World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:3, 2016

Knowledge-Driven Decision Support System Based on Knowledge Warehouse and Data Mining by Improving Apriori Algorithm with Fuzzy Logic

Pejman Hosseinioun, Hasan Shakeri, Ghasem Ghorbanirostam

***Abstract*—**In recent years, we have seen an increasing importance of research and study on knowledge source, decision support systems, data mining and procedure of knowledge discovery in data bases and it is considered that each of these aspects affects the others. In this article, we have merged information source and knowledge source to suggest a knowledge based system within limits of management based on storing and restoring of knowledge to manage information and improve decision making and resources. In this article, we have used method of data mining and Apriori algorithm in procedure of knowledge discovery one of the problems of Apriori algorithm is that, a user should specify the minimum threshold for supporting the regularity. Imagine that a user wants to apply Apriori algorithm for a database with millions of transactions. Definitely, the user does not have necessary knowledge of all existing transactions in that database, and therefore cannot specify a suitable threshold. Our purpose in this article is to improve Apriori algorithm. To achieve our goal, we tried using fuzzy logic to put data in different clusters before applying the Apriori algorithm for existing data in the database and we also try to suggest the most suitable threshold to the user automatically.

***Keywords*—**Decision support system, data mining, knowledge discovery, data discovery, fuzzy logic.

I. Introduction

DECISION SUPPORT SYSTEMS (DSS) increasingly become more critical to the daily operation of organizations [1]. DSS is an equivalent synonym as management information systems (MIS). Most of imported data are used in solutions like data mining (DM). Successfully supporting managerial decision-making is critically dependent upon the availability of integrated, high quality information organized and presented in a timely and easily understood manner [2]. Since the mid-1980s, data warehouses have been developed and deployed as an integral part of a modern decision support environment [1]. Therefore, Data Warehouse (DW) provides an infrastructure that enables businesses to extract, cleanse, and store vast amounts of corporate data from operational systems for efficient and accurate responses to user queries [3]. DW is one of the solutions for decision-making process in a business organization. But it only stores data for managerial purpose and it has no intelligent mechanism for decision making. This raises the issue of knowledge storage in organization for high capability decision support [4]. Knowledge in the form of procedures, best practices, business rules, expert knowledge, facts within a context and processed data can be stored in logical structures accessible by computers. The logical structures in the knowledge warehouse to store knowledge are analogous to the system of tables that implement data storage in the DW. Knowledge is applied through a layered representation that is readable by both humans and machines this representation is also a system executable that is portable and can be run on a computer to help make decisions and take actions [5]. The enterprise-wide information delivery systems provided in a DW can be leveraged and extended to create a knowledge warehouse (KW). A framework of KW is introduced, which is enhanced form of DW to provide a platform/ infrastructure to capture, refine and store consistent and adequate knowledge along with data to improve decision making in an organization [4]. The primary goal of a KW is to provide the decision-maker with an intelligent analysis platform that enhances all phases of the knowledge management process. KW architecture will not only facilitate the capturing and coding of knowledge but also enhance the retrieval and sharing of knowledge across the organization [3]. In order to understand, analyze, and eventually make use of a huge amount of data, Enterprises use mining technologies to search vast amounts of data for vital insight and knowledge. Mining tools such as DM, text mining, and web mining are used to find hidden knowledge in large databases or the Internet [6]. DM is the process of identifying interesting patterns from large databases. DM has been popularly treated as a synonym of knowledge discovery in databases, although some researchers view DM as an essential step of knowledge discovery [7]. In this paper, mining tools are automated software tools used to achieve decision making process by finding hidden relations (rules), and predicting future events from vast amounts of data.

Pejman Hosseiniun is with the Sama Technical and Vocational Training College, Islamic Azad University, Islamshahr Branch, Islamshahr, Iran (phone +98-912-372-5474; e-mail: p_hosseiniun@yahoo.com).

Hasan Shakeri is with the Department of Computer Engineering, Mashhad Branch Islamic Azad University Mashhad, Iran (e-mail: shakerie@hotmail.com).

