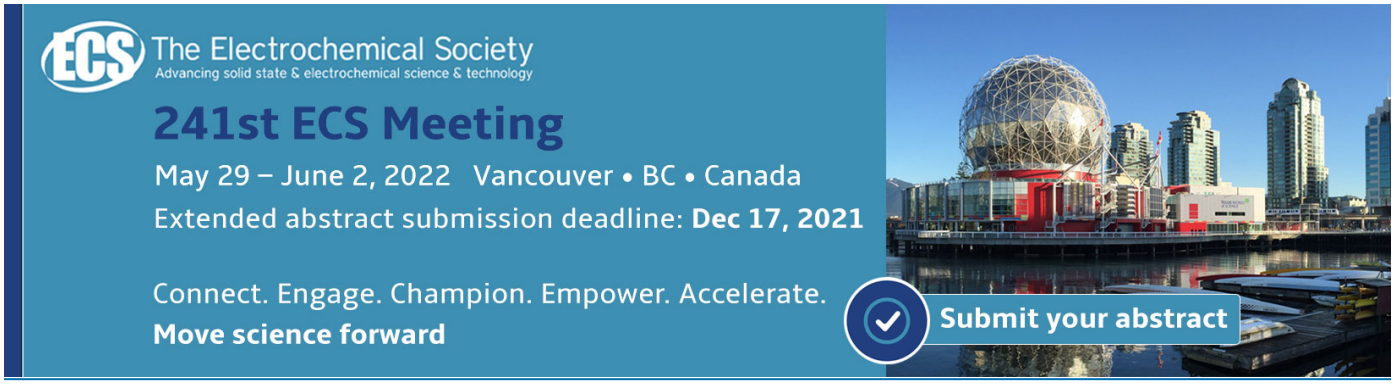
Research on Clustering Algorithm Based on Web Log Mining



Research on Clustering Algorithm Based on Web Log Mining

Haifei Xiang[1, *]

[1]Department of public education, Wenzhou Polytechnic, Wenzhou, 325000, China

[*]Corresponding Author's e-mail: xhf_1980@163.com

Abstract: People's online business behaviour has become more and more frequent. Internet service providers are beginning to find ways to obtain the interests and hobbies of users in order to provide targeted services to users. Analysis of user behavior based on Web logs can obtain valuable information from users. User clustering based on Web logs can cluster users according to user behavior, and then analyze user access patterns, providing a good solution for problem solving. This article introduces the concept and process of data mining, the classification and process of Web data mining, and then analyzes the K-Means clustering algorithm. The class-centered algorithm avoids clustering and only obtains the local optimal solution, and it can reduce the algorithm iteration time and improve the clustering quality.

1. Introduction

A large number of e-commerce websites have emerged based on e-commerce. Too many commodities have also troubled the user's choice. The user is only interested in a small part of the commodities, and in order to select these commodities, it often takes much time. Each user may need different information. Therefore, classify users and push targeted services according to their preferences. Through data mining, users' interests and hobbies are classified according to server logs, and different information is pushed for different users [1].

Web logs are the source of user data. Cluster analysis is an unsupervised learning method and one of multivariate statistical analysis methods. It performs clustering based on a certain similar feature of objects. The user 's behavior is reflected in the web log as a click record on the page. We can encapsulate the user 's click action into a sequence, and use the clustering algorithm to mine the user access sequence to accurately and quickly implement user clustering, and then proceed Personalized service [2]. The rapid development of the Internet has had a huge impact on our lives. Various commercial websites are providing services for us. They need to understand the needs and interests of the user group to convert them into commercial value, so they gather based on Web logs. Class analysis is becoming more and more important

2. Web log mining

2.1. Data mining

It generally refers to the process of automatically searching for effective, unknown, and hidden useful information hidden in noisy and fuzzy data sources (text, databases, Web, etc.) [3]. The relevant overview is shown in Figure 1. It achieves the above goals through many methods such as artificial intelligence and visualization.

