Research on Model and Algorithm of User Access Pattern Data Mining

Feng Pan

Southwest Jiaotong University Hope College, Chengdu, Sichuan, China

Keywords: Component, Big data, Data mining, Kohonen, Gaussian-shaped

Abstract: The era of big data has come. Today, all kinds of information and data are showing explosive growth. Internet activities of different scales are struggling to catch up with the pace and pace of the development of “big data.” Strictly follow the steps of data mining, we use time series mining algorithms, and use Microsoft's data mining tools to model the data sets collected from data halls, so as to discover the user's online behavior patterns and potential online rules within a certain period of time. This paper made reasonable suggestions for the scientific management of the campus network. A new clustering method based on the improved Kohonen self-organizing feature mapping neural network is proposed. A Gaussian-shaped membership function has introduced to output several neurons with a degree of membership greater than the threshold, thereby solving the problem of mining users' multiple interests.

1. Introduction

The study of online user behavior analysis and modeling is mainly divided into two methods. The first one is based on the social perception of user behavior in a standardized scenario. The second method focuses on the analysis of user behavior records and products. The author [1] adopts the second method, from the perspective of time evolution, by comparing and analyzing Internet navigation traces (URLs relative to keywords) and characterizing them as individual user or group user behavior, proposes an access redundancy. Degree is used as a global static parameter user online behavior analysis model. Tsuyoshi Murata and Kota Saito [2] introduced a method based on keyword analysis of website keywords to clarify the user interest degree; this method is mainly used to extract the sub graphs that embody the user's main interests in Web log data. The specific characterization of the user session is reflected in [3- 5]. At the same time, it also provides preliminary results in different aspects, including the request for each session, the number of pages for each session request, the length of the session, and the internal session time. . Through the application research of the entropy mixing model and Markov hybrid model, the paper discusses the requirements, steps and system framework for establishing a network user behavior analysis system, as well as some key technologies in the process of designing and implementing the system. Based on the above, several commonly used network user behavior analysis system models or their ideas are proposed. The application of network user behavior in Web click flow analysis, computer, and network security, and user structure analysis in Intranet network are listed. In addition, a path clustering method based on SIODATA algorithm has also proposed. The paper [7] mainly studied the network user behavior classification system and several common network user behavior analysis system models, combined with the example of the campus network of Southwest University of Science and Technology for analysis, through data analysis, and on this basis combined with network broadcast Structural theory proposes reform proposals for existing networks. In the dissertation [8], Zhang Jing used the cluster analysis and discriminant analysis function of MATLAB software to analyze the user's online log in a certain period in a certain university, and discovered the behavior pattern of the user's access to the Internet during the sampling period, for scientific purposes. Network management provides the basis.

WEB mining is a hot research topic in the current computer science. Applying neural network

technology can enhance the intelligence of WEB mining. Kohonen neural network is an unsupervised, self-organizing feature map network. Because it trains training weights through competitive learning and automatically derives the center of each cluster, it is widely used in the fields of pattern recognition, mode control, and so on. Based on its advantages in pattern clustering, this paper applies it to user access patterns mining. The user access mode represents the user's interest in accessing the website. By mining user access patterns, you can improve the performance of Web servers, improve the site structure, identify potential customers in e-commerce, and improve the quality of user services. However, there are still some deficiencies when Kohonen neural network is used for user access mode mining. Since the Kohonen neural network only outputs the neurons with the smallest Euclidean distance between the input samples and the output neurons, the optimally matched output neurons, it is applied to the user access pattern mining, that is, only one of the users is reflected. Interest, but ignore the user's other interests, so it is not suitable for users digging a variety of interests. In view of this, this paper introduces the Gaussian membership function to improve the Kohonen neural network algorithm and outputs several neurons whose degree of membership is greater than the threshold, to solve the problem of mining users' multiple interests.

Because of previous research, this paper conducted in-depth mining analysis of the http request log from November 2016 to May 2017 of the Boston University School of Computer Science and Technology extracted [9], and extracted the most visited users. Some websites and their content, and access to relatively concentrated periods and other related information from different time granularities, provide support for the scientific management of school network resources, optimizing network configuration, and guiding students to better study.

