A new web information fusion tool for web mining

Zuyi Chen [1, a], Taixiang Zhao [1,b]

[1]Department of foundation First Aeronautic Institute of the Air Force Xinyang, China

[a]zychen7410@163.com, [b]ztx304@126.com

Keywords: Web warehouse; Web mining; Data warehouse; Data cube

Abstract: Most data warehouse systems have some limitations in terms of flexibility, efficiency, and scalability.In particular, the sizes of these data warehouses are forever growing and becoming overloaded with data, a scenario that leads to difficulties in data maintenance and data analysis.this reserch introduce a tool – web warehouse, for web mining and knowledge discovery.To formulate a web warehouse, a four-layer web warehouse architecture for decision support is firstly proposed. In the web warehouse process model, a series of web services including wrapper service, mediation service,ontology service and mapping service are used. Particularly, two kinds of mediators are introduced to fuse the heterogeneous web information.

1.introduction

In designing a data warehouse, one would first integrate into one another various data sources from an enterprise's heterogeneous databases. Whenever data from multiple sources has to be consolidated, developers must analyze the structure and the content of the source before defining the rules for merging. To execute these rules, researchers in the field must develop a satisfactory process. The integration should ensure results that exhibit data consistency acrossthe entire enterprise. Ideally, end user should be able to access data from the data warehouse without knowing either where data resides or the form in which it is stored. The architecture of data warehouses falls into three categories [6,7,8]:

1.Virtual view approach: The repository of the data warehouse contains only the data schema, and the local databases store the physical data. This approach uses query pre-processing and query shipping to answerqueries that queries make against the integrated view. The disadvantage of this architecture is its poor performance.

2.Materialized view approach: The repository of the data warehouse contains the data schema and the physical data. It collects all relevant information in the data warehouse. The disadvantage of this architecture is the difficulty it poses in the managing and the maintaining of a huge data bank.

3.Datamart approach: This approach extracts data from a primary data warehouse for Datamart application.The extraction has a special purpose. Hence, the repository of the data warehouse contains the subject of information, which is in Datamart format. The disadvantage of the architecture is that a data warehouse contains only limited knowledge. This approach cannot manage huge amounts of data.

Designing a data web warehouse entails transforming the schema that describes the source data into a multidimensional schema for modeling the information that will be analyzed and queried by users [2]. Actually, the construction of web warehouse is a process of web information fusion that integrating web information from different sources into web warehouse. Thus, a web warehouse used in this paper refers to a single, subject-oriented, integrated, time-variant collection of web information that supports the web mining and knowledge discovery. Due to the fact that web information does not only include structural data but also semi-structured text, web warehouse can be seen as a federated warehouse integrating data warehouse [1,3,4] and text warehouse [5].

In this study we propose a generic framework architecture to design a web warehouse, similar to the construction process of data warehouse [1]. In the proposed framework for web warehousing, we view the Internet or the WWW as data sources. The information including structured data and semi-structured text is extracted from the web, refined, transformed, and placed into the web

warehouse for later use. In our proposed framework, a configurable extraction program presented by [2] is first used for converting a set of hyperlinked HTML pages into database objects. After extraction, the converted results are refined, integrated and transformed to appropriate forms or formats and then loaded into the web warehouse for web mining and knowledge discovery purposes.

2. Related research

Although much research has been done on data warehouse design, work on specific methods for the improvement of data warehouse efficiency is scarce in the literature. This research here improves data warehouse efficiency by identifying the semantic relationships that exist between data warehouse data cubes.

A multidimensional database or data cube consists of (1) a huge amount of attributes in the fact table and(2) a relatively small set of dimensions with respect to which the data is analyzed. For example, a car cube may store sales amounts, sales models, and so forth. Many research papers have discussed the implementation of data cubes in data warehouses.

In the 1970s, Codd developed a relational model to organize data into databases [7], but it supported only text, the numeric data types that are insufficient for complex applications. The object-oriented model emerged and reflected researchers' attempts to solve the problem in the relational model . At present,most data warehouse systems use a relational data model that conveniently transforms relational databases into data warehouses. During the query process, it is difficult to represent some information in a relational data model, especially in terms of abstract semantics. The semantics of the OO model are much richer than the semantics of the relational model. Most research has suggested that the OO model is appropriate for the development of a data warehouse. The OO techniques are therefore widely used in data[8].

3. The generic formulation process for web warehousing

3.1. The general web warehouse architecture for decision support

The emergence of web warehouse architecture is to respond the evolving data and web information requirements.Initially, the classic data warehouse was used toextract transactional data from operational systems to perform on-line analytical processing (OLAP). Because there are different data types, such as structured data and semi-structured text, among web information, the traditional data warehouse, which only handles the structured data, does not deal with semi-structured texts. Therefore, a new warehouse architecture should be presented to meet the practical requirements. In such situations, web warehouse is proposed in response to the increasing web information.For this point, data warehouse will be evolved into a federated data warehouse, i.e., web warehouse. the general web warehouse architecture for decision support consists of four layers:data source layer, warehouse construction layer, web mining layer and knowledge utilization layer. The data sources layer of web warehouse is composed of the organization's sinternal data sources including daily operation data, internal files and OLTP data, etc. and external web text repositories and electronic messages.

