IACSIT International Journal of Engineering and Technology, Vol.2, No.5, October 2010
ISSN: 1793-8236

Mining Web Usage within a Local Area Network

Hsin-Chih Lin, *Senior Member, IACSIT* and Mei-San Choi

***Abstract*—In this study, we propose a novel method of mining web usage within a local area network (LAN). The proposed method makes use of Taiwanese Yahoo! (i.e. Yahoo! Kimo) to create a topic hierarchy and classifies the most visited websites for users within a LAN into the hierarchy. Each of the websites will have a unique category path. According to the category path, web usage statistics can be accumulated from the bottommost level, i.e. a single website, to the topmost level, i.e. the set of all websites. Thus, web surfing demands of users within the LAN can be analyzed at various levels of detail. Experimental results strongly support the effectiveness of the proposed method.**

***Index Terms*—Web usage mining, Yahoo! Directory, Yahoo! Search.**

I. Introduction

Once Tim Berners-Lee of the Conseil European Pour Recherches Nucleaires (CERN) announced the World Wide Web (WWW) in 1991, the amount of web pages is growing exponentially. A new information world that exceeds the limit of time and space forms gradually as a result of the World Wide Web. According to the report of R. H. Zakon [1] at July 2009, the amount of web hosts on the Internet have reached to about 250,000,000. Each of the web hosts may have several websites, thus exhibiting an abundance of and a variety of information. As a result, retrieving useful information quickly and effectively from WWW has become a key lesson in the information world. To provide more precise and relevant information, Internet service providers (ISPs) are required not only to provide web surfers with effective navigation services, but also to develop powerful tools for gathering intelligence from web surfing activities.

Web navigation services are generally classified into two types, including search engines and web directories. A search engine allows users to pose a keyword (or a phrase) and then replies a list of relevant websites sorted according to their similarities to the keyword. A web directory is a tree or hierarchical structure of web topic categorization. The homepage of the web directory may consist of several main topic categories. Each main category is then separated into smaller topic categories called sub-categories. This tree structure extends until each website belongs a category of the website itself. Undoubtedly, Google has gained the lion's share of search engine market in the world. And according to the study of Baeza-Yates and Ribeiro-Neto [2], the other largest search engines on the Internet are AltaVista [3], HotBot [4], Northern Light [5], Excite [6], and so on; and the largest web directories are Yahoo! [7], LookSmart [8], eBLAST [9], NewHoo [10], and so on. Generally speaking, most of portals that provide a web directory also provide a search engine. As a user pose a keyword (or a phrase), the search engine will reply a title, a description, and the URL of each relevant website. Besides, the search engine will also reply the category path of each website. Accordingly, the user can understand the categorization of website topics and the characteristics of each website. Although both search engines and web directories can improve the efficiency of web surfing, it is not facile sometimes for users to find a target website from such a long list of websites. Moreover, most of web navigation services mainly support users from the Internet or users speaking the same language. These services have not considered access patterns and special interests of users within a local area network (LAN).

E-commerce (EC) has become a necessary and a valuable element of doing business. Competitions among businesses or ISPs are increasing rapidly; search engines and web directories are no more capable of satisfying information demands required by various professionals. For example, end users may need a prompt and effective guide for searching more relevant websites, and for extracting more implicit but more important intelligence. ISPs will have an urgent demand to understand usage patterns of their customers, so as to personalize their services and perform mass customization. The understanding of usage patterns can also be used as a reference to design web pages and contents. Furthermore, it can be used in reducing the load of networks and enhancing the quality of services. Business analysts are seeking for tools to predict special demands of their customers, and to automatically perform knowledge management and decision making. The above-mentioned tools can help business analysts to gain competitive advantages for their organizations. Because of the EC trend, web mining is becoming one of the most popular and important research topics in these years.

In 1996, Etzioni [11] issued the term “Web Mining,” which means the use of data mining techniques to automatically discover and extract useful information from web documents and services [12-13]. According to Kosala and Blockeel [13], web mining can be categorized into three areas, including web content mining, web structure mining, and web usage mining. Web content mining describes the automatic discovery of useful information from web contents, data (including texts and multimedia objects), and documents. Web structure mining, based on the topology of hyperlinks, discovers the link structure within a website or among websites. Web usage mining describes the utilization of data

Manuscript received August 11, 2010.

