IUCStream: A Novel Increment Update Clustering Algorithm for Commodity Stream Data Analysis in e-Commerce and Logistics

Peihua Fu
Col. of Comp.& Inf. Eng., Zhejiang Gongshang Univ., Hangzhou, China
Email: fph@mail.zjgsu.edu.cn

Yangfei Chen and Hongbo Yin
Col. of Comp.& Inf. Eng., Zhejiang Gongshang Univ., Hangzhou, China
Email: faye.chen1988@163.com

***Abstract*—With China's rapid development of e-commerce and logistics, many large electronic business enterprises start to establish large volume warehouses. It takes a long time to distribute the goods every time, so the optimal distribution link can save a lot of time and it has important practical significance. In order to optimize goods inventory and delivery, the electronic commerce goods shopping cart stream data need to be analyzed. In this paper, a novel increment update clustering algorithm, named as IUCStream for commodity stream data analysis in e-commerce and logistics is proposed. In this algorithm, the correlation between goods is calculated and an efficient algorithm processing incremental updating of the data stream of goods is used to cluster different goods into groups. Finally, the algorithms' superiority and effectiveness are verifying by an example.**

***Index Terms*—Data stream, Clustering algorithm, Commodity correlation, e-commerce, logistics**

I. INTRODUCTION

Recently, China's e-commerce industry is growing explosively. Traditionally, e-commerce dealers take charge of inventory and logistics, so that the e-commerce service providers need not establish large volume warehouses. However, in the past two years, China's main electronic business enterprise such as Jingdong, Amazon, Taobao and so on are interesting in establishing their own large scale warehouses and logistics centers. Some smaller online retailers like XIU, OkayBuy also would link to arrange inventory and logistics themselves. Generally, the electronic business enterprises service as third party logistics enterprises.

In a traditional manufactory or warehousing enterprise, the goods dealt with are always in few types and can be arranged in uniform goods shelves in automatic warehouses. But, logistics enterprises have a wide variety of goods, to process this commercial information always is a very hard work. All the goods are needed to be stored in categories according to size property, turnover frequency and cargo owner. How to increase efficiency of inventory management? Clustering algorithms are always used to solve this problem.

The traditional clustering algorithm can be divided into almost seven types, that is, partitioning-based methods, hierarchical-based methods, density-based methods, grid-based methods and model-based methods and so on, refer to reference [1]. The data stream analysis model was proposed by Henzinger in the first time in 1988 [2]. Incremental DBSCAN algorithm was the first incremental updating clustering algorithm [3], used to process data of warehouse. It shows efficient practical on the analysis of the stability of the data stream, but can do little about data stream that changing with real time.

In 2011, Niu improved the fuzzy c-means clustering algorithm based on particle swarm optimization (PSO) [4]. Optimization strategy was presented that the optimal particle can be guided to close the group effectively. In 2012, Li presented an improved clustering algorithm based *K*-means and self-organizing model (SOM) [5]. It is not very sensitive to the initial cluster center and the algorithm has a higher accuracy and better stability.

In order to solve the problem, in 2003, Aggarwal put forward a data stream clustering framework----CluStream [6]. Then in 2006, Feng Cao proposed a DenStream algorithm which mainly aimed at the data stream dynamic evolution [7]. It is based on the merits of the Density-based Method and made much improvement on the CluStream algorithm.

In 2007, Bhatnagar put forward the ExCC algorithm which based on the grid and the density of data stream clustering algorithm [8]. It put forward the concept of completeness clustering. In 2008, Aggarwal proposed UMicro algorithm [9], which proposed the concept of uncertainty on the basis of CluStream. In 2009, Luehr et al. presented an incremental graph-based clustering

Manuscript received December 10, 2011; revised January 6, 2012; accepted July 1, 2012.
This work was supported in part by the Commonweal Technology Project of Zhejiang Province, China under Grant No. 2011C23076.
Corresponding author: Peihua Fu, fph@mail.zjgsu.edu.cn.

doi:10.4304/jcp.8.3.795-802

algorithm whose design was motivated by a need to extract and retain meaningful information from data streams produced by applications [10].

In 2009, Bhatnagar et al. performed a comparative study of different approaches for existing stream clustering algorithms and presented a parameterized architectural framework and two assembled algorithms G-kMeans and G-dbscan [11].

In 2009, Li et al. proposed an effective bit-sequence based, one-pass algorithm, called MFI-TransSW to mine the set of frequent item sets from data streams within a transaction-sensitive sliding window which consists of a fixed number of transactions [12].

In 2010, Cao et al. proposed a generalized clustering framework for categorical time-evolving data, which is composed of three algorithms [13]. In the same year, Bae et al. introduced a new clustering similarity measure, known as ADCO, which aims to address some limitations of existing measures, by allowing greater flexibility of comparison via the use of density profiles to characterize a clustering [14].

In 2011, Skala and Kolingerova presented a novel approach to handle large amounts of geometric data and proposed a method for removal multiple points from Delaunay triangulation [15]. And Brice et al. leveraged existing clustering techniques for static categorical data sets to capture dynamic data streams based on the CT models using an information-theoretical approach [16].

