

Cluster Analysis of E-Commerce Sites with Data Mining Approach*

Yongyi Cheng^{1,2}, Yumian Yang³, Jianhua Jiang³ and GaoChao Xu¹

¹College of Computer Science and Technology, Jilin University

²Jilin Nongxin Information Technology Service Co., Ltd.

³School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun, China

echojlu@163.com, 1083827461@qq.com, jianhuajiang@foxmail.com, xugc@jlu.edu.cn

Abstract

With the rapid development of E-Commerce, how to evaluate the E-Commerce sites accurately has become an important issue. However, to cluster E-Commerce sites correctly and accurately is not an easy thing based on characteristics of high dimensions and uneven density for E-Commerce sites. This leads to bad performance of the cluster result. To analyze 100 E-Commerce demonstration enterprises in 2013-2014 named by the Ministry of Commerce People's Republic of China, this paper adopts a data mining approach of DBSCAN method. In the data preprocessing phase, it adopts factor analysis to reduce dimensionality. In the cluster phase, this paper implements an improved DBSCAN algorithm to process the uneven density data. Finally, this paper gives suggestions to these 100 E-Commerce enterprises based on experiment results.

Keywords: Commerce sites; Factor analysis; DBSCAN; Cluster analysis

1. Introduction

According to the report released by the Ministry of Commerce People's Republic of China in 2013, by this year, the national E-Commerce transactions are more than 10200 billion, and the number of E-Commerce service enterprises is more than 41303[1]. E-Commerce sites as an important platform for enterprises to carry out a variety of business activities have become the focus of the development of enterprise and an important symbol that evaluates the development level of E-Commerce service companies. Therefore, E-Commerce sites evaluation is an urgent issue that should be solved.

To solve the issue of E-Commerce sites evaluation, scholars at home and broad evaluate the E-Commerce sites with several kinds of methods, such as factor analysis[2], fuzzy comprehensive evaluation method[3], network measurement[4], analytic hierarchy process (AHP)[5], and so on. Factor analysis approach focuses on getting the influencing factors that determine the development level of E-Commerce site. Fuzzy comprehensive evaluation method[3] adopts mathematic method to generate a reasonable model to evaluate E-Commerce sites. However, these above methods don't take their similarity into account when evaluating E-Commerce sites. Data mining approach[6] is considered as an effective method to solve the issue of E-Commerce similarity in their evaluation. DBSCAN algorithm[7-8] is an effective approach to solve cluster issue for different E-

*Jianhua Jiang is corresponding author. This research has received financial support by the National Natural Science Foundation of China (No. 61202306, 61170004, 61472049, 61402193), by the Foundation of Education Bureau of Jilin Province (No. 2012188), and by the Foundation of Jilin University of Finance and Economics (No. XJ2012007, 2013006).

Commerce sites. And because of the advantages of DBSCAN, it is also used in other fields. In 2010, American scholars Kisilevich S et al. adopted P-DBSCAN algorithm to cluster and analyze its location and events according to the collected pictures[9]. In 2011, Chinese scholars Ou Yangjia clustered and analyzed the content of wage page with DBSCAN[10]. In 2012, DBSCAN algorithm was applied to the classification of telecom customers by Chinese scholars Zuo Guocai et al.[11]. In 2014, Australian scholars Kieu L M et al. unraveled travel regularity of per passenger with DBSCAN algorithm[12].

The data of selected 100 E-Commerce sites have characteristics of high-dimension and uneven density. Though DBSCAN algorithm can cluster E-Commerce sites and process high-dimensional data, it cannot get effective result when handling data of uneven density. And for high-dimensional data, every factor ingredient is not important equally. If the weight of each factor is ignored when E-Commerce sites are clustered by DBSCAN, then that will affect the accuracy of cluster result. However, factor analysis can reduce dimension, and some scholars adopts this algorithm in some other fields. In 2010, European scholars Hannu Verkasalo, Carolina Lopez -Nicolas carried out the analysis of the influencing factors when clients select application program of smart phone with factor analysis method and found out the main factors urging the clients to use the application[13]. In 2010, Chinese scholar Chen Wencheng studied the development imbalance of regional economy with factor analysis[14]. In 2013, American scholars Taufiq Hasan, John h. I. Hansen et al, adopted factor analysis method to analyze and verify the voice[15]. Besides dimensionality reduction, this paper adopts the methods to get reasonable weights of Euclidean distance. Therefore, a new distance formula is proposed and shown as following.

$$D(i, j) = \sqrt{\alpha(F_1(i) - F_1(j))^2 + \beta(F_2(i) - F_2(j))^2 + \gamma(F_3(i) - F_3(j))^2}$$

For uneven density, the paper divides data into pieces based on different densities, and then clusters respectively.

