International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 7 (2017), pp. 1765-1775

http://www.ripublication.com

Extracting Structural Data For Business Intelligence Using Cluster Data Mining

Stibu Stephen[1] and Dr. R.ManickaChezian[2]

PhD. Scholar, Department of Computer Science[1],
Associate Professor, Department of Computer Science[2],
N.G.M College, Pollachi, Tamil Nadu, India.

Abstract

Data mining is a knowledge extraction field that attempt to discover and store the related pattern from the large dataset. Extraction and storing the information is useful for many intellects. Storing of data has been enormously increasing day by dayin many business administrations, so extracting patterns and drifts out of immense data is anexcessive challenge. For the persistence of information finding and information retrieval from such written data text mining is used.. 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. This large amount of information can lead to the development of new opportunities for the organization. For these issues extracting the unstructured data and form a structured data, here this paper uses the cluster data mining techniques to find out the structured data from the large datasets and provide an effective one to the business Intelligence.

Keywords: Data Mining, Business Intelligence, Textual Data and Clustering.

I. INTRODUCTION

Data Mining (DM) is defined as the process of analyzing large databases. DM scrutinizes datasets of rational databases, in multiple magnitude and angles, manufacture a summary of the common trends commencement the dataset, relationships and models that hysterics the dataset. DM is a quite new interdisciplinary field about computer science, statistical modeling, artificial

intelligence, information science, and machine erudition. One of the main uses of DM is business astuteness and risk management. Business Intelligence (BI) refers to knowledge and applications for accumulating storing and investigates business data that finally helps the enterprise to make enhanced decision. That judgment based on business-critical decisions like large datasets stored in their databases, DM straight affect decision-making [1]. DM is relied on inretail, telecommunication, rumor; indemnity, schooling, and healthcare engineering them are data-driven. Other uses of DM contain natural investigate such as DNA and the person genomeproject, geospatial and conditions investigate for study raw data. Analyzing of raw data which is generally paying notice on textual data which is related to Text Mining (TM). TM is related field to DM, but differs in its practice and methodologies used. TM is also an interdisciplinary field around computational linguistics, statistics, and machine learning. TM uses multipart Natural Language Processing (NLP) techniques. It involves a training period for the TM tool to realize patterns and hidden relations. The process of mining text documents involves linguistically and semantically analysis of the plain text, thus organization the text. Finally relates and induces some secreted traits found in the text, like frequency of use for some words, article extractions, and documents summarizations. TM is used, aside from business use, for methodical research, specifically medical and biological. TM is very levelheaded in finding and identical proteins' names and acronyms, and finding concealed relations among millions of documents.

In business submission plays an imperative role on mining software solutions. Mining tools are now an integral part of undertaking decision-making and risk administration. Obtaining information all the way through mining isreferred to as Business Intelligence BI. Project datasets are increasing rapidly, thanks to use ofInformation 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 and

Mobile machinist as they mount up to 100 million user accountsproduce thousands of millions of data per year

As these numeral increases up, investigativeprocessing such as OLAP and labor-intensive comprehension seems ineffective. With BI such tasks arewithin reach [2]. BI is put into practice through mining tools; these tools produce findings that are in due course used to gainbloodthirstygain over rivals, better and resourceful business operations, and bettersurvivability and risk administration. Mining tools provide better customers' relationshipadministration CMR, through mining real practice, patterns, and even consumers churn. Customerschurn is definite as the per cent of customers that have left the endeavor, most likely to otherchallenger or due to the inability to keep your aggressive advantage and customer pleasurelevels. Behavior and trends in customers' data help in conclude the customers' segmentations,

whatclientele to target, especially alpha customers. Alpha clientele are those that play a key role ina product achievement, thus judgment what they want is essential. This funds that, mining tools are essential for catalogue advertising industries and advertising agencies. In addition, mining tools, especially DM, make accessible market basket analysis that helps the unearthing of products that arebought usually together. As modern wealth around the world today are driven by information, fetching information and knowledge based economies [3]. This leads to extensive BI development of the CRM data. Companies can achieve aggressive compensation in providing value intention to its customers; these value propositions embrace lowest cost, bunch and offers, special discounts, and trustworthiness and partnership constitutional rights. Lowest cost nevertheless is not easily achieved, as the WWW had fetch new ways for the endeavor customers to search for other rivals maybe offering superior prices. Simplicity in worth is another factor that clients would appreciate and look for. BI tools help in stay away from cost transparency problems by offering the project the choice of price favoritism. Price favoritism means that the business offers the same product with dissimilar prices for dissimilar people and district; this of course depends on each individual's willingness to pay for how much. This energetic trait permits the project to application to its diverse customers with present that are matching to the lowest cost offers [5]. This ability is not probable without prior information of how much each personality is prepared to pay of course. Here is whereBI tools help in appraise your customer's import and expenses habits.

