On Data Warehouse-Based Campus Network User Behavior Analysis of the Decision System

Xiaobing Xian, Jinyi Chang , Feng Chen

School of Computer Science and Engineering, Changshu Institute of Technology, Changshu 215500, P.R.China

*Corresponding Author: xianxb@cslg.edu.cn

Abstract

This paper tends to give an insight into campus network user behaviors with the aid of research tools peculiar to natural science such as data warehouse and data mining in combination of principles for human and social science, namely, psychology and sociology. What's more, the author tries to base webpage classification criteria on ODP with a view to exploring specific patterns and rules found in network user behaviors. On the basis of those pioneering studies, he's put forward inspiration, procedures and framework concerning the campus network user behavior analysis and decision system, and kernel technologies may face in this process as well.

Keywords: data warehouse; OLAP; ODP; RDF; network behavior analysis

1. Overview

Surfing online has increasingly become an essential lifestyle for most people and thus related studies targeting at network behaviors aroused international concern. It is no surprise to research institutions and learned societies to find vast stores of relevant studies has been in full swing, among which The Cooperative Association for Internet Data Analysis and The IP Provide Metrics Affiliated to International Engineering Task Force are noteworthy. The essence of these studies was centralized in performance, strategies and influencing factors of varieties of network user behaviors as well as specific patterns found.

A much greater insight into behavior characteristics of campus network group users, including network behavior analysis of large quantities, incomplete, noisy, vague and random network data as well as a full-scale probe into potential, hidden user information, behavior pattern and

trend via data mining, and further improvement of respective behavior analysis system of campus network group users will definitely help give a comprehensive and rational proof of decision-making for school administrators. This makes it possible to effectively take precautions against undesirable network behaviors of college students, predict and correct in case of any misconduct and raise students' self-consciousness of network behaviors. While creating favorable conditions for students' development, this will exert a profound influence on the education, management and training of universities and colleges.

Therefore, we choose to take specific behaviors of campus network users as the subject of this study, which form essential theoretical basis for many related researches such as student management, design and construction of campus network and network management.

2. Basic concept

2.1 Group Network Behavior

Group network behavior refers to specific group behaviors demonstrated on the net, which has gotten growing attention owing to its distinctive features. Aiming at shedding light on respective group behaviors from different angles and obtaining overall feature by analyzing single group behavior, it is necessary to classify groups rationally under different circumstances.

Behavior analysis of network users, fallen into the category of network information retrieval, tends to study their components, characteristics and rules found in network practice as well through multi-disciplinary approaches. Technically speaking, crossing the boundaries of various disciplines such as philosophy, economics, sociology, psychology and computer science, network behavior analysis has been at the cutting-edge of social

science studies in the information era and network virtual space on which it rely has unquestionably become one brand-new social form.

2.2 Open Directory Project and Resource Description Framework

Open Directory Project (ODP), also termed as dmoz, is one of the largest artificial classified retrieval systems online at the present time. Its initial purpose is to bridge the gap between the demand of utilizing resources in the widest, fastest and most common way and the challenge of tackling rapidly increasing amount of network information. ODP offers a practical solution for most search engines such as Google, Netscape, Dogpile, thunderstone and Linux as well. Keeping up to date with an increasingly expanding pool of network information, it enables network users to experience themselves the thrill of drawing up and managing the directory and harness widely distributed information in a diverse and growing Internet environment to update database records and directory as well by setting up and further improving a complete set of management system and index normalization.

Resource Description Framework (RDF), which was proposed by The World Wide Web Consortium, is one type of mark-up languages applied for describing resources on the web by symbolizing them only as Uniform Resource Identifier. It is used as the criterion of accessing metadata collected from the resources on the web and also the standard protocol for reflecting specific resources.

In light of their own advantages and long-term practice, we choose open directory project as the criterion for webpage classification and have tried to process RDF-based ODP classified data.

Fig.1 RDF-based ODP Classified Structure

3. System Design & Application

We attempts to shed light on general appearance of

user behavior patterns concerning the group and individual campus network users as well as behavior differences between diversified groups by acquiring source data from backbone data flow on the campus network and studying group user behaviors such as the duration and preference of net use.