2. Related Work

2.1 Data Transformation

The data set used in this paper has three characteristics. 1. Log file is not a common txt file, but consists of three fields that contain the user name, machine name, log generation time (these three fields uniquely identify a log file) and by space Separated files. 2. The data set is very large with 9633 log files, 1,143,839 records. Each log record consists of six fields: machine name, online time, user ID, access URL, response time, and file size. Among them, the time format is UNIX time stamp. 3. The data record is incomplete. In the log records, some data were incomplete and incorrectly formatted. Therefore, before we analyze the data, we need to preprocess the data. All the time in the log file is in the form of a Unix timestamp. To facilitate observation and processing, we need to write a timestamp conversion function to process the format and convert it to the local system time at Boston University. For example, the Unix timestamp 797704525 is converted from the timestamp conversion function: 1995-04-13 00:35:25.000. At the same time, for those cases in which the suffix of the domain name is missing and the suffix name is incorrect, we need to adopt the means of data transformation to uniformly format the database before storing it in the database. For suffix-missing items, we fill it with NULL.

2.2 Kohonen Neural Network Overview

The Kohonen neural network is a two-layer feed-forward neural network, namely the input layer and the output layer. The input layer contains m neurons, the number of neurons equals the dimension of the input sample vector, and the output layer neurons are competing output neurons. Each neuron in the input layer aggregates external information into each neuron of the output layer through a weight coefficient vector, and through the self-organizing learning of the input pattern, the classification result has expressed in the output layer. Kohonen neural network through repeated learning of the input mode, so that the distribution density of the connection weight vector and the probability distribution of the input pattern tends to be consistent, that is, the connection weight vector space distribution can reflect the statistical characteristics of the input mode. The improved Kohonen neural network algorithm is a

combination of the improved Kohonen neural network algorithm and membership function. Its working process is divided into three stages: coarse adjustment learning stage, fine adjustment learning stage and application stage, and coarse adjustment of the learning stage to roughly determine the input mode. In the mapping position of the competitive layer, in the fine-tuning learning phase, the network learning set adjusts the connection weights of neurons in a smaller range. The learning principle is the same as the principle of the competitive learning process of Kohonen neural network, through the calculation of the competition layer. The Euclidean distance between the input sample and the output neuron is used for similarity judgment [10- 13]. The Euclidean distance-minimizing neuron is considered as the winning neuron, the weight coefficient of the winning neuron itself and other neurons in the neighboring domain is modified, and the neighboring neuron is modified. Mutual stimulation, while the more distant neurons have little effect. Finally, the weight coefficient connected to each output neuron is adaptively adjusted to have a certain distribution, that is, a vector composed of weight coefficients connected to the same output neuron is a class center, representing a class of features. Stores weight coefficients connected to all output neurons. In the application phase, when a new sample is input, the Euclidean distance between the new sample and each output neuron has calculated by the connection weight coefficient stored in the learning phase, and these distances are then input into the membership function to calculate each input sample to the output nerve. The degree of membership reflects the degree of closeness of the sample to the center of the class. The closer the sample is to the center of the class (that is, the smaller the Euclidean distance), the higher the degree of membership relative to it and the farther from the class center (the Euclidean distance is longer). A large sample has a lower degree of membership relative to this class [14]. A threshold is set based on experience. If the degree of membership of the input sample to an output neuron is greater than the threshold (the number of such neurons may be one or more), then the corresponding class of the neuron is output.

