Focused Access to Sparsely and Densely Relevant Documents

Paavo Arvola
Dept. of Information Studies and Interactive Media
University of Tampere, Finland
paavo.arvola@uta.fi

Jaana Kekäläinen
Dept. of Information Studies and Interactive Media
University of Tampere, Finland
jaana.kekalainen@uta.fi

Marko Junkkari
Dept. Of Computer Science
University of Tampere, Finland
marko.junkkari@cs.uta.fi

ABSTRACT

XML retrieval provides a focused access to the relevant content of documents. However, in evaluation, full document retrieval has appeared competitive to focused XML retrieval. We analyze the density of relevance in documents, and show that in sparsely relevant documents focused retrieval performs better, whereas in densely relevant documents the performance of focused and document retrieval is equal.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Measurement, Performance, Experimentation.

Keywords

XML retrieval, tolerance to irrelevance, focused retrieval.

1. FOCUSED ACCESS TO A DOCUMENT

Ideal information retrieval (IR) systems would return only relevant information to the user. In traditional document retrieval, returned documents typically include both relevant and non-relevant content. Approaches like passage and XML retrieval aim at returning the relevant content more accurately: the user should be guided directly to the relevant content inside the document instead of having to browse through the whole document. Surprisingly, in recent studies document retrieval has been found a competitive approach to focused XML retrieval according to retrieval effectiveness [e.g. 7]. However, some essential features in focused access to a document have been overlooked: the order of browsing, the user's reluctance to browse non-relevant information and the proportion of relevant text in documents. In the present study this proportion is referred to as the *density* of relevance of the document.

An XML retrieval system provides retrieved and assumedly relevant passages first to the user. If a returned passage turns out to be non-relevant, the user will not necessarily browse it through but rather continues with the next result. This user behavior is combined to effectiveness evaluation in the tolerance to irrelevance (T2I) metric [4], which models the user interrupting to browse after a given amount of non-relevant information is encountered. The sooner T2I is reached, the less the document benefits the effectiveness in evaluation.

In this study, we follow a browsing model with the given assumptions: A focused retrieval system guides a user to the relevant content, and the user starts browsing the document from the passages indicated by the system [1,2]. Returned passages are browsed first and the browsing continues until T2I is reached. With this model, effectiveness measures like precision and recall can be calculated for the document. These measures are calculated based on the proportion of relevant text browsed at the point where T2I is reached.

We compare focused XML retrieval with document retrieval by taking into account the access point to the document, browsing order and T2I. We analyze the effectiveness of retrieval at *the level of a retrieved relevant document*. More specifically, we examine the effectiveness by the density of relevance in documents. Our hypothesis is that focused retrieval provides a more effective access to the relevant content of a relevant document than full document retrieval, especially when it comes to sparsely relevant documents.

2. DENSITY AND DISTRIBUTION OF RELEVANCE

As the test collection we use the frozen Wikipedia collection [3] of more than 650,000 documents covering various subjects. The collection is used with topics and relevance assessments of the INEX 2008 initiative [6], where relevant passages (in 4,887 relevant documents) for each topic (totally 70) are assessed. All the relevant documents are sorted according to their ratio of relevant text to all text, i.e. how many percent of the document's text is relevant. Then the sorted list of documents is split into deciles, each covering 10% of the documents. The (rounded) lower boundaries of the density (relevance ratio) for the deciles are 0.005%, 2.4%, 6.6%, 12.1%, 24.4%, 58.4%, 94.9%, 99.3%, 99.7% and 99.9% (dec 1, dec 2,…, dec 10 respectively). That is, the last 4 deciles i.e. 40% of the relevant documents have a very high relevance density. Obviously, focused access to those documents does not bring any improvements.

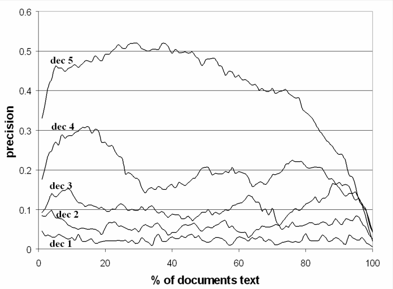

Figure 1: Average distribution of document's relevant text on five smallest deciles

