MINING CLICKSTREAM-BASED DATA CUBES

Ronnie Alves and Orlando Belo
Departament of Informatics,Schooç of Engineering, University of Minho
Campus de Gualtar, 4710-057 Braga, Portugal

Keywords: Clickstream Analysis, OLAP systems, Multidimensional Databases, On-line Analytical Mining.

Abstract: Clickstream analysis can reveal usage patterns on company's web sites giving highly improved understanding of customer behaviour. This can be used to improve customer satisfaction with the website and the company in general, yielding a great business advantage. Such information has to be extracted from very large collections of clickstreams in web sites. This is challenging data mining, both in terms of the magnitude of data involved, and the need to incrementally adapt the mined patterns and rules as new data is collected. In this paper, we present some guidelines for implementing on-line analytical mining engines which means an integration of on-line analytical processing and mining techniques for exploring multidimensional data cube structures. Additionally, we describe a data cube alternative for analyzing clickstreams. Besides, we discussed implementations that we consider efficient approaches on exploring multidimensional data cube structures, such as DBMiner, WebLobMiner, and OLAP-based Web Access Engine.

1 INTRODUCTION

The concepts (and techniques) of data mining and knowledge discovery could be applied efficiently on web sites (or e-commerce sites). Recently, web usage mining has attracted much attention from researchers and e-business professionals, because it offers many benefits to an e-commerce website such as:

- Targeting customers based on usage behaviour or profile (personalization).
- Adjusting web content and structure dynamically based on page access pattern of users (adaptive web site).
- Enhancing the service quality and delivery to the end user (cross-selling, up-selling).
- Improving web server system performance based on the web traffic analysis.
- Identifying hot area/killer area of the web site.

The data needed to accomplish such tasks is derived normally from a Web server log file – almost all e-commerce applications are Web based. Clickstream files are generated in order to represent information that is specific to each Web access attempt.

The recent progress and development of data mining and data warehousing technologies contribute effectively to the emergence of new data mining and data warehousing systems (Fayyad et al., 1998) (Chen et al., 1996), opening new doors to the handling of very large data files like clickstreams.

In this paper we propose a new computational platform specially design to support the exploration of multidimensional data cube structures. We also describe some of the most relevant mechanisms for exploring data cubes, and discuss the best practices in data cubes analysis and exploration.

2 EXPLORING DATA CUBE STRUCTURES

The main idea of a data warehouse is to provide decision makers with integrated information that is organized according to their requirements. Online Analytical Processing (OLAP) systems are the predominant front-end tools used in these environments. The main focus of OLAP tools is to provide multidimensional analysis over decision-support oriented data. To achieve this goal, these tools employ multidimensional models for the storage and presentation of data. In these systems,

Alves R. and Belo O. (2004).
MINING CLICKSTREAM-BASED DATA CUBES.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 583-586
DOI: 10.5220/0002631305830586

data is organized in data cubes (or hypercubes), which are defined over a multidimensional space involving several dimensions of analysis. Consequently, mining can be performed in different portions of data cubes and at different levels of abstraction (Zaiane et al., 1998).

2.1 Mining Functions of Data Cube Engines

Building a clickstream data cube allows for the application of OLAP operations − such as drill-down, roll-up, and slice and dice −, to view and analyze clickstreams from different angles, derive ration and compute measures across many dimensions (Kimbal, 2000). This greatly facilitates the exploration process, since such a process should be investigative in nature, that is, mining should be performed at different portions of data at multiple levels of abstraction improving Knowledge Discovery Process in Database systems.

We believe the following mining functions are essential for successful implementation of data cube engines, and its uses are extremely desired for clickstream analysis: data characterization, class comparison, association, prediction, classification, and time-series analysis.

2.2 Data Cube Engines

There are several projects and implementations on data cube engines. Most of them try to accomplish the features mentioned in the previous section.

DBMiner

DBMiner has been developed for interactive mining of multiple-level knowledge in large relational databases and data warehouses (Han et al., 1997). The system implements a wide spectrum of data mining functions, including characterization, comparison, association, classification, prediction, and clustering. By incorporating several interesting data mining techniques, including OLAP and attribute-oriented induction, statistical analysis, progressive deepening for mining multiple-level knowledge, and meta rule guided mining, the system provides a user friendly, interactive data mining environment with good performance.

WebLogMiner

In the WebLogMiner project, the data collected in the web logs goes through four stages. In the first stage, the data is filtered to remove irrelevant information and a relational database is created containing the meaningful remaining data. This database facilitates information extraction and data summarization based on individual attributes like user, resource, user's locality, day, etc. In the second stage, a data cube is constructed using the available dimensions. On-line analytical processing (OLAP) is used in the third stage to drill-down, roll-up, slice and dice in the web log data cube. Finally, in the fourth stage, data mining techniques are put to use with the data cube to predict, classify, and discoverer interesting correlations (Zaiane et al., 1998).