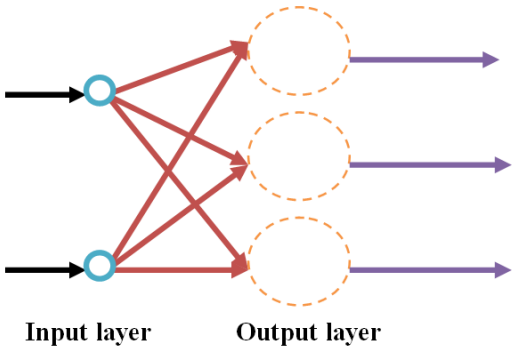

Fig.1 Improved Kohonen Neural Network Structure

$x_1, x_2, ..., x_m$ is the m-dimensional input of the Kohonen neural network, which is denoted as $X = [X_1(t), X_2(t), \ldots, X_m(t)]^T$, $w_{11}, w_{12}, ..., w_{mn}$ is the m*n weight coefficient vector of the input neuron and the output neuron, denoted as $W_j = [w_{1j}, w_{2j}, ..., w_{mj}]^T$, $j = 1,2,\ldots,n$, $\mu_1, \mu_2, ..., \mu_n$ means for n output neurons, which is the n fuzzy output values for the neural network.

3. User Access Pattern Data Mining Model and Algorithm Construction

3.1 Kohonen Learning Rate and Neighborhood Selection

Considering that the standard Kohonen algorithm only uses linear decrement for learning rate adjustment, the convergence time of clustering is slow, and only the square or circle domain is used for the neighborhood. When the weight adjustment is performed, all neurons in the neighborhood are also activated. The final clustering effect is not ideal, so we must think of ways to improve the performance

of the network from the aspects of improving the learning rate, the neighborhood, and so on. In the SOFM algorithm, the learning rate and the declining manner of the neighborhood are critical to the learning speed and clustering accuracy [15]. Through experiments, it has found that when the learning rate and the neighborhood range of the standard Kohonen algorithm decrease linearly, although the clustering result is better, the learning convergence is slower. When the index form decreases, the clustering result is not ideal or even leads to non-convergence. The function is decremented, the clustering result is best, and the convergence is fast. The decreasing method used is as follows.

$$\eta(t+1) = \eta_0 (a_1 / \eta_0)^{t/T_1} \tag{1}$$

In the formula, a1 is a constant, generally can take 0.05.

According to a neurobiological perspective, the strength of the lateral feedback should be related to the distance between the neuron i in the neighborhood and the neuron c that wins the victory. The typical Gauss function is best suited as a neighborhood function.

$$N_{ic}(t) = e^{(-|W_i - W_c|^2)/2\sigma(t)^2} \tag{2}$$

In the formula, $\sigma(t)$ is the effective width of the neighborhood, and $|W_i - W_c|$ is the distance between the neurons in the competition layer. In the two-dimensional network, $|W_i - W_c| = \sqrt{(i_x - c_x)^2 + (i_y - c_y)^2}$ represents the coordinates on the array between the neurons. Moreover, we use the power function to decrease.

$$\sigma(t+1) = \sigma_0 (a_2 / \sigma_0)^{t/T_1} \tag{3}$$

In the formula, the parameter a2 is a constant, generally takes 0.5.

3.2 Gaussian Membership Function

According to the characteristics of the practical problems dealt with in this paper, the value range of Euclidean distance is between [0, 2], and when the Euclidean distance is 0, the degree of membership is 1. When the Euclidean distance is 2, the degree of membership is 0. The closer the Euclidean distance is to 0, the closer the membership degree is to 1. The closer the Euclidean distance is to 2, the closer the membership degree is to zero. The Gaussian membership function can satisfy this requirement. Therefore, Gaussian membership function is used as the membership function of the neural network.

Gaussian-shaped membership function, with two characteristic parameters, is a Gaussian function. Its function expression is as follows.

