International Journal of u- and e- Service, Science and Technology
Vol.7, No.1 (2014), pp.9-18
http://dx.doi.org/10.14257/ijunesst.2014.7.1.02

Data Mining Research on Time Series of E-commerce Transaction

Xiao Qiang[1,2], He Rui-Chun[1] and Liao Hui[2]

[1]*School of Traffic and Transportation, Lanzhou Jiao tong University, Lanzhou, China*
[2]*School of Economics and management, Lanzhou Jiao tong University, Lanzhou, China*
Lzjt_xq@126.com, herc@mail.lzjtu.cn, lzjt_liaohui@126.com

Abstract

In E-commerce, data mining can help the online customers accurately grasp the sellers' product sales, to improve the online product purchase rate. In this paper, the mining algorithm of E-commerce transaction based on time series is proposed, which analyzes the relationship between the density of E-commerce transaction recorders and product sales records in E-commerce sites by use of the method of Guass density function and sliding-window. To examine the approach by MATLAB, illustration is provided to demonstrate the effectiveness of the algorithm

Keywords: *data mining, E-commerce, time series, density*

1. Introduction

E-commerce development accelerates the E-commerce commodity boom. According to statistics, turnover of E-commerce has reached 7.85trillion Yuan in china, up by 30.83% percent year-on-year, turnover of nets retail market has reached 1.3205 trillion Yuan, up by 64.7% percent year-on-year, the number of online shoppers have reached 247million people, up by 21.7% percent year-on-year, the number of online shops have reached 13.65 million [1].

It can be seen from E-commerce transaction in China, the number of net purchase products and transaction turn over are also increase by degree in year after year, net purchase have become popular consumption.

Yet there were an enormous growth in the networking products information, it has made the net purchase the difficult to the online shoppers, only rely on credit rating and customer evaluation which were provided by the shop [2]. Customer rapid analysis and cognitive related information from the amount information, it needs to solve imminently in net purchase [3], so if the problem is solved, the net purchase rate will be promoted.

In order to change the present situation that network customer purchased the goods depend on the credit of seller and evaluation of buyer. In this paper, we proposed data mining research on time series of E-commerce transaction. Through analysis the shop transaction recorders, we can be master of the transaction and sales about the shop, which can provide purchase basis, and help seller to improve sales rate.

The paper is organized as follow, in the next section we present some related research to E-commerce trade with analytics and the associated problems with time series. Section 3 describes model structure design and the proposed algorithm. Section 4 then present details of our approach including these problem: getting data, preprocessing data and calculating data

ISSN: 2005-4246 IJUNESST

and analysis data. Section 5 presents the evaluations for method with an experiment and its test result, and finally section 6 presents summary and conclusion this stray.

2. Related Works

In recent times, the increasing use of time series data has activated various researches in the field of data and knowledge management [4, 5]. Time series data are described as large, with high dimensionality and that needs continuous update. Moreover, the time series data are usually considered as a whole instead of individual numerical fields. Time series research includes these tasks such as indexing, classification, clustering and representation of time series.

Time series have a widely used in fields of signal processing, automatic control, financial, meteorological, information management, hydrological and so on in recent years. It is an important that time series were process in real life, such as stock analysis [6], products sales analysis [7], bank data analysis, seismic data analysis, traffic data analysis [8] and so on. Through the analysis of time series, we can reveal the inherent law of change and development, to cognize thing and make scientific decision, which have important significance [9].

Based on the time series analysis, different mining tasks can be found in the literature and they can be roughly classified into four areas: pattern discovery and clustering, classification, rule discovery and summarization [10]. Some research issues concentrate on one of these areas, while the others may focus on more than one of the above processes. The fundamental problem is how to represent the time series data [11, 12]. Mostly, there are many kinds of time series data related research, such as finding similar time series , subsequence searching in time series [13], dimensionality reduction and segmentation [14] .One of the common approaches is transforming the time series to another domain for dimensionality reduction [15] followed by an indexing mechanism. These researches have been studied in considerable detail by both database and pattern recognition communities for different domains of time series data.

