BUILDING MULTIMEDIA DATA WAREHOUSES FROM DISTRIBUTED DATA

e-Gnosis, año/vol. 2
Universidad de Guadalajara
Guadalajara, México

2004
Tania Cerquitelli / Genoveva Vargas Solar / José Luis Zechinelli Martini
BUILDING MULTIMEDIA DATA WAREHOUSES FROM DISTRIBUTED DATA

CREACIÓN DE DEPÓSITOS DE DATOS MULTIMEDIA A PARTIR DE DATOS DISTRIBUIDOS

Tania Cerquitelli¹, Genoveva Vargas-Solar², José Luis Zechinelli-Martini³
taniacerquitelli@tiscali.it / Genoveva.Vargas@imag.fr / zechinel@mail.udlap.mx

Recibido: noviembre 12, 2003 / Aceptado: enero 30, 2004 / Publicado: febrero 18, 2004

ABSTRACT. Multimedia data mediation is characterized by three important aspects: multimedia data integration, system architecture and global query evaluation. Multimedia mediation problem can be seen from an architectural point of view in which suitable mediation architectures, adapted to multimedia data characteristics (e.g. volume vs. communication and bandwidth costs) must be identified. Given multimedia data characteristics (distributed, homogeneity, volume, etc), the integration process requires time and resources. The contribution of our work is associated to multimedia data exploitation. We provide mechanisms for analyzing multimedia data collections coming from heterogeneous and distributed sources. Our solution proposes a mediated query service for distributed multimedia data, adapted to build multimedia data warehouses.

KEYWORDS: Computer Science, multimedia data, multimedia mediation, multimedia data análsis.

RESUMEN. La mediación de datos multimedia está caracterizada por tres aspectos importantes: la integración de datos multimedia, la arquitectura de sistema y una evaluación global de las búsquedas. El problema de la mediación multimedia puede ser visto desde un punto de vista arquitectónico en el que deben identificarse arquitecturas de mediación adecuadas, adaptadas a las características de los datos multimedia (como por ejemplo, volumen vs. costos de comunicación y ancho de banda). Debido a las características de los datos multimedia (distribución, homogeneidad, volumen, etc.), el proceso de integración requiere de tiempo y recursos. La contribución de este artículo se asocia con la explotación de los datos multimedia. Nosotros proporcionan los mecanismos para analizar las colecciones de datos multimedia provenientes de fuentes heterogéneas y distribuidas. Como solución se propone un servicio de búsqueda para datos multimedia distribuidos, adaptados para crear depósitos de datos multimedia.

PALABRAS CLAVE: Computación, datos multimedia, mediación multimedia, análisis de datos multimedia.

Introduction

Data mediation is a technique that enables applications and users to access transparently, distributed, autonomous and heterogeneous data sources giving the illusion of a single, homogeneous and centralized system. Different mediation architectures have been proposed that can be classified as virtually materialized systems (e.g. federated database systems, multi-databases) [1, 2, 8], materialized systems [7, 9, 4] (e.g., data warehouses) according to the strategy used to retrieve and integrate distributed data.

Two aspects are important in building a mediation system: data integration and mediation system architecture. Data integration refers to the problem of combining data residing in different sources, and
providing the user with a unified view of them. Such a unified view is structured according to a so-called global schema that represents the intentional level of the integrated and reconciled data. The integration can be logic and physic. Logic integration of multimedia data is based on a strongly coupled strategy that uses a global schema built according to Local as view (LaV) or Global as View approaches (GaV) [3]. Physic integration is done by materializing, data stored in different sources, in a single repository. Such approach is based on logic integration that guides the way heterogeneous data must be homogenized ("cleaned") in order to be stored a repository (data warehouse).

This paper proposes a mediated query service which is a system used for configuring mediation systems that can be used for building and maintaining multidimensional multimedia data warehouses. Multimedia data exploitation (retrieval, consolidation, analysis, visualization) requires adaptable mechanisms that organize data according to different user analysis needs. Accordingly, the remainder of the paper is organized as follows. Section 2 discusses problems to be considered for analyzing multimedia data. Section 3 describes our approach for building adaptable multimedia data warehouses. Sections 4 and 5 give respectively an overview of how a multimedia data warehouse can be queried and built. Finally Section 6 concludes the paper and discusses research perspectives.