$$\mu_A(x, \sigma, x_1) = e^{-((x-x_1)/(\sqrt{2}\sigma))^2} \tag{4}$$

Since the range of Euclidean distances calculated in this paper is between [0, 2] and is a monotonically decreasing function, $x_1 = 0$, $\sigma^2 = 0.2$, the Gaussian membership function can be simplified as:

$$\mu(C_j(X)) = \mu(d_j(X), 0, \sqrt{0.2}) = e^{-\frac{d_j(X)^2}{0.4}} \tag{5}$$

3.3 Data Preprocessing

The improved Kohonen neural network model is a two-layer neural network model, namely the input layer and output layer (competition layer). The input layer contains 22 neurons, representing 22 pages on the site that reflect the user's interest in the visit. The output layer contains 10 neurons, representing the site's desire to classify all users into 10 categories. W_{ij} is the connection weight coefficient of the

input layer and the output layer, initialized to a small random number. After a period of training, the same output neuron W_{ij} represents the characteristics of the class (neuron). $\mu_1, \mu_2, ..., \mu_n$ is the output of the model, indicating the extent to which the input sample belongs to each class. The value $\mu_1, \mu_2, ..., \mu_n$ is between [0, l]. The website structure used in this article comes fromWEB and is appropriately modified for the needs. The 22 pages of the input layer have represented by $X_1, X_2, ..., X_{22}$, and the specific page $X_1, X_2, ..., X_{22}$ represented in the website is as follows:

X1	X2	X3	X4	X5
Humanities	Social	Sciences	Management	Economics
X6	X7	X8	X9	X10
Mathematics	Physics/Mechanics	Chemistry/Chemicals/Materials	Earth	Science/Astronomy
X11	X12	X13	X14	X15
Biology	Electronic	&	Electrical	Computers/Networks
X16	X17	X18	X19	X20
Mechanical	Civil/Construction	Environment/Energy	Transportation/Aviation/Aerospace	Music
X21	X22			
Movie	TV			

This article does not use the entire website page as the input of the model. Since there are too many pages in the website, a few pages can reflect the user's visit interest. Therefore, in this article, the web pages are categorized and all the web pages are divided into tables. The 22 directions shown. In the table, X13 represents books accessing computers and networks [16]. When a user accesses the book “Java programming ideas” in the computer book, adding 1 to the X13 count will not make the model too large, but it will also benefit the site [17].

The data in this article comes from the Xinhua Bookstore in Jiangxi Province, where user logs with user sessions>5 were selected for mining, making the mining access mode more meaningful. The preprocessed data is as follows: The data shows the number of visits to each page since each user was registered. For example, the first data in User1 is 4, indicating that User 1 accesses the X1 page and its linked page 4 times. $X = [x1x2x3...x22]$.

4. Results and Discussion

4.1 Results

In this article, five users were randomly selected for testing. The data was preprocessed: User1: [4015000200010002212700], User2: [1500100302302000300204], User3: [3500000001100000100103], User4: [1200100302100000200000], User5: [0000100101000000300002], when training step is the output of the model, the numbers are 50 and 1000 are as follows.

Table 1 Result of Model Output

Input	Training steps=50	Training steps =1000
1	X[20]= 0.92512	X[1]= 0.87100 X[4]= 0.60215 X[20]= 0.77313
2	X[2]= 0.84727	X[2]=0.71417 X[8]= 0.89328 X[11]= 0.63147 X[22]= 0.61635
3	X[2]= 0.83238	X[1]= 0.600332 X[2]= 0.62417 X[22]= 0.60245
4	null	X[8]= 0.83786
5	null	X[17]= 0.80620604

According to the table, when the number of training steps is 50, since there are too few training samples, the access behavior of some users cannot be recognized, and therefore these users do not have the data output satisfying the conditions. For the analysis of The user data already output, it can be seen that both the user 2 and the user 3 output are x[2], indicating that both users are interested in social science. When the number of training steps is 1000, the user 1, the User 2 and User 3 has a number of interests. User 1 has a wider interest, that is, both like to look at the humanities books, but also like the economy and television drama, it can be accurately divided into 3 categories, namely X [1], X[4], X[20]. Analyze all the data and concluded that, whether the user is interested in a single user or a user with a wide range of interests. The model can more accurately cluster the user after 1000 steps of training.

4.2 Analysis of Overall Traffic/Request Times in Comparison Period

Through the comparative analysis of the user access trends on March 17, 2017 (0 to 24 o'clock) and March 18, 2017 (0 to 24 o'clock), the following analysis conclusions can be obtained as follows.

