DOCTORAL ABSTRACT

Content and structure summarisation for accessing XML documents

Zoltán Szlávik
Department of Computer Science
Queen Mary University of London
UK, E1 4NS
zolley@dcs.qmul.ac.uk
http://www.dcs.qmul.ac.uk/~zolley/

Abstract

As the availability of structured documents is constantly increasing, retrieval systems able to return document portions are being developed. Structured documents, usually formatted in XML, may consist of large numbers of document portions, often organised into a hierarchical logical structure. With the high number of document portions, it is necessary to direct the attention of users of retrieval systems towards the most important document portions, and also, to give overviews of the structure of documents, in other words, to show document portions in context. This thesis investigates summarisation as a means to help searchers of XML retrieval systems in the process of accessing the contents of document portions. Two types of summarisation are investigated.

First, summaries of the textual contents of document portions, called XML elements, are studied in a user-based environment. Traditionally, summarisation is associated with whole documents or document sets, but rarely with document portions. As summaries of documents have been proved to be useful in whole document retrieval, it is considered worthwhile to investigate summaries of document portions in XML element retrieval. Summaries of elements are presented to searchers in the context of other elements from the document. The textual summaries of elements also reflect the searchers' information needs: they are query based.

The second type of summarisation investigated in this thesis is called structure summarisation. The automatic generation of tables of contents, as structure summaries, is described and examined. ToC generation is studied either when searchers' queries are available (query based structure summarisation) or otherwise (query independent structure summarisation).

The work presented in this thesis has made several contributions to the fields of summarisation and interactive XML retrieval.

The thesis is available online at http://www.dcs.qmul.ac.uk/~zolley/thesis.html