An Exhaustive Study on Data Mining Techniques in Mining of Multimedia Database

Pramod Kumar Yadav
Dept. of Computer Science & Engineering
Krishna Institute of Engineering & Technology, Ghaziabad.
pramodyadavster@gmail.com

SAM Rizvi
Dept. of Computer Science
Jamia Millia Islamia University, New Delhi
samsam_rizvi@yahoo.com

Abstract - **Multimedia database can be define as a collection of storage and retrieval systems, in which large amount of media objects are created, modified, searched and retrieved ,where as Multimedia is the combination of text, image, graphics, animations, audio and video. The extension of database application to handle multimedia objects requires synchronization of multiple media data streams. Multimedia data mining refers to the extraction of implicit knowledge, data relationships, or other patterns which are not stored in multimedia files explicitly. The system's overall performance in retrieval can be increase by indexing and classification of multimedia data with efficient information fusion of the different modalities is mandatory. Apart from text retrieval, the current waves in web searching and multimedia data retrieval are the search for and delivery of 3D scenes, images, music and video. The content-based multimedia information retrieval provides new techniques and methods for searching various multimedia databases over the world.**

Keywords: Query Optimization, Spatial data mining, web mining, Text Mining, Image Mining, Audio and Video Mining, Map Reduce.

I. INTRODUCTION

The search result performance of the given query is predicted by Query difficulty estimation. It is a powerful tool which is used for multimedia retrieval and now it is becoming more popular. There are several techniques proposed to estimate the query difficulty in the textual information retrieval, but directly they cannot be apply for image search, since it will result in poor performance. Existing research on query difficulty estimation focuses on the text-based queries, while the difficulty of image and video retrieval related to multimedia queries has not been yet studied so far. In current years, the prevalence of social media systems, e.g., Flicker, Face book, and YouTube, has largely increase Internet's multi- media database. These enrich database triggers may leads to the growth of large number of multimedia research scenarios .The success of these social media system also benefits the Content-based image retrieval [1]. Various content-based multimedia retrieval methods have been introduced by a large number of researchers. Beyond the methods for content-based image retrieval, audio retrieval and video retrieval, there also have been a wide variety of innovative content-based retrieval methods for new media types, such as content-based retrieval of 3D model, culture artefacts, motion data, biological data, etc. [9]. Data mining can be defined as a method of extracting or "mining" knowledge from large amount of data. It refers to the process of discovering interesting knowledge from huge amounts of data stored in databases, data warehouses, or other information repositories [21]. A multimedia data mining is the process that includes the construction of a multimedia data cube which facilitates multiple dimensional analyses of multimedia data, primarily based on visual content, and the mining of multiple kinds of knowledge, including summarization, comparison, classification, association, and clustering [9]. Thus, in multimedia system the knowledge discovery deals with non-structured information. In order to improve the results of the multimedia files, a database must be first pre-processed, followed by feature extraction. The significant patterns may be discovered with the help of generated features, using various data mining techniques [10].

II. MINING OF MULTIMEDIA DATABASE

Multimedia data mining is the method of discovering, interesting patterns from multimedia databases that store and manage large collections of multimedia objects, such as video data, audio data, image data, sequence data, and hypertext data which includes text, text markups, and linkages [21]. Some of the important issues in multimedia data mining include similarity search, content based retrieval, and multidimensional analysis. In mining of multimedia process, data collection is the first step of a learning system, as the overall achievable performance depends upon the quality of raw data. The next steps are the data pre-processing which is responsible to discover important features from raw data. It includes data cleaning, normalization, transformation, feature selection, etc. Learning largely depends on the informative features identified at pre-processing stage. The output of data pre-processing is the training set. Thus a training set has

978-1-4799-2900-9/14/$31.00 ©2014 IEEE

to select a learning model to learn from it and make multimedia mining model more iterative [2].

