ATLANTIS PRESS

Advances in Economics, Business and Management Research, volume 100

International Conference of Organizational Innovation (ICOI 2019)

A Novel Method of Applying Big Data for Analysis Model of Library User Behavior

Kaijun Yu (Library, Shanghai University of Medicine & Health Sciences, Shanghai, China)
Song Luo (Department of Equipment, The Affiliated Hospital of Guizhou Medical University, Guizhou, China)
Xuejun Zhou (Department of Equipment, The First People's Hospital of Nantong, Jiangsu, China)
Rui Wang (Micro Mechatronics Laboratory, Yamaguchi University, Ube, Japan),
Longjie Sun (Library, Shanghai University of Medicine & Health Sciences, Shanghai, China)

Email：sunlj@sumhs.edu.cn

Abstract—A large number of library user behaviour data generated in real time in the era of big data artificial intelligence requires more efficient and scientific analysis technology to help libraries improve the level and quality of personalized services, while the increasingly popular campus Internet of Things system needs to be more Active network security precautions, proactively detect unreliable abnormal behavior of the network and feedback users to improve security awareness. Explores a big data analysis model using traditional data mining and classification learning, which combines user personality analysis and abnormal behavior detection.

Keywords—*Data mining, Supervised learning, User portrait*

I. INTRODUCTION

In the era of Big Data and Internet of Things, information interaction means are more abundant, convenient and personalized. Mobile devices such as PC, smart phones, iPad and Kindle have long been the main tools for people to read. The main way for users to acquire, recognize, utilize and communicate knowledge information is gradually transferred to major Internet platforms. With the continuous development of artificial intelligence technology, various intelligent interactive devices in libraries are emerging[1]. Every day, library users need to carry out frequent operations and data transmission when using service products. A large number of user behavior logs are generated and continuously transmitted to the server storage background to form massive data. This, on the one hand, makes it difficult for existing network data mining technologies to analyze and process these data in a timely manner, leads to the lack of lag in user behavior analysis, and makes it difficult to launch personalized and rich service products. On the other hand, frequent user behaviors bring security risks to users' own account information, and also put forward higher requirements for library digital resource network security.It is necessary to detect the network abnormal harmful behavior in time to avoid the loss.

This paper analyzes the behavior of library users through the method of literature analysis. The development process of user behavior data analysis technology was analyzed, and the core technologies of user behavior personality analysis and credibility analysis were defined.Seven core data analysis technologies were selected as the framework to build the big data analysis model of user behavior in this paper.

II. EVOLUTION OF DATA ANALYSIS TECHNOLOGY

A. Mathematical statistics analysis

In China in the early 1990s, papers on mathematical statistics analysis of user behavior through borrowing records, questionnaires and other methods have been published[2].With the rapid development of Internet technology in the early 21st century, the website statistics and custom software industry developed based on the principle of data statistics gradually began to be commercialized[3-4]. For example, eXTReMe Tracking provides URL real-time tracking service and user website browsing statistics. Web Site Traffic report sends user access traffic statistics in the form of email, MiniTab software statistical analysis user usage habits questionnaire, etc. The preset firmware only provides partial statistical parameter data, the scope and accuracy of predicting user behavior is limited, but it is of great significance for digital libraries entering the web2.0 era. In-depth data mining analysis research began to gradually extend to the various businesses of the library.

B. Data mining

In recent years, with the development of the Internet of Things, the library-related business data has also shown a trend of increasing over the years. The diversity and complexity of user behavior information data also provides an excellent opportunity for data mining technology in library application research. The following is a brief introduction to data mining.

