



The effect of granularity and order in XML element retrieval

Nils Pharo *

Oslo University College, Pb 4 St. Olavs plass, N-0130 Oslo, Norway

ARTICLE INFO

Article history:

Received 30 November 2007

Received in revised form 25 March 2008

Accepted 15 May 2008

Available online 26 June 2008

Keywords:

Interactive information retrieval

XML

User studies

ABSTRACT

The article presents an analysis of the effect of granularity and order in an XML encoded collection of full text journal articles. Two-hundred and eighteen sessions of searchers performing simulated work tasks in the collection have been analysed. The results show that searchers prefer to use smaller sections of the article as their source of information. In interaction sessions during which articles are assessed, however, they are to a large degree evaluated as more important than the articles' sections and subsections.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

XML-coding of semi-structured documents facilitates the retrieval of document elements as an alternative to retrieving full documents (Luk et al., 2002). There are many arguments for partial document retrieval, e.g., it makes it possible to filter out only those parts of the documents that specifically treat the issues related to searchers' queries. In addition to explicating the logical structure of documents XML also makes it possible to add extra semantics to elements thus combining the mark-up and content for more precise retrieval. Little, however, is known about how searchers react to such features from the XML retrieval systems. For example, we do not know much about the effect on searchers of presenting them with document parts in addition to full documents.

This article presents an analysis of the granularity of documents and the order in which searchers prefer to examine parts of documents and whole document. The study is based on data collected in the INEX 2005 interactive track, of which there were three different sub tracks. The current study is based on Track A data, which was compulsory for all track participants. It was performed in a data set consisting of articles in computer science journals from the IEEE computer society using the Daffodil¹ retrieval system.

We have investigated two research questions in this article:

1. What level of document granularity do searchers judge to be the most relevant?
2. In what order do searchers interact with and judge the relevance of elements of different granularity?

The article is built up in the following way: in the next section we address previous research in XML retrieval, in particular the work being done as part of the INEX conferences, and relate it to our approach in the next section. Section 3 contains a description of our method. In Section 4 we present our findings and Section 5 contains the discussion and our conclusions.

* Tel.: +47 22452684; fax: +47 22452605.

E-mail address: nils.pharo@jbi.hio.no

¹ <http://www.is.inf.uni-due.de/projects/daffodil/>.

2. Background

The initiative for the evaluation of XML retrieval (INEX) started in 2002 in order to provide “an infrastructure to evaluate the effectiveness of content-oriented XML retrieval systems” (Kazai, Lalmas, Fuhr, & Gövert, 2004). INEX builds its experimental design on the TREC model, with a test collection which consists of topics/tasks (submitted by the participating groups), documents (approximately 12,000 articles from a selection of IEEE Computer society’s journals²) and relevance assessments provided by the participants, thus making it possible to compute the retrieval effectiveness of different matching algorithms. Since its beginning several tracks have been introduced to the initiative in order to explore topics such as relevance-feedback, heterogeneous collections, natural queries, document mining, multimedia and interactive information retrieval (INEX, 2005).

The interactive track was first introduced in 2004 (Tombros, Larsen, & Malik, 2005) and aims at studying how end-users interact with XML-encoded data. The IEEE collection was used for the 2004 and 2005 (Larsen, Malik, & Tombros, 2006) main tracks and has resulted in some new knowledge. In 2005 an additional data set was added to the track from the Lonely Planet travelling guides.

The data from the INEX 2004 and 2005 interactive track have been used to evaluate and redesign the user interface (Malik, Klas, Fuhr, Larsen, & Tombros, 2006). The questions related to system design have been analysed showing that searchers in general were positive to both the 2004 and 2005 systems, but that the 2005 system had better “learnability”.

Pharo and Nordlie used the INEX 2004 (Pharo & Nordlie, 2005) IEEE data to investigate whether users, when presented with the opportunity to interact with relevant parts of documents instead of or in addition to the full document, prefer to work with the elements in or out of context with the full text. Their findings are not conclusive, but they indicate that giving users access to the most relevant text elements on lower levels of granularity is valuable, but only if the full article is present to provide context.

Hammer-Aebi, Christensen, Lund, and Larsen (2006) used the INEX Lonely Planet dataset in a similar study. They found that users prefer elements of depth 2–4 rather than full documents. Their study also involved comparing systems with and without context, which, surprisingly, showed not much preference of searchers for the context providing system.

Kim and Son (2006) performed a small scale study to compare the presentation format of the search engines used at the 2005 and 2004 INEX interactive tracks. Their findings indicate that the searchers has good use of the embedded table of content in the 2005 interface.

The interactive track organisers (Larsen, Tombros, & Malik, 2006; Malik, Larsen, & Tombros, 2007) have performed initial studies of the 2005 interactive search logs which suggest that users seem to prefer document elements rather than whole documents and that elements of short length are often assessed as not relevant. They also suggest that searchers prefer to use document metadata as their initial entry point.

Ramirez and de Vries (2006) have analysed the effect of three different contextual factors on users interacting with the INEX 05 IEEE data. They found that users with higher factual knowledge about the topic were better able to perform compound tasks, and also that such users made more use of larger elements, such as whole articles, than users performing simpler tasks.

Our study is related to the study reported by Larsen, Tombros, et al. (2006), which it to a certain extent overlaps and also to the study by Hammer-Aebi et al. (2006) who use a different data set, the Lonely Planet collection. Our study is, however, more specific in identifying the context of searching, i.e. how the order of interaction with elements of different granularity affects the relevance judgements of said elements.

3. Method

The data used in our analysis stems from the logs collected as part of the INEX 2005 interactive track, in this section we will present the data collection procedures used in the experiment. There were 11 institutions from around the world participating in the interactive track, each group was instructed to collect data from minimum six end-users (hereafter called searchers). In our data we have analysed 218 sessions performed by 73 different searchers.

The experiments were conducted using the Daffodil IR system with an interface developed specifically for the interactive track.

The IR system is designed with a quite simple search interface where searchers can enter their query terms. In the result list relevant documents and elements are presented hierarchically with the relevant elements presented within the documents they belong to (see Fig. 1 captured from Larsen et al., 2006). The documents are presented at four levels; sections, subsections and sub-subsections which can be accessed directly and a link within the document’s title which appears to be to the whole document, but which actually leads to the document’s metadata.

Having selected an element from the result list searchers are presented with an interface containing the requested element and the ability to browse the full document upwards and downwards (see Fig. 2, also captured from Larsen et al., 2006). Also the interface contained the relevance judgement tool. Relevance are judged on a three-point scale:

² From 2006 the collection consists of approximately 600,000 Wikipedia articles.

Download English Version:

<https://daneshyari.com/en/article/515724>

Download Persian Version:

<https://daneshyari.com/article/515724>

[Daneshyari.com](https://daneshyari.com)