International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 14008-14013
© Research India Publications. http://www.ripublication.com

Explorative Data Visualization Using Business Intelligence and Data Mining Techniques

G S Ramesh [1]
Assistant Professor, Department of Computer Science and Engineering,
Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology,
Bachupally, Hyderabad, Telangana, India.
Orcid Id: 0000-0002-3308-535X

Dr. T.V. Rajinikanth[2]
Professor, Department of Computer Science and Engineering,
Sreenidhi Institute of Science and Technology, Yamnamet,
Ghatkesar, Hyderabad, Telangana, India.

Dr. D Vasumathi [3]
Professor, Department of Computer Science and Engineering,
Jawaharlal Nehru Technological University Hyderabad, Kukatpally,
Hyderabad, Telangana, India.

Abstract

Today the Business data is exploding and posing challenges to Business community in terms of effective analysis and decision making process. Business Intelligence (BI) uses various strategies and techniques that will help enterprises to perform data analysis effectively. It provides legacy, current and futuristic views of business operations.BI technologies are capable of processing huge amounts of data. The data can be either structured or unstructured. Data Mining (DM) Techniques will help the business community for making potentially useful decisions to develop their business in profitable manner apart from withstanding the challenges. DM techniques consist of Classification, Clustering, and Association algorithms which are used to identify hidden patterns. This Paper aims at using both BI and DM techniques together in visualizing and analyzing the business data. Explorative data visualizations were also made to identify the hidden patterns for effective decision making suitable to their needs

Keywords: Business Intelligence (BI), Data Mining (DM), Classification, Clustering, Association Mining.

INTRODUCTION

The amount of information available to Enterprises is growing rapidly; the data is doubling every two to three years. It caused difficulties in report simplification which spread over multiple transaction system. BI is the method of turning data into information and then into knowledge. It supports business analysis and decision support system. BI will make sure are we doing things right for the short term and are we doing right things for the long term. BI is used for the reasons like the position of the organization when compared to it's competitors in market, customer behavior change, time and money sending patterns, market condition, capabilities of organization demographic information, social and political environments, local conditions and other information. The advantage of BI is that it converts business knowledge using analytics and give solutions to the business problems like decrease response time and increase response rate using e-mail, phone etc., BI is useful in the cases where firms wants to identify reasons in change in customer behavir and recognize future customers by comparing the previous patterns. It is also useful in identifying money laundering activities and to detect fraudulent behavior like spikes in money sending when credit/debit card is stolen. Data mining technique is a collection of a set of processes that relates to analyze data apart from finding useful, actionable knowledge hidden beneath large volumes of data stores or data sets. The task of knowledge discovery is to find out patterns in the data which helps to make decisions that leads to profits in business. Data mining techniques are applied on current data and historic data which may be in large volumes to find out the hidden knowledge. Upon collecting the data from various sources, it is preprocessed, validated and stored in Data Warehouse/Data Marts. BI tools extract required information from this data and use for knowledge discovery process. Data Mining can be done using classification, clustering Market-Basket analysis, Association rule mining, link based analysis and so on.

