Entropy-Based Collective Spatial Data Mining for Simple Sequence in Experimental Machine Learning

R.Kalidasa Krishnaswami [1], Dr. Ashish Chaturvedi[2]
Research Scholar, Department of Computer Science and Engineering,
Himalayan University, Arunachal Pradesh, India,
Professor and Associate Director, Arni School of Computer Science and Applications,
Arni University, Indora (Kathgarh), Himachal Pradesh, India,

***Abstract*—**The tasks of improving access to information is receiving more and more attention specially in modern business The discovered useful information can be used directly by the teacher or the author of the course to improve the instructional/learning performance. This can be an arduous task and therefore educational recommender systems are used in order to help the teacher in this task. In this paper we describe a recommender system oriented to suggest the most appropriate modifications to the teacher in order to improve the effectiveness of the course. We propose to use a cyclical methodology to develop and carry out the maintenance of web-based courses in which we have added a specific data mining step. We have developed a distributed rule mining system in order to discover information in the form of IFTHEN recommendation rules about the web courses. We have used an iterative and interactive association rule algorithm without parameters and with a weight-based evaluation measure of the rule interest.

***Index Terms*—** Data mining, clasification, clustering, Rule mining

I. INTRODUCTION

In recent years we have witnessed a great increased education systems on-line or e-learning systems. More and more centers public or private teaching available to students learning management systems (Learning Management Systems, LMS) web based. the first systems of this type were commercial WebCT Top Class Virtual-U or if, in the Today, increasingly dominated systems charge freely distributed. Data Mining and Knowledge Discovery in Databases is an interdisciplinary field [9] merging ideas from statistics, machine learning, databases, and parallel and distributed computing. It has emerged due to the phenomenal growth of data in all spheres of human endeavor, and the economic and scientific need to extract useful information from the collected data. The key challenge in data mining is the extraction of knowledge and insight from massive databases. Data Mining refers to the overall process of discovering new patterns or building models from a given dataset. There are many steps involved in the KDD enterprise which include data selection, data cleaning and preprocessing, data transformation and reduction, data mining task and

algorithm selection, and finally post-processing and interpretation of discovered knowledge [16]. Typically data mining has the two high level goals of prediction and description [16]. In prediction a model is built that will predict unknown or future values of attributes of interest, based on known values of some attributes in the database. In KDD applications, the description of the data in human understandable terms is equally if not more important than prediction. Two main forms of data mining can be identified [37]. In verification-driven data mining the user postulates a hypothesis, and the system tries to validate it. The common verification-driven operations include query and reporting, multi-dimensional analysis or On-Line Analytical Processing (OLAP), and statistical analysis. Discovery-driven mining, on the other hand automatically extracts new information from data. The typical discovery driven tasks include association rules, sequential patterns, classification and regression, clustering, similarity search, deviation detection etc. While data mining has its roots in the traditional fields of machine learning and statistics, the sheer volume of data today poses the most serious problem. For example, many companies already have data warehouses in the terabyte proportions.

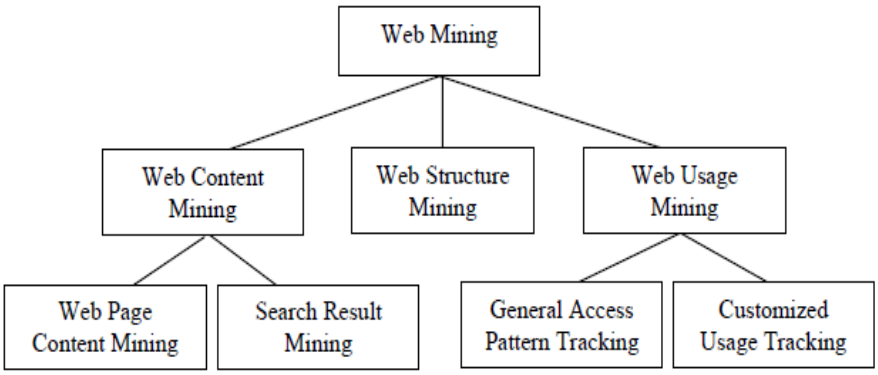

Fig 1. Taxonomy of Web Mining

Data-mining is a tool for increasing the productivity of people trying to build predictive models. Data-mining offers pragmatic values across a broad spectrum of industries. Retail Industries perform data-mining on transactional database for their business promotions. Telecommunications and credit card companies are leaders in applying data-mining to detect fraudulent use of their services. Insurance companies and stock exchanges are also interested in applying this technology to reduce fraudulent activities.