For the disadvantages of DBSCAN, the paper improves DBSCAN. Improved DBSCAN can process uneven data effectively and gain an accurate cluster result. The main idea of the paper: firstly, we adopt factor analysis to reduce dimension and get suitable weights for distance formula; secondly, we improve DBSCAN by dividing data into pieces based on densities, which makes the improved DBSCAN handle data of uneven density effectively and solve the issue of similarity measurement. Therefore, the key contributions in the paper are as follows:

(1) In the data preprocessing phase, principal component analysis is adopted to find the major influencing factors to different E-Commerce sites. These major influencing factors will lead to the effectiveness of similarity measurement among different E-Commerce sites.

(2) In the DBSCAN algorithm, uneven density data are classified into different classes with different densities. Compared with traditional DBSCAN algorithm, the proposed DBSCAN algorithm can handle uneven density data.

(3) Based on the characteristics of E-Commerce sites, a new similarity measurement function is proposed to evaluate the similarity of different E-Commerce sites.

The rest of this paper is organized as follows: Section 2 expands the construction of index system and data preprocessing with factor analysis; Section 3 introduces the traditional DBSCAN algorithm and confirms the shortage that it processes the uneven density data. Based on the disadvantage, an improved DBSCAN is designed, which expands the application field of the algorithm. The main part of this paper is section 4 which describes the evaluation of the E-Commerce sites in detail, points out their disadvantages and then proposes relative solutions. At last, the summary is made and future work is given.

2. Data Collection and Factor Analysis

2.1. Data Collection

In the September of 2013, the Ministry of Commerce People's Republic of China announced that there are 100 E-Commerce demonstration enterprises from 2013 to 2014. The 100 enterprises are the leading E-Commerce companies in China. This paper selects the 100 demonstration enterprises as samples and evaluates their development situation.

2.2. Index Variables Selection

If all the factors which have influence on the E-Commerce sites are taken into account, this is not a good choice. Based on HOU Zhi-Ping[16] and LI Jun-Jun[17]'s suggestion, this paper selects 5 index variables as the E-Commerce sites evaluation criteria: *Access per Million Users, Page Views Per User, Site Linking In, Speed, and Website Size*[16-17].

2.3. Factor Analysis

In April 2014, this paper got the data of the 100 E-Commerce demonstration enterprises by visiting Alexa[18]. Details are shown in the Table 1.

In order to get the standardized data that DBSCAN can process, this paper applies the SPSS13.0 to conduct the factor analysis of the raw data which comes from the E-Commerce sites.

2.3.1. Normalization of Raw Data: This paper applies the method of calculating Z-score value to normalize the raw data. The normalized data are shown in Table 2.

2.3.2. Factor Analysis in SPSS: According to Table 2, this paper selects 5 index variables to conduct KMO and Bartlett's Test of Sphericity. The results are shown in Table 3. KMO statistic is 0.539(>0.5), and the observed Bartlett's Test of Sphericity statistic whose significant level is 0.0000(<0.01), and Aprox Chi-square is 183.467. Therefore, the data of 100 E-Commerce sites can be analyzed by factor analysis.

Table 1. The Raw Data of 100 National E-Commerce Sites (Some Samples)

| ID | Website Address | Access Per Million Users | Page Views Per User | Site Linking In | Speed | Website Size |
|-----|-----------------|--------------------------|---------------------|-----------------|-------|--------------|
| 1 | jd.com | 3828000 | 15.03 | 7913 | 0.357 | 743000000 |
| 2 | dangdang.com | 534000 | 7.54 | 15306 | 0.224 | 13400000 |
| 3 | hc360.com | 414000 | 3.5 | 23103 | 0.862 | 39300000 |
| 4 | xiaomi.com | 474000 | 5.4 | 3187 | 0.172 | 6880000 |
| 5 | vancl.com | 158400 | 6.93 | 8239 | 0.366 | 1400000 |
| 6 | gome.com.cn | 10800 | 4.18 | 4243 | 0.877 | 12000000000 |
| ... | ... | ... | ... | ... | ... | ... |
| 100 | xj359.com | 4 | 1 | 3 | 1.346 | 274 |