In order to analyze the time series of E-commerce transaction to affect relationship between the seller and buyer, in this paper the time series of E-commerce were calculating by Guass density function, and were analysis by sliding-window. It reveal the sales of shop according to the density of transaction, to help customer know the sales of shop, to help improved the sales rate.

3. The Model of the Time Series of E-commerce

In order to help buyer master the transaction of shop, and analyze the time series, we constructed the model of the time series of E-commerce. The model explains the process of calculating and analyzing. Figure 1 show this model.

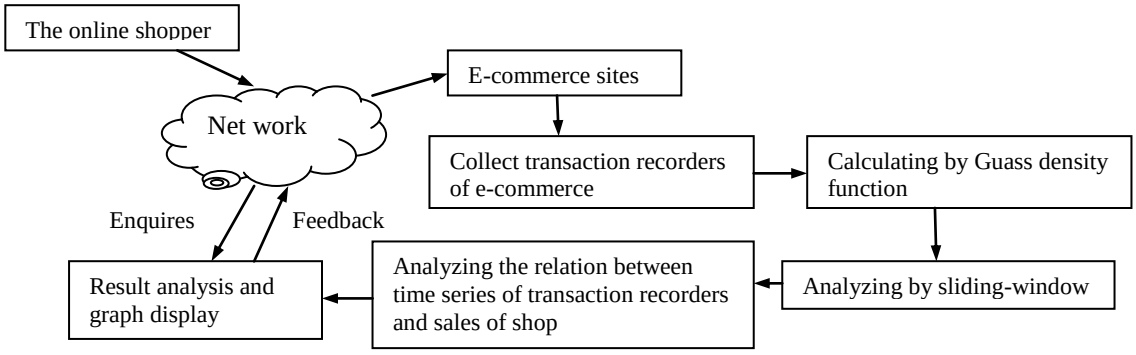


Figure 1. This Model of Time Series of E-commerce

It can be seen from figure 1, this model include three parts, namely extract time series from E-commerce transaction, time series of calculate and process and analysis result of time series.

3.1. Generate Time Series

The research of object is E-commerce transaction time in this model, so we construct a set T of transaction time series from online shop.

Definition 1: a set T of transaction time series is defined as a sequence of variable, t_1, t_2, t_3…and t_n, where the variable t_i marks the transaction time of seller, the set T will dynamic change with the time of the shop opened and trading volume of the shop.

3.2. Calculating and Processing of Time Series

We calculating the transaction time internals to a set T, and construct a set D of the transaction the internal.

Definition 2: a set D of transaction time internals is defined as a sequence of variables, d_1, d_2, d_3 … and d_n, where the variable d_i means transaction time internals of purchase product, that is $d_1 = t_2 - t_1$, $d_2 = t_3 - t_2$,... $d_n = t_{n+1} - t_n$

The set D are calculated by Guass density function, so we can construct a set M of density. The Guass density function as a research methods, because the Guass density function may be used to approximate describe density distribution of arbitrary and density of data point [16].

Definition 3: a set M of density of transaction recorder point is defined as a sequence of variables, m_1, m_2 … and m_n, where the variable mi means density value of transaction recorder point.

In order to keep the analysis result accurate and effective to a set M of density data. We process the set M by sliding window method,. The sliding window algorithm have given a window with length |W|, we slide it over this density sequence to see a set of overlapping sequences, where is each sequence is called the time-sensitive sliding window.

Definition 4: a set W of window length is defined as a sequence of variables, w_1, w_2 … and w_n, where the variable w_i means the length of window, that is $w_1 = w_2 = \cdots = w_n$, in general, the number of windows k may be equal to n/w, it is

required that the length of the windows be equal. Figure 2 show the sliding window method of our approach.