II. EXTRACTING KNOWLEDGE BASES DATA AND DATA MINING

Knowledge discovery in databases data (KDD) is defined as the process of identify meaningful patterns in the data that are valid, novel, potentially useful and understandable to a user [4, 8, 10, 12]. The overall process is to transform information low-level high-level knowledge. The KDD process is interactive and iterative containing the following steps Understand the application domain: This step includes the relevant knowledge and goals prior to application. Remove the target database: data collection, evaluate data quality and use exploratory data analysis to become familiar with them. Preparing Data: includes cleaning, processing, integration and data reduction. It tries to improve the quality of the data while reducing the time required for the learning algorithm applied subsequently. Data Mining: as noted above, this is the critical phase of the process. It consists of one or more of the following functions, classification, regression, clustering, summary, retrieval, rule extraction, etc. Interpretation: explaining the discovered patterns as well as the ability to visualize. Use knowledge discovered: do using the model created.

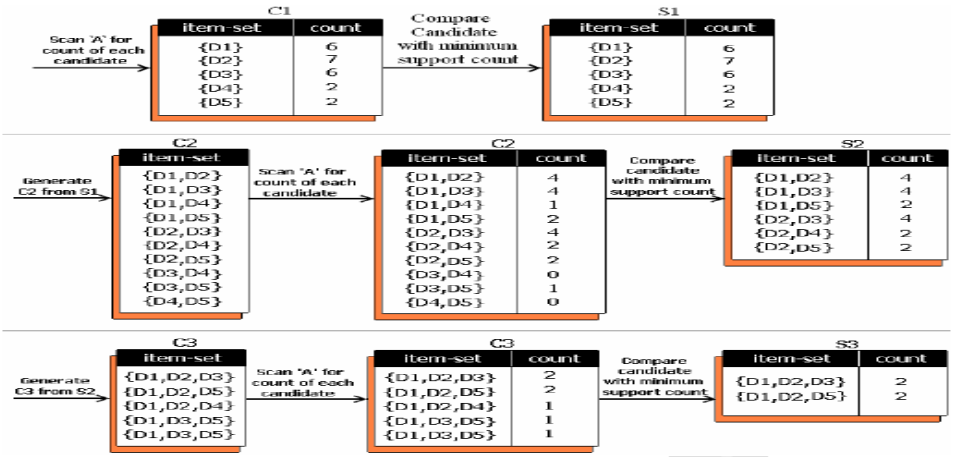

Fig 2. A Priori Algorithm to find the frequent item-set in Database

The key step of the process is marked with number 4. Then discussed briefly the most common tasks of mining data, with an example of use. In the remainder of this article will consider mining data from different perspectives. In the next section bases are provided knowledge discovery and mining data. Section 3 lists some of the frequent applications in business. In the Section 4 briefly describes the techniques most used. Trends in data mining shown in Section 5. And finally it the main conclusions are drawn.

III. METHODOLOGY

Data-mining refer to extending traditional data-mining techniques with the data modified to mask sensitive information. The key issues there of are how to modify the data and how to recover the data-mining result from he modified data. The solutions obtained are often tightly coupled with the relevant data-mining algorithms. Two most important questions related to the release of actual data and the privacy of the individuals are: (i) Can general trends across individuals be determined without revealing information about individuals? And (ii) can highly private associations be extracted from the public data? In the former case, there is a need to protect individual data values while revealing associations or aggregation. In the latter case, there is a need to protect the associations and correlations between the data. In this research work, transformation methods are proposed for preserving the privacy of the individuals. Two different types of transformation methods are proposed for two distinct data types. It is proposed to follow a categorical grading based transformation for numerical sensitive data and mapping-table based transformation for alphanumeric nominal sensitive data. It is proposed to develop a tool that performs the categorical grading and mapping-table based transformations on the micro data table. The transformed or publishable table can be released for research purpose without any information loss. Any data-mining algorithm can be applied on the released table without any modification in the algorithm. And the results hence obtained are, as if mined from the original micro data table. Yet the privacy of the individual is preserved across the released table as well as in the mining results obtained. Also, in real-life data publishing, a single organization often may not hold the complete data. Organizations need to share data for publishing to a third party for analysis. Data-mining performed on this kind of data is called collaborative data-mining. Mining this kind of collaborative data should preserve privacy of individual organization without disclosing sensitive information to other organizations involved in the collaboration. The proposed methods for micro-data release can be used for collaborative data mining also. Two different frame works are proposed to preserve the privacy of the horizontally and vertically partitioned data.

