ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 4, July 2013

An Architecture for Creation of Multimedia Data Warehouse

[1]Meenakshi Srivastava, [2]Dr. S.K.Singh, [3]Dr. S.Q.Abbas
[1]Assistant Professor, Amity University ,Lucknow Campus, India, [2] Professor, Amity University Lucknow Campus, India, [3]Director, Professor, Ambalika Institute of Management and Technology,Lucknow, India

Abstract— Increasing trend in storage of multimedia information on worldwide web servers has resulted in growing tendency of searching and capturing data on Internet. Interest of scientific, academic, industrial and business communities has been growing many folds for searching multimedia information. Reciprocating to this trend various Organizations are willing to make their data available on www. Organizations store multimedia data in a multimedia data warehouse. Complexity in organizations storage methodology and the quantity of information stored cause delay in retrieval of information. Also often the information desired and the search result received has a major disconnect. Author proposes architecture for storing multimedia content in a format which facilitates retrieval of data in an efficient, accurate and fast manner. This architecture has developed a multimedia warehouse, in which the concept of Content Server has been incorporated. Content Server will provide a Service-centric architecture and storage technology-centric architecture.

***Index Terms*—Content Server, Multimedia Data Ware House, OLAP Engine, Knowledge Server**

I. INTRODUCTION

Data warehouses are dedicated to collecting heterogeneous and distributed data in order to perform decision analysis. Based on multidimensional model, OLAP commercial environments such as they are currently designed in traditional applications are used to provide means for the analysis of facts that are depicted by numeric data (e.g., sales depicted by amount or quantity sold). However, in numerous fields, like in medical or bioinformatics, multimedia data are used as valuable information in the decisional process. One of the problems when integrating multimedia data in a warehouse is to deal with dimensions built on descriptors that can be obtained by various computation modes on raw multimedia data and the speed of information retrieval. Taking into account that computation modes makes possible the characterization of the data by various points of view depending on the user's profile, his best-practices, his level of expertise, and so on.

Extracting information of text files, sounds and images comprising a multimedia object is comparatively easier for Human beings, but step by step extraction of information by automated process requires deployment of sufficient techniques pertaining to varied types of multimedia objects available from which the information is extracted i.e. availability of large amount of multimedia data requires deployment of system for information extraction. As manually it is not possible to search_and extract entire information of a multimedia data. To overcome these constraints processes used for information extraction are systematic and automated or semi automated as per requirement. Although data warehouse technology for numerical and symbolic data is considered to be mature [1], there is much to do in regard to complex, multimedia data warehousing [2]. Defense research, Geological research, Weather forecast, Marine research are areas which have 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, use of Data warehouses is most common to cater to above mentioned requirement. DBMS systems are traditional and are not designed to fit in the requirement of handling complex multimedia data, since relational databases store structured data only on the contrary multimedia data is often semi structured, hence deployment of new techniques to store, retrieve and process the multimedia data is essential and imperative.

"Analyst project that 60% of data warehouse projects over the next year will incorporate Web-enabled technology *(King).* In addition, The Data Warehouse Institute (TDWI) predicted that the market for data warehousing hardware, software and services would grow from $16.9 billion in 1996 to $40.5 billion in 2001 *(Ubois).* Industry data warehouse experts also believe that the Web is critical to a successful data warehouse implementation. Ralph

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 4, July 2013

Kimball believes that data warehouse results should be published everywhere, preferably over the Internet *(Kimball)."*[3].

At the beginning, there were mostly numerical and textual data warehouses, primarily in business operations such as economics, marketing and sales. The large amount of data generated by these businesses was successfully integrated using the dimensional model. The relational model, used for regular storage in OLTP databases, was not suited for the decisions that had to be pulled from the existing data. This is because it not optimized for aggregating such large amounts of data, which are stored in data warehouses. The fact tables store summarized, aggregated data, which is later used by OLAP tools to aid in the decision making process[4]. The basic concept of a data warehouse is originated from the traditional business information systems, and is extended to multimedia information retrieval. In this paper, we propose architecture for multimedia data warehouse to facilitate fast and effective searching for multimedia information retrieval. Beyond the difficulties encountered in extracting and modeling multimedia data, lays the difficulty in manipulating, interpreting the data and the need for transforming raw data into useful information. This expertise, which is user-dependent as well as analysis-dependent, is fully part of the decisional process [5].

