IRM PRESS
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.irm-press.com
ITB9194

Chapter I

System of Information Retrieval in XML Documents

Saliha Smadhi
Université de Pau, France

ABSTRACT

This chapter introduces the process to retrieve units (or subdocuments) of relevant information from XML documents. For this, we use the Extensible Markup Language (XML) which is considered as a new standard for data representation and exchange on the Web. XML opens opportunities to develop a new generation of Information Retrieval System (IRS) to improve the interrogation process of document bases on the Web.

Our work focuses instead on end-users who do not have expertise in the domain (like a majority of the end-users). This approach supports keyword-based searching like classical IRS and integrates structured searching with the search attributes notion. It is based on an indexing method of document tree leafs which authorize a content-oriented retrieval. The retrieval subdocuments are ranked according to their similarity with the user's query. We use a similarity measure which is a compromise between two measures: exhaustiveness and specificity.

INTRODUCTION

The World Wide Web (WWW) contains large amounts of information available at websites, but it is difficult and complex to retrieve pertinent information. Indeed, a large part of this information is often stored as HyperText Markup Language (HTML) pages that are only viewed through a Web browser.

This research is developed in the context of the MEDX project (Lo, 2001) of our team. We use XML as a common structure for storing, indexing and querying a collection of XML documents.

Our aim is to propose the suited solutions which allow the end-users not specialized in the domain to search and extract portions of XML documents (called units or subdocuments) which satisfy their queries. The extraction of documents portion can be realized by using XML query languages (XQL, XML-QL) (Robie, 1999; Deutsch, 1999).

An important aspect of our approach concerns the indexation which is realized on leaf elements of the document tree and not on the whole document.

Keywords are extracted from a domain thesaurus. A thesaurus is a set of descriptors (or concepts) connected by hierarchical relations, equivalence relations or association relations. Indexing process results are stored in a resources global catalog that is exploited by the search processor.

This chapter is organized as follows. The next section discusses the problem of relevant information retrieval in the context of XML documents. We then present the model of XML documents indexing, followed by the similarity measure adopted and the retrieval strategy of relevant parts of documents. The chapter goes on to discuss related work, before its conclusion. An implementation of SIRX prototype is currently underway in Python language on Linux Server.

INFORMATION RETRIEVAL AND XML DOCUMENTS

The classical retrieval information involves two principal issues, the representation of documents and queries and the construction of a ranking function of documents.

Among Information Retrieval (IR) models, the most-used models are the Boolean Model, Vector Space Model and Probabilist Model. In the Vector Space Model, documents and queries are represented as vectors in the space of index terms. During the retrieval process, the query is also represented as a list of terms or a term vector. This query vector is matched against all document vectors, and a similarity measure between a document and a query is calculated. Documents are ranked according to their values of similarity measure with a query.

XML is a subset of the standard SGML. It has a richer structure that is composed mainly of an elements tree that forms the content. XML can represent more useful information on data than HTML. An XML document contains only data as opposed to an HTML file, which tries to mix data and presentation and usually ignores structure. It preserves the structure of the data that it represents, whereas HTML flattens it out. This meta markup language defines its own system of tags representing the structure of a document explicitly. HTML presents information and XML describes information.