International Journal of Intelligent Information and Management Science
ISSN: 2307-0692, Volume 7, Issue 2, April, 2018

Simulation and Analysis of Dynamic Trend of Tourism Big Data based on Feedback Constraint Association Rule

Heqing Zhang[1], Jiahao Xiang[1], Leilei Wang[2], Xiaobo Su[3]*
[1]Tourism College of Guangzhou University, Guangzhou, 510006, China
[2]Guangzhou Panyu Polytechnic, Guangzhou, 511483, China
[3]Tourism College of Guangzhou University, Guangzhou,510006, China

Abstract: In present, with the rapid development of economy and science and technology, the tourism industry is leading the trend of development. Meanwhile, the era of big data is also one of the hot topics in media. How to apply the big data of tourism in the tourism industry is always one of the research focuses in the field of tourism. Aiming at the bottleneck problem of traditional data mining algorithm in big data processing, a dynamic trend mining method of tourism big data based on feedback association rule is proposed. With the proposed method, tourism data is collected, sorted, cleaned, filtered and statistically analyzed. The image specification and format of the data set are unified, and the feature extraction and fusion of the tourism image are carried out. The latent association information between tourism image features is mined by using feedback constrained association rule. The re-ranking of tourist images is carried out with the method of re-ranking based on graph model. Finally, the dynamic trend of tourism big data is analyzed according to the results of re-ranking. Through the use of massive real tourism data, the proposed method can achieve accurate and efficient analysis of the dynamic trend of tourism big data, which has certain economic value and practical significance for the development of the tourism industry.

Keywords: Tourism; Data; Association rule; Dynamic trend; Re-ranking

1. Introduction

The development of tourism has gone through the artificial stage, and gradually developed into the current electronic and digital era. The service of the tourism industry also transfers from the traditional offline to online, providing more diversified tourism services, which are closely related to the internet technology (Zhou et al., 2016; Schäfer, 2017). In recent years, the wisdom tourism is shown in the status and role of tourism in China. Tourism is also an activity with mobile, real-time, and variant. For the serious problems of overload information and the development of the internet, how to mine from big data out of order in wealth is the largest area of data mining the challenge (Liu et al., 2015).

At present, as tourism industry of China is in a critical period of construction of wisdom tourism, the construction of tourism information is essential. However, the accurate and satisfied service cannot be provided to users, which leads to poor experience for users. With the arrival of big data, how to make full use of the original tourism information and the characteristics of big data is the new subject in the application of big data, which is through the analysis of data mining of user history to rapid and accurate providing efficient and satisfied tourist information service for users (Chen & Schintler, 2015).

The traditional dynamic trend analysis method of tourism big data based on the concept hierarchy theory cannot provide better experience for users (Liu et al., 2016). To address this problem, a dynamic trend mining method of tourism big data based on feedback association rule is proposed for the bottleneck problem of traditional data mining algorithm in big data processing.

2. Research on Dynamic Trend of Tourism Big Data based on Constraint Feedback Association Rule

2.1. Tourism data collection and preprocessing

2.1.1.Tourism data collection

To analyze user behavior, it is necessary to describe the user behavior first. User behavior is stored in the tourism log database in the form of a record. Users can browse the three kinds of contents, which are tourism video, tourism text, and tourism picture. The record of the browse of three kinds of information is in form of [type, length of time]. The record of tourism video is the number of video and the length of browse time. The record of tourism text is the number of text and the length of stay time. The record of tourism picture is the number of the

browsed picture set and the length of browse time. The format of the record is as follows.

Record:{
recordID: [R0000001-..],
userID:[U000000[..]],
type:[V0001 V0538,T0001 T0538,P0001 P0328]
long(ms):[],
time:[Date]
}

Figure 1. Format of the tourism record

In Figure 1, V denotes video, T denotes text, and P denotes picture. The analysis of user behavior generates each record in the tourism log database, and the next step is to clean and filter the tourism data.

2.1.2. Subject filtering of tourism plotting data

As the amount of tourism plotting data is large and quality of data is varied, the important mean is to extract the keywords contained in the text of tourism plotting data for further filtering. Chinese word segmentation is the main method to extract keywords. By segmenting the text of the tourist plotting data and calculating the weight, the keywords can be extracted effectively. The extracted keywords can be used to filter the tourism data effectively for further data preprocessing. By using NLPIR word segmentation system, the text information in the drawing data can be effectively analyzed. The key words can be extracted to obtain the relatively clean tourism data.

