A Survey on Association Rules in Case of Multimedia Data Mining

[1]Vishwadeepak Singh Baghela, [2]Dr. S. P Tripathi
[1]Uttrakhand Technical University, Dehradun, Uttrakhand, India
[2]HOD, IET, Lucknow, UP, India

Abstract
Data mining is the extraction of hidden predictive information from large database. Due to rapid application of computer and electronic devices and tremendous growth in computing power and storage capacity, there is explosive growth in data collection. Data mining was done primarily on numerical set of data, now days as large multimedia datasets such as text, audio , video, image and combination of several types are becoming , which is almost unstructured or semi structured data by nature, which makes it difficult to extract the information without powerful tools. Association is a powerful data analysis technique that appears frequently in data mining literature. The main advantages of association are simplicity, intuitiveness and freedom from model-best assumption. This survey paper aims at giving an overview to some of the previous researches done in this topic, evaluating the current status of the field and envisioning possible future trends in this area.

Keywords
Multimedia Data Mining, Association Rules, Text Mining, Unstructured Data

I. Introduction

Most international organization produces more information in a week than many people could read in a life time. The situation is even more alarming in world wide networks like the internet. Most organization has large database that contain wealth of potentially accessible information however it is usually very difficult to access this information.

A. What is Data Mining

Data mining seeks to extract hidden knowledge from large amount of data. Data mining is the process of extracting patterns from data. Data mining can be used to uncover patterns in the data but it is often carried out only on the samples of data. This mining process will be ineffective if the samples are not a good representation of the larger body of the data. Therefore an important part of the process is the verification and validation of patterns on others samples of data. In principle, the knowledge discovery process consists of six states –Data selection, Cleaning, Enrichment, Coding, Data Mining and Reporting. The fifth stage, data mining is the phase of real discovery [17].

B. Data Mining Method

Any techniques that helps extract more, out of your data is useful. So, data mining techniques from quite a heterogeneous group. All the various different techniques are used for different purpose are –query tools, statistical techniques, decision tree, neural network, Genetic algorithm association rules [17].
In principle, Data mining should be applicable to any kind of data repository, as well as to transient data, such as data streamed. Thus the scope of data mining will include relational data base, transactional database, advance database systems and the World Wide Web. Advance database system includes object relational databases and specific application-oriented database, such as spatial database, Time series databases and Multimedia Databases.

II. What is Multimedia Data Mining?

Multimedia Data refers to data such as text, image, video, audio, graphical, relational and categorical data [13]. Current data mining tools operate on structure data, the kind of data that reside in large relational databases whereas data in multimedia data bases are semi structured or unstructured. Often compared with data mining. Multimedia Mining has higher complexity.

A. Unstructured Data

Unstructured data is simply a bit string .Example include pixel level representation for image, video and audio, and character level representation for text [18]. Fig. 1, illustrates various aspect of multimedia data mining.

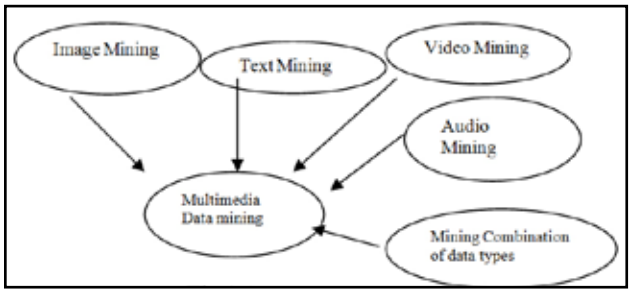

Fig. 1:

III. Survey of Related Works

A handful of multimedia data mining approaches are available for many potential information and association from large amount of multimedia data. A brief survey of some recent researches related to mining association from multimedia data is presented here. A generalized Affinity – based association rules Mining approach to discover the set of Quasi – equivalent Media objects from a network of databases have proposed by Shyu,Chen and Kashyap[2] The recent progress in high speed communication networks and large capacity storages devices has led to a tremendous increase in the number of data base and the volume of data in them. This has created a need to discover structural equivalence relationship from the databases since queries tend to access information from structurally equivalent media object residing in different database. The more databases, there are the more query processing. Performance improvement can be achieved when the structural equivalent relationship are automatically discovered. In response to such a demand association rule mining has emerge and proven to be highly successful techniques for discovering knowledge from large database. A generalized affinity –based association rule mining approach to discover the quasi equivalence relationship from a network of database are conducted in this paper. The result show that the proposed Algorithm not only correctly exploit the set of quasi equivalent media object from the database but also outperform the basic association rule mining approach in the discovery of the Quasi- equivalent media object pairs.