Ghasem Ghorbanirostam is with the Sama Technical and Vocational Training College, Islamic Azad University, Islamshahr Branch, Islamshahr, Iran.

II. Knowledge-Driven DSS

A knowledge-driven DSS provides specialized problem solving expertise stored as facts, rules, procedures, or in similar structures and it suggests or recommends actions to managers [8]. A KD-DSS is a knowledge driven DSS, which has problem solving expertise. The KD-DSS can give

suggestions or recommendations based on several criteria's. These systems require human-computer interaction.

Advanced analytical tools like DM can be integrated with the KD-DSS to find hidden patterns. Knowledge Driven DSS is also called as Intelligent Decision Support methods, and it is analogues to the KW strategy work. We choose KD-DSS model, because it has capacity to self-learn, identify associations between the data, and perform heuristic operations, if required. These abilities turn the DSS system into intelligent, increase the capacity of problem solving and improve suggestion accuracy. It is important to mention that the Knowledge representation play key role in KD-DSS. Well-defined knowledge representations include rule-based systems, semantic web and frame systems. A rule-based system contains rules in the database [9].

III. Knowledge Warehouse

KW can be thought of as an "information repository". The KW consists of knowledge components (KCs) that are defined as the smallest level in which knowledge can be decomposed. Knowledge components (objects) are cataloged and stored in the KW for reuse by reporting, documentation, execution the knowledge or query and reassembling which are accomplished and organized by instructional designers or technical writers. The idea of KW is similar to that of DW. As in the DW, the KW also provides answers for ad-hoc queries, and knowledge in the KW can reside in several physical places [10].

A KW is the component of an enterprise's knowledge management system. The KW is the technology to organize and store knowledge. The KW also has logical structures like Computer programs and databases to store knowledge that are analogous to the system of tables that implement data storage in the DW [5]. The primary goal of a KW is to provide the knowledge worker with an intelligent analysis platform that enhances all phases of the knowledge management process [3], [1]. Like the DW, the KW may be viewed as subject oriented, integrated, time-variant, and supportive of management's decision making processes. But unlike the DW, it is a combination of volatile and nonvolatile objects and components, and, of course, it stores not only data, but also information and knowledge [11].

The KW can also evolve over time by enhancing the knowledge it contains [3]. KW provides the infrastructure needed to capture, cleanse, store, organize, leverage, and disseminate not only data and information but also knowledge [4].

IV. Knowledge Discovery Process

Knowledge discovery in databases (KDD) is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economical needs. The term KDD is used to denote the overall process of turning low-level data into high-level knowledge. A simple definition of KDD is as follows: Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [12].

Knowledge Discovery has also been defined as the 'non-trivial extraction of implicit, previously unknown and potentially useful information from data'. It is a process of which DM plays an important role to extract knowledge from huge database (data warehouse) [13]. DM is the core part of the knowledge discovery in database (KDD) process as shown in Fig. 1 [13].

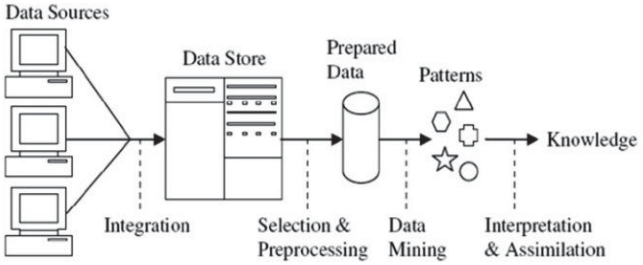

Fig. 1 Typical knowledge discovery process

The KDD process may consist of the following steps: 1) data integration, 2) data selection and data pre-processing, 3) DM as it will be explained in section 5; 4) interpretation & assimilation. Data comes on; possibly from many sources therefore it is integrated and placed in some common data store like DW. Part of it is then selected and pre-processed into a standard format. This 'prepared data' is then passed to a DM algorithm which produces an output in the form of rules or some other kind of 'patterns'. These are then interpreted to give new and potentially useful knowledge. Although the DM algorithms are central to knowledge discovery, they are not the whole story. The pre-processing of the data and the interpretation of the results are both of great importance [13].