Hsin-Chih Lin is with the Department of Information and Learning Technology, National University of Tainan, Tainan, Taiwan 70005. (E-mail: hclin@mail.nutn.edu.tw).

Mei-San Choi received her bachelor’s degree in Information Management in 2004 from Chang Jung Christian University, Tainan, Taiwan 711.

mining techniques to discover and analyze access patterns and special interests of users while interacting with the website. Web usage mining can improve the design or management of websites as well as web contents and the way of interaction. Generally speaking, web content and structure mining utilize the primary data of websites, whereas web usage mining makes use of logs resulted from interactions of users with websites. These logs are secondary data, which may include web server logs, proxy server logs, browser logs, user profiles, and so on. The analysis of web server logs can be used to interpret web surfing activities of users who visit the same website from any hosts on the Internet. Proxy server logs can be analyzed and used to understand web surfing activities of users within a LAN or those who are using the same ISP. The analysis of browser logs can be utilized to understand web surfing behaviors of a single user.

Web usage mining analyzes logs of web surfing activities and interprets usage patterns and special interests of users. In the information age, we believe in “Customers are Paramount” and “Knowledge is Power”. Web usage mining is becoming more important for academic researchers and business analysts. Detailed surveys on web usage mining can be referred to [12-16]. Web mining techniques, which are derived from the traditional techniques of data mining, can be separated into categories as follows: statistical analysis [14], association rules [17-21], clustering [22], classification [23-27], similar sequences [28-29], sequential patterns [30-32], dependency modeling, etc. Web mining applications may include personalization or mass customization [33], marketing intelligence discovery [34], intelligent agents [35], adaptive web sites [36], and so on.

In this study, we propose a novel method of mining web usage within a LAN. The proposed method makes use of Taiwanese Yahoo! (i.e. Yahoo! Kimo [37]) to create a topic hierarchy and classifies the most visited websites for users within a LAN into the hierarchy. Each of the websites will have a unique category path. According to the path, web usage statistics can be accumulated from the bottommost level, i.e. a single website, to the topmost level, i.e. the set of all websites. Thus, web surfing demands of users within the LAN can be analyzed at various levels of detail.

The proposed method consists of three phases, including data cleaning, hostname lookup, and statistics computation. In the data-cleaning phase, our method is responsible to check every request sent to the proxy server and then removes unavailable requests, so as to improve the effectiveness of the method. In the hostname-lookup phase, our robot program carries the hostname of each available request to Yahoo! Search and fetches the title as well as the category path of the requested website. In the statistics-computation phase, our system computes the web usage statistics from the bottommost level category to the topmost level category. Accordingly, we are able to know the most visited websites under a certain category, as well as the number of times visited and the sub-category of each website. We can also view the most visited sub-categories under a certain category, and the number of times that each sub-category was visited. Experimental results strongly support the effectiveness of the proposed method.

The organization of this paper is as follows. In Section 2, we introduce the proxy server and the format of a log file. In Section 3, we introduce the search engine and the web directory of Yahoo! Kimo. In Section 4, we describe the three phases of the proposed method. In Section 5, we demonstrate parts of achievements of the proposed method. Finally, we conclude this study and discuss further research.

II. PROXY SERVER

A proxy server, also called a cache server, is usually located between servers on the WWW and end-user browsers within a LAN. The proxy server is used to store web pages that have been downloaded through it in advance. When a browser requests to download a web page, the proxy server checks if the web page is already stored. If yes, the proxy server replies the web page to the browser; otherwise, the request is passed to the web server. After that, the web page is downloaded, stored at the proxy server, and sent to the browser. The goal of a proxy server is to improve the efficiency in fetching web pages and to filter web pages according to their contents.

When a browser fetches a web page from a proxy server, each of the requested information will be stored in a text file “access.log,” as shown in Fig. 1. The native format of each request is as follows.

