International Journal of Knowledge and Language Processing
Volume 7, Number 3, **2016**

www.ijklp.org
KLP International ©2016 ISSN 2191-2734
pp.1–27

Review of User Behavior Analysis and Application Based on Web Logs

Tu Meng
School of Software and Microelectronics
Peking University
No. 24, Jinyuan Road, Daxing District
Beijing, 102600, China
tumeng@pku.edu.cn

Received May 2016; revised June 2016

ABSTRACT. *User log analysis is an important part of the web data mining. This review summarizes the currently research status of user log analysis, including the analysis process, main methods, representative open source tools and its application researches, laying the foundation for further study.*
Keywords: user log analysis, user behavior, web logs

1. **Introduction.** Behavior analysis based on user log is an important method for pattern acquisition of user behavior on Internet. In January 2016, China Internet Network Information Center (CNNIC) published the 37th Statistical Report on Internet Development in China, which shows that as of December 2015, the number of Chinese Internet users reached 688 million, of which 527 million are mobile phone users, the Internet penetration rate reached 50.3%[1]. Due to such a large user base, there are a large number of user logs, which record the most direct users' operational behaviors in the Internet, even including the network users' patterns, they are of a great value for study. Research on the intelligence of the group contained in web user behaviors is becoming a research hotspot in the current information retrieval field.

1.1. **User Logs and User Behaviors.** Network user log, refers to a large number of user operation log generated by user while surfing the Internet, is a part of network data mining[2]. These user data represents all operations by user, including a lot of user access information, such as user's IP address, visited URL, date and time of visit, access path, etc[3]. which attracted many researchers' attentions in recent years.

Network user behavior, refers to a series of activities generated by user interacted with the Internet environment and services. User behavior analysis refers to the statistics and analysis of valuable data selected from the basic network access data, concluding the rules

of user access to webs, and combining these rules with network resources and network marketing strategy, so as to find out the possible problems in the current network resources integration and marketing activities, and provide the basis for further revising or reforming the network resources and marketing strategies.

1.2. **Objective and Meaning.** User log, as the truest data of user on Internet, has a great importance. User's intention and needs can be caught by user log mining. Through the analysis of network user behavior monitoring data, Internet service providers can have more detailed and clear understanding of user behavior, which contributes to improving user experience and meeting more user needs, ultimately enhancing the benefits of Internet service providers. At the same time, user behavior analysis helps users understand themselves better, master misuse behavior or inadvertently malicious behavior, etc. to provide statistical data.

As of the end of 2015, the most popular communications APP in China is WeChat, which has more than 650 million active users. Obviously, such a large user group led to a large user log, containing a lot of user's information, such as user behavior, preferences, interests, age, geographical information and so on. By studying user logs, companies can adjust their operating strategies, so as to develop a more reasonable direction of development.

China's Internet giants, Ali, Tencent, Baidu and Sina, release their user data “white paper” each year. Each company will develop more appropriate development planning and provide better services to meet user needs according to the unique characteristics of their own user groups, such as gender, age, geographical distribution, etc.

As for university, the most widely used technology of user log mining is the establishment of university library. Based on user log and user association mining, some services like personalized library, personalized recommendation of resources, can be provided for users. In addition, the online education industry can also carry out user log analysis to identify the user's points of interest, in order to improve their teaching system and programs. Besides, they can increase the course based on user interests. In healthcare industry, user log of some online medical sites can be used to understand user needs, even identify the types of information that users are looking for on Internet, meanwhile learn more about users’ habits and develop more rational medical plans in the future. .

The purpose of this paper is to summarize the current research of user log analysis, to help researchers understand the behavior characteristics of users in the Internet, and to dig out the relationship and practical application of user log in order to understand user needs and behavior better.

2. **User Logs Analysis.**

2.1. **Main Statistical Targets.** User log analysis and mining have multiple levels, each level contains a number of statistical indicators and a variety of methods. According to the existing research, we summarize and analyze the indexes of log analysis and their relationship. According to the "research level", the users' log analysis is divided into basic analysis, deep analysis and comprehensive analysis. The indicator of Basic analysis is the

basic user data, including basic statistics and analysis of the term, query, click and clickstream; deep analysis is based on basic analysis, each user and session is regarded as a unit for statistical analysis, the contents of analysis is a series of basic data collections; comprehensive analysis is more complex, adding models and contrast, such as contrast between different years of data features.

2.1.1. **Term level.** Counting and analyzing of "single word" entered by users on website search box, including the word itself, multi-language word and vocabulary spelling [4]. For example, counting the number of characters contained in each word, and mapping distribution; counting the proportion of Chinese and foreign language input, respectively [5]. The deeper research results include: frequency distribution of English lexical items accord with the characteristics of power-law distribution (or Zipf distribution) [6].

2.1.2. **Query level.** A query string is a set of strings that the user submits in the search box at one time. The length of the query, and the complexity of the query, are studied in the query string input by the user. Generally, query string analysis includes:

1) **length of query:** query string contains the number of terms; query string contains the number of characters; study the input string length and input time, the number of bytes of the query string.
2) **boolean queries,** such as the proportion of the string contained "AND, OR, NOT" and its content characteristics.
3) **query diversity:** the proportion of unique queries to the totals; extracting TOP-N query strings with the highest query frequency N, generating their frequency accumulation graphs; the number of query duplications is observed, making sure whether the graph is subject to power-law distribution. Entropy-percent[7] is introduced to quantitatively examine the diversity of intentions of a user in all query sessions.
4) **query topic and content:** the number of occurrences of the top N different query string, according to the theme of classification. There are two main classification methods: artificial classification and automatic machine classification. Among them, the choice of categories can refer to the traditional PC search log mining research topic classification.
5) **voice query:** studying the length of input voice query string, content features.

The main results are as follows[5,8-10]: The query string input by English search engine contains 2.2 ~ 2.4 English words, most of which are English words. The number of English words in the query string obeys the Poisson distribution. Most Chinese users input query strings contain only one word and Chinese characters, most of which has 2 to 4 Chinese character. Users who use complex queries account for a smaller proportion.

2.1.3. **Click and clicksteam level.** The URL characteristics that users click during the search is observed here, such as the distribution of the first N feedbacks on the result list, the total number of clicks and their distribution in a single session.

Clickstream, also known as path analysis, is used to discover the behavior of a user clicking on a link, as well as the jumping of user between different pages. The analysis items include: the conversion rate of query and view, the proportion of reference sources, high-frequency click path and so on. Click rate analysis can reveal the user interest path,

and help optimize the website topology.

The main results are as follows[11−12]: the number of users clicking different URLs is in accordance with Heaps' law, the frequency of clicking URL is same as the Zipf distribution, the URL clicking is related to the size of page, URL clicking has time locality and click process has self-similarity.

2.1.4. **User level.** During a period (such as a month, a week or a day, the data source determines the time period), user is treated as a statistical unit for the corresponding analysis. These include:

1) Analyze the number of visits, queries and page views, such as the average number of queries (views) submitted per day, average query (views) duration, query (views) time distribution, and query (views) content in different time periods.
2) user segmentation, such as the distinction between old and new users, distinction of geographical location, distinction of terminal device, etc.

2.1.5. **Session level.** The same conversation may contain one or more query strings, or one or more clicks. The number of submitted queries, the number of topics, duration, number of bytes sent, query modification times, and their characteristics, in the user search session are all counted.

2.1.6. **Behavior level.** From a more macro perspective on the operation of user, the log data can be divided by a classification method "different operations", which can be established on the basis of session, user and so on. There are three main methods of analysis: Method 1, according to the different log, the behavior is divided into "query behavior" and click (views) behavior, then explore the duration and operation content of these behaviors. Method 2, according to the different URL clicking in the file (or different visited pages), the user's behavior is divided into login, browse, query, download, exit and so on. Method 3, according to user interface layout of the site or search engine, distinguishing between actions according to "the purpose of mining and analysis".

2.1.7. **Results pages viewed.** Look at the number of results pages (such as pages turning, etc.), the number of viewing the snapshot, and the time interval between viewing the results page. The main results**[13]**: the majority of search engine users only view a few results pages, usually 1 to 2; time interval of viewing the pages is 2 to 3 minutes; few users view page snapshots, as Skynet data shows, the number of snapshots-clicking is only 3.5% of total traffic.

2.2. **Process of User Logs Analysis.** User log analysis mainly has four procedures as follows:

2.2.1. **Source Data Capture (User Logs Collection).** Web log data analysis tools, also known as user behavior analysis tools, are frequently used by webmaster and operators. There are many well-known domestic and international tools, such as Google Analytics, Baidu statistics, Tencent analysis and so on. The first step of these analysis tools is collecting web access logs. Besides, there are some specialized web log collection tools such as WebTrends, FastStats Analyzer, and Happy Log.

At present, the main source of log analysis data is from the server-side. Every user’s visit to Web will form a new access log on the server. For different servers, the log file record

format is almost the same. The W3C (World Wide Web Consortium) organization defines two formats for server logs: the Common Log Format (CLF) and the Extended Log Format (ECLF).

The current mainstream data collection method is based on Javascript, details of this method as follows.