The analysis of group user behaviors, vertical and horizontal, mentioned here mainly covers three aspects: time of using campus network, time and trends of employing network services and types of services provided, and the ratio of the number of students employing typical services and flow ratio, trends and types of services provided in the same period. To put it more concretely, vertical analysis focuses on the network use behaviors of one specific group. Emphasis is placed upon the network use in different periods per day, per week, weekday or weekend and even much longer period. By comparing the group user behaviors in the same period and different period, we tend to find out the relationship between network services and time of network use. As for horizontal analysis, we attempt to make a comparison of network use behaviors between different groups based on the previous vertical analysis for each group.

Fig.2 Architecture of the Decision System for Campus Network User Behavior Analysis

As is clearly shown in the architecture above, this system is mainly made up of four modules. The first module focuses on the acquisition of source data, including ODP, basic information and IP information of registered students, router mirror data, domain name region and the department or college concerned. The second is responsible for processing those source data the first module acquired, that is, ruling out noise and redundant data, with the purpose of building a normalized data warehouse. The third module is in charge of arranging and dealing with the data warehouse mentioned for multi-dimensional data after SSAS analysis. The last one designs to demonstrate the multi-dimensional data obtained with the aid of the present

client software and meanwhile browse them through the web interface developed by ASP.NET.

(1)Data Acquisition tends to acquire source data flow of campus network users, namely, to acquire, receive, restore and tackle the clean data.

(2)Data Pre-processing aims to preprocess network user behaviors in correspondence with the requirements of OLAP and data mining to obtain target data suitable for analysis, process RDF-based ODP classified data, design and set up data in various dimensions for analysis and then build data warehouse concerning campus network user behaviors.

(3)Behavior Analysis is, on the basis of data warehouse concerning campus network user behaviors, to set up target data and online analyze user behaviors aiming at yielding data unveiling their behavior characteristics.

(4)Data Mining attempts to analyze definite data property and transfer source data to valuable information.

3.1 Source Data Acquisition

With the aid of SAM Identification System, HR Management System, System of Education Administration and Service and related information system, we tend to acquire network data and user information and then explore the relationship between IP address and user information.

Data flow arising from network group users paves the way for behavior analysis. We adopt the techniques of real-time data acquisition and port mirroring for key network facility (e.g. Cisco 9509) for the purpose of acquiring dynamic network data and then capturing IP packet. As a result, preprocessing vast stores of data with a view to ruling out unnecessary data and obtaining those the system needs has become an unavoidable task for us.

student	country	hostip	time	ipadress	hosturl
ALL	ALL	ALL	ALL	ALL	ALL
Sname	Name	Country	Year	Sip	Host

Fig. 3 Table for Classification Dimension

3.2 To construct multi-dimensional data and hierarchical organization.

In addition to classification and statistical analysis of network behaviors of individual users and group users by ODP approach, we attempt to classify and identify user groups based on their IP address and user features in combination of basic information of network users such as name, discipline, major, age, gender, class and grade.

Furthermore, we are enable to take a comprehensive analysis on network group user behaviors from a vertical or horizontal viewpoint, including website type, time of network use, network flow, network service and user preference for websites. Based on these, several technological dimensions, namely, the dimension of ODP classification, user classification, time classification, IP address classification and classification of country of origin according to IP address, have been established and the multi-dimensional data for network behavior analysis have been finally constructed.

3.3 To set up data warehouse for campus network user behaviors.

Based on established multi-dimensional data, we tend to analyze online network user behaviors and launch an in-depth probe into data, that is, store and access rapidly, consistently and interactively the information transferred from ready-to-process ones, much easier to understand and true to user group behavior characteristics from different angles. OLAP will help satisfy needs of inquiry and report for decision support or specific multi-dimensional environment, better demonstrate the results of data analysis and produce valid data of user group behavior characteristics as well.

3.4 Results analysis

Data resulting from Internet use are found in a massive and unstructured state and out of order. By means of data mining, we are capable of obtaining potential, unknown and useful information, pattern and trend among massive, incomplete, noisy, vague and random data. Furthermore, we strives to analyze network user behavior characteristics with a follow-up in depth study of individual and group characteristics of campus network users and then predict the behaviors and mental state of student users, which will definitely provide a clue to network use behaviors and psychological features of college students as a whole for college administrators.