Timestamp Elapsed Client Action/Code Size Method URI Ident Hierarchy/From Content

Each of the fields in the format is described as follows:

1) Timestamp: The time when the request is completed (socket closed). The format is “Unix time” (seconds since Jan. 1, 1970) with millisecond resolution.
2) Elapsed: The elapsed time of the request, in milliseconds. This is the time between accept() and close() of the client socket.
3) Client: The IP address of the connecting client.
4) Action: This field describes how the request was treated locally (hit, miss, etc).
5) Code: The HTTP reply code taken from the first line of the HTTP reply header. For ICP requests this is always “000.” If the reply code was not given, it will be logged as “555.”
6) Size: For TCP requests, the amount of data written to the client. For UDP requests, it is the size (in bytes) of the request.
7) Method: The HTTP request method (GET, POST, etc), or ICP_QUERY for ICP requests.
8) URI: The requested Uniform Resource Identifier.
9) Ident: The result of the RFC931/ident lookup of the client username. If RFC931/ident lookup is disabled (default: “ident_lookup off”), it is logged as - .
10) Hierarchy: A description of how and where the requested object was fetched.
11) From: The hostname of the machine where we got the object.
12) Content: The content-type of the object (from the HTTP reply header).

```
989830333.057  74872 r311h4.dorm1.cju.edu.tw TCP_MISS/000 0 POST http://sinamail.sina.com.tw/cgi-bin/mail/login.cgi - DIRECT/sinamail.sina.com.tw -
989830333.557  74545 r311h4.dorm1.cju.edu.tw TCP_MISS/000 0 POST http://sinamail.sina.com.tw/cgi-bin/mail/login.cgi - DIRECT/sinamail.sina.com.tw -
989830334.187   9302 210.70.156.144 TCP_MISS/304 103 GET http://www.kimo.com.tw/wmadcmd_www.js - DIRECT/www.kimo.com.tw -
989830348.971  30391 210.70.147.203 ERR_CLIENT_ABORT/000 0 GET http://zuojun.hypermart.net/ - TIMEOUT_DIRECT/zuojun.hypermart.net -
989830359.401   3073 210.70.147.203 TCP_MISS/200 304 GET http://zuojun.hypermart.net/images/top-bg.gif - DIRECT/zuojun.hypermart.net image/gif
989830370.112    193 r532h4.dorm2.cju.edu.tw TCP_MISS/304 103 GET http://www.kimo.com.tw/ - DIRECT/www.kimo.com.tw -
989830371.763    346 210.70.160.127 TCP_MISS/200 1677 GET http://ms1.mail2000.com.tw/adv/images/1000.gif - DIRECT/ms1.mail2000.com.tw image/gif
989830372.096    678 210.70.160.127 TCP_MISS/200 4511 GET http://www.openfind.com.tw/img/logo.gif - DIRECT/www.openfind.com.tw image/gif
989830372.377    546 210.70.160.127 TCP_MISS/200 3062 GET http://www.openfind.com.tw/images/free12.gif - DIRECT/www.openfind.com.tw image/gif
989830372.450    284 210.70.160.127 TCP_MISS/200 1691 GET http://ms1.mail2000.com.tw/adv/images/1300.gif - DIRECT/ms1.mail2000.com.tw image/gif
989830372.951    279 210.70.160.127 TCP_MISS/200 1648 GET http://ms1.mail2000.com.tw/adv/images/600.gif - DIRECT/ms1.mail2000.com.tw image/gif
989830372.992    326 210.70.160.127 TCP_MISS/200 1669 GET http://ms1.mail2000.com.tw/adv/images/1600.gif - DIRECT/ms1.mail2000.com.tw image/gif
989830373.650    235 210.70.160.127 TCP_MISS/304 143 GET http://cia.openfind.com.tw/image/cia0.gif - DIRECT/cia.openfind.com.tw -
989830373.747    145 r532h4.dorm2.cju.edu.tw TCP_MISS/200 513 GET http://www.kimo.com.tw/b5.css - DIRECT/www.kimo.com.tw text/plain
989830375.841    453 r532h4.dorm2.cju.edu.tw TCP_MISS/200 3504 GET http://www.kimo.com.tw/wmadcmd_www.js - DIRECT/www.kimo.com.tw text/plain
```

Fig. 1. An example of “access.log” file.