1) Principle of Data Collection[14]

Web analysis tools need to collect the behavior of the user when they visited the target site (such as opening a Web page, click a button, add goods to the shopping cart, etc.) and behavior extra data (such as an order amount generated). Earlier web statistics often collected only one kind of user behavior: the page opened. But the user's behavior in the page cannot be collected. However, with the widespread use of ajax technology and the growing demand for statistical analysis of e-commerce sites, this kind of collection strategy can meet the basic analysis of traffic analysis, source analysis, content analysis and visitor attributes. So, traditional collection strategy has become unable to reach.

Later, Google added customizable data collection script in Google Analytics. With a small amount of javascript code, user could use the scalable interface defined by Google Analytics to realize the tracking and analysis of customized events and indicators. At present, Baidu statistics, Sogou analysis and other products copied the Google Analytics model. In fact, the basic principles and processes of two data collection models are similar, but the latter could collect more information through javascript.

2) Overview of Process

The process of data collection is as shown below:

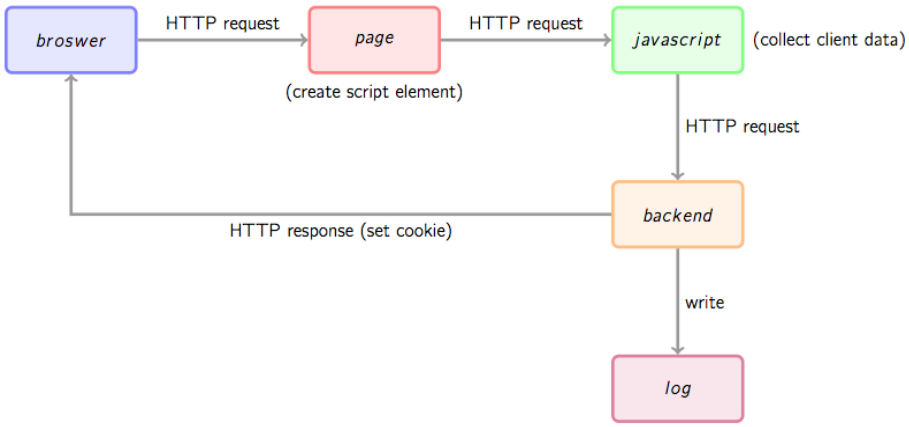


FIGURE 1. WEB SITE LOG DATA COLLECTION BASIC PROCESS

First of all, the user's behavior will trigger a http request from browser for the target page, here we suppose that user's behavior is "open the page." When target page is opened, the page in the buried javascript fragments will be executed. General site statistics tools will require users to add a small piece of javascript code in the page, the code snippet will dynamically create a script tag, and src points to the single js file, this time the js file (green node) will be requested by browser and then be executed. This js file is usually the real data collection script. After data collection is completed, js will request a back-end data

collection script (the figure backend), the script is generally a dynamic script disguised as a picture. js will send the data to back-end script through the http parameters, then back-end script will parse parameters and record them as a fixed format in access log. At the same time, some tracking cookies may be put in the http response for clients.

2.2.2. **Data Pre-processing.** Due to the uneven quality of the data obtained in the network, and some can not directly include the pattern mining, which need to obtain data on the basis of the need for data pre-processing.

In general, although the Internet environment has some of its own rules, but there are still uneven levels of the site, a different URL format will give a great deal of mining problems, the main purpose of preprocessing is to solve these format problems, while accurate The distinction between the different sessions, which is conducive to the pattern mining process. At present, the data preprocessing for network log is divided into four stages: data cleaning, user identification, session identification and path supplement.

As shown in the figure is Zidrina Pabarskaite & Aistis Raudys[15] proposed data preprocessing flow chart, data preprocessing, including the following processes:

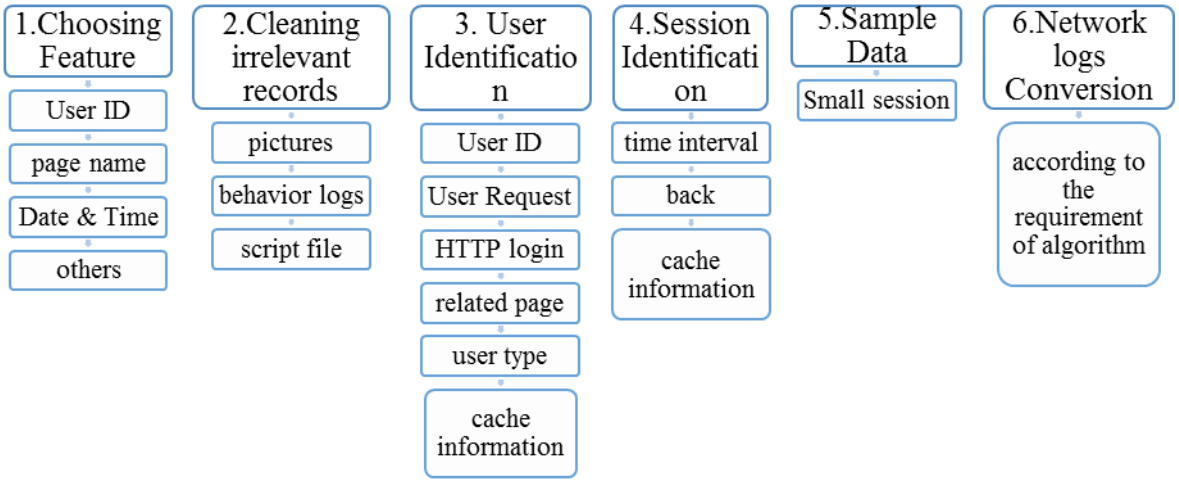


FIGURE 2. DARA PRE-PROCESSING

1) Data Cleaning

Data cleansing is the initial processing of the data. Data cleansing is to have noisy, inconsistent, irrelevant data (mainly gig, jpeg, jpg, etc. at the end of the URL to clean up, these data are generally in the form of pictures and video content, does not mean that the user access intent, therefore You need to remove the data from the Web log data source.

As shown in the figure is to show the need to clean up the original user access log:

```
-
172.25.171.112 - - [10/Mar/2015:09:21:45 +0800] "GET /images/bg.gif HTTP/1.1" 304 -
172.25.171.112 - - [10/Mar/2015:09:21:45 +0800] "GET /images/all.gif HTTP/1.1" 304 -
172.25.171.112 - - [10/Mar/2015:09:21:45 +0800] "GET /images/ss_9.gif HTTP/1.1" 304
-
172.25.171.112 - - [10/Mar/2015:09:21:45 +0800] "GET /images/nj_06.gif HTTP/1.1" 304
-
172.25.171.112 - - [10/Mar/2015:09:21:45 +0800] "GET /images/nj_02.gif HTTP/1.1" 304
-
172.25.171.112 - - [10/Mar/2015:09:21:45 +0800] "GET /images/nj_04.gif HTTP/1.1" 304
-
172.25.171.112 - - [10/Mar/2015:09:21:45 +0800] "GET /images/1.png HTTP/1.1" 304 -
172.25.171.112 - - [10/Mar/2015:09:21:45 +0800] "GET /images/2.png HTTP/1.1" 304 -
172.25.171.112 - - [10/Mar/2015:09:21:45 +0800] "GET /images/lvse_01.gif HTTP/1.1"
304 -
172.25.171.112 - - [10/Mar/2015:09:21:45 +0800] "GET /images/nj_05.gif HTTP/1.1" 304
-
```

FIGURE 3. THE USER ACCESS LOGS TO BE CLEANED UP

2) User Identification

The second stage is user identification. The purpose is to identify a single user, to determine the access path of a single user. User identification can be carried out by four methods:

a) Through HTTP access to identify, this method by the user login name to determine the difference.

b) By the type of customer to determine the user, that is, by the user to access the IP address to determine, Piroll, P and Pitkow, J[16] in the same IP can be judged by the user's browser and operating system to determine the operation.

c) Through the site site to determine the difference, Cooley[17] proposed, if not through the existing Web page to open a new page, you can determine for different users.

d) This is the most widely used and most authoritative judgments, that is, through the use of the user's site to achieve some of the cache records to determine the user. Roberts[18] suggested that the first time a user sends a cached record to the system using the browser, a second user will not have a new cache sent to the system when he or she opens the page again for user identification.

3) Session Identification

The third stage is dealing with the conversation, Spiliopoulou, M[19] pointed out that a session is a user at a particular time on the same site access sequence.

Conversation judgment can also be determined by four methods:

a) According to the user access time to determine, from 1994 onwards, different scholars have different views on the timing of the session, the first to make the earliest time to determine the session time is 1994 Pitkow J., & Margaret R.[20] proposed that if a user browsing a Cathi, LD, & Pitkow, JE[21] in the second year will be clear time for 25.5 minutes; Paliouras, G. et al[22] that the user visits an hour can be judged as a session; He D, & Goker A.[23] claimed that 10-15 minutes is the best choice.

b) According to the time interval to determine the pages of the visit, the user for a page residence time of more than a corresponding value to determine the end of the session. Zhuang Like[24] and so put forward by the analysis of the user's overall

situation, to determine the time interval will be spent.

c) By judging whether the user performs a return operation when browsing the Web page, Chen, and M.S. et al[25] found that if the user performs a return operation during browsing, then a session is judged to be ended. But now the user can directly browse the page itself is returned, so this judge is not a good way to judge.

d) Berendt, B[26] and so on that by opening the user to determine whether the new page is opened through the original page links, and if so, then determine a session, or a new session and proposed several exploration methods to determine the user session.

The first is to set a limit α, all sessions can not exceed this time, the initial session time is set to If the access time t, t-t0 <α, then determine a session.

