International Journal of Research
Available at https://edupediapublications.org/journal

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 07
June 2017

Extracting features and content of images to model a data warehouse

Manvi Breja [1], Gunjan Chandwani[2]

[1,2] Manav Rachna University, Faridabad

Abstract:

Data Warehouse is an information system where recent and historic data, related to a topic are unified at a common place. The data is used by companies for their strategic decision making. With the increase in growth of data in the form of images over World Wide Web (WWW), there is a need to organize such data that can help us for decision making. This motivates us to build an image data warehouse. The paper proposes an architecture to model an image warehouse based on the content, metadata and features extracted from images. The metadata as well as images description used as sources of the warehouse are specified by various standards like MPEG-7, JPEG, GIF.

Keywords

Warehouse, OLAP, multimedia, ETL

1. Introduction

Data warehouses is a system that collects heterogeneous and distributed data from various sources and stores them at a common place, so that it can be used to perform decision analysis. OLAP environment designed for traditional applications, are used for the analysis of facts which contain mainly numeric data (e.g., sales depicted by amount or quantity sold). But today, in many fields, like in medical or bioinformatics, multimedia data is used, which provides valuable information and which involves in the decision process. But, integrating multimedia data in a warehouse arise many problems. We have to deal with dimensions which are built on descriptors and can be obtained by various computation modes. And the speed with which the information from multimedia data is retrieved, also creates problem and thus makes, difficult to handle such data.

It is easy for human beings to extract information of text files, sounds, and images, but automating the process of step by step extraction of information requires a refined methodology. Various techniques are deployed to extract information from various types of multimedia objects. Because, manually searching and extracting the entire information of multimedia data requires various efforts. To overcome these constraints, various processes are used as per the requirement, some of which are systematic, automated or semi-automated. Although there is a very matured data warehouse technologies present to handle numerical and symbolic data [1], but there is a lot more to do, to handle complex, multimedia data warehousing [2]. Various areas like Defense research, Geological research, Weather forecast, Marine research etc. use complex data, comprising of various formats like audio, video and text. This complex data needs to be stored and has to be retrieved and processed on requirement. Presently, data warehouses are most commonly used to handle the above requirement. Since relational databases store structured data only and multimedia data is often semi-structured, therefore DBMS systems can't handle these multimedia data. Thus, there is a need for the deployment of new techniques to store, retrieve and process the multimedia data.

2. Related Work

Although, a lot of work has been done in the area of multimedia information retrieval and data warehouse, but still, the field of developing multimedia data warehouse need a lot of attention. As multimedia information is being used in almost in every field, whether economics, medicines, education, etc.. there is a need to find a proper solution for storage and retrieval of multimedia information. When dealing with multimedia, media files are separated during data retrieval or processing [4]. In [7] a hierarchical method for storing these media files has been described. In order to enhance the data ware house semantically, a system is implemented in which medical related data has been extracted from the database and analyzed [8]. In the existing methodology to retrieve multimedia information from a warehouse, data is directly extracted from warehouse. There are many activities like creation of fact table [4], data cubes [5], multidimensional data cube [10], mapping between different types of data which are semantically same, summarization [10] or designing a schema for representation of multimedia data [5] which interacts directly with the warehouse and thus, it increases the time to retrieve data for these activities. Like in [4], each fact table of each data mart is stored in a different XML file; and hierarchical structure is created of the data marts, which contains a fact table as a dimension for another fact table. Such a fact file contains the result of the aggregation and at least one set of references for each dimension is used in the aggregation process. To speed up the process, they have developed a procedure to automatically create new queries, creates an XML file, containing technical terms associated with every existing fact table. Similarly [10] generates summary tables by mapping different media data sources to the data warehouse, dynamic indexing is used to speed up the retrieval task.

3. Proposed Architecture

This section describes the proposed architecture for modeling a data warehouse which comprises four modules such as ETL process, low-level feature extraction, high-level feature extraction, Data staging and OLAP tools and report generation is shown in figure1.

3.1. The ETL Process:

ETL stands for extraction, transform and load. This process extracts the image metadata, content and its features. The metadata includes the name, file length, format, author, date of creation/modification etc., of image ; the content comprises the size, width, brightness etc.; and features such as detecting corners, edges, textures, color, and shapes etc. from the disparate image sources (MPEG, JPEG, GIF etc.). This data is then transformed into a common image format. The feature extraction is a two level process, viz. low-level and high-level.

3.2. Low-level feature extraction:

Low-level feature extraction involves finding corresponding points between images, finding edges or lines in an image. All features dealing with pixel intensities or colors are considered as low-level features.

3.3. High-level feature extraction:

High-level feature extraction involves body pose classification, face detection, classification of human actions, object detection and recognition and so on.

3.4. Data Staging:

The Data Warehouse Staging Area is a temporary intermediate storage area that stores the results of extract, transform and load (ETL) process. It is volatile in nature, as its contents get automatically erased before running or successful completion of an ETL process. They hold data for historical or troubleshooting purposes.