RELATED WORK

Greg Nelson et al. [1] stated about landscape of business intelligence and how it has been implemented in SAS apart from what things can accomplish with BI, discussed the tools and technologies interact in some common architectural scenarios. Demonstration of the capabilities of the SAS suite of BI tools made. Brojo Kishore Mish et al. [2] presented a review on the influence of DM in BI. It involves three steps: explorations, pattern identification and deployment. BI is the hot topic among all industries aiming for relevance. BI emphasizes on detail integration and or organizing of data. DM and BI work together to reduce work load on end user and organization. It also explains Business Analytics (BA) which is a part of BI. There are various sectors in business to which BA has proved to be a powerful tool to obtain effective results. Mohiuddin Ali Khan et al [3] stated that educational data mining is useful for evaluating academic environment. They have used Apriori algorithm to the student result data for analysis. Academic planners can monitor the student academics performance there by make effective decisions to improve the results which lead to increase in profits of private educational institutions. Memorie Mwanza et al [4] developed a model for identification of fraud on tax and taxpayer data. They concentrated on baseline study which is focused on fraud detection and challenges for tax payers and development of automated tool to detect fraud based on fundamental study. Random audits and undercover operations were used to detect fraud. Distance based outliers queries were developed for continuous monitoring. They used these algorithms to find out both under and over payments based on business rules. They are marked as outliers. Vitri Tundjungsari [5] stated that social media is one of fastest growing medium across the world which available 24 hours which has generated extraordinary volumes of social data. Social media has latent content which can be used for mining. Therefore mining social media can yield to patterns which are useful for business as well as customers. They made reviews based on data mining and social media, presented various techniques which have potential to be chosen as methods of mining the social media, and explained few applications of data mining, particularly for telecommunication industry, to provide customer satisfaction and preserve customer relationship in the industry of telecommunication . Mrs. Smita Bhanap et al [6] stated that social networks are providing a platform for users to pass and use information. Mining social networks has worth to extract patterns, behaviors which are beneficial business organizations and customers. Different mining tools and techniques can be used to achieve better decision making in the field of Electronic Business Intelligence. The review on the basics of Data Mining, Social Network Analysis and its applications in Business Intelligence with data mining techniques suggest how this survey and study of the data mining approaches can benefit the importance of social network analysis and mining for business intelligence. Arti J. Ugale1 et al [7] stated that Data Mining can be used to find out patterns within a database. BI aims at integration and organization. DM assists BI’s objectives. DM along with BI process the data and evaluates reports such that it reduces workload of end user and assists in grasping findings. It can be carried out by identifying relationships in the data and opportunities, risks of the company. BI assists managers by integrating the result report to make operational strategic business decisions. DM can be used to achieve business intelligence objectives. Customer data can be used for pattern recognition by using classification. The extracted information can be used in decision making. Identifying pattern and new trends within time is very crucial for business success. Companies have huge volume of data but unable to turn it into knowledge. The target is to find out low risk customer with high profit using customer clustering. Rina Fitriana et al [8] stated majority of research work is focused on using single approach for BI. Integrating DM with BI is most widely used(67%). They integrated supply chain management, Data Mining, Enterprise Resource Planning Customer relationship Management, Data Warehouse, Decision Support System, Business Process Management, Knowledge Management, AI, ETL, OLAP, Quality Management System, Strategic Management. Gintarė Vizgaitytė et al [9] stated that their study aimed at exposing the part of business intelligence in upcoming years using the analysis of ongoing investigation and real trends. The recent/ongoing trends in BI technologies are software as a service, BI in mobiles, BI based on location, big data and predictive analytics. They stated that in BI the role of human is underestimated. Considering human role in BI will provide additional information for more effective BI systems. problem-solving support, coordination of BI activities are some of the important human factors. Lian Duan [10] stated that new technologies play vital part to connect enterprise and industry informatics apart from improvement of application in enterprise. This paper focused on outline of BI with the importance of basic algorithms and modern progress. In addition, they explained the challenges and chances to easily connect industrial informatics to enterprise systems for BI investigation. Jayanthi Ranjan [11] stated that Companies realized importance of achieving targets set by their business strategic using BI. By identifying BI needs, a real tie BI is required. This paper see the sights of BI and its components, emerging trends in BI , benefits of BI, influencing factors of BI, BI requirements, designing and implementing BI, and several BI techniques. Deepti Sindhu et al [12] stated that Data mining techniques in real are applicable on only digital data. Business Intelligence can be aided very efficiently using data mining techniques and tools. Data mining accomplishes the objectives of business intelligence by classification, pattern recognition, clustering, prediction and decision making. The concepts of Data Digitalization, Data Mining, DM Techniques, Knowledge Discovery and Business Intelligence were introduced. Dmitry Brusilovsky et al [13] stated that by using data mining techniques, an enterprise can acquire competitive information and thereby gain advantage which cannot be acquired in any other way.