2.2. Feature fusion of tourism image

The image specification and format of the tourism dataset are unified, and the feature extraction of tourism image is carried out. In the description of tourism image, the BovW feature representing visual feature and the color moment feature are distributed on local details and colors, respectively. Both of them have visual representation. In this paper, the BovW feature (visual word bag model) and color moment feature are used to achieve feature fusion of tourism image (Zhao, Sun & Yuan, 2016).

For the BovW feature of tourism image, feature vector is extracted. Firstly, the Scale-invariant feature transform (SIFT) feature of each tourist image is extracted, where the dimension of SIFT feature is 128. K-Means method is used to cluster SIFT feature of tourism image. The clustering center is regarded as the visual word of tourism image, the distance between the SIFT feature points of each tourism image to the cluster center is calculated, and the SIFT feature is mapped into the visual words of the cluster center. The tourism image is represented as the feature vector based on the visual words. For the color moment features of tourism image, the color distribution can be represented by low order moments (the first order, the second order, the third order) on three color channels for the he tourism image. So the tourism image is represented on low order moments, where the dimension of the color moment is 255.

Singular value decomposition (SVD) is used to solve latent semantic space of visual features of tourism image (Djenouri et al., 2015). Singular value decomposition is obtained by

$$B_{LSA} = TSD^T \tag{1}$$

where the matrices T and D are unit orthogonal, that is, $T^T T = D^T D = I$. The diagonal matrix S_{LSA} contains the singular values of the matrix B_{LSA}. When selecting k maximum singular values, the approximate matrix $\hat{S}$ of S multiplies the matrix T and D to obtain the approximate matrix $\hat{B}_{LSA}$. Then

$$B_{LSA} \approx \hat{B}_{LSA} = T_0 \hat{S}_0 D_0^T = \tilde{\sum}_{LSA} D_0^T \tag{2}$$

where the matrix $T_0 \hat{S}_0$ equals to the matrix $\hat{\Sigma}_{LSA} \in R^{n \times k}$. The matrix $\sum_{LSA}$ is regarded as the low dimensional representation of the matrix B_{LSA}. By using $\sum_{LSA} D_0^T$ as the approximation of B_{LSA}, the optimization problem is given by

$$\min_{\sum_{LSA}, D_0} \left\| B_{LSA} - \sum_{LSA} D_0^T \right\|_F^2 + g\,LSA \|D_0\|_F^2 \tag{3}$$

where $\|\cdot\|_F$ is F-norm, $g\,LSA$ is a very small positive value, and $D_0 \in R^{m \times k}$.

The sharing matrix of two visual feature matrices of tourism image is obtained by the standard conjugate gradient optimization method. The gradient of the objective function is represented by

$$\begin{cases} \dfrac{\partial Q}{\partial D_B} = \left(D_B \Sigma^T \Sigma - B_{BovW} \Sigma \right) + g\, D_B \\ \dfrac{\partial Q}{\partial D_C} = l \left(\left(D_C \Sigma^T \Sigma - C_{CM} \Sigma \right) \right) + g\, D_C \\ \dfrac{\partial Q}{\partial \Sigma} = \Sigma D_B^T D_B - B_{BovW} D_B + l \left(\Sigma D_C^T D_C \right) - C_{CM} D_C \end{cases} \tag{4}$$

where the sharing matrix Σ is the obtained hybrid feature matrix based on feature fusion. It can not only retain the important information of the matrix B_{BovW} and the matrix C_{CM}, but also mine the latent association information between them.

2.3. Dynamic trend mining of tourism big data based on constraint feedback association rule

Input: Tourism transaction dataset, support threshold *Support*, and confidence threshold *Confidence*.

Output: Frequent itemset, association rule

Steps: 8 steps for dynamic trend mining of tourism big data based on constraint feedback association rule

1)During the first iteration, all the travel events are scanned, the candidate frequent 1 itemset are counted, and the support counter $C1$ of the frequent 1 itemset is calculated.