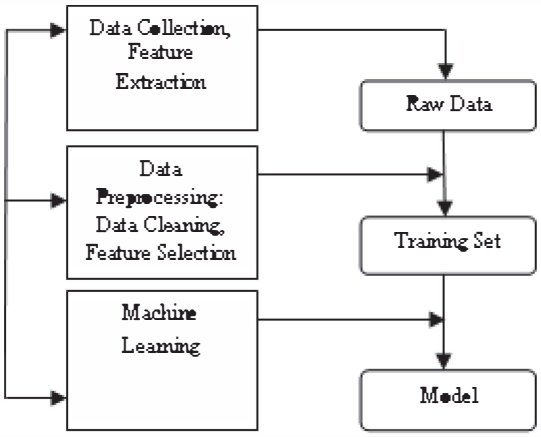

Figure1: Multimedia Mining Process

Data collection is the very first step in multimedia mining process. It acts as a raw data which are further input to the data preprocessing stage, which includes several task such as data cleaning and feature selection. After data preprocessing, a training set is obtained, which are further refine by applying machine learning to obtain a realistic model.

III. SPATIAL DATA MINING

Spatial Data Mining refers to the extraction of knowledge, spatial relationship, or other interesting patterns not explicitly stored in spatial databases. Such type of mining requires an integration of data mining with spatial databases technologies. In spatial database huge amount of space-related data, such as preprocessed remote sensing or medical imaging data, maps, and VLSI chip layout are stored. The main task of spatial databases is to discover the interesting patterns from large geospatial databases. Spatial data cubes are those which contain spatial dimensions and measures. Spatial online analytical processing can be implemented to facilitate multidimensional spatial data analysis. Spatial data mining includes clustering, classification and spatial trend and outlier analysis, mining spatial association and co-location patterns. Some of the important challenging issues regarding the construction and utilization of spatial data warehouses include the integration of spatial data from heterogeneous source and systems. Since these spatial data are stored in different industry firms and government agencies using various data formats, which are not only structure specific but also vendor specific. A lot of work has to be done on the integration and exchange of heterogeneous spatial data, which lay down the path for construction of spatial data integration and spatial data warehouse. Another important challenge is the realization of fast and flexible on –line analytical processing in spatial data warehouses, for which the star schema model is a good choice, since it provides a concise and organized warehouse structure and facilitates OLAP operations [21].

IV. WEB MINING

The Internet has now become a tool to support the planning of activities that require information gathering and reasoning. A major activity on the Internet is the retrieval and browsing of multimedia information [12]. It consist of a huge, widely, distributed, advertisements, consumer information, education, government, e-commerce, and many other services. The main task of web mining includes mining of web contents, web access patterns and web linkage structures. This involves mining the web page layout structure, mining the web's link structures to identify authorize web pages, mining multimedia data on the web, automatic classification of web documents, and web usage mining[2].

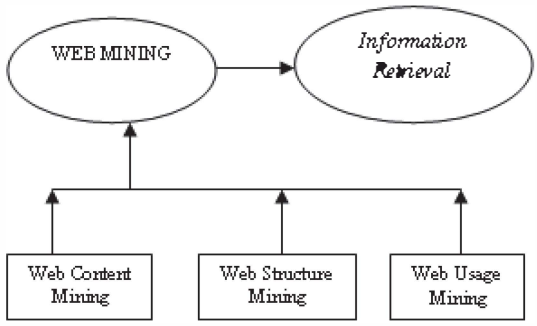

Figure 2: Process of Web Mining.

A. Web Content Mining: The process of extracting useful information from the contents of web documents is called web content mining. It may consist of text, images, audio, video information which is used to convey to the users about that documents [2].

B. Web Structure Mining: The process of discovering structure information from the Web is known as web structure mining. It consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages [2].

C. Web Usage Mining: It is the method of discovering interesting usage patterns from web data, in order to understand and better serve the needs of web-based applications [2].