PROPOSED SYSTEM

Initially raw data set namely Products data set was taken and preprocessed for removal of inconsistency; filling of missing values etc, was done and transformed into a refined and suitable data set. Business Intelligence consists of data visualization, Reporting, Online Analytical Processing and statistical Analysis. On the refined suitable data set data visualization techniques were applied as part of BI techniques. Data analysis was done using Exploratory Data Analysis techniques. Later Hybrid Data mining techniques were applied in which initially it was preprocessed and later attribute relevance analysis was done using Cfs-Subset-Evaluator algorithm along with Best First Algorithm. Only Relevant attributes were considered for further analysis. Then K-means clustering algorithm was applied and made in to 5 clusters. Finally on the resultant clustered data set decision tree C4.5 (J48) classifier algorithm was applied and later performance of classifier was evaluated and it found to be good.

EXPERIMENTAL RESULTS

Tree map drawn between Products Name versus Units in Stock and which is shown below in Fig.1.

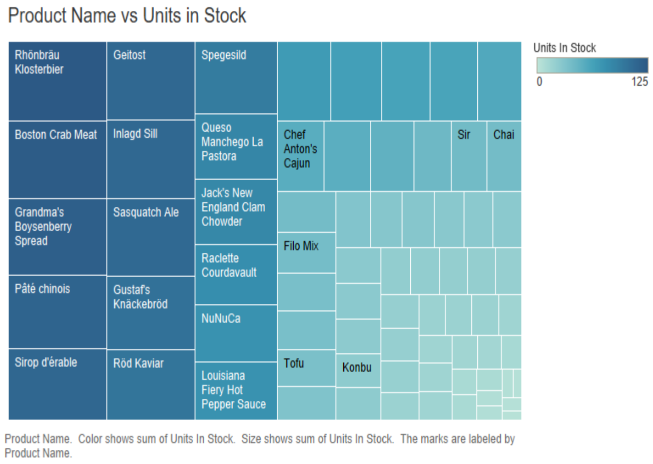

Figure 1: Tree Map of Products Name vs Units in stock

It is observed from the Tree Map shown by Fig.1 that highest Units of stocks of the products were arranged in decreasing order in Column wise as we move from Top to bottom like RhonBrau Klosterbier, Boston Crab Meat, Grandma's Boysenberry Spread, Sasquatch Ale etc. Highest numbers of units were available for RhonBrau Klosterbier and one of the products with lowest stock Chef Anton's Gumbo Mix etc. Normally Tree maps were used when space is constrained and you have a large amount of hierarchical data that you need to get an overview of. Tree maps should primarily be used with values that can be aggregated. The benefit of Tree maps is that they can be used within a limited space and yet display a large number of items simultaneously. The rectangles in the Tree map oraznized in size from the top left corner of the chart to the bottom right corner, with the biggest rectangle placed in the top left corner shows with peak number of units of stock and the smallest rectangle in the bottom right corner shows lowest number of units of stock. The detailed list of units in stock against the products was shown in Fig.2.

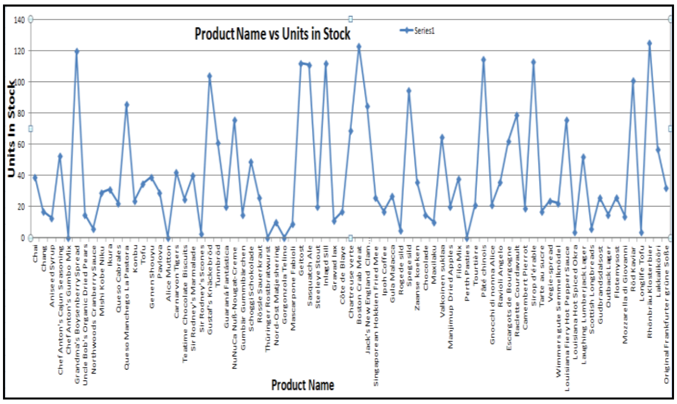

Figure 2: Bar chart for Products vs Units in Stock

The following Fig.3 shows another way of representing the scattered plot of Products vs Units in Stock along with values of Stock in Hand. End user can simply classify which products stocks are more and which products stock are less in an efficient manner. Compare to other visualization techniques this may be the best.