V. DM Technique

DM is one of the most important techniques that are used to discover required knowledge for intended enterprise. DM derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Since mining for gold in rocks is usually called "gold mining" and not "rock mining", thus by analogy, DM should have been called "knowledge mining" instead [14]. DM is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information [15].

DM is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies, and significant structures from large amount of data stored in databases, DW, or other information repositories [16]. DM refers to discover useful, previously unknown knowledge by analyzing large and complex" data sets. DM is defined as the extraction of patterns or models from observed data [12].

DM, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While DM and knowledge discovery in databases (or KDD) are frequently treated as synonyms, DM is actually part of the knowledge discovery process [14].

The goal of DM is to allow a corporation to improve its marketing, sales, and customer support operations through a better understanding of its customers. DM, transforms data into actionable results [17]. Other similar terms referring to DM are: data dredging, knowledge extraction and pattern discovery [14].

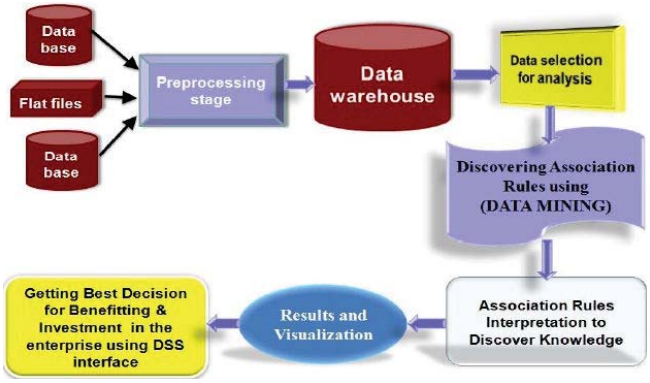

Fig. 2 The proposed knowledge-driven DSS system

VI. The Proposed and Designed System

In this paper, we proposed a knowledge-driven DSS and it consists of several phases as shown in Fig. 2. These phases are:

1- Collect data from different sources, these sources can be different files such as (Excel, Access, Word, Text files, etc.)
2- Data pre-processing

This phase consists of the following three steps:

a- Data integration
b- Data reduction
c- Data consistency

3- Loading the cleaning data after performing preprocessing steps into the DW
4- Data selection for knowledge discovery phase
5- Knowledge discovery by applying DM and association rule mining task in particular.
6- Interpret the association rules to discover and gain knowledge as output.
7- Represent the result which is knowledge using one of the visualization tools
8- Make decisions by investment and benefit from the output (knowledge) of the system through the DSS system interface.
9- In the first step of the proposed system which is *Data Gathering and Integrating phase*, we have collected data about items sales of a building items market from several sources and files such as (text file, excel, access, …etc) that have been existed in multiple sales departments of the market. Where collecting data from different sources usually presents many challenges, because different departments will use different styles of record keeping, different conventions, different time periods, different degrees of data aggregation, different primary keys, and will have different kinds of error. So the data must be assembled, integrated in to one unified file which is (Microsoft Access file) in our system to be ready for importing in to the C# environment for other data pre-processing techniques like resolving inconsistency and reduction.

In our proposed system, integration step led to emerging duplicated records (transactions) and inconsistent attributes which are processed in the data pre-processing phase by applying proposed algorithms of reduction and consistency *techniques* that are (Removing Duplication (Reduction) Algorithm) and (Resolving Inconsistency Algorithm). The cleaned and prepared data from pre-processing phase are loaded into the DW which is a wide data store of the market that contains historical data and complete information about building items and has capability of modifying its data and ready for processing phase. In order to mine vast amounts of data in the DW for discovering knowledge, part of the data should be selected and customized in the *Data Selection phase*, where we use the concept of data mart to select and customize the data for processing phase depending on the technique used for knowledge discovery.

