International Journal of Advance Research in Science and Engineering
Vol. No.4, Issue 11, November 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

PERFORMANCE ANALYSIS OF ASSOCIATION RULE MINING ALGORITHMS USING TUMOR DATASET

Subas Chandra Mishra[1], Prof.(Dr.) Subhendu ku.Pani[2]

[1]PG Dept. of CSA, DDCE, Utkal University, Odisha (India)

[2]Dept. of CSE , OEC, Odisha (India)

ABSTRACT

In data mining and knowledge discovery, association rule mining is considered as one of the major basic research topics that discovers interesting relationships between itemsets and predicted the associative and correlative behaviour for new data.. It shows all interesting relationships, called associations, in a potentially large database. However, how interesting a rule is depends on the problem a user wants to solve. Existing approaches employ different parameters to guide the search for interesting rules.The key strength of association rule mining is that all interesting rules are found. The number of associations present in even moderate sized databases can be, however, very large – usually too large to be applied directly for classification purposes. Therefore, any classification learner using association rules has to perform three major steps: Mining a set of potentially accurate rules, evaluating and pruning rules, and classifying future instances using the found rule set.In this work, we make a comparision of association rule mining algorithms. We use two most popular algorithms namely Apriori and filtered Associator using tumor dataset which is available at UCI machine learning repository.

Keywords: Association Rule Mining, Apriori, Predictive Apriori

I. INTRODUCTION

Data mining is considered to be an emerging technology that has made revolutionary change in the information world. The term `data mining' (often called as knowledge discovery) refers to the process of analyzing data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system. Technically, "data mining is the process of finding correlations or patterns among dozens of fields in large relational databases". Therefore, data mining consists of major functional elements that transform data onto data warehouse, manage data in a multidimensional database, facilitates data access to information professionals or analysts, analyses data using application tools and techniques, and meaningfully presents data to provide useful information. According to the Gartner Group, ``data mining is the process of discovering meaningful new correlation patterns and trends by sifting through large amount of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques"[3] . Thus use of data mining technique has to be domain specific and depends on the area of application that requires a relevant as well as high quality data. More precisely, data mining refers to the process of analysing data in order to determine patterns and their relationships. It automates and simplifies the overall statistical process, from data source(s) to model application. Practically analytical techniques used in data mining include statistical methods and mathematical modeling. However, data mining and knowledge discovery is a rapidly growing area of research and application that builds on techniques and theories from many fields, including statistics, databases, pattern recognition, data visualization, data warehousing and OLAP, optimization, and high performance computing [1] . Worthy to mention that online analytical processing (OLAP) is quite different from data mining, though it provides a very good view of what is happening but cannot predict what will happen in the future or why it is happening. In fact, blind applications of algorithms are not also data mining. In particular, ``data mining is a user-centric interactive process that leverages analysis technologies and computing power, or a group of techniques that find relationships that have not previously been discovered" [4] . So, data mining can be considered as a convergence of three technologies -- viz. increased computing power, improved data collection and management tools, and enhanced statistical algorithms. Data and information have become major assets for most of the organizations. The success of any organisation depends largely on the extent to which the data acquired from business operations is utilised. In other words, the data serves as an input into a strategic decision making process, which could put the business ahead of its competitors. Also, in this era, where businesses are driven by the customers, having a customer database would enable management