OLAP with UDFs in Digital Libraries

Carlos Garcia-Alvarado
University of Houston
Houston, TX 77204, USA

Zhibo Chen University of Houston Houston, TX 77204, USA

Carlos Ordonez
University of Houston
Houston, TX 77204, USA

ABSTRACT

Queries on digital libraries generally involve the retrieval of specific documents, but most techniques lack the ability to efficiently explore these collections. The integration of OLAP techniques with digital libraries allows users to navigate throughout these collections on multiple levels. In order to accomplish this, we propose the creation of OLAP networks, a complex data structure that contains summarized representations of the original collection of metadata to enrich traditional retrievals and allow the users to quickly explore the collection. We developed a system that enables OLAP-based exploration on the metadata of digital libraries through the use of a combination of efficient UDFs and optimized SQL queries. In addition, we also incorporated visualization methods into our system to allow fast navigation and exploration.

Categories and Subject Descriptors: H.2.7 [Database Administration]: Data warehouse and repository

General Terms: Algorithms, Management

Keywords: OLAP, Information Retrieval, SQL

1. INTRODUCTION

Enriching digital libraries using Online Analytical Processing (OLAP) can help users to better understand such digital collections. Typically, digital libraries have been explored through the use search engines, while OLAP has generally been used for data summarization in various fields. With current searching mechanisms, we are restricted in the type of information that we can obtain. With our proposal, we will be able to answer questions such as "Rank all authors who have published more than five papers in a range of years" and "List all conferences in which a specific author published papers." As one can see, these queries are suitable for a data cube based algorithm.

We developed an OLAP system with User-Defined Functions (UDFs) that extended the algorithm presented in [1], as well as with a pure SQL approach, to build data cube summarization in the database that can explore digital library metadata, such as conference, year, author, among others. Since document metadata can be stored and managed by the DBMS, we are able to efficiently retrieve results.

This work was partially supported by NSF grants CCF 0937562 and IIS 0914861.

http://doi.acm.org/10.1145/1645953.1646307

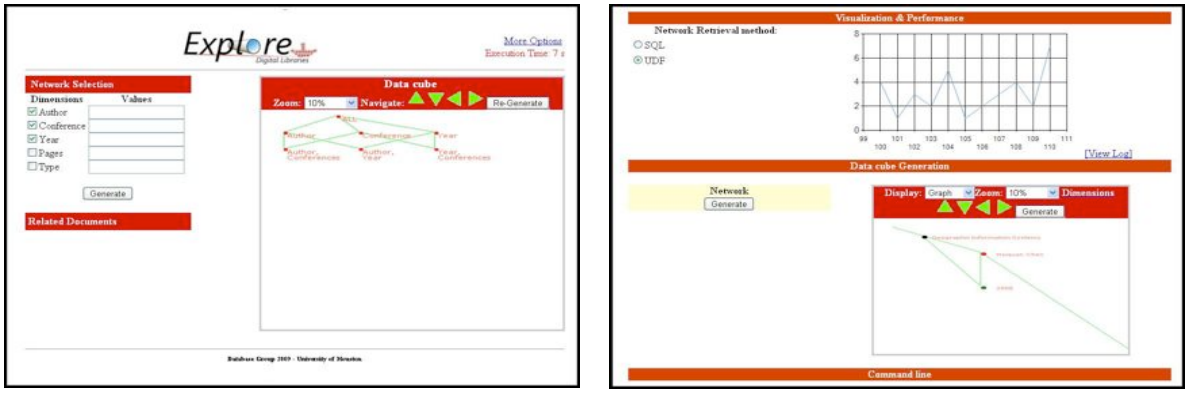
Additionally, OLAP algorithms are highly valuable for data exploration when the data sets are not high-dimensional, which is the case in digital library metadata. In such cases, the combinatorial bottleneck is reduced, and the visualization of the results is much easier. Both the UDF and SQL approaches are able to answer iceberg queries, which allow the user to narrow or expand the amount of results.

The main challenge into these problems is the ability to efficiently generate the lattice of desired attributes as well as storing the resulting lattice. We decided to build an OLAP network structure, which is new and not well-studied [2].

2. TECHNICAL DETAIL

To perform OLAP operations in the relational DBMS without the use of an OLAP server, one would normally use pure SQL. In our previous research [3], we conducted experiments and feasibility studies of using a UDF approach and found that this approach allowed for the development of more complex data structures. In order to efficiently generate the lattice of metadata within the UDF, we proposed a special data structure. While normally the most challenging problem with OLAP is the combinatorial complexity as related to the number of dimensions, the situation is different with metadata. For this type of data, the dimensionality is usually low, while the number of records is often enormous. Therefore, the difficulty for OLAP becomes the size of the data set instead of the number of dimensions.

