A Method for Data Stream Processing Based on Curve Fitting

Yixu Song，Jing Hu
State Key Laboratory of Intelligent Technology and
System Department of Computer Science and
Technology
Tsinghua University
Beijing, China
e-mail:songyixu@hotmail.com

Xiaokui Yang
Department of Computer Science and Technology
University of Science and Technology Beijing
Beijing, China
e-mail:yangxk-2000@163.com

Jie Fu, Xiufen Xie
Sany Intelligent Control Equipment CO.LTD
Changsha, Hunan, China
line 4: e-mail: fujie74@gmail.com

***Abstract*— The sampling storage method which used in the current data stream could not respond data tendency effectively. For the problem, this paper presents a new processing method based on curve fitting. A weighted least-square principle is used to fit the cached stream data and better model description is obtained. Then the fitting results are analyzed by clustering algorithm, which serves as a classifier for polynomial fitting parameters. According to the clustering result, the appropriate window size will be given to fit the periodic stream data. Comparing the function solutions with the actual data, the different methods are adopted to store data according to the comparison result. The experimental results indicate that the proposed method has better fitting accuracy and compression ratio, could meet the requirement of data stream processing. And the data tendency could be responded effectively by the fitting results.**

Keywords-curve fitting; data stream; clustering; least - square principle

I. INTRODUCTION

In recent years, stream data processing is gaining extensive attention. The reason is that this class of data-intensive application is more and more popular with the development of network technology, such as electronic commerce, network monitoring, stock shares, and wireless communications network, etc. Stream data is a special data type which composed of large-volume, quick, continuous and real-time data sequences [1]. As the data stream is arriving continuously and indefinitely, limited data processor can not store all the information. On the other hand, for some systems, such as equipment or safety monitoring applications, such data is often rendered multi-point concurrent, large scale and so on. In majority situation the data stream contains a lot of correlated data according the timing sequence; even the follow-up data are redundancy or the same as the former data. In order to guarantee the storage data accuracy, adopt appropriate method to describe the data stream, that is not only an effective way to find out data association, but also has great significance in data compression, reducing the pressure of the system memory and increasing query speed.

At present, for the data stream processing, a general method is to open a sliding window in memory unit, which used to store the recent stream data. Through this approach, real-time queries could be supported. With the fact that the data stream flows into the sliding window unceasingly, some old data will be out from the sliding window when the sliding window has been full. These parts of data that flow out of the sliding window are called the historical data. There have been some researches on compression algorithm of the historical data. In overseas, a sketching sampled method is presented by Rusu F. and Dobra A. for data stream processing [2]. The literature [3] presents a sampling data stream algorithm for wireless sensor networks. Manoranjan Dash and Willie Ng propose a distance based sampling (DSS) for transactional data streams [4]. The multi-layer recursive sampling method is used in historical streaming data processing in China [5]. These methods are based on sampling. But in many cases, the continuous data often need to be preserved. So the sampling method may not meet this demand.

The swinging door compression algorithm [6] [7] and the dead band limit compression algorithm [8] are the main compression algorithms in the industrial processing of data stream. The swinging door compression algorithm is a patent compression technology which used in Plant Information System developed by American OSI Software Company. The data stored by the swinging door compression algorithm are the actual data, so the data tendency must be obtained through secondary treatment. Besides, some data are abandoned by the swinging door algorithm, and difficulties would be encountered when this data needed to be restored. For the dead band limit compression algorithm, the data is retained when its value greater than the dead zoneband limit, otherwise the data

978-1-4244-6893-5/$26.00 © 2010 IEEE

would be discarded. So the accuracy of the dead band limit compression algorithm can not be guaranteed.

To overcome the above-mentioned disadvantages, in this paper, the historical data of data stream are compressed by the way of the weighted piecewise curve fitting, and the k-means cluster algorithm is used to analyze the fitting results for selecting the sliding analysis window. The data are processed in appropriate window size according to the clustering result after the periodicity of the stream data has been found out. As the fitting results are stored, the data tendency could be responded.

The rest of the paper is organized as follows:

The principle of the least - square principle curve fitting is showed in the Section II. The Section III introduces the basis knowledge of clustering analysis, analysis of the k-means algorithm based on the typical division and the improvement of this paper. The Section IV describes the proposed data stream processing method composed of the weighted curve fitting algorithm and clustering algorithm. Finally, experimental results and some conclusions are given in this paper.

