A Crowdsourcing-based Approach for Efficient XML Keyword Search

Leila M. Amini
Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
Leila.Amini@qiau.ac.ir

MohammadReza Keyvanpour
Department of Computer Engineering
Alzahra University
Tehran, Iran
keyvanpour@alzahra.ac.ir

Abstract—XML keyword search, due to its importance and applicability, is known as a well-attentoned topic in the context of XML (query) processing. Over the years, many different approaches have been proposed in order to perform the process in an effective way. The analysis of the literature in this field addresses the lack of benefiting users’ potential and capabilities to improve the process. However, there are some user-centric studies that incorporated users in the process in some way; to the best of the authors’ knowledge, there is not any focused research work the respected literature that leverages human computation ability in this regard. On the other side, taking advantages of humans computing and processing abilities is currently known as a well-experienced strategy to solve computing problems in different contexts. Specifically, in the field of database and query processing there are several seminal works that leveraged users’ participation in order to improve the process. Following such an idea, in this paper a crowdsourcing-based XML keyword searching approach is proposed. In addition to explaining the conceptual architecture and details of different parts of the proposed approach, the efficiency of the introduced approach is proved through conducting some experiments. The authors believe that the presented solution, as the first step, can provide a roadmap for the future works in the field.

Keywords—Component; Crowdsourcing; XML Keyword Search; XML Processing; Query Processing; Human Computation; Usability

I. INTRODUCTION

Due to its capabilities, XML (documents) as a general-purpose means for storing and managing information, is among of most popular and applicable data formats ever. To manage such data, over the years, many methods and mechanisms have been proposed and leveraged [1-4]. Among them, finding desired information (on both content and/or structure of XML data) from a large amount of data is of high importance. The issue will be more important and critical when we know that most of users, i.e. information searchers, have not any (sufficient) knowledge of target data, its underlying structure and connections. For this reason, searching XML data or better to say XML keyword search has become an important research field [5-9]. In this regard, many works have been done in this context in order to improve the process in some way, including [10-12]. However, the lack of leveraging humans’, i.e. users, capabilities and potential for enhancing the process is obvious. In fact, this is a well-experienced strategy in the context of relational data [13, 14].

The rationale behind this approach is laid on the concept of from humans for humans. In other words, the queries should be processed and interpreted so that provide users (who issued queries) with the most accurate results (matches between input keyword and stored records). So, the users themselves are the most qualified ones who can interpret the input, match the keywords to the answers and validate the final results. Relying on this methodology, as experienced in other domains [15, 16], can almost guarantee more accurate and efficient results. In a general view, this approach falls under category of user-centric computing. Following this trend, in this paper a crowdsourcing-based approach for XML keyword processing and answering is proposed. To our knowledge this is the first of such works in the field. The authors believe that the present work, as the first step, can provide the researchers with a roadmap for the future works in the field. The rest of the paper is organized as follows: the background and motivation behind the study are presented in section 2. The contextual requirements and essentials of the suggested solutions are introduced in sections 3 and 4, respectively. Finally, the conceptual architecture of the proposed approach is presented in section 5 and its efficiency is studied through some experiments conducted in section 6.

II. BACKGROUND AND MOTIVATION

XML keyword search as a well-studied topic has been gained many attention over the years. In this regard, several efficient methods have been proposed, including works done in [17 - 19]. Moreover, in order to summarize the literature, the surveys reported in [20 - 22] are worth reading. On the other side, the notion of crowdsourcing was officially introduced in 2006 [23] and it has become a main stream in different context from information searching [24, 25] to image processing [26, 27] and results ranking [28, 29]. Among others, an interesting application area that some notable works have done around that is query processing in specific [30, 31].
and database related tasks [32, 33], in general. The rationale behind the idea of leveraging humans’ potential for such data manipulation tasks is providing users with more accurate and precise results. The most of queries which issued by users are in natural language and ambiguity is intrinsic to such medium. Therefore, interpretation of such query is considered a difficult task. The case will be more critical when such imprecise interpretation regarded as the foundation for generating results. Following such a manner will definitely result in inaccurate search results. From another point of view, there are usually some types of inherent semantics in the user-submitted queries. Deriving such semantics is another hard-to-perform task for the machines. The inefficiency in this phase will greatly affect the query processing and answer generation. Another important issue in this regard is determining users’ intention and aim(s) in order to find the most appropriate answer for the submitted query. As mentioned earlier, such issues are essentiality hard-to-tackle by machines due to their intrinsic complexity.

To cope with such issues, i.e. dealing with human-oriented computing challenges, the silver bullet is to take advantage of human abilities, knowledge and participation [34-36]. Following this concept, some interesting and notable works have been done in the field. To name some, the following references are worth mentioning [37- 39]. From another point of view, In the context of XML keyword search, besides current remarkable achievements, to the best of the authors’ knowledge there is only one—partially related- research paper [40]. In this study, application of crowdsourcing for relevance assessment in XML retrieval was examined. Specifically, the conducted experiments shown that general crowdworkers could provide functionality equal to what INEX experts provides. As a matter of fact, incorporating humans’ ability and participation in XML keyword search process is mainly to leverage their cognitive and problem-solving abilities for providing user-friendly results.

