Structured Document Retrieval

Mounia Lalmas, Queen Mary, University of London, UK, mounia@acm.org

Ricardo Baeza-Yates, Yahoo! Research Barcelona, Spain, rbaeza@acm.org

SYNONYMS

structured text retrieval, querying semi-structured data, passage retrieval, XML retrieval, focused retrieval

DEFINITION

Structured document retrieval is concerned with the retrieval of document fragments. The structure of the document, whether explicitly provided by a mark-up language or derived, is exploited to determine the most relevant document fragments to return as answers to a given query. The identified most relevant document fragments can themselves be used to determine the most relevant documents to return as answers to the given query.

MAIN TEXT

The aim of this entry is to clarify different terminologies that have been used to refer to or are strongly related to structured retrieval and semi-structured data.

The term "structured document retrieval", which was introduced in the early to mid 90s in the information retrieval community, refers to "passage retrieval" and "structured text retrieval". In passage retrieval, documents are first decomposed into passages (e.g. fixed-size text-windows of words, fixed discourses such as paragraphs, or topic segments through the application of a topic segmentation algorithm). Passages could themselves be retrieved as answers to a query, or be used to rank documents as answers to the query.

Structured text retrieval is concerned with the developments of models for querying and retrieving from structured text, where the structure is usually encoded with the use of mark-up languages, such as SGML, and now predominantly XML. Indeed, text documents often display structural information. For example, a scientific article will have a so-called logical structure, such as an abstract, several sections and subsections, each of which composed of paragraphs. A book will have a so-called layout structure, such as pages and columns.

Structured text retrieval is to be contrasted to traditional text retrieval, where the latter is concerned with the retrieval of unstructured text – so-called "raw text" or "flat text". The use of the term "structured" in "structured text retrieval" is there to emphasize the interest in the structure. Furthermore, structured text retrieval aims to exploit the available structural information to return text fragments (e.g. XML elements) as opposed to entire text documents.

The term "semi-structured" comes mainly from the database community. Traditional database technologies, such as relational databases, have been concerned with the querying and retrieval of highly structured data (e.g. from a student table, find the names and addresses of those with a grade over 80 in a particular subject). Text documents marked-up, for instance, in XML are made of a mixture of highly

structured components (e.g. year, author name) typical of database records, and loosely structured components (e.g. abstract, section). Database technologies are being extended to query and retrieve such loosely structured components, called semi-structured data. Databases that support this kind of data, mainly in the form of text with mark-up, are referred to as semi-structured databases, to emphasize the loose structure of the data and use "querying data" instead of "data retrieval".

From a terminology point of view, structured text retrieval and querying semi-structured data, in terms of end goals, are the same. The difference comes from the fact that in information retrieval, the structure is added, and in database, the structure is loosened. It should however be pointed out that research in information retrieval and databases with respect to accessing structured text (or semi-structured data in the form of text) have been concerned, because of historical reasons, with different aspects of the access process, e.g. ranking in information retrieval versus efficiency in databases. Nowadays there is a convergence trend between the two areas (e.g. [1]).

In the late 1990s, the interest in structured document retrieval grew significantly due to the introduction of XML in 1998, which has now became the de-facto format standard for structured documents (or structured text, semi-structured data). Research on XML retrieval was further boosted with the set-up of INEX in 2002, the Initiative for the Evaluation of XML Retrieval, which allowed researchers to compare and discuss the effectiveness of models specifically developed for XML retrieval [2]. Nowadays, XML retrieval is almost a synonym for structured document retrieval, structured text retrieval and querying semi-structured data.

Structured document retrieval, passage retrieval, structured text retrieval, querying semi-structured data, XML retrieval, all belong to what has recently been called "focused retrieval" [3]. Focused retrieval is concerned with returning the most focused results to a given query. Such focused results include passages, XML elements, and factoid answers (e.g. London being the capital of the UK).

CROSS REFERENCES

Document databases
INitiative for the Evaluation of XML retrieval (INEX)
Integrated DB&IR
Semi-structured data
Structured text retrieval models
XML retrieval

RECOMMENDED READING

[1] S. Amer-Yahia, P. Case, T. Rölleke, J. Shanmugasundaram and G. Weikum. In Report on the DB/IR panel at SIGMOD 2005, *SIGMOD Record*, 34(4):71-74, 2005.

[2] G. Kazai, N. Gövert, M. Lalmas and N. Fuhr. The INEX Evaluation Initiative. In *Intelligent Search on XML Data, Applications, Languages, Models, Implementations, and Benchmarks*, Springer, pp 279-293, 2003.

[3] A. Trotman, S. Geva and J. Kamps. Report on the SIGIR 2007 workshop on focused retrieval. In *SIGIR Forum*, 41(2):97-103, 2007