1) Cluster analysis

Cluster analysis is to group the elements in a set according to a certain degree of similarity, and then form the clustering classes[5]. The inner elements of a class differ less from each other (that is, they are more similar). The more rigorous mathematical description of

clustering is as follows[6]. Studied sample sets K, class M is defined as a set of the loophole of K, M ⊂ K, and M ≠ K. Different classes M_i(i=1, 2, 3, 4, ...) that satisfy the following two conditions are clusters.

$$M_1 \cup M_2 \cup \cdots\cdots \cup M_n = \mathrm{K} \quad (1)$$

$$M_i \cap M_j = \emptyset \quad (\mathrm{i} \neq \mathrm{j}) \quad (2)$$

According to condition 1, each sample must belong to one of the clusters. Condition 2 means that each sample belongs to no more than one class. Clustering is a very important part of data mining. Clustering itself is not a specific algorithm, but a universal task. The main clustering algorithms have six categories, which are segmentation based algorithms, layer-based, density-based, grid-based, model-based[7]. Experts and scholars at home and abroad have continuously improved the main clustering algorithms based on hierarchy and density through long-term unremitting efforts. Pei Jifa et al. proposed a sample distribution density function as the initial membership matrix of FCM clustering algorithm for density-based clustering[8-9].

2) Correlation analysis

Association analysis is also called association mining. The goal of association analysis is to find strong association rules. Support and confidence are the important basis to determine whether an association analysis method is successful. Most existing algorithms based on association rules need to use support and confidence to filter out association effects. There are mainly 6 association algorithms: Apriori algorithm and its optimization algorithm, multi-dimensional association mining, multi-level association mining, constrained association mining, statistical association, and unstructured complex type association[10].

3) Time series analysis

Time series refers to a sequence that ranks the data values of a certain statistical indicator according to the time sequence in which it occurs[11].Its typical characteristics are large data size, high data dimension and noise. Time series analysis technology has been widely used in the development of all walks of life, the technology is very mature. Time series analysis techniques are currently divided into two categories based on stage development. The first category is a time series analysis method based on mathematical statistics, which focuses on the stochastic process of statistical analysis of discrete indicators. The second category is the data mining based time series analysis technology adopted in this paper. The research hotspots mainly focus on the approximate representation of time series, similarity measure, classification, clustering, pattern mining, anomaly detection, etc[12].

C. Supervised learning analysis

Machine learning algorithms are the core areas of artificial intelligence applications. The main classification algorithms are Naive Bayes, Support Vector Machine, Integrated Learning, etc.

1) Naive Bayesian analysis

Bayesian analysis is an algorithm that uses prior probabilities for classification and prediction[13]. It is based on Bayes' theorem, calculating the probability that a data sample of an unknown category belongs to each category, and selecting the most likely one as the final category. Naive Bayes classification requires feature attributes to be conditionally independent or substantially independent. This classification work firstly calculated the conditional probability and occurrence frequency of each category corresponding to each data feature in the data sample of the training set, and then applied bayes theorem to the data feature of the test set to calculate the possibility of occurrence of each category.

2) Support vector machine

Support Vector Machine (SVM) is a new learning method based on statistical VC theory and structural risk minimization criteria[14]. The SVM method maps the low-dimensional space of training data samples into a high-dimensional feature space (Hilbert space) through a nonlinear mapping K, which transforms the original linear indivisible problem into a linear separable problem in high-dimensional space. A simple mathematical of SVM is introduced in the following.

The SVM separating hyperplane:

$$w^T x - b = 0 \quad (3)$$

Eq. (3) is calculated by solving the quadratic optimization problem:

$$\min_{\omega,b} \frac{1}{2} w^T w \quad (4)$$

Subject to $y_i(w^T \phi(x_i) + b) \geq 1$. Where x_i represents the *i*th training vectors, y_i represents the class value (±1), $\Phi(x_i)$ maps the input data to the feature space, i.e., applying a given function to the input data, for instance a polynomial function. Then, the classification of a given input *x* is made by finding the sign of the Eq. (3):

$$f(x) = \mathrm{sgn}(w^T x - b) \quad (5)$$

Where the class of the input vector x_i is:

$$\begin{cases} y_i = -1 \cdots f(x_i) \leq 0 \\ y_i = +1 \cdots f(x_i) \geq 0 \end{cases} \quad (6)$$

This method is referred to as soft classifier. Actually, the impact of misclassified vectors does not appear in the Eq. (4). For approach that is more rigorous Eq. (4) is rewritten as:

$$\min_{\omega,b,\xi} \frac{1}{2} w^T w + C\sum_i \xi_i \quad (7)$$

Subject to $y_i(w^T\phi(x_i)+b) \geq 1-\xi_i$. Here the impact of misclassification is counted in the optimization step. ξ_i define slack variable that represent the degree of misclassification of a training vector and C the regularization parameter represents a bound on the Lagrange multipliers α such as.