Choupo, Equille, Morin[4], have presented multimedia indexing and retrieval with features association rules mining methods that describes a new technique applied to very large still image databases that combines too data mining techniques clustering and association rules mining in order to better organized image collection and to improve the performance of querries.The objective of carried work is to exploit association rules discovered by mining global MPEG-7 features data and to adapt the query processing. K-Mean algorithm is used for initially determine several clusters of images. The interesting point is this method can applied to all kind of large imaged databases by exploiting association rules extracted from mining the clusters of global image features.

Yaik, Yong and Haron [5], have proposed a model to perform Time series prediction using adaptive association rule. This model uses the idea that if a segment of repeatable Time series pattern has occurred, it has the possibility that the following segment of the repeatable pattern will appear. Data mining and pattern matching techniques are being applied to mine for repeatable Time series patterns. This model has the ability to provide confident label for each prediction it may and perform continuous adaption. The result from the experiment shows that model is able to capture repetitive Time series patterns and perform prediction using those patterns. However this model has some drawback such as it require high computational power and required large storage.

Zhao, Chen and Rubin [6], have proposed a new approach that can automate system training by evaluating user feedback in real time. User feedback is widely deployed in recent multimedia research to refine retrieval performances. However most of the exiting online learning algorithm handles interaction of a single user, which may pose restrained performance due to limited size of positive feedback and alternative solution, is to learn general user perception via collecting feedback from different users. The training process is initiated only when the number of feedback reaches a certain threshold. This could improve the performances but it becomes a manual process to decide the threshold and initiate the training process. To address this challenge and advance training method by adopting the association rule mining technique, which can effectively evaluate feedback and automatically evoke the training process. The proposed approach utilize the HMMM mechanize couple with the ARM-based feedback evaluation to support both offline training and online learning and eliminates the need for manually initiating the training process.

Selvi and Tamilarasi [7], have proposed a new method for generation of class association rules and for selecting the best rules for classifying the new instance. For generation the class association rules, this method use the calculated minimum support instead of user specified minimum support. The new approach use the best run time minimum support for each item set generation and for rule generation. The first and foremost task in any associated classification algorithm is mining of the association rules. Many studies have shown that the minimum support measures play a key role in building and accurate classified. Without the knowledge of the items and their frequency, user specifies support measures are inappropriate. In this paper a new approach called DASApriori that is dynamic adaptive support Apriori to calculate the minimum support for mining class association rules and to build a simple and accurate classifier. The running time of the new algorithm will be improved if it adopts the FP-Growth method for the generation of association rules instead of the classic Apriori method.

Martinet and Satoh [8], present a study of the use of association rules in document representation. The proposed method is based on the discovery and the study of intra –model association rules between individual objects in the context of a spatial – temporal window. The discovered relation is used to build midlevel objects capturing the most frequently occurring patterns in the database. The approach showed in the experiment that note only obtained representation is more compact, but it also makes is possible to both decrease the density of documents space and increase the discriminating power of individual objects.

Somodevilla, Torres and Zecua [10], present a framework for the extraction of association rules from a spatial fuzzy data cube. First of all, spatial query are executated to filter the information to be to be loaded in the data warehouse considering the spatial relationship among data. Later on, using the mondrinian tools a classic data tool is created and a Fuzzy Data Cube {FDC} is generated from the first cube by selecting the linguistic variables, the fuzzy sets are defined and a threshold that allow us determine whether the transactional value belong to a fuzzy set or other.

Anwar and Naftal [11], described a video event modeling, detection and mining framework for multimedia surveillances. Event representations model provide a framework in which we can reason about events so as to interpret the collective behavior of objects overtime and space domains. This paper proposes a comprehensive event modeling framework for multimedia surveillances system. An event detection model in corporate multimedia strings and a new predicate set for describing more complex event scenarios.