II. RELATED WORK

In the last few years lots of work has been done in the area of multimedia information retrieval, and data warehouse but the field of Multimedia data warehouse still need a lot of attention, as the multimedia information is being used in almost in every field whether economics, medicines, education there still exist a requirement for a proper solution for storage and retrieval of multimedia information. In [6] the authors describe a data warehouse for complex objects. The semi-structure format of these objects is captured via XML files, which are then parsed and validated against a minimum requirements pattern. The communication between the user and the data warehouse is accomplished using the xQuery language, a language for XML-like structured data [5]. In [1] the author has mentioned the two types retrieval methods from a multimedia database, the first one is by content and the second is via description. Description based retrieval uses attribute descriptions of data (color, audio/video duration, number of instances a particular word is used), while content based retrieval uses the actual data inside the files (clouds, ideas, theories). When dealing with multimedia, it is helpful to separate the types of media files during data retrieval or processing [4]. In [7] a hierarchical method for storing these files has been described. In order to enhance the data ware house semantically, a system in which medical related data has been extracted from the database and analyzed is implemented in [8]. The traditional multidimensional models have a static structure where members of dimensions are computed in a unique way. However, multimedia data is often characterized by descriptors that can be obtained by various computation modes. [9] define these computation modes as "functional versions" of the descriptors, a Functional Multiversion Multidimensional Model is proposed by integrating the concept of "version of dimension". This concept defines dimensions with members computed according to various functional versions. This new approach integrates a choice of computation modes of these members into the model, in order to allow the user to choose the best representation of data.

III. HOW TO CREATE A MULTIMEDIA WAREHOUSE

Data warehouse can be viewed as a technology which not only functions as a data superstore, but also processes data to create a data warehouse, operational data store, or data mart stored on traditional servers, Intranet servers, or Internet servers. In other words, data warehouses are not just large databases; they are large, complex environments that integrate many technologies. A major gain of using a data warehouse consists of being able to store data at different levels of granularity along different dimensions such as data type, time, etc [10]. A data warehouse is a single site repository of information collected from multiple sources. Information in the data warehouse is organized around major subjects and is modeled so as to allow pre-computation and fast access to summarized data. "OLAP" as defined in warehouse [11,12] refers to analysis techniques used to explore the data[9]. Specific extraction processes, specification of metadata on the basis of content and description, indexing schemes, adapted tools for visualization and definition of specific multimedia aggregation function are required when multimedia data (text, images, videos, audios) are stored in warehouse e. The process of creation of a multimedia data warehouse can be defined in four phases. First Phase: Extract, Transform, Load. This process involves extracting data from various resources, then transforming it to quality levels (operational need), and finally loading them to the warehouse. The Extraction process is basically parsing of data, which is checked for expected pattern. The Transformation process deal with rules which are applied over the extracted data before loading them into the warehouse, like ignoring null columns. The job of Load process is to finally load the data in

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 4, July 2013

to the warehouse according to the requirement of the organization. In case of multimedia data ware houses its XML format. Various ETL tools are available now days. The fastest ETL record is currently held by Syncsort[13]. The second phase is The Multimedia Data Ware House. A Data Warehouse is a "subject-oriented, integrated, non-volatile and time-variant collection of data in support of management's decisions" [14]. Information in the data warehouse is organized around major subjects and is modeled so as to allow pre-computation and fast access to summarized data. "OLAP" as defined in warehouse [11, 12] refers to analysis techniques used to explore the data, however, this can restrict the analysis of data, particularly in the case of multimedia data. Indeed, multimedia data are huge data, with large information stored in them (text, graph, video, sound). These data are generally described by content-based or description-based descriptors [9]. For storing these data about data i.e. the meta data, a separate sub-subsystem is implemented by various researchers in various ways[4],[5],[9],[14],[15],[16]. The third phase : OLAP Server. The OLAP server models the data in a multidimensional way and precalculates aggregates in order to optimize queries. In the data warehouse, the measures are stored in a fact table. The dimension data can be stored in tables following a “star model,” a “snowflake model,” or a “galaxy model.” The “star schema” models the dimension data as a unique table, in which the hierarchical relationship of a dimension is not explicit but is rather encapsulated in its attributes. The “snowflake schema" normalizes dimension tables, and makes it possible to explicitly represent the hierarchies by separately identifying a dimension in its various granularities. When multiple fact tables are required, the “galaxy model," or “fact constellation" models allow the design of a collection of star schemas [5]. The fourth phase is Queries & Reports. This phase basically deals with End User Software, which provides interface for query, analysis tools and reporting tools.