The second is to set a page stay time β, if the residence time is less than β is considered a session.

The third kind is if the request Q comes from the current conversation, join Q to the current conversation, and otherwise begin as new conversation.

Some scholars use the session time to weight the page[27], by judging the session time to determine the importance of the page, when mining, the important information of the page weight greater, with more important mining value.

1) Path Completion

Pabarskaite Z & Raudys A final processing of the data, access to data mining patterns can be achieved, that is, the fourth stage - the path complement. When the user accesses the web page, the user can not record the user information in the log because the local cache is returned when the user clicks the button to return to the previous page. The main function of the path supplement is to supplement these missing information, so as to supplement the user's access behavior information.

Yan LIa, b[28] and other scholars put forward the use of user session path set for path complement.

If the URL of a record is not the URL of the previous record, it is assumed that the user has returned to the previous page and needs to supplement the path. In this case, the path of the user session is determined.

Munk M, JKapusta J[29] proposed that if there is no hyperlink between the last two items of an access sequence, then it may be determined that the user may have clicked the return button on a page of the access sequence and reached the last item, To trace back to the path complement, you can use forward-looking way to find the path.

2.2.3. **Pattern Mining.** Pattern mining is the use of data mining algorithms through the processing of pre-processed log mining effective, novel, potential, useful and ultimately understandable information and knowledge. Using Patterns Data mining techniques are commonly used in the path analysis technology and data mining commonly used in the field of association rules, classification and clustering technology, the next section will be described in detail.

2.2.4. **Pattern Analysis.** After the pattern mining process is finished, the user can obtain the information he needs according to the acquired user pattern. The pattern analysis is mainly applied by the researcher through the mining algorithm to obtain the information.

This is the specific application part of the log mining. Analysis of the role of the researchers can be based on the results obtained to improve the relevant measures. Section IV of this article is described in detail.

2.3. **Methods of User Behavior Analysis.** In the use of a variety of user behavior logging tools to obtain user behavior data, we must first choose the appropriate method for analysis of these data. The existing analysis methods include statistical analysis, cluster analysis, association analysis, decision tree, neural network and time series data mining methods, the following are introduced separately.

2.3.1. **Statistical Analysis Method.** Statistical analysis is the most basic behavior analysis method, mainly on the user behavior decomposition, classification after the number of statistics, the amount of data obtained a type of behavior, and by means of the total amount of data analysis of user behavior[30]. In recent years, there are some studies at home and abroad have adopted a statistical analysis method, Cheng Peng[31] through the collection of user clicks and access logs and other data, the use of statistical analysis of data technology for user behavior analysis; You Ting[32] and Deng Xiawei[33] also statistical analysis of social Web site user behavior characteristics were studied.

2.3.2. **Clustering Analysis Method.** Cluster analysis is a kind of exploratory analysis. In the process of classification and clustering, it does not set the classification criteria in advance, but automatically classifies it based on the sample data. In view of the complexity of user behavior data, cluster analysis has a wider range of use than classification statistics, and it is also the most commonly used analysis method in user behavior analysis. SangHyun Oh[34] and so on through user behavior modeling, proposed a kind of clustering analysis method which is applied to user's abnormal behavior monitoring, especially discussed the feature value selection problem for clustering. The experiment proved that the clustering analysis is more accurate than the statistical analysis Describe user behavior. Marcelo Maia[35], etc. by capturing a large number of Youtube user data, clustering algorithm used to aggregate the same behavior of the user, so as to better characterize different types of user behavior. SUN Yan-hua [36] (School of Computer Science and Technology, Nanjing University, Nanjing 210096, China) Based on the Netflow flow information, the IP is taken as a source and destination statistical model. The fast clustering K-Medoids algorithm is improved, and an anomaly anti-selection criterion is designed. The clustering anomaly. LIU Peng[37] This paper proposes a fast hierarchical clustering algorithm based on entropy for data grouping and merging multiple data samples at one time based on data characteristics to study the regularity of the change of customer service preference with time and to analyze the behavior of network users on and off line . CAI Yue[38], a search engine algorithm based on user behavior clustering, mining user intent from the user behavior log, and based on user feedback to locate user intent information. Chen Min [39] introduced the concept of roughness and proposed a new path similarity calculation method. In the calculation of similarity degree, not only the user's browsing pattern was considered as a sequence pattern, but also the fullness The time factor of users browsing on the Internet is considered to realize the clustering of Web browsing behaviors. Based on traffic monitoring, an entropy-based access state transition matrix (CWSTM)

clustering algorithm is proposed by Ting Hao[40] for the analysis of user session behavior, the analysis of web-time preference, the analysis of Web preferences and the state transition of user Web access. Zhang Xia and others[41] proposed a clustering analysis algorithm based on user query intention. By retrieving the user keywords, the data information was stored in a tree form so as to obtain the upper and lower information of the adjacent tree. The key information was used to analyze the users of the query intent.

2.3.3. **Association Analysis Method.** In the analysis of user behavior, user behavior habits and other behavioral habits are often analyzed by association rules algorithm such as Apriori, so as to discover the relationship and regularity of different behavioral habits, so as to achieve the purpose of user behavior prediction. Wang Aiping[42] introduced the algorithm of mining association rules in data mining from the aspects of width first, depth first, data set partitioning, sampling, incremental updating, constraint association and multi-layer multi-dimension association. WANG Yong-li[43] (School of Information Science and Engineering, Central South University, Changsha 410083, China) In the process of Web user behavior mining, we find the shortcomings of Apriori, a classical algorithm of association rules, and propose an algorithm for mining association rules based on Web mining. LUO Qiang[44] (School of Computer Science and Technology, Beijing University of Aeronautics and Astronautics, Beijing 100875, China) This paper proposes Apriori algorithm for social network user behavior association analysis algorithm to divide virtual communities and improve the relevance of user information push for virtual community.

2.3.4. **Decision Tree Method.** Decision tree uses information and data tree to provide decision-makers with decision support for follow-up questions. In the process of network user information behavior analysis, the decision tree is also used in the analysis of network user information behavior because of the prediction of user's follow-up information behavior. Xu Xiaojuan et al[45]. Used Ethnographic decision tree to analyze the behavior of science blog users and use questionnaire to verify the prediction effect. He Lu[46] used the decision tree method combined with linear regression analysis to predict the social network user's personality traits. LI Xian-peng et al[47]. (2003) analyzed the ID3 decision tree algorithm widely used in classification prediction, and pointed out the disadvantages such as the bias of the algorithm and the low computational efficiency. On this basis, an improved ID3 algorithm was proposed and applied to some Mobile churn prediction. Zou Jing, et al[48]. Clarified the decision tree algorithm is an important means to improve customer loyalty and prevent customer churn in telecom industry, and introduce the method, steps and concrete realization process of decision tree algorithm in telecommunication industry customer loss analysis.

2.3.5. **Neural Network Method.** Neural network, also called connection model, is an algorithmic data model that mimics the behavioral characteristics of animal neural networks and performs distributed parallel information processing. The neural network is mainly used to predict the information behavior of users in the analysis of network user information behavior, dynamically adjust the corresponding strategies to provide more personalized service. Liu Rong et al[49]. Proposed an adaptive neural network modeling

method combining manual customization and system automatic modeling, which can dynamically adjust the parameters of neural network and modify the user model so that the output of neural network can be changed according to the user's interest. Based on the improved LPCA neural network method, a new user behavior analysis model is proposed by Zuo Lin[50], which reveals the influence of the user region attributes on the usage patterns and application habits of the neural network.

2.3.6. **Time Series Data Mining.** Timing data refers to time-related or time-dependent data or time series represented by numbers or symbols. Since the relevant data changes continuously over time, it can reflect the state or performance of a certain process in a certain period of time. In view of the obvious timing characteristics of network user information behavior, this method can learn the past time characteristics of network users and can predict the future behavior of users. Hutchins[51] collected data from more than 600 million unique user logins from May to September 2000 from a US dial-up Internet service provider (RADIUS) server, starting with 1 minute intervals for holiday, weekend, and weekday log-in changes The description then analyzes the online time length distribution, the geographical distribution of users, and the average logon time in different regions, and also makes an estimation model for the number of users. Subsequently, Hutchins[52] uses the same data source to describe the change in the number of log entries in 5-minute intervals over several months, as well as the online time-length distributions for several months, and the log-in interval of individual users distributed.

3. **Log Collection and Analysis Tools.** From the perspective of information security risk management, enterprises need to run for all types of systems log and user network access behavior audit system, through the network Zhiji management system to understand the behavior of Internet users, trends. Log management and analysis plays an important role in network security as an integral part of network security defense system.

Log Management Infrastructure consists of hardware, software, networks, and media used to generate, transfer, store, analyze, and process log data. Log management, usually include the following modules [53]:

a) Log Collection
b) Centralized log Aggregation
c) Long-term Log Storage and Retention
d) Log Rotation
e) Log Analysis (real-time and off-line)
f) Log Visualization （search and reporting）

3.1. **Open Source Log Collection System.** User log analysis is based on log collection, want to carry out in-depth log data mining, it must be established on the efficient and large-scale log collection. Many of the company's platform will generate a lot of daily log (usually streaming data, such as the search engine pv, queries, etc.), these logs require a specific log system, in general, these systems need to have the following characteristics:

1) build the bridge between the application system and the analysis system, and decouple the association between them;

2) support near-real-time on-line analysis systems and off-line analysis systems like Hadoop;
3) has a high scalability. That is, when the amount of data increases, it can be extended horizontally by adding nodes.