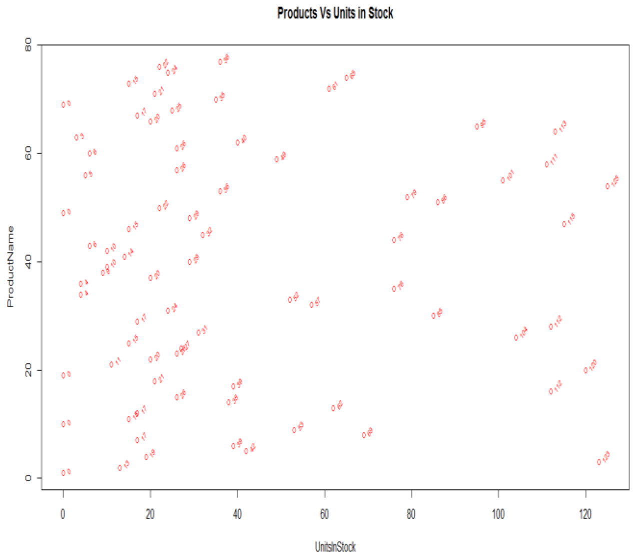

Figure 3: Products vs Units in stock

The following Fig.4 shows highest Unit Prices category ID and Product name and also lowest Unit Price.

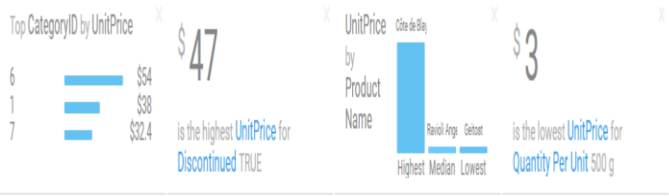
Top CategoryID by UnitPrice
6 $54
1 $38
7 $32.4
$ 47
is the highest UnitPrice for Discontinued TRUE
UnitPrice by Product Name
Côte de Blay
Ravioli Ang
Geitost
Highest
Median
Lowest
$ 3
is the lowest UnitPrice for Quantity Per Unit 500 g

Figure 4: Highest Unit Price vs Category ID and Product Name

The following Fig.5 shows lowest Regression curve between Units In Stock vs Reorder Level

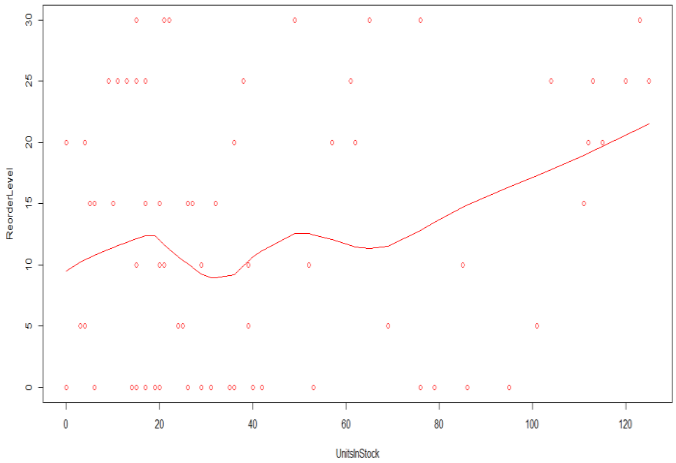

Figure 5: lowest Regression curve between Units In Stock vs Reorder Level

The Simple Linear regression equation on Reorder Level vs Units In Stock subjected to 10 fold cross validation is given by