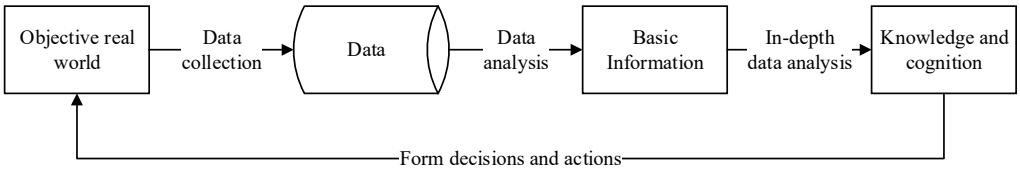


Figure 1. Overview of data mining

The general process of data mining is described as follows:

- Data cleaning: Complete data integrity and consistency checks according to mining tasks, eliminate noisy data unrelated to data mining, and improve data quality.
- Data integration: There are many types of data, and the sources and formats of the data are different. Data integration can complete the work of unified data format.
- Data selection: Choose different data according to different mining tasks.
- Data transformation: Perform certain data operations to transform the data organization form, so that the data transformation is suitable for mining expression.
- Data mining: Select appropriate algorithms and techniques to model the transformed data according to business needs, and discover hidden patterns and knowledge.
- Model evaluation: Evaluate the patterns and knowledge gained from mining based on reasonable interest metrics that meet business needs.
- Knowledge representation: Use techniques such as data visualization to visualize knowledge in an easy-to-understand form and present it to users [4].

According to the above description, the general process structure of data mining can be represented as Figure 2.

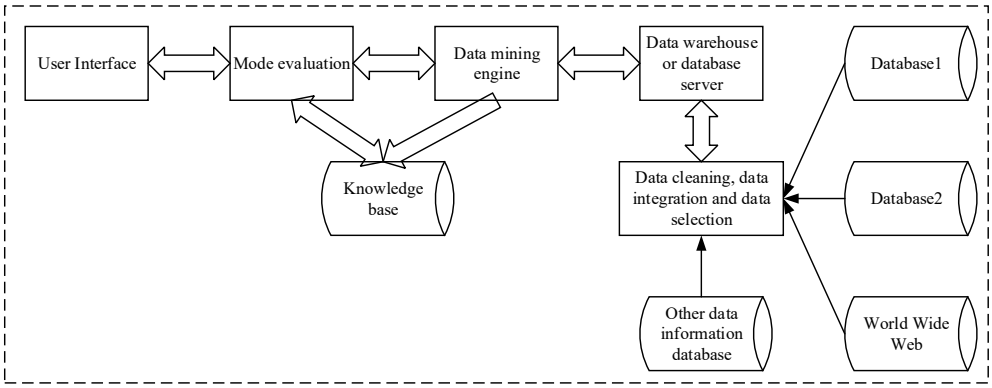


Figure 2. Data mining process

2.2. Web log data mining

There are many classification methods for Web mining, such as classification by mining site attributes and classification by Web text language. At present, most researchers divide mining into three categories according to the different data objects. The data objects that are mined are content data, structural data, and user access data on Web. Its structure is shown in Figure 3.

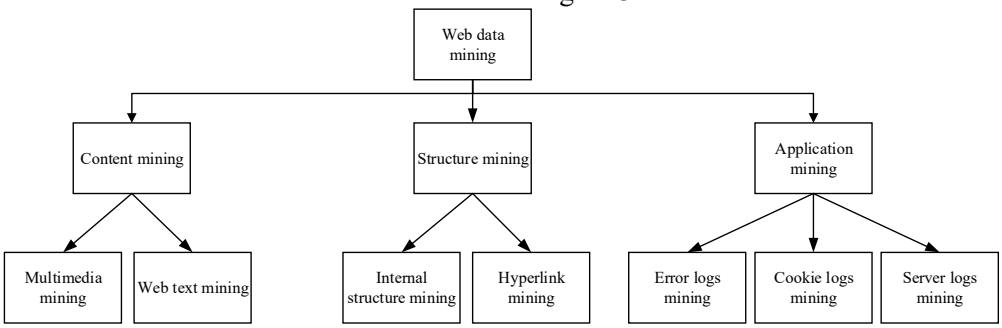


Figure 3. Web mining classification

Web log mining is the use of Web mining in Web logs, mainly through preprocessing and analysis of Web log data, an analysis mode to obtain information of interest from it [5]. The server log records the user's various website browsing information. By analyzing these data, we can know what the user is interested in, and the website can push related services for the user in a targeted manner. Figure 4 shows the typical process.