The following four aspects of the design architecture, load balancing, scalability and fault tolerance, a brief introduction to today's four major open-source log collection system: including Facebook scribe, Apache chukwa, Linkedin kafka and Cloudera flume and so on.

3.1.1. **Scribe of Facebook.** Scribe is a facebook open source log collection system, has been a large number of applications within the facebook. It collects logs from a variety of log sources and stores them on a central storage system (NFS, distributed file system, etc.) for centralized statistical analysis. It provides a scalable, highly fault-tolerant solution for the "distributed collection, unified processing" of logs.

Its most important feature is fault tolerance. When the back-end storage system crash, scribe will write data to the local disk, when the storage system back to normal, scribe will reload the log to the storage system.

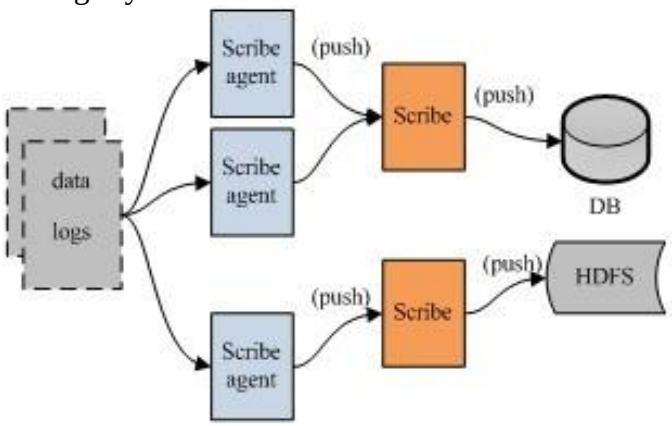


FIGURE 4. ARCHITECTURE OF SCRIBE

Architecture: Scribe architecture is relatively simple, mainly consists of three parts, namely scribe agent, scribe and storage systems.

1) **Scribe agent.** The scribe agent is actually a thrift client. The only way to send data to the scribe is to use the thrift client, which scribe internally defines a thrift interface that the user uses to send data to the server.
2) **Scribe.** Scribe received thrift client sent over the data, according to the configuration file, the topic will be different data sent to different objects. Scribe provides a variety of store, such as file, HDFS, etc., scribe data can be loaded into these stores.
3) **Storage systems.** Storage system is actually scribe in the store, the current scribe support a lot of store, including file (file), buffer (double storage, a main storage, a storage), network (another scribe server), bucket Multiple stores, stores the data in different stores via hash), null (ignores data), thriftfile (writes to a Thrift TFileTransport file), and multi (stores data in different stores at the same time).

3.1.2. **Chukwa of Apache.** Chukwa is a very new open source project that uses a number of hadoop components (with HDFS storage and mapreduce data) because it is part of the

hadoop family. It provides a number of modules to support hadoop cluster log analysis. Demand:

1) flexible, dynamically controllable data source

2) high-performance, high scalable storage system

3) appropriate framework for the collection of large-scale data analysis

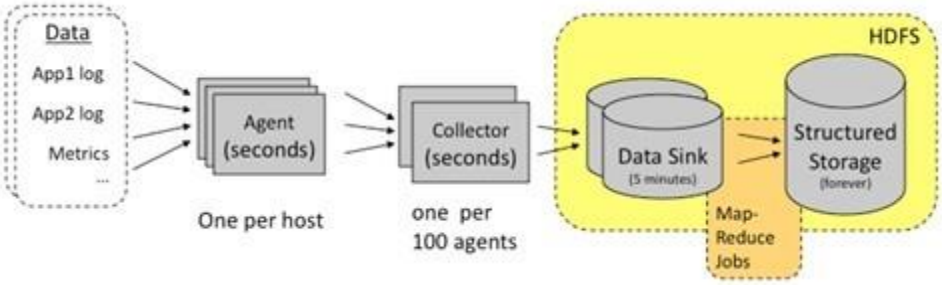


FIGURE 5. ARCHITECTURE OF CHUKWA

Architecture: Chukwa has three main roles, namely: adaptor, agent, and collector.

1) **Adaptor data source.** Can be packaged with other data sources, such as file, unix command line tools. Currently available data sources are: hadoop logs, application metrics data, and system parameter data (such as linux cpu use flow rate).

2) **HDFS storage system.** Chukwa uses HDFS as a storage system. HDFS is designed to support large file storage and small concurrent high-speed write applications, while the log system features the opposite, it needs to support high concurrent low-speed write and a large number of small file storage. Note that a small file written directly to HDFS is not visible until the file is closed, in addition, HDFS does not support the file to re-open.

3) **Collector and Agent.** In order to overcome the problem in (2), the agent and collector phases are added.

Agent role: to adaptor to provide a variety of services, including: start and close the adapter, the data will be passed to the Collector via HTTP; regularly record the adapter state, in order to resume after the crash.

Collector role: the data from multiple data sent over the merger, and then loaded into the HDFS; hide HDFS implementation details, such as, HDFS version replacement, simply modify the collector can be.

4) **Demux and achieving.** Direct support for processing data using MapReduce. It has two built-in mapreduce jobs, which are used to retrieve data and transform data into structured log. Stored in the data store (which can be a database or HDFS, etc.).

3.1.3. **Kafka of LinkedIn.** Kafka is an open source project in December 2010. It is written in scala and uses a variety of efficiency optimization mechanisms. The overall architecture is relatively new (push / pull) and is more suitable for heterogeneous clusters. Design goals:

1) the data on disk access cost O (1)

2) high throughput, in the ordinary server can handle hundreds of thousands of messages per second

3) distributed architecture, able to partition the message

4) to support the parallel data load hadoop

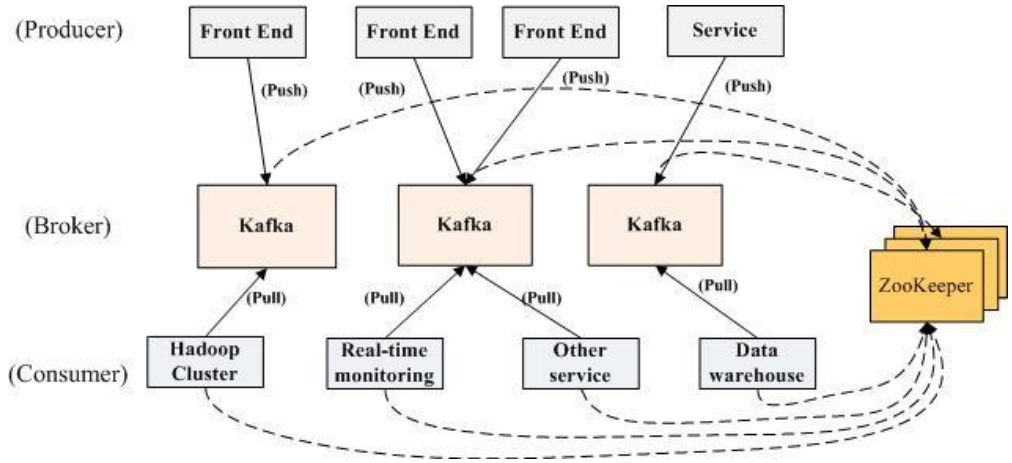


FIGURE 6. ARCHITECTURE OF KAFKA

Architecture: Kafka is actually a news release subscription system. A producer posts a message to a topic, while a consumer subscribes to a topic message, and as soon as there is a new message about a topic, the broker passes to all consumers subscribing to it. In kafka, the message is organized by topic, and each topic will be divided into multiple partitions, so easy to manage data and load balancing. At the same time, it also uses zookeeper for load balancing. Kafka has three main roles, namely, producer, broker and consumer.

1) Producer. Producer's job is to send data to the broker. Kafka provides two producer interfaces, one is the low_level interface, the use of the interface to a particular broker under a certain topic to send data partition; the other is a high level interface, the interface supports synchronous / asynchronous send Data, zookeeper based broker automatic identification and load balancing (based on Partitioner).

Among them, based on the zookeeper broker automatic recognition is worth. The producer can get the list of available brokers through zookeeper, or the listener in zookeeper, which will be awakened in the following cases:

a) Add a broker
b) Delete a broker
c) Register a new topic
d) Broker Register the existing topic

When the producer learned of the above time, according to the need to take some action.

2) Broker. Broker has taken a variety of strategies to improve data processing efficiency, including sendfile and zero copy technologies.

3) Consumer. The role of consumer is to load the log information to the central storage system. Kafka provides two consumer interfaces, one is low level, it maintains a connection to a broker, and the connection is stateless, that is, each pull data from the broker, it is necessary to tell the broker data bias Shift. The other is the high-level interface, which hides the details of the broker, allowing the consumer to push data from the broker without worrying about the network topology. More importantly, for most of the log system, the consumer has access to the data stored by the broker, and in kafka, by the consumer to maintain their own access to data information.

3.1.4. **Flume** of **Cloudera.** Flume is cloudera in July 2009 open-source log system. Its built-in various components are very complete, users almost do not have any additional development can be used. Design goals:

1) Reliability. When a node fails, the log can be transferred to other nodes without loss. Flume provides three levels of reliability protection, from strong to weak in turn, respectively: end-to-end (received data agent event written to the disk first, when the data transfer is successful, then delete; if the data transmission failure , You can re-send.), Store on failure (this is scribe strategy, when the data receiver crash, the data written to the local, to be resumed, continue to send), Best effort (data sent to the receiver, not Will be confirmed).