$$ReorderLevel = 0.07 * UnitsInStock + 9.63 \text{ ----} \quad (1)$$

The Linear regression equation on Reorder Level vs UnitsInStock subjected to training data set is given by

$$ReorderLevel = 0.07 * UnitsInStock + 9.631 \text{ --} \quad (2)$$

The Equations (1) and (2) are representing the same relation between Reorder Level vs Units In Stock whether they are subjected to 10 fold cross validation or simply using training data set. The association relation of Reorder Level and Unit Price values were shown in the following Fig.6. Most of the Reorder Levels were zero for different Unit prices followed by reorder level 15. Reorder point the threshold that activates the replacement. Reorder points are the points that are tested for their correctness.

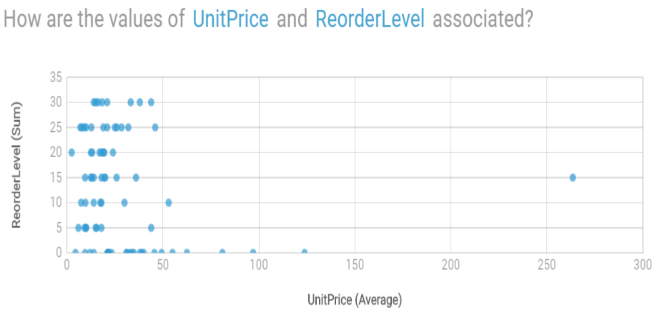

Figure 6: The association of Unit Price vs Reorder Level values

The Fig.7 shows Unit price over Rows by Reorder Level. It is witnessed that number of Reorder Levels with 0-6 is the peak with Unit price followed by 25-30 level compared to other Reorder Level.

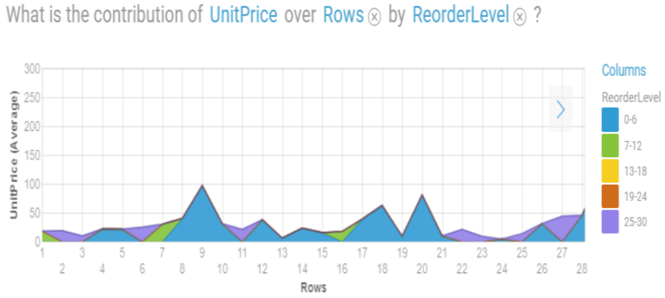

Figure 7: Unit Price over Rows by Reorder level

The following Fig.8 shows that there is a relation between Product Name vs Supplier ID. This shows which supplier supplies which product.

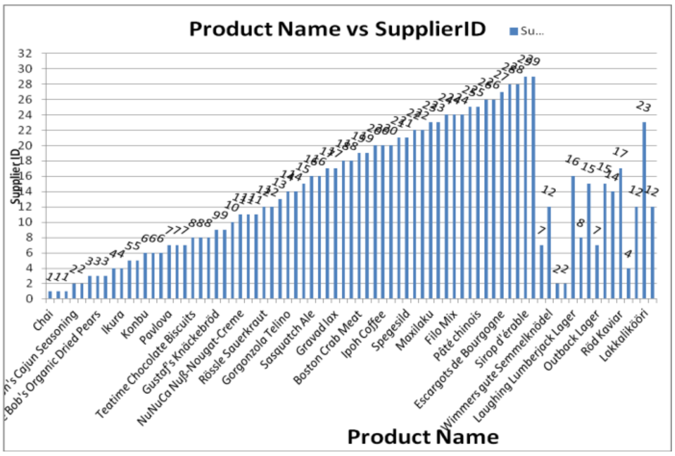

Figure 8: Product Name vs Supplier ID

The following Fig.9 shows Units In Stock vs Units On Order shows that most of the products are in sufficient quantity so Ordering of Units are not required. Same is observed in the following graph.