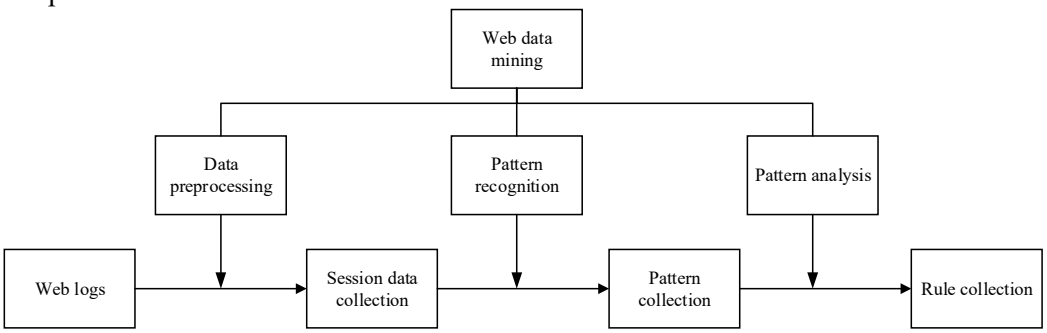


Figure 4. Web log mining process

Web data preprocessing. A lot of raw data information is contained in the web log records. The data information is stored in the form of text. The log data information includes the basic information accessed by the user. A lot of information such as customer name and address, date and time, resource type, number of bytes sent and received, time spent, etc. are recorded in each piece of access information [6]. Due to the large differences between the servers and parameter settings, different log records are obtained. These log data cannot be used directly, and they need to be preprocessed.

Web data collection. Users need to connect with the provider's server first, and then access the provider's server to connect the client machine to the Internet. If the client needs to access the website, it needs to connect with the website server. Under normal circumstances, large-scale websites use a three-level work mode, specifically set up a server to store data. Each link in the above process will generate a corresponding log file. The data obtained from the agent, client and server are all valid data [7]. User identification. The identifier used to distinguish the user in the log is the unique identifier written by the server site to the user's browser. When the user requests the page again, the identifier is appended to the request and returned to the server so that the user's identity can be identified. If multiple users use a client at the same time, or the user deletes the logo, the server will treat the user as the first login at the next login. In addition, the user may refuse to be written because of privacy issues.

3. Research on cluster analysis algorithm

3.1. Clustering Algorithm

The basic idea of the K-Means algorithm is to first randomly select K data objects as the initial clustering center, then calculate the distance of all remaining objects from each center point, and assign it to the cluster of the nearest center point locates, and finally Iterate the division until the resulting cluster no longer changes [8].

In the data set $D\{x_1, x_2, \dots, x_n\}$, the sparsity of the object F is defined as:

$$S(x_i) = \frac{1}{K}\sum_{x_i \in N_i} d(x_i, x_j) \tag{1}$$

Among them, Ni represents an object set composed of K neighbor objects of x_i; $d(x_i, x_j)$ represents the distance between x_i and x_j.

Given a data set $D\{x_1, x_2, \dots, x_n\}$, each data object is m-dimensional, and the data set D is divided into K clusters $\{S_1, S_2, \dots, S_k\}$, using K-means algorithm The definitions needed for the process are as follows:

Definition 1: Euclidean distance.

$$d(x_i, x_j) = \sqrt{(x_i, x_j)(x_i, x_j)} \tag{2}$$

Among them, x_i and x_j are two data objects.

Definition 2: Clustering center.

$$Z_c = \frac{1}{n_c}\sum_{i=1}^{n_c} x_i \tag{3}$$

Among them, n_c is number of data objects contained in cluster C; x_i represents the i-th data object of cluster C.

Definition 3: Termination conditions for clustering.

$$E = \sum_{i=1}^{K}\sum_{j}^{n_j} d(x_j, Z_j) \tag{4}$$

Among them, K is the number of clusters; Z_j is center of the j-th cluster; n_j represents the number of data objects contained in the jth cluster.

Definition 4: M inputs value.

$$Minpts = beta \times \frac{N}{K_{max}} \tag{5}$$

$$K_{max} = \sqrt{N} \tag{6}$$

The above formulas are all empirical rules. The M inputs points adjacent to x_i form N_i, K max represents the maximum number of clusters, and beta is a user-defined parameter.