The “Client” field in the “access.log” file shows the IP address of the connecting client; the “From” field represents the hostname of the host that provides objects; and the “Content” field indicates the content-type of the object. Regarding the proposed method, information of the above-mentioned three fields is useful in identifying user sessions and category paths of the requested websites. In Section 4, we will explain how the proposed method tackles the “access.log” file.

III. Yahoo! Kimo

Yahoo! Kimo is one of the most famous portals in Taiwan. It provides seven types of services, including (1) Web Contents, (2) Transaction Services, (3) Communication Services, (4) Yahoo! Anywhere, (5) Marketing Services, (6) Business Services, and (7) Value-Added Services. Since the proposed method makes use of the web directory (also called Yahoo! Kimo Directory) and the search engine (also called Yahoo! Kimo Search) of Yahoo! Kimo, we discuss these two web navigation services in this section.

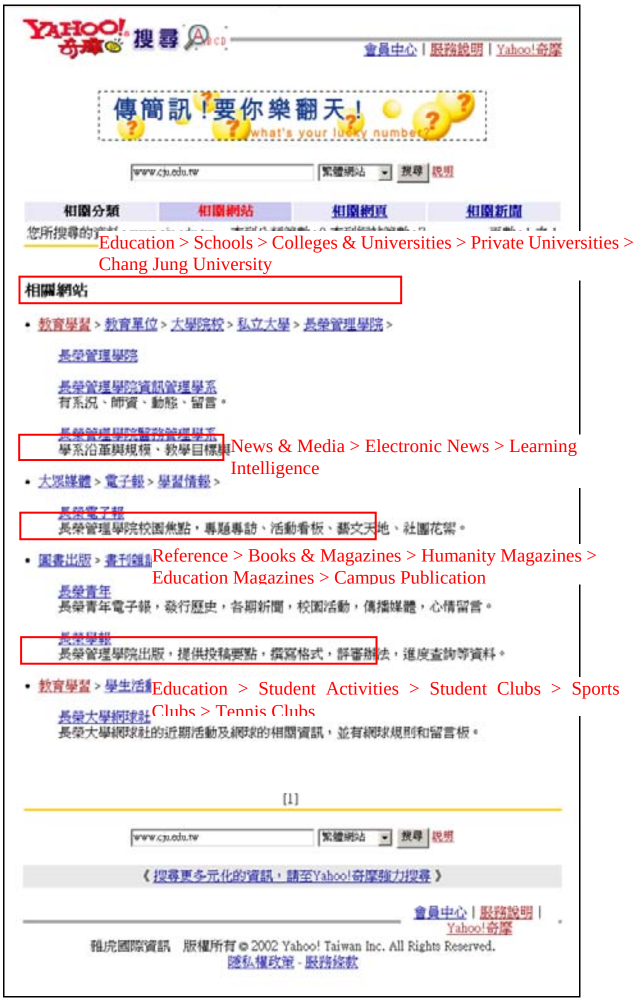
Yahoo! Kimo Directory is an online navigation guide to websites; it has a topic-based tree (or hierarchical) structure that can be browsed. The homepage of Yahoo! Kimo Directory contains links to 16 main topic categories, including (1) Arts & Culture, (2) Entertainment, (3) Sports, (4) Education, (5) Science & Technologies, (6) Regional, (7) Reference, (8) Internet Guides, (9) Life Information, (10) Business & Finance, (11) Computers & Telecommunication, (12) Healthcare, (13) News & Media, (14) Society & Humanities, (15) Government, and (16) Recreation. Each of the main topic categories is separated into smaller topic categories, i.e. their sub-categories. For example, the “Arts & Culture” category is separated into Companies, Artists, “Museums, Galleries & Centers”, and so on. Next, the “Arts & Culture > Companies” category is separated into Handicrafts, Business Design, Recording Companies, and so on. This hierarchical structure extends until each website becomes a category itself. Each website stored in Yahoo! Kimo Directory has its own category path such as “Arts & Culture > Companies > Handicrafts > … ” If a user stays at a certain category, he/she can make use of the search engine, i.e. pose a keyword (or a phrase) and retrieve relevant websites from either “All Categories” or “Just This Category.” Notice that Yahoo! Kimo Directory is created manually. A staff of editors on Yahoo! Kimo visits and evaluates websites daily, and then organizes these websites into topic-based categories. In Yahoo! Kimo Directory, some websites or small categories may reside in more than one category. Although these exceptions do not totally meet the tree (or hierarchical) structure, we consider that Yahoo! Kimo Directory has a tree-based taxonomy for websites.