II. LITERATURE REVIEW

KondaSreenu [5], study and analyze the Data mining, a field at the junction of computer science and information, is the procedure that challenge to discover model in large data sets. The generally goal of the data mining process is to mine information from a data set and convert it into an comprehensible structure for supplementary use and it is division of a technological, social, and monetary rebellion that is creation the world smaller, more linked, more service driven, and as long as unparalleled levels of prosperity. At the same time, more information is recognized, amass and transmitted about us as persons than ever. Racking systems and hit contradict are commanding tools to conclude if your customers are judgment your site. However, they don't help you conclude the opportunity of growth. That's where a good online industry astuteness data service comes in. It is deal with the business; where business enclose lot of data from years. Related on query condition it should be excavation and exhibitcorrectresult. It provides a succinct indication of some of the privacy issues and disagreement that surround the use of data mining in business today.

Palak Gupta, BarkhaNarang [6] includes the shared study of business intelligence and

text mining of hesitant data. Instead, data is measured uncertain and therefore this indecision is propagated to the consequences fashioned by Business Intelligence (BI). Through use of a semantic web, text mining can find comfortable based on meaning and framework rather than just by an unambiguous word. This recovers our search and BI consequences and makes possible social networks psychotherapy or counter-intelligence. A text articulates a vast, rich range of information, but encodes this information in an appearance that is difficult to remove routinely specially from shapeless data. May be for this motive, there has been little work in text data mining to date, and nearly everyone people talk about it as in sequence admission or many have not used text unswervingly to determine unknown information. Through this paper, theywish to unearth such neighborhood and define Business Intelligence, structured or unstructured data, text mining, and then talk about the potential submission and limitations of text mining. The idea behind converse this is to draw attention to stimulating new kinds of problems and BI trends like green computing, social networking, data hallucination, mobile BI, predictive analytics, amalgamated applications, cloud computing, multi-touch and Software as-a-Service (SaaS).

SonalTiwari[7]introduces the basic ideas of Recommender system and significance of web usage mining in business intelligence. Recommender systems have materialized as powerful tools for plateful customers find items of attention. The explore work obtainable in this paper makes quite a few donations to the construction of recommender systems linked research. First of all, they offer a new scaffold based on web mining equipment for configuration a Web-page recommender system. Furthermore, we demonstrate how web withdrawal knowledge can be successfully applied in a production intellect environment. There is some probable conservatory to this work. Research for investigate customers‟ past acquire pattern will enable to discover a suitable. Also, it will be fascinating research area to demeanor a real marketing encouragement to target clientele using our slant and then to estimate its presentation. It developed web mining for business intelligence. For this we have developed a structure that uses connection rule, web mining and on hallucination of web services log data incarcerate in business intellect environment.

JayanthiRanjan et al [8] had given a successful way of amalgamate venture applications in real time by assume 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. The data that concern the business, its development and the characteristics of data eminence that are used in Business intellect and the tackle that are used for Business similar to Data mining, OLAP.)

From Rajender Singh Chhillar[9], presented a review of TM techniques. The authors evidently stated that TM faces confront as natural language processing NLP techniques are not willingly made for mining actions. The paper demonstrates quite a few TM technique that included information mining topic tracking, summarization, cataloging, gathering, concept linkage, information hallucination and problem answering to name a few. Lastly, the author confirmed that TM is used in media, reservoir politics, and smooth in insurance. How business intelligence is derivative from web mining. WM or web usage mining is portraying as an intellect tool to aid endeavor in the intense antagonism originate in ecommerce. This work obtainablean evaluation of present WM techniques used as well as commence a novel approach called intellectual miner. Intelligent miner (called i-Miner) is a mixture scaffold for WM; it uses amalgamation of algorithms for finding and dispensation log files from web servers. Its then is appropriate rules and structure to find secreted patterns originate in the log files.

From Judy Redfearn [10] the Internet and the WWW upheaval in the early 1990s, numerous research and expansion were made to mechanize the search and exploration of the net, particularly text, found in the URLs. Developments in NLP, neural networks and text giving out led eventually to explore engines development. They requires for better search algorithms led to textual searching of web pages. These expansions greatly superior the search engines and opened the entry for text mining to be functional in several other submission Search engines' technologies were centered on manager that might map the vast WWW and compare keywords and comparable other feasible keywords. These developments will lead to the more intellectual agents that search the WWW for not only keywords but also place visitors' patterns. Ultimately, the expansion in both DM and TM lead to the notion of WM, were the WWW is used as a foundation for appear for new knowledge, hidden away anywhere. WM agents are small standalone software, that crawl the WWW, attain logging data, cookies, and site visits behavior found on the servers and other machines attached to the WWW.