III. THE CONTEXT-SPECIFIC REQUIREMENTS

The related works and ideas in the previous section are mainly related to the relational data. Nonetheless, when it comes to XML context, the situation will be more complicated. As a matter of fact, XML data (documents), due to their features, in addition to the mentioned ones introduce some context-specific challenges. For example, dealing with semantic ambiguity in XML documents is a challenging issue [41, 42]. Moreover, when a user issues a query, the answer may be resided either in the content or structure of documents. For such reasons, XML keyword search and query processing—as a hot topic—has gained many attention from researchers [5-7]. Further, in order to compensate lack of human specific knowledge and intervention, several works have been focused on user-centric XML processing [31, 32]. The aim of such studies was to leverage users’ participation in the query interpretation and result generation process.

Despite the partial efficacy of such user-mediated methods, to take the most of users’ potential, a more inclusive and general consideration is needed. In fact, adhering to case-specific users’ participation may not provide general solutions. Therefore, to achieve such level of performance, the ultimate solution may be taking advantages of collective human intelligence in the form of crowdsourcing.

IV. THE SOLUTION DEFINED

In order to improve the XML keyword searching process, in this paper a crowdsourcing-based solution is proposed. Doing so the conceptual architecture and technical details will be presented and discussed in the rest of the paper. At the first step, in order to clarify the context on which the proposed solution will be introduced, the general XML keyword search process is recalled.

As illustrated in the Fig.1., there are three main phases for a generic XML keyword search [32] as follows:

1. **Query Processing**: as the first step, the submitted queries will be interpreted, analyzed and demystified in this component. Doing so, several sub-processes including grammatical and semantic validation and disambiguation, in order to identity users’ intention of issued queries, are performed. The important and determining outputs of this component will navigate rest of the process [32].

2. **Result Generation**: based on the output(s) of the first step, the users’ query will be matched with appropriate records (node or its content). The major challenge in this step is the appropriateness of query translation. In fact, there are several barriers that may affect this process including inaccurate results from the previous step, imprecise matching procedure, ill-structured, ambiguous data, incorrect query, etc [32].

3. **Representation**: the result list generated in the second step, regardless of its accuracy, will be organized and prepared in order to be presented to the user. There are two major considerations in this step. The first one is how to arrange the results through which the user can easily find the most appropriate and relevant results. This is an important issue that is regarded in designing search engines. The rationale behind this is laid on information representation and usability concepts. In other words, an improper search result representation so that users could not find them in a user-friendly way may make the entire process ineffective. The second important
According to the explained issues and requirements, the authors propose a state-of-the-art, crowdsourcing-based approach for improving XML keyword search process. The nuts and bolts of the suggested approach are described in the following sections.

V. THE CONCEPTUAL ARCHITECTURE

According to the general architecture of XML keyword search process presented in the previous section, a comprehensive solution is introduced. The proposed approach provides the process with several possibilities in order to take benefits of user’s participation.

It is worth mentioning that, however, the different components of the proposed approach are represented in a unified solution; depending on the case, those one may be leveraged separately and independently. The conceptual architecture of the proposed approach is as illustrated in Fig.2.

As a user-centric approach, the architecture is designed in such a way that user interaction and feedback can be leveraged in all the steps. The building blocks of the system’s architecture are described as follows:

- **Crowdsourcing-based query processing**: The major aim of this component is to analyze users’ query submission patterns when targeting same data item. In fact, different users may use different keywords for retrieving something identical. Therefore, having adequate knowledge of how the users search a specific item can be very useful in order to retrieve the appropriate search results. In other words, the users searching intention, from a general point of view, is identified in this step. In addition to feeding the result generation phase, the obtained insights will be used to reorganize the dataset in order to make the search, join, retrieval and similar data manipulation processes more efficient and optimized. Doing so, when the underlying data being adapted to the users requirements and needs, it is expected to retrieve more accurate results. In order to acquire large amounts of user-submitted queries for sake of further analysis, two main approaches may be employed. In one side, as an indirect strategy, regular queries issued by users over the time may be classified and analyzed. On the other side, a specific-purpose data collection procedure can be leveraged through which some thoughtfully-designed targets (e.g. ask for finding some data) will be presented to the users. Then, they asked to submit their queries to retrieve such data. Such a collective intelligence when aggregated and analyzed can provide an invaluable source of information for efficient query processing. As a well-experienced method, to improve the interaction between users and machines, the alternative or improved queries will be suggested to the users based on their initial submission.