An OLAM based Web Access Analysis Engine

In (Chen et al., 1999), was described a scalable framework developed on top of an Oracle-8 based data warehouse and a commercially available multi-dimensional OLAP server, Oracle Express, which they have used to develop applications for analyzing customer calling patterns from telecom networks and shopping transaction from e-commerce sites. In (Chen et al., 2000), they have described a web access analysis engine implemented on this framework to support the collection and mining of web log records at the high data volumes typical of large commercial web sites.

3 A DATA CUBE ALTERNATIVE ENGINE FOR CLICKSTREAM ANALYSIS

Although there are many studies, implementations and proposals for efficient and effective data mining algorithms, data cube engines require fast response due to its nature of interactive mining, which poses new challenges on efficient implementation.

The main goal of this system is to develop a data cube engine based on data mining and OLAP techniques with abilities to analyze specific clickstreams from specialized data cubes. In addition, using this engine, it will be possible to:

- create efficient data cube structures for effective pattern behaviour analysis;
- create efficient data cube structures for pattern usage analysis;

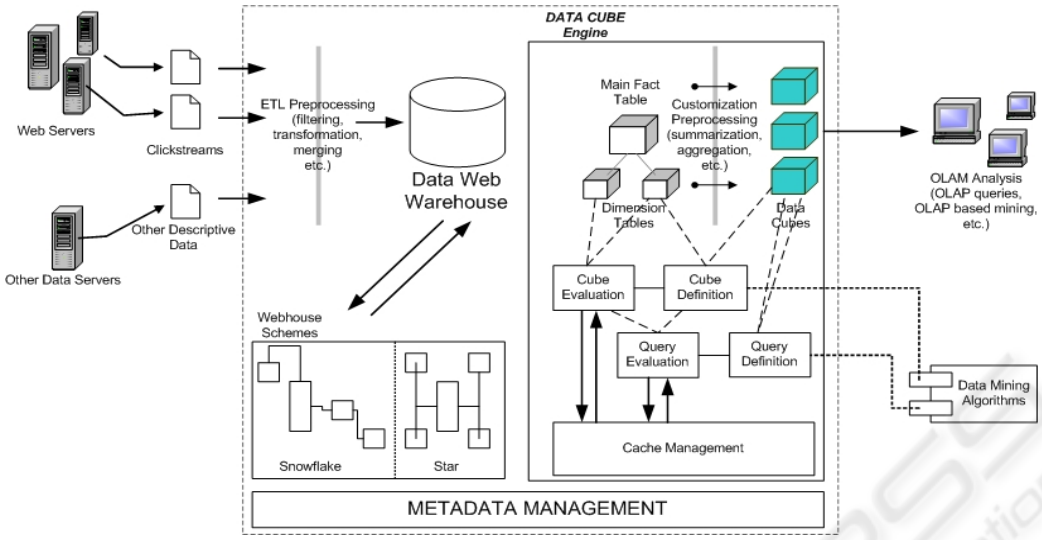

Figure 1: An overall perspective of analysing clickstreams using data cubes.

- perform efficient OLAP operations for effective exploration of data cubes;
- perform efficient mining techniques for effective discovery and understanding of clickstream data cubes.

In fact, we believe that these features will allow users to carry out several web usage mining tasks such as mentioned previously. Our motivation in this data cube approach relies not only on the implementation issues behind such integration of OLAP and mining techniques, but on the configuration, handling and deployment of data cubes for e-commerce purposes.

3.1 Implementation Guidelines

It is expected that special attention should be paid to the following implementation considerations to the successful of this approach, as mentioned in (Han et al., 1998).

- Modularized design and standard APIs.
- Support of OLAM by high performance data cube technology.
- Constraint-based on-line analytical mining.
- Progressive refinement of data mining quality.
- Layer-shared mining with data cubes.
- Book marking and backtracking techniques..

Based on these considerations in the next section we introduce our data cube alternative for analyzing clickstreams.