2) Scalability. Flume uses a three-tier architecture, were asked agent, collector and storage, each level can be extended horizontally. All agents and collectors are managed by the master, which makes the system easy to monitor and maintain, and the master allows multiple (using ZooKeeper for management and load balancing), which avoids single point of failure.

3) Manageability. All agents and colletors are managed uniformly by the master, which makes the system easy to maintain. The user can view various data sources or data flow execution status on the master, and can configure and dynamically load each data source. Flume provides web and shell script command in two forms of data flow management.

4) Functional scalability. Users can add their own agent, colletor or storage as needed. In addition, Flume comes with a lot of components, including a variety of agents (file, syslog, etc.), collector and storage (file, HDFS, etc.).

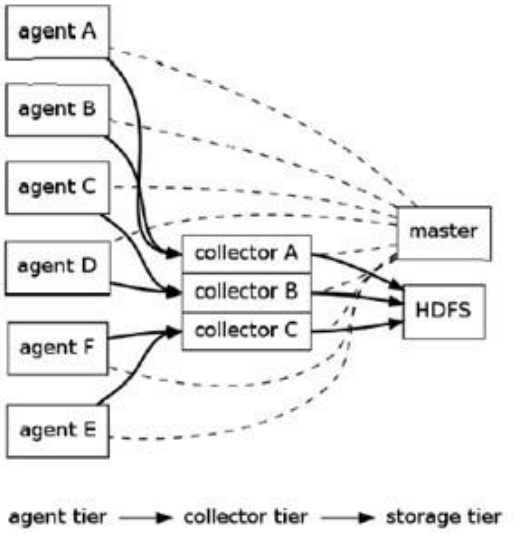


FIGURE 7. ARCHITECTURE OF FLUME

Architecture: As mentioned earlier, Flume uses a layered architecture, composed of three layers, namely, agent, collector and storage. The agent and collector are composed of two parts: source and sink, source is the data source, sink is the data going.

1) Agent. The role of the agent is to send the data source data to the collector, Flume comes with a number of directly available data source (source), such as:

Text ("filename"): The file filename as a data source, sent by line
Tail ("filename"): detect the filename of the newly generated data, sent out by line
FsyslogTcp (5140): listening to the TCP port 5140, and the received data sent out
While providing a lot of sink, such as:
Console [("format")]: The data will be displayed directly on the desktop
Text ("txtfile"): Write the data to the file txtfile
Dfs ("dfsfile"): Writes data to the dfsfile file on HDFS
SyslogTcp ("host", port): Passes data to the host node over TCP

2) collector. The role of collector is to aggregate the data of multiple agents, loaded into the storage. Its source and sink are similar to agents.

3) storage. Storage is a storage system, can be a common file, it can be HDFS, HIVE, HBase and so on.

3.2. **Open Source Log Analysis System.** At present the open source Web log analysis tools are many, some tools for some of the advantages of the user's recognition and widespread. Most Web log analysis tools attempt to extract more information from the log records, but the software that performs the analysis stably, and the data analysis has readability, and can display the analysis results graphically is rare. Here are three of the most representative of the system:

3.2.1. **AWstats.** AWStats is an open source Web analytics reporting tool, suitable for analyzing data from Internet services such as web, streaming media, mail, and FTP servers. AWStats is a free powerful and featureful tool that generates advanced web, streaming, ftp or mail server statistics, graphically. This log analyzer works as a CGI or from command line and shows you all possible information your log contains, in few graphical web pages. It uses a partial information file to be able to process large log files, often and quickly. It can analyze log files from all major server tools like Apache log files (NCSA combined/XLF/ELF log format or common/CLF log format), WebStar, IIS (W3C log format) and a lot of other web, proxy, wap, streaming servers, mail servers and some ftp servers. Small and medium-sized Web site suitable for this log analysis tool for analysis.

Advantages： The use of cross-platform Perl CGI language, can be very good in Windows, Unix, Linux and other operating systems to run; Graphical interface, excellent, able to provide a very fine analysis of reports and graphics; The report structure is logical and readable. Configuration is relatively simple to install, do not need to compile and install the local; Can be well supported in Chinese, to provide multi-language support; Software version upgrades faster, the latest version of the updated Chinese search engine, browser and so on.

Disadvantages: The lack of content (such as columns, channels) and other in-depth analysis; Lack of analysis of China's geographic information; Insufficient search engine and keyword analysis (compared to WebTrend); Slightly slower (compared to Webalizer);The report only a simple bar graph, do not use the more representative of the line graph, pie chart and area map; Lack of visits, the number of pages per visit click on a number of key statistics.

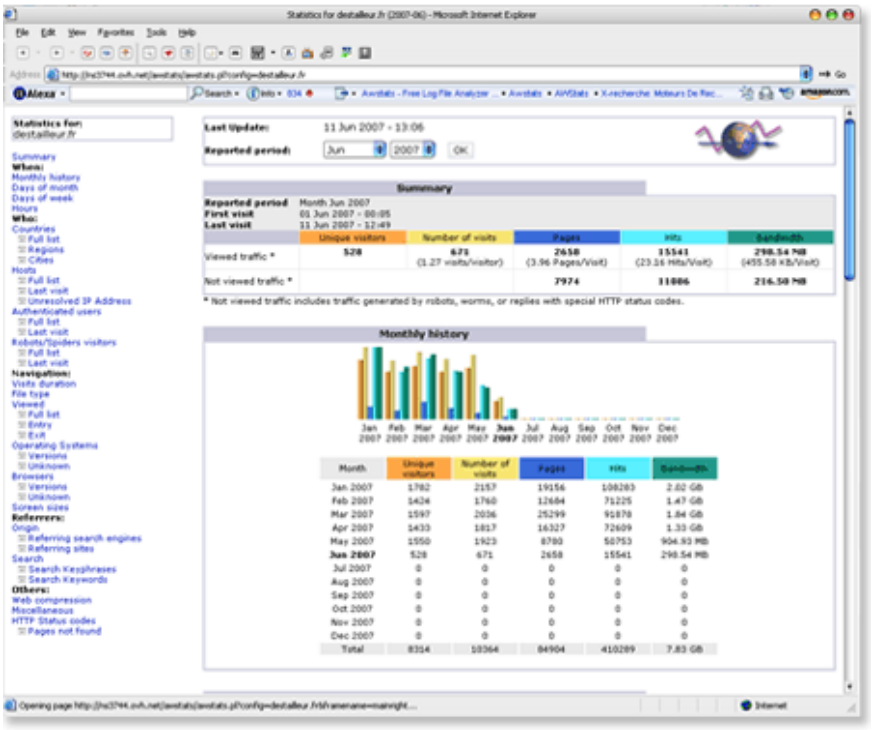
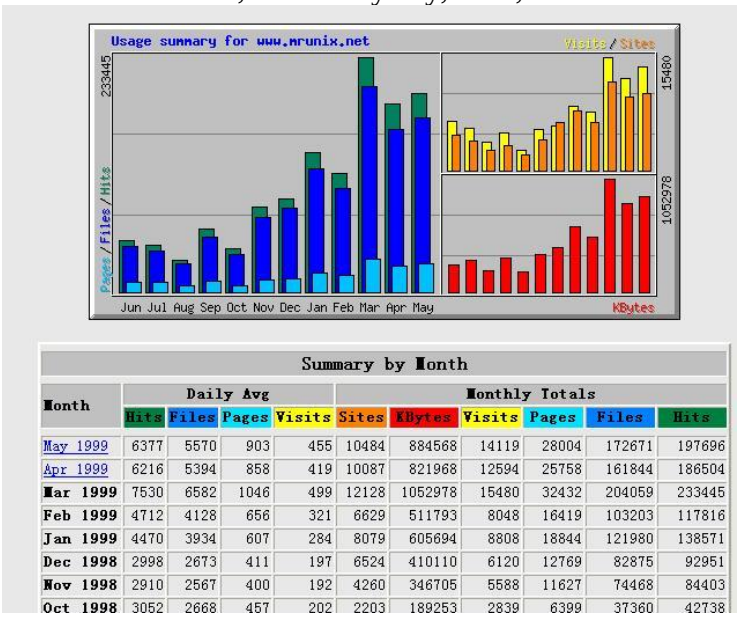


FIGURE 8. INTERFACE OF AWSTATS

3.2.2. **Webalizer.** The Webalizer is an application that generates web pages of analysis, from access and usage logs, i.e. it is web log analysis software. It is one of the most commonly used web server administration tools. It was initiated by Bradford L. Barrett in 1997. Statistics commonly reported by Webalizer include hits, visits, referrers, the visitors' countries, and the amount of data downloaded. These statistics can be viewed graphically and presented by different time frames, such as by day, hour, or month.



