page performances, respectively. The most important finding is that the increased involvement of Wikipedia terms seems to dramatically improve retrieval performance at the low recall region in the runs labeled (a), while it seems to deteriorates the performance of the runs labeled (b). The increase in (a) is not linearly proportional to the number of terms added, and we see that 5 added terms draw the curve down at the low recall region. This may be due to an interaction between the basic query term (topic title) and the Wikipedia text draws down the result at this point. The low number of relevant judgments may also be a factor here. The tendency as a whole, and the difference between the two modes may, again, be explained by the fact that a group of many terms will tend to co-occur in a book as a long chunk of text, but will have a lower probability of co-occurrence in a page. The correlation with the percentage of nouns in the queries (table 2) is not clear, and is difficult to judge on the basis of the current results.

Book-specific (structure-supported) retrieval In sub-figures 3(a) and 3(b) we are showing precision-recall curves of some of the runs, based on each book as a long chunk of text - and combination of single page performances, respectively. Here the increased involvement of Wikipedia terms does not seem to improve retrieval performance, and the higher the involvement, the worse the performace gets.

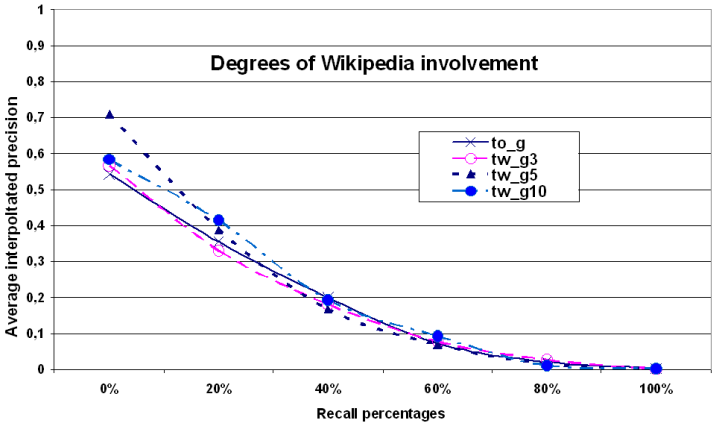
Also here, looking at a book as a long chunk of text is more beneficial for the results than looking at each page. Even if the potential is there, we are doubtful whether the results in the current setting actually indicates better experienced retrieval on the side of the user. More research and experimentation will be needed.

8 Conclusion

The results we obtain indicate that Wikipedia article texts have potential as retrieval aid for digitized books that are topically relevant for the subject of the articles. There is, however, little doubt that more experiments in laboratory conditions along this line of research, as well as conditions resembling real life usage of Wikipedia in combination with digitized books, will be necessary in order to approach combinations and query expansion settings that will be beneficial in real life situation. It is doubtful whether the best results we get in the current experiments also predict the user-experienced retrieval performance.

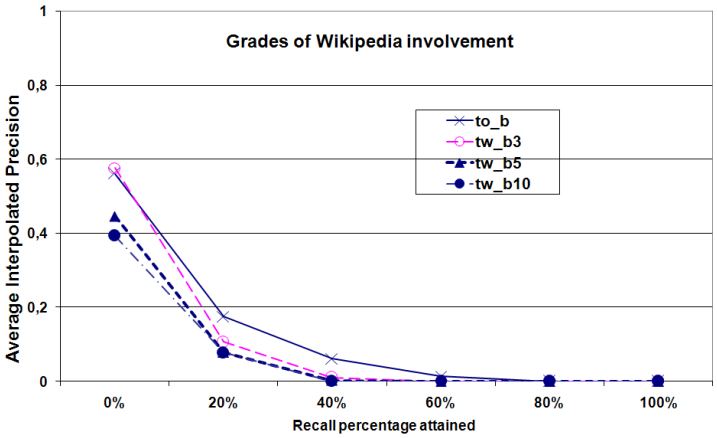


(a) Each book regarded as a large chunk of text

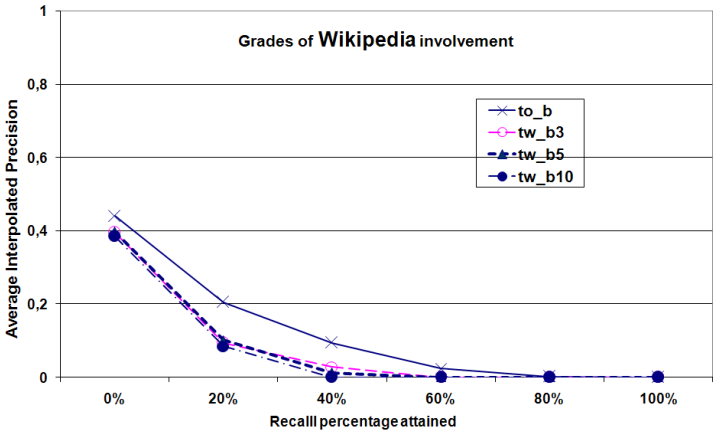


(b) Each book regarded as a combination of pages

Fig. 2. Precision-recall curves of generic book retrieval



(a) Each book regarded as a large chunk of text



(b) Each book regarded as a combination of pages

Fig. 3. Precision-recall curves of book-specific (structure supported) book retrieval

Trying to find book-specific (structure supported) methods for book retrieval, the page as a unit seems to have some disadvantages. The page is quite a little partition, and, in addition page breaks are often a consequence of physical attributes of the book rather than a conscious or structural choice taken by the author. It will be interesting to repeat our book-specific experiments with larger, structurally more coherent partitions of books.

References

1. Kazai, G., Koolen, M., Landoni, M.: Summary of the book track. In: INEX 2009. (in press)
2. Luyt, B., Tan, D.: Improving wikipedia's credibility: References and citations in a sample of history articles. Journal of the American Society for Information Science and Technology **43**(3) (2010)
3. Wu, M., Scholer, F., Thom, J.A.: The impact of query length and document length on book search effectiveness. In: INEX 2008. (2009)
4. unknown. Retrieved 5 Marchy, 20010 from the World Wide Web: `http://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines` (2010)
5. Kazai, G., Koolen, M.: Inex book search topics for 2009 (2009)
6. Chowdhury, A., McCabe, M.C.: Improving information retrieval systems using part of speech tagging. Technical report, ISR, Institute for Systems Research (1998)