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07

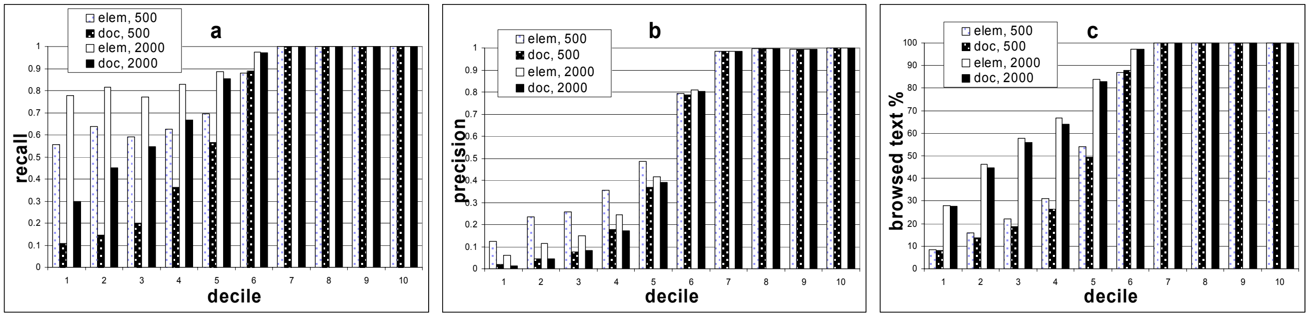

Figure 2: Average Recall (a), Precision (b), % browsed content (c) of relevant documents on each decile

Figure 1 shows the average precision of the five smallest deciles at percentages of the relevant documents' running text. The relevant content in the lowest deciles is somewhat steady across the average document, which means the relevant content may be at any location, whereas the fourth and fifth decile shows a slight bias towards the beginning of a document. The remaining deciles especially from 7 upwards draw a high, relatively straight line and are left out for the readability of the lowest curves.

3. PRECISION AND RECALL WITHIN A DOCUMENT

To study the benefit of focused retrieval strategy for the retrieval within a document on each decile, we selected the retrieved passages of each relevant document. These passages were provided by the best performing run at INEX 2008 (RiCBest, University of Waterloo [5]). Then we compared these focused results with a document retrieval baseline, where each relevant document is browsed sequentially. Figure 2 shows the average recall, precision and the percentage of browsed content in the relevant documents for each decile. The *elem* column refers to the focused retrieval strategy (i.e. RiCBest) while the *doc* column refers to the document retrieval baseline. We report figures on two T2I points: 500 and 2000 characters. In other words the browsing is expected to end when the user has bypassed 500 or 2000 non-relevant characters. The amount of 500 characters corresponds approximately to the next paragraph.

Figure 2c shows that the amount of browsed content is about the same for both of the strategies when assuming the same T2I. That is, the focused retrieval strategy does not reduce the amount of browsed content. However, with that amount the precision (Figure 2b) and especially the recall (Figure 2a) of the browsed content are notably higher with the focused strategy for half of the relevant deciles (dec1-5). The documents after sixth decile are uninteresting since they are densely relevant and neither browsing order matters nor T2I is reached.

4. DISCUSSION AND CONCLUSIONS

Focused retrieval is beneficial in locating the relevant content in between non-relevant material. Therefore documents with high relevance density are not interesting in the scope of focused retrieval. The results show that the less relevant content, the better the focused retrieval performs. In plain document retrieval, the user is responsible for finding the relevant content within a sparsely relevant document. This leads into poor performance with documents having only some relevant content, when T2I is assumed. Namely, in many cases the browsing ends before the relevant content is met. This leads to zero recall. On the other hand, if the relevant content is pointed out accurately, the recall for the document is typically 1 (100%). However, due to the nature of T2I, where the browsing goes on with the non-relevant material until the T2I is met the precision is always less than 1.

While we assume the T2I and browsing order in focused retrieval, our findings differ from the previous studies, where the full document retrieval has been a competitive approach [7]. This is due to the densely relevant documents, where the focused retrieval systems tend to retrieve only parts for why the recall per document remains low. This has led into overemphasizing precision, which is taken into account four times more than recall in the official metrics (i.e. the F-Measure).

5. ACKNOWLEDGEMENTS

The study was supported by the Academy of Finland under grants #115480 and #130482.

6. REFERENCES

[1] Arvola, P. 2008. Passage Retrieval Evaluation Based on Intended Reading Order. LWA 2008, 91-94.

[2] Arvola, P., Kekäläinen, J., Junkkari, M. 2010. Expected Reading Effort in Focused Retrieval Evaluation. To appear in Information Retrieval.

[3] Denoyer, L., and Gallinari, P. 2006. The Wikipedia XML Corpus. Sigir Forum, 40:64-69.

[4] de Vries, A.P., Kazai, G., and Lalmas, M. 2004. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In Proceedings of RIAO 2004, 463-473.

[5] Itakura, K.Y., and Clarke C.L.A. 2009. *University of* Waterloo at INEX 2008: Adhoc, Book, and Link-the-Wiki Tracks, In Advances in Focused Retrieval, 132-139.

[6] Kamps, J., Geva, S., Trotman, A., Woodley, A., and Koolen, M. 2009. Overview of the INEX 2008 ad hoc track. In Advances in Focused Retrieval, 1-28.

[7] Kamps, J., Koolen, M., and Lalmas, M. 2008. Locating relevant text within XML documents. In Proceedings SIGIR '08. ACM, New York, NY, 847-848.