Summary by Month

Month	Daily Avg				Monthly Totals					
	Hits	Files	Pages	Visits	Sites	KBytes	Visits	Pages	Files	Hits
May 1999	6377	5570	903	455	10484	884568	14119	28004	172671	197696
Apr 1999	6216	5394	858	419	10087	821968	12594	25758	161844	186504
Mar 1999	7530	6582	1046	499	12128	1052978	15480	32432	204059	233445
Feb 1999	4712	4128	656	321	6629	511793	8048	16419	103203	117816
Jan 1999	4470	3934	607	284	8079	605694	8808	18844	121980	138571
Dec 1998	2998	2673	411	197	6524	410110	6120	12769	82875	92951
Nov 1998	2910	2567	400	192	4260	346705	5588	11627	74468	84403
Oct 1998	3052	2668	457	202	2203	189253	2839	6399	37360	42738

FIGURE 9. INTERFACE OF WEBALIZER

Advantages： Using C language, running fast; Can be cross-platform operation; Can provide illustrated basic report.

Disadvantages: Need to be compiled in the local installation; Update is not timely, the last update is April 2002; The analysis is not deep enough to provide 15 basic reports.

3.2.3. **Analog.** Analog is a free web log analysis computer program that runs under Windows, Mac OS, Linux, and most Unix-like operating systems. It was first released on June 21, 1995, by Stephen Turner as generic freeware; the license was changed to the GNU General Public License in November 2004. The software can be downloaded for several computing platforms, or the source code can be downloaded and compiled if desired.

Analog's interesting list includes how many hits from each country, which search engine queries the user has brought to the site, and which browser and which operating system the visitor uses. This software can display all the information in the web server log. The software has a slightly improved graphical interface compared to the Webalizer software based on the GD graphics library. However, the pie icon and the bar chart are far from ideal.

Advantages: Using C language, running fast; Can be cross-platform operation; Report navigation is more chic, convenient; Can provide illustrated basic report, analysis depth slightly better than Webalizer.

Disadvantages: Need to be compiled in the local installation; Update is not timely, the last update is November 2004; Chinese reports are not supported.

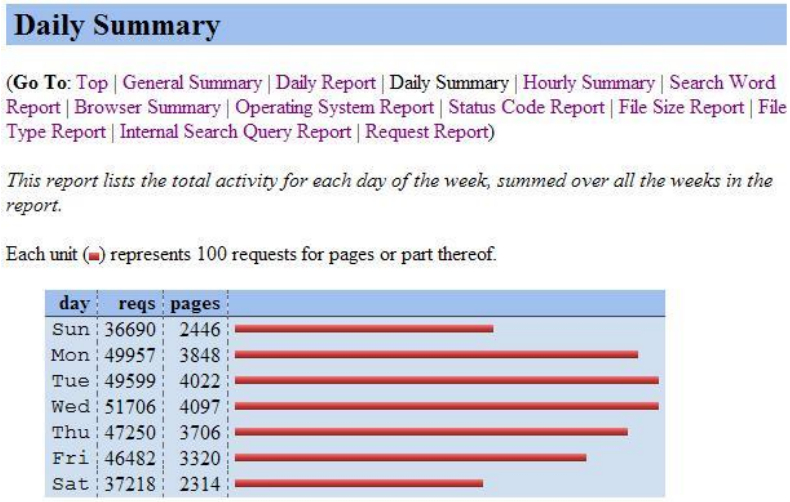


FIGURE 10. INTERFACE OF ANALOG

4. **Application Research on User Logs.** The analysis of network user logs plays an important role in various fields, the following are several representative applications.

4.1. **Optimization of Site Navigation.** Web site optimization using the site's pre-caching method, according to the user's operation log, the use of machine learning algorithms, analysis of user logs, can help the site decision cache content, and then solve the slow loading problem.

Baskaran KR and others[54] through the cache in the log information, the use of SVM-LRU method to determine the content of the cache page to enhance the site user experience.

The following figure shows the process of site navigation optimization[55]:

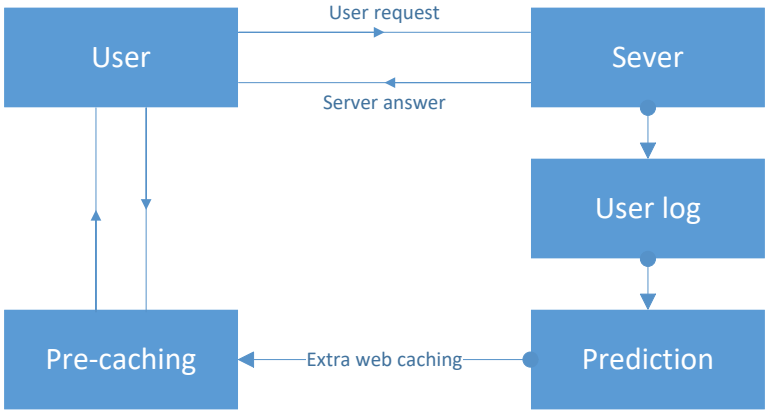


FIGURE 11. PROCESS OF SITE NAVIGATION OPTIMIZATION

4.2. **Optimization of Web Organization.** The application is to optimize the user log mining, site design, including the organization and display of information through the construction site to enhance user satisfaction.

CUI Rong-wei, LIU Yi-qun, ZHANG Min, et al[56]., Carried out user behavior analysis on the search engine, using 7056 real users to obtain the user usage patterns, to improve the search engine optimization, the study found that users prefer to type phrases or short sentences To achieve the query.

Oakes M and Yan Xu[57] put forward the search sequence based on predecessors, when users search the same problem, based on the same search purposes in front of the search statement, the user's search for standardized processing, so as to optimize the search function. Jiann-Cherng Shieh[58] in his article also proposed based on the user log to improve the search engine search capabilities.

Jiang Tingting[59] and others through the study of Wuhan University Library Log, depth excavation OPAC system user's search behavior, through the keyword, query W and search session log analysis, the final user to retrieve the different literature of some behavior.

Li Gang[60], who first digging for the real estate industry, through the real estate website user data analysis, to understand their user behavior preferences and browsing preferences for the site to enhance the user experience made a constructive proposal.

4.3. **Personalized Web and Push Service.** Through the user log mining user usage patterns, in order to achieve personalized website construction and service push service is a major application of log mining.

Personalized recommendation in the commercial field of the most widely used, the largest e-book Web site Amazon collaborative filtering algorithm to provide users with relevant books recommended services for their own development provides a great impetus. The related collaborative filtering algorithm and its optimization algorithm are widely used

in electronic business websites. Wan-Yu, Deng et al[61]. Proposed a collaborative filtering-based optimization algorithm to reduce the system resource consumption, personalized recommendation to enter the public site has laid a good foundation. ZHANG Wan-shan[62], etc. In the personalization recommendation of Web resources, tracking user's behavior data of browsing and retrieving Web resources, proposing a personalized recommendation algorithm based on topic clustering, and realizing the dynamic recommendation of Web resources.

LiGeWei[63] put forward the use of log mining can library website organizational structure and other aspects to improve, at the same time increase the personalized service, discovery of potential readers.

In addition to these areas of application, log mining for the study of user psychology research also has important reference value.

4.4. **Improvement of Ranking Quality for search Engine.** In the search engine, users submit a query request, if you click on the system to return to the results page of a URL, the user generally said that the URL of a recognition, and in most cases contains information related to the poor URL is not concerned about. In 2002, Zhang Dell proposed a method to improve the quality of the result sorting by using the user click record[64]. This method was firstly applied to the Chinese image information retrieval, and then Baeza-Yates extended it to the text information retrieval[65] and achieved good experimental results. Zhang said that this method is MASEL (Matrix Analysis on Search Engine) algorithm.

The MASEL method tries to find the relationship between the user, the query, and the click URL. The basic assumption is that good users submit good queries, good queries return good URLs, and good URLs are clicked by good users. Recursively defining these three basic quantities according to the user click record is very similar to the recursive relationship established between the Authority and the Hub of the page in the Hits algorithm[66].

Baeza-Yates found that the MASEL method has a good effect on the small-scale experiment of the Chilean Todo search engine log, and pointed out that the selection of the log cycle has certain influence on the accuracy of the experimental results, especially for the polysemy words, When a longer period of time to select the log accuracy of the results instead of declining.

In order to sort the results of the meta search engine, Joachims[67] proposed a method that uses click data as the training set and learns to retrieve functions, and the results of the experiments are superior to those of Google's search results.

4.5. **Suggestion of Related Web Query.** As the search engine users to enter the query string is usually relatively short, and short words can express the theme of a broad, easy to ambiguity, and users often can not accurately express their information needs. Therefore, sometimes the user needs to constantly modify their query requests in order to find their own satisfaction with the information. Therefore, query suggestion technology is widely used in major mainstream search engines. For the convenience of users to modify the query, some search engine systems such as Google, Baidu, etc. have been submitted for the first time in the user's request, the results returned in the system page contains a list of related

queries for users to modify the reference, which The user to more accurately express their information needs to provide a convenient.

The traditional information retrieval system uses query expansion to discover related queries. The main methods are: query expansion based on user feedback, query expansion based on part or all information[68]. These methods generally rely on the specific content of each document in the document set, the realization of the more complex in the actual retrieval system is not used much.

It is a practical method to discover related queries based on search engine logs. Based on the contents of the log, the log-based query recommendation method can be divided into three categories: based on the query string, based on the click URL, based on user session. The method based on the query string calculates the query relevance by using the similarity between the input query contents. The query content can include information such as the anchor text, the summary and the like of the corresponding user click result. Based on the click URL method, the same or similar and the user's session based on the number of times that the two queries are co-occurring in the same session, and calculates the correlation degree between the two queries by using the URL as a feature.