II. Variable-order Weighted Piecewise Curve Fitting Algorithm

Closely related with the time, the data stream is a set of data sequences which arrive sequentially, largely, quickly and continuously. The data stream could be regarded as infinite multiset in which each element has the form <*s*, *t*>, where *s* is a tuple and *t* is the time stamp of *s*. The value of *t* can be the time that *s* enters the data stream system or the time that *s* is produced by data source [9]. Take one-dimensional data in the tuple and the time to consider separately, then the data stream is a function with respect to the time *t*. Therefore, the curve fitting method can be used in the data stream processing. And the data tendency could be obtained by this function. The disadvantage of sampling storage method is overcome.

The curve fitting method is one kind of solution method that finds out the approximate expression (1) of the function $y=f(x)$, according to a set of experimental data points (x_i, y_i) $(i=1, 2, \ldots,k)$.

$$y=\varphi\left(x\right)=a_0+a_1x+a_2x^2+\ldots+a_nx^n=\sum_{i=0}^{n}a_ix^i\ . \qquad (1)$$

Numerous fitting curves may be solved depending on different definitions. The solved fitting curve $y=\varphi(x)$ is required to close to the every data point as much as possible. For the data point *i*, let the error between the function value $\varphi(x_i)$ and the actual value y_i is ε_i ,

$$\varepsilon_i=\varphi\left(x_i\right)-y_i\ . \qquad (2)$$

The least - square principle curve fitting is a method that uses "minimum variance" as a judgment principle, i.e. eq. (3) is to be maximum:

$$e=\sum_{i=1}^{k}\left(\varepsilon_i\right)^2=\sum_{i=1}^{k}\left(\varphi\left(x_i\right)-y_i\right)^2\ . \qquad (3)$$

In practical applications, not all data points are the same importance, especially in the case of mutation curve, in which the mutation points indicate special meanings usually. Therefore, different weights should be arranged according to importance of the data points and more important data point will be provided with greater weight. This algorithm is called the weighted least- square curve fitting. The weighted least-square curve fitting function of $y=\varphi(x)$ is an n-order polynomial (1). In (1), *x* and its coefficients can be obtained by solving the normal equation. Let the weight is w , then the corresponding normal equations are

$$\begin{bmatrix} \sum_{i=1}^{k}w_i & \sum_{i=1}^{k}w_ix_i & \sum_{i=1}^{k}w_ix_i^2 & \cdots & \sum_{i=1}^{k}w_ix_i^n \\ \sum_{i=1}^{k}w_ix_i & \sum_{i=1}^{k}w_ix_i^2 & \sum_{i=1}^{k}w_ix_i^3 & \cdots & \sum_{i=1}^{k}w_ix_i^{n+1} \\ \sum_{i=1}^{k}w_ix_i^2 & \sum_{i=1}^{k}w_ix_i^3 & \sum_{i=1}^{k}w_ix_i^4 & \cdots & \sum_{i=1}^{k}w_ix_i^{n+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{k}w_ix_i^n & \sum_{i=1}^{k}w_ix_i^{n+1} & \sum_{i=1}^{k}w_ix_i^{n+2} & \cdots & \sum_{i=1}^{k}w_ix_i^{2n} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{k}w_iy_i \\ \sum_{i=1}^{k}w_iy_ix \\ \sum_{i=1}^{k}w_iy_ix^2 \\ \cdots \\ \sum_{i=1}^{k}w_iy_ix^n \end{bmatrix} . \qquad (4)$$

Where

$$\boldsymbol{X}=\begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2^2 & x_2^2 & \ldots & x_2^n \\ 1 & x_3^2 & x_3^2 & \ldots & x_3^n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_k^2 & x_k^2 & \ldots & x_k^n \end{bmatrix}, \boldsymbol{W}=\begin{bmatrix} w_1 & 0 & 0 & \ldots & 0 \\ 0 & w_2 & 0 & \ldots & 0 \\ 0 & 0 & w_3 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & w_k \end{bmatrix},$$

$\boldsymbol{A}=\left(a_0,a_1,a_2,\cdots,a_k\right)^T$, $\boldsymbol{Y}=\left(y_0,y_1,y_2,\cdots,y_k\right)^T$, and the normal equation can be denoted

$$\boldsymbol{X^TWXA}=\boldsymbol{X^TWY}\ . \qquad (5)$$

For curve fitting, the high accuracy of the fitting results is demanded. According to the literature [10], the accuracy of curve fitting is related with the following three aspects: the segment number of piecewise fitting, the order of fitting curve and the distribution of data point weights. The fitting accuracy is better with more fitting segments, higher fitting order and more reasonable data point weight distribution. Therefore, the various factors should be integrated with piecewise fitting approach for better accuracy. In order to improve the accuracy of curve fitting, the variable-order weighted piecewise method is adopted in this paper.

