# Research on Applications of Data Mining in Electronic Commerce

Xiuping YANG[1, a]

[1] Computer Science Department, Guangdong AIB Polytechnic College, Guangzhou 510507, China

[a]yangxiuping@126.com

**Keywords:** Data Mining; Electronic Commerce; Logistic Regression; Service Evaluation

**Abstract.** With the fast development in the subject of computer science and technology, the combination of machine learning algorithms and electronic business is needed. There may be some unpredictable but frequent problems such as delay in shipment, shipping errors caused by E-commerce participants' low efficiency. There are problems will have a negative impact on enterprises participants end up. The efficiency of e-commerce is an important way for a proper evaluation of improving management. In this paper, we propose the theory of knowledge mining based on Rough Set Theory to handle the vague and inaccurate information about the evaluation of supplier and mine the law knowledge that exists between input variables and adverse position. The RST output is then used as the feature and sent to the Logistic regression (LR) to the electronic commerce website product grade. The proposed method, called RST-LR, the discretization process by the attribute values; the minimum attribute set filtering; evaluation criteria; the establishment of calculation accuracy of ranking and assessment system. We simulate and experiment the algorithm and illustrate the accuracy.

## Introduction

According to the fact that wide application of computer networks and the rapid development of the Internet, therefore, more and more people access the internet and the great improvement of internet transaction security technology, E-commerce has gradually been accepted by the public and it has become a mainstream business model [1]. At the same time, through business to business applications (business to business) and (business to consumer) business to consumer electronic commerce practice, many companies realize their dreams of opening an online store, paid for by real time online, supply chain management (SCM) and other mechanisms. They can manage their logistics and capital efficiency, and to provide a safe and substantive online transaction environment, so as to attract more people to shop online. The global e-commerce online shopping market is increasing year by year. Data mining is an interdisciplinary topic. According to the different requirements of specific tasks, we can use statistical methods, neural networks, online analytical processing, genetic algorithms and decision trees, rough sets and so on. (1) Statistical analysis methods. Statistical analysis methods mainly refer to the relevant statistical methods for data correlation analysis, regression analysis, cluster analysis and principal component analysis. (2) Neural network based methods. Neural network is similar to the data processing method which caused by simulating human brain neurons process information, which is more common Back Propagation (BP) neural network. (3) Online processing method. Online processing and traditional online transaction processing (OLTP) is different, is mainly used to deal with affairs, such as the civil aviation calibration system, bank storage systems, etc. (4) Genetic algorithm (GA) is a kind of evolutionary algorithm, its basic principle is evolution simulation of biological, coding parameter problem, mark them as chromosome, and then use an iterative method to select, crossover and mutation, and ultimately meet the optimization objectives have chromosome. (5) Decision tree algorithm is a kind of inductive learning methods.

The main purpose of this paper is for the service quality and efficiency evaluation of E-commerce, and we introduce RST theory in the course. The effective use of mining knowledge rule set and Logistic regression, we can provide appropriate counseling service quality and efficiency, the ultimate ecommerce. In the experiment, there are many procedures as follows: data collection, data preprocessing, discrimination, attribute reduction, reduction filtering, rule generation, rule filtering, identification prediction, accuracy calculation and diagnostic system. And the results show that web data mining can greatly improve the service quality and efficiency when used in r-commerce.

## Our Proposed Methodology

**General Structure of Proposed Algorithm.** The main purpose of this paper is to evaluate the service quality and the service efficiency of E-commerce, and we introduce RST theory in the course. In the folowing we use flowcgat to discribe the procesure[3].

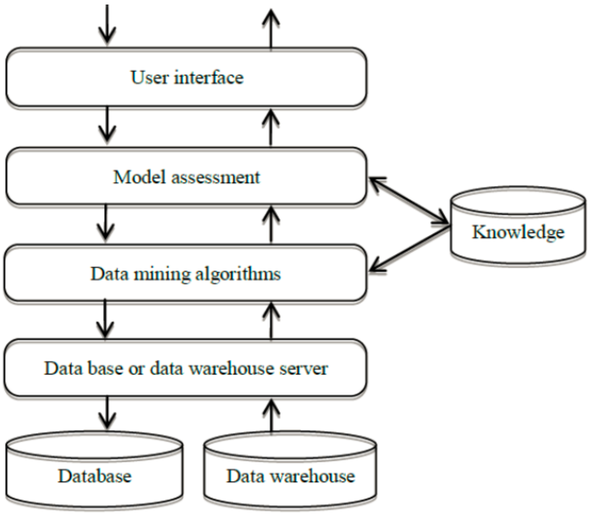
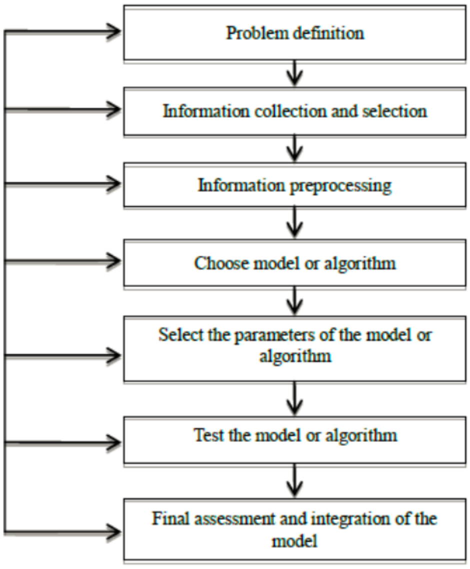