According to the technical methods used, log-based query recommendation can be divided into the following categories:

Method one: manually mark some training data, the establishment of regression model to determine the relevance of the size of the relevant web query. Regression analysis methods commonly used are multiple linear regression and support vector regression. In general, the parameters of the linear regression method are simple and can get the prediction results of each query very quickly, but the prediction precision is slightly worse. The SVR method involves more parameter selection, longer training time, Higher. The results show that there are significant differences in the accuracy of the prediction results for different types of queries (information, navigation, and transaction[69]).

Method two: only through the user click log, the query string clustering to find the relevant query. The method first constructs a bipartite graph, which joins the query set with some corresponding elements in the click URL set according to the user's click record, and then uses the agglomerative iterative algorithm to merge the two queries and the two URL until the end of the iteration[70]. One disadvantage of this approach is that it can not effectively deal with noise data, that is, if a user delays a URL, the two irrelevant queries may be forever together[71].

Method 3: Use the association rules to determine the relevant Web query[72]. A query is considered as an item in an association rule, and a session in the query log is treated as a transaction. That is, a set of queries submitted by a single user is considered as a transaction within a certain time interval. Then the association rules mining algorithm is used to find the strong association rules, and then find the related Web queries.

4.6. **Detection of Information Security.** Through the analysis of the user behavior in the user log, it can detect the information security of the network or the system, discover the security loophole in time, and better protect the information security.

The current research results, such as: Pan Lei et al[73] for network monitoring system

needs to address the extraction of user access mode information in the multi-dimensional multi-value association rules, the traditional association rule algorithm has been expanded and improved to extract effective association rules, Reflect the behavior of the user model. DAI Zhen et al. [74]proposed a new user association rule mining algorithm based on Apriori algorithm, which can obtain the maximum frequent itemsets by recursively mining the pattern tree, according to the requirement of user access patterns in intrusion detection system. Zhou Yunxia et al[75] in the database intrusion detection system user behavior mining improved FP-Growth algorithm to improve the efficiency of mining. In Li Yuhua's view[76], neural network, fuzzy system and subtractive clustering are used to monitor intrusion behavior. A fuzzy neural network user behavior analysis system which can effectively monitor the attack behavior is developed.

4.7. **Network Opinion Monitoring.** Search engine is an important access to Internet users' access to information, when users are interested in a social event, it is possible to search the network for the event information. Therefore, the search engine user log records for you to use the query, including the query time, query location, query results and click results, to some extent reflects the purpose of the user's query. Analysis of user logs, users can find the content and time of frequent inquiries, and then find public opinion hot spots, according to public opinion and the source of change, reflecting the more real-time and comprehensive public opinion information. Li Lei[77] and others micro-blog users will be divided into general concern, active participation and information dissemination of the three types, based on a micro-blog theme of user clustering algorithm for network public opinion detection.

At present, the major search engine companies to strengthen research and development through query log analysis of network public opinion monitoring applications for government decision-making and the relevant management departments to provide an important reference. Such as the two typical applications: Google Trend and Baidu Index.

1) Google Trends. Google Trends is a Google company launched an analysis of users in Google search keywords and display the attention of the keyword service. Users can visually see the trend of each keyword in the Google global search volume index and related news citation changes, and can be a number of different search terms to compare the search behavior.

2) Baidu Index. Baidu Index and Google Trends similar to Baidu's official launch in 2006 to Baidu Web search and Baidu news search-based free mass data analysis services to reflect the different keywords in the past period of time the "user attention" and "Media attention", can vividly reflect the trend of different keywords every day, directly and objectively reflect the social hot spots and Internet users interest.

5. **Conclusion.** This paper mainly introduces the main statistical indexes of Internet user log analysis, and summarizes the main processes of log analysis. The six log analysis methods are introduced: statistical analysis method, clustering analysis method, association analysis method, decision tree method, neural network method and time-series data mining method. According to different purposes, different methods will be used to explore

different user behavior; then introduction of the current representative open source log collection and analysis tools, which have some shortcomings in function optimization. Finally, this paper explains the practical application of user log analysis, which can be applied to various fields, including website navigation optimization, website design and organization optimization, personalized website and push service, improvement of ranking quality for search engine, suggestion of related Web query, detection of information security and network opinion monitoring.

Network user behavior contains a lot of meaningful information and has many valuable research directions. User data collection, statistical analysis of data, user behavior modeling, analysis behavior model, the mining of network user behavior is of great significance. With internet services become more and more personalized and differentiated, digging deeper into the status of user log and user behavior analysis is increasingly important, it is foreseeable that the related research work and application development will continue to develop.

REFERENCES

[1] China Internet Network Information Center (CNNIC). The 37th Statistical Report on Internet Development in China [EB/OL]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201502/P020150203548852631921.pdf,2015

[2] Srivastava J, Cooley R, Deshpande M, et al. Web usage mining: discovery and applications of usage patterns from Web data [J]. *Acm Sigkdd Explorations Newsletter,* vol.1, no.2, pp.12-23, 2000.

[3] Sun Jianjun, Li Jiang. Theory, Tools and Applications of Network Information Measurement [M]. Science Press, 2009.

[4] Wang Ji-min, Lilei Mingzi, Meng Tao. A Research Framework of Web Search Usage Mining [J]. *Digital Library Forum*, no.8, pp.25-31, 2011.

[5] Wang Ji-min, Chen Chong, Peng Bo. Analysis of the User log for a Large-scale Chinese Search Engine [J].*Journal of South China University of Technology (Natural Science Edition)*, vol.32, no.s1, pp.1-5, 2004.

[6] Xie Y, O'Hallaron D. Locality in search engine queries and its implications for caching [J]. *Proceedings - IEEE INFOCOM*, vol.3, pp.1238-1247, 2001.

[7] Kamvar M, Kellar M, Patel R, et al. Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices[C]// *International Conference on World Wide Web. ACM*, pp. 801-810, 2009.

[8] Spink A. Web Search: Public Searching of the Web [M]. *Springer Netherlands*, 2005.

[9] Baldi P, Frasconi P, Smyth P. Modeling the Internet and the Web: Probabilistic Methods and Algorithms [J]. *Wiley & Sons*, vol.42, no.1, pp.325-326, 2003.

[10] Jansen B J, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the web [J]. *Information Processing & Management an International Journal*, vol.36, no.2, pp.207-227, 2000.

[11] Wang Ji-min, Peng Bo. User Behavior Analysis for a Large-scale Search Engine [J]. *Journal of the China Society for Scientific and Technical Information*, vol.25, no.2, pp.154-162, 2006.

[12] Park S, Lee J H, Bae H J. End user searching: A Web log analysis of NAVER, a Korean Web search engine [J]. *Library & Information Science Research*, vol.27, no.2, pp.203-221, 2005.

[13] Jansen B J, Spink A. How are we searching the World Wide Web? A comparison of nine search engine transaction logs [J]. *Information Processing & Management*, vol.42, no.1, pp.248-263, 2006.

[14] Zhang Yang. Theory and realization of website statistical data collection [EB/OL]. http://blog.codinglabs.org/articles/how-web-analytics-data-collection-system-work.htm

[15] Pabarskaite Z, Raudys A. A process of knowledge discovery from web log data: Systematization and critical review [J]. *Journal of Intelligent Information Systems*, vol.28, no.1, pp.79-104, 2007.

[16] Huberman B A, Pitkow J E, Lukose R M. Strong regularities in World Wide Web surfing[J]. *Science*, vol.280, no.5360, pp.95-97, 1998.

[17] Cooley R, Mobasher B, Srivastava J. Data Preparation for Mining World Wide Web Browsing Patterns [J]. *Knowledge and Information Systems*, vol.1, no.1, pp.5-32, 1999.

[18] Roberts S. Users are still wary of cookies [J]. *Computer Weekly*, 2002.

[19] Spiliopoulou M. Managing Interesting Rules in Sequence Mining[C]// *European Conference on Principles of Data Mining and Knowledge Discovery.* Springer-Verlag, pp.554-560, 1999.

[20] Pitkow J E, Recker M. Integrating Bottom-Up and Top-Down Analysis for Intelligent Hypertext [J]. *National Institute of Standard Technology*, 1994.

[21] Catledge L D, Pitkow J E. Characterizing browsing strategies in the World-Wide web [J]. *Computer Networks & Isdn Systems*, vol.27, no.6, pp.1065-1073, 1995.

[22] Paliouras G, Papatheodorou C, Karkaletsis V, et al. From Web usage statistics to Web usage analysis[C]// *IEEE International Conference on Systems, Man, and Cybernetics, 1999. IEEE Smc '99 Conference Proceedings. IEEE*, vol.2, pp.159-164, 1999.

[23] He, Daqing, Göker, et al. Detecting Session Boundaries from Web User Logs [J]. *Lannée Psychologique*, vol.100, no.4, pp.585-627, 2005.

[24] Zhuang L, Kou Z, Zhang C. Session identification based on time intervals in Web log mining[J]. *Journal of Tsinghua University*, vol.163, pp.389-396, 2005.

[25] Chen M S, Park J S, Yu P S. Data mining for path traversal patterns in a web environment[C]// *International Conference on Distributed Computing Systems. IEEE*, pp.385-392, 1997.

[26] Berendt B, Mobasher B, Nakagawa M, et al. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis[C]// *Webkdd 2002 - Miningweb Data for Discovering Usage Patterns and Profiles, International Workshop, Edmonton, Canada, July 23, 2002, Revised Papers. DBLP*, pp.159-179, 2002.