In *Data Selection phase* the set of items is selected for DM and as input of the proposed (Index-based Apriori Algorithm) because the used technique is *DM* and specifically the *Association* functionality. In the *discovering knowledge phase*, we use DM and apply its *Association* functionality. The selected set of items is entered to the proposed algorithm (Index-based Apriori) for mining association rules. The number of mining association rules are different based on specified and entered min. count threshold for generating supported item sets and min. confidence threshold for generating interesting association rules. The market manager to be able of taking decisions and managing the market resources, these rules must be interpreted for discovering knowledge to support the process of decision making.

In the Association Rules Interpretation phase, we proposed and used an algorithm named (Association Rules Interpretation Algorithm) applying a simple statistical method which is represented by substituting and counting the items in the antecedent and consequent of the association rules. The results of this system represent the discovered knowledge which is the predicted ratios of items sales for the next year. *The results and visualization phase* which we explain and discuss in the next section, visualizes the results graphically using Line Chart tool to provide the decision maker or the market manager with conceptual values (knowledge) supporting him in managing the market easily and in a perfect way. Fig. 3 illustrates the flow chart of the proposed system.

VII. Research Methodology

A suitable recommendation to users is always related to extraction of suitable regulations that exist in data base. On the other hand, extraction of such suitable regulations is related to a suitable threshold which is specified by the user. Many algorithms, similar to Apriori algorithm, have been suggested for extraction of regulations.

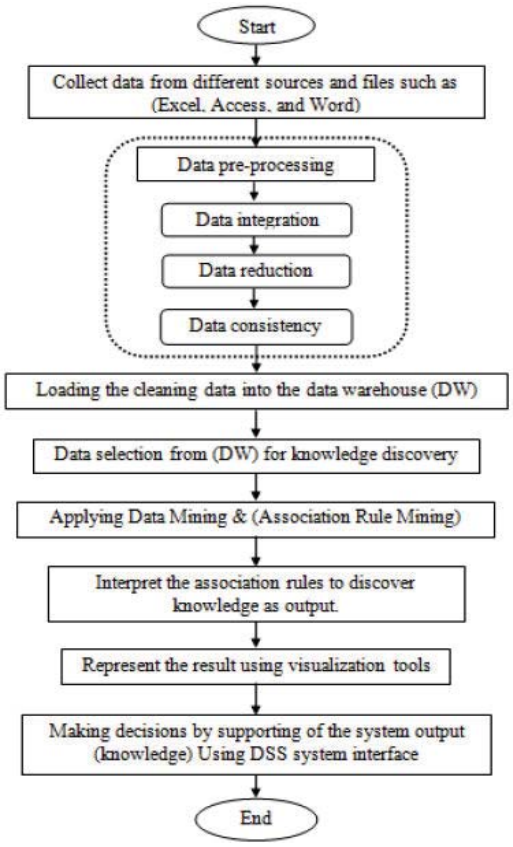

Fig. 3 The flowchart of the proposed system

The Apriori algorithm assumes that the user can specify a suitable threshold as the minimum supporting domain for regularity. However, search of various databases in real world applications requires enough knowledge of existing data in those databases. It is almost impossible that users adopt a suitable minimum supporting threshold for regularity in database. If the user specifies a small number as the threshold, it will lead to production of a large number of inappropriate regulations, and if the user specifies a large number as the threshold, there is a possibility of losing many interesting regulations.

A large gap exists in all studies and that is none of them consider the fact that sometimes a user can not specify a suitable threshold, so it is necessary to calculate a suitable threshold from existing entries in database automatically and introduce it to the user. This article suggests a fuzzy approach to identify Association Rules that are related to the minimum support of regularity. It also allocates different membership levels to database entries by using the membership function that exists in fuzzy logic, so it is expected to obtain more acceptable results in comparison with the classical Apriori algorithm.

A. Fuzzification

Regarding the explanations mentioned about problems of definite sets in this article, we should define definite data as fuzzy data without losing any information. Therefore, we use fuzzy clustering to insert definite data into fuzzy sets. Among the existing algorithms for fuzzy clustering, we use the favorite fuzzy clustering algorithm C-means This algorithm receives the set of definite data as an input from the user. Then asks the user to enter number of clusters and after that distributes all data with different membership levels between clusters.