3.5.The OLAP Tools and Report Generation:

OLAP Tools comprises of OLAP server, Content Server and Knowledge Server. OLAP server is used for strategical analysis on the ware house. Content Server provides image content to front end tools which act as a friendly interface to applications and external services. Content Server stores the domain specific metadata and manages its lifecycle. It provides a query interface for retrieving the content while hiding the details of how and where files and metadata are stored. The Content Server manages both the retrieval of image content and their metadata internally. The OLAP tool is a kind of Knowledge Server which provides the intelligence to the image warehouse [17]. The Knowledge Server stores all the results of SQL, OLAP and DM. OLAP reporting and data mining capabilities allows the analyst to create a web-based report, which provides clients a multi-dimensional view of vital data. We can perform a variety of analytical operations like consolidation, drill-down, slicing and dicing, and nesting on the data warehouse that can provide increased query performance.

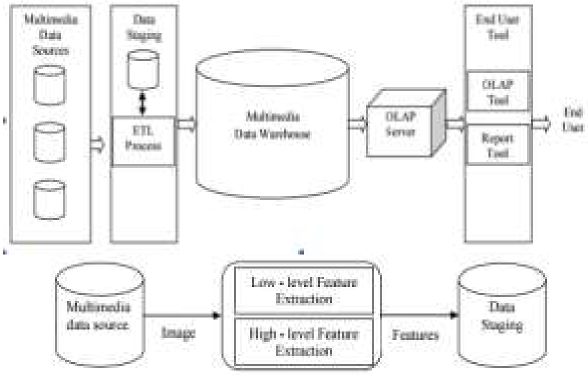

Figure 1: Proposed architecture

4. Conclusion

Our paper has proposed an organized approach to model an architectural framework for image data warehouse in a generalized manner. Compression and partitioning techniques are applied to improve the storage efficiency. Representing image by multilevel features, applying indexing techniques, and partitioning techniques improve access and analysis efficiency. Information retrieval from image ware house is a time consuming process and most of the time, the results obtained don't fulfill the expectations and requirement of the user. This paper has tried to address the concern by introducing the concept of Content Server along with Image feature extraction. The Content Server will maintain the indexes for metadata of the stored image content as well as will have a repository of image content which will not only speed up the multimedia information retrieval but will also make it more specific, generating results faster which are as per the requirement of user. A possible extension to this could be to develop a generalized scheme for storing the metadata as well as to develop query language to support free text query, query by description specified by various standards like MPEG -7.

5. Second and Following Pages

The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5 x 11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

References

[1] A. M. Arigon, M. Miquel, A. Tchounikine, Multimedia data warehouses: a multiversion model and a Medical application, Multimedia Tools and Applications, vol. 35, 2007.

[2] H. Mahboubi, J.C. Ralaivao, S. Loudcher, O. Boussaid, F. Bentayeb, J. Darmont, X-WACoDa: An XML-based approach for Warehousing and Analyzing Complex Data, Advances in Data Warehousing and Mining, IGI Publishing, 2009.

[3] Hamid R. Nemati, Sherrie Cannoy, Robert Delk "Data Warehousing and Web Enablement Opportunities, Issues, and Trends".

[4] Andrei Vanea, "A Hierarchical Semantically Enhanced Multimedia Data Warehouse",978-1-4244-8230- 6/10 © 2010 IEEE.

[5] Anne-Muriel Arigon, Anne Tchounikine and Maryvonne Miquel, "Handling Multiple Points of View in a Multimedia Data Warehouse", ACM Transactions on Multimedia Computing, Communication and Applications, Vol.2, No.3, August 2006, Page 199-218.

[6] H. Mahboubi, J.C. Ralaivao, S. Loudcher, O. Boussaid, F. Bentayeb, J. Darmont, "X-WACoDa: An XML-based approach for Warehousing and Analyzing Complex Data", Advances in Data Warehousing and Mining, IGI Publishing, 2009.

[7] J. You, Q. Li, On hierarchical content-based image retrieval by dynamic indexing and guided search, Proceedings of the 8th IEEE International Conference on Cognitive Informatics, 2009.

[8] Meenakshi Srivastava, Dr. S. K. Singh, Dr. S. Q. Abbas, "An Architecture for Creation of Multimedia Data Warehouse" IJESIT, volume 2, issue 4, pp. 309-315, July 2013.

[9] Mohd. Fraz, Ajay Indian, Hina Saxena, Saurabh Verma, "Improving Compression Efficiency of Data Warehouse," International Journal of Scientific & Engineering Research (IJSER), pp. 1575-1578 Volume 4, Issue7, Jul 2013.

[10] Stephen T C Wong, Kent Soo Hoo Jr, Robert C Knowlton, Kenneth D Laxer, Xinhau Cao, Randall A Hawkins, William P Dillon, Ronald L Arenson,"Design and Applications of a Multimodality Image Data Warehouse Framework," Journal of the American Medical Informatics Association Volume 9 No. 3, pp. 239-254, May / Jun 2002.