III. PROBLEM DEFINITION

In existing work, we find the problem of extracting labeled data, the problem were provide unstructured data problem with classification algorithm, that we challenge to address in this paper. As previously declare the classification method learns consequent functions focus on the erroneous occurrences (which provide the erroneous label). Since these consequent functions are adapted on a moderately smaller number of occurrences, they can often envisage the correct labels for beforehand incorrect occurrences. Now, while this existing process has been successful on many data sets and applications, there are two boundaries that help

explain deprived results with label noise and complex functions. Many techniques were proposed to answer this problem but in everywhere this happen because of immaterial data and noise data and also a duplicate data, so we can't extract the data in proper manner [10].

IV. PROPOSED SYSTEM

For extracting the labeled data this paper proposed the clustering method to overcome the above said issues. 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 Mainfunctions of business intelligence in data mining are coverage, analytical processing, event dispensation, business performance management, etc., for this occupation Clustering techniques work by make out groups of consumers who appear to have similar partiality Once the clusters are created, averaging the estimation of the other consumers in her come together can be used to make predictions for an entity. Some clustering techniques represent each user with partial contribution in several clusters. The prophecy is then a standard across the clusters, weighted by scale of participation. For this it have to collect the data, then administration process, storing in warehouse and extract the data for BI. This is shown in the diagrammatical representation in Fig 1.

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. 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. The most centrally situated instance in a cluster is considered as centroid in place of taking mean value of the objects in K-Means clustering. This centrally located object is called reference point and medoid. It minimizes the distance between centroid and data'spoints means minimize the squared error. KMedoids algorithm performs better than K-Means algorithm when the number of data points increases to maximum [11]. It is robust in presence of noise and outlier because medoid is less influenced by outliers, but processing is more expensive. This is used to find the attackers in wireless network by the instance of the center point. It act like the K-Means cluster algorithm and then it assigns each item to the nearest point or node with the help of similar data and also it choose the references object of the related data.

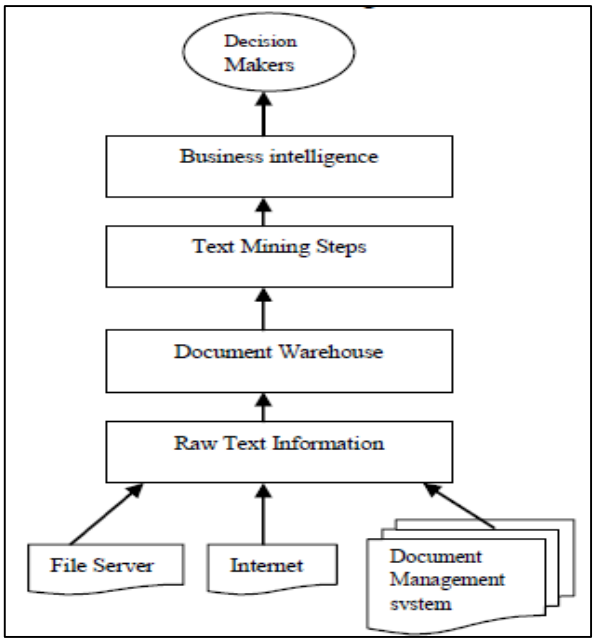

Fig 2: Business Intelligence using text mining

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. However, the traditional feature selection methods are not effective for selecting text features for solving the relevance issue because relevance is a single class problem [12]. Analyzing and solving the problem SVM is used. It is one of the simplest methods is the centroid algorithm.

Aneditorial is easily catalog by determine the centroid-vector closest to its characteristic vector. The technique is also improper when the numeral of grouping is very large. It needconstructive guidance documents and also a certain number of negative teaching documents so that we can effortlessly investigate the data.

c. **Dependency modeling** consists of finding a model that describes significant dependencies between variables. Dependency models exist at two levels: (1) the *structural level* of the model specifies (often in graphic form) which variables are locally dependent on each other and (2) the *quantitative level* of the model specifies the strengths of the dependencies using some numeric scale [13]. Probabilistic dependency networks are increasingly finding applications in areas as diverse as the development of probabilistic health expert systems from databases, data retrieval, and representation of the human genome.

d. 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, or association amongdiverse products. The knowledge of customers‟ behavior will help to recover customer relationships and make business strategies. 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. The real assumption is highly based in the idea of likeness between users or between products, with the similarity being expressed as a function of agreement between past ratings or preferences. Two basic variants of CF approach can be classified as user-based and item-based. It helped a lot to define the business aspect to recommend the items to the others.

V. EXPERIMENTAL RESULT

Our proposed system helps a lot to find out the accurate data from the large dataset, collectedinformation from the host webserver and collect as much information from analyzing the web page itself. Mainly they look forth hyperlinks, cookies, and the traffic patterns.