Support: The rule $A \Rightarrow B$ is established in the tourism transaction dataset D. It means that A and B often appear together, and the ratio of the number of occurrences to all transactions refers to support $Support$, denoted by $P(A \cup B)$, which is given by

$$Support(A \cup B) = P(A \cup B) = \frac{D(X)}{|D|} \quad (5)$$

2)Comparer the counter of the candidate 1 itemset with the set support threshold, the candidate itemset not conforming to the requirement is excluded and the candidate 1 itemset $L1$ conforming to the support threshold is written into the file.

3) The Cartesian product of $L1$ obtained in step 2) is used to obtain the candidate 2 itemset. The 2 itemset is obtained as $L2$.

4) Repeat step 3) to obtain the candidate itemsets in each phase. According to the property of the Boolean association rule frequent itemset algorithm (Apriori), all subsets of frequent itemset must also be frequent. Using the layer-by-layer search technique, given the candidate K itemset, it is only necessary to check whether the $K-1$ subset is frequent.

5) The algorithm terminates until the candidate K itemset CK is empty set. Then all frequent itemsets are found.

6) The tourism data association rule is generated by using the rule generation function on the obtained frequent K itemset.

7) The confidence of the generated association rule is calculated and compared with the confidence threshold. The tourism data association rule conforming to the condition is called strong association rule, which is the mined dynamic trend rule of tourism big data (Dng & Peng, 2016).

8) Apriori algorithm is end of run and output the results.

2.4. Re-ranking based on graph model

Assume given query term q, the image set returned by the initial ranking result is denoted as $M = \{m_1, m_2, \mathbf{L}, m_i, \mathbf{L}\}$. The rank of initial ranking of tourism image m_i is r_i', after re-ranking based on visual feature, its rank is r_i. Since the single visual effect of tourism image can only describe tourism image unilaterally, it is necessary to use latent space analysis to learn a hybrid sharing feature $\sum$ with two features after extracting the visual features of tourism image. $G = (V', E)$ denotes connection graph, where the node V' is the tourism image, the edge E is the similarity of the tourism image m_i and m_j. Assume $W = w_{ij}$ is weight connection matrix with $e \times e$ dimensions, where w_{ij} is the weight of the similarity of the tourism image m_i and m_j. Finally, tourism images are re-ranked based on graph model (Mcdonald, 2017).

As the visual consistency hypothesis, that is, the similar tourism images are ranked together as far as possible, the Laplician regularization is used to calculate the similarity of tourism images and the difference of the rank, which is defined as energy function $y(r, m)$.

Feature matrix $\sum$ of the tourism images obtained by fusion of two visual features is used to construct a Laplician regularization graph based on hybrid feature. The weights of the edges are calculated by Gaussian kernel to obtain weight connection matrix W.

$$w_{ij} = \exp^{-\|\Sigma_i - \Sigma_j\|^2 / 2s^2} \quad (6)$$

where s is the width parameter of Gaussian kernel function, $\|\Sigma_i - \Sigma_j\|$ is hybrid visual feature of tourism image.

The energy function is given by

$$y(r, m) = \frac{1}{2} \sum_j w_{ij} (r_i - r_j)^2 \quad (7)$$

where Laplician matrix is $L = D - W$, D denotes diagonal matrix and $d_{ii} = \sum_j w_{ij}$, the constructed graph is denoted as G and equals to $D^{-\frac{1}{2}} W D^{\frac{1}{2}}$.

The Laplician regularization is normalized and weight matrix W is normalized by $\sum^{-\frac{1}{2}} W \sum^{\frac{1}{2}}$. Then the energy function is finally given by

$$y(r, m) = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(\frac{r_i}{\sqrt{d_{ii}}} - \frac{r_j}{\sqrt{d_{ii}}}\right)^2 \quad (8)$$

In re-ranking the rank, only part of nodes of the constructed graph is selected, and a subgraph G' is extracted to improve the efficiency of re-ranking. Then re-ranking is rewritten by

$$\hat{r} = (I - eG')^{-1} \hat{r}' \quad (9)$$

where $\hat{r}$ denotes $t \times 1$ dimensional vector, t denotes the first t tourism images in the initial ranking.

According to the association rules of tourism data and the re-ranking results of tourism images, the dynamic trend of tourism big data is analyzed.