in any organisation to determine customer behaviour and preference in order to offer better services and to prevent losing them resulting better business. The data needed that will serve as an input to organizational decision-making process is generated and warehoused. It is being collected via many sources, such as the point of sales transactions, surveys, through the internet logs – cookies, etc. This has resulted in huge databases which have valuable knowledge hidden in them and may be difficult to extract. Data mining has been identified as the technology that offers the possibilities of discovering the hidden knowledge from these accumulated databases. Techniques such as pattern recognition and classification are the most important in data mining [4,5]. The task of recognition and classification is one of the most frequently encountered decision making problems in daily activities. A classification problem occurs when an object needs to be assigned into a predefined group or class based on a number of observed attributes, or features, related to that object. Humans constantly receive information in the form of patterns of interrelated facts, and have to make decisions based on them. When confronted with a pattern recognition problem, stored knowledge and past experience can be used to assist in making the correct decision. Indeed, many problems in various domains such as financial, industrial, technological, and medical sectors, can be cast as classification problems. Examples include bankruptcy prediction, credit scoring, machine fault detection, medical diagnosis, quality control, handwritten character recognition, speech recognition etc. Pattern recognition and classification has been studied extensively in the literature. In general, the problem of pattern recognition can be posed as a two-stage process: Feature selection which involves selecting the significant features from an input pattern Classification which involves devising a procedure for discriminating the measurements taken from the selected features, and assigning the input pattern into one of the possible target classes according to some decision rule. Research efforts dedicated to data mining, which focused on improving the classification and prediction accuracy, have recently been undergoing a tremendous change [6,7]. The continuous development of more and more sophisticated classification models through commercial and software packages have turned out to provide some benefits only in specific problem domains where some prior background knowledge or new evidence can be exploited to further improve classification performance. In general however, related research proves that no individual data mining technique has been shown to deal well with all kinds of classification problems. Awareness of these imperfections of individual classifiers has called for the emergence of careful development and evaluation strategies of data mining classification models. Association rule mining is a widely-used approach in data mining. Association rules are capable of revealing all interesting relationships in a potentially large database. The abundance of information captured in the set of association rules can be used not only for describing the relationships in the database, but also for discriminating between different kinds or classes of database instances. However, a major problem in association rule mining is its complexity. Even for moderate sized databases it is intractable to find all the relationships. This is why a mining approach defines a interestingness measure to guide the search and prune the search space. Therefore, the result of an arbitrary association rule mining algorithm is not the set of all possible relationships, but the set of all interesting ones. The definition of the term interesting, however, depends on the application. The different interestingness measures and the large number of rules make it difficult to compare the output of different association rule mining algorithms. There is a lack of comparison measures for the quality of association rule mining algorithms and their interestingness measures. Association rule mining algorithms are often compared using time complexity. That is an important issue of the mining process, but the quality of the resulting rule set is ignored. On the other hand there are approaches to investigate the discriminating power of association rules and use them according to this to solve a classification problem. This research area is called classification using association rules. It has to deal with a large number of rules.