Table 2. The Normalized Data (Partial Data)

| ID | Website Address | Access Per Million Users | Page Views Per User | Site Linking In | Speed | Website Size |
|----|-----------------|--------------------------|---------------------|-----------------|---------|--------------|
| 1 | jd.com | 3.43429 | 3.20657 | 0.53679 | 0.42916 | 0.45551 |
| 2 | dangdang.com | 0.29416 | 1.00364 | 1.30587 | 0.46682 | -0.14824 |
| 3 | hc360.com | 0.17976 | -0.18459 | 2.11698 | 0.28614 | -0.12680 |
| 4 | xiaomi.com | 0.23696 | 0.37423 | 0.04515 | 0.48155 | -0.15363 |

| | | | | | | | |
|-----|-------------|----------|----------|----------|-----|---------|----------|
| 5 | vanc1.com | -0.06390 | 0.82423 | 0.57070 | - | 0.42661 | -0.15817 |
| 6 | gome.com.cn | -0.20460 | 0.01541 | 0.15501 | - | 0.28190 | 9.77064 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 100 | xj359.com | -0.21490 | -0.91988 | -0.28607 | - | 0.14908 | -0.15932 |

Table 3. KMO and Bartlett's Test

| Kaiser-Meyer-olkin Measure of Sampling | Bartlett's Test of sphericity | | |
|--|-------------------------------|---------|-------|
| | Aprox square | Chi- df | Sig. |
| 0.529 | 183.467 | 10 | 0.000 |

Table 4.Total Variance Explained

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | |
|-----------|---------------------|---------------|--------------|-------------------------------------|---------------|--------------|--|
| | Total | % of variance | Cumulative % | Total | % of variance | Cumulative % | |
| 1 | 2.077 | 41.548 | 41.548 | 2.077 | 41.548 | 41.548 | |
| 2 | 1.112 | 22.249 | 63.797 | 1.112 | 22.249 | 63.797 | |
| 3 | 0.979 | 19.587 | 83.383 | 0.979 | 19.587 | 83.383 | |
| 4 | 0.742 | 14.837 | 98.221 | | | | |
| 5 | 0.089 | 1.779 | 100.000 | | | | |

From Table 3, we can see that this paper gets the common factor's character root contribution rate and accumulated contribution rate by applying the SPSS software, as is shown in Table 4. From the Table 4, the first three common factors can explain the 83.383% of the total variance. So the first three common factors can be used to replace the original 5 index variables. In other words, three common factors are extracted. It can be observed that the first three common factor's construction rates are 41.54%, 22.249% and 19.587%. By rotating them with the method of Varimax, the loading matrix of the three common factors can be calculated, as is shown in Table 5. And then, we design the factor analysis model of index system of the E-Commerce sites evaluation , as is shown below.

$$\begin{aligned}
 X_1 &= 0.968F_1 + 0.111F_2 - 0.010F_3 \\
 X_2 &= 0.963F_1 + 0.099F_2 + 0.026F_3 \\
 X_3 &= 0.048F_1 - 0.851F_2 - 0.018F_3 \\
 X_4 &= 0.10F_1 + 0.031F_2 + 0.999F_3 \\
 X_5 &= 0.259F_1 + 0.697F_2 + 0.022F_3
 \end{aligned}$$

According to the Table 5, the first common factor summarizes the *Access Per Million Users* and *Site Linking In*; the second common factor results from the *Page Views Per User* and *Speed*; the third common factor represents the *Website Size*. The data of the principal components are calculated based on the normalized data of 100 demonstration E-Commerce sites and is shown in Table 6.

Table 5. Component Matrix

| | Component | | |
|---|-----------|--------|--------|
| | 1 | 2 | 3 |
| <i>Access Per Million Users</i> (X_1) | 0.968 | 0.111 | -0.010 |
| <i>Site Linking In</i> (X_2) | 0.963 | 0.099 | 0.026 |
| <i>Speed</i> (X_3) | 0.048 | -0.851 | -0.018 |
| <i>Website Size</i> (X_4) | 0.10 | 0.031 | 0.999 |
| <i>Page Views Per User</i> (X_5) | 0.259 | 0.697 | 0.022 |