In this paper, correlation of data stream and data stream clustering algorithm is analyzing. The correlation of the data stream is the level of similarity between the data, different data types determining the specific meaning. Data stream clustering algorithm is extended and improved to the traditional clustering algorithms under the environment of the data stream. In order to clustering the continuous emerging commodity stream data of e-commerce and logistics, a novel algorithm named as IUCStream (Increment Update Clustering Algorithm) is proposed.

The outline of the rest of this paper is as follows. In Section II, some basic concepts of data stream clustering algorithm are reviewed. In Section III, a measure of the commodity correlation algorithm and the IUCStream algorithm are proposed. Experimental studies on a real dataset are conducted in Section IV.

II. Data Stream Clustering Algorithm

A. The Correlation Measure of the Data Object

Calculating the similarity or diversity general uses the method of obtaining the relevant among data. The similarity means to measure similar between single or multiple attributes of two data objects. The similarity usually denotes with the non-negative number, and value range is between 0 and 1, where 0 means dissimilarity, while 1 means completely similar. Diversity means the distance between objects, using numerical to measure the distance between two data. Lower diversity means more similar. Diversity generally takes values between 0 and1, but it can also be in $[0,\infty]$.

(1) The similarity between data

If data have k attributes, then may regard each object as a single spot in k-dimensional space, and the whole data describe n data point of the entire spatial. Easy to comprehend, multiple data points exist in the same cluster should be near each other in this space, while exists in any two of the data points' distance should be the bigger the better. Therefore, our most immediate idea to judge similarity between the two is to measure the distance between the data points.

First, let $i=(X_{i1},X_{i2},...,X_{ik})$ and $j=(X_{j1},X_{j2},...,X_{jk})$ are two dataset with k attributes. So the distance between the data must have the following properties:

a. For any i and j, $d(i,j)\geq 0$, distance range must be non-negative numeric value. When $d(i,j)=0$, it means $i=j$.

b. For any i, the data of its own distance is defined as zero. Namely, $d(i,i)=0$, or did not express.

c. For any i and j, $d(i,j)=d(j,i)$. This demonstrates the symmetry of any distance between data.

d. For any i, j and m, $d(i,j)\leq d(i,m)+d(m,j)$, represents the direct distance between the two objects will never bigger than through a third object. It reflects the principle that straight line is the shortest between two points.

Considering the specific situation the differentiation of diversity, we use the Euclidean distance to measure similarity of data points.

(2) The similarity coefficient between the data object

In contrary, the similarity coefficient and the data similarity is proportional. Let

$$\overline{x_i}=\frac{1}{m}\sum_{k=1}^{m}x_{ik}$$

$$\overline{x_j}=\frac{1}{m}\sum_{k=1}^{m}x_{jk}$$

where x_{ik} means whether customer i purchase commodity k. x_{jk} means whether customer j purchase commodity k.

The similarity coefficient r_{ij} between X_i and X_j can denoted in following ways.

a. Scalar product method

$$r_{ij}=\begin{cases}1 & i=j\\ \dfrac{1}{M}\displaystyle\sum_{k=1}^{m}x_{ik}\times x_{jk} & i\neq j\end{cases} \tag{1}$$

where M is an integer and can be calculated by formula of

$$M\geq\max_{i,j}\left(\sum_{k=1}^{m}x_{ik}\times x_{jk}\right).$$

b. Included angle cosine method:

$$r_{ij}=\frac{\left|\displaystyle\sum_{k=1}^{m}x_{ik}\times x_{jk}\right|}{\sqrt{\left(\displaystyle\sum_{k=1}^{m}x_{ik}\right)^2\left(\displaystyle\sum_{k=1}^{m}x_{jk}\right)^2}} \tag{2}$$

c. Correlation coefficient method:

$$r_{ij}=\frac{\sum_{k=1}^{m}\left(x_{ik}-\overline{x_i}\right)\left(x_{jk}-\overline{x_j}\right)}{\sqrt{\sum_{k=1}^{m}\left(x_{ik}-\overline{x_i}\right)^2\times\sum_{k=1}^{m}\left(x_{jk}-\overline{x_j}\right)^2}} \quad (3)$$

d. Index similar method:

$$r_{ij}=\frac{1}{m}\sum_{k=1}^{m}e^{\frac{3\left(x_{ik}-x_{jk}\right)^2}{4S_k^2}} \quad (4)$$

Besides, there are a lot of similarity coefficients which used little. Such as minimum arithmetic average method, bottom absolute value method, non-parametric test method.

B. Binary Data Correlation

The correlation measure between data objects which only contain binary attributes also knows as computing similarity coefficient. These data typically values between 0 and 1 and the value 0 indicates that two data objects are completely different, while the value of 1 means completely similar.

Suppose x and y are two data which contain n binary attributes. When measure x and y, the following four conditions may occur:

f_{00} denotes the number of objects that $x=0, y=0$.

f_{01} denotes the number of objects that $x=0, y=1$.

f_{10} denotes the number of objects that $x=1, y=0$.

f_{11} denotes the number of objects that $x=1, y=1$.