IV. EXISTING METHODOLOGY

The existing methodology for retrieving multimedia information from a warehouse includes direct extraction of data from warehouse. For creation of fact table [4], data cubes [5], multidimensional data cube [10], mapping between different type of data which are semantically same, summarization [10] or designing a schema (ex. Straflake schema) for representation of multimedia data [5] all such activity interacts directly with the warehouse for required data directly from the ware house, which increases the time involved in above mentioned activities. Like in [4] each fact level of each data mart is stored in a different XML file; this creates hierarchical structure of the data marts, in which a fact table may become 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 used in the aggregation process. To speed up the process they have developed a procedure to automatically creating new queries, an XML file is created, 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. Here we argue that by enabling proper and generalized metadata for, metadata indexing on the warehouse itself and introducing the concept of Content Server in the architecture will provide a faster mechanism for retrieval of multimedia data. The proposed architecture is described in next section.

V. PROPOSED ARCHITECTURE FOR CREATING MULTIMEDIA DATA WARE HOUSE

We have structured our architecture in six subsystem (Fig 1), the ETL subsystem, the meta data extractor and manager sub system, the data ware house sub system, metadata indexer sub system, the OLAP engine and finally the front end tools sub system.

A. The ETL Tools Sub System

The ETL tools extract the data and transform them in to the structure required by the organization. The features such as name, file length, format, author, date etc are extracted and stored in XML file.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 4, July 2013