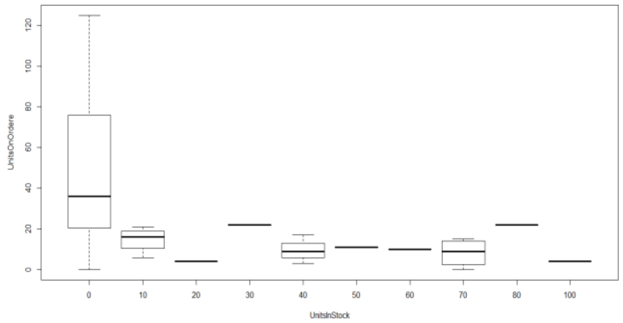

Figure 9: UnitsInStock vs UnitsOnOrder

The Products data set was subjected to data cleaning process and after that it was subjected to attribute relevance analysis and selected only the required attributes using algorithms Cfs-Subset-Evaluator and Best-First. Cfs-Subset-Evaluator algorithm assess the value of a subset of attributes by looking the individual foretelling ability of each characteristic along with the amount of duplication between them. Subsets of characteristic that are extremely correlated with the class while ensuring low inter-correlation are chosen. Best-First Algorithm explores the space of attribute subsets by using greedy hill climbing method enhanced with a backtracking ability. Setting the number of successive non-improving nodes permitted the level of backtracking done. Best first begin with the void set of attributes and investigate forward, or begin with the full set of attributes and investigate backward, or begin at any point and investigate in both directions. Then k-means cluster algorithm was applied and found 5 clusters with their values namely cluster0, cluster1 etc. Later the classifier decision tree J48(C4.5) algorithm was applied on the clustered data set which is shown in Fig.10. and Fig.11.

J48 pruned tree:

```
ReorderLevel <= 10
|  UnitsInStock <= 20
|  |  ReorderLevel <= 5: cluster0 (12.0)
|  |  ReorderLevel > 5: cluster2 (2.0)
|  UnitsInStock > 20
|  |  UnitsInStock <= 61: cluster2 (18.0)
|  |  UnitsInStock > 61: cluster4 (7.0)
ReorderLevel > 10
|  UnitsInStock <= 57: cluster3 (25.0)
|  UnitsInStock > 57: cluster1 (13.0)
```

Figure 10: Decision Tree J48 algorithm

It is observed from the Fig.10 and Fig.11 that it has 6 leaves with size of the tree is 11 and took time to build the model is 0.08 seconds.

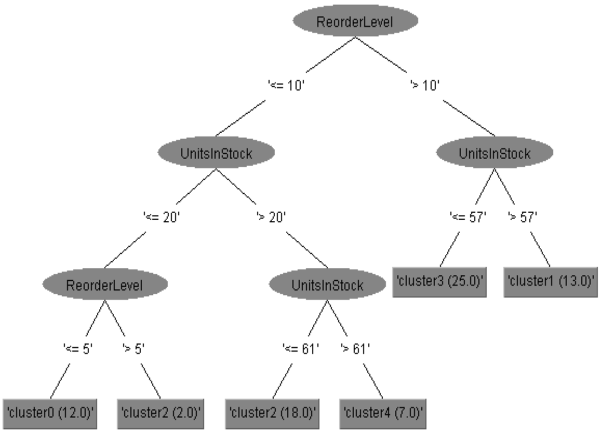

Figure 11: J48 Decision Tree

The performance parameters are correctly classified instances are 100%, Kappa statistic is 1, Mean Absolute error is 0 and Root Mean square error is 0. TP rate, Precision, Recall and F-measure of all cluster variables is 1. The performance of the C4.5 is shown by Confusion Matrix represented by Fig.12. It shows that classifier performance is 100%.