Fig. 1 The General Web Mining



Problem definition
Information collection and selection
Information preprocessing
Choose model or algorithm
Select the parameters of the model or algorithm
Test the model or algorithm
Final assessment and integration of the model

Fig. 2 The Proposed Methodology

**Data Collection and Pre-processing.** UCI database is maintained by the University of California at Irvine California Irvine, which are widely used in machine learning Our Proposed Model and Analysis. The database is currently a total of 264datasets, and the number is still growing. UCI data set is used standard test data set. This is entirely the design standards of international famous. In addition, combining with the actual situation China financial market, puts forward the empirical study, the completion of the entire design guide. In this paper, we use the Amazon business review data set UCI as the main data resource. The data set is used for on-line write-spring is in pattern recognition field a new author identification. The characteristics of the dataset are plural. The data sets are from the Amazon e-commerce website customer identification comments. Previous studies conducted experiment in two to ten author identification parts. But in the online context, reviews to be identified usually have more potential authors, usually recognition algorithm is not suitable for the target class a lot. Robust detection and recognition algorithm, we identified the 50 most active user (through a unique ID and username donation) posting comments often in these news group. We collected 30 each author's comment number. This paper used the database and by using the Rough Set it can deal with the vague and imprecise information during the process of evaluation of electricity.

**Rough Set for Feature Representation.** When the rough set approach deals with data, it defines the table of information as below $S=<U,A,V,F>$ where $U=\{x_1,x_2,\cdots x_{10}\}$ is for search set. For example: $\{\text{weight},\text{gender},\cdots,\text{blood}-\text{type}\}$, $A=\{a_1,a_2,\cdots a_3\}$ is the attribute set. The information function expression is $f:U\times A\rightarrow V$. It also can be described as $f(x,a)\in V_a$. Let $S=<U,A,V,F>$, and $P\subseteq A$ is the subset [2]. So the relationship without distinction must meet the following formula:

$$f(x,a)\in f(y,a),\quad a\in P \tag{1}$$

Based on the above definition, the lower limit can be defined as $PX_{down}=\left\{x_i\in U\,|\,[x_i]_{Ind(P)}\subset X\right\}$.

The critical value is described as:

$$PNX = PX - PX_{down} \tag{2}$$

There are four comments relating to this formula. If X meets the following $PX_{down} \neq \varphi$, X can be considered as RST element in the data set. The calculation of identification error rate is:

$$\mu_p(X) = \frac{card(PX_{down})}{card(PX)}, \quad 0 \leq \mu_p(X) \leq 1 \tag{3}$$

## Experiment Analysis and Simulation

This section will empirically validate our proposed RST-LR based on rough set theory (RST) and logistic regression (LR) for product ranking. The experiment procedures are as follows. First, collect data according to experiment design; second, extract feature from the processed data; third, model training and test. The experimental steps are shown in Figure 3.

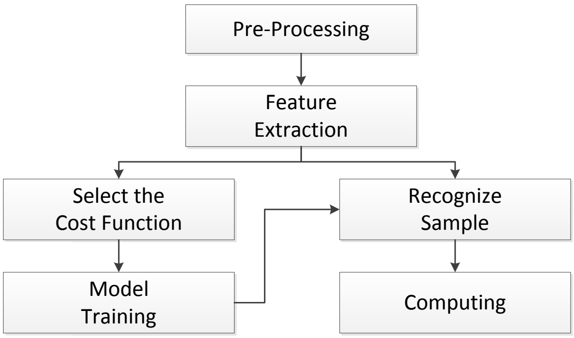


Fig. 3 The Flowchart of the Proposed Method

**Experimental Dataset.** The Amazon Commerce reviews data used in our experiments was collected by National Engineering -Center for E-Learning. The data sets are from the Amazon e-commerce site identification customer reviews. It is used to identify the author reviews from amazon. Most previous works carried out two to ten author identification experiment. But in the network environment, there are usually more potential authors to be recognized, and the recognition method is usually is not suitable for the target class of a lot of. The data set is from customer reviews on Amazon e-commerce site identification. The data set includes 10000 properties of 1500 samples, used to identify the author reviews from amazon. Through it, we divide the data set into two parts, 1000 samples of the training data set, the prediction data of 500 sample sets.