3. Simulation and Analysis of Dynamic Trend of Tourism Big Data based on Constraint Feedback Association Rule

3.1. Collection of tourism data

3.1.1. Collection of tourism data from tourism community website

In order to better reflect the tourism data and mining valuable information from massive data, the well-known tourism community Mafengwo is chosen in this paper. In this website, the registered users are 25 million. There are 100 thousand photos for each scenic spot. More than 2000 blogs are newly increased on daily average. Equivalently, on average, 8000 people fly around the world every day, taking photos and having delicious food. 87% of them will contribute to travel information and 30% of people will sort out travelling blogs every day. The information will be replied and discussed by the 500 thousand active users. The website updates fast and has a large number of information, which can be used as the target website for experimental tourism data collection.

In this paper, data collection is obtained with the tool of LocoySpide. This tool is programed with Visual C# and can run under the operating system of Windows2008, Windows2000, Windows XP, and Windows 7.

3.1.2. Collation of the collected tourism data

Taking inbound tourism of the city of Xi'an as an example, the related tourism data is collected in Mafengwo website. These data can be divided into two categories: tourism note and tourism photo. Tourism notes can quickly determine the time and location of tourism from the text description. Photos focus on confirming the rationality of text expression. They complement each other. Meanwhile, the information of tourist is also collated and counted.

For a large number of tourist data collected, preliminary filtering should be carried out. The standard of filtering is as follows.

In tourism notes, a certain number of photos should be equipped, while a complete travel itinerary should be clearly expressed for the tourism notes with no photos.

The complete record of tourist activities in chronological order is required in the tourism note.

Tourism notes should be equipped with text or photos for tourist activities.

After collation and analysis, the number of tourists conforming to the research standard is 1308, the number of tourism notes is 1416 (a visitor may continue to issue 2 to 3 notes), the number of tourism photos is 74671. Statistical data is shown in Table 1. The corresponding tourist information has also been collated, and input into the database, part of the collation is shown in Table 2.

Table 1. Collection of collected tourism data

	Tourism note and photo	Tourism note	Photo	Total
Persons (per)	1115	180	36	1331
Ratio (%)	83.77%	13.52%	2.70%	100%

Table 2. Part of tourism information after collation of the collected tourism data

User	Time	Person	Days	Form	Per capita spending	PageUr1
Purple onion(Shanghai)	2013-11-18	Single	5	Walk	2800	http//www.mafengwo.cn/i/288
Take me on a trip(Guangdong)	2013-10-18	Friends	4	Tourist group	3500	http//www.mafengwo.cn/i/132
Helen(Beijing)	2014-07-29	Couple	6	Independent travel	5000	http//www.mafengwo.cn/i/295
Deserted city (Shenyang)	2014-9-02	Parent-child	4	Self-driving	6000	http//www.mafengwo.cn/i/298
Silent Separation(Shandong)	2014-5-03	Couple	3	Leisure	4000	http//www.mafengwo.cn/i/133

3.2. Analysis of tourism data

The spatial and temporal distribution pattern of filtered information is statistically analyzed. Based on the number of tourists, the time distribution characteristics of inbound tourist flow in a city are researched. The statistics of the changes of tourist amount with month, tourist amount with stay time, and tourist amount with the number of accompanying persons is obtained.

For the change of tourist amount with month, the number of persons conforming to the condition is 1331. The histogram of tourist number with month is shown as Figure 2.

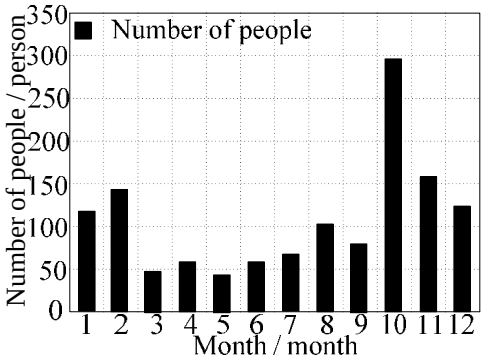

Figure 2. Histogram of tourist number with month

For the distribution of tourist amount with stay time, the number of persons conforming to the condition is 1288. The distribution of tourist amount with stay days is shown as Figure 3.

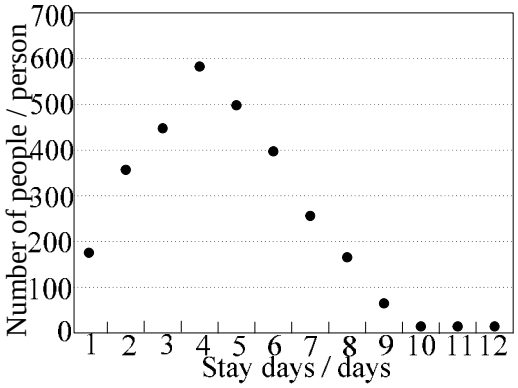

Figure 3. Distribution of tourist capacity with stay days

For the distribution of tourist amount with the number of accompanying persons, the number of persons conforming to the condition is 1046. The distribution of tourist amount with the number of accompanying persons is shown as Figure 4.

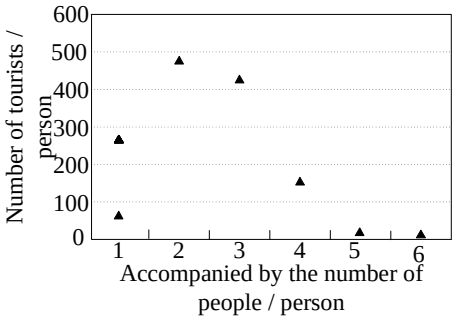

Figure 4. Distribution of tourist amount with the number of accompanying persons

From statistics, the number of tourists in the National Day holiday to the region is the largest, accounting for 22.6% of the total annual population. The number of accompanying persons is mainly from 2 to 3, and most people will choose 3 days for mid-term travelling. From the results, it can be obtained that, the tourists in the collected data have the characteristics of individual tourists, and the seasonal intensity index can be used to analyze the uneven time distribution of tourism demand in the city. Therefore, the dynamic trend analysis of tourism big data is mainly aimed at the domestic tourists who come to this region.

3.3. Experimental results

The collated tourism data is uploaded to HDFS file system. By executing shell instruction, #hadoop fs-putregulation.dat regulation.dat will package and run the program. The program execution is shown in Fig. 5. The results are shown in Figure 6.

Figure 5. Program execution obtained with the presented method

Figure 6. Program execution results obtained with the presented method

In the experiments, the minimum number of support is set to 5(As the frequency of occurrence of some item is too low, in order to make the dynamic trend analysis results of tourism big data as comprehensive as possible, so the minimum number of support is set to 5). Then a large number of constraint feedback association rules are generated. In order to prevent the high frequency of item generating too many rules, the confidence is set in the program and the rule with confidence less than 0.15 is removed, so the reduced association rules are obtained. For the case of j13(Great Wild Goose Pagoda), the filtered rules are shown in Figure 7.

[j2] => j13 : confidence=0.397;
[j2,x19] => j13 : confidence=0.268;
[j11] => j13 : confidence=0.421;
[j11,x1] => j13 : confidence=0.227;
[j11,x1,z51] => j13 : confidence=0.219;
.....

Figure 7. Filtered association rules with the presented method

From the results, it can be seen that, the rules include the combination of several tourism services. There are more repetitive items in the rules needed to be further merged. The merged association rules are transferred into the tourism relational database through the tool of Sqoop. Dynamic trend analysis of related tourism big data achieve by using web service.

Comparison of performance between the presented method and the traditional method

The performance of the presented method and the traditional method is tested under the condition of same amount of tourist affairs. The number of clusters for performance test is 4, 6, and 9. Different number of tourist affairs is executed and the running time is recoded. Comparison is shown in Table 4.

Table 3. Cluster for performance test

Cluster	Namenode	Datanodes
number		
4	10.2.192.47	10.2.192.48.....10.2.192.52
6	10.2.192.47	10.2.192.48.....10.2.192.54
9	10.2.192.47	10.2.192.48.....10.2.192.55

Table 4. Comparison of performance between the presented method and the traditional method

Running time /s Method	Number of affairs with the presented method /10^4			Number of affairs with the traditional method /10^4		
	3	6	9	3	6	9
4	322	468	512	582	789	930
6	487	496	579	763	882	997
9	501	568	637	1529	1633	2352

From the results, it can be seen that, for Hadoop cluster, the performance of the program obtained with the presented method is better than the traditional method. With the increase of the number of clusters, the execution time will be shorter and the effect is obvious. However, two points should be noticed. First, when the number of tourist affairs is small, the presented method does not have much advantage. When the number of tourist affairs is small, the execution time with the presented method is not smaller than the traditional method. Because it takes time to assign tasks from Tasktracer to Jobtracker. Second, the running time of the same number of affairs for 6 clusters and 9 clusters is the same. The reason is the number of tourist affairs is not large enough. In 9 clusters, 2 datanodes are not assigned to job, so although 9 clusters are running algorithms, actually only 7 datanodes work. Limited to the experimental conditions, only the above experiments have been carried out for data and cluster.

