Seminar for Computer Engineering Ph.D.

1st Semester 2000

Assignment 3 : Selected Thesis Discussion

Lecturers : Dr. Prabhas Chongstitvatana amd Dr.Boonserm Kijsirikul

This report presents the discussion on the selected Ph.D. thesis. Its title is *Integrated Query and Search of Databases, XML, and The Web* [Goldman00]. This thesis describes works towards unifying and integrating query techniques for structured data such as traditional databases, semistructured and unstructured data, search engines.

Thesis Motivations

Data published directly from traditional, well-structured databases, to the huge number of unstructured hand-written HTML pages, to the increasing amount of semistructured data, the web brings it all together into one huge amalgamation of information.

Traditional databases store large amounts of structured, typed data. Query language such as SQL can be used to get the relevant data and enable express results. However, it is extremely difficult to capture the data relationships with unstructured document such as HTML documents through a traditional database system. Therefore, information retrieval (IR) effective technologies can be used to support keyword-based searches over document data. However, these searches are less precise than SQL queries in term of relevant data.

As a bridge between these two approaches to information exchange, semistructured data was proposed to be able to handle data that has imprecise structures. However, it is either too irregular or changes too often to be constrained by the table schemas or object class required by the traditional Database Management Systems (DBMSs).

Thesis Contributions

The contributions of this thesis can be concluded as followed:

- **Lore** is the Database Management System for managing semistructured data such as XML. It has been developed at Stanford University. Semistructured data is defined as *schemaless* or *self-describing* in term of no separated description of the type or structure of data. Lore manages data in the Object Exchange Model (OEM). In OEM, structure and data are combined into a simple, graph-based object model. The primary query interface to Lore is Lorel, a declarative query language based on OQL (Object Query Language). This thesis describes the contributions to Lore DBMS for managing semistructured data, focusing on ways to enhance system usability for effective querying and searching.

- **Data Guides** is a concise, accurate, structural summary of a semistructured database as a given data graph. The Data Guide is generated dynamically from the database and is modified dynamically as the database structure evolves. The Data Guides serves the role of schema is a semistructured database. It is a tool for guiding query formulation. Not only can it be used for query optimization but it also can be used within a graphical user-interface for interactive specification of Lorel queries in Lore DBMS.

- **Proximity Search** is a searching techniques that enable effective keyword-based over traditional and semistructured databases. While query languages such as Lorel or SQL (in the relational world) can be used to exploit precise data relationships, there are some limitations on expressiveness of queries in finding data based on non-precise relationship or keywords. To be able to handle such queries, it has to search the OEM graph for objects containing some keywords and rank the resulting objects based on *proximity* to other objects. The proximity search is measured in distance on the OEM graph.

- **WSQ/DSQ (Web-Supported Database Queries/Database-Supported Web Queries)** is a new approach that combines the existing strengths of traditional databases and web searches into a single query system. It exploits existing search engines to augment SQL queries over a relational database (WSQ) and for using a database to enhance and explain web searches (DSQ). On WSQ, this thesis shows how web search engine can be modeled through virtual tables. The virtual table is similar to normal tables in a query processor but its tuples are actually dynamically computed rather than physically stored in the database.

Discussion

Strength Points

- Query capabilities have optimization of select-project-join queries

- Algorithm of the proximity search produces the shortest distance using the *best-first* search to compute the shortest path.

- Searching considers interrelationships between different fields apart from the specified fields through a query.

- *Asynchronous iteration* is a new query processing system that supports a high number of document web searches.

Weak Points

- Computing Data Guides for some databases are extremely expensive and can cause poor data guide performance.

- *Asynchronous* technique in the querying process assumes external sources continually and returns tuples immediately. So the return tuples are incomplete although they allow the query processor to confirm working without dynamic rescheduling.

- The thesis should present the comparison of the performance of asynchronous iteration with a traditional query processor.

- The thesis claims that it has WSQ/DSQ as a new approach. However, the focus on DSQ has not been explored yet because the keyword-based search still dominates searching on the Internet.

- Each chapter is not related and there are not many illustrated examples. It is also hard to follow and be clearly understood.

References

[Goldman00] R Goldman "Integrated Query and Search of Databases, XML, and The Web," *Ph.D. Thesis*, Department of Computer Science, Stanford University, May 2000.