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

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 04
February 2018

Search Aspects and Participant Major Datasets Using Cluster Mining

J. Malathi & Bikkina Lalitha Bhavani

Assistant Professor, Department of IT, Sir.C.R.Reddy College of Engineering , Eluru, AP.

Email: malathi.komma@gmail.com

[2]Assistant Professor, Department of IT, Sir.C.R.Reddy College of Engineering , Eluru, AP.

Email:karuturi.lalithacse@gmail.com

Abstract: *The data mining of association rules between items in a large database is an essential research aspect in the data mining fields. Discovering these associations is beneficial to the correct and appropriate decision made by decision-makers. Fast retrieval of the relevant information from databases has always been a significant issue. Clustering is a main task of exploratory data analysis and data mining applications. The selection of right and appropriate text mining technique helps to enhance the speed and decreases the time and effort required to extract valuable information. The evaluation of competitiveness always uses the customer opinions in terms of reviews, ratings and abundant source of information's from the web and other sources. We include platform and framework for managing and processing large data sets. We also discuss the knowledge discovery process, data mining, and various open source tools. . User generated text data is intrinsically noisy, with misspellings, informal language, and digressions. Because of the many variations in spelling and expression, the data is also very sparse. The Business Intelligence (BI) system is an effective and ancient data with systematic tools to present valuable and inexpensive information to business developers and decision makers. Many administrations have huge amounts of data in the method of formless text.*

Index Terms: Data Mining, Business Intelligence, Textual Data and Clustering, Knowledge discovery, Data mining tools, Competitor Mining, Firm analysis, Electronic commerce, Mining Association Rules, Large Item sets.

1. INTRODUCTION

Number of data mining algorithms is developed that greatly facilitate the processing and interpreting of large stores of data [1]. One example is the association rule mining algorithm which discovers correlations between items in transactional databases. An association rules mining is motivated by decision support problems faced by most business organizations and is described as an important area of research

[2]. One of the main challenges in association rules mining is developing fast and efficient algorithms that can handle large volumes of data because most mining algorithms perform computation over the entire database and often the databases are very large. Mining association rules may require iterative scanning of large databases, which is costly in processing [3]. The need of big data generated from the large companies like Face book, yahoo, Google, YouTube etc for the purpose of analysis of enormous amount of data which is in unstructured form converted into structured form [4]. The need of big data analytics which is stored in relational database systems in terms of five parameters-variety volume, value, veracity and velocity [5]. Text mining is a process to extract interesting and significant patterns to explore knowledge from textual data sources [6]. Text mining is a multi disciplinary field based on information retrieval, data mining, machine learning, statistics, and computational linguistics [7]. We have access to large scale user-generated data today much of the user-generated feedback is in the form of very noisy text from which it is difficult to extract information [8]. The motivation of our proposed method comes from the following observation about user reviews. When users review a service or a product, not only do they express their overall opinion on the subject but they also demonstrate their likes and dislikes over various attributes and functionalities of the service or product in question [9]. Mining tools are now an integral part of undertaking decision-making and risk administration. Business Intelligence (BI) project datasets are increasing rapidly, thanks to use Information Systems IS and data warehousing. On average Credit Card Company frequently has millions of business logged per year. Major data sets are frequently generated by great telecommunications [10].

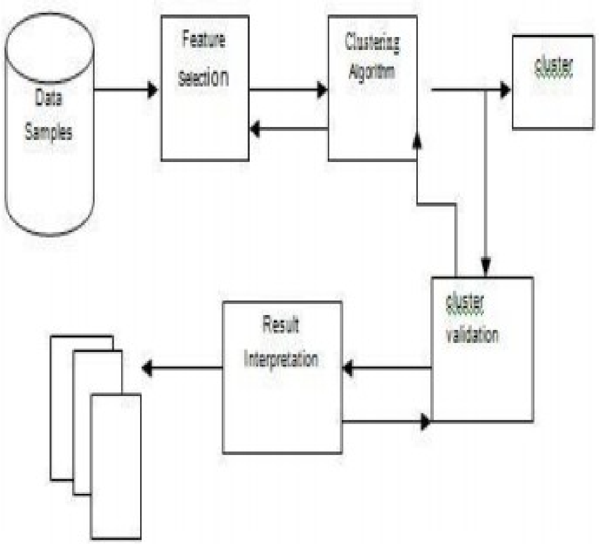

Fig1: Cluster Analysis

PROBLEM STATEMENT:

The aim of the study is directed by the need of a clear understanding of the functionality of data mining tools, and the domain of their application. This understanding idea about different tools is formulated by the comparative study carried on all the considered important tools for mining. The work done will represent improvements and suggestions to the current deficiencies while comparing some data mining tools with different datasets. The current application of these tools deals with limited data size and is restricted to particular organizational datasets where as tools are required there to support data mining on such abruptly increased and still expanding databases with dynamic data. With such distributed and huge data there is also a need for a tool to provide features that of integrated multi agent mining tools, with enhanced job scheduling, support for multiple data structures, compatibility with various decision making algorithms and ease of use for the novice user.

RELATED WORK

Research in the area of aspect-based dataset analysis is down into topic modeling based approaches and machine learning based approaches [11]. Outside of the topic modeling framework, Parts-of-Speech (POS) tagging is a widely used method for this problem [12]. The methods proposed apply POS tagging to identify nouns and noun phrases, based on the observation that aspects or features are generally nouns [13]. Business Intelligence (BI) tools and systems and about a variety of Business intelligence machinery and techniques that are utilized in text mining and a mixture of function of Text mining in diverse business intelligence standpoint proposed a work on Business Intelligence sphere and make available a few stimulating and innovate speculation and practices related to the future trends and challenges of Business Intelligence as well as the neighboring technologies, such as data warehousing and cloud computing [14]. In addition different widely used text mining techniques clustering, categorization, decision tree categorization, and their application in diverse fields are surveyed [15]. They discussed that dealing with unstructured text is difficult as compared to structured or tabular data using traditional mining tools and techniques [16]. Customer data for competitor mining is collected through several methods, which is usually unstructured most data mining technologies is only handle structured data. Therefore during competitor mining process unstructured data is not taken into account and much valuable service information is lost [17]. Structured systems are those where the data and the computing activity is predetermined and well-defined. Unstructured systems are those that have no predetermined form or structure and are usually full of textual data [18].

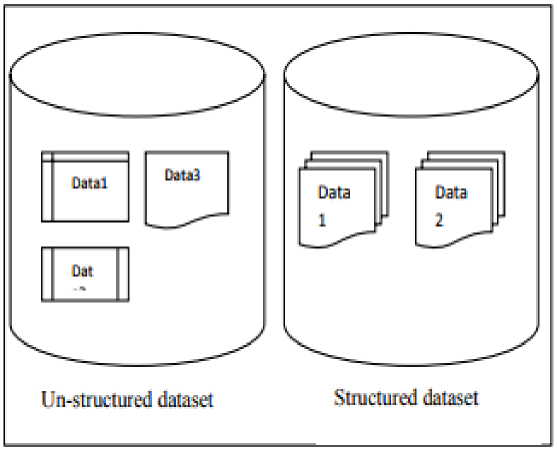

Fig 2 Structured and un-structured systems

2. ASSOCIATION RULES MINING

The mining association rules basically are to find important associations among items in a given database of sales transactions such that the presence of some items will imply the presence of other items in the same transaction. Let I = {i1, i2, …, im} be a set of binary attributes, called items [19]. Let D a set of transactions and each transaction T is a set of items such that T⊆I. Let X be a set of items. A transaction T is said to contain X if and only if X⊆T. An association rule is an implication of the form X⇒Y, where X⊂ I, Y⊂ I, and X∩Y = ∅. Furthermore, the rule X⇒Y is said to hold in the transaction set D with confidence c if there are c% of the transaction set D containing X also containing Y. The rule X⇒Y is said to have support s in the transaction set D if there are s% of transactions in D containing X∪Y. The confidence factor indicates the strength of the implication rules; whereas the support factor indicates the frequencies of the occurring patterns in the rule [20].

It has been shown that the problem of discovering association rules can be reduced to two sub-steps:

1. Find all frequent item sets for a predetermined support

2. Generate the association rules from the frequent item sets.

Different text mining techniques are available that are applied for analyzing the text patterns and their mining process.