Definition 5: a set S of item set is defined as a sequence of variable, s_1, s_2 … and s_n, where s_i means item time sensitive sliding window. The item set will increase an item and delete an item, when the slide window over the density sequence, where T_i and T_{i-n} means the time point respectively, the increase item as S_i that is $S_i = (S_{i-1} - T_{i-n}) \cup T_i$ The delete item as S_{i-1} that is $S_{i-1} = S_{i-1} \cap S_i = S_{i-1} - T_{i-n} = S_i - T_i$.

To realize the data acquisition by sliding window method [17-18], and static and analysis the density distribution of transaction time series, we can conclusion the relationship between the sales of shop and time of transaction to help the customer purchase the products.

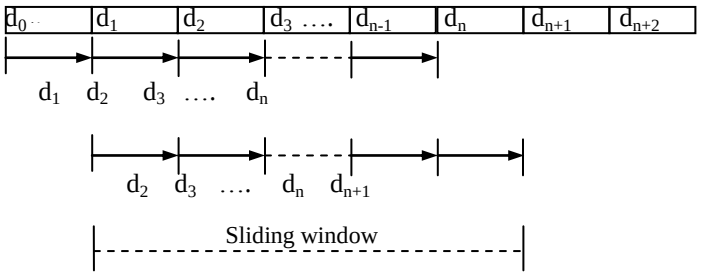


Figure 2. Time Sensitive Sliding Window Model

3.3. Result Analysis and Graph Display

We process the programming design to this algorithm, to design the information inquiring and result feedback in online shop, when the customer inquired the sales of shop, they will obtain an advice about purchase product of online shop, to help customer know the sales of shop, and help customer purchase product, and improved the purchase rate.

4. Our Approach

4.1. Calculating the Density of Time Series

To construct data base, these data of transaction time recorder were put into the database, so we can call and process these data in algorithm in time. To call the data of transaction time from the database, and construct the set T, that is $T = [t_1 - t_0, t_2 - t_1, \ldots t_n - t_{n-1}]$.

Let x and t be object or point in F^d, a d-dimensional input space, the influence function of data object t on x is a function, it can be used to compute a Guassian influence function [16]:

$$f_{Guass}^{x}(t) = e^{-\frac{d(x,t)^2}{2\delta^2}}$$

Where closeness is determined by parameter δ, the distance d(x,t) is Eudiden distance, which is defined as:

$$d(x_i, t_j) = [\sum_{k=1}^{n} (x_{ik} - t_{jk})^2]^{\frac{1}{2}}$$

The density function at object or point $x \in F^d$ is defined as sum of influence function of all data point, that is, it is the total influence on x of all of the data point. Given n data objects, $D = \{x_1, x_2 \ldots x_n\} \in F^d$, (that is $x_1=t_1$, $x_2=t_2$ … $x_n=t_n$) the density function at x is defined as:

$$F_B^D(X) = \sum_{\substack{i=1 \\ j=1}}^{n} e^{-\frac{d(x_i - t_j)^2}{2\delta^2}} = \sum_{i=1}^{n} e^{-\frac{((x_i - t_1)^2 + (x_i - t_2)^2 + \ldots + (x_i - t_n)^2)^{\frac{1}{2}}}{2\delta^2}}$$

A set D of density of time series is defined as sequence of variables $f_B^D(x_1)$, $f_B^D(x_2)$,... $f_B^D(x_n)$, where the variable $f_B^D(x_i)$ means the total influence on x of all of the data point.