$$\min_{\alpha} \frac{1}{2} \alpha^T L\alpha + e^T\alpha \quad (8)$$

Subject to $y^T\alpha = 0$, $0 \leq \alpha_i \leq C$, $\forall i\ \xi_i \geq 0 \forall i$. Where $y^T=[y_1, y_2, \ldots, y_N]$ is the vector of class values (±1), $e^T=[1, 1, \ldots, 1]$ is a vector of ones, $L_{ij}= y_i y_j K(x_i, x_j)$ and $K(x_i, x_j)= \Phi(x_i)^T\Phi(x_j)$ is the kernel function that performs the nonlinear mapping of the input data into the feature space. In our study, we used polynomial kernel of degree d=5 and the regularization parameter C=∞. The polynomial kernel is of the form.

$$K(x_i, x_j) = (x_i^T x_j + 1)^d \quad (9)$$

Through the above transformation, a linear hyperplane can be found for classification analysis tasks, so the kernel function is the most important part of SVM. In the machine supervised learning model, support vector machine and neural network are very practical, can analyze data, identify patterns, and perform efficient classification and regression analysis.

3) Integrated learning

The basic idea of integrated learning is to continuously call multiple learning algorithms to gain stronger learning ability. However, so far there is no consistent classification of integrated learning, and most scholars tend to fall into four categories through their research results.

1. Bagging: It is currently the only useful method for unstable nonlinear models.
2. Boosting: This method is used in the classification prediction explored in this paper.
3. Stacked Generalization: Although it has not been widely accepted so far, as the research progresses, it will continue to tap the potential.
4. Random Subspace Method: The training data set uses non-traditionally randomly selected input subspaces, such as the feature space of the training data set, and the output is combined by majority voting.

III. USER BEHAVIOR BIG DATA ANALYSIS MODEL

The above provides a variety of practical and efficient analysis techniques for massive data generated by library users in the era of big data, which can analyze user behavior personalization.

A. Personalized analysis of user behavior

Personalized analysis of library user behavior is based on human-computer interaction log records, web browsing records, digital resource downloads, platform interactive information, etc., through the above data collection and supervised learning, analysis and prediction of user behavior. Two training data samples can be constructed according to the user's response time to each module of the library - coarse-grained training data and fine-grained training data[15]. By extracting and building training data based on image features and text features posted by users on post bar, message board and chat platform, the integrated learning classifier of support vector machine and gradient promotion can be used to analyze and predict users' publishing habits. Finally, collaborative filtering constitutes the complete user behavior portrait. Analysis flow chart is shown below.

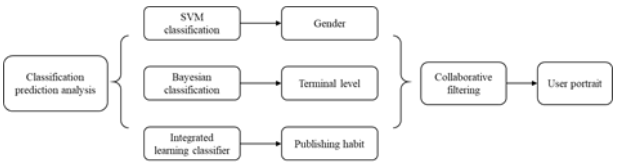

Fig. 1. User Personality Information Analysis

B. User behavior credibility data analysis

The smart library itself is in the Internet all the time, inevitably suffering from various network security issues. The interaction of various applications of interactive devices, the negligence of personal account settings, and the lag of software updates can easily lead to the loss and tampering of user information. How to prevent micro-duration and timely detection and find that identifying untrustworthy user behavior has become another focus of this paper. The user behavior credibility data analysis flow chart is as follows.