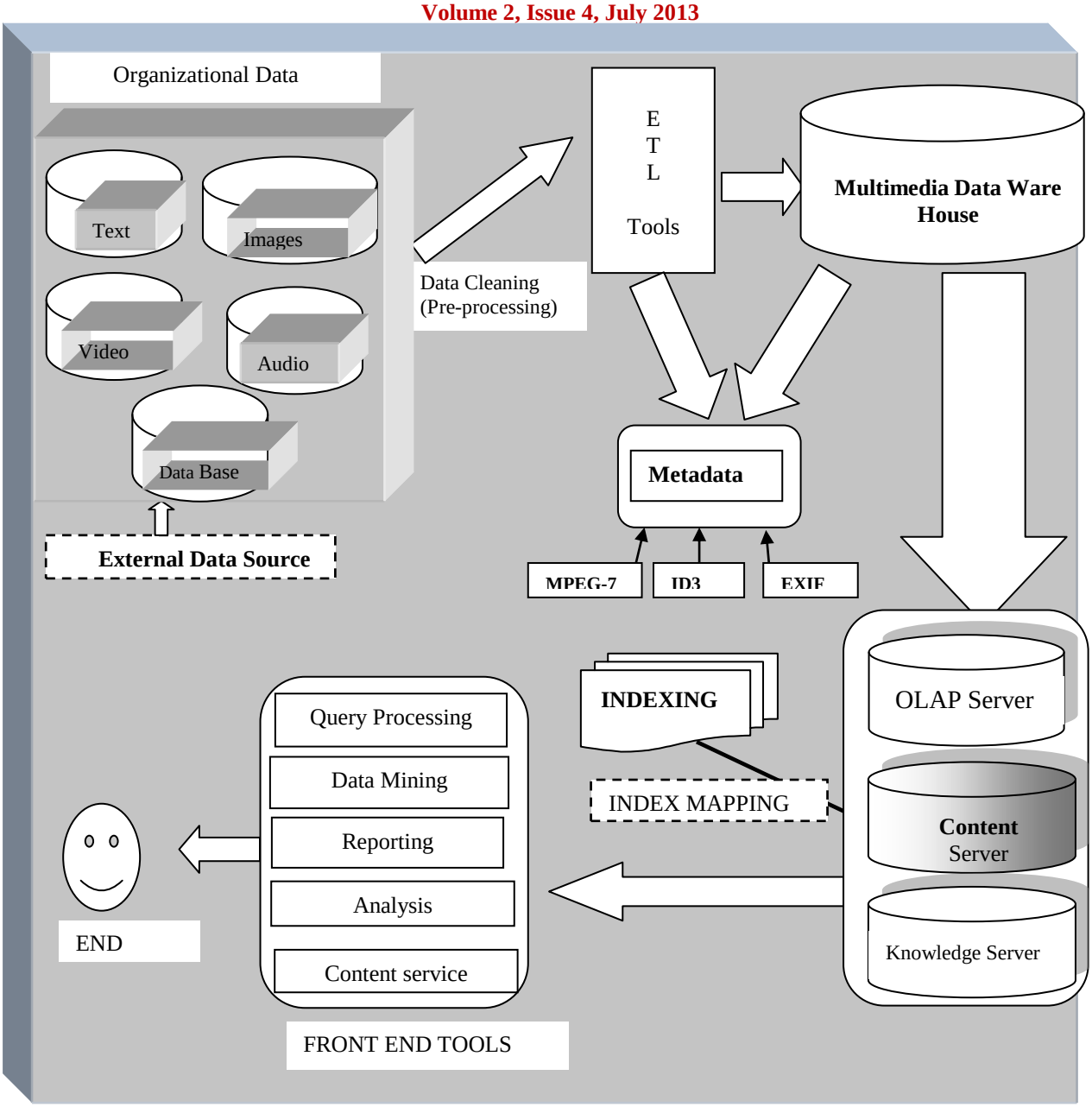

Fig 1 Architecture of multimedia data ware house

B. The Meta data extractor and manager Sub system

The metadata extractors sub system analyzes a given multimedia data to extract the desired metadata. As Multimedia data are found in many forms, like text, image, audio, video, their related metadata vary and so is the extractor. Each metadata extractor supports a specified file format and has its own unique functionalities. We have used MPEG-7 standard for storing images, audio, and video but various other standards like ID3 for music, EXIF standard for JPEG file format can also be used. The interface for any new metadata extractor can also be used. MPEG -7 uses descriptors (Ds), description schemes (DSs) and a description definition language (DDL). The metadata manager subsystem uses aggregation tools which operate on the different types of media supported by the data warehouse. The metadata manager allows editing the metadata repositories and also the mappings between the various files formats which belong to the same domain metadata.

C. The Multimedia data ware house subsystem

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 4, July 2013

The traditional multidimensional structure of multimedia data ware house is enhanced by adding the concept of OLAP Engine. The data are stored in the warehouse in XML file format, and once the indexing is done over the stored data no further direct interaction will occur with those files. The physical records will be retrieved by comparing their indexes with the domain specific indexes stored in Content Server . We have designed a multimedia ware house for education system; the modeling of multimedia database is done on the basis of descriptors that define the data. The user may select a subject, then a specific topic and further a sub-topic of that subject. Fig. 2. Represents domain knowledge stored in Content Server in metadata format.

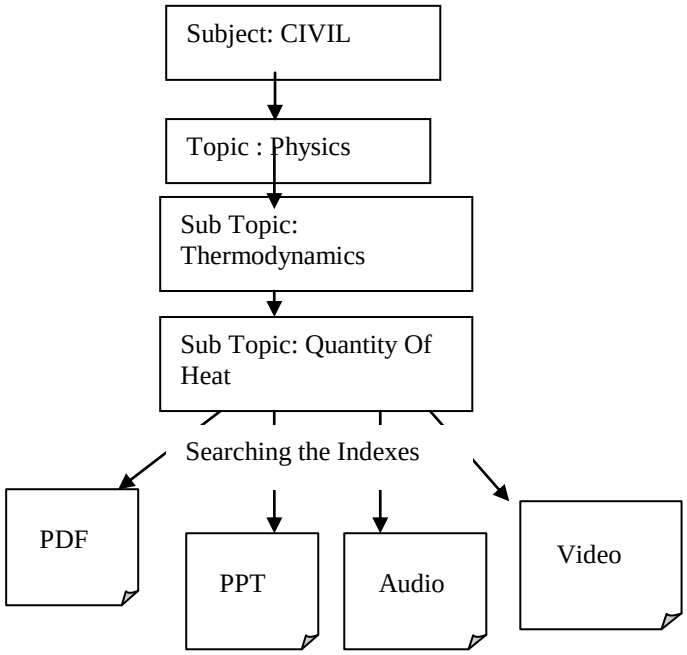

Fig 2. Representation of Domain Knowledge in Content Server

D. The Meta data indexer Subsystem

The metadata indexer will maintain a dynamic indexing schema, on the extracted metadata from the various multimedia sources. The metadata indexer will map the indices to the Content Server (described in next section). The Content server stores the domain specific multimedia metadata repository. The indices will be mapped to the indexes which are created over the domain specific. By this mapping the retrieval process will be faster.

E. The OLAP Engine

OLAP Engine is a combination of OLAP server, Content Server and Knowledge Server. OLAP server is used for analytical operations over the ware house.

