Chapter 3

Automated Ranking of Relaxing Query Results Based on XML Structure and Content Preferences

Wei Yan
Northeastern University, China

Li Yan
Northeastern University, China

Z. M. Ma
Northeastern University, China

ABSTRACT

This paper proposes a contextual preference query method of XML structural relaxation and content scoring to resolve the problem of empty or too many answers returned by XML. This paper proposes a XML contextual preference (XCP) model, where all the possible relaxing queries are determined by the users’ preferences. The XCP model allows users to express their interests on XML tree nodes, and then users assign interest scores to their interesting nodes for providing the best answers. A preference query results ranking method is proposed based on the XCP model, which includes: a Clusters_Merging algorithm to merge clusters based on the similarity of the context states, a Finding_Orders algorithm to find representative orders of the clusters, and a Top-k ranking algorithm to deal with the many answers problem. Results of preliminary user studies demonstrate that the method can provide users with most relevant and ranked query results. The efficiency and effectiveness of the approach are also demonstrated by experimental results.

INTRODUCTION

Nowadays, XML is ubiquitous in retrieving and exchanging information over the Internet (Cho & Balke, 2008). XML as a data format differs from other document formats in that it has rich structure besides content. As a result, XML is often represented as tree model, and it is often queried on both structure and content (Liu, Wan, & Chen, 2009). In general, users search results from the XML document simply either matches or it does not. In such context, users may be confronted with the following two problems:

DOI: 10.4018/978-1-4666-2470-2.ch003
1. **Empty answers**: When the query is too selective, the answer may be empty or too little. In this case, it is desirable to have the option of relaxing the original query for presenting more relevant answers that can meet user’s needs and preferences closely.

2. **Many answers**: When the query is not too selective, too many results may be in the answers. In such a case, it will be desirable to have the option of order the matches automatically that ranks more globally important answers higher and returning only the best matches.

In the first case, several approaches have been proposed to deal with this issue (Amer-Yahia, Cho, & Srivastava, 2002; Cho & Balke, 2009). The basic idea of these approaches is based on considering XML relaxing queries to return closest or the most relevant results to the users, but most of them does not consider the user’s preferences when relaxing the original query. However, in the real application the efficiency of the query relaxation is affected greatly by the user’s preferences. In this paper, we automatically generate preferences by using association rules mining in the profile tree (Paik et al., 2009).

In order to avoid empty results and to further personalize users’ queries, a preference query considers nodes relaxation to the preferred query structure. Moreover, preferred structure in the query can be relaxed to all still relevant query structure. To enhance the expressiveness of the preference model, preference may depend on context (Agrawal, Rantzau, & Terzi, 2006). Context is a general term used to express the situation of the time of the submission of a query, including the surrounding environment, time or location (Stefanidis, Pitoura, & Vassiliadis, 2007).

In this paper, we focus on both the relaxation of structural preferences and content scoring for XML, inspired by structural relaxation techniques to capture scoring and ranking queries (Cho & Balke, 2009). Due to the structural heterogeneity of XML data, queries are usually interpreted approximately and Top-k answers are returned ranked by their relevance to the query (Amer-Yahia, Lakshmanan, & Pandit, 2004; Polyzotis, Garofalakis, & Ioannidis, 2004).

However, after relaxing original queries, another problem faced by the users will be that there are usually many answers returned to the users’ preference queries. To resolve the many answers problem, we use efficient Top-k ranking algorithm to rank the results. We propose the method of content scoring for XML tree nodes based on contextual preferences. Users assign an interest score to each preferred node, preference expressed by the users to be an indicator of interest degree. Assuming that the XML document is large and only a few nodes are interested by the users, sorting the whole XML document for each user’s preference query will result in both wasting resources and slow query responses. Thus, we propose a preference queries results ranking approach to improve efficiency of the user’s processing query.

Specifically, we compute the scores of representative orders of the clusters and then take into account to provide ranked Top-k results. We also consider the contextual preferences based on the similar context states for all XML tree nodes. For an incoming preference query, we first evaluate the user’s interest degree, and then quickly provide the ranked results that agree with the scores of representative orders of the clusters.

Our contributions are summarized as follows:

- We propose a XML contextual preference model, and then assign interest scores to the XML tree nodes and quickly provide best results for the users.
- In order to resolve problem of the empty results, we propose a method of the XML structural relaxing queries. We consider nodes relaxation to the user’s preferred query structure.