4.2. Analyzing and Calculating by use of Sliding Window

With the increase of transaction times, the time series can increase gradually, in order to decrease the computational complexity and enhance the effect of analyzing. We calculated and analyzed the density of time series by use sliding window algorithm. Given a window with length |W| [17], and extract item set D from the density of time series [18]. An item set P as a sequence of variable p_1, p_2 … and p_n, where the variable pi means point that is lower than threshold value $f_B^D(x_0)$. The threshold value $f_B^D(x_0)$ is compare with the data of window, the point data which lower than $f_B^D(x_0)$ is stored in set P. If the data set increases the data of transaction time recorder, began to slide window back words. Figure 3, given a description algorithm.

```
1:   Dim D(n)          'density of time series
2:   Dim M(n)          ' the point which lower than f_B^D(x_0) is stored
3:   Dim f_B^D(x_0)     ' threshold value
4:   i=1
5    t=0
6:   s=1
7:   flag=0            ' set flag ,judge new data
8:   if flag<>0
        t=s
     else
        t=1
     End if
9:   for j=t to 1*i
10:  if D(j)< f_B^D(x_0)
        M(t)=D(j)
     End if
11:  s=1*i
     i=i+1
     End for
12:  if mod(n/1) >0
13:  for j=s to n-(1*i)
14:  if D(j)< f_B^D(x_0)
     m(t)<D(j)
     end if
     T=t+1
   End for
```

Figure 3. The Algorithm of Sliding Window

4.3. Calculating and Analysis of Result

In order to verified the validly of the model algorithm, in this paper, we simulated the transaction data of online shopping. It includes transaction-intensive, transaction-convention and transaction-sparse. The transaction-intensive is defined as high sales. The transaction-convention is defined as normal sales. The transaction-sparse is defined as low sales. Figure 4, Figure 5 and Figure 6 show the result for this model algorithm.

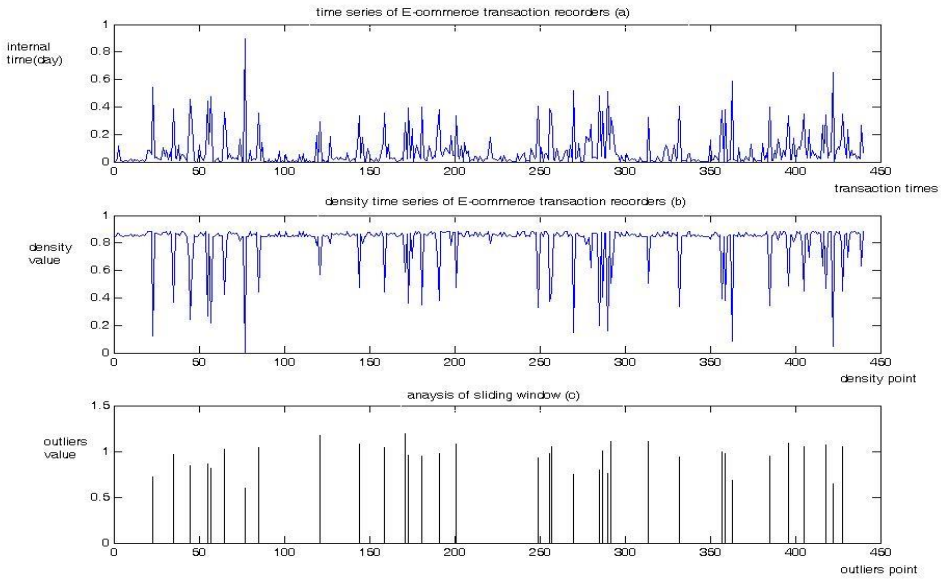


Figure 4. The Result of Transaction- intensive is Calculated by this Model Algorithm