Confusion Matrix

```
 a  b  c  d  e   <-- classified as
 12  0  0  0  0 |  a = cluster0
0 13  0  0  0 |  b = cluster1
 0  0 20  0  0 |  c = cluster2
0  0  0 25  0 |  d = cluster3
 0  0  0  0  7 |  e = cluster4
```

Figure 12: Confusion Matrix

CONCLUSION

The product RhonBrau Klosterbier has highest UnitsInStock. Most of the Reorder Levels were zero for different Unit prices followed by reorder level 15. Reorder point the threshold that activates the replacement. There is an association of Unit Price vs Reorder Level values. It is observed that number of Reorder Levels with 0-6 is the peak with Unit price followed by 25-30 level compared to other Reorder Level. It is observed that there is a relation between Product Name vs

Supplier ID. It is found that most of the products are in sufficient quantity so Ordering of Units is not required. The data set was clustered in to 5 clusters. J48 Decision tree Algorithm after the clustering algorithm has yielded good performance.

REFERENCES

[1] Greg Nelson ThotWave Technologies, Chapel Hill, North Carolina, Introduction to the SAS® 9 Business Intelligence Platform: A Tutorial, Sesug Proceedings.

[2] Brojo Kishore Mish, Deepannita Hazra, Kahkashan Tarannum, Business Intelligence using Data Mining techniques and Business Analytics, System Modeling & Advancement in Research Trends (SMART), International Conference, IEEE Xplore, DOI: 10.1109/SYSMART .2016.7894496.

[3] Mohiuddin Ali Khan, Wajeb Gharibi, Sateesh Kumar Pradhan, Data mining techniques for business intelligence in educational system: A case mining, Computer Applications and Information Systems (WCCAIS), 2014 World Congress, DOI: 10.1109/WCCAIS.2014.6916559.

[4] Memorie Mwanza and Jackson Phiri, Fraud Detection on Bulk Tax Data Using Business Intelligence Data Mining Tool: A Case of Zambia Revenue Authority, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016, ISSN (Online) 2278-1021 ISSN (Print) 2319 5940.

[5] Vitri Tundjungsari, Business Intelligence with Social Media and Data Mining to Support Customer Satisfaction in Telecommunication Industry, International Journal of Computer Science and Electronics Engineering (IJCSEE) Volume 1, Issue 1 (2013) ISSN 2320–4028 (Online).

[6] Mrs. Smita Bhanap, Dr. Seema Kawthekar, Data Mining for Business Intelligence in Social Network: A survey, International Advanced Research Journal in Science, Engineering and Technology Vol. 2, Issue 12, December 2015, ISSN (Online) 2393-8021 ISSN (Print) 2394-1588.

[7] Arti J. Ugale1 , P. S. Mohod, Business Intelligence Using Data Mining Techniques on Very Large Datasets, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438.

[8] Rina Fitriana, Eriyatno, Taufik Djatna, Progress in Business Intelligence System research: A literature Review, Vol: 11 No: 03, International Journal of Basic & Applied Sciences IJBAS-IJENS. 118503-6464 IJBAS-IJENS © June 2011 IJENS.

[9] Gintarė Vizgaitytė, Rimvydas Skyrius, Business Intelligence in the Process of Decision Making: Changes and Trends, ISSN 1392-1258. ekonomika 2012 Vol. 91(3).

[10] Lian Duan, Li Da Xu, “Business Intelligence for Enterprise Systems: A Survey”, IEEE Transactions on Industrial Informatics (Volume: 8, Issue: 3, Aug. 2012).

[11] Jayanthi Ranjan, Business Intelligence: Concepts, Components, Techniques and Benefits Journal of Theoretical and Applied Information Technology © 2005 - 2009 JATIT.

[12] Deepti Sindhu, Anupma Sangwan, Optimization of Business Intelligence using Data Digitalization and Various Data Mining Techniques, International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 8 (2017), pp. 1991-1997 © Research India Publications http://www.ripublication.com.

[13] Dmitry Brusilovsky & Eugene Brusilovskiy, White Paper: Data Mining: The Means to Competitive Advantage April 2008