However the web also poses great challenges for effective resource and knowledge discovery, some of them are as follows:

i. The web seems to be too huge for effective data warehousing and data mining.

ii. The complexity of web pages is far greater than that of any traditional text document collection.
iii. The web is a highly dynamic information source.
iv. The web serves a broad diversity of user communities.
v. Only a small portion of the information on the web is truly relevant or useful.

These challenges have promoted research into efficient and effective discovery and uses of resources on the internet.There are many search index –based web search engines. These search the web, indexs web pages, and build and store huge keyword-based indices that help locate sets of web pages containing certain keywords [21].

V. TEXT MINING

Text mining is the process of finding useful or interesting patterns, models, directions, trends, or rule from unstructured text [2]. Latent Semantic Indexing is the latest and efficient approach to access databases that contain text. LSI associates with each document a limited-size vector vec (*d*) that contains frequency terms. Thus, the storage of documents in database becomes equivalent with storing the associated vectors. The basic idea of the technique is that similar documents have similar frequencies of words. The technique allows both the elimination of words and phrases that do not allow distinguishing between various documents and the identification of the ones that do so. It can identify also the similar words [3]. In text mining there are two open problems: Polysemy, synonymy. Polysemy means that a single word can have multiple meanings where as Synonymy means that multiple words can have the same/similar meaning [2].

VI. VI. IMAGE MINING

Image mining is the technique to detect unusual patterns and extract implicit and useful data from images which are stored in the large databases. It deals with making associations between different images from large image databases. The applications areas of Image mining are medical diagnosis, remote sensing, agriculture, industries and space research and also handling hyper spectral images. Images database include maps, geological structures, and biological structures. The primary challenge in image mining is to reveal out how low-level pixel representation enclosed in a raw image or image sequence can be processed to recognize high-level image objects and relationships[2] .In today's databases, images are represented as relationships, or as spatial data structures or with help of image transformations[3]. Existing query difficulty estimation methods play an important role in the text based information retrieval. However, it cannot be directly applied to content-based image retrieval, due to the complex structure of image queries and the well-known semantic gap [1].

Image mining performs several operations before generating Knowledge. Firstly the images stored in the databases are pre-process and then they undergo transformation and feature extraction. After this the mining is performed on the collected data and then after Interpretation and evaluation the Knowledge is generated.

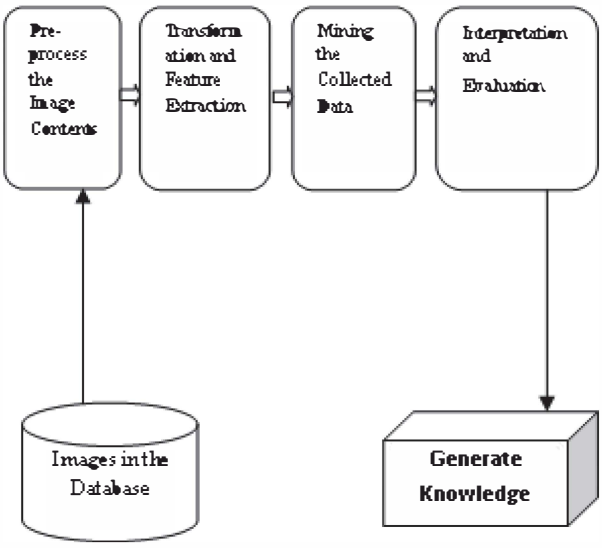

Figure 3: Process of Image Mining.

VII. VIDEO MINING

A video can be defined as scene sequences, which are composed by shots that are frame sequences. A frame is a static image, where as a shot is a clip that presents a continuous action in time and space .Video data mining is more typical than mining of image data. It is a collection of moving images, much like animation. Developing query and retrieval techniques for video databases are an important area which includes optimization strategies, video indexing, and query languages. The produced video (movies), the raw video (traffic video) and the medical video (ultra sound videos) are the three types of video [3]. Conventional techniques for video indexing are based on characterizing the video content using a set of computational features. These techniques often do not exploit the commonly available high-level information conveyed by textual captions. Relevant videos retrieval from large video databases has vast applications in various domains. Retrieval from broadcast news video database is vital for effective information access. Video mining techniques can also be used for mining digital libraries of video clips of lectures for distance learning applications. The two important issues in Video mining are to develop a representational scheme for the content and a Human friendly query/interface [15].