[27] Malarvizhi S P, Sathiyabhama B. Frequent pagesets from web log by enhanced weighted association rule mining [J]. *Cluster Computing*, vol.19, no.1, pp.269-277, 2016.

[28] Li Y, Feng B, Mao Q. Research on Path Completion Technique in Web Usage Mining[C]// *International Symposium on Computer Science and Computational Technology. IEEE Computer Society*, pp.554-559, 2008.

[29] Munk M, Kapusta J, Švec P, et al. Data advance preparation factors affecting results of sequence rule analysis in web log mining [J]. *E A M Ekonomie A Management*, vol.13, no.4, pp.143-160, 2010.

[30] Five General Methods of Website User Analysis [EB/OL].

http://www.admin5.com/article/20100726/254679.shtml.2014.

[31] Cheng Peng. User Behavior Analysis Platform of Large and Medium Websites [D]. *Fu Dan University*, 2012.

[32] You Ting. The Research on Social-networking Users' Behavior Characteristics and Interior Mechanism——Take Renren.com for Example [D]. *University of Posts and Telecommunications*, 2012.

[33] Deng Xiawei. Use Behavior Analysis Based on Social Network Service——User Behavior Analysis and User Influence Modeling [D].*Beijing Jiaotong University*, 2012.

[34] Oh S.H., Lee W.S., An anomaly intrusion detection method by clustering normal user behavior [J]. *Computers & Security*, vol.22, no.7, pp.596-612, 2003.

[35] Maia M, Almeida J, Almeida, Virg&#. Identifying user behavior in online social networks[C]// *The Workshop on Social Network Systems*. ACM, pp.1-6, 2008.

[36] Sun Yanhua. Analysis of Network Users' Behavior Based on Clustering [D]. *Central South University*, 2011.

[37] Liu Peng. Behavior of Network Users Research Based on Time-varying & Services [D]. *University of Posts and Telecommunications*, 2010.

[38] CAI Yue, YUAN Jin-Sheng. User Activities Clustering of Search Engine [J]. *Computer Systems & Applications*, vol.19, no.4, pp.94-97, 2010.

[39] Chen Min, Miao Duoqian, Duan Qiguo. Clustering Web Users Based' Browsing Action [J]. *Computer Science*, vol.35, no.3, pp.186-187, 2008.

[40] Yan Hao. Network User Behavior Analysis Base on Traffic Monitoring and Measurement [D]. *University of Posts and Telecommunications*, 2011.

[41] Zhang Xia, Ma Yining, Chen Jingru. A Clustering Algorithm Based on User Query Intention [J]. *Computer Knowledge and Technology*, vol.08, no.5X, pp.3388-3390, 2012.

[42] WANG Aiping, WANG Zhan feng, TAO Si-gan, YAN Fei-fei. Common Algorithms of Association Rules Mining in Data Mining [J]. *Computer Technology and Development*, vol.20, no.4, pp.105-108, 2010.

[43] Wang Yongli. A Study on Association Rules Mining Algorithm and Its Application on Web Mining [D]. *Harbin Engineering University*, 2003.

[44] Luo Qiang. The Research on Key Technology of Analysis User Behavior Association in Social Network [D]. *University of Electronic Science and Technology of China*, 2013.

[45] Xu Xiaojuan, Zhao Yuxiang, Zhu Qinghua. Explore User's Behavior of Academic Blog Based on EDTM [J]. *New Technology of Library and Information Service.* no.1, pp.79-86, 2014.

[46] He Lu. User Personality and Behavior Analysis Based on Social Networking Services [D]. *University of Posts and Telecommunications*, 2014.

[47] LI Xian-peng，HE Song-hua，ZHAO Xiao-min，et al. Improved ID3 algorithms applying in forecasting customer's churn[J]. *Computer Engineering and Applications*， Vol.45, no.10, pp.242-244, 2009.

[48] ZOU Jing and XIE Kun, Application of C4.5 Algorithm on Customer Churn Analysis in Mobile Communication Industry, *Computer Technology and Automation*, vol.28, no.3, pp.98-101, 2009.

[49] Liu Rong, Chen Peng and Zhang Xingyan. Research on neural network based adaptive user model in automatic Web personalization, *Electronic Measurement Technology*, vol.30, no.4, pp.165-168, 2007.

[50] Zuo Lin, The Study of Neural Networks with Application in Analyzing of Users' Online Behaviors. *University of Electronic Science and Technology of China*, 2011.

[51] Hutchins R, Zegura E W, Kolesnikov O, et al, Usage Characteristics of Dial-in Internet Users: A National Study, 2001.

[52] Hutchins R, Zegura E W, Liashenko A, et al, Internet user access via dial-up networks-traffic characterization and statistics, *Network Protocols Ninth International Conference on ICNP*, pp.314-322, 2001.

[53] US Department of Commerce, NIST, Guide to Computer Security Log Management, *Nist Sp*, 2006.

[54] Baskaran K R and Kalaiarasan C, Improved Performance by Combining Web Pre-Fetching Using Clustering with Web Caching Based on SVM Learning Method, *International Journal of Computers Communications & Control*, vol.11, no.2, pp.67, 2016.

[55] Nandini N, Yogish H K and Raju G T, Pre-fetching techniques for effective web latency reduction — A survey, *Africon*, pp.1-6, 2013.

[56] CEN Rongwei, LIU Yiqun, ZHANG Min, et al, Search Engine User Behavior Analysis Based on Log Mining, *Journal of Chinese Information Processing*, vol.24, no.3, pp.49-54, 2010.

[57] Oakes M and Xu Y, A search engine based on query logs, and search log analysis at the university of Sunderland, *Water Supply Paper*, 2009.

[58] Shieh and Jiann Cherng. Mining Website Log to Improve Its Findability, *International Conference DBLP*, pp.239-247, 2010.

[59] Jiang Tingting, Wang Miao, Gao Hui, et al, A Search Log Analysis of OPAC Users' Searching Behavior —— A Case Study of Wuhan University Library, *Documentation, Information & Knowledge*, pp.46-56, 2015.

[60] Li Gang, Li Chunya, Hu Rong, Hai Lan, Research and Application of Web Mining Based on Association Rule, *Journal of Information Resources Management*, pp.28-36, 2015.

[61] Deng W Y, Zheng Q H, Lian S, et al, Adaptive personalized recommendation based on adaptive learning, *Neurocomputing*, vol.74, no.11, pp.1848-1858, 2011.

[62] Zhang Wan-shan, Xiao Yao, Liang Jun-jie and Yu Dun-hui, Personalized Recommendation of Web Resources Based on Topic Clustering, *Microelectronics & Computer*, pp.35-39, 2015.

[63] Li Gewei, Data Preparation in Web Log Mining and Individual Service of Digital Library, *Journal of Intelligence*, vol.26, no.8, pp.90-91, 2007.

[64] Zhang D and Dong Y, A novel Web usage mining approach for search engines, *Computer Networks*, vol.39, no.3, pp.303-310, 2002.

[65] Baeza-Yates R. Query usage mining in search engines, *Web Mining: Applications & Techniques*, *Anthony Scime, Editor Idea Group*. IGI Global, pp.307—321, 2004.

[66] Chakrabarti S, Mining the Web: Discovering Knowledge from HyperText Data, *Science & Technology Books*, pp.275–276, 2003.

[67] Joachims T, Optimizing search engines using clickthrough data, *Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM*, pp.133-142, 2003.

[68] Cui H, Wen J R, Nie J Y, et al. Query Expansion by Mining User Logs, *IEEE Transactions on Knowledge & Data Engineering*, vol.15, no.4, pp.829-839, 2003.

[69] Kang I H and Kim G C. Query type classification for web document retrieval, *International ACM SIGIR Conference on Research and Development in Informaion Retrieval. ACM*, pp64-71, 2004.

[70] Beeferman D and Berger A, Agglomerative clustering of a search engine query log, *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM*, pp.407—416, 2000.

[71] Chan W S, Leung W T and Lee D L, Clustering search engine query log containing noisy clickthroughs, *International Symposium on Applications and the Internet, 2004. Proceedings*, pp.305-308, 2004.

[72] Fonseca, B. M, et al, Using association rules to discover search engines related queries, *Web Congress, 2003. Proceedings. First Latin American IEEE*, pp.66-71, 2003.

[73] Pan Lei, Su Jin and Xu Tingrong, Research on Mining Multi-Dimension Association Rules About Network Accessing Behavior, *Computer Applications and Software*, vol.25, no.3, pp.189-191, 2008.

[74] Dai Zhen, Fei Hongxiao, Xie Wenbin and Xiaoxinhua, The Algorithm of users behavior associate rules mining Based on Specific Pattern Tree, *Computer Systems & Applications*, vol.16, no.5, pp.56-59, 2007.

[75] Zhou Yunxia and Li Lei. An Improved FP-Growth Algorithm Based on Database User Behavior, *Science Technology and Engineering*, vol.18, no.18, pp.4380-4383.2011.

[76] Li Yuhua, Fuzzy Neural Network Based User Behavior Analysis, *Huazhong University of Science and Technology*, 2013.

[77] Li Lei and Liu Ji, Empirical Analysis of Micro-blogger Behavior Clustering on Public Opinion Topics, *Journal of Intelligence*, pp.118-121, 2014.