Content Server serves content to front end tools which in turn provide friendly interfaces 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 proposed Content Server consists of services to store domain specific metadata, i.e. *metadata repository*, and a domain metadata manager which will provide interface and tools to represent the metadata about a specific domain. The Content Server manages both the retrieval of multimedia data and their metadata internally. A Multimedia Warehouse may contain multiple Content Servers as necessary. Fig 3 depicts the functional diagram of content Server.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 4, July 2013

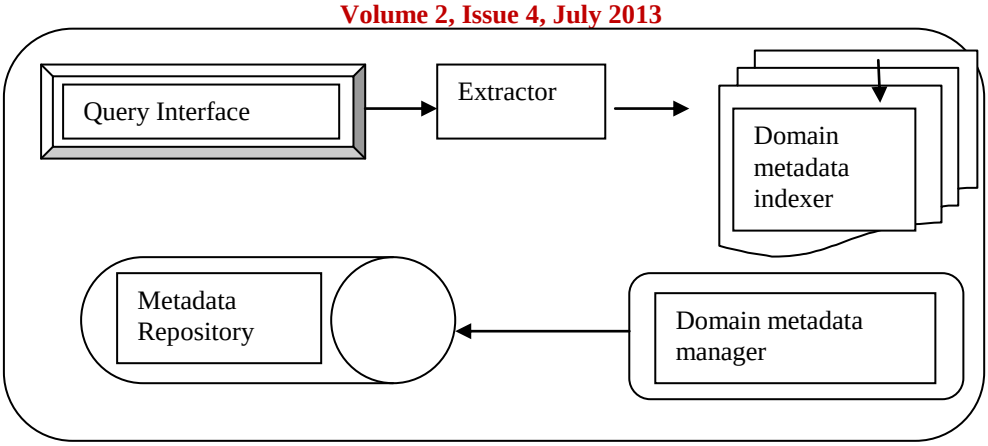

Fig 3. Components of Content Server

The third component of OLAP Engine is Knowledge Server which will provide the intelligence to the multimedia warehouse [17]. The Knowledge Server stores all the results of SQL, OLAP and DM. The purpose is instead of extracting the results from warehouse databases by using an extraction tool (Data Mining, SQL, or OLAP), time is saved by searching the stored results of previous analysis to check if the desired analysis is extracted and stored previously. We have applied this concept on the results which has been got by Content Server. It gives another level of refinement to the search.

VI. IMPLEMENTATION

Implementation of the Content Server can be done as a custom application deployed on an Application Server to provide the above mentioned services as well as JAVA standard Server APIs can also be used to implement it.

VII. CONCLUSION AND FUTURE WORK

As has been discussed at length that information retrieval from Multimedia ware house is time consuming process and the results obtained are many a times not in line with the expectations and requirement of the user. This paper has made an effort to adequately address the concern by introducing the concept of Content Server. The proposed Content Server will maintain the indexes for metadata of the stored multimedia content as well as will have a repository of multimedia 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.

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.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 4, July 2013

[8] M. L. Antonie, O. R. Zaiane, A. Coman, *Application of Data Mining Techniques for Medical Image Classification*, Proceedings of the Second International Workshop on Multimedia Data Mining, 2001.

[9] Anne-Muriel Arigon, Anne Tchounikine, Maryvonne Miquel, "A multiversion model for multimedia data warehouse", MDM/KDD 2005 Chicago, August 21, Chicago, Illinois, USA © 2005 ACM-MDM 2005-1-59593-216.

[10] Jane You, Tharam Dillon, James Liu, Edwige Pissaloux, "On Hierarchical Multimedia Information Retrieval", 0-7803-6725-l/01 ©2001 IEEE.

[11] Chaudhuri S., Dayal U. An Overview of Data Warehousing and Olap Technology. In SIGMOD Record 26(1).

[12] Vassiliadis P.: Modeling multidimensional databases, *cubes and cube operations* Proc. of 10th SSDBM 1998, Capri.

[13] Extract, transform, load- Wikipedia the free encyclopedia.

[14] Hurtado, C., Mendelzon, A.O. and Vaisman, A.: Updating OLAP Dimensions. Proceedings of the ACM Second Int. Workshop on Data Warehousing and OLAP, USA, 1999.

[15] Jane You, Tharam Dillon, James Liu " An Integration of Data Mining And Data Warehousing For ierarchical Multimedia Information Retrieval, Proceddings of 2001 International Symposium on Intelligent Multimedia, Video and speech Processing, May 2-4 2001 Hong kong.

[16] Maija Koivusaari, Jaakko Sauvola and Matti Pietikainen, "Automated document content characterization for a multimedia document retrieval system.

[17] Ala'a H. AL-Hamami, Soukaena Hssan Hashem "An Approach for Facilating Knowledge Data Warehouse", International Journal of Soft Computing Applications ISSN: 1453-2277 Issue 4 (2009), pp.35-40 © Euro Journals Publishing, Inc. 2009.