VIII. AUDIO MINING

Since audio data is similar to video, the techniques for information processing and mining of audio data are also similar to video information retrieval and mining. Mining of audio data requires conversion of audio data into text by using speech transcription techniques. By using audio information processing techniques audio data can also be mined directly [2]. As for images or videos, audio data can be characterized in two ways: by using metadata to explain the content (objects or activities) or by extracting specific features with signal processing techniques (for instance, frequency, amplitude, vibratory period etc.). For audio data mining the most common content-based indexing technique is to segment the signal in time, to get small windows in which it can be considered homogeneous (amplitude, speed and wave length are constant). The segmentation is made with one constant step (window length) or by using a homogeneity predicate [3].

IX. CONCLUSION AND FUTURE WORKS

The multimedia data mining, knowledge extraction plays a vital role in multimedia knowledge discovery. The major challenges in multimedia mining are lost of knowledge when the sequence of multimedia is ignored. Polysemy and Synonymy are the two open problems in text mining. Polysemy means that a word can have multiple meanings whereas Synonymy means that multiple words can have the similar meaning. The combination of different types of image data is a great challenge in Image mining. In audio and video mining, the combination of information across multiple media fundamental open problem [2].The collections of multimedia objects are stored in digitized forms in content-based retrieval systems. Content-based retrieval queries are characterized by fuzziness and ambiguity. The two kinds of errors encountered during content-based retrieval are the false results and the absent results. [3]. since multimedia data are unstructured by nature therefore mining the multimedia data is more difficult than that of traditional business. The main difficulty in mining of multimedia data is the *heterogeneous* nature of the multimedia data [23]. Huge amount of multimedia information are generated day by day in different domains. The need for efficient techniques for accessing the relevant information from these multimedia databases is on the rise. Content based information retrieval techniques are one of the methods to access minimum relevant information present in the multimedia database. Search and retrieval techniques have reached a maturity for textual data resulting in powerful search engines. Active research has been seen in the area of image retrieval and video retrieval during the last decade. Large video collections, which were thought of impossible at one stage, are becoming increasingly common due to the advancement of storage technologies [15]. Map Reduce is emerging as one of the important parallel data processing tool and is gaining significant momentum from both industry and academia as the volume of data to analyze grows rapidly [24].

X. REFERENCES

1. Yangxi Lian , Bo Geng , Linjun Yang , Chao Xu ,Wei Bian ,"Query difficulty estimation for image retrieval", *Elsevier Journal on Neurocomputing* , 2012, 48-53.
2. Pravin M. Kamde, Dr. Siddu. P. Algur, "A survey on web multimedia mining", *The International Journal of Multimedia & Its Applications (IJMA) Vol.3, No.3, 2011.*
3. Catalina Negotia ,Monica Vladoiu , "Querying and Information Retrieval in Multimedia Databases", *Journal of Seria Matematica- Informatica- Fizica,* Vol. LVIII, No. 2/2006, 73 -78.
4. Remco C. Veltkamp, "Multimedia Retrieval Algorithmics", *Springer-Verlag Berlin Heidelberg* 2007, 138–154.
5. Alfredo Cuzzocrea, Elisa Bertino, "Privacy Preserving OLAP over Distributed XML Data: A Theoretically-Sound Secure-Multiparty-Computation Approach", *Elsevier Journal on Computer and System Sciences 77 (2011), 965–987.*
6. David Rebollo-Monedero, Jordi Forne´ , and Josep Domingo-Ferrer, Fellow, IEEE , "Query Profile Obfuscation by Means of Optimal Query Exchange between Users" , IEEE *transactions on dependable and secure computing, VOL. 9, NO. 5, 2012, 641-654.*
7. Michael s. Lew, Nicu Sebe, Chabane Djeraba Lifl,Ramesh Jain , "Content-based Multimedia Information Retrieval: State of the Art and Challenges", *In ACM Transactions on Multimedia Computing, Communications, and Applications, 2006* .
8. Vagelis Hristidis , Shu-Ching Chen, Tao Li, Steven Luis, Yi Deng , "Survey of data management and analysis in disaster situations", *Elsevier Journal on Systems and Software* (2010) 1701–1714.
9. Yi Yang , Fei Wu,, Dong Xu , Yueting Zhuang , Liang-Tien Chia , "Cross-media retrieval using query dependent search methods", *Elsevier Journal on Pattern Recognition* (2010) 2927–2936.
10. Manuel Barrena, Elena Jurado, Pablo Márquez-Neila, Carlos Pachón, "A flexible framework to ease nearest neighbor search in multidimensional data spaces", *Elsevier Journal on Data & Knowledge Engineering* (2010) 116–136.
11. Samir Kumar Jalal, "Multimedia Database: Content and Structure", *Multimedia and Internet Technologies, 2001.*