The resulting association patterns which predict user behavior are used for the construction of web personalization. Then, personalization's are generated by certain rule generation algorithms, such as association rule algorithms, clustering algorithms, classifications and so on. The amount of personalization can be huge when first generated; therefore, post-processing for the generated results is performed in this stage. There are several rule evaluation techniques available towards rule post processing of rules The rule interestingness measure is used to find important rules and to rank the rules. In an online shopping application, individuals' online purchasing patterns and online browsing experiences may be personalized as well. Such personalization is helpful to predict customers' interests and to recommend relevant advertisements of interested products to facilitate customers' online shopping experiences. However an online web user normally browses hundreds of web pages before making a purchase online, and different users visit different websites. Personalization based on other people's past histories may not be very interesting to another user.

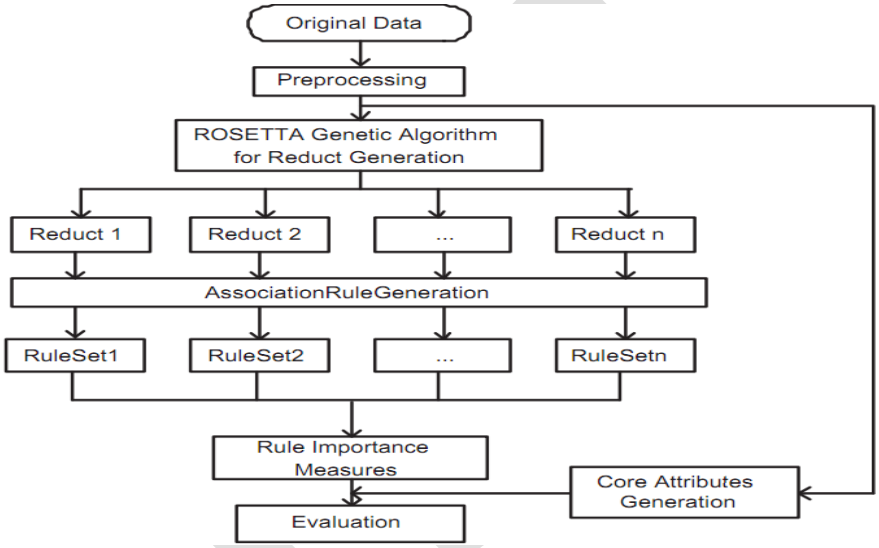

Fig 3. Experimental procedure

Data Mining is the process of analyzing data from different perspectives and summarizing the results as useful information. It has been defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

The process of data mining uses machine learning, statistics, and visualization techniques to discover and present knowledge in a form that is easily comprehensible. The word "Knowledge" in Knowledge Discovery Database refers to the discovery of patterns which are extracted from the processed data. A pattern is an expression describing facts in a subset of the data. Thus, the difference between KDD and data mining is that "KDD refers to the overall process of discovering knowledge from data while data mining refers to application of algorithms for extracting patterns from data without the additional steps of the KDD process".

IV. RESULT

Data mining could be inferred as building a model set to about data provides an understanding. Therefore we can distinguish two steps in a task MD, first the choice of the model, other end of it to fit the data. The choice of model is determined basically by two factors: the type of data and the objective to be obtained.

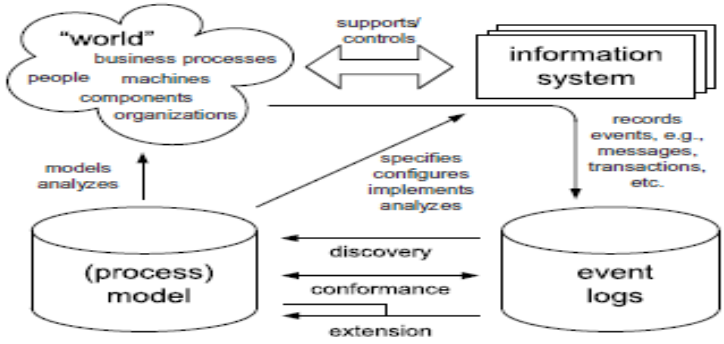

Fig 4. Schematic overview of process mining.