**Multimedia data analysis**

Consider an application providing information about Italy and Mexico as multimedia documents. Information concern different topics, for example tourism, economic situation and investment policies available in both countries, cultural places, and geographic description of regions. Information can be organized according to intervals or points in time, for example, cultural activities available in July or in summer. Assume that two users exploit environment multimedia data. The first one searches images and videos of South Mexico Beaches, with information about cities, from May to July; the second, searches videos of the first ten Italian's and Mexico's beaches better tourist in the last year. How can required data be efficiently retrieved for without doing search operations into a very large set of data? To satisfy these needs it is necessary to have a broker capable to provide transparent access to multimedia data and that provides mechanisms for visualizing results.

However, our users may require analyzing data about Mexico's beaches. The first may want to analyze images about beaches in Mexico with likeness average equal to 80%; the second may want to analyze images about beaches in Mexico with a given resolution average. Users would not like to receive a very large quantity of data and analyze them “by hand”. Non automatic analysis on a large data collection can be complex but once multimedia data are involved, the process can be almost impossible! Existing systems are limited for enabling multimedia data analysis. Given different set of analytical requirements associated to the same collection of multimedia data, it is necessary to have a system that provides different views of the same data. Such requirements influence the way multimedia data are analyzed and exploited (how to synthesize data?) and they concern observation criteria needed to analyze data and visualization format (how to present multimedia data analysis results?).

From the point of view of multimedia data analysis, design strategies based on multidimensional models must be explored; and mechanisms must be specified for helping to the expression of analytical queries and for supporting their processing. Provide mechanisms adapted for analyzing of multimedia data implies considering their inherent characteristics such as distribution, heterogeneity, and volume, semantic
heterogeneity of their content, spatial and temporal characteristics.

**Adaptable multimedia data warehouse**

An adaptable data warehouse provides the illusion of a single and homogeneous database adapted for different applications needing to reason about the same collection of historicized multimedia information. Applications express analytical queries using their own languages and interact with the data warehouse through an adapter. Each adapter provides an application schema. An application schema is a view on the data warehouse multidimensional schema. It specifies the set of terms that can be used by the application and the set of multimedia data types that it requires organized according to dimensions. It also associates each multimedia type with aggregation functions, a default presentation that specifies how to present analytical query results.

We based our data warehouse schema in the multidimensional model defined in [6]. Then we define dimensions and measures in according to user needs and media requirements. For example, given three dimensions: place, environment and text, it is possible define measure of similarity between documents that could be the measure of the fact table. To calculate which documents satisfy the possible user queries we use a vector space model where a vector is used to represent each document in the collection, as is specified in [5]. To reduce the space it is frequently used the Single Value Decomposition and after calculate the measure of similarity as the cosine of the angle between the query and the documents vectors.

The proposed solution is adaptable because it provides access to multimedia data to heterogeneous sets of users with different needs. All users interact with the same data collection (data are not replicated) and system without knowing its characteristics, functionalities and complexity, but they customize it according to their information needs.

**Exploiting multimedia data warehouses**

Query expression. A multimedia data warehouse provides "user friendly" interfaces where users can express their queries without having to know details about analytical query languages. In our approach, an application schema representing available data is visualized and browsed through an animated hyperbolic tree. Nodes represent data types of a specific context (i.e., environment) and the graph represents a classification of those types. In our example, classification terms are natural resources, sea, mountain, plain, deserted, river and natural reserve. Each of them has its associated subtypes, for instance, natural resources has the associated subtypes: mines, forest and atmospheric agents. A query can be specified by defining three elements: (i) the topic within a given domain. In our example INFORMATION TYPE -> TOURIST; TIME. RANGE (for specifying range from May to July). (ii) A domain within the classification graph. In our example, the user is interested in ENVIRONMENT INFORMATION -> THE BEAUTY OF NATURE -> SEA -> BEACH -> STATE (for specifying State Mexico); and BEACH -> HOTEL. (ii) The types of data corresponding to a given topic: DATA TYPE -> SIMPLE -> VIDEO, IMAGE, TEXT. Users specialize queries by specifying filters concerning descriptive attributes (i.e. author, resolution, duration and dimension) and content (i.e. likeness average about colour, sound and form). For example, a user might like to see images and videos concerning wonderful beaches in Mexico from May to July and associated to average hotel prices organized according to the season of the year.
Query processing. Similar to the unfolding phase in mediation systems, given a query expressed by an application, the adapter transforms it into an analytical query according to the data warehouse schema so that it can be processed. Interaction rules between the application schema and the multidimensional schema define instances of concepts and relationships between them. In this context rules associate words used in application graph with dimension and measures used to define multidimensional data warehouse schema. Considering that analysis results involve multimedia data, mechanisms are proposed for presenting such results according to application execution environments.