Table 6. Extract the Data of Principal Ingredients (Partial Data)

| ID | Website Address | F ₁ | F ₂ | F ₃ |
|-----|-------------------|----------------|----------------|----------------|
| 1 | jd.com | 2.06768 | 1.88019 | 0.33422 |
| 2 | dangdang.com | 0.77844 | 0.80859 | -0.16085 |
| 3 | made-in-china.com | 0.67351 | 0.11248 | 0.39347 |
| 4 | 1688.com | 9.09133 | -0.07975 | -0.23032 |
| 5 | vip.com | 0.18982 | 1.37171 | 0.27111 |
| 6 | gome.com.cn | -0.13083 | -0.01901 | 9.77718 |
| ... | ... | ... | ... | ... |
| 100 | xj359.com | -0.30975 | -0.36937 | -0.14981 |

3. DBSCAN for Uneven Density Data Processing

Firstly, this section makes a brief introduction on the traditional DBSCAN and clusters these 100 demonstration E-Commerce enterprises. Since the data of these enterprises have characteristics of uneven density, the traditional DBSCAN algorithm results in a large number of outliers which are not reasonable. Then for the disadvantages of the traditional DBSCAN algorithm, an improved DBSCAN is proposed. Finally, this paper clusters 100 demonstration E-Commerce sites with the improved DBSCAN algorithm.

3.1.Traditional DBSCAN Algorithm and Its Application

DBSCAN algorithm is a density-based spatial clustering algorithm. It can discover clusters of arbitrary shape and handle outliers effectively[19]. The basic idea is that the number of data points must be greater than the minimum number *MinPts* in a sphere of given radius (*Eps*)[20].

The clustering process of the algorithm can be listed as follows: firstly find an arbitrary point *P* from a database *D*, and find all the points which are density-reachable from the point of *P* on *Eps* and *MinPts*. If *P* is the core point, we will find a cluster about the parameters of *Eps* and *MinPts*. If *P* is a border point, then no point is density-reasonable from *P*. *P* is temporarily labeled as noise points. Then DBSCAN handles next point in database *D*[21]. To classify the DBSCAN with our proposed DBSCAN algorithm, we name this DBSCAN algorithm as traditional DBSCAN algorithm.

In Weka software[22], by several trials, when setting the parameter values of *Eps* is 1 and of *Minpts* is 4, we can get a good clustering resut. The result that 100 national demonstration E-Commerce sites are clustered by traditional DBSCAN (*Eps*=1, *MinPts*=4) is shown in Table 7.

Table 7. Clustering Result of the Original DBSCAN

| Cluster ID | Number of Sites | E-Commerce Sites |
|------------|-----------------|---|
| 1 | 65 | womai.com mi.com vancl.com oysd.cn yhd.com suning.com maimaibao.com hodo.cn iaijiayue.com inzuol100.com...xj359.com |
| 2 | 4 | 0316366.com eipin365.com 1872.net xjguoguo.com |
| Outliers | 31 | 6666688888.com vip.com 1688.com, gome.com.cn dangdang.com 58.com jd.com ... made-in-china.com |

According to Table 6 and 7, in terms of outliers, 1688.com and gome.com.cn have an index data which are more than 9 and apparently higher than the other index data. And the two are different index. So the two points are labeled as obvious outliers. However, According to the data in Table 6, there are some other outliers such as jd.com, vip.com,

made-in-china.com, and they have been found that their data are quite similar. If points labeled as outliers are singly clustered, some points with similar data can be divided into the same cluster. Therefore, the clustering results are not satisfactory, when the traditional DBSCAN processes the uneven density data. Thus, if the traditional DBSCAN is still applied to deal with uneven density data, it must be improved to ensure the accuracy of clustering and the reasonability of the analysis.

To further prove the traditional DBSCAN algorithm unable to deal with the uneven density data, the figure of the E-Commerce sites distribution is given by Matlab software and the result is shown in Fig.1. To indicate different density in the 100 E-Commerce sites, we label each site with similar density with same color. Based on fig. 1, we can see that data distribution is uneven. If this paper set only global parameters (*Eps* and *MinPts*) based on the density of the data labeled in green, the data will be divided into three categories by the traditional DBSCAN, and its result can be seen in Table 7. However, if this paper set a global parameter with an extreme parameter, such as the density of the data in blue in fig. 1, there will be a large number of outliers. Thus for the uneven density data, we can't get desired clustering results by the traditional DBSCAN. In terms of the problem, it is a practical way to set different local parameters based on the different density. On this basis, this paper can improve the traditional DBSCAN so that it may handle the uneven density data effectively.