On the other hand, the fitting accuracy and the processing time can be influenced by the fitting window size. If the size of fitting window is too large, data processing time is longer and the fitting accuracy also decline. However, the tendency of the data stream can not be effectively grasped if the size of fitting window is too small. In the literature [11] a scheme that unifies the characteristic of two ways of window selection is introduced. In this paper the window size is set by method in [11], that is, to establish a standard data window as the minimum fitting window and a big data window as the ultimate fitting window. The treatment process is as follows:

(1) Accept the data stream, if the cached data are more than standard size, fit the data and calculate the maximum fitting error.

(2) If the maximum error is less than the error limit, then continue to accept the data, repeat step (1). Otherwise, take the maximum error point as the segmentation point, and fit the data before this point. Then store the fitting result, and

set the first data point after the segmentation point as a new starting point.

(3) If the cached data are more than the limit size, fit the data, store the fitting result, and take the next time read-in data as a new starting point, repeat step (1).

III. K-MEANS CLUSTER ALGORITHM

Clustering is a certain type of process that divides data set into several groups or several classes, according to an established way, making the data in same categories with high similarity and the data in different categories with lower similarity [12]. In this paper the k-means clustering algorithm is used to analyze the fitting results for the inherent law of data stream.

The treatment of the k-means clustering algorithm is that assign a data sample, enter the number *k* of obtained cluster, divide the data into *k* parts, adjust the division through a renewal of the cluster center and finish the processing when whole diversity function converges. The differences among categories are the representation of the cluster center, the adjustment strategy of division and the definition of whole diversity function. The treatment process of k-means clustering algorithm is as follows:

(1) Choose arbitrary k-data as the initial cluster centers.

(2) Calculate the distance of each data point to center of those clusters and reclassify the data in accordance with the minimum distance.

(3) Recalculate the center of each cluster.

(4) Repeat steps (2) and (3) until the clusters no longer change.

In this paper k-means clustering algorithm is used to find out the law of the data stream, which is identifying the data cycle by clustering the fitting results. The k-means clustering algorithm need to be given the number *k* of clusters in advance and is sensitive to initial value. But regarding the data stream, we do not know how several suitable categories that the given data set should divide into. So the k-means clustering algorithm is improved to meet the need of the data stream in this paper. The concrete procedure is that set an appropriate initial distance to determine the initial cluster center. Treatment processes are as follows:

(1) Set the initial distance.

(2) Calculate the distance of new data point to each cluster center. If the minimum distance is greater than the initial distance，then this point is the center of new cluster. Otherwise carry on the classification to the data according to the minimum distance.

(3) Recalculate the center of each cluster.

(4) Repeat steps (2) and (3) until the cluster no longer changes.

(5) Repeat steps (2), (3) and (4) until the completion of data processing.

Clustering algorithm implementation requires large amounts of data. Therefore, the data can be cached for a period until data is sufficient for processing.

IV. A PROCESSING METHOD FOR STREAM DATA BASED ON THE CURVE FITTING AND THE CLUSTERING ANALYSIS

For reasonable compression of historical data of the data stream and retaining maximally all the information of the data stream, the variable-order weighted piecewise curve fitting algorithm is used to fit the data stream firstly. Then the k-means clustering algorithm is adopted as classifier of the coefficients of the polynomials. If the stream data is periodic, the optimum window size for fitting will be selected by quasi data periodicity given by clustering. If the data stream is non-regular, then the curve fitting result can be stored directly.

The steps of the algorithm are as follows:

(1) In given analysis window size, use variable-order weighted piecewise curve fitting algorithm for received stream data.

(2) Calculate the maximum fitting error.

(3) If the maximum fitting error is greater than the threshold, the fitting weight of the maximum error points plus 1. Otherwise, continue to receive stream data to analysis window.

(4) If the fitting weight of the error point is greater than the given weight, the fitting order plus 1. If the fitting order is greater than the maximum fitting order and the given error bound is not satisfied, the error point will be reserved as a segmentation point.

(5) Fit the data before segmentation point with appropriate fitting order and store the fitting results.

(6) If the cached data is greater than the maximum number, fit the data in the temporary memory and store the fitting result.