Construction. Adaptable multimedia data warehouse construction is based on a mediation system, configured for accessing sources needed to build it. Such a mediation system implements a global schema that represents the universe of information provided by sources. Given an (i) application schema that represents data types and analytical criteria needed by an application; and (ii) a data warehouse schema which is a multidimensional view on the global schema, the mediator retrieves data from heterogeneous, distributed and autonomous multimedia sources. Then, it integrates results according to the data warehouse schema and builds (refreshes) the data warehouse.

Building multimedia data warehouses

The construction of data warehouse requires to retrieve data from different sources, to integrate them, to express views on the global schema according to user needs, to organize data according to data warehouse multidimensional model and store data in the data warehouse.

Mediation system specification. A mediation system can be specified giving a (i) mediator schema, (ii) an application schema, (iii) a data warehouse schema, (iv) data sources exported schemata; and (v) a set of transformation rules. Schemata are expressed under a semi-structural pivot data model, transformation rules are expressed by first order logic expressions. Using this information, data types specified in an application data schema are mapped with the data warehouse schema. Data types in the application data schema are associated with an aggregation function that computes analysis measures; and they are organized with respect to dimensions. Transformation is expressed by transformation, generation and interaction rules. Transformation rules specify mappings between sources exported schemata and the global schema. Generation rules describe the mapping between data warehouse analytical criteria and multimedia data types of the global schema. Interaction rules describe the mapping between a data type specified in an application schema and a type of the data warehouse schema.

Configuration. According to transformation rules, adapters are configured for interacting with the data warehouse. Then, the data warehouse is configured for communicating with mediator for refreshing data. Finally, the mediator is configured for using specific wrappers to retrieve objects from sources. Interaction and generation rules specify how to configure adapters so that they can be used by applications to communicate with the data warehouse and specify how to configure the data warehouse so that can communicate with mediator. Three application programming interfaces are generated wrapper API and adapter API, that includes the data warehouse API.

Refreshment. We assume that a multimedia data warehouse integrates data that can be used to answer a given set of queries. However new analytical query types can trigger the extraction of new data (i.e., data
warehouse refreshment). In our approach, multimedia data warehouse refreshment is triggered by data sources updates, periodically (e.g., daily or weekly) and by queries needing new data (e.g., average of the number of visitors during the last hour).

**Conclusion**

Physical integration is well adapted for enabling transparent access to distributed multimedia sources. Having data materialized in a single repository can be more efficient for applications needing to access multimedia data, even if maintenance and space costs can be elevated. A data warehouse centralizes data needed for analytical operations. In the case of multimedia data this can increase query processing performance since costs associated to distributed data retrieval and transport, and aggregated values computation, are solved a priori (during the data warehouse construction). Data warehouse maintenance (i.e., construction and refreshment) is done independently from the analytical process. Our research contributes to the construction of multimedia data warehouses, considering adaptability and the inherent characteristics of multimedia data.

The main result of our investigation is the definition of an approach that enables the specification mediation systems adapted for multimedia data analysis and exploitation. With the proposed approach it is possible to configure each wrapper in according to the source needs and solve problems related to mapping between relational model used in the source and XML-exported schema. This solution provides transparently access to the source. It is possible configure each adapter in according to application needs. Each of them implements associated rules, based in the first order logic, to solve the mapping between the data warehouse schema and the application schema. Last, but certainly not least, the proposed approach reduces the time of the exploitation data and provide materialized views defined all right to user analytical criteria used in the analysis of multimedia data.

**References**