Therefore, rule selection and rule weighting are essential for these approaches in classification. An important aspect of classification using association rules is that it can provide quality measures for the output of the underlying mining process. The properties of the resulting classifier can be the base for comparisons between different association rule mining algorithms. A certain mining algorithm is preferable when the mined rule set forms a more accurate, compact and stable classifier in an efficient way. In the next section, we provide an overview of data mining concepts, its process, different techniques and their potential applications.In section 3, we describe our study on finding the best set of class association rules for higher predictive accuracy. Finally the paper concludes in section 4 with a glimpse to our future work.

III. TECHNIQUES AND ALGORITHMS

Classification approach can also be used An Empirical Study on Class Association Rules Mining for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. Some commonly used clustering methods are: a) Partitioning Methods b) Hierarchical Agglomerative (divisive) methods c) Density based methods d) Grid-based methods e) Model-based methods 2.1 Association Rules An Association Rule is a rule of the form milk and bread =>butter where 'milk and bread' is called the rule body and butter the head of the rule. It associates the rule body with its head.In context of retail sales data, our example expresses the fact that people who are buying milk and bread are likely to buy butter too. This association rule makes no assertion about people who are not buying milk or bread.We now define an association rule: Let D be a database consisting of one table over n attributes {a1, a2, . . . , an}. Let this table contain k instances. The attributes values of each ai are nominal1. In many real world applications (such as the retail sales data) the attribute values are even binary (presence or absence of one item in a particular market basket). In the following an attribute-value-pair will be called an item. An item set is a set of distinct attribute-value-pairs.Let d be a database record. d satisfies an item set X _ {a1, a2, . . . , an} if X _ d.An association rule is an implication X) Y where X, Y _ {a1, a2, . . . , an}, Y 6= ; and X \ Y = ;. The support s(X) of an item set X is the number of database records d which satisfy X.Therefore the support s(X) Y) of an association rule is the number of database records that satisfy both the rule body X and the rule head Y . Note that we define the support as the number of database records satisfying X \ Y , in many papers the support is defined as s(X\Y) k . They refer to our definition of support as support count. The confidence ˆc(X) Y of an association rule X) Y is the fraction ˆc(X) Y) = s(X\Y) s(X) . From a logical point of view the body X is a conjunction of distinct attributevalue-pairs and the head Y is a disjunction of attribute-value-pairs where X \ Y = ;.Coming back to the example a possible association rule with high support and high confidence would be i1) i2 whereas the rule i1) i3 would have a much lower support value. 2.2 Class Association Rules The use of association rules for classification is restricted to problems where the instances can only belong to a discrete number of classes. The reason is that association rule mining is only possible for nominal attributes. However, association rules in their general form cannot be used directly. We have to restrict their definition. The head Y of an arbitrary association rule X) Y is a disjunction of items. Every item which is not present in the rule body may occur in the head of the rule. When we want to use rules for classification,we are interested in rules that are capable of assigning a class membership. Therefore we restrict the head Y of a class association rule X) Y to one item. The attribute of this attribute-value-pair has to be the class attribute.According to this, a class association rule is of the form X) ai where ai is the class attribute and X _ {a1, . . . , ai−1, ai+1, . . . , an}. The Apriori algorithm [8,14] has become the standard approach to mine association rules. We have adapted it to mine class association rules in the way explained by Liu et al. [9,13].The second algorithm, Predictive Apriori, has been recently proposed by Scheffer [10,12].Both algorithms have their first step in common. They generate frequent item sets in the same way. An item set is called frequent when its support is above a predefined minimum support. Data mining is considered to be an emerging technology that has made revolutionary change in the information world. The term `data mining' (often called as knowledge discovery) refers to the process of analysing data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system. Technically, “data mining is the process of finding correlations or patterns among dozens of fields in large relational databases”. Therefore, data mining consists of major functional elements that transform data onto data warehouse, manage data in a multidimensional database, facilitates data access to information professionals or analysts, analyse data using application tools and techniques, and meaningfully presents data to provide useful information. According to the Gartner Group, ``data mining is the process of discovering meaningful new correlation patterns and trends by sifting through large amount of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques"[3] . Thus use of data mining technique has to be domain specific and depends on the area of application that requires a relevant as well as high quality data. More precisely, data mining refers to the process of analysing data in order to determine patterns and their relationships. It automates and simplifies the overall statistical process, from data source(s) to model application. Practically analytical techniques used in data mining include statistical methods and mathematical modeling. However, data mining and knowledge discovery is a rapidly growing area of research and application that builds on techniques and theories from many fields, including statistics, databases, pattern recognition, data visualization, data warehousing and OLAP, optimization, and high performance computing [1] . Worthy to mention that online analytical processing (OLAP) is quite different from data mining, though it provides a very good view of what is happening but cannot predict what will happen in the future or why it is happening. In fact, blind applications of algorithms are not also data mining. In particular, ``data mining is a user-centric

interactive process that leverages analysis technologies and computing power, or a group of techniques that find relationships that have not previously been discovered" [4] . So, data mining can be considered as a convergence of three technologies -- viz. increased computing power, improved data collection and management tools, and enhanced statistical algorithms.

III. EXPERIMENTAL STUDY AND ANALYSIS

3.1 WEKA Tool

We use WEKA (www.cs.waikato.ac.nz/ml/weka/), an open source data mining tool for our experiment. WEKA is developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art tool for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, feature reduction, classification, regression, clustering, and association rules. It also includes visualization tools. The new machine learning algorithms can be used with it and existing algorithms can also be extended with this tool.

3.2 Dataset Description

Tumor.arff dataset is selected for this work. Dataset contains 18 attributes, one class attribute and 339 instances. It contains total 22 classes of primary tumor. The rest of the attributes indicate the areas from where primary tumors start. The dataset has been collected from the University Medical Centre, Institute of Oncology, Ljuljana, Yugoslavia.