Thus by example would not be appropriate to apply regression to or consist of text-based data models distance symbolic data. Regarding the relationship target model, literature presents a catalog of different models for different objectives. So, if you have a problem classification is used vector machines support or decision trees, if it is a problem regression can be used or regression trees neural networks, if it wishes to clustering You can opt for hierarchical models or interrelated, etc. Level is also important in this election understandability to be obtained from the model end, as there are easy models to "explain" the Users such as association rules and clear difficulties involving other networks as or neuronal support vectors. The second step is to perform a "phase learning "with the data available to adjust the previous model to our particular problem. So if we have a neural network will have to define its architecture and adjust the values of the weights of its connections. If we are to obtain a regression line must find the values of the coefficients and if we use the k-nearest neighbors we need to set a metric and k, etc.

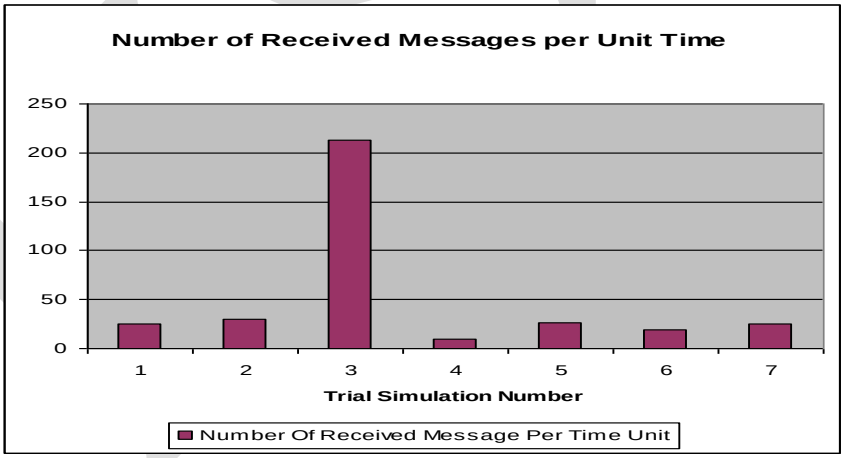

Fig 5. High e-mails messages can potentially trigger false virus alert.

This learning phase looking squares pattern values that attempt to maximize the "goodness" thereof. This question brings us back to raise two problems: How do you define a goodness model for some data? And two, how to perform this search? Regarding the first, normally base any model must be accompanied by a function of adapted to be capable of measuring the setting (in English concept of fitness function is used). This is easy in many cases, for example in classification or regression problems, however can pose serious challenges in others such as clustering.

V. DISCUSSION AND CONCLUSION

This work may be further extended to improve the performance of systems that involve real time data in the form of audio, video and other multimedia objects which are stored as data items in a database with valid time constraints i.e. a temporal database of transactions containing complex form of It would be helpful to compose fixed length vectors in data preparation process. This enables executing various prediction tasks to understand user's behavior. Therefore data analyst can identify situations typical for browsing termination and succession, and web site designer can reorganize web site information to be more convenient for web site visitors. The rule evaluation approach discussed Data. Automatic assignment of membership function could be considered as enhancements to the future method. This approach will yield important results for the decision makers in the fields where customer preferences are strongly interrelated with the time. Other problems for future research discovery of Association Rules, and determining the right time interval size which is application dependent

Rule templates can be used towards the construction of web personalization as rule interestingness measures. The rule templates can be used during the rule generation process to limit both the type of rules expected and the quantities of rules. This approach is a subjective rule interestingness measure, which can be combined together with other rule interestingness measures for rule evaluation purposes.

A rule Importance Measure which is an automatic and objective approach extracts and ranks important rules. The Rule Importance Measure differentiates rules by indicating which rules are more important than other rules. One can use the Rule Importance Measure upon the data set directly, and obtain a list of ranked rules by their importance. Evaluations with Human Users. To study effectiveness of these measures for rule evaluation it is planned to perform experiments with human users who are experts in the domain. Certain user satisfaction studies may be conducted for real people's evaluations with appropriate measures from across a sufficiently large sample of users in a restricted domain.

Integration of ontology knowledge of Web pages into Web recommendation. The current research is mainly based on analysis of Web usage knowledge, not taking other Web data sources into account. With the development of semantic Web and ontology research, it is believed that ontology knowledge of Web pages can provide deeper understanding or semantic linking of Web pages as a result of conveying the conceptual information.