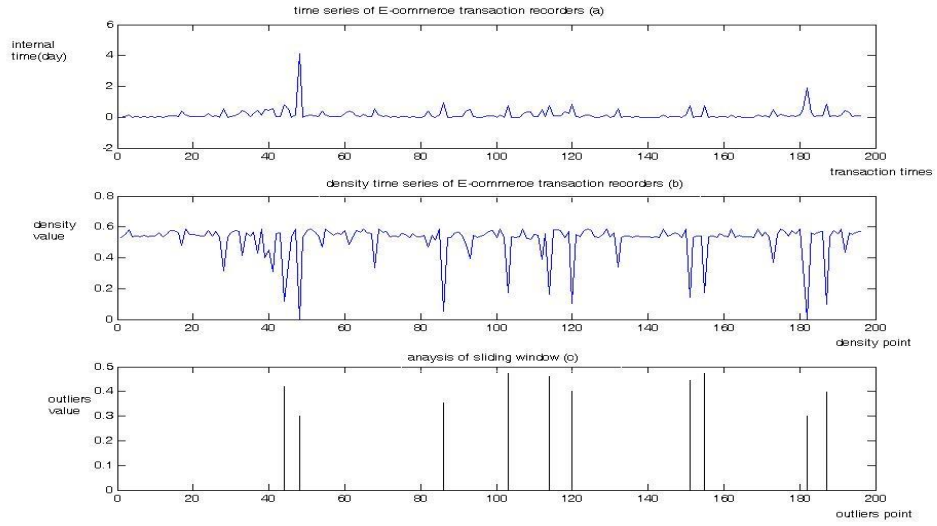


Figure 5. The Result of Transaction- convention is Calculated by this Model Algorithm

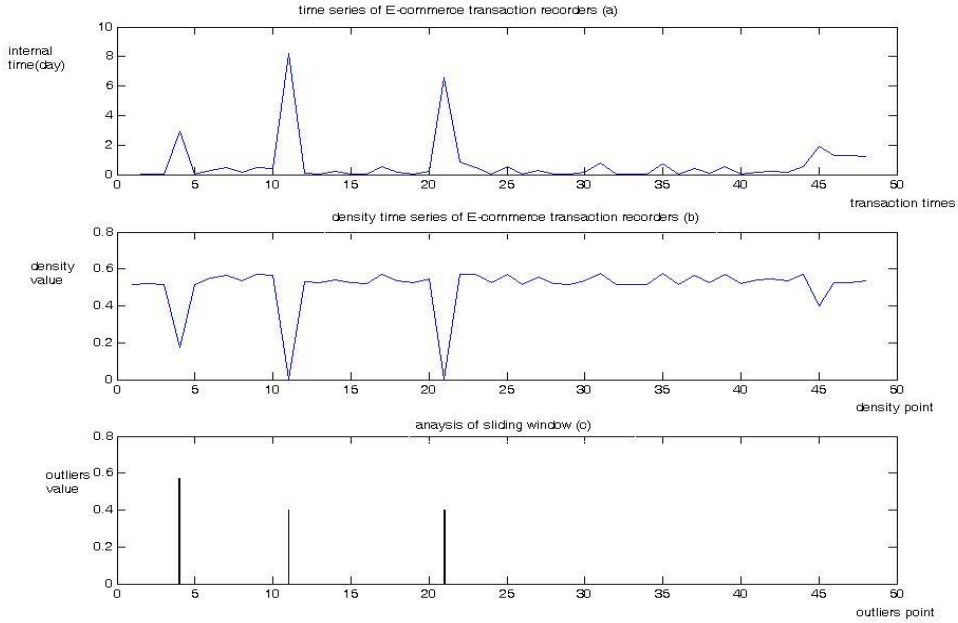


Figure 6. The Result of Transaction- sparse is Calculated by this Model Algorithm

Form Figure 4, when the online shopping transaction turnover is high, it can obtain many outliners of density series, through the model algorithm, the outlier is distributed mainly over 0:00-7:00, because the customer purchase is lower than other time. In according to the theory of Guass density function, the point that aren't density attracted by x*, but for which the density function value is less than $f_E^D(x_0)$, we considered outlier, if the outlier is none, it explains the sales of online shopping is high, the quality and service of product and user satisfaction, it has a stable passengers

From Figure 5, the transaction turnover is lower than transaction-intensive. The outlier is lower, because the generation outliers are accompanied by abnormal change of the transaction, which explain the quality and service of product and user satisfaction are lower than transaction-intensive, and its passenger are lower, if the customer decide purchase the online shopping product. The customer should think the shop opened time. If the shop opened time is long, we should abandon the online shopping product.