3.3 Results Analysis

The class association rules generated by Apriori algorithm on the original dataset is given in Figure-3.1 and rules generated by Filtered Associator is shown in Figure-3.3.. Result is derived from the below properties[11,15].Minimum Support is 0.1, Minimum Metric is 0.9, Number of Cycles Performed 10.

Apriori:

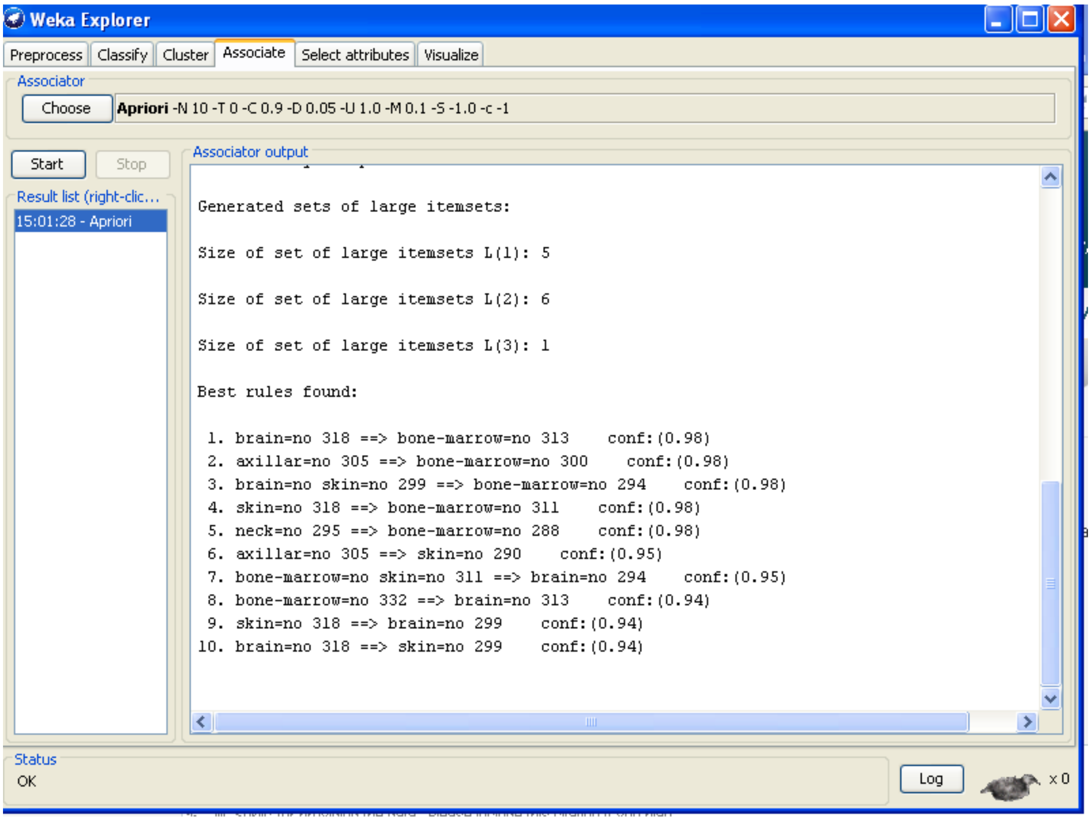

Figure3.1: Rules Generated By Apriori From Original Dataset

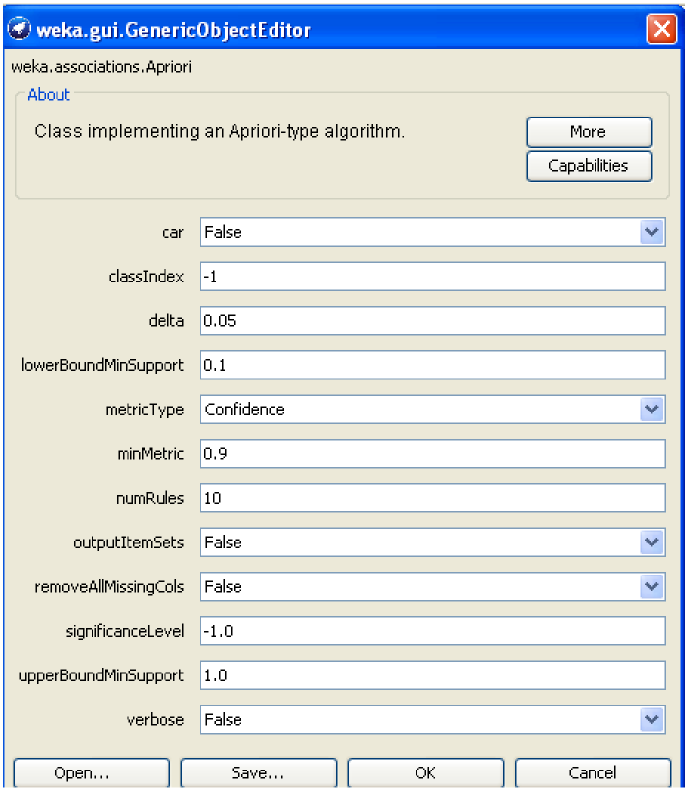

Figure 3.2:Properties of Apriori

Filtered Associator:

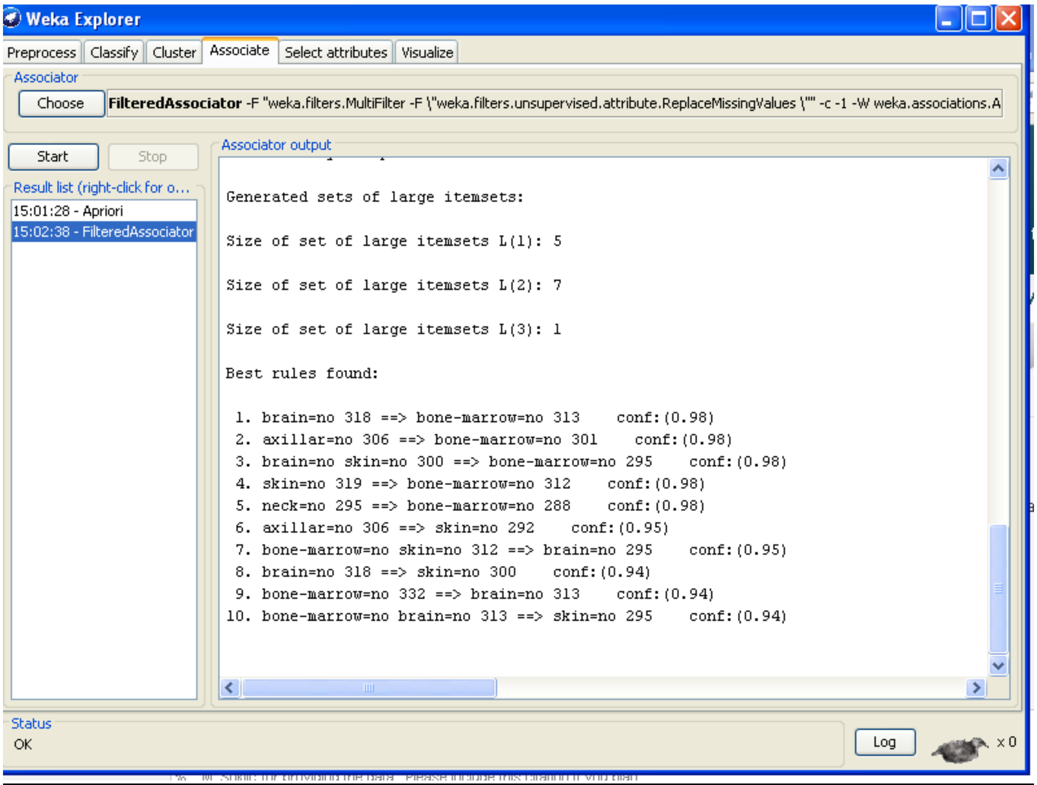

Figure3.3 : Rules generated by Filtered Associator from original dataset

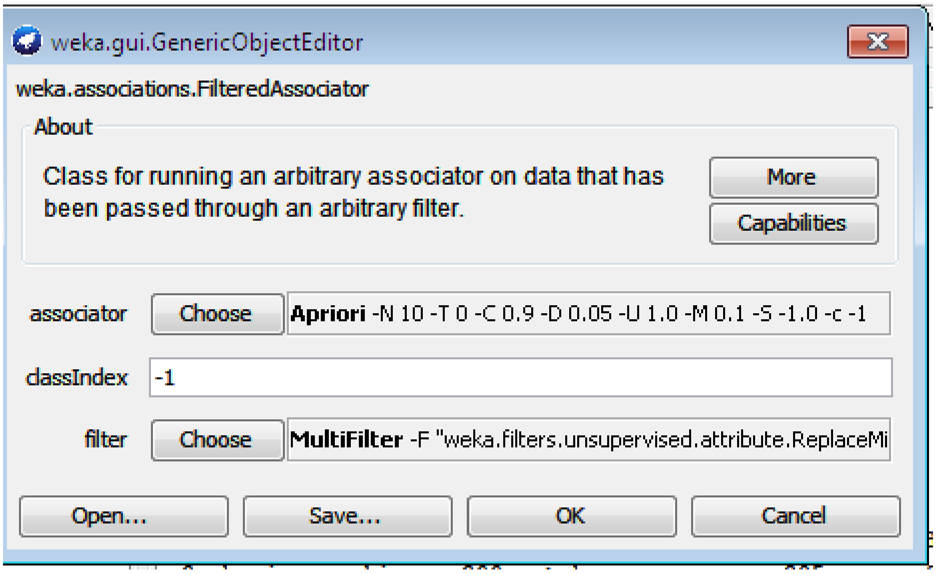

Figure 3.4: Properties of Filtered Associator

IV. CONCLUSION

In this paper we have compared two association rule algorithms i.e. Apriori algorithm and Filtered Associator We have analyzed the frequent itemsets generation and number of cycle performed over the Apriori algorithm and Filter Associator in the context of association analysis.

REFERENCES

[1] Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.

[2] Provost, F., & Fawcett, T., Robust Classification for Imprecise Environments. Machine Learning, Vol. 42, No.3, pp.203-231, 2001.

[3] Larose D T, Discovering knowledge in data: an introduction to data mining, John Wiley, New York, 2005.

[4] Kantardzic M, Data mining: concepts, models, methods, and algorithms, John Wiley, New Jersey, 2003.

[5] Goldschmidt P S, Compliance monitoring for anomaly detection, Patent no. US 6983266 B1, issue date January 3, 2006, Available at: www.freepatentsonline.com/6983266.html

[6] Bace R, Intrusion Detection, Macmillan Technical Publishing, 2000.

[7] Smyth P, Breaking out of the Black-Box: research challenges in data mining, Paper presented at the Sixth Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD-2001), held on May 20 (2001), Santra Barbara, California, USA.