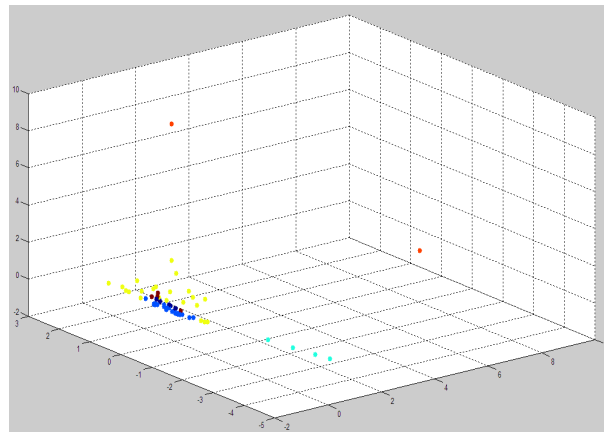


Figure 1. 100 E-Commerce Sites Data Distribution

3.2. Proposed Improved DBSCAN and its Application

Since the 100 demonstration E-Commerce sites data have the characteristics of uneven density, firstly these E-Commerce sites need to be filtered based on density. Then data of different levels of density are clustered by DBSCAN algorithm with enhanced similarity measurement. Therefore, the basic steps of the improved DBSCAN can be listed as follows:

- (1) Partition E-Commerce sites by density detection method. Density variability of 100 demonstration E-Commerce sites should be detected with density detection method.
- (2) Cluster respectively based on different density with enhanced similarity measurement. After partition of these 100 E-Commerce sites, proposed DBSCAN algorithm with enhanced E-Commerce site similarity measurement is applied to get a reasonable cluster result.

3.2.1. Partition of E-Commerce Sites with Density Detection Method

This paper adopts two DBSCAN holistic Parameters (*Eps* and *MinPts*) to filter the different density data. Firstly, parameter value of *MinPts* is set to 4. For a random E-Commerce site, we measure the distance of nearest 4 neighbor E-Commerce sites. For these nearest 4 distances, we get *Eps* of this E-Commerce site. Secondly, we check its

nearest 3 neighbor sites to get their *Eps*. Finally, *Eps* of each E-Commerce site is calculated, and then E-Commerce sites can be divided into different classes with their *Eps*. These above steps can be summarized as follows.

Algorithm 1 Improved DBSCAN algorithm

Input: 100 E-Commerce sites data, *MinPts* = 4

Output: the clustering results

1. Calculate distance from *k* to its 3 neighbors $\{m,n,p\}$, $Distance(k,y)$, $y \in m,n,p$.
2. Set the *Eps* value for node *k*: $Eps = \max(Distance(k,y))$, $y \in m,n,p$.
3. Put the *Eps* value in descending order.
4. Classes of E-Commerce sites are determined by different density areas which are merged based on the ratio of the nearby two *Eps* values. If its ratio is large enough, it suggests that different density data are captured.
5. Different clusters are generated from each class of E-Commerce sites by DBSCAN with enhanced similarity measurement.

3.2.2. Cluster Respectively based on Different Density with Enhanced Similarity Measurement

In traditional DBSCAN algorithm, similarity can be measured with Euclidean distance function. However, Euclidean distance neglects the weights of three influencing factors. Based on Table 5, the first common factor summarizes the *Access Per Million Users* and *Site Linking In*; the second common factor results from the *Page Views Per User* and *Speed*; the third common factor represents the *Website Size*. For E-Commerce sites, we can know that *Access Per Million Users* and *Site Linking In* play the most important influencing factor. That's to say, the weight of F_1 should be much higher than F_2 and F_3 . To overcome this problem, new similarity measurement is given as follows.

$$D(i, j) = \sqrt{\alpha(F_1(i) - F_1(j))^2 + \beta(F_2(i) - F_2(j))^2 + \gamma(F_3(i) - F_3(j))^2}$$

Where, $\alpha + \beta + \gamma = 1$, $D(i,j)$ is the distance of E-Commerce site *i* and *j*. The value of α , β and γ is determined by factor analysis.

According to five indexes' contribution to three common factors, the paper sets the value of α , β and γ . From table 5, we can know the rate that every factor contributes to the common component. We get the value of α , β and γ as the following .