The similarity coefficient between data can be express in following two ways [17, 18]:

a. Similarity matching coefficient (SMC):

$$SMC=\frac{f_{11}+f_{00}}{f_{01}+f_{10}+f_{11}+f_{00}} \quad (5)$$

So, SMC denotes the matching attributes ratio.

b. Jaccard coefficient:

$$J=\frac{f_{11}}{f_{01}+f_{10}+f_{11}} \quad (6)$$

III. ALGORITHM ANALYSIS

This section improves the traditional data stream algorithm slightly and proposes a novel increment update clustering algorithm (IUCStream) in view of the features of the products of e-commerce. The algorithm mainly includes two steps: the correlation analysis of the commodity data stream and the data stream clustering algorithm based on the correlation.

A. Commodity Correlation Analysis

Commodity correlation analysis can help the businessman to sell more goods. There are two ways to deal with it: one is handling shopping cart data, finding the association rules between goods. Another method is making cluster processing according to the different commodity relevant measure results, lays aside the high correlation in the same cluster, and to ensure that the correlation between the cluster and cluster as low as possible.

In addition, in measure correlation between goods most research literature using the computation the Euclidean distance method, while it is only suitable for the continual data. It is infeasible to judge whether several kinds of commodities appear in the same shopping cart.

Suppose the relevant analysis objects are m kind of different commodity. In view of i^{th} customer's purchase business, may use the vector $u_i=(x_{i1},x_{i2},...,x_{im})$ to denote. So in view of j^{th} kind of commodity, if has been purchased by the i^{th} customer, obtains $x_{ij}=1$, otherwise $x_{ij}=0$. If customer's quantity is n, then may use the vector $v_j=\left(x_{1j},x_{2j},...,x_{nj}\right)$ to denote the complete purchase business.

The vector represents purchases of the n^{th} customer for the j^{th} kinds of merchandise. If there is a lot of 1 in the vector, the information given is the probability of the j^{th} customer was bought is very high. Conversely, if there are a lot of 0 in the vector, then the j^{th} commodities are rarely purchased.

The above is for a commodity j to the analysis, the same token, the commodities j and k in data stream affairs, we can also study and the similarity between two vectors v_j and v_k to measure the similarity between j and k.

But this is only the easiest way to measure the correlation between two commodities. To e-commerce data mining, we want to get the number 1 in the two vectors, but not the number 0. Therefore, further metrics will be given:

$$s_{jk}=\frac{1}{n}\sum_{i=1}^{n}I\left\{x_{ij}=x_{ik}\right\} \quad (7)$$

Where $I\left\{x_{ij}=x_{ik}\right\}$ is the indicative function.

If $x_{ij}=x_{ik}$, then $I\left\{x_{ij}=x_{ik}\right\}=1$, otherwise $I\left\{x_{ij}=x_{ik}\right\}=0$. It can be seen s_{jk} is reflecting the ratio of commodity j and k is purchased or not purchased at the same time is occupied total number of people n. Therefore, if the value is large, it means big correlation. Otherwise express the correlation is small.

This method seems good. But if in view of e-commerce data, the value of s_{jk} can be very large. So we cannot simply rely on s_{jk} to determine the correlation between two commodities. And the situation both items have not been purchased can often happen.

These may cause $s_{jk}=\frac{1}{n}\sum_{i=1}^{n}I\left\{x_{ij}=x_{ik}\right\}=1$.

Thus, it should be improved.

$$s_{jk}=\frac{\sum_{i=1}^{n}I\left\{x_{ij}=x_{ik}=1\right\}}{\sum_{i=1}^{n}I\left\{x_{ij}+x_{ik}>0\right\}} \quad (8)$$

where x_{ij} and x_{ik} belongs to the binary data variable, the condition $x_{ij}+x_{ik}>0$ explain it must ensure at least one value of x_{ij} and x_{ik} is 1. In other words, one kind is purchased at least in commodities j and k.

$\sum_{i=1}^{n} I\{x_{ij} + x_{ik} > 0\}$ means the quantity of customers who has bought j or k or both. So s_{jk} express a ratio. It can be concluded that s_{jk} can reflect the correlation between j and k. Then through formula $d_{jk} = 1 - s_{jk}$ to measure the diversity distance between j and k.

B. The Correlation Analysis of Commodity Data Stream

In the large-scale e-commerce retail sites, relying solely on the correlation metrics is not enough. Because of real-time dynamics for e-commerce shopping data, we need a data stream algorithm with the dynamic update function, and then proposed an algorithm which in view of e-commerce data stream to calculate the correlation dynamically.

The related construction of data's construction is as follows:

(1) Shopping business matrix:

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

Where m stands for the number of commodity type, n stand for the number of customers.

(2) Similarity matrix:

$$\begin{bmatrix} 0 & & & \\ s_{21} & 0 & & \\ \dots & \dots & 0 & \\ s_{m1} & s_{m2} & \dots & 0 \end{bmatrix}$$

It is m line of m row, reflects the degree of similarity between each other in m attributes.

