A Genetic Algorithm-Based XML Information Retrieval Model

Fatma Zohra Bessai-Mechmache, Karima Hammouche
Research Center on Scientific and Technical Information
Ministry of Higher Education and Scientific Research
Algiers, Algeria
fatmazohrabm@gmail.com

Zaia Alimazighi
Computer Science Department
University of Science and Technology Houari Boumediene
Algiers, Algeria
zalimazighi16@gmail.com

***Abstract*— Finding the valuable relevant information continues to be the major challenges of Information Retrieval Systems owing to the explosive growth of online web information. Among these challenges, we consider the XML Information Retrieval challenges as XML has become a de facto standard over the Web. In this paper, we tackle the issue of content-based XML information retrieval. We formulate the retrieval issue as a combinatorial optimization problem in order to generate the best set of relevant XML elements for a given keywords query. In our proposal, we define a genetic algorithm which maximizes similarity between a set of XML elements and the user query. The results based on the precision measure are very promising.**

Keywords*— *Genetic algorithm; result merging; XML information retrieval.

I. INTRODUCTION

The development of electronic document and the Web have emerged and imposed structured data formats such as XML (eXtensible Markup Language) to represent information in a richer form adapted to specific needs. These formats allow representing jointly the textual information and the information of structure of a document. The logical structure or hierarchy of XML documents contains content and structural elements. The purpose of an XML information retrieval system is refined to retrieval strategies, which aim at returning document components, i.e. XML elements, instead of whole documents in response to a user query [1-6].

Content-based XML information retrieval considers just content-only queries. Content-only queries make use of content constraints only, i.e. they are made of keywords and are suitable for XML retrieval scenarios where users do not know, or are not concerned, with the document's logical structure when expressing their information needs.

Content-based XML information retrieval is a challenging task. In this researcher work, XML information retrieval is concerned with a refined strategy which involves searching a huge search space for the better selection and combination of relevant XML elements. As Genetic Algorithm (GA) is well suited for searching huge search spaces, in this paper, a content-based XML information retrieval model is proposed for efficient information retrieval systems using genetic algorithm [7-11]. In the proposed model, Genetic Algorithm generates the best combination of XML elements from a set of selected XML elements.

The roadmap to the remaining part of the paper is as follows: Section 2 reviews some related work. Section 3 describes the proposed model with an approach for finding the optimal solution among all possible solutions. The results of the experiments are presented in Section 4. Finally, section 5 concludes this work and lists some perspectives.

II. RELATED WORK

Content-based XML information retrieval addresses issues related to granularity, diversity and relevance of search result. So, authors in [12, 13] assume that search result is not necessarily a sorted list of independent XML elements; it could also be a meaningful aggregation of XML elements from various XML documents providing thereby more complete and less noisy information for users.

In literature, there are many works on generating meaningful query results for keyword XML search. Liu and Chen [14] address the query results display. Other works focus on how to identify keyword matches [15, 16]. In XML information retrieval, Polyzotis and Garofalakis [17] were the first to propose a solution to merge search results by representing XML summaries. They grouped some XML elements and used small space to store their abstracts. Huang [18] focused on the problem of result snippet generation to generate meaningful small snippets. The model presented in [19, 20] is close to our work. It is a content-based XML information retrieval model using Possibilistic networks to merge XML elements to provide a result including all relevant information for user query.

978-1-7281-8855-3/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: Consortium - Algeria (CERIST). Downloaded on February 22,2024 at 07:18:35 UTC from IEEE Xplore. Restrictions apply.