Yahoo! Kimo Search releases a group of robot (or spider) programs into the WWW for crawling websites. The crawled websites will be compared with the Yahoo! Kimo Directory and organized into topic-based categories. When a user poses a keyword (or a phrase) for making queries, Yahoo! Kimo Search immediately replies a title, a description, and the URL of each relevant website. Also, Yahoo! Kimo Search provides every website with a category path. Although other search engines may support the same function, we found that Yahoo! Kimo Search provides an interesting function that other search engines do not support. If we pose a hostname in Yahoo! Kimo Search, each relevant website together with its category path will be replied. For instance, if we pose “www.cju.edu.tw” (the hostname of URL of Chang Jung University, Taiwan) in Yahoo! Kimo Search, four category paths will be replied, as enclosed by red rectangles in Fig. 2. These four paths are:

1) Education > Schools > Colleges & Universities > Private Universities > Chang Jung University
2) News & Media > Electronic News > Learning Intelligence
3) Reference > Books & Magazines > Humanity Magazines > Education Magazines > Campus Publication
4) Education > Student Activities > Student Clubs > Sports Clubs > Tennis Clubs

Among the four paths, the exact website of CJU is located at the first one. Others are websites in which their URLs may contain “www”, “cju”, “edu”, or “tw”. Besides, these websites are resided in the administration offices, academic departments or student clubs at CJU. According to our survey, we would like to emphasize that Yahoo! Kimo Search is the only search engine in Taiwan that allows users to pose a hostname and replies relevant websites together with their category paths.

To sum up, Yahoo! Kimo Directory can classify websites to topic-based categories. And Yahoo! Kimo Search is able to cooperate with the web directory. Even if the user poses a hostname, Yahoo! Kimo Search can reply each relevant website with a category path. Therefore, we believe that



會員中心 | 服務說明 | Yahoo!奇摩

傳簡訊！要你樂翻天！ what's your lucky number?

www.cju.edu.tw 繁體網站 搜尋 說明

相關分類 相關網站 相關網頁 相關新聞

您所搜尋的資

Education > Schools > Colleges & Universities > Private Universities > Chang Jung University

相關網站

- 教育學習 > 教育單位 > 大學院校 > 私立大學 > 長榮管理學院 >

  長榮管理學院

  長榮管理學院資訊管理學系
  有系況、師資、動態、留言。

  長榮管理學院醫務管理學系
  學系沿革與規模、教學目標與

News & Media > Electronic News > Learning Intelligence

- 大眾媒體 > 電子報 > 學習情報 >

  長榮電子報
  長榮管理學院校園焦點，專題專訪、活動看板、藝文天地、社團花絮。

- 圖書出版 > 書刊雜誌

Reference > Books & Magazines > Humanity Magazines > Education Magazines > Campus Publication

  長榮青年
  長榮青年電子報，發行歷史，各期新聞，校園活動，傳播媒體，心情留言。

  長榮學報
  長榮管理學院出版，提供投稿要點，撰寫格式，評審辦法，進度查詢等資料。

- 教育學習 > 學生活動

Education > Student Activities > Student Clubs > Sports Clubs > Tennis Clubs

  長榮大學網球社
  長榮大學網球社的近期活動及網球的相關資訊，並有網球規則和留言板。

[1]

www.cju.edu.tw 繁體網站 搜尋 說明

《搜尋更多元化的資訊，請至Yahoo!奇摩強力搜尋》

會員中心 | 服務說明 | Yahoo!奇摩

雅虎國際資訊 版權所有 © 2002 Yahoo! Taiwan Inc. All Rights Reserved.
隱私權政策 - 服務條款

Fig. 2: Results of posing “www.cju.edu.tw” in Yahoo! Kimo Search.