(3) Indicative function matrix:

$$\begin{bmatrix} 0 & & & \\ b_{21} & 0 & & \\ \dots & \dots & 0 & \\ b_{m1} & b_{m2} & \dots & 0 \end{bmatrix}$$

It is m line of m row, describe the indicative of elements in the similarity matrix.

After examination, we find the Divide-and-Conquer method is good when the dynamic analysis the similarity of the commodity.

First divide the entire shopping data stream according to the hypothesis time-gap. Then get a group of the data tuple which expands along with the time $x_1, x_2, ..., x_n, ...$

When meet a time-gap, integrate the new arrived data with the historical data, like this can obtain the similarity result effectively.

For the m commodities, involved in the purchase of the customer number was n by the time t_h.

Vector $u_i = (x_{i1}, x_{i2}, ..., x_{im})$ indicates the purchase data of the i^{th} customer.

If i buys goods j, then $x_{ij} = 1$, otherwise $x_{ij} = 0$.

Vector $v_j = (x_{1j}, x_{2j}, ..., x_{nj})$ describes whether goods j is purchased or not.

If it purchased by i, then $y_{ij} = 1$, otherwise $y_{ij} = 0$.

Combining u_i and v_j, making up shopping data stream matrices at time t_h:

$$X_{n\times m} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

If i^{th} customer buy j^{th} goods, then $x_{ij} = 1$, otherwise $x_{ij} = 0$.

Again structure the following matrix $B_{m\times m}$ and $S_{m\times m}$:

$$B_{m\times m} = \begin{bmatrix} 0 & & & \\ b_{21} & 0 & & \\ \dots & \dots & 0 & \\ b_{m1} & b_{m2} & \dots & 0 \end{bmatrix}$$

$$S_{m\times m} = \begin{bmatrix} 0 & & & \\ s_{21} & 0 & & \\ \dots & \dots & 0 & \\ s_{m1} & s_{m2} & \dots & 0 \end{bmatrix}$$

Then calculate the element values in $B_{m\times m}$ and $S_{m\times m}$.

where

$$b_{jk} = \sum_{i=1}^{n} I\{x_{ij} = x_{ik} = 1\} \tag{9}$$

$$s_{jk} = \frac{\sum_{i=1}^{n} I\{x_{ij} = x_{ik} = 1\}}{\sum_{i=1}^{n} I\{x_{ij} + x_{ik} > 0\}} \tag{10}$$

Matrix $S_{m\times m}$ means similarity matrices between every two commodities at time t_h.

Suppose increasing n' new purchase data from time t_h to time t_{h+1}, so the computation of commodities similarity at time t_{h+1} was as following:

(1) Build matrix $X'_{n\times m}$, $B'_{m\times m}$ and $S'_{m\times m}$:

$$X'_{n\times m} = \begin{bmatrix} x'_{11} & x'_{12} & \dots & x'_{1m} \\ x'_{21} & x'_{22} & \dots & x'_{2m} \\ \dots & \dots & \dots & \dots \\ x'_{n1} & x'_{n2} & \dots & x'_{nm} \end{bmatrix},$$

$$B'_{m\times m} = \begin{bmatrix} 0 & & \dots & \\ b'_{21} & 0 & \dots & \\ \dots & \dots & 0 & \dots \\ b'_{m1} & b'_{m2} & \dots & 0 \end{bmatrix},$$

$$S'_{m\times m} = \begin{bmatrix} 0 & & \dots & \\ s'_{21} & 0 & \dots & \\ \dots & \dots & 0 & \dots \\ s'_{m1} & s'_{m2} & \dots & 0 \end{bmatrix}.$$

The expression method of matrix $X'_{n\times m}$ is similar with $X_{n\times m}$. The computation principle of b'_{jk} and s'_{jk} is the same with b_{jk} and s_{jk}.

(2) Matrixing:

$$s_{jk} = \frac{b_{jk} + b'_{jk}}{b_{jk} / s_{jk} + b'_{jk} / s'_{jk}} \quad (11)$$

The new matrix $S_{m\times m}$ is the correlation matrix at the time t_{h+1}.

(3) Matrix operation: $b_{jk} = b_{jk} + b'_{jk}$.

Likewise may get the similarity matrix at any time spot.

C. Data Stream Clustering Algorithm which based on the Correlation

Aiming at the continuously increased data flow of shopping affairs, according to the specific time interval, the data stream will be divided into a series of units. This is the thought of divide-and-conquer.

When the new tuple X_i inflows, we can bring out the correlation between every two goods at this moment, using the algorithm given in the previous section. And then by initial clustering the unit, a cluster unit of waiting to update will be produced.

Realizing incremental updating clustering based on the existing clustering results R_{i-1}. The specific operation can be shown in Fig 1.