1. Calculate the sum of the contribution, Sum_1 , Sum_2 , Sum_3 and Sum .

$$Sum_1 = 0.968 + 0.93 + 0.048 + 0.10 + 0.259 = 2.338$$

$$Sum_2 = 0.111 + 0.099 + 0.851 + 0.031 + 0.697 = 1.789$$

$$Sum_3 = 0.010 + 0.026 + 0.018 + 0.999 + 0.022 = 1.075$$

$$Sum = Sum_1 + Sum_2 + Sum_3 = 5.202$$

2. $\alpha = Sum_1 / Sum = 0.449$, $\beta = Sum_2 / Sum = 0.344$, $\gamma = Sum_3 / Sum = 0.207$

The contribution of the formula is that it takes the weight of each factor into account and improves the precision of similarity measurement. To some degree, the accuracy of the cluster result depends on the distance formula.

4. Results and Analysis

The 100 national demonstration E-Commerce sites are clustered by the improved DBSCAN, and the result is shown in Table 8.

Comparing Table 7 with Table 8, it can be seen that the improved DBSCAN is more effective than the traditional DBSCAN when processing the uneven density data, because the improved DBSCAN algorithm can set several different local Parameters based on the

density of data and meanwhile, it takes the similarity measurement into account. Through setting different local Parameters, some E-Commerce sites with high similarity labeled outliers in Table 7 can be divided into a class. Taking dangdang.com jd.com and vip.com for example, they are labeled as outliers in Table 7. However, they are divided into a cluster by setting two local parameter *Eps* and *MinPts* in Table 8. Therefore, the improved DBSCAN can deal with data with different densities, but the traditional DBSCAN fails to do it. Besides, the improved DBSCAN can improve largely the clustering accuracy and guarantee the objective evaluation on E-Commerce sites.

Table 8 shows that 100 E-Commerce sites are divided into six clusters and detailed data are shown in Table 9. From vertical and horizontal aspects, this paper calculates mean and standard deviation of their index data and evaluates these sites based on the results.

The index of cluster 1 have significant difference horizontally: 1688.com has obvious advantages in *Site Linking In* and *Access Per Million Users*; and the *Website Size* of gome.com.cn obviously outweighs *Access Per Million Users*, *Page Views Per User*, *Site Linking In*, *Speed*, that is to say, the difference between index is conspicuous; vertically, comparing 1688.com with gome.com.cn, the difference in *Page Views Per User* and *Speed* is smaller than others. Though there is a big development gap among index, the overall strength of such websites is strong compared with other sites. Compared with cluster 1, though development of the index of cluster 2 is unbalanced, the difference is relatively small. And different sites focus on different index of sites construction, which results in the development gap under the same index within different sites. However, the overall strength of such sites is strong. Compared with cluster 2, though the index development level of cluster 3 is different, the gap is relatively small. However, the overall strength is weak. Vertically, the characteristics of cluster 4 are reflected in the small development gap of *Site Linking In*, *Access Per Million Users* and the big gap of technique. Horizontally, the distinction between indexes is obvious, especially *Site Linking In* and *Access Per Million Users* outweighing other index. The overall strength of this kind of sites is greatly reduced due to the development imbalance among index. Cluster 5 focuses on the construction of technique holistically but ignore the construction of *Website Size*. The development level of cluster 6 on each index is low, so the overall strength is poor. Various sites have their comparative advantages, but their focus should be targeted in the development process. Among cluster 1 1688.com should focus on the construction of technology and website size. And gome.com.cn should pay attention to the construction of *Site Linking In*, *Access Per Million Users* and *Speed* improvement. Cluster 1 and cluster 2 should focus on the weak points, and focus on the balanced development of each index to further enhance the overall strength of such websites; Cluster 4 should focus on the construction of *Website Size* and *Speed* improvement; Cluster 5 should focus on the construction of *Website Size* and increase site propaganda to enhance the number of *Site Linking In* and *Access Per Million Users*; various index should be taken into consideration for the sake of expanding to cluster 6.

5. Conclusion and Future Work

Based on the sites' data characteristics of high dimension and uneven density in terms of 2013-2014 national demonstration enterprise E-Commerce sites, these sites are studied and evaluated in the view of combining factor analysis with improved DBSCAN. Since DBSCAN algorithm ignores weights when calculating the Euclidean distance, the result of the similarity measurement is not accurate. However, factor analysis is a good means to deal with weights. This paper improves the clustering accuracy and reasonableness of the evaluation by combining factor analysis with DBSCAN. However, the data processed by factor analysis have characteristics of uneven density. The traditional DBSCAN is improved to partition the data with different densities and cluster these sites. This paper

puts forward a new processing idea on E-Commerce sites evaluation: a new DBSCAN algorithm combining factor analysis with different densities. Compared with the traditional DBSCAN algorithm, the results of evaluating websites are more reasonable and interpretable with the improved DBSCAN algorithm. In the future work, the scale of the evaluation object will be further expanded and more research needs to be done.