Yahoo! Kimo Directory possesses great reference value for academic researches. The hostname lookup function in Yahoo! Kimo Search inspires us to develop the proposed method. In Section 4, we will explain how the proposed method makes use of Yahoo! Kimo.

IV. Proposed Method

The proposed method is composed of three phases, including data cleaning, hostname lookup, and statistics computation, as shown in Fig. 3. In this section, we illustrate the three phases respectively.

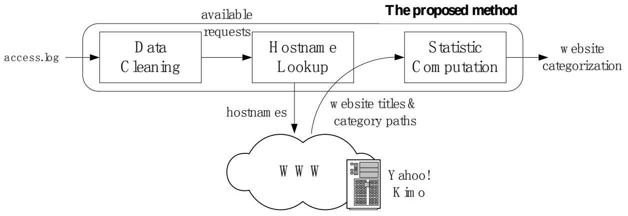


Fig. 3: Diagrams of the proposed method.

In the data-cleaning phase, our method parses the fields of each request in the “access.log” file, and identifies the availability of the request. To improve the effectiveness of the proposed method, we must remove unavailable requests, which may contain invalid URL syntax, unreachable remote sites, non-text data (possibly the attached multimedia objects in a web page), etc. The size of the “access.log” file will be then dramatically reduced. Furthermore, we have made a statistics beforehand of the number of requests sending to the same “From” field, i.e. the field that records the original hostname of the website containing the requested web page. This step is to reduce the load in the next phase and the load of Yahoo! Kimo.

In the hostname-lookup phase, we designed a robot program which carries the “From” field, i.e. the hostname, of each available request to Yahoo! Kimo and carries back the “optimal” title as well as the category path of the website. The above-mentioned task is similar to the previous one described in Section 3, as shown in Fig. 2. That is, when a user poses the hostname “www.cju.edu.tw” in Yahoo! Kimo manually, four category paths are shown on the replied web page. Instead of doing the task manually, our robot program performs each available request automatically. Besides, what the robot program carried back is definitely a web page, as shown in Fig. 2, and there may be more than one website and category path. We have designed a heuristic matching method which allows the robot program to select the “optimal” website and category path automatically as follows: We compare the URLs of all the websites with the “From” field of the request. The website that has a high degree of similarity will be the “optimal” website, and the category path that the website belongs to will be the “optimal” path.

The robot program carries the hostname of a website to Yahoo! Kimo and carries back an “optimal" website and an “optimal” category path. In the statistics-computation phase, we can build a topic hierarchy of web categorization according to the category path gradually. Also, from the bottommost level category, i.e. the website, to the topmost level category, i.e. the set of all websites, along the category path, we can add up the number of times that the visited websites. As a result, we will know the number of times that each category or website of the topic hierarchy was visited. In addition, we would like to emphasize that although related statistical methods of counting the number of times that each website was visited have been proposed in some of the literatures, the topic hierarchical method in this study is definitely an innovative idea of statistical analysis to mine web usage at various levels of detail.

V. Experimental Results

The effectiveness of the proposed method is demonstrated with examples in this section. The proposed method was implemented on a PC (CPU: AMD Athlon 1G, RAM: 512MB, and OS: Microsoft Windows 2000, DBMS: Microsoft SQL 2000). The data-cleaning program was written in C language; the hostname-lookup program, i.e. the robot program, was written in VB language; and the statistics-computation program was written in ASP language. The user interface program was written in VB language; its appearance was further designed by using Ulead PhotoImpact and Adobe Flash. The “access.log” file was obtained from the Computer Center at CJU; it recorded one-year access information from May 2001 and its size was 484MB. After the data-cleaning phase was performed, its size shrank to 47MB and contained 3,234,517 available requests to 18,938 hosts. In the hostname lookup phase, we made use of two PCs to tackle the "access.log" file in parallel to improve the efficiency of the proposed method. The time required to perform all the three phases was about 12 hours.