From Figure 6, the transaction turnover is bad in transaction-sparse, the outlier is little, which explain nothing much has changed for the sales of online shopping. Its sales is lower, it hasn't a stable passenger, they should carefully consider, when the customer purchase product of online shop.

5. Experiment and Result

This section talks about the experiment and results in this research to evaluate the proposed the model algorithm. We present the data, which took from Taobao.com. The transaction data involves transaction the high sales and the low sales at online shopping in one month, the data set were processed by MATLAB7.0 in windows XP SP3. Figure 7 and Figure 8 show the result for the experiment.

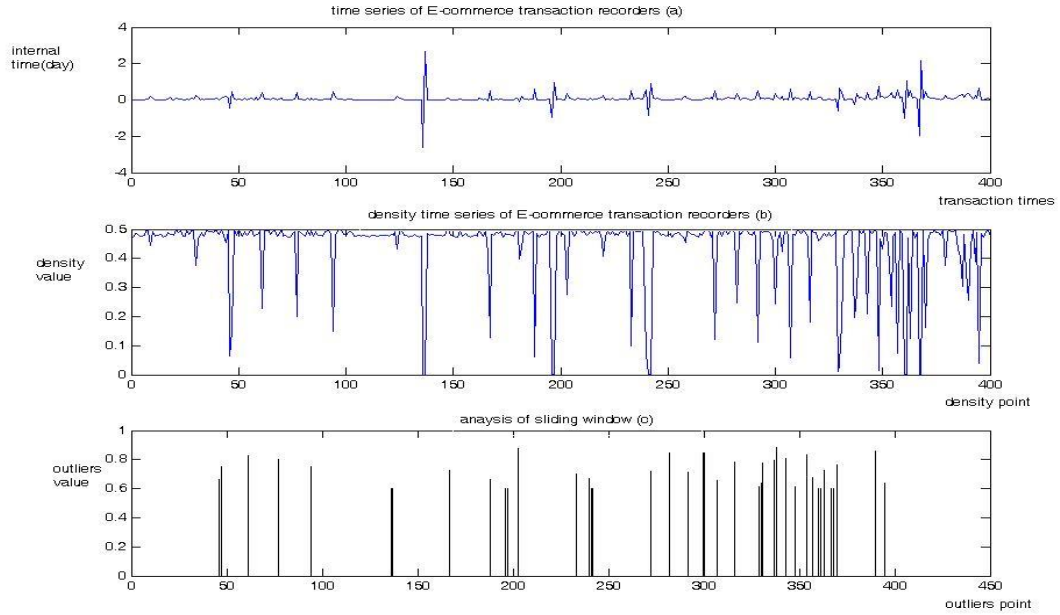


Figure 7. The High Sales Data which Processed by this Model Algorithm

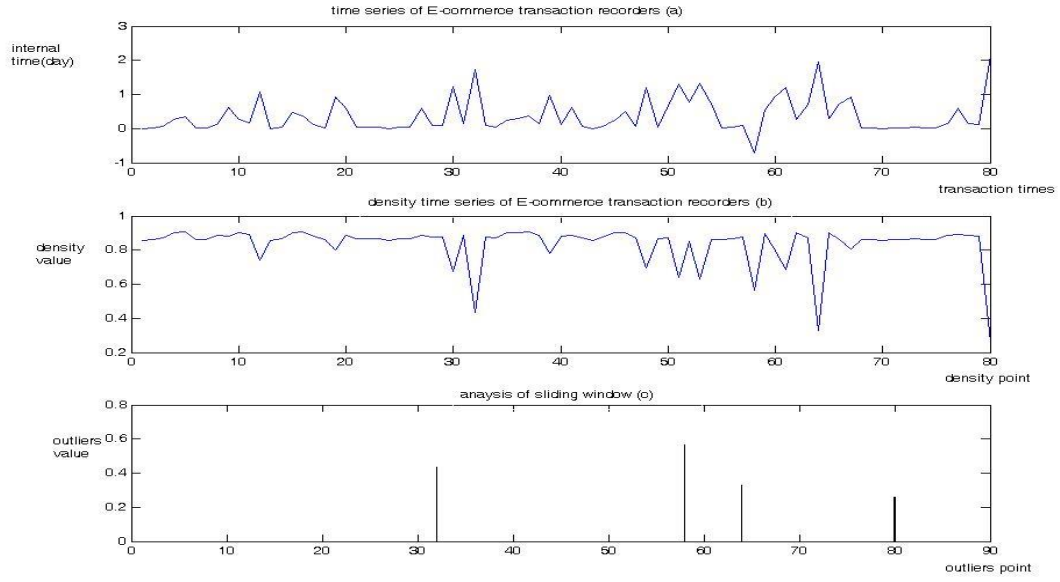


Figure 8. The Low Sales Data which Processed by this Model Algorithm

Form Figure 7, this time series have many outliers by calculate of density and analysis of sliding window to transaction recorders, which explains the online shopping have a stable sales, it sales are distributed mainly over the online shopping have a stable sales, its sales are distributed mainly over the midmorning, mid afternoon and around 9 pm, the quality and service of product and user satisfaction are better, customer may rest that the goods are guaranteed.

From Figure 8, this sales is lower about this online shopping, the outlier is lower by calculating of density and analysis of sliding window, which explain that the product is rarely buy by customer, when you buy goods in this online shopping, you should think change other shop.

6. Conclusion and Future Work

In this paper, we have shown the model algorithm can provide an efficient and effectives analysis of the transaction; it can be explain the relationship between the outlier and sales by use of Guass density and function and sliding window. The analysis results have certain instructive significance to purchase product in online shopping. This model algorithm is a certain practicality. But this approach was only used in the environment where the data set is large, if the data is little, the analysis accuracy will reduce.

In the future, we propose to provide the complete system for the low data set, and we plan to extend the current work towards purchase behavioral analysis of online buyer.

Acknowledgment

The work is supported by the National Natural Science Foundation of China (Grant no. 6106402 and 61364026) and young scientific research Foundation of LAN Zhou Jiao Tong University (Grant no. 2011044).

