Vector-Oriented Retrieval in XML Data Collections

Jaroslav Pokorný

Faculty of Mathematics and Physics, Charles University,
Malostranské náměstí 25, Prague, Czech Republic
pokorny@ksi.mff.cuni.cz

Many modern applications produce and process XML data, which is queried in its both structural and textual component. This is especially useful if we consider a casual user who looks for information in web-based database systems or intranets containing XML data, like online shops, airline reservations, digital libraries catalogues or any other, and does not expect an exact answer. Many websites are built from document-centric XML documents [3]. A remarkable characteristic of such XML data collections is that they are mostly heterogeneous, i.e. they contain domain-focused data, possibly valid w.r.t. various DTDs or XML schemes. XML documents can come from various sources. These collections can be managed as XML databases [5] as well as collections, providing an approximate way for users to search their contents. To ensure such functionality, it is required to approach these collections with both database and information retrieval (IR) methods.

Current XML query languages like XPath and XQuery are applicable rather for data-centric than for document-centric XML data. Moreover, XML schemes are often necessary for their use. In other words, the languages are not longer appropriate for searching in such environments because they can not cope with the diversity of data. Hence, a research of integration of database querying and IR in context of XML is undoubtedly interesting and promising trend. Despite of the fact that a variety of systems that support such methods have been proposed, conventional IR techniques [2], e.g. vector space model, can be employed only restrictedly. The reason for it is that two types of queries should be dealt with: content-only (CO) queries, i.e. the traditional ones in IR, and content-and-structure (CAS) queries.

A number of techniques to extend the vector space model have been designed, e.g. [6], [7], [8], [9], [11], and [12]. A usual critique of the mentioned approaches is that they not sufficiently reflect the structure of XML documents. A more advanced, two-phase evaluation schema is proposed in [1]. First, a modified vector space model is employed to obtain similarity scores for the textual nodes of XML trees. Then, the scores are propagated upward in the XML-trees with a possible modification and possibly new scores of other nodes are generated.

In [13] we described a matrix model based on an extension of the vector space model for XML data. A document D in a collection of XML documents C is represented by a matrix $\boldsymbol{D}$, whose each row vector w_t associated with a term t contains the weights of t for each path occurring in C. A query Q considered also as an XML tree is expressed as a matrix $\boldsymbol{Q}$. The matrix model proposes to evaluate the degree of similarity of D with regard to the Q as the correlation between the matrices $\boldsymbol{D}$ and $\boldsymbol{Q}$.

V. Snášel, K. Richta, J. Pokorný (Eds.): Dateso 2008, pp. 74–76, ISBN 978-80-248-1746-0.

Experiments have shown that it is not possible to rely only on this score. Instead we adjust the matrix ***D*** by an additional data structure, so called a path transform matrix, which reflects relationships among paths. The same is done for the matrix ***Q***. Then, the resulted transformed matrices ***TD*** and ***TQ*** are used for query processing. First experiments have been done with the well-known collection of Shakespeare's plays [4] and synthetic data generated by a widely used database benchmark XBench.

In next development of the matrix model we found its critical points and proposed its new version based on the approach [7]. In experimental implementation (called MAMEX in [14]) we used INEX collection [10] as input data. We have compared vector model and renewed matrix model and explored cases in which precision of results are comparable and cases where the latter model wins. The experiments confirmed that the matrix model is mostly not worse than vector model and is significantly better in the cases of queries with more terms. This can be of an importantance for Web querying where a page is a query unit and a collection of pages is relatively stable.

References

1. Anh, V.N., Moffat, A.: Compression and an IR Approach to XML Retrieval. In: Proc. of the First Workshop of INEX, Dagstuhl, Germany, December 2002, pp. 99-104.
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval. NY: ACM Press, 1999.
3. Barbosa, D., Mignet, L., Veltri, P.: Studying the XML Web: Gathering Statistics from an XML Sample. World Wide Web 8(4): 413-438, Springer Business + Media; 2005.
4. Bosak, J.: Shakespeare 2.00. Los Altos, California, http://www.ibiblio.org/bosak/, 1999.
5. Bourret, R.: XML and Databases, http://www.rpbourret.com/xml/XMLAndDatabases.htm.
6. Bremer, J.-M., Gertz, M.: XQuery/IR: Integrating XML Document and Data Retrieval In: Proc. of the 5th Int. Workshop on the Web and Databases (WebDB), June 2002, pp. 1-6.
7. Carmel, D., Efraty, N., Landau, G.M., Maarek, Y., Mass, Y.: An Extension of the Vector Space Model for Querying XML Documents via XML Fragments. In: Proc. of XML and Information Retrieval (Workshop) Tampere, 2002, pp. 14-25.
8. Crouch, C.J., Apte, S., Bapat, H.: Using the Extended Vector Model for XML Retrieval. In: Proc. of the 1st INEX 2002 Workshop, Dagstuhl, December 2002, pp. 95-98.
9. Fuhr, N., Großjohann, K.: XIRQL: A Query Language for Information Retrieval. In: Proc. of ACM-SIGIR, New Orleans, 2001, pp. 172-180.
10. Gövert, N., Kazai, G.: Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In: Proc. of the first Workshop of the INitiative for the Evaluation of XML Retrieval (INEX), Dagstuhl, 2002, pp. 1-17.
11. Grabs, T., Schek, H.: Generating vector spaces on-the-fly for flexible XML retrieval. In: Proc. of XML and Information Retrieval (Workshop), Tampere, ACM Press, 2002, pp. 4-13.
12. Kakade, V., Raghavan, P.: Encoding XML in vector spaces. In: Proc. of the 27th European Conf. in Information Retrieval (EPIC). LNCS 3408. Springer, NY, 2005, pp. 96-111.
13. Pokorný, J., Rejlek, V.: A Matrix Model for XML Data. Chap. in: Databases and Information Systems, Volume 118 Frontiers in Artificial Intelligence and Applications, Eds. J. Barzdins and A. Caplinskas, IOS Press, 2005, pp. 53-64.
14. Vávra, J.: Matrix model in context of XML IR methods. Master Thesis, Faculty of Mathematics and Physics, Charles University, Praha, Czech Republic, 2005. (in Czech)