1. Producing Homogeneous Fuzzy Association Rules

After transferring data to fuzzy environment, we try to extract homogenous fuzzy Association Rules from these data by Apriori approach. As we mentioned before, two measurement scales called "supporting domain of regularity" and "level of reliance on regularity" are used to extract homogenous Association Rules in fuzzy environment.

2. Supporting Domain of Fuzzy Regularity

At first, to identify number of entry repetition in a fuzzy database, we sum up membership levels for each entry and introduce it as the number of entry repetition as:

$$\text{Fuzzy sum} = \sum_{i=1}^{n} \mu(x) \tag{1}$$

Then to produce a set of repeated entries, we should compare all membership levels of every 2 entries and choose the minimum one and finally, introduce sum of these minimums as the number of entry repetition of those 2 entries. Then to produce set of 3 entries, we should compare all membership levels of 3 entries and choose the minimums one and finally introduce sum of these minimums as the member of entry repetition of those 3 entries.

We continue this process until we cannot produce any new sets with multi repetition.

$$\text{Fuzzy sup } (A \to B) = \sum_{i=1}^{n} \min(f_A(x), f_B(x)) \tag{2}$$

3. Level of Reliance on Fuzzy Regularity

After extracting fuzzy algorithms with multi repetition we try to produce homogeneous Association Rules which need using measurement scale to measure level of reliance on fuzzy regularity. To calculate level of reliance on fuzzy regulation we use (3). Those Association Rules that can meet the minimum level of reliance on fuzzy regularity are introduced as interesting homogenous Association Rules.

B. Calculation of Minimum Supporting Domain of Regularity by Suggested Technique

We try to make the algorithm introduce a suitable minimum threshold for measurement scale of supporting domain of regularity to user. When exports want to compare a set of data with an index and classify the data that are greater than or less than the index, they use statistical techniques and using formulas such as average, standard deviation, median, mode and variance, they define an index by which they can compare data and extract their favorite information.

Since there is not any suitable index we also use the average formula to define a suitable minimum threshold to extract frequent patterns we should consider that this minimum

threshold should not be very low, because it leads to producing a large number of useless patterns.

The minimum threshold should not be very high, because it leads to losing useful and suitable patterns.

We calculated sum of all membership level and divide it by C-means (using algorithm of fuzzy clustering) to obtain the average of number of entry repetition, then we introduce the result as the minimum supporting domain of regularity.

$$\text{Fuzzy min sup (A, B)} = \frac{\sum[sum(f_x(A), f_x(B))]}{|T|} \quad (3)$$

VIII. Comparing the Suggested Technique with Basic Algorithm

In this part, we compare some of homogenous Association Rules produced by the suggested technique with the basic algorithm presented in this article. As we have used fuzzy logic and clustering technique, all Association Rules have been analyzed more accurately.

Fig. 4 illustrates the number of homogenous Association Rules produced by these techniques.

Since we have defined the minimum threshold for measurement scale of supporting domain of regularity in an automatic way, we claim that we extract all frequent patterns.

Fig. 5 illustrates a comparison between number of frequent entries produced by the suggested technique and different minimum threshold.