(7) Use k-means clustering algorithm to analyze the fitting results when the fitting results is sufficient.

(8) If the clustering results are stable after the cluster analysis has completed, use the appropriate size to fit the data, store the fitting results.

Fig.1 shows the flow chart of above-mentioned algorithm. In the chart *m* is the standard window length; *tempMaxNum* is the maximum window size; *errMax* is the maximum limit error; *wMax* is the maximum weight and *PolyMaxNum* is the maximum fitting order.

In this paper we use the clustering algorithm to analyze the fitting results and to find out the law according to the fitting results, which indicate the data periodicity. If the data stream is periodic, then the best data size may be obtained. Set the optimum data size as the fitting window size, fit the data, process data according to the fitting results. The flow chart of clustering algorithm is shown in the Fig.2.

V. EXPERIMENTS AND ANALYSIS

In order to verify the validity of the proposed method in the paper, an experimental platform has built. The hardware setup of test platform is: Pentium（R）4 CPU 3.00Ghz 2.99Ghz 2.00GB memory. And software setup is Microsoft Visual Studio 2005 and Microsoft SQL Server 2005 under Window XP. The programming language is Microsoft c#. The experimental data set is a class of GIS data.

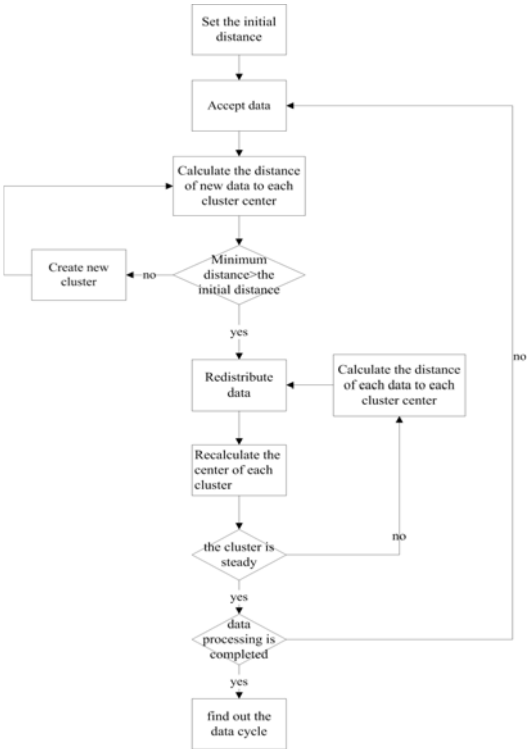
Set the initial distance
Accept data
Calculate the distance of new data to each cluster center
Create new cluster
no
Minimum distance>the initial distance
yes
Redistribute data
Calculate the distance of each data to each cluster center
no
Recalculate the center of each cluster
the cluster is steady
no
yes
data processing is completed
yes
find out the data cycle

Figure 2. The flow chart of Clustering Algorithm

Experiment 1: To test description ability for the stream data with the given accuracy. The variable-order weighted piecewise curve fitting algorithm is used to describe test sample. The parameters of the algorithm set as follows: The size of sample is 34, the maximum limit of error is 0.001, the standard window size is 10, the maximum data window size is 20, the maximum weight is 10 and the maximum fitting order is 3. The fitting results are composed of four polynomials, including three sections of third-order polynomial and one section of second-order polynomial.

The first section:

$$y = 112.410801003947 + 0.00436141102121959x- 0.00661338917509281x^2 + 0.000955684979106687x^3$$

The second section:

$$y=112.3951341468-0.0130302752127297x+ 0.0171521622678368x^2-0.00855948918011805x^3$$

The third section:

$$y = 112.385175925918 + 0.000539381733469711x- 0.000114091549429529x^2 + 4.26095748333338E-06x^3$$

The fourth section:

$$y=112.393699540921+0.0122177559194992x- 0.000793078562202373x^2$$

Fig.3 shows the comparison of the reductive data and the original data.The reductive data are restored from compressed data. From the fitting results, the majority of data are well fitted and little the partial data present some errors. The maximum error appears in the 26th point, which error is 0.0009042, less than the maximum error limit.

Experiment 2: In this experiment, the efficiency of fitting algorithm will be verified. The size of test sample is 48.5kb.