Fig.4 network access rate

As can be seen from above graph, website traffic after class is apparently much heavier than that during regular class periods and meanwhile students in school of foreign languages may seem share stronger preference for surfing the web than those in other departments after class. It might be surprising to notice a general tendency of Internet addiction in students of school of chemistry and materials engineering during regular class periods. Another noteworthy fact is that the vast majority of students of school of computer science and engineering and school of arts and garments engineering also show great enthusiasm for Internet surfing on campus after class.

Based on the statistical analysis aiming at data of three months we've acquired in a row, the suggestions for education administration and students management can be listed as follows. To begin with, office of academic affairs is expected to encourage students to learn to be actively involved in a variety of extra-curricular activities rather than indulge in Internet. What's more, members of teaching staff are also supposed to check for students' daily attendance on a regular basis in case of any skipping behaviors. As for the website type students interested, those books or references concerning world are strongly recommended to be on the priority list of book purchase for the school library. Last but not the least, it can be inferred that students used to hang out on weekends in terms of remarkable decrease in the flow of data.

4. Conclusions

By means of research tools peculiar to natural science combined with principles for human and social science, decision system for campus network user behavior analysis offers sufficient, reliable and visible data, laying a sound foundation in the decision of education and administration for teachers of universities and colleges. No doubt this will contribute to greater efficiency and cost effectiveness of decision-making in school administration.

Of course some problems are likely to remain unsolved, for instance, a mismatch between address in the present ODP data and network address in web log, inaccuracy of website classification and low update frequency of classified information table owing to absence of ODP approval. We've prepared ourselves for the attempt to analyze campus network user behaviors in a more accurate, effective and scientific way through other classification criteria.

5. Reference

- [1]<http://news.xinhuanet.com/it/2004-04/01/content/39658g.htm> China Internet Network Information Center. Investigation Report of Internet Information Resources. 2004. <http://news.xinhuanet.com/it/2004-04/01/content/39658g.htm>
- [2]Fuqiang Dong. Study on Network Users' Behaviors Analysis and Its Application. Computer Software and Theory. 2005 01 01
- [3]Guang Cheng ,Jian Gong ,Wei Ding. Research on Network Measurement & Behavior. Computer Engineering and Applications. 2004 (27)6:1-5.
- [4]Shaohua HUANG ,Yupeng WU. On the Status of Internet Behavior. Journal of Lanzhou University (Social Sciences). 2007, Vol. 35 No. 2: 32-40
- [5]Keim Daniel, Andrienko Gennady, Fekete Jean-Daniel et al. Visual analytics: Definition, process, and challenges // Kerren A. Information Visualization. LNCS 4950. Berlin: Springer, 2008: 154-175
- [6]Card S K, Mackinlay J D, Shneiderman B. Readings in Information Visualization: Using Vision to Think 1st ed. San Francisco: Morgan Kaufmann, 1999
- [7]Dongxing Teng, Zilu Wang, Haiyan Yang, Hong'an Wang, Guozhong DAI. Research on Decision Analysis System Based on Interactive Visual Components. Chinese Journal of Computers. 2011, Vol. 34 No. 3: 555-565
- [8]Dongya Wang. Research of Campus Network Behavior Analysis. Beijing: Beijing Jiaotong University. 2007
- [9]Mingsheng HE. The Philosophic Sense of Net-Behavior. Studies In Dialectics of Nature. 2000, Vol. 16 No. 11: 56-68
- [10]Thomas James J, Cook Kristin A. Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press, 2005
- [11]Zhe Sui. Research and Implication of Group Network Behaviors Model. Sichuan: Sichuan University. 2006
- [12]Shen Zhong Lu, Kai Yu. Organizational Behavior-Theories and Practice. Zhejiang: Zhe Jiang Education Publishing House. 1993. 8-12
- [13]GOMS Models: An Approach to Rapid Usability Evaluation <http://www.cecs.unvch.edu/kieras/goms>

2011 Jiangsu Social Science Fund (11SHB004)