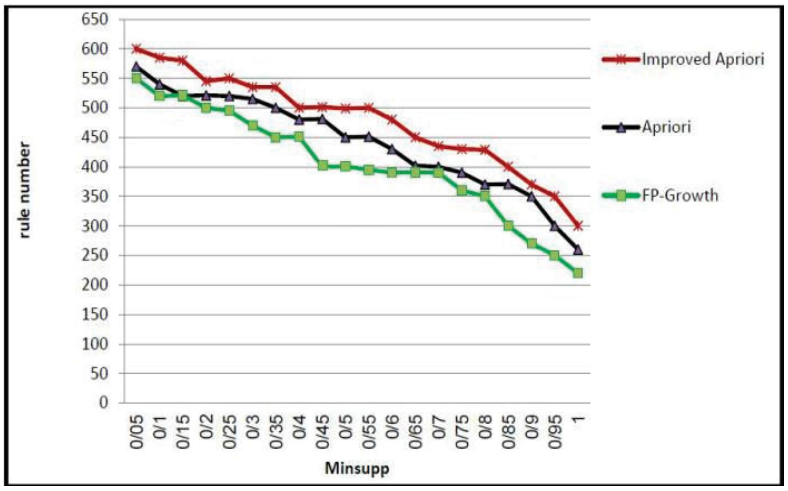

Fig. 4 The number of homogenous Association Rules

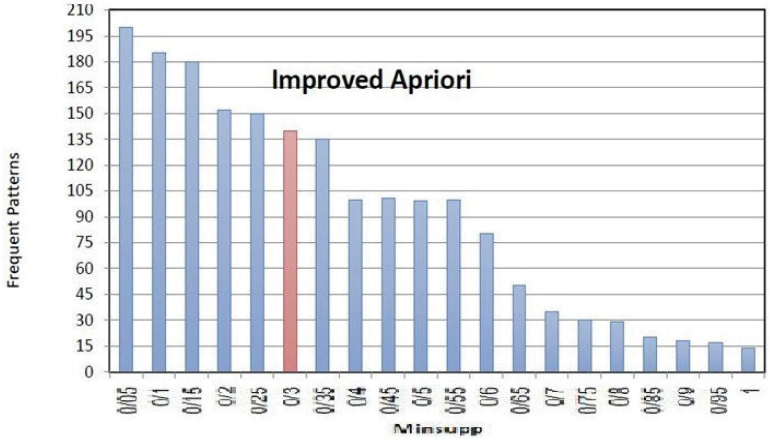

Fig. 5 Improved Apriori

IX. Conclusion and Suggestions

Using DSS system and running the fuzzy algorithm to extract communication Association Rules and coding algorithm of communication Association Rules and obtained results, we obtained the following issues:

1) Running our system, it is clear that KW is less than DW but it is more accurate and more regulated than DW, because the knowledge that is discovered in KW in form of Association Rules or patterns, has obtained from various data in DW.
2) Accuracy of discovered knowledge depends on specified threshold used in fuzzy algorithm. Accuracy of knowledge grows up when the minimum numeric threshold and minimum reliance threshold decrease,

because use of lower threshold will increase number of supported sets and new communication Association Rules. The manager or decision makers to make proper decisions.

3) Decrease in number of data sets will decrease production of communication Association Rules which leads to low quality knowledge.
4) Decrease in production of data set and communication Association Rules will lead to short run time and will occupy less space in the memory. To decrease occupation of memory space and non-time without any decrease in data set and communication regulation, we have used the fuzzy technique to have quick access to information by use of an algorithm based on suggested list.

In this article, a user can not specify a suitable threshold to extract homogenous Association Rules from a big database. We tried to improve Apriori algorithm by suggesting a suitable threshold according to existing entries in database, automatically and using techniques of fuzzy logic. We tried to use fuzzy logic techniques in order to define the dependency between entries in form of membership function and extract all homogenous Association Rules from the database.

Results show that Apriori algorithm has been improved hopefully and extract all homogenous Association Rules.

If we can specify the number of clusters in a systematic and accurate way and define the boundary of each cluster move accurately, we get more accuracy in distribution of data entries between clusters. It leads to production of more interesting homogenous Association Rules. On the other hand, it is suggested to use statistical techniques to specify the minimum threshold of supporting domain of regularity and compare the results with suggested technique in this article.

As the last suggestion, it seems that definition of measurement scale of minimum reliance on regularity in an automatic way is effective in improvement of extracting the Association Rules.

REFERENCES

[1] Hamid R. Nemati, David M. Steiger, Lakshmi S. Iyer, Richard T. Herschel, "Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing", http://www.elsevier.com/locate/dsw, Decision Support Systems, Volume 33, pages 143– 161, 2002 .