Table 8. The Clustering Result of Improved DBSCAN Algorithm

| Cluster ID | Number of Sites | E-Commerce Sites |
|------------|-----------------|--|
| 1 | 2 | 1688.com gome.com.cn dangdang.com jd.com hc360.com dhgate.com 58.com bestbang.cn lvjinzaixian.com |
| 2 | 23 | gogo-le.com e365.com ctrip.com mysteel.com ly.com shoes.net.cn s.cn ju51.com vip.com a8.com lightinthebox.com sf-express.com ap88.com jc001.cn you369.com |
| 3 | 4 | 1-day.cn ebay.com shopin.net vancl.com |
| 4 | 4 | 6666688888.com kakw.cn hq53.com shncp.com |
| 5 | 7 | mi.com yhd.com suning.com chemnet.com moonbasa.com wanggou.com zhubajie.com |
| 6 | 60 | womai.com leyou.com.cn trade2cn.com sd.cn hinabuy.net.cn jiajiayue.com inzuo100.com ehaier.com weipin365.com liqunshop.com ... xj359.com |

Table 9. Detailed Data of Six Categories Sites (Partial Data)

| ID | Parameters | E-Commerce Sites | F ₁ | F ₂ | F ₃ | Mean | Std |
|-----------|------------|--------------------|----------------|----------------|----------------|-------|------|
| Cluster 1 | Mean | 1688.com | - | - | 9.777 | 3. | 5.33 |
| | | gome.com.cn | 0.131 | 0.019 | - | 928 | 9 |
| | | | 9.09 | - | - | 3. | 5.68 |
| | | | 4.48 | 0.080 | 0.230 | 209 | 8 |
| Cluster 2 | Mean | jd.com | 2.06 | 1.88 | 0.334 | 1. | 0.95 |
| | | ... | ... | ... | ... | ... | ... |
| | | jc001.cn | 0.015 | 1 | 0.196 | 490 | 5 |
| | | | 0.28 | 0.61 | 0.024 | 7 | |
| Cluster 3 | Mean | ebay.com | 0.099 | 6 | 0.330 | 228 | 1 |
| | | ... | 0.22 | 0.74 | - | 0. | 0.45 |
| | | vancl.com | 2 | 3 | 0.170 | 264 | 8 |
| | | | - | 0.32 | 0.089 | | |
| Cluster 4 | Mean | 6666688888.c om | 0.26 | - | - | - | 1.68 |
| | | ... | 0.57 | - | - | - | 2.81 |
| | | shncp.com | 5 | 4.562 | 0.015 | 1.334 | 1 |
| | | | 0.42 | - | - | | |
| Cluster 5 | Mean | mi.com | 0.08 | 1 | 0.175 | 0.156 | 5 |
| | | ... | 0.02 | 0.61 | - | - | 0.36 |
| | | zhubajie.com | 5 | 0 | 0.167 | 0.153 | 8 |
| | | | 0.13 | 0.76 | 0.024 | | |

| | | | | | | |
|-----------|-------------|-------|-------|-------|-------|------|
| | Mean | - | 0.25 | - | | |
| | | 0.003 | 0 | 0.166 | | |
| | Std | 0.05 | 0.29 | 0.010 | | |
| | | 2 | 0 | | | |
| | | - | 0.18 | - | - | 0.25 |
| | okorder.com | 0.308 | 6 | 0.164 | 0.095 | 4 |
| | ... | ... | ... | ... | ... | ... |
| | xj359.com | - | - | - | - | 0.11 |
| Cluster 6 | | 0.310 | 0.369 | 0.150 | 0.276 | 4 |
| | Mean | - | - | - | - | - |
| | | 0.280 | 0.035 | 0.152 | | |
| | Std | 0.04 | 0.37 | 0.018 | | |
| | | 8 | 0 | | | |