We developed a new UDF data structure that can store OLAP values and was shown to be more efficient than SQL for most cases [3]. In this paper, we built upon this data structure to created a new structure for dealing with digital library metadata. This data structure is also an extension and abstraction of the idea presented by the authors in [2]. The UDF structure involves the creation of a network of connected vertices that represent the values held within the dimensions. These vertices are further connected by edges that represent the summation amounts, such as the number of documents, that satisfy the two vertices. For example, an edge from author to year may store the number of documents that an author produced in that year. The novelty of our work is that we are able to generate the entire OLAP dimensional lattice using only the network. To create individual nodes or cuboids of the lattice, we would need to sum together multiple edges. However, if we only store the vertex values and counts, we would not be able to create all possible subsets. This is because an edge in the network structure can represent multiple rows of the input data set. As a result, when combining edges to create nodes, we need



Cube Exploration. Lattice Generation & Options.

Figure 1: OLAP Exploration tool.

to also know the specific rows that the edge represents. We accomplish this by including the unique identifiers (IDs) of the documents, along with the number of documents, in each edge of the network. Our UDF is efficient because it is able to create the entire OLAP network using only one-pass on the metadata data set. For each row, the UDF will create d vertices connected by multiple edges. If a vertex does not already exist in the network, meaning this is the first time the specific dimension value has appear, one is created and the edges are appropriately connected to other vertices. If a particular edge does not currently exist, then one is created and initially filled with the values of that particular tuple. Finally, if the edge does exist, then the additional information from the tuple is included with the current edge. Once the network is created, it is written to the database in the form of a table with four columns: vertex1, vertex2, weight, and IDs. Further OLAP processing is conducted on this network as opposed to the original data set. The rationale is that in most data sets, there are quite a few repeated tuples that would be removed in the network.

The retrieval of the cube exploration is performed using either a UDF aggregation or a set of pure SQL queries using GROUP-BY statements. From the network, we can efficiently create the entire OLAP cube because we do not have to return to the original data set. In addition, the network is structured in such a way that it can also efficiently retrieve individual or sets of subcubes without computing the entire OLAP cube. For example, we can easily isolate all edges connected to a particular vertex by using a predicate on the network table. Once the edges are isolated, the subcube can be computed by summing all the edges together according to the stored unique IDs and values.

3. SYSTEM DEMONSTRATION

We developed a system with a Web GUI (see Figure 1), where the user is capable of building and exploring a digital library metadata lattice. The metadata collection is the well-known DBLP data set, which contains more than 1000000 entries, from which we took the author names, conferences, year, pages and publication type and stored them in tabular format in a DBMS.

Our presentation will show the generation of the OLAP network using the DBLP data set. One of the features of our application is the ability to view the network creation in real-time, which will be interesting for small data sets. We will also show how our application can graphically view the produced network in multiple ways by choosing different dimensions, for example, the user can choose to only view the network between authors and years or view the connections between authors, conferences, and years. In addition, we also provide the capability to view specific values of a dimensions, such as only viewing the connections formed by a specific author or in a specific year. These features of the application will allow the user to better understand the data set and formulate additional queries to explore the metadata collection. Additionally, the user will be able to query the network using two methods: command-line and in an interactive manner. For the command line approach, we will demonstrate how the user can input text commands into a command line to obtain the needed results. For example, we will input "authors=ALL, years=2005, conference=CIKM" to retrieve all authors in 2005 who have a paper in CIKM. This approach also allows the user to enter more complex and customized queries than the ones shown. Additionally, we will show that the user can also interactively arrive at these results by choosing to view authors and years only. Then, we will apply a constraint to fix years to 2005, and, finally, we will constrain the number of documents to be above two. We will also show how we can also choose ranges of values for the dimensions by retrieving from the years 2000-2005. One intriguing feature of our application is the ability to retrieve not only a count of the documents, but also the documents themselves. For the above queries, we can not only use a UDF to obtain the results, whose times are shown in a performance graph in the option panel, but we can also use pure SQL to extract the same information.

4. REFERENCES

[1] Z. Chen and C. Ordonez. Efficient OLAP with UDFs. In DOLAP, pages 41–48, 2008.

[2] K. Morfonios and G. Koutrika. OLAP Cubes for Social Searches: Standing on the Shoulders of Giants? In WEBDB, 2008.

[3] C. Ordonez and Z. Chen. Evaluating statistical tests on OLAP cubes to compare degree of disease. IEEE TITB, 13(5), 2008.