3.2 The Engine

In way to attend the requirements previously mentioned, the data cube engine has been built in five modules (Figure 1):

- *Cube Definition*. In this module is defined the data source which it will be used for creating the data cube. Then, it is possible to model the cube, which means designing dimensions and measures. Also, some hierarchies are defined as well as some OLAP operations.
- *Cube Evaluation*. It is responsible for evaluating cubes generated, which means to verify if all the constraints and the requirements on the cube definition engine are satisfied.
- *Query Definition*. This module is responsible for describing the query for exploring data cube using OLAP operations. And sometimes, interchange these OLAP operations with mining techniques.
- *Query Evaluation*. This module acts in the same way as the cube evaluation module. But, in this case, it analyzes the query with constraints and definitions of the data cube that would be explored, checking the consistency and presenting query results.
- *Cube Mining*. This system's module must be used for mining data cube using the techniques available inside the engine. It is also used when some mining query is required by the query evaluation module.

The process of analyzing clickstreams, using data cube structures, begins with some ETL tasks on the

clickstream. Next, a ROLAP database is used for storing the clickstreams. As long as the clickstream are available on the database it is possible to perform exploratory analysis using data cubes. Besides, before any kind of exploration over data cubes a multidimensional structure needs to be built. This is supported using the data cube definition module. It includes the following attributes: page_dimension, time_di-mension, date_dimension, user_agent, referrer_di-mension, request_dimension, and session_di-mension.

Building this clickstream data cube allows the application of OLAP operations, to view and analyze the clickstreams from different angles, derive ratios, or compute measures across many dimensions. The data cube structure offers analytical modelling capabilities, including a calculation engine for deriving various statistics, and a highly interactive and powerful data retrieval and analysis environment. It is possible to use this engine to discover implicit knowledge in the clickstream data cube. The knowledge that can be discovered is represented in the form of rules, tables, charts, graphs, and other visual presentation forms for associating or classifying data from clickstream data cube (Figure 2).

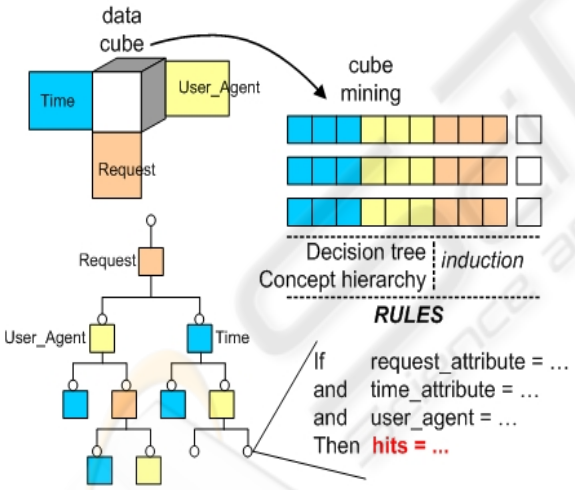

Figure 2: Discovering implicit knowledge on clickstream data cubes.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we have discussed some implementation issues on on-line analytical mining, specifically on the data cube engines and its desired functions. In fact, the observations mentioned in (Han et al., 1998) (Chen et al., 1999) (Chen et al., 2000) (Han et al., 1997) motivated us to study the desired way to perform data cube mining and its efficient implementation. As a result, we present our data cube mining engine proposal, which contains some guidelines and perspectives of research in applying data cube techniques for analyzing clickstreams.

REFERENCES

Chen, S., M., Han, J. and Yu, S., P., 1996. Data Mining: An overview from database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-8883.

Chen, Q., Dayal, U. and Hsu, M., 2000. An OLAP-based Scalable Web Access Analysis Engine". HP Labs, Hewlett-Packard, 1501 Page Mill Road, MS 1U4, Palo Alto, CA 94303, USA.

Chen, Q., Dayal, U. and Hsu, M., 1999. A Distributed OLAP Infrastructure for E-Commerce. Proc. Fourth IFCIS Conference on Cooperative Information Systems (CoopIS'99).

Fayyad, U., M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy., R., 1998. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

Han., J., 1998 Towards on-line analytical mining in large databases. ACM SIGMOD Record, 27:97-107.

Han, J., Chee, S., and Chiang, J., Y., 1998. Issues for On-line Analytical Mining of Data Warehouses, SIGMOD'98 Workshop on Research Issues on Data Mining and Knowledge Disvovery (DMKD'98).

Han, J., Chiang, J., Chee, S., Chen, J., Chen, Q. , Cheng, S., Gong, W., Kamber, M., Liu, G., Koperski, K., Lu, Y., Stefanovic, N., Winstone, L., Xia, B., Zaiane, O., R., Zhang, S. and Zhu H. 1997. DBMiner: A system for data mining in relational databases and data warehouses. In Proc. CASCON'97.

Kimbal. R., 2000. The Data Webhouse Toolkit, Wiley.

Zaiane, O., Xin, M., and Han. J., 1998. Discovering web access patterns and trends by applying olap and data mining technology on web logs. In Proceedings of Advances in Digital Libraries Conference (ADL), pages 19—29.