- **Crowdsourcing-based result generation**: Users’ participations in this component, in the form of presenting feedback, are leveraged in order to improve the result generation. Specifically, the goal of this step is to best match between users’ intentions and possible results. Due to its close relationship with the first step, to provide an interactive interface to collect users’ feedback, it is variable to generate real-time answers. Then, according to what users have selected as the appropriate response, more filtered and similar ones will be provided for the user. In order to acquire more diverse and user-specific feedback, a working strategy is to present common queries as challenges or surveys to the crowds. Then, analyzing the collective answers can provide an appropriate criterion for producing calibrated search results. Further, when the crowd-contributors are experts in the field, or who are familiar with the target data, the collective intelligence can be considered gold standard. Such a criterion may be leveraged to evaluate other crowdworkers answers. When the answering patterns are revealed, the structure and organization of target data, i.e. XML tree, can be improved so that search and retrieval processes become more efficient. For example, an XML tree may be divided into two instances or several trees may be combined.

- **Crowdsourcing-based representation**: Besides the generic features of common search result lists, providing users with the options to rearrange the list, filter results based on some measures, ranking results, etc. is regarded as a powerful means to implicitly leverage crowds’ opinions to improve the representation. Moreover, the representation method
may be adopted with respect to type of results and/or users requirements. It is worth mentioning that all the processes introduced in the aforementioned steps are continuously changing according to the user’s needs and submitted queries. Therefore, some of features, including data organization and representation may be changed case by case.

- **User Management:** In order to manage the users involved in the previously mentioned processes, in this architecture, this additional and important component, relative to the classic XML keyword search process, is introduced. In fact, all the user-related issues, including call for participation, assigning tasks, connection with different parts of the system and so on will be handled in this component.

- **Task Management:** As the sibling of user management component, task manager unit is in charge of managing and organizing aforementioned tasks. After the user performed the designated tasks, the results will get back to the task manager for sake of aggregation, integration and validation.

- **Data Analysis:** The log data for the user interaction with the system and the performed tasks and their details will be directed to the data analysis component. In order to extract some useful and applicable rules to be applied in the respective component, appropriate data fusion and analysis techniques will be employed. Based on the specific tasks and the obtained results, the analyzed crowdsourced data will be returned to the respected component.

VI. PRELIMINARY EXPERIMENT

In order to evaluate the efficiency of the proposed approach, in this section a preliminary experiment, inspired by the what performed in [40, 43-44] is conducted. The major goal of this experiment is to study how crowdsourcing-based keyword search could improve users’ satisfaction levels with generated results. In this regard, three keywords (namely Nolan, Bat, Spider) were searched on the IMDB dataset [45] (i.e. a limited version with 5000 records in 28 fields (Fig.3)). Firstly, the results are generated based on LCA [46] semantics. Moreover, as the second attempt, some 20 crowdworkers asked to search for the keywords in the dataset. Then, satisfaction scores of 40 users as judges (in percentage format) with two group of results were collected. The obtained results could also be investigated from another point of view. Specifically, the number of crowd-generated results – due to incorporation of their domain knowledge and cognitive abilities- were reduced and as a result such calibrated outputs could gain higher scores (Table 1 and Fig.4-9). Such a result can be regarded as a best practice to incorporate (expert) crowdworkers in similar scenarios.

As a preliminary example, the conducted experiment shows that leveraging humans' cognitive and problem solving, specifically in the case of natural language processing-related topics could provide more reliable, accurate and user-friendly results. However, as the major drawback that all crowd-powered search engines face with, following such an approach...
will scarify time. For example, in this experiment the first approach introduces better efficiency in term of time (an average of 100 ms against average of 8 minutes for crowdsourced process). Nonetheless, it is proved that in such cases users prefer to employ a time-consuming yet more accurate process [47]. However, such remarkable time gap may be considered a drawback in real-time applications and contexts. Therefore, in order to deal with such a challenge, there is need to a large group of active crowdworkers to provide quick and accurate answers. On the other side, the crowd-provided answers could be leveraged to train machines or evaluate/filter machine-generated answers. Doing so, can be regarded as a working strategy to diminish the intrinsic time-related deficiencies.

As illustrated in the Fig.10, the average satisfaction scores for LCA-based and crowdsourced answers present a meaningful
difference. Particularly, there are 22.2%, 29.75% and 21.25% difference between users' scores on LCA-based and crowdsourced results for keywords 1-3, respectively. Such a remarkable distinction can be regarded as an important index to support efficiency (user-friendliness) of the proposed crowdsourcing-based approach.

VII. CONCLUSION AND FUTURE WORKS

Due to applicability of XML data, over the years many research works have been done regarding its different aspects. Since exploring such data is of high importance, XML keyword search has been gained remarkable attentions. Despite many works have been done in the literature around the topic, lack of leveraging human intelligence is considered as a short coming for the current works. In fact, in order to provide users with most appropriate and precise search result, incorporating their participation may be the silver bullet. In this regard, in this paper, the conceptual architecture of a crowdsourcing-based XML keyword search system is introduced and discussed. To our knowledge, the work reported in this paper is the first study in which crowdsourcing applications have been leveraged for improving XML keyword search process. As the future work, the authors are going to conduct some real-world and large-scale experiments in order to verify the efficiency and performance of the presented approach under various conditions and in different scenarios.

REFERENCES