Shah and Mahajan [12], have proposed a new efficient formulation for the generation of frequent item- sets that are used for computing association rules. It addresses the shortcoming of the serial Apriori algorithm and also its parallel formulation based on Count Distribution (CD). It improves on the time taken during every pass in the CD algorithm and helps improving the scalability. The scalability and efficiency of serial Apriori and CD algorithm is analyzed with respect to the proposed formulation. As the database size increases the efficiency of the CD algorithm decreases. Due to the large number of scans' through the entire database. This approach addresses this problem and suggests improvement so that better scalability and efficiency are achieved. The algorithm also has excellent scale –up properties with respect to the transaction size and the number of items in the dataset. The new formulation shows approx 15% performance improvement for generating frequent item sets over serial Apriori algorithm. The result also proves the scalability of algorithm to the dataset.

Manjunath , Hegadi and Ravi Kumar [13], presented a survey on multimedia data mining and its relevance today. This paper explores on survey of the current states of multimedia data mining and knowledge discovery, data mining efforts aimed at multimedia data, current approaches and well-known techniques for mining multimedia data. Multimedia refers to data such as text, Numeric, Images, Video, Audio, Graphical, Temporal, Relational and Categorical data. It is well-known that multimedia information is ubiquitous and often required in digital libraries, analysis of traffic video, medical image classification and analysis, media production and broadcasting and for World Wide Web. Following techniques can be applied for multimedia datamining – classification, clustering, Association and Statistical modeling. Among these classification, Association and Statistical modeling are supervised framework and clustering is unsupervised learning mythology.

Changchun and Li [14], introduce a framework for multimedia data mining system. They have analyzed the method which is associated with the model: Data cube, Clustering, Classification and association rule. Multimedia data mining is the combination

of data mining and multimedia database; it is not only emerging research direction but is also a challenging field of study.

Despande and Thakare [15], have been critically reviewed data mining system and application. The paper present that due to vast use of computers and electronics device and tremendous growth in computing power and storage capacity, there is explosive growth in data collection. To analyze this vast amount of data and fruitful conclusion it needs the special tools called data mining tools. The different methods of data mining are used to extracts the patterns and thus the knowledge from this varieties databases. Selection of Data and methods for Data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop generic data mining system but no system found completely generic. Thus for every domain the domain experts assistant is mandatory. The result yield from the domain specific application is more accurate and useful. This is very difficult to design and develop and a data mining system, which can work dynamically for any domain.

Yan and Hui [16], have proposed a new algorithm named as BOFP –V for frequent item set mining. The existing Apriori algorithm produce a lot of candidacy sets and needs scanning database many times but the proposed algorithm needs scanning database only once.

Hany Mahgoub[19], has presented a system for discovering association rules from collection of unstructured document called EART (Extract Association Rules from Text). The EART System has treated text only not images or figures. The main characteristic of EART is that the system has integrated XML technology (to transform unstructured documents into structured documents) with information retrieval scheme (TF-IDF) and data mining techniques for association rules extraction. EART is consisted of four phases: Structure Phase, Index Phase, Text Mining Phase and visualization phase. Experiments applied on a collection of scientific documents selected from MEDLINE that are related to the outbreak of H5N1 avian influenza virus.

VI. Conclusion

On the basis of survey carried out so far, it is found that in spite of the increasing effort in multimedia data mining during the last 10 years, extracting relevant and accurate knowledge from multimedia data sets remains a very difficult task. Almost all techniques are based upon Apriori algorithm to find frequent item set, however some of them are capable to improve performance but scalability is still a big problem. The problem with Apriori is that it generates too many two-item set that are not frequent. Some new techniques give better performance result in comparison with Apriori algorithm. It is also seen that no algorithm gives concentration on selection of items for making association rules. In real life, frequency and quantity of different items are different .There are no criteria for selecting parameters value i.e confidence, support and interestingness for association rules. Therefore to assign a fix parameters value for support and confidence to all items can not generate a good association rules. Almost all association rules deals only on a single attributes of items, but considering two or more than two attributes can generate a more powerful association rules for data mining.

References

[1] S.J.Simoff, O. R. Zaiane,"Multimedia Data Mining for the second time", The first International Work Shop on Multimedia Data mining, SIGKDD exploration ,Vol. 2, pp. 103-105, 2000.