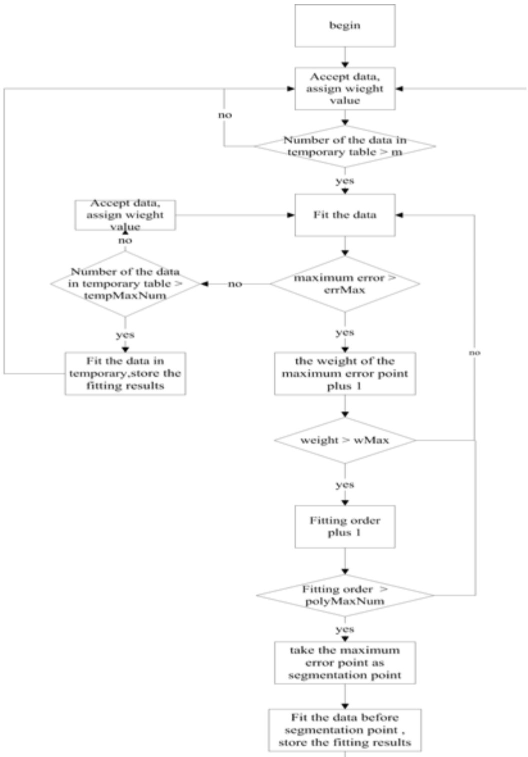
begin
Accept data, assign wieght value
no
Number of the data in temporary table > m
yes
Accept data, assign wieght value
Fit the data
no
Number of the data in temporary table > tempMaxNum
no
maximum error > errMax
yes
yes
Fit the data in temporary,store the fitting results
the weight of the maximum error point plus 1
no
weight > wMax
yes
Fitting order plus 1
Fitting order > polyMaxNum
yes
take the maximum error point as segmentation point
Fit the data before segmentation point , store the fitting results

Figure1. The program flow chart

The variable-order weighted piecewise curve fitting algorithm and the swinging door compression algorithm are respectively used to compress the test sample, with the accuracy of 0.003. In the storage region the parameters of swinging door compression algorithm are: the serial number (No), the system number (SystemId), time (Time) and data values (Data). The parameters of the variable-order weighted piecewise curve fitting algorithm are set to the maximum weight of 5, the maximum order of 4, the standard window size of 15 and the maximum window size of 20. In the storage region the parameters of the proposed method are: serial number (No), the system number (SystemId), start time (StartTime), parameter 1 (A0), parameter 2 (A1), parameter 3 (A2), parameter 4 (A3), parameter 5 (A4) and fitting number (Num). The comparison of compression results are shown in the Table 1.

From the comparison of the processing results, the compression ratio of the variable-order weighted piecewise curve fitting algorithm is higher than the swinging door compression algorithm with appropriate order and fitting window size under the same compression accuracy. Since the parameters of variable-order weighted piecewise curve fitting algorithm are more, the algorithm can even more manifest the superiority when the data description scale is quite large and complicated.

Experiment 3: the clustering algorithm is applied to analyze periodic data. The size of test sample which has certain periodicity is 22.8kb. The scatter diagram of test sample is shown in the Fig 4. In the storage region the parameters of the swinging door compression algorithm are: the serial number (No), the system number (SystemId), time

(Time) and data values (Data). The parameters of variable-order weighted piecewise curve fitting algorithm are set to the maximum weight of 10, the maximum order of 4, the standard window size of 20 and the maximum window size of 30. In the storage region the parameters of the proposed method are: serial number (No), the system number (SystemId), start time (StartTime), parameter 1 (A0), parameter 2 (A1), parameter 3 (A2), parameter 4 (A3), parameter 5 (A4) and fitting number (Num).

The curve described by polynomials is composed of 20 sections. The clustering result divides into 2 categories; each category has 10 segments with the initial distance of 0.003. The period of test sample is 40 by calculating. The proposed algorithm with and without cluster algorithm and the swinging door compression algorithm are respectively used to compress the test sample, with the accuracy of 0.003. The processing results are shown in table 2. From the processing result, we find that the compression ratio of this article's algorithm with cluster algorithm is higher than the other.