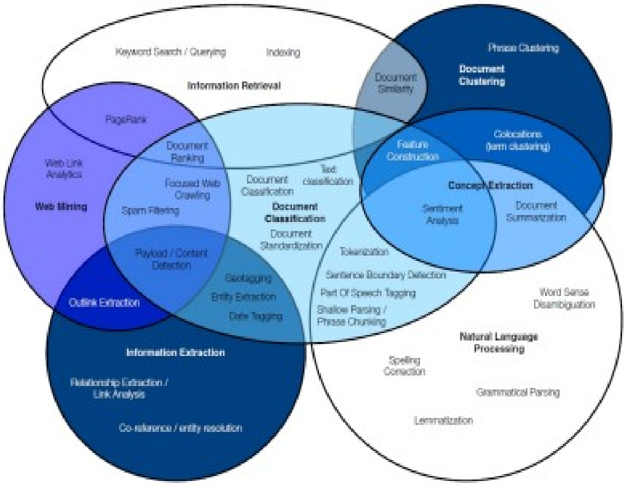

Fig. 3. Text mining Techniques

3. PROPOSED SYSTEM

For extracting the labeled data this paper proposed the clustering method to overcome the Business intelligence keeps you informed of your market trends, alert you to new boulevard of produce revenue and helps you determine how your antagonism is doing. Without that knowledge, you may endure false growth or impede Main functions of business intelligence in data mining are coverage, analytical processing, event dispensation, business performance management [21].

A. Handling Metadata using K-Medoids

Businesses store volumes of data in the form of web pages, emails, video and image files, news and reports which are called semi structured or unstructured data. In practice, such data leads to wastage of time in searching and leads to poor decisions as volumes of unstructured data are stored in variety of formats and referred by different technologies [22]. By the techniques of information extraction and automatic categorization, metadata can be generated in the form of summaries or topics it uses theK-Medoids is clustering by partitioning algorithm as like as K-means algorithm.

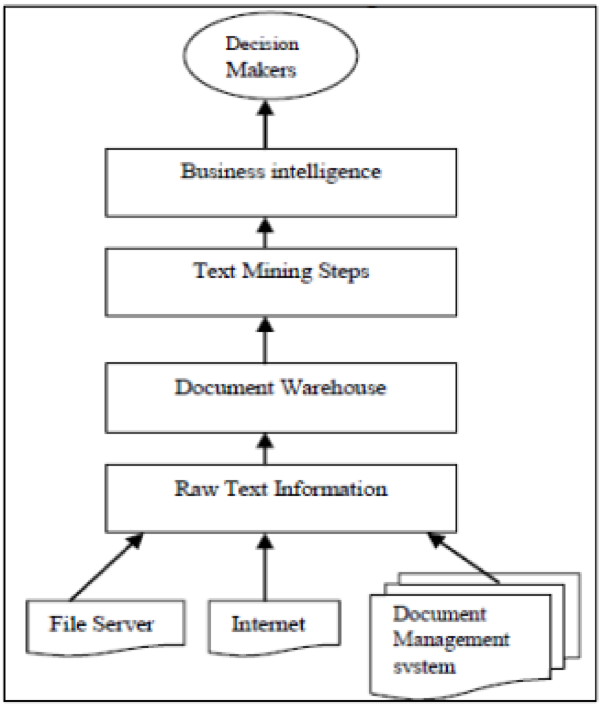

Fig 4: Business Intelligence using text mining

B. Feature extraction using SVM

Feature selection is the process of selecting a subset of feature used to represent the data. In text classification, it focuses on identifying relevant information without affecting the accuracy of the classifier. In text documents feature, can be term, pattern, sentence the traditional feature selection methods are not effective for selecting text features for solving the relevance issue because relevance is a single class problem [23]. Analyzing and solving the problem SVM is used. It is one of the simplest methods is the centroid algorithm.

C. Recommendation for BI

Using the user profiles and content profiles, the businesses apply data mining techniques to identify appropriate business rules. These rules could involve a simple classification of the users using their profiles and the website click-streams relationship between comfortable profiles and user behavior association products. The knowledge of customer's behavior will help to recover customer relationships and make business strategies [24]. To solve this problem, here use the Collaborative Filtering (CF) methods share a capability to utilize the past ratings of users in order to predict or recommend new content that an individual user will like.

4. CLUSTERING ALGORITHM:

Clustering algorithm design or selection: Patterns are grouped according to whether they resemble one another. The construction of a clustering criterion function makes the partition of clusters an optimization problem. Clustering is ubiquitous, and a wealth of clustering algorithms has been developed to solve different problems in specific fields. It is important to carefully investigate the characteristics of the problem on hand, in order to select or design an appropriate clustering strategy [25].