1) There is no obvious change in the overall trend of the user's request. The slope of the linear trend analysis curve is greater than 0 (and the linear fit is high), which means that the number of requests for analyzing the morning hours in days is less than other time periods.

2) The two-day user's request situation has a better fit from 20 to 23 o'clock, while the number of requests at other times on March 17, 2017 is slightly higher than March 18, 2017 (only 15 The point and 20 points are slightly lower than the amount of petitions on March 18, 2017. This can be clearly reflected from the two-day linear trend curve. The overall analysis of the number of user requests on March 17, 2017 is higher than that on March 18, 2017.

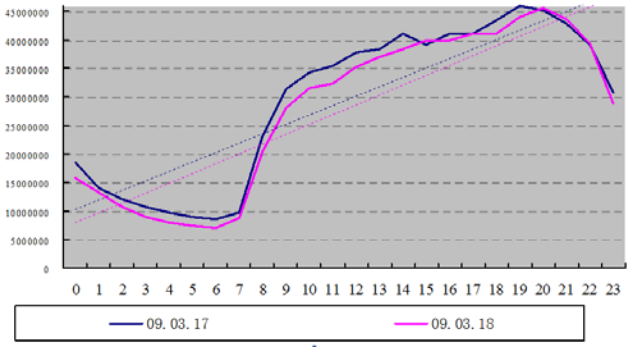

Fig.2 Power Market Experiences to Date

4.3 Comparing and Fitting Analysis of Traffic/Visits of Similar Websites (Resources)

By comparing the same type of websites in the region, it is possible to quickly discover the characteristics of such websites (or value-added services) in the region. At the same time, the same types of websites can provide differentiated competition while providing mutual reference and reference. The strategy (also provides an important reference for designing new MMS or WAP services).

Through the user's popularity index at different times, you can understand the user's access patterns and habits, stickiness, and service performance differences; through the mid-to-long-term tracking and analysis can also understand the site (business) users of the loss (increase), found The user's migration model further extends the user's satisfaction and loyalty.

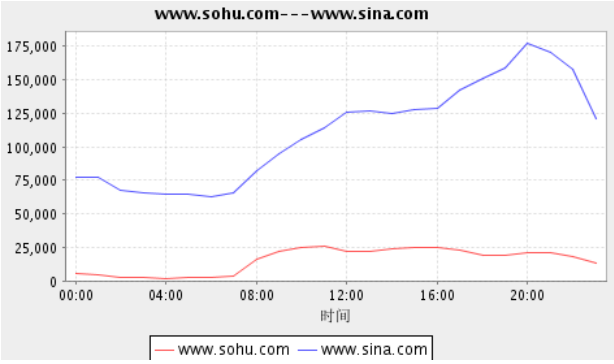

Fig.3 Power Market Experiences to Date

5. Conclusion

Exploring meaningful user access patterns and related potential customer groups from a large number of customer data and log data plays an important role in the business decisions of merchants implementing e-commerce strategies. This article proposes a comprehensive consideration of servers based on the Kohonen model. The user access patterns of multiple data sources, such as logical design, page topology, and user's browsing path, and the mining algorithms of potential customer groups. The clustering accuracy of Kohonen network is greatly improved. At the same time, the user access pattern mining model based on the improved Kohonen neural network algorithm has a memory function and can easily cluster new users. Because the fuzzy Gaussian membership function is used as the output function of the model, the model can provide recommendations that are more accurate for a variety of different interests of users.

References

[1] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107.

[2] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 47, 98-115.

[3] Dutt, A., Aghabozrgi, S., Ismail, M. A. B., & Mahroeian, H. (2015). Clustering algorithms applied in educational data mining. International Journal of Information and Electronics Engineering, 5(2), 112.

[4] Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Applied Computing and Informatics, 12(1), 90-108.

[5] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.

[6] Su, Q., & Chen, L. (2015). A method for discovering clusters of e-commerce interest patterns using click-stream data. electronic commerce research and applications, 14(1), 1-13.