[2] M.L. Shyu, S. C. Chen, R.L. Kashyap,"Generalized Affinity –Based Association Rule Mining for Multimedia Database Queries", Knowledge and Information System, Vol. 3, pp. 319-337, Springer, Verlag, London Ltd., 2001.

[3] S. J. Simoff, C. Djeraba, O. R. Zaiane,"Multimedia Data Mining Between Promises and Problems", Proceedings 3rd Int. Workshop MDM/KDD2002, Vol. 4, pp. 118-121, Canada, 2002.

[4] A.K. Choupo, L.B. Equille, A. Morin,"Multimedia indexing and Retrieval with features Association Rules Mining", Proceedings of International Conference on Multimedia and Expo(ICME'04),Vol. 2, pp. 1299-1302, IEEE 2004 .

[5] O.B. Yaik, C.H. Yong, F.Horon,"Time series prediction using Adaptive Association Rule", Proceeding of the 1st International conference on Distributed Framework For Multimedia Applications (DFMA'05) P.P 310-314, IEEE 2005.

[6] N. Zhao, S. Chen, S. H. Rubbin,"Automated Multimedia Systems Training Using Association Rule Mining", Proceedings of International Conference on Information Reuse & Integration (IRI'07), P.P 373-378, IEEE 2007

[7] C.S. Selvai, A.Tamilarasi, "Association Rule Mining With Dynamic Adaptive Support Thresholds for Associative Classification", Proceedings of International Conference on Computational Intelligence & multimedia Application (ICCIMA'07),Vol. 2, pp. 76-80, IEEE 2007.

[8] J. Martinet, S. Satoh,"A Study of Intra Model Association Rules For Visual Modality Representation", International Work Shop on Content Based Multimedia Indexing (CBMI'07), pp. 344-350, IEEE 2007

[9] Fayyad, Usama, G.P.Shapiro, P Smyth,"From Data mining to Knowledge Discovery in Databases", Fayyad. pdf, pp. 12-17, 2008

[10] M.J. Somodevilla I.H.P Torres, J.T Zecua," Framework for Discovering association rule in a fuzzy data cube", Proceedings of Mexican International conference on computer science (ENC'08), pp. 126-131, IEEE 2008.

[11] F. Anwar, A. Naftal,"Video Event Modeling and Associa-tion Rule Mining In Multimedia Surveillance System", VIE'08, pp. 426-431, IET Press, 2008.

[12] K.D. Shah, S.Mahajan,"A New Efficient Formulation for Frequent Item -Set Generation", Proceedings of International conference on Advances in Computing, Communication & Control (ICAC3'09), pp. 198-201, 2009.

[13] T.N. Manjunath, R.S. Hegdai, G.K Ravi Kumar,"A Survey on Multimedia Data Mining and Its Relevance today", IJCSNS, Vol. 10, pp. 165-169, November 2010.

[14] Y. Changchun, Y. Li,"A Data mining Model and Methods Based on Multimedia Database", Proceedings of International conference on Internet Technology & Application (ITA'10), pp. 1-4, IEEE 2010.

[15] S.P. Deshpande, V.M Thakare,"Datamining System and Applications: A Review", IJDPS, Vol. 1, pp. 32-44, September 2010.

[16] Z.H.Yan, Q.Hui,"Association Rule Mining With Establishment of Frequent Item Set Vectors", Proceedings of International Conference on Multimedia Information Networking And Security (MINES'10), pp. 696- 699, IEEE 2010.

[17] J. Han, M.Kamber,"Data Mining Concepts and Technique", Second edition, Morgan Kaufmann Publishers, pp. 1-40, 2008

[18] Sanjiv Purba,"Data Management Handbook", Published by CRC Press, 1999.
[19] Hany Mahgoub,"Mining Association rules from unstructured documents", Word Academy of Science, Engineering and Technology, Vol. 20, No. 1, pp. 1-6, 2006.

Mr. VDS Baghela received his graduation degree in statistics from BHU, Varanasi in 1999 and completed MCA Degree from AAIDU, Allahabad in 2004. He obtained M.Tech (CSE) degree from UPTU, Lucknow in 2010. Now he is Persuing Ph.D (CSE) under supervision of Dr. S.P. Tripathi. Currently he is working as an Associate Professor in IT Department at IIMT College of Engg, Greater Noida, UP, India. His research area is Data Mining.