A. K-means Algorithm:

This method is a type of hierarchical clustering method using K-means. The algorithm starts by putting all the documents in a single cluster. It partitions the original clusters into two clusters by using K-means i.e. K=2. It makes the cluster which has highest intra cluster similarity as permanent & recursively split the other cluster into two more clusters using K-means with K=2& continue this until the desired number of clusters are created [26].

Step1: Pick a cluster to split.

Step2: Find two sub-clusters using the basic K-mean algorithm

Step3: Repeat Step2 The bisecting step for ITER times and takes the split that produces the clustering.

Step4: Repeat Step1, 2, & 3 until the desired number of Clusters are reached.

B. Key Drivers of Sentiment

We aim to identify the aspects that users base their reviews on, as well as the sentiment associated with the aspects. Thus, we propose the identification of the following two groups of words from the reviews:

• **Aspects:** Aspects are the features or attributes of the restaurant under review, such as food, service, ambience, price, etc. They form the key elements of the reviews about which users express their likes or dislikes.

• **Descriptors:** Descriptors are words that occur in the neighborhood of Aspects, and either describe the Aspect, or contain underlying sentiment associated with the Aspect. Examples include tasty, good, disgusting, expensive.

5. EXPERIMENTAL RESULT

Our proposed system helps a lot to find out the accurate data from the large dataset, collected information from the host web server and collect as much information from analyzing the web page itself. Mainly they look forth hyperlinks, cookies, and the traffic patterns. Using this collected knowledge enterprises can establish better customer relationships, offers and target potential buyers with exclusive deals. Accuracy calculation is the main one in the dataset.

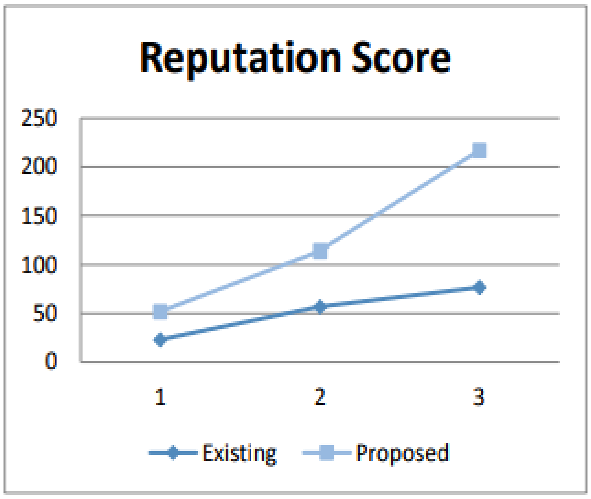

Fig No 5: Reputation score calculation

BI is also the main tools for decision support in modern enterprises. BI tools provide competitive advantages, better customer relationships managements, and enhanced management of risk in reserves. Mining tools provides predictive profiling; this means that using current and historical behaviors' of your customers, possible future behaviors of purchase are predicted.

6. CONCLUSION AND FUTURE WORK

We demonstrate the effectiveness of our algorithm using sample databases. We develop a visualization module to help the user manage and understand the association rules. Text mining plays an important role BI. It presents valuable and inexpensive information to business developers and decision makers. Many administrations have huge amounts of data in the method of formless text. This paper provide the SVM method to classify the data and stored in a well formatted thing and also using the K-Metroid algorithm for meta data handling, it clears the noise data from our dataset and finally using the CF ethos to recommend to the BI for their growth. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, the growth of data volumes and item relationships should help from legitimate patterns. Future work includes applying these algorithms to real data like retail sales transaction, medical transactions, WWW server logs, etc. to confirm the experimental results in the real life domain, since the proposed algorithm is evaluated only with test cases.

8. REFERENCES

[1] M. E. Porter, Competitive Strategy: Techniques for Analyzing Industries and Competitors. Free Press, 1980

[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD, May 1993.

[3] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. of Very Large Data Bases, Sept. 1994.

[4] BABU, G.P. and MARTY, M.N. 1994. Clustering with evolution strategies Pattern Recognition, 27, 2, 321-329.

[5] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996.

[6] N. Padhy, D. Mishra, R. Panigrahi et al., “The survey of data mining applications and feature scope,” arXiv preprint arXiv:1211.5723, 2012.