**The Evaluation Criterion.** To validate the advantages of the proposed approach, we employ the identification accuracy as the evaluation criterion which can be defined as follows, where TP indicates the true positive; TN indicates the true negative; FP denotes false positive; FN represents false negative.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{4}$$

**The General Result.** In the first experiment, we take advantage of the RSTLR approach for product ranking, and use the Amazon Commerce reviews collected from Amazon Commerce Website for authorship identification as the samples to run experiment. In this experiment the data set was collected by the national network of Engineering Research Center of distillate. The data set was collected from customer reviews on Amazon business identification. It uses the accuracy and precision as the evaluation standard to verify the product ranking RST-LR effect. In the experimental process uses standard methods to determine the parameters of RST-LR. Then use the RST-LR operation of the trained recognition. In the experiment, the parameters of RST-LR are configured as the default. In the experiment, it targets to validate the advantage and robustness of RST-LR method for product ranking, in comparison with others. The Amazon Commerce reviews is collected from Amazon Commerce Website for authorship identification and is randomly partition to training set and test set. The dataset are collected from the customer's reviews in Amazon Commerce Website for authorship identification.. We take advantage of accuracy and precision as the assessment criterion for assessment. As show in previous section of this paper, the parameters of RST-LR is solve using the standard algorithm. The learn RST-LR to is then employ to run identification. The test is performed for multiple rounds, with default set. The reasons for these results are mainly three aspects. (1) The RST-LR method can be applied to the conditions that sample data is large scale, complex dimension, containing a large number of heterogeneous information. (2) The learning method is according to the data distribution of the input data to select the model parameters of the RSTLR, which makes the RST-LR having better adaptability. (3) The framework of the proposed algorithm is composed of some comprehensive procedures which sequentially maximizes the identification ability. The result is shown in the figure4.

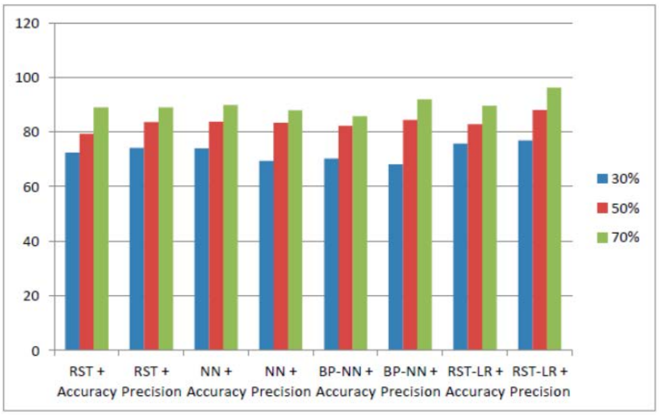


Fig. 4 The Experimental Illustration

## Summary

The quality of the E-commerce service efficiency has relationship with the survival of E-commerce, so this paper is about RST based E-commerce service efficiency evaluation system. But in the application of the model, there are some aspects should be considered. There are several simplifying attribute methods. Here, we adopt the latter genetic algorithm, the design principles for the fitness function of the genetic algorithm is the rate of the cases included in the attribute set. The answer got from the simplified attribute is not unique; we can obtain many minimal set of attributes. But which set should be adopted, should not only compare the accuracy of the identification results but also refer to the advice of experts in the field. We can also filter rules if it is necessary. The principles whose ability of predicting is less should be filtered, and the others are kept. In the future, we plan to use deep learning algorithms to optimize the proposed method. For example, Wang [4] use the deep learning method to classify the data.

## References

[1] Chen, F. -L. and Liu, S. -F., "A Neural-Network Approach To Recognize Defect Spatial Pattern In Semiconductor Fabrication", IEEE transactions on semiconductor manufacturing, Vol. 13, No. 3, pp. 366-373, 2010.

[2] Han, J, and Kamber, M., Data Mining: Concepts and Techniques, San Francisco, CA: Morgan Kaufmann Publishers, 2001.

[3] Johnson, D. S., "Approximation Algorithrns for Combinatorial Problerns." Journal of Computer and System Sciences, Vol. 9, pp. 258-278, 1974.

[4] Haoxiang Wang. "CLASSIFYING GRAY-SCALE SAR IMAGES: A DEEP LEARNING APPROACH", Machine Learning and Applications: An International Journal (MLAIJ) Vol.1, No.1, September 2014.