Chart 1 shows the result of the existing and proposed system using the clustering techniques.

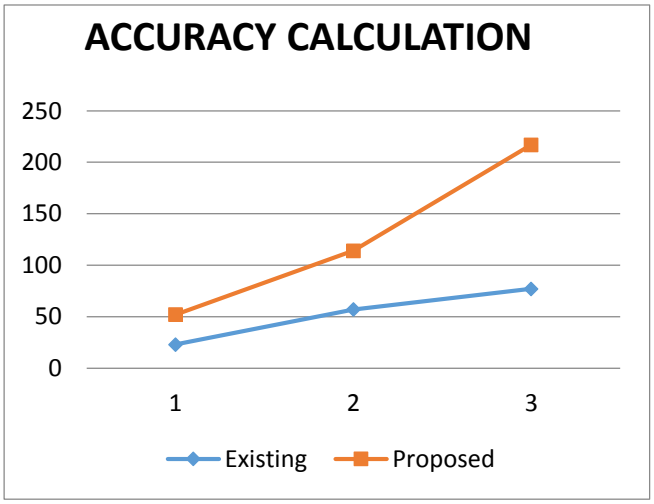

Chart 1: Accuracy Calculation

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.

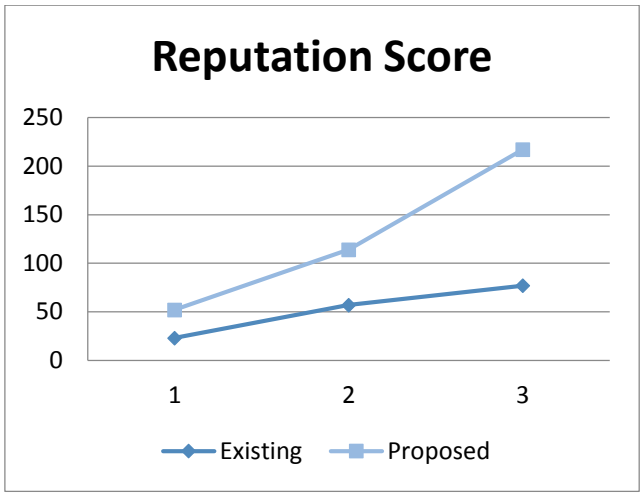

Chart 2: Reputation score calculation

Chart 2 shows the reputation among the data from proposed and the existing system. Reputation score is one of the function calculations of the data status.

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.

VI. CONCLUSION

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 large amount of information can lead to the development of new opportunities for the organization. handling those data is a crucial thing, for this process 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. With the help of the cluster data mining techniques it provides a structured data from the large datasets and provide an effective method for business Intelligence.

REFERENCES

[1]. Abdul-Aziz Rashid Al-Azmi, "Data, text, and web mining for business Intelligence: a survey". (IJDKP) Vol.3, No.2, March 2013

[2]. De Ville, B.. Microsoft Data Mining Integrated Business Intelligence for E-Commerce and Knowledge Management. Boston: Digital Press; 1st edition (17 May 2001)

[3]. ChidanandApte, Bing Liu, Edwin P.D. Pednault, Padhraic Smyth, "Business Applications of DataMining," Communications of the ACM, Vol 45 Issue 8, August 2002.

[4]. 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

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

[6]. Palak Gupta1, BarkhaNarang, "Role of Text Mining in Business Intelligence". Volume 1, Issue 2 (Jan – Mar 2012)

[7]. SonalTiwari, "A Web Usage Mining Framework for Business Intelligence".

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

[9]. 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.

[10]. Judy Redfearn and the JISC Communications team, (2006) "What Text Mining can do" Briefing paper, 'Joint Information Systems Committee' JISC.

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

[12]. C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[13]. Glymour, C.; Madigan, D.; Pregibon, D.; and Smyth, P. 1996. Statistics and Data Mining. *Communications ofthe ACM* (Special Issue on Data Mining). November 1996. Forthcoming.

Sarwar, B., Karypis, G., Konstan, J.A., &Reidl, J. Item-based Collaborative Filtering Recommendation Algorithms. Proceedings of the Tenth International Conference on World Wide Web, pp. 285 -295, 2001.

Authors:

* **Stibu Stephen** received MSc(C.S) from Calicut University, Calicut. He completed his M.Phil Degree from Bharathiar University, Coimbatore. Currently he is doing Ph.D in Computer Science at NGM College Pollachi. India. He has 5 years of Teaching Experience. His research interest includes in the areas of Data Mining and Knowledge management.

** **Dr.R.ManickaChezian** received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph.D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published one-fifty papers in international/national journal and conferences: He is a recipient of many awards like Desha Mithra Award and Best Paper Award. Recently he received the award "Best Computer Science Faculty of the Year 2015" from Association of Scientists, Developers and Faculties. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.