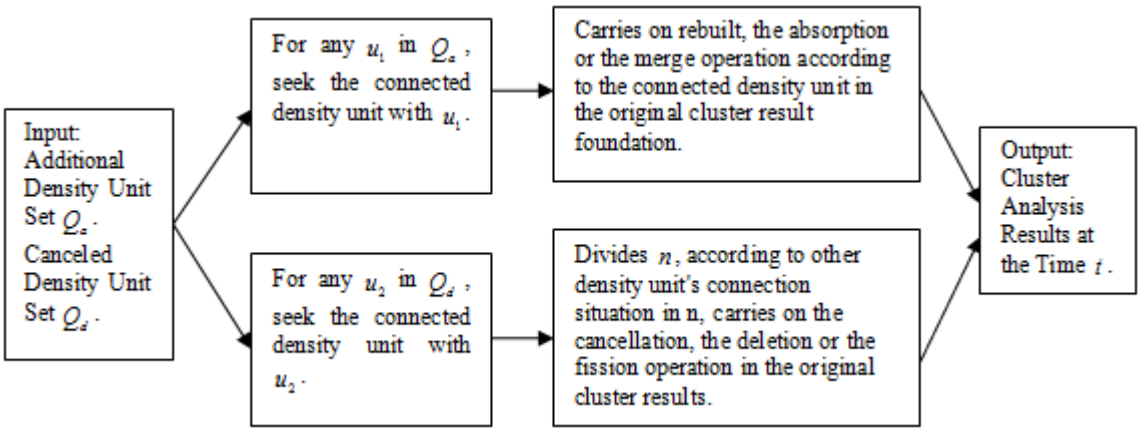

Figure. 1. The Flowchart of Algorithm of Increase Renewal of Clustering

(1) Cluster renewal algorithm I

Aimed at different situation, there are three operation methods for new cluster u:

a. Rebuilt:

Rebuilt a new cluster if belongs to no cluster.

b. Absorption:

If it belongs to cluster w, then put it in cluster w.

c. Merge:

If it belongs to cluster $w_1, w_2, ...w_k (k>1)$, then merge clusters $w_1, w_2, ...w_k$, and put u into the merged cluster.

When there is new cluster u arrives, repeat the above operation, concrete operation shows in algorithm I:

Input: A new density unit set $\{u_{i1}, u_{i2}, ..., u_{ik}\}$. The result $R_{i-1} = \{t_1, ..., t_s\}$, s is the quantity of clusters, clusters' represent letter $c_j = a_n ... a_1$. If $u_{im} \in cluster_j$, then $a_{im} = 1$, otherwise $a_{im} = 0$.

Output: The new result $R_i = \{c_1, ..., c_{s'}\}$, s' is the quantity of clusters.

The pseudo code of the algorithm is shown as follows:

for any u_{ij} do
  $flag = false$
for each c do
  if there is one or more clusters that similar with u_{ij} in c
  then
  if $flag = false$
  then /*case 2 absorption*/
    put u_i in c
    $flag = true$
    $temp\_c = c$
  else /*case 3 merge*/
    merge $temp\_c$ and c into a new cluster c'
    $temp\_c = c'$
if $flag = false$
then
  rebuilt a cluster which including u_{ij}

(2) Cluster renewal algorithm II (based on the canceled density unit set)

Aimed at canceling density unit set u, existing three circumstances:

a. Fission:

There are disjunct density units in the cluster, then divide it.

b. Cancellation:

There is no density units in the cluster, then cancel it.

c. Deletion:

If other units within the cluster is associated, do not make any other changes except delete u.

The operation based on the canceled cluster shows in algorithm II:

Input: for canceled cluster $\{u_{i1}, u_{i2}, ..., u_{ik}\}$, when the result $R_{i-1} = \{t_1, ..., t_s\}$, and s is the quantity of clusters,

clusters' represent letter $c_j = a_n...a_1$, if $u_{im} \in cluster_j$, then $a_{im} = 1$, otherwise $a_{im} = 0$.

Output: the new result $R_i = \{c_1,...c_{s'}\}$, s' is the quantity of clusters.

The pseudo code of the algorithm is shown as follows:

```
for any u_ij do
        flag = false
for each c do
        delete unit from c
if c's represent letter equal 0
then
        cancel c
else
        divide c into n clusters
if n>1 then
        divide c into n clusters
else
        delete c
```

We can get the renewal cluster R_i through the above two steps.

IV. Experimental Analysis

This paper used the example of data from Dangdang nearly a week sales record on September 9. In order to fit this paper, we select the best-selling of 20 pieces IT products (see table 1), because involves the confidential commercial information to the website, it could only be referenced related literature data simulation. Trade data this paper needs is all the shopping cart data for some periods, namely commodities a shopping cart contains.

The top 20 high frequency IT merchandises customers buy are numbered as TABLE I.

TABLE I.

THE TOP 20 KINDS OF HIGH FREQUENCY IT MERCHANDISES CUSTOMERS BUY

Number	1	2	3	4	5
Commodity	USB Flash Disk	Wireless Router	Mouse	Radiator	Mobile HDD
Number	6	7	8	9	10
Commodity	Earphone	Keyboard	Memory Card	DVD	IPAD
Number	11	12	13	14	15
Commodity	Printer	Headset	Laptop	Bluetooth Headset	Digital Camera
Number	16	17	18	19	20
Commodity	Original Battery	Smartphone	Audio	MP4	Photographic Paper

Through the shopping cart information about related these 20 products to analyze the correlation between the 20 products, calculate similarity of simulated transaction data with the algorithm given by section III. Finally, we get the similarity between various products which is shown as TABLE II and TABLE III.