12. Nadia Bianchi-Berthouze, Naoto Katsumi, Harutaka Yoneyama,Subhash Bhalla, Tomoko Izumita , "Supporting the Interaction between User and Web-based Multimedia Information", *Proceedings of the IEEE/WIC International Conference on Web Intelligence ,2003* IEEE.
13. Ira Assent, Andrea Wenning, Thomas Seidl , "Approximation Techniques for Indexing the Earth Mover's Distance in Multimedia Databases" , *Proceedings of the 22nd International Conference on Data Engineering ,2006 IEEE.*
14. Shu-Ching Chen, Mei-Ling Shyu, Na Zhao , "An Enhanced Query Model for Soccer Video Retrieval Using Temporal Relationships", *Proceedings of the 21st International Conference on Data Engineering ,2005 IEEE.*
15. C. V. Jawahar, Balakrishna Chennupati, Balamanohar Paluri, Nataraj Jammalamadaka, "Video Retrieval Based on Textual Queries", , *Proceedings of the International Conference on Data Engineering ,2006* .
16. Dong Wang , Zhikun Wang , Jianmin Li , Bo Zhang , Xirong Li , "Query representation by structured concept threads with application to interactive video retrieval" , *Elsevier Journal on Visual Communication and Image Representation (2009) 104–116*.
17. Ramesh Natarajan, Radu Sion, Thomas Phan, "A grid-based approach for enterprise-scale data mining", *Elsevier Journal on Future Generation Computer Systems* (2007) 48–54.
18. Gang Wu, Huxing Zhang, Meikang Qiu, Zhong Ming, Jiayin Li, Xiao Qin, "A decentralized approach for mining event correlations in distributed system monitoring", *Elsevier Journal on Parallel Distrib. Comput (2012) 330–340.*
19. Peter Paul Beran, Werner Mach, Erich Schikuta, Ralph Vigne," A Multi-Staged Blackboard Query Optimization Framework for World-Spanning Distributed Database Resources", *Proceedings of the International Conference on Computational Science, ICCS 2011 Elsevier Journal , 156–165.*
20. Xin Chen, Chengcui Zhang, and Wei-Bang Chen ,"A Multiple Instance Learning Framework for Incident Retrieval in Transportation Surveillance Video Databases", 2007 IEEE.
21. Han and Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann publishers-An Imprint of Elsevier, 2006.
22. Korth, Silberschatz, and Sudarshan, "Database system Concepts", Singapore, Mc Draw Hill, 2006.
23. Chidansh Bhatt and Mohan Kankanhalli, "Probabilistic Temporal Multimedia Data Mining", *ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 2, Article 17, 2011.*
24. Kyong-Ha Lee, Yoon-Joon Lee, Hyunsik Choi, Yon Dohn Chung, Bongki Moon, "Parallel Data Processing with Map Reduce: A Survey", *SIGMOD Record, December 2011 (Vol. 40, No. 4).*