[7] W. Fan, L. Wallace, S. Rich, and Z. Zhang, “Tapping the power of text mining,” Communications of the ACM, vol. 49, no. 9, pp. 76–82, 2006.

[8] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, “Learning to classify short and sparse text & web with hidden topics from large-scale data collections,” Proceedings of the 17th International Conference on World Wide Web, pp. 91–100, 2008.

[9] I. S. Dhillon and D. S. Modha, “Concept decompositions for large sparse text data using clustering,” Machine Learning, vol. 42, no. 1-2, pp. 143– 175, 2001.

[10]. De Ville, B.. Microsoft Data Mining Integrated Business Intelligence for ECommerce and Knowledge Management. Boston: Digital Press; 1st edition

[11] L. V. Subramaniam, S. Roy, T. A. Faruquie, and S. Negi, “A survey of types of text noise and techniques to handle noisy text,” Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data, pp. 115–122, 2009.

[12] D. Lopresti, S. Roy, K. Schulz, and L. V. Subramaniam, “Special issue on noisy text analytics,” International Journal on Document Analysis and Recognition, vol. 12, no. 3, pp. 139–140, 2009.

[13] A. B. Wilcox and G. Hripcsak, “The role of domain knowledge in automating medical text report classification,” Journal of the American Medical Informatics Association, vol. 10, no. 4, pp. 330–338, 2003

[14]. ThiagarajanRamakrishnan, Mary C. Jones, Anna Sidorova, “Factors Influencing Business Intelligence and Data Collection Strategies: An empirical investigation”, Decision Support Systems. Vol 52, Issue 2, January 2012

[15]. KondaSreenu, “Web Data Mining Based Business Intelligence and Its Applications”. Vol. 4, Issue Spl - 4, Oct - Dec 2013

[16] W. He, “Examining students online interaction in a live video streaming environment using data mining and text mining,” Computers in Human Behavior, vol. 29, no. 1, pp. 90–102, 2013.

[17] Saxena, Prateek, David Molnar, and Benjamin Livshits. "SCRIPTGARD: automatic context-sensitive sanitization for largescale legacy web applications." Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011.

[18] Ghamisi, Pedram, Jon Atli Benediktsson, and Johannes R. Sveinsson. "Automatic spectral–spatial classification framework based on attribute profiles and supervised feature extraction." IEEE Transactions on Geoscience and Remote Sensing 52.9 (2014): 5771-5782.

[19] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. editors, Advances in Knowledge Discovery and Data Mining, pages 307–328. MIT Press, 1996.

[20] Generating Non-Redundant Association Rules. Mohammed J. Zaki Computer Science Department, Rensselaer Polytechnic Institute, Troy NY 12180.

[21]. JayanthiRanjan, “Business Intelligence: Concepts, Components, techniques and Benefits”, Journal of Theoretical and Applied Information Vol 9, Issue 1, Nov 2009.

[22]. Rajender Singh Chhillar, (2008) “Extraction Transformation Loading, A Road to Data Warehouse,” Second National Conference Mathematical Techniques: Emerging Paradigms for Electronics and IT Industries, India, pp. 384-388.

[23]. Judy Redfearn and the JISC Communications team, (2006) “What Text Mining can do” Briefing paper, ‘Joint Information Systems Committee’ JISC.

[24]. Velmurugan, T., Santhanam. T ,” Computational Complexity between KMeans and K-Medoids Clustering Algorithms for Normal and Uniform istributions of Data Points”. Journal of Computer Science ,pp 363–368, 2010.

[25] Patel, A.B., Birla, M. and Nair, U. (2012) Addressing Big Data Problem Using Hadoop and Map Reduce, NIRMA University Conference on Engineering, pp. 1-5.

[26] Wu Yuntian, Shaanxi University of Science and Technology, “Based on Machine Learning of Data Mining to Further Explore”, 2012 International Conference on Machine Learning Banff, Canada.

Authors profile:

1) J. Malathi,

Assistant Professor, Department of IT, Sir.C.R.Reddy College of Engineering , Eluru, AP.

Email: malathi.komma@gmail.com

2) Bikkina Lalitha Bhavani,

Assistant Professor, Department of IT, Sir.C.R.Reddy College of Engineering , Eluru, AP.

Email:karuturi.lalithacse@gmail.com