Fig. 4 shows the user interface of the system. There are totally 16 buttons displayed at the top and bottom of the user interface. These 16 buttons indicate the 16 main topic categories of Yahoo! Kimo Directory. A user can click on the button to browse a topic category or search for rankings of websites under that specific category. There are two buttons in the middle of the user interface, indicating "Websites Billboard" (left) and "Category Billboard" (right). In what follows, we demonstrate the function and characteristics of the two billboards with examples.

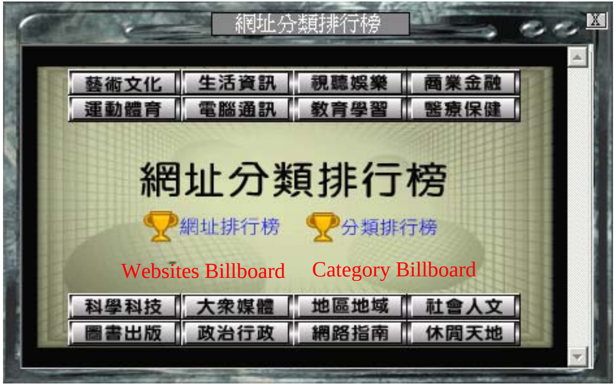


Fig. 4. User interface of the proposed method.

When a user clicks on the "Website Billboard" button, the most visited websites under all topic categories of Yahoo! Kimo, as well as the number of times visited and the category of each website are displayed, as listed in Table 1. For the sake of conciseness, only the top 10 visited websites are displayed in this table. Notice that by using a robot program to search websites on Yahoo! Kimo, the proposed system is capable to retrieve the category of each website. Table 1 shows that "Yahoo! Kimo", "Chang Jung University", "PChome Online", "Sina Taipei", "Taiwan.com", and so on are the most visited websites for CJU users. This table may imply:

TABLE 1. Under all Categories of Yahoo! Kimo, the Most Visited Websites for CJU Users, The Number of Times Visited and the Categories they Belong to.

Website	Number of Visits	Category
Yahoo! Kimo	89,141	Computers & Telecom.
Chang Jung University	51,233	Education
PChome Online	29,563	Computers & Telecom.
SINA Taipei	26,511	Computers & Telecom.
Taiwan.com	19,562	Internet Guides
Tennis Club in CJU	14,900	Education
SINA Fortune-Telling	12,949	Recreation
GeoCities	12,246	Internet Guides
PC home Online Chat Room	11,363	Recreation
Department of Information Management in CJU	10,449	Education

1) once CJU users get online, they first visit a portal (such as "Yahoo! Kimo", "PChome Online", "Sina Taipei", and so on), and make use of their web navigation services to find target websites;
2) most of CJU users set a portal or the CJU website as their browser homepage;
3) the CJU website is often visited because of the daily affairs at the university.

Furthermore, when a user clicks on any of the websites displayed on the table, our system starts a new session and downloads that website. When the user clicks on a topic category, the most visited websites under that category for CJU users are listed. Also, the number of times visited and the topic category each website belongs to are displayed. For example, if the user clicks on the "Education" category, the most visited websites under the "Education" category for CJU users, as well as the number of times visited and the category that each website belongs to are displayed, as listed in Table 2. Under the "Education" category, "Chang Jung University", "Tennis Club in CJU", "Department of Information Management in CJU", "WuKoon Digital School", and so on are the most visited websites for CJU users. When the user clicks on the "Schools" category, the most visited websites under the "Education > Schools" category for CJU users, the number of times each website was visited, and the category each website belongs to are displayed, as listed in Table 3.

TABLE 2. Under the "Education" Category, the Most Visited Websites for CJU Users, the Number of Times Visited, and the Categories They Belong to.

Website	Number of Visits	Category
Chang Jung University	51,233	Schools
Tennis Club in CJU	14,900	Student Activities
Department of Information Management in CJU	10,449	Schools
Wukon Digital School	3,422	Online Learning
Special Education in Long Shan Elementary School	3,280	Special Education
Shih Hsin University	2,038	Schools
National Cheng Kung University	1,756	Schools
ALE Online	1,503	Supplement Education
ERP Center in Chung Yuan University	1,207	Schools
Yuan Ze University	1,196	Schools

TABLE 3. Under the "education > Schools" Category, the Most Visited Websites for CJU Users, the Number of Times Visited, and the Categories They Belong to.