Using the new proposed algorithm IUCStream to deal with the above data stream, get the clustering results of 20 kinds of IT commodities, as showed in figure 2.

In Fig. 2, the highest similarity was Mouse and Laptop, tally with the actual situation. Then the Wireless Router and Headset, both items are accessory products of the laptop. So the probability of purchase is both large. Mobile HDD and DVD also reflect a high similarity is most remarkable. Diversity between the two can be 0.88. That is the relevance $s_{jk} = 1 - 0.88 = 12\%$. It means if a customer buys mobile HDD or DVD, there is 12% of the probability will also buy another.

TABLE II.

THE SIMILARITY AMONG 20 IT MERCHANDISES (I)

Number	1	2	3	4	5	6	7	8	9	10
1	–	0. 004	0. 002	0. 0021	0. 0022	0. 0019	0. 0019	0. 0025	0. 0026	0. 0023
2	0. 004	–	0. 0021	0. 0021	0. 002	0. 0018	0. 0019	0. 0025	0. 0026	0. 0024
3	0. 002	0. 0021	–	0. 033	0. 033	0. 003	0. 0031	0. 0038	0. 0037	0. 0038
4	0. 0021	0. 0021	0. 033	–	0. 007	0. 0033	0. 0032	0. 0035	0. 0036	0. 0037
5	0. 0022	0. 002	0. 033	0. 007	–	0. 0032	0. 0034	0. 0035	0. 0037	0. 0036
6	0. 0019	0. 0018	0. 003	0. 0033	0. 0032	–	0. 12	0. 0025	0. 0027	0. 0029
7	0. 0019	0. 0019	0. 0031	0. 0032	0. 0034	0. 12	–	0. 0025	0. 0026	0. 0029
8	0. 0025	0. 0025	0. 0038	0. 0035	0. 0035	0. 0025	0. 0025	–	0. 02	0. 0058
9	0. 0026	0. 0026	0. 0037	0. 0036	0. 0037	0. 0027	0. 0026	0. 02	–	0. 0057
10	0. 0023	0. 0024	0. 0038	0. 0037	0. 0036	0. 0029	0. 0029	0. 0058	0. 0057	–
11	0. 0023	0. 0022	0. 0037	0. 0037	0. 0035	0. 0028	0. 0027	0. 0057	0. 0057	0. 005
12	0. 0023	0. 0022	0. 0036	0. 0038	0. 0037	0. 003	0. 0028	0. 006	0. 0058	0. 005
13	0. 0026	0. 0026	0. 0034	0. 0035	0. 0034	0. 0033	0. 0031	0. 0053	0. 0055	0. 02
14	0. 0026	0. 0025	0. 0033	0. 0036	0. 0032	0. 0034	0. 0031	0. 0054	0. 0054	0. 019
15	0. 0025	0. 0026	0. 0034	0. 0034	0. 0032	0. 0033	0. 0032	0. 0053	0. 0055	0. 019
16	0. 0031	0. 0031	0. 0032	0. 0034	0. 0030	0. 0026	0. 0023	0. 0060	0. 0059	0. 015
17	0. 0032	0. 0032	0. 0033	0. 0034	0. 0031	0. 0026	0. 0024	0. 0061	0. 0060	0. 0149

18	0. 0030	0. 0033	0. 0032	0. 0032	0. 0032	0. 0025	0. 0023	0. 0059	0. 0059	0. 0152
19	0. 0034	0. 0035	0. 0038	0. 0038	0. 0037	0. 0023	0. 0022	0. 0036	0. 0036	0. 0033
20	0. 0034	0. 0036	0. 0038	0. 0039	0. 0036	0. 0022	0. 0021	0. 0036	0. 0035	0. 0034

TABLE III.

THE SIMILARITY AMONG 20 IT MERCHANDISES (II)