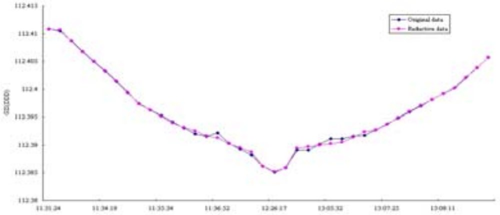

Figure3. The comparison diagram of curve fitting

TABLE I. COMPARISON OF TWO COMPRESSION ALGORITHMS

algorithm	Comparison	
	processing result	*compression ratio*
The swinging door compression algorithm	10.8kb	22.26%
The variable-order weighted piecewise curve fitting algorithm	10.8kb	14.56%

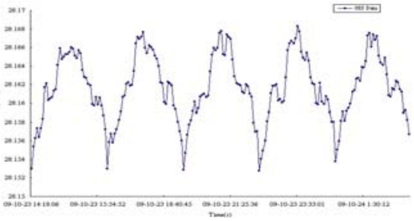

Figure4. The scatter diagram of test sample

TABLE II. COMPARISON OF COMPRESSION RATIO OF THREE DIFFERENT METHODS

algorithm	Comparison	
	processing result	*compression ratio*
The swinging door compression algorithm	3.90kb	17.11%
The proposed algorithm without clustering algorithm	2.91kb	12.76%
The proposed algorithm with clustering algorithm	1.60kb	7.02%

VI. CONCLUSIONS

This paper presents a new processing method to store the data stream. The data can be well compressed by using the variable-order weighted piecewise curve fitting algorithm, combined with clustering algorithm. The experimental results show that the variable-order weighted piecewise curve fitting algorithm has better fitting accuracy and compression ratio. The compression ratio will be more significant and only small amounts of data are stored, if the stream data is periodic and appropriate analysis window is adopted by clustering algorithm. By the fitting polynomial equations the description of the stream data is accurately obtained, and the problem of storage and query for large-volume data is effectively solved. Furthermore, the data model is also useful for tendency prediction of the stream data.

This work is supported by the National High Technology Research and Development Program of China (Grant No.2008AA042802).

REFERENCES

[1] Tomoya Saito, Takuya Kida, and Hiroki Arimura, An Efficient Algorithm for Complex Pattern Matching over Continuous Data Streams Based on Bit-Parallel Method, IEEE International Workshop on Databases for Next Generation Researchers(SWOD 2007), 2007, pp. 13 – 18.

[2] Rusu F., and Dobra A., "Sketching Sampled Data Streams," IEEE 25th International Conference on Data Engineering(ICDE '09), April. 2009, pp. 381 – 392, doi: 10.1109/ICDE.2009.31.

[3] de Aquino A.L.L. , Figueiredo C.M.S. et al., "A Sampling Data Stream Algorithm For Wireless Sensor Networks," IEEE International Conference on Communications (ICC '07), June. 2007, pp. 3207 – 3212, doi:10.1109/ICC.2007.532

[4] Manoranjan Dash , and Willie Ng , "Efficient Reservoir Sampling for Transactional Data Streams," Sixth IEEE International Conference on Data Mining Workshops(ICDM Workshops 2006.), Dec. 2006, pp. 662 – 666, doi: 10.1109/TMM.2006.879875.

[5] Zhang Dong-Dong, Li Jian-Zhong, Wang Wei-Ping, and Guo Long-Jiang, "Algorithms for Storing and Aggregating Historical Streaming Data," Journal of Software, vol. 16, no. 12, 2005, pp. 2089-2098.

[6] Rafael S Parpinelli, "Data mining with an ant colony optimization algorithm," IEEE Transactions on Evolutionary Computation, vol. 6, no. 4, 2002, pp. 321-322.

[7] Bristol E. H., "Swinging door trending : Adaptive Trend Recording[C]," In : ISA National Conference Proceedings, 1990, pp.749-753.

[8] Gao Ningbo, Jin Hong, and Wang Hongan, "Study on Real-Time Compression of Historical Data," Computer Engineering and Applications, vol. 28, no. 8, 2004, pp. 167 – 173.

[9] Araru A , Babu S , and Widom J, An abstract semantics and concrete language for continuous queries over and relations Technical Report , Stanford University Database Group , 2002, Available at http : //dbpubs.Stanford.edu/pub/2002-57.

[10] Han Wen-Qing, "The Method of Polynomial Fitting in Sections by changing Weights to Suit Data Processing," Journal of Data Acquisition and Processing , vol. 2, no. 2, Sep. 1987, pp.38-44.

[11] Wang Cheng-Liang, Lu Zhi-Jian, and Pang Xu, "Research and Application of an Algorithm for Trend Analysis of Data Streams," Computer Systems & Applications, vol. 19, no. 1, 2010, pp.152-156.

[12] Zhou Xin, and Zhang Hua-Xiang, "Study and Improve on K-means Algorithm," Microcomputer Information, vol. 24, no. 10, 2008, pp.269-271.