Employing the latest progress of other related research areas into Web data management. The successes and contributions from data mining, machine learning, information retrieval domains always bring in new data models and algorithms to Web data research. It is believed these progresses will produce a big potential for Web researchers to address the open research problems not solved yet.

REFERENCES

[1] Liu, D.R., and Liou, C.H., 2011, Mobile commerce product recommendations based on hybrid multiple channels, Electronic Commerce Research and Applications, Vol. 10. No. 1, pp. 94-104, doi:10.1016/j.elerap.2010.08.004.

[2] Ismail, R., Othman, Z., and Bakar, A.A., 2010, Associative prediction model and clustering for product forecast data, Intelligent Systems Design and Applications, 10th International Conference, pp. 1459-1464.

[3] Gang Cui., 2010, A methodologic application of Customer Retention based on back propagation Neural Network prediction, Computer Engineering and Technology, 2nd International Conference, V3-418, V3-422.

[4] Yeh, J. and Hsu, P. "HHUIF and MSICF: Novel Algorithms for Privacy Preserving Utility Mining", Expert Systems with Applications, Vol. 37, No. 7, pp. 4779-4786, 2010.
[5] Koh, Y.S., Pears, R. and Yeap, W-K. "Valency based weighted association rule mining", PAKDD No. 1, pp. 274-285, 2010.
[6] Zhou Li., Wu Qi Zong., 2011, Study of influence factors of customer satisfaction based on BP neural network, Software Engineering and Service Science (ICSESS), IEEE 2nd International Conference, pp. 409-411
[7] Rad, A., Naderi, B., and Soltani, M., 2011, Clustering and ranking university majors using data mining and AHP algorithms: A case study in Iran, Expert Systems with Applications, Vol. 38, No. 1, pp. 755-763.
[8] Erkan Bayraktar., Ekrem Tatoglu., Ali Turkyilmaz., and Dursun Delen., 2011, Measuring the efficiency of customer satisfaction and loyalty for mobile phone brands with DEA, Expert Systems with Applications, Vol. 39, No. 1, pp. 99-106.
[9] Koh,Y.S., Pears, R. and Dobbie, G. "Automatic assignment of item weights for pattern mining on data streams", PAKDD No.1, pp.387-398, 2011.
[10] Dawn, E. Holmes., Jeffrey Tweedale., and Lakhmi Jain, C., 2012, Data Mining Techniques in Clustering, Association and Classification, Data Mining : Foundations and Intelligent Paradigms, Intelligent Systems Reference Library, Springer, Vol. 23, pp. 1-6.
[11] Beomsoo Shim., Keunho Choi., and Yongmoo Suh., 2012, CRM Strategies for A Small-Sized Online Shopping Mall Based on Association Rules and Sequential Patterns, Expert Systems with Applications: An International Journal, Vol. 39, No. 9, pp. 7736-7742.
[12] Glady, N., Baesens, B., and Croux, C., 2009, A modified Pareto/NBD approach for predicting customer lifetime value, Expert Systems with Applications, Vol. 36, No.2, pp. 2062-2071.
[13] Sublaban, C.S.Y., and Aranha, F., 2009, Estimating cell phone providers' customer equity, Journal of Business Research, Vol. 62, No. 9, pp. 891-898.
[14] Ngai, E.W.T., Li Xiu and Chau, D.C.K., 2009, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications, Vol. 36, No. 2, pp. 2592-2602
[15] Mishra, A., and Mishra, D., 2009, Customer Relationship Management: Implementation Process Perspective, Acta Polytechnica Hungarica, Vol. 6, No. 4, pp.83-99.
[16] Huang, S.C., Chang, E.C., and Wu, H.H., 2009, A case study of applying data mining techniques in an outfitter's customer value analysis, Expert System with Applications, Vol. 36, No. 6, pp. 5909–5915.
[17] Ching-Hsue Cheng., and You-Shyang Chen., 2009, Classifying the segmentation of customer value via RFM model and RS theory, Expert Systems with Applications, Vol. 36, No. 3, pp. 4176-4184.
[18] Chang, P.C., Liu, C.H., and Fan, C.Y., 2009, Data clustering and fuzzy neural network for sales forecasting: A case study in printed circuit board industry, Knowledge-Based Systems, Vol. 22, No. 5, pp. 344-355.
[19] Cao, Q., and Parry, M.E., 2009, Neural network earnings per share forecasting models: A comparison of backward propagation and the genetic algorithm, Decision Support Systems, Vol. 47, pp. 32-41.