Reference

- [1] "The monitoring report of Chinese E-Commerce market data in [EB/OL]", <http://www.100ec.cn/data>, vol. 3, no. 19, (2014).
- [2] G. Jin-hua, G. Qi-Sheng, Z. Si-Ying, "Community Satisfaction Evaluation System's Research Based on AHP and Factor Analysis", *Journal of System & Management*, (2007), pp. 673-677.
- [3] X. Xue-Yan, "Research and application of fuzzy comprehensive evaluation model", Southwest Petroleum University, (2011).
- [4] P. Kang-Rong, L. Chun-Hui, "Network measurement: our new opportunities and challenges", *China Metrology*, (2001), vol. 7, pp. 27-28.
- [5] G. Jin-yu, Z. Zhong-bin, S. Qing-yun, "The Research and Application of Analytic Hierarchy Process", *China Safety Science Journal*, (2008), vol. 5, pp. 148-153.
- [6] Z. Yang-Yong, X. Yun, "DNA sequence data mining technique", *Journal of Software*, (2007), vol. 18, no. 11, pp. 2766-2781
- [7] L. Ertoz, M. Steinbach, V. Kumar. "Fiding clusters of different sizes, shapes, and densities in noise", high dimensional data, Technical Report, (2002).
- [8] M. Ester, H. P. Kriegel, J. Sander, "A density-based algorithm for discovering cluster in large spatial databases with noise", *Proceeding the 2nd International Conference on Knowledge Discovery and Data Mining KDD*, Portland, (1996), pp. 226-231.
- [9] S. Kisilevich, F. Mansmann, D. Keim, "P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos", *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application. ACM*, (2010).
- [10] O. Yang- Jia, L. Pi-Yuan, "Webpage content extraction based on DBSCAN", *Computer Engineering*, (2011).
- [11] Z. Cuo-Cai, Z. Rong-Hua, F. Kai-Yao, "DBSCAN algorithm based on telecom customer classification application research", *Journal of Beijing Union University (Natural Sciences)*, vol. 3, (2012), pp. 52-55
- [12] L. M. Kieu, A. Bhaskar, E. Chung, "Transit passenger segmentation using travel regularity mined from Smart Card transactions data", (2014).
- [13] H. Verkasalo, C. López-Nicolás, F. J. Molina-Castillo, "Analysis of users and non-users of smartphone applications", *Telematics and Informatics*, vol. 27, no. 3, (2010), pp. 242-255.
- [14] C. Wen-Cheng, "Region economy unbalanced development research based on factor analysis", *Journal of Applied Statistics and Management*, vol. 3, (2010), pp. 490-501.
- [15] T. Hasan, J. H. L. Hansen, "Acoustic factor analysis for robust speaker verification", *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, (2013), pp. 842-853.
- [16] Z.-P.Hou, "Analysis on Application and Research of Factor and Clustering in E-Commerce Website Evaluation", *Science and Technology Management Research*, (2011), pp. 144-147.
- [17] J.-J. Li, H.-M. Chen, "E-Commerce Web Site Evaluation Based on Factor Analysis and Correspondence Analysis", *Information science*, vol. 8, (2008), pp. 1252-1256.
- [18] Alexa website, aviable at: <http://www.alexa.com>
- [19] H. Xu, J. Ma, Q.Wu, "Application Research of DBSCAN Algorithm Based on High-Performance Computing Center Users Classification", *Journal of Jilin University (Information Science Edition)*, vol. 5, (2013), pp. 528-534
- [20] Z. Wen, G.-S. Xia, Z.-J. Gou, Z.-X. Yan, "An improved DBSCAN algorithm", *Journal of Sichuan Normal University (Natural Science)*, vol. 2, (2013), pp. 312-316.
- [21] Q.-S. Rong, J. B. Yan, G.-J. Guo, "Research and Implementation of Clustering Algorithm Based on DBSCAN", *Computer application*, vol. 4, (2004), pp. 45-46.
- [22] M. Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, vol. 11, no. 1, (2009).

Authors



Yongyi Cheng, he was born in 1982, Ph.D. Candidate. His major research interests include cloud computing, big data, data mining, etc. His email is echojlu@163.com.



Yumian Yang, she was born in 1990, Master Candidate. Her major research interests include cloud computing and business intelligence. Her email is 1083827461@qq.com.



Jianhua Jiang, he was born in 1979, Ph.D., Associate Professor. His major research interests include business intelligence, data mining, etc. His email is jianhuajiang@foxmail.com.



Gaochao Xu, he was born in 1966. Professor and PhD supervisor of Jilin University. His main research interests include cloud computing, information security, software testing and software reliability.