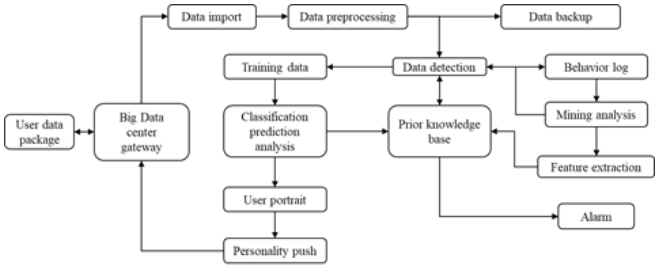

Fig. 2. User behavior big data analysis model

Firstly, all data generated by user behaviors in real time are collected in the gateway of big data center. After the preliminary data preprocessing and other steps, the data detection based on the prior knowledge base is started. If normal is determined, the next step is to train the data set of classification learning technology. If abnormal is determined, data mining is carried out for its behavior log. This is mainly based on cluster analysis,

with multi-dimensional association rules and time series analysis to accelerate the deep global optimization search of large-scale data. If the new abnormal behavior is confirmed, the feature is extracted and added to the prior library knowledge and the user is warned. If it cannot be confirmed, conduct the second test. Similarly, the ability of machine learning to detect virus variants can also be used in the classification prediction of user personality analysis. The new feature categories are added to the prior knowledge base, and the remaining normal behaviors are predicted to form the user personality information through a series of analysis. After collaborative filtering, the user portrait is sketched, and finally the personalized recommendation information is sent to the user terminal to complete the entire user behavior analysis process.

IV. CONCLUSION

At present, the library is gradually transforming into a smart library under the leadership of the development of Big Data and AI technology. With the new intelligent interactive devices and analytics technology, many of the coveted library service concepts are truly realized. From the perspective of user behavior analysis, this paper studies the related literature technology to obtain the big data analysis model with both network security detection and user personality behavior analysis. In addition to enabling libraries to deliver personalized services to users in a more timely and accurate manner, this model also maintains users' security privacy and library network security at all times, and minimizes the impact of harmful behaviors.

REFERENCES：

[1] Thimm M. The Tweety Library Collection for Logical Aspects of Artificial
[2] Intelligence and Knowledge Representation. KI - Künstliche Intelligenz, 2016:1-5.
[3] Leon A Jakobovits, Diane Nahl Jakobovits, Lu Bing. Using the Library: User Behavior Analysis. Journal of Agricultural Sciences of Henan Province, 1990(3):131- 134.
[4] J.W. Hsieh, L.W. Huang, Y.S. Huang. Multiple-Person Tracking System for Content Analysis. Springer Berlin Heidelberg, 2001, 2195(4):897-902.
[5] P. Alpar, M. Porembski, S Pickerodt. Measuring the Efficiency of Web Site Traffic Generation. International Journal of Electronic Commerce, 2001, 6(1):53-74.
[6] A.K. Jain， M.N. Murty， PJ Flynn. Data clustering: a review. Acm Computing Surveys, 1999, 31(3):264-323.
[7] A.K. Jain， R.C. Dubes. Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice Hall.1988.
[8] Yin Ruifei. Clustering Method in Data Mining and Its Application—Based on Statistical Perspective. Xiamen University.2008.
[9] G. Karypis ， E.H. Han ， V. Kumar. CHAMELEON A hierarchical clustering algorithm using dynamic modeling. Computer, 2008, 32(8):68-75.
[10] Pei Jifa, Xie Weixin. Clustering Density Function Method. Journal of Xidian University, 1997(4): 463-467.
[11] R. Agrawal, T. Imielinski, A. Swami. Mining Association Rules Between Sets of Items in Large Databases. Proc.1993 ACM SIGMOD Int Conf. Management of Data. Washington, D.C, 1993: 207-216.
[12] Tunnicliffe-Wilson G. Non-Linear and Non-Stationary Time Series Analysis. Journal of Time, 2010, 10(4):385-386.
[13] Chuang A. Time Series Analysis: Univariate and Multivariate Methods. Technometrics, 2006, 33(1):108-109.
[14] Wen Zhicheng, Cao Chunli, Zhou Hao. Network security situation assessment method based on naive Bayesian classifier. Journal of Computer Applications, 2015, 35(8): 2164-2168.
[15] Czermiński R, Yasri A, Hartsough D. Use of Support Vector Machine in Pattern Classification: Application to QSAR Studies. Qsar & Combinatorial Science, 2015, 20(3):227-240.
[16] Willis C J. Hyperspectral image classification with limited training data samples using feature subspaces// Defense and Security. International Society for Optics and Photonics, 2004.