Searching XML Element Using Terms Propagation Method

Samia Berchiche-Fellag[1] and Mohamed Mezghiche[2]

[1] Université Mouloud Mammeri de Tizi-Ouzou, 15000 Tizi-Ouzou, Algérie
samfellag@yahoo.fr
[2] Université M'Hamed Bougara Boumerdes, Algérie
mohamed.mezghiche@yahoo.fr

Abstract. In this paper, we describe terms propagation method dealing with focussed XML component retrieval. Focussed XML component retrieval is one of the most important challenge in the XML IR field. The aim of the focussed retrieval approach is to find the most exhaustive and specific element that focus on the user need. These needs can be expressed through content queries composed of simple keyword. Our method provides a natural representation of document, its elements and its content, and allows an automatic selection of a combination of elements that better answers the user's query. In this paper we show the efficiency of the terms propagation method using a terms weighting formula that takes into account the size of the nodes and the size of the document. Our method has been evaluated on the «Focused» task of INEX 2006 and compared to XFIRM model which is based on relevance propagation method. Evaluations have shown a significant improvement in the retrieval process efficiency.

Keywords: Structured Information Retrieval (SIR), XML, terms propagation, CO query, terms weighting, element, INEX.

1 Introduction

XML documents are semi-structured documents which organize text through semantically meaningful elements labelled with tags. Hierarchical document structure can be used to return specific document components instead of whole documents to users.

Structural information of XML documents is exploited by Information Retrieval Systems (IRS) to return to users the most exhaustive[1] and specific[2] [1] documents parts(i.e. XML elements, also called nodes) answering to their needs. These needs can be expressed through Content queries (CO: Content Only) which contain simples keywords or through Content And Structure queries (CAS) which contain both keywords and structural information on the location of the needed text content. Most

[1] An element is exhaustive to a query if it contains all the required information.

[2] An element is specific to a query if all its content concerns the query.

T. Andreasen et al. (Eds.): ISMIS 2014, LNAI 8502, pp. 395–404, 2014.

of the retrieval models used for structured retrieval are adaptation of traditional retrieval models. The main problem is that the classical IR methods work at the document level. This does not perform well at the node level due to node nesting in XML as explained in [2] [3] [4].

The challenge in XML retrieval is to return the most relevant nodes that satisfy the user needs. Of most interest is the class of CO queries where the user doesn't know anything about the collection structure and issue her query in free text. The IRS exploits the XML structure to return the most relevant XML nodes that satisfy the user needs. Besides being relevant, retrieved nodes should be neither too large nor too small. In this aim we present our method which consists of searching the relevant nodes to a CO query composed of simple keywords in a large set of XML documents and taking into account the contextual relevance. The search process that we propose is based on a method of terms propagation.

Our method has already proved its effectiveness in [5] using the weighting formula usually used in IRS; In this paper we show that our method remains efficient using a weighting formula that takes into account the number of terms and the average number of terms in both the nodes and the document.

The rest of this paper is organized as follows: We present the state of the art in section 2. In section 3 we describe our baseline model, which uses a terms propagation method; we also present our weighting formula which uses node size and document size. Finally we present in section 4 results of our experimentations.

2 Related Work

The IR community has adapted traditional IR approaches to address the user information needs in XML collection. Some of these methods are based on the vector space model [6], [3] , [7] , or on the probabilistic model [8]. Language models are also adapted for XML retrieval [9], [10], as well as Bayesian networks in [11].

The aim of IRS dealing with XML documents is to retrieve the most relevant nodes the user need. For this purpose several approaches based on propagation methods were proposed by authors. Relevance propagation, terms propagation and weights propagation. In the relevance propagation approach, relevance score of leaf nodes in xml document tree is calculated and propagated to ancestors. Authors in [12] used linear combination of children's scores called "maximum-by-category » and « summation ».While the relevance propagation in [13] using XFIRM system is function of the distance that separates nodes in the tree. In [14], [15] authors used a method of weights propagation. For computing the weights of inner nodes, the weights from the most specific nodes in the document multiplied with an augmentation factor are propagated towards the inner nodes. Cui et al.[16], Benaouicha [17] , and Fellag[5] [18] used terms propagation method. In this case, textual content of leaf nodes in Xml document is propagated to their ancestor considering some conditions. In[16] and [5] [18] authors exploited both structural