Number	11	12	13	14	15	16	17	18	19	20
1	0. 0023	0. 0023	0. 0026	0. 0026	0. 0025	0. 0031	0. 0032	0. 003	0. 0034	0. 0034
2	0. 0022	0. 0022	0. 0026	0. 0025	0. 0026	0. 0031	0. 0032	0. 0033	0. 0035	0. 0036
3	0. 0037	0. 0036	0. 0034	0. 0033	0. 0034	0. 0032	0. 0033	0. 0032	0. 0038	0. 0038
4	0. 0037	0. 0038	0. 0035	0. 0036	0. 0034	0. 0034	0. 0034	0. 0032	0. 0038	0. 0039
5	0. 0035	0. 0037	0. 0034	0. 0032	0. 0032	0. 0030	0. 0031	0. 0032	0. 0031	0. 0036
6	0. 0028	0. 003	0. 0033	0. 0034	0. 0033	0. 0026	0. 0026	0. 0025	0. 0024	0. 0022
7	0. 0027	0. 0027	0. 0031	0. 0031	0. 0032	0. 0023	0. 0024	0. 0023	0. 0022	0. 0021
8	0. 0057	0. 006	0. 0053	0. 0054	0. 0053	0. 0060	0. 0061	0. 0059	0. 0036	0. 0036
9	0. 0057	0. 0058	0. 0055	0. 0054	0. 0055	0. 0059	0. 0060	0. 0059	0. 0036	0. 0035
10	0. 005	0. 005	0. 02	0. 019	0. 019	0. 015	0. 0149	0. 0152	0. 0033	0. 0034
11	–	0. 013	0. 019	0. 02	0. 021	0. 0155	0. 0148	0. 0152	0. 0031	0. 0034
12	0. 013	–	0. 021	0. 0019	0. 02	0. 0158	0. 0155	0. 015	0. 0031	0. 0034
13	0. 019	0. 021	–	0. 085	0. 084	0. 015	0. 015	0. 0153	0. 0032	0. 0031
14	0. 02	0. 019	0. 085	–	0. 14	0. 0158	0. 0159	0. 0148	0. 0031	0. 0032
15	0. 021	0. 02	0. 084	0. 14	–	0. 0139	0. 016	0. 0146	0. 0031	0. 0032
16	0. 0155	0. 0158	0. 015	0. 0158	0. 0139	–	0. 015	0. 014	0. 0037	0. 0038
17	0. 0148	0. 0155	0. 015	0. 0159	0. 016	0. 015	–	0. 042	0. 0038	0. 0037
18	0. 0152	0. 015	0. 0153	0. 0148	0. 0146	0. 014	0. 042	–	0. 0038	0. 0036
19	0. 0031	0. 0031	0. 0032	0. 0031	0. 0031	0. 0037	0. 0038	0. 0038	–	0. 031
20	0. 0034	0. 0034	0. 0031	0. 0032	0. 0032	0. 0038	0. 0037	0. 0036	0. 031	–

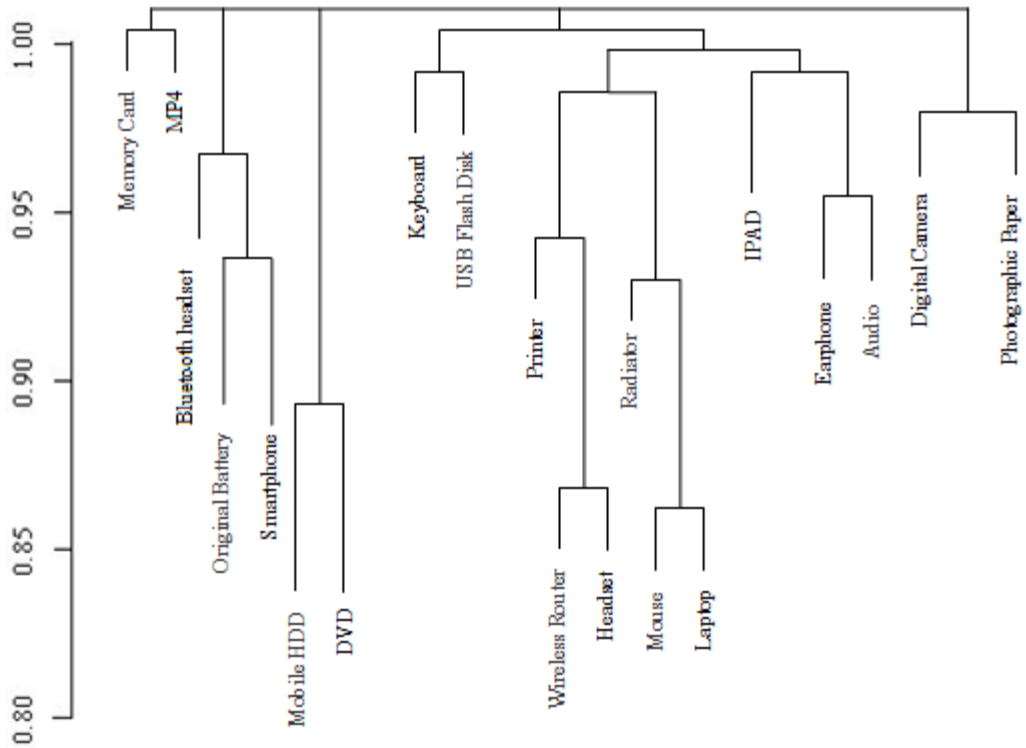

Figure 2. Tree View of Clustering Results, the 20 Top Frequency of IT Merchandises are Divided into 5 Clusters

V. CONCLUSIONS

Based on the network retail goods as the research object, this paper puts forward a dynamic update goods similarity algorithm. The proposed algorithm is quantitative calculate the correlation of the goods to the data stream processing methods into the algorithm. Then put forward a new kind of incremental data stream clustering algorithm for the purchase of the continuously update data stream of goods affairs clustering. The

algorithm adopts the idea of Divide-and-Conquer, and compared with the traditional algorithm, it is an incremental updating dynamic clustering analysis, saving time and memory, and other resources. The proposed algorithm can effectively help logistics enterprise arrange the goods warehouse reasonably and save storage time.