[2] Ahmad Bahgat El Seddawy1, Dr. Ayman Khedr2 and Prof. Dr. Turky Sultan, "Adapted Framework for Data Mining Technique to Improve Decision Support System in an Uncertain Situation", International Journal of Data Mining & Knowledge Management Process (IJDKP) Volume 2, Issue 3, Pages 1-9, May 2012.

[3] Hamid R. Nemati, David M. Steiger, Lakshmi S. Iyer, and Richard T. Herschel, "Knowledge Warehouse: An Architectural Integration of Knowledge Management, Decision Support, Data Mining and Data are housing", University of North Carolina at Greensboro, 2009.

[4] Mir Sajjad Hussain Talpur, Hina Shafi Chandio, Sher Muhammad Chandio, Hira Sajjad Talpur, "Knowledge Warehouse Framework", International Journal of Engineering Innovation & Research, ISSN: 2277 – 5668, Volume 1, Issue 3, Pages 262-270, 2012 .

[5] Anthony Dymond, Dymond and Associates, LLC, Concord, CA, "The Knowledge Warehouse: The Next Step Beyond the Data Warehouse", Data Warehousing and Enterprise Solutions \ SUGI 27 \ Paper 144-27, 2008 .

[6] Abdul-Aziz Rashid Al-Azmi, Kuwait University,"Data, Text, and Web Mining for Business Intelligence: A Survey", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.2, March 2013 .

[7] Yongjian Fu, "Data Mining: Tasks, Techniques, And Applications", Potentials, IEEE, ISSN 0278 ,6648 -Volume 16, Issue 4 Pages 18 - 20, Oct/Nov 1997 .
Daniel J. Power, Frada Burstein, and Ramesh Sharda, "Reflections on the Past and Future of Decision Support Systems: Perspective of Eleven Pioneers "\ chapter two, © Springer Science+Business Media, LLC, 2011 .

[8] S. S Suresh, Prof. M. M. Naidu, S. Asha Kiran, "An XML Based Knowledge-Driven Decision Support System For Design Pattern Selection", International Journal of Research in Engineering and Technology (IJRET) ISSN 2277 4378–- Vol. 1, No. 3, 2012 .

[9] Michael Yacci, "The Knowledge Warehouse: Reusing Knowledge Components", ©Performance Improvement Quarterly \Volume 12, Issue 3, pages 132-140, September 1999, provider: citeseer 2008 .

[10] Joseph M. Firestone, Ph.D. Executive Information Systems, "Knowledge Base Management Systems and The Knowledge Warehouse: A (Strawman)", http://www.dkms.com, eisai@home.com, ©1999-2000 Executive Information Systems, Inc., Provider: citeseer 2009 .

[11] Michael Goebel, Le Gruenwald, "A Survey of Data Mining and Knowledge Discovery Software Tools", SIGKDD Explorations. Copyright © 1999 ACM SIGKDD, June 1999, Volume 1, Issue 1, pages 20-33, provider: citeseer .2009

[12] Max Bramer, the book "Principles of Data Mining", Printed on acid-free paper © Springer-Verlag London Limited .2007

[13] CMPUT690, the book "Principles of Knowledge Discovery in Databases"\Chapter I: Introduction to Data Mining, © Osmar R. Zaïane, 1999 .

[14] Rupali, Gaurav Gupta, "Data Mining: Techniques, Applications and Issues", International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE), ISSN: 2277 – 9043, Volume 2, Issue, 2 February 2013 .

[15] Slavco Velickov and Dimitri Solomatine, "Predictive Data Mining: Practical Examples", Artificial Intelligence in Civil Engineering. Proc. 2nd Joint, Workshop, Cottbus, Germany. ISBN 3-934934-00-5, March 2000 .

[16] Radhakrishnan B, Shineraj G, Anver Muhammed K. M, "Application of Data Mining in Marketing", IJCSN International Journal of Computer Science and Network, ISSN (Online): 2277-5420 http://www.ijcsn.org, Volume 2 Issue 5, October 2013 .

[17] Bezdek, J. C. (1998). Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers Norwell, MA, US.