[8] Agrawal R. and Srikant R. Fast Algorithms for Mining Association Rules. In M. Jarke

[9] J. Bocca and C. Zaniolo, editors, Proceeedings of the 20th International Conference on Very Large Data Bases (VLDB’94), pages 475–486, Santiago de Chile, Chile, Sept 1994 . Morgan Kaufmann.

[10] Scheffer T. Finding Association Rules That Trade Support Optimally against Confidence. Unpublished manuscript.

[11] Scheffer T. Finding Association Rules That Trade Support Optimally against Confidence. In L. De Raedt and A. Siebes, editors, Proceeedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD’01), pages 424–435, Freiburg, Germany, September 2001. Springer-Verlag.

[12] UCI Machine Learning Repository, Available at http://archive.ics.uci.edu/.

[13] SAS Institute Inc., Lie detector software: SAS Text Miner (product announcement), Information Age Magazine, [London, UK], February 10 (2002), Available at: http://www.sas.com/solutions/fraud/index.html.

[14] Berry M J A and Linoff G S, Data mining techniques: for marketing, sales, and relationship management, 2 nd edn (John Wiley; New York), 2004.

[15] Delmater R and Hancock M, Data mining explained: a manager's guide to customer-centric business intelligence, (Digital Press, Boston), 2002. Fuchs G, Data Mining: if only it really were about Beer and Diapers, Information Management Online, July 1, (2004), Available at: http://www.information-management.com/ news/1006133-1.html.