ACKNOWLEDGMENTS

This paper was partly supported by the Commonweal Technology Project of Zhejiang Province, China under Grant No. 2011C23076. We are interested in empirical researching of e-commerce and logistics optimization. And the authors also wish to thank Dangdang Company for the data source.

REFERENCES

[1] Han J. and Kamber M. Data Mining: Concepts and Techniques (Second Edition). Morgan Kaufmann, Elsevier Inc., 2006, pp. 467-589.
[2] Henziger M. R. Raghavan P., Rajagopalan S. Computing on data streams. SRC Technical Note 1998-011. Digital systems research center: Palo Alto. California.1998.
[3] Ester M., Kriegel H-P., Sander J., Incremental Clustering for Mining in a Data Warehousing Environment, in: Proceedings of the 24th International Conference on VLDB, New York, 1998, pp. 323-333.
[4] Qiang Niu, Xinjian Huang. An Improved Fuzzy C-means Clustering Algorithm based on PSO, Journal of Software, vol. 6, No. 5, May 2011, pp. 873-879.
[5] Li Xinwu. A New Text Clustering Algorithm Based on Improved *K*-means, Journal of Software, vol. 7, No. 1, Jan. 2012, pp. 95-101.
[6] Aggarwal C. C., Han J, Wang J, et al. A Framework for Clustering Evolving Data Streams. Proceedings of the 29th International Conference on Very Large Data Bases, 2003, pp. 81–92.
[7] Cao F, Ester M, Qian W, et al. Density-Based Clustering over an Evolving Data Stream with Noise. Proceedings of the 6th SIAM International Conference on Data Mining, 2006, pp. 328–339.
[8] Bhatnagar V., Kaur S. Exclusive and Complete Clustering of Streams. Springer-Verlag Berlin Heidelberg, 2007, pp. 629-638.
[9] Aggarwal C. C, Yu P. S. A Framework for Clustering Uncertain Data Streams. Proceedings of the 24th International Conference on Data Engineering, 2008. pp. 150–159.
[10] Luehr Sebastian, Lazarescu Mihai. Incremental Clustering of Dynamic Data Streams Using Connectivity Based Representative Points. Data & Knowledge Engineering. 2009, 68 (1) : pp. 1-27.
[11] Bhatnagar Vasudha, Kaur Sharanjit, et al. A Parameterized Framework for Clustering Streams. International Journal of Data Warehousing and Mining, 2009, 5 (1) : pp. 36-56.
[12] Li Hua-Fu, Lee Suh-Yin. Mining Frequent Item Sets over Data Streams Using Efficient Window Sliding Techniques. Expert System with Applications, 2009, 36 (2) : pp. 1466-1477.
[13] Cao Fuyuan, Liang Jiye, Bai Liang, et al. A Framework for Clustering Categorical Time- Evolving Data. IEEE Trans. On Fuzzy Systems, 2010, 18 (5) : pp. 872-882.
[14] Bae Eric, Bailey James et al. A Clustering Comparison Measure Using Density Profiles and Its Application to the Discovery of Alternate Clusterings. Data Mining and Knowledge Discovery, 2010, 21(3) : pp. 427-471.
[15] Skala J., Kolingerova I. Dynamic Hierarchical Triangulation of a Clustered Data Stream. Comptures & Geosciences, 2011, 37 (8) : pp. 1092-1101.
[16] Brice Pierre, Jiang Wei, et al. A Cluster-Based Context-Tree Model for Multivariate Data Streams with Applications to Anomaly Detection. Informs Journal on Computing, 2011, 23 (3) : pp. 364-376.
[17] L. Kaufman, P. J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics. John Wiley and Sons, New York, November, 1990.
[18] P. H. A. Sneath, R. R. Sokal. Numerical Taxonomy. Freeman, SanFranciSCO, 1999.

Peihua Fu (1966-) received the BS degree in applied electronics from Zhejiang University, China, in 1987, and the Master Degree in power electronics from Zhejiang University, China, 1989. He is currently a full professor of applied computer and logistics management at Zhejiang Gongshang University of Hangzhou, China.

His research interests are mainly in using computer simulation technologies and computing algorithms to solve logistics and supply chain management problems. Recently, He has published 4 research books and more than 20 research publications.

Professor Fu is a member of Association of Automation of Zhejiang Province, China, and the vice-president of College of Computer and Information Engineering of Zhejiang Gongshang University, China.

Yangfei Chen is currently a graduate student in computer and information engineering at Zhejiang Gongshang University, Hangzhou, P.R.China. She receives the BS degree in E-business from Zhejiang Gongshang University, China in 2010.

She has published one paper in CICA in 2010. Her research interests include data stream, clustering algorithm, complex network and supply chain management.

Hongbo Yin received the MS degree in logistics technology and management in College of Computer and Information Engineering of Zhejiang Gongshang University, China, in 2011. His research interests were mainly in supply chain management and logistics management.