References

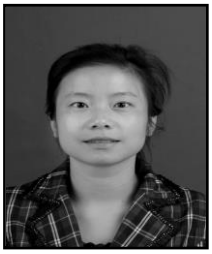
[1] China Electronic Commerce Research Center, The 2012 annual China network retail market data monitoring report, July **(2012)**
[2] Ji SHu-xian, Zhao Dong-mei, Validity of the trust model in online reputation feedback based on time value. Journal of Applied statistics and management, vol.30, no. 6, **(2011)**, pp. 1061-1066.
[3] L iu Zh uo-jun, Li Xiao-ming, An App roach for Unusual Transaction Detection Based on Time Series Modeling and Control Chart. Mathematics in practice and theory, vol.43, no.10, **(2013)**, pp. 89-96.
[4] Van Vo, Luo Jiawei1, Bay Vo, Stream Time Series Approach for Supporting Business Intelligence, International Journal of Database Theory and Application, vol. 6,no. 2,**(2013)**, pp.1-17.
[5] SUN Mei-yu, Research on discords detect on time series based on distance and density,Computer Engineering and Applications, vol. 48, no. 20, **(2012)** , pp. 11-17.
[6] N. Arora and J. R. Saini, "Time Series Model for Bankruptcy Prediction via Adaptive Neuro-Fuzzy Inference System", International Journal of Hybrid Information Technology, vol. 6, no. 2, **(2013)**, pp. 51-63.
[7] H.-Y. Lin, H.-Y. Chiu, C.-C. Sheng and A-P. Chen, "Hybrid Intelligence Approaches for Designing a Dynamic FinancialTime-series Predictive Model Based on Web-Architecture Home Finance Learning Environment", International Journal of Smart Home, vol. 2, no. 2, **(2008)**, pp. 13-31.
[8] R. G. Maravilla, Jr., E. R. A. Tabanda, J. A. Malinao and H. N. Adorna, "Data Signature-based Time Series Traffic Analysis on Coarse-grained NLEX Density Data Sets", International Journal of Database Theory and Application, vol. 5, no. 1, **(2012)** March.
[9] W. Xiang-yu and Z. Yin-qiong, "Discussion on Time Series Abnormal Detection Methods Based on MATLAB", Computer Knowledge and Technology, vol. 8, no. 4, **(2012)**, pp. 866-872.
[10] T. Fu, "A review on time series data mining", Engineering Applications of Artificial Intelligence, vol. 24, no. 1, **(2011)**, pp. 164-181.
[11] S. Rouhani, M. Ghazanfari and M. Jafari, "Evaluation model of business intelligence for enterprise systems using fuzzy TOPSIS", Expert Systems with Applications, vol. 39, no. 3, **(2012)**, pp. 3764-3771.
[12] A. Camerra, T. Palpanas, J. Shieh and E. Keogh, "iSAX 2.0: Indexing and mining one billion time series", In Proceedings of the IEEE 10th International Conference on Data Mining (ICDM), **(2010)**, pp. 58-67.
[13] W. K. Wong, E. Bai and A. W. C. Chu, "Adaptive time variant models for fuzzy time series forecasting", IEEE Transaction on Systems, Man and Cybernetics-Part B: Cybernetics, vol. 40, no. 6, **(2010)**, pp. 1531-1542.
[14] S.-H. Park, J.-H. Lee, S.-J. Chun and J.-W. Song, "Representation and clustering of time series by means of segmentation based on PIPs detection", Proceedings of the 2nd International Conference on Computer and Automation Engineering (ICCAE), vol. 3, **(2010)**, pp. 17-21.

[15] A. Camerra, T. Palpanas, J. Shieh and E. Keogh, "iSAX 2.0: Indexing and mining one billion time series", Proceedings of the IEEE 10th International Conference on Data Mining (ICDM), **(2010)**, pp. 58-67.
[16] C. Na, "Automatic image annotation method based on Gaussian mixture model", Journal of Computer Applications, vol. 30, no. 11, **(2010)**, pp. 2986-2987+2993.
[17] J. Hyuk Chang and W. Suk Lee, "estWin: Online data stream mining of recent frequent item sets by sliding window method", Journal of Information Science, vol. 31, no. 2, **(2005)**, pp. 76-90.
[18] K. Xiangxia, R. Yonggong and S. Kuiyong, "An algorithm for mining frequent itemsets in data streams over sliding window", Computer Applications and Software, vol. 30, no. 1, **(2013)**, pp. 143-146.

Authors

Xiao Qiang, he received the master degree from the school of information and electrical engineering, Lan Zhou Jiao tong University, in 2007. He is currently working toward the PhD degree in the school of Traffic and Transportation, Lan Zhou Jiao tong University, at Lanzhou in china. His research interests include data mining, E-commerce and information system.

He Rui-chun, she is currently a professor and Ph.D. adviser at the School of Traffic and Transportation, Lanzhou Jiao tong University. Her major research focuses on Analysis and optimization of transportation system, Analysis of traffic network complexity, Management decision analysis.



Liao Hui, she received the BS degree from the school of Economics and Management LanZhou Jiao tong University, in 2011. She is currently working toward the master degree in the school of Economics and Management; Lan Zhou Jiao tong University, Lanzhou. Her research interests include business management, E-commerce and information system.