3.2. Fusion

When the data are extracted through the wrapper services,many structured data with different schemas can beobtained. In order to formulate a unified view for these extracted data from various sources, it is very important to maintain an integrated schema. In this study, we use two kinds of mediation services to integrate web information from different sources. The two kinds of mediation services are designed to fuse schemas without and with structural heterogeneities. an integrated schema that reconciles structural heterogeneous information can be generated.

In order to solve data conflicts in information fusion we follow the approach presented in, which is based upon a conceptual representation of the data warehouse application domain. The main idea is to declaratively specify suitable matching and reconciliation operations to be used in order to solve possible conflicts among data in different sources.

Another solution to information fusion is ontologybased services. The goal of using ontology services is to resolve mainly the heterogeneity problem by performing mediation processes. These processes exploit some formal ontologies which take important part in the architecture. Ontology services aims to define the semantic description of services using ontological concepts. According to Gruber, an ontology is defined as an explicit and formal specification of a conceptualization. In general, the construction of domain-specific ontology is of utmost importance to providing consistent and reliable terminology across the warehouses. Similarly, hierarchical taxonomies are an important classification tool and here it is used as information integration tool. It can assist analysts in identifying similar, broader or narrower terms related to a particular terms thereby increasing the likelihood of fusing similar information from different information sources.In addition, active rules and heuristics associated with object types as well as their attributes and functions can also be used for information integration. Content fusion in heterogeneous environments can make use of ontologies, particularly in the area of catalog integration.

3.3 Metadata and loading

After extraction, fusion and transformation, the information is classified and indexed, and metadata is created in terms of domain concepts, relationships and events. In web warehouse, the domain contexts and domain usage constraints are specified. Data pedigree information is also added to the metadata descriptors, for example, intellectual property rights, data quality and source reliability etc. In addition, web mining and data analysis techniques can be applied to discover patterns in the data, to detect outliers and to evolve the metadata associated with object descriptors.

Usually, the metadata of the web warehouse consists of five kinds of information: (1) web pages classification criteria, (2) web warehouse design criteria, (3) mappings betweenthe web pages and integrated global schema and between the integrated schema and the web data warehouse , (4) the semantic and structural correspondences that hold among the entities in different sources schemas which is used for information integration or fusion), and (5) the classification of the web data warehouse schema structures according to the dimensional model.

In order to explore the hidden knowledge, the transformed or refined data, metadata, and knowledge should be loaded into web warehouse and stored in the web warehouse. For fast retrieval purpose, the loaded information should be indexed using multiple criteria, for example, by concept, by keyword, by author, by event type or by location.In the case where multiple users are supported, these should be indexed by thread, and additional summary knowledge may be added and annotated.So far, a whole EFML process for web warehouse construction is completed. To give a direct view and understanding to this process model, an illustrative example ispresented in the next section.

3.4 The EFML process for web warehousing

Since the quotation information is displayed by HTML format and it cannot be queried directly by users. Therefore,we first have to extract the contents of the quotationtable from the underlying HTML page.we can extraction the quotation information from the web. Typically, the extraction process in this case is performed by five commands. After the file has been fetched and its contents are read into root, the extractor will filter out unwanted data such as the HTML tags and extra uninteresting text.

4. Conclusions

In this paper, a new web information fusion tool web warehouse, which is suitable for web mining and knowledge discovery is proposed. In the proposed web warehouse, a layered web warehousing architecture for decision support is introduced. In terms of the four-layer web warehouse architecture, an EFML process model for web warehouse construction is then proposed. In the web warehouse process model, a series of web services, including wrapper services, mediation service, ontology service and mapping service are used. Particularly, two kinds of mediators are introduced to fuse the heterogeneous web information. Finally, an illustrative example is presented to interpret the web warehouse construction process. The experiment implementation process implies that such a web warehouse can not only increase the efficiency of

web mining and knowledge discovery on the web, but also provide an effective web information fusion platform.

References

[1] R. Kimball, The Data Warehouse Toolkit: Practical Techniques for BuildingDimensional Data Warehouses, John Wiley & Sons Inc, New York, 1996.

[2] B. Vrdoljak, M. Banek, S. Rizzi, Designing web warehouses from XML schemas,in: Proceedings of 5th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2003), Prague, 2003, pp. 89–98.

[3] W.H.Inmon, R.D. Hackathorn, Using the Data Warehouse, John Wiley & Sons Inc., 1994.

[4] W.H. Inmon, Buliding the Data Warehouse, John Wiley & Sons Inc.,1996.

[5] M.Z.Bleyberg, K.Ganesh, Dynamic multi-dimensional models for text warehouse, Computing and Information Sciences Department, Kansas State University, 2000.

[6] C.M.Chao, Incremental maintenance of object-oriented data warehouses, Information Sciences 160 (2004) 91–110.

[7] Gutierrez, A. Marotta, An overview of data warehouse design techniques, ReporteTe´cnico INCO-01-09. InCo-Pedeciba, Facultad deIngenier a, Universidad dela Republica, Montevideo, Uruguay. Noviembre, 2000.

[8] Y.Liu, S.Y.Sung, H.Xiong, A cubic-wise balance approach for privacypreservation in data cubes, Information Sciences 176 (2006) 1215–1240.