Fast Similarity Search in XML Data

Dissertation

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

eingereicht von

Torsten Schlieder

am 9. Dezember 2002

Betreuer:
Prof. Dr. Heinz Schweppe
Prof. Dr. Myra Spiliopoulou

Datum der Disputation: 14. April 2003

Abstract

The eXtensible Markup Language (XML) is a widely accepted standard for the representation of data. The more data is stored in XML documents, the more important become methods for effective and efficient searching. An important characteristics of XML documents is their self-describing structure. Queries that specify selection conditions for the structure promise to greatly improve the precision of the search. However, the use of the structure can also be problematic, because it is hard for users to learn all of the details of the often complex and heterogeneous structure required to phrase a query, and because structural selection conditions often lead to overspecified queries that miss relevant results.

In this thesis, we propose an innovative method for searching in XML data, which uses the descriptive structure as a guide to locate the requested information. A user needs only partial knowledge of the structure to formulate queries that specify conditions on both the content and structure of documents. A query is interpreted in such a way that it retrieves not only exact matches, but also results considered to be similar to the query. To find the similar results, sequences of transformations are applied to the query so that its structure is adapted to the structure of each document in the collection. Each transformation within a sequence has a cost; the total cost of a sequence measures the similarity between the original query and a document matched by the transformed query. This total cost is assigned to the document and determines its position in the list of results, which is sorted by decreasing similarity. By adjusting the costs, the interpretation of queries can be tailored to the needs of different users, and also to the varied characteristics of XML documents.

We present all necessary algorithms and data structures to implement a query processor that answers a query in polynomial — typically sublinear — time with respect to the size of the database. For a given query, the query processor creates a compact query-execution plan that represents all possible query transformations. It evaluates the plan by executing operators that successively calculate the transformation costs for each document in the collection. We present techniques to effectively optimize the evaluation of query-execution plans by exploiting equivalences between operators. To reduce the query-evaluation times even more, we propose a method to retrieve the best n results, without computing similarity scores for all documents in the collection. This method uses a structural summary of the data to estimate the best k transformed queries, which are successively evaluated until the best n results are found.

The theoretical concepts are validated by a prototypical implementation. We describe the architecture of the prototype, and discuss the results of systematic tests carried out to analyze the evaluation times for a representative set of queries with respect to various collections of real and synthetic XML documents.

Acknowledgements

Many people have accompanied my long journey to this dissertation. To all of them I give my deepest thanks, especially

My first advisor, Prof. Heinz F. Schweppe, for his continuous guidance, support, and his belief in my work.

My second advisor, Prof. Myra Spiliopoulou, for her critical yet constructive remarks, which substantially improved this dissertation.

The members of the Database Group at the Freie Universität Berlin for many controversial discussions, helpful feedback, and for the great time we had together.

Among them, in particular Annika Hinze, for working as my closest colleague, sharing literally all my joys and sorrows throughout this effort.

The students and professors of the Graduate School in Distributed Information Systems for their useful suggestions and all the great discussions at our meetings.

Among them, in particular Ulf Leser and Felix Naumann, for all the help and support they gave, leading the path I was to follow.

Prof. Klaus U. Schulz from the Universität München, for enabling a very close and successful cooperation with his group.

Holger Meuss from the Universität München, for the very enjoyable, interesting, and fruitful work on our joint research project.

My family, for their patience and understanding.

Last but not least, Ella, for her encouragement and belief in me, for her warmth and love.

This research was supported by the German Research Society, Berlin-Brandenburg Graduate School in Distributed Information Systems (DFG grant no. GRK 316).

Contents