Effective keyword query structuring using NER for XML retrieval

ABSTRACT

Purpose: A more effective way for searching XML database is to use structured queries. However, using query languages to express queries prove to be difficult for most users since this requires learning a query language and knowledge of the underlying data schema. On the other hand, the success of web search engines has made many users to be familiar with keyword search and therefore they prefer to use a keyword search query interface to search XML data. The purpose of this paper is to propose and evaluate XKQSS, a query structuring method that relegates the task of generating structured queries from a user to a search engine while retaining the simple keyword search query interface. Design/methodology/approach: Existing query structuring approaches require users to provide structural hints in their input keyword queries even though their interface is keyword base. Other problems with existing systems include their inability to put keyword query ambiguities into consideration during query structuring and how to select the best generated structure query that best represents a given keyword query. To address these problems, this study allows users to submit a schema independent keyword query, use named Entity Recognition (NER) to categorize query keywords in order to resolve query ambiguities and compute semantic information for a node from its data content. Algorithms were proposed that find user search intentions and convert the intentions into a set of ranked structured queries. Findings: Experiments with Sigmod and IMDB datasets were conducted to evaluate the effectiveness of the method. The experimental result shows that the XKQSS is about 20% more effective than XReal in terms of return nodes identification, a state-of-art systems for XML retrieval. Originality/value: Existing systems do not take keyword query ambiguities into account. XKSS consists of two guidelines based on NER that help to resolve these ambiguities before converting the submitted query. It also include a ranking function computes a score for each generated query by using both semantic information and data statistic as opposed to data statistic only approach used by the existing approaches.

Keyword: Query languages; XML database; Searching; Structured queries; Web search engines; Keyword search; Named Entity Recognition (NER)