Website	Number of Visits	Category
Chang Jung University	51,233	Colleges & Universities
Department of Information Management in CJU	10,449	Colleges & Universities
Shih Hsin University	2,038	Colleges & Universities
National Cheng Kung University	1,756	Colleges & Universities
ERP Center in Chung Yuan University	1,207	Colleges & Universities
Yuan Ze University	1,196	Colleges & Universities
Fu Jen Catholic University	1,186	Colleges & Universities
Ming Shuan University	1,112	Colleges & Universities
National Sun Yat-Sen University	1,076	Colleges & Universities
Chinese Culture University	963	Colleges & Universities

When a user clicks on the "Category Billboard" button, the most visited websites under all topic categories in Yahoo! Kimo for CJU users, and the number of times that each category was visited are displayed, as listed in Table 4. Table 4 shows that "Computers & Telecommunication", "Education", "Entertainment", "Business & Finance", etc.

are the most visited websites for CJU users. Once the user clicks on a category, the most visited sub-categories under that specific category for CJU users, and the number of times visited are displayed. For example, if a user clicks on the “Education” category, he/she will be able to view the most visited sub-categories under the “Education” category for CJU users, and the number of times that each of those sub-categories were visited, as listed in Table 5. Table 5 shows that under the “Education” category, “Schools”, “Student Activities”, “Online Learning”, “Further Education”, and so on are the most visited categories for CJU users. Furthermore, if a user clicks on the “Schools” category, he/she will be able to view the most visited sub-categories under the “Education > Schools” category for CJU users, and the number of times that each of them was visited, as listed in Table 6.

TABLE 4. Under all Categories of Yahoo! Kimo, the Most Visited Categories for CJU Users, and the Number of Times Each of Those Categories Were Visited.

Category	Number of Visits
Computers & Telecom.	276,892
Education	118,116
Recreation	100,610
Business & Finance	94,670
Internet Guides	88,935
Life Information	78,029
News & Media	56,918
Regional	32,154
Reference	20,420
Government	14,967
Sports	14,626
Arts & Culture	14,527
Society & Humanities	12,758
Entertainment	12,725
Healthcare	9,636
Science & Technologies	4,247

TABLE 5. Under the “Education” Category, the Most Visited Sub-categories for cju users, and the number of Times Each of the Sub-categories was Visited.

Category	Number of Visits
Schools	88,582
Student Activities	14,931
Online Learning	5,813
Further Education	4,271
Special Education	3,763
Skill Training	461
Organizations	192
Examinations	57
Teaching Resource	32
Internet	9
Regional	3
Finance & Investment	1
Shopping Guides	1

TABLE 6. Under the “Education > Schools” Category, the Most Visited Sub-categories for CJU Users, and the Number of Times Each of the Sub-categories Was Visited.

Category	Number of Visits
Colleges & Universities	88,452
Junior Colleges	119
Foreign Schools	11

VI. Conclusions and Further Research

A novel method of mining web usage for users within a LAN has been proposed in this study. The contribution of this study is threefold. (1) We point out an innovative idea of utilizing the search engine and the web directory of Yahoo! Kimo. (2) We proposed a topic hierarchy that is able to compute not only the number of times a specific website was visited, but also the number of times any topic category was visited. (3) We analyzed usage patterns and special interests of users at various levels of detail. The analytical results will be very useful for LAN management, the establishment of regional policies, and the planning of group activities. Finally, experimental results strongly support the above-mentioned claims.

Further research may be directed to the following issues. (1) Combining the proposed method with data mining techniques, such as association rules or sequential patterns, to analyze web usage patterns at various levels of detail. (2) Designing an intelligent web browser to recommend users for the next possible websites or categories, and assist users in organizing the websites in their bookmark of a web browser under a specific topic. (3) Designing content-based tools that can filter and control network flows according to web categories.

References

[1] R. H. Zakon, Hobbes' Internet Timeline, Available: http://www.zakon.org/robert/internet/timeline/.

[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, New York, USA, 1999.

[3] AltaVista, Available: http://www.altavista.com/.

[