Generally, the efficiency of algorithm tends to be proportional to the number of cluster nodes. Thee advantage of the presented method is the greater the amount of tourism data in computation, the better the performance. In addition, the memory space is limited for the traditional method. As recursion operation requires a lot of memory with large-scale I/O operation, it will inevitably cause waiting or blocking. However, in the presented method, the data is assigned to multiple Datanodes by Namenode, and the Job is assigned to each node, and then the working pressure of the program is dispersed. Additionally, Map/Reduce is two separate running processes. For the I/O operation, the cost of the presented method is relatively small, the running time is reduced, and the accuracy of dynamic trend analysis of tourism big data is ensured.

4. Conclusions

With the rapid development of tourism and transportation and the explosive growth of tourism information, tourism data has formed a huge mass of information space. How to analyze the massive historical data of the tourism information fast, accurately, and conveniently, which is to find out the overall behavior with a specific area, a specific group of people or a specific need from the data, is of great significance to analyze the operation status of tourism market, predict the impact of tourism on related industries, and adjust the macro policy of tourism.

For the above problems, constraint feedback association rule is researched according to the feature of tourism data. A method of dynamic trend analysis of tourism big data based on feedback constraint association rule is proposed in this paper. This method can solve the problem that the existing data mining analysis methods are difficult to be applied because of the frequent itemsets of massive tourism data. According to the feature of tourism data, the tourism data is mined to obtain more valuable information for a specific area, a specific group of people or a specific need. In this way, the dynamic trend of tourism big data can be more accurately analyzed.

5. Acknowledgment

National Natural Science Foundation of China : Reserch on Constructing Ecological Gene Identification of Ethnic Traditional Settlement and "Ecological Gene Information Atlas"(71473051)

References

[1] Chen, Z., & Schintler, L. A. (2015). Sensitivity of location-sharing services data: evidence from American travel pattern. Transportation, 42(4), 669-682. DOI: 10.1007/s11116-015-9596-z.

[2] Dng, B. Q., & Peng, J. J. (2016). Abnormal Data Mining Algorithm in Complex Network Data Flow Simulation.

Computer Simulation, 33(1), 434-437. DOI: 10.3969/j.issn.1006-9348.2016.01.095.

[3] Djenouri, Y., Bendjoudi, A., Mehdi, M., et al. (2015). GPU-based bees swarm optimization for association rules mining. Journal of Supercomputing, 71(4), 1318-1344. DOI: 10.1007/s11227-014-1366-8.

[4] Liu, M. M., Zhao, S. L., Chen, M., et al. (2015). Scaling-up mining algorithm of multi-scale association rules mining. Application Research of Computers, 32(10), 2924-2929. DOI: 10.3969/j.issn.1001-3695.2015.10.010.

[5] Liu, M. M., Zhao, S. L., Han, Y. H., et al. (2016). Research on Multi-Scale Data Mining Method. Journal of Software, 27(12), 3030-3050. DOI: 10.13328/j.cnki.jos.004924.

[6] Mcdonald, N. C. (2017). Trends in Automobile Travel, Motor Vehicle Fatalities, and Physical Activity: 2003-2015. American Journal of Preventive Medicine, 52(5), 598-605. DOI: 10.1016/j.amepre.2016.12.012.

[7] Schäfer, A. W. (2017). Long-term trends in domestic US passenger travel: the past 110 years and the next 90. Transportation, 44(2), 293-310. DOI:10.1007/s11116-015-9638-6.

[8] Zhou, S., Li, Z., Cotter, C., et al. (2016). Trends of imported malaria in China 2010–2014: analysis of surveillance data. Malaria Journal, 15(1), 1-8. DOI: 10.1186/s12936-016-1093-0.

[9] Zhao, X. J., Sun, Z. X., & Yuan, Y. (2016). An Efficient Association Rule Mining Algorithm Based on Prejudging and Screening. Journal of Electronics & Information Technology, 38(7), 1654-1659. DOI: 10.11999/JEIT151107.