WWW Exploration Queries

David Konopnicki and Oded Shmueli*

Computer Science Department
Technion, Haifa, 32000, Israel
konop, oshmu@cs.technion.ac.il

Abstract. The World-Wide Web presents new challenges to database researchers, especially in the area of query processing. Currently, querying the World-Wide Web is done by using *online indices*. These sites employ search engines, known as "robots", that scan the network periodically and form text based indices. A severe limitation of these search services is that the structural information, namely the organization of documents into parts pointing to each other, is lost. Several tasks, ranging from data mining to Intranet management, require the analysis of the hypertext structural organization.

In this paper, we propose a simple graph based query language. In this language, both the query and its target are graphs. We present and evaluate the efficiency of a general class of algorithms for answering graph queries. The algorithms' definitions take into account two important facts of the WWW: (1) efficient algorithms must minimize the communication needed to answer a query and (2) query evaluation involves a process of data graph exploration.

1 Introduction

The World-Wide Web presents new challenges to database researchers, especially in the area of query processing. Currently, querying the World-Wide Web is done by using *online indices* such as Lycos, Infoseek, AltaVista1 and others. These sites employ search engines, known as "robots", that scan the network periodically and form text based indices. A severe limitation of these search services is that the structural information, namely the organization of documents into parts pointing to each other, is lost. Several tasks, ranging from data mining to Intranet management, require the analysis of the hypertext structural organization.

Many data organizations resemble directed graphs, especially hypertext. This has resulted in the design of query languages that view their target data as graphs [6, 11]. In the context of object oriented databases, path expressions can

---

* Work supported by grant 8528-95-1 of the Israeli Ministry of Science and the Arts and by the Fund for the Promotion of Research at the Technion.

© Springer-Verlag Berlin Heidelberg 1999
also be viewed as traversing graphs [9]. Such languages are also closely related to the semi-structured databases of [1].

In this paper, we propose a simple graph based query language. In this language, both the query and its target are graphs. We present and evaluate the efficiency of a general class of algorithms for answering graph queries. The algorithms’ definitions take into account two important facts of the WWW: (1) the dominating cost is that of communication, (2) the structure of the data graph is usually unknown prior to query evaluation. The consequences of these facts are that: (1) efficient algorithms must minimize the communication needed to answer a query and (2) query evaluation involves a process of data graph exploration.

Consider the following example. In a CS Dept., the faculty WWW pages are organized as follows. There is an index page containing a hypertext link for each faculty member. These links lead to pages that contain general information about the faculty members. These pages point, in turn, to the faculty members’ home pages. Home pages point to pages of the courses taught by the faculty members. An example of a graph query is: find pairs of faculty members’ home pages that point to the same course page. Given the WWW data graph, the query graph is shown in Figure 1. The solution of the query is a set of mappings from the query graph to the data graph such that: node $I$ is mapped to the site’s index page, nodes $a$ and $b$ are mapped to the information pages of the two faculty members, nodes $a'$ and $b'$ are mapped to their home pages and $C$ is their common course page.

**Related Work** There exist several query languages for hypertext systems that address the problem of analyzing hypertext structural information. *Graphlog* [5] can be used for the graphical specification of search patterns. In [3], Beeri and Kornatzky define a logic-based language to state queries on hypertext structures. A query language on a dynamically changing hypertext is proposed in [15]. W3QL [11] and WebSQL [14] are two languages defined especially for the WWW. Weblog is a language, based on Datalog, to query and restructure WWW information [12]. However, to date, little has been done in defining what are the requirements from a general hypertext query language, what are the basic problems involved in answering such queries and what optimization techniques should be used.

We have built W3QS [11], a query system for the World-Wide Web which uses W3QL, a graph based language. W3QS is limited to simple queries and uses naive algorithms for query processing. Extending W3QL’s query processing capabilities was the primary motivation for this paper. It is also conceivable that the techniques developed here will be useful for handling queries on semi-structured data [1, 10] and queries on XML [4] data.

In [13], the authors propose an algorithm to answer a particular class of graph queries: the regular simple path queries. Their algorithm does not take