The K-means algorithm is shown in Figure 5.

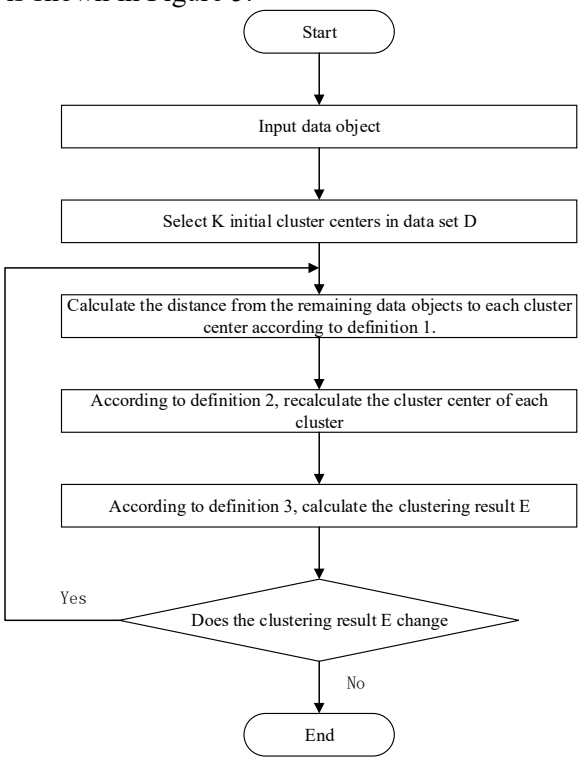


Figure 5. The basic flow of K-means algorithm

3.2. Web log mining based on clustering

The technology based on clustering is to use cluster analysis algorithm to cluster Web log data, which is mainly divided into web page clustering, session clustering and user clustering. Get together to quickly mine user access patterns and user characteristics with specific access rules. The data objects to be mined include the number of users visiting website pages, the number of users, and user visits to each web page [9].

3.3. Improved K-means algorithm

For the selection of the initial clustering center, not only should we consider the tightness around the initial clustering center, but also to ensure the initial clustering center is as discrete as possible. Therefore, while using sparsity to judge the sparsity around the data objects, in order to ensure that the clustering center points are as discrete as possible, a new measurement method combining sparsity and distance is adopted. Firstly, judge the tightness around the data by sparseness, and then construct a new evaluation function combining the principle of maximum and minimum distance. The evaluation function can effectively avoid the shortcomings of artificially determining the parameters, making the selection of the initial clustering center more stable. The evaluation function is as follows:

$$S_i = \frac{D(x_i) - S(x_i)}{\max\left[D(x_i), S(x_i)\right]} \tag{7}$$

Among them, S (xi) represents the sparsity of data object xi; D (xi) is the minimum distance between data object xi and other selected initial cluster centers; the value of Si is between -1 and 1. When Si is close to 1, it means that the periphery of data i is compact and far away from other cluster centers.

The improved algorithm flow is as follows:

Input: K, D, maximum number of iterations T and termination conditions.

Output: K clusters that meet the termination condition and the number of iterations

Step1: According to equations 1 to 5, calculate the sparsity of the set of objects consisting of M inputs points for each data object, and initial first clustering center by select the minimum sparsity point;

Step2: Calculate the remaining data objects and the selected initial clustering center value according to Equation 7, select the point with the largest value as the remaining initial clustering center, and loop in turn until satisfy K centers;

Step3: K initial clustering centers are used for clustering.

3.4. Experimental platform and data

The goal of this improved algorithm is to improve accuracy and efficiency, explore and research new algorithms and provide a re-optimized thought framework for its further development and application. By building an experimental environment on the Windows system, coding is completed to verify the feasibility, correctness and operating efficiency of the improved algorithm. The experimental simulation data in this article mainly come from the Internet, and the data are obtained from several commonly used websites in daily life to verify the algorithm, which can better explain the versatility of the algorithm. The data registered by the user on the website can be imported with verified data content. On the one hand, the data comes from the current popular travel network data. From these data, quickly mine users with similar interests, similar content, and similar travel items. Analyze groups of people with similar concepts, so as to assist enterprises to build travel recommendation webchat groups for different groups of people, categorize and push travel promotion information, and carry out targeted travel-related advertising and promotion. The other part of the data refers to the transaction data of the current popular shopping websites. For the data of the untransacted access items, the user's interest trends are found, the transaction and the untransacted data are analyzed separately, and the corresponding shopping recommendation promotion is targeted.

3.5. Experimental results

Table 1. Comparison of results of different algorithms

Feature	Traditional algorithm			Max distance method			Improve algorithm		
E	A	R	F	A	R	F	A	R	F
0.015	44.00	43.9	43.89	46.00	45.88	45.92	48.56	44.41	46.37
0.020	44.59	43.9	44.22	43.39	43.19	43.27	47.45	46.51	46.99
0.025	54.20	53.79	53.98	45.79	44.05	44.89	58.43	58.26	58.36
0.030	48.81	41.58	42.18	37.47	37.36	37.42	45.87	43.45	44.63
0.035	45.67	44.45	45.05	47.68	45.92	46.79	51.35	54.98	53.11

E: Extraction rate; A: Accuracy; R: Recall rate; F: F measure;

Aiming at the principle that the cluster mean cannot replace the new cluster center in the clustering process of Web logs, new center calculating method in the K-center algorithm is used to obtain the center point in the iteration. It can be seen from Table 1 that the improved algorithm has a certain improvement in clustering quality compared to the traditional algorithm. It can be seen from the accuracy and recovery data that the improved algorithm has certain advantages over the original algorithm in both accuracy and effectiveness. Therefore, the algorithm is more suitable for clustering Web logs than the traditional algorithm.

4. Conclusion

From the above analysis, we can see that this article first introduced data mining and Web log mining in detail, and gave a detailed description of the entire process of data mining. The application of clustering methods is more and more extensive, and the research on clustering algorithms is more and more. Among the many clustering algorithms, K-Means is currently the most popular clustering algorithm. This paper analyzes the basic principles and processes of the algorithm, and based on the algorithm Disadvantages, an algorithm for selecting initial clustering centers based on sparseness is proposed to avoid clustering and obtain only local optimal solutions, which can effectively reduce algorithm iteration time and improve clustering quality. Due to various reasons, some issues need to be further improved, and further research is needed in the follow-up work. There is a large gap between the amount of data processed in the current experiment and the amount of log data in the actual site. It is necessary to provide distributed computing capabilities to process massive log data and user clustering, thereby improving processing efficiency. Although the improved K-Means algorithm has been verified from the perspective of correctness, subsequent researches are expected to improve the effectiveness of the results by introducing density clustering knowledge.

References

[1] He, Y. (2013) Research on decision support system based on WEB data mining [D]. Changsha: Hunan University, no.58.
[2] Chen, G.Y. (2019) Research on computer information processing technology in the era of big data. Network Security Technology and Application, no.3, pp.44 + 52.
[3] Pan, L. (2017) Research on Personalized Recommendation System Based on Web Mining [D]. Jiangsu University of Science and Technology.
[4] Xiong, F.Q. (2016) Summary of Web Data Mining. Electronic World, no.18, pp.98-99.
[5] Zhai, D.H., Yu, J., Gao, F., et al. (2014) Research on K-means Text Clustering Algorithm for Selecting Initial Cluster Center by Maximum Distance Method. Journal of Computer Applications, vol.31, no.3, pp.713-715.
[6] Zhang, R.X. (2014) An Enhanced Agglomerative Fuzzy K-Means Clustering Method with Mapred uce implementation on Hadoop Platform[A]. IEEE Beijing Section.Proceedings of 2014 IEEE International Conference on Progress in Informatics and Computing[C]. IEEE Beijing Section, no.5.
[7] Charu C. Aggarwal. (2017) Introduction to Special Issue on the Best Papers from KDD 2016. ACM Transactions on Knowledge Discovery from Data (TKDD), vol.11, pp.4.
[8] Lu, J., Diao, Y.J. (2012) Research on Data Preprocessing in Web Log Mining. Journal of Jiangsu University of Science and Technology (Natural Science Edition), vol.26, no.1, pp.81-85.
[9] Yang, J.W. (2019) Data mining technology from the perspective of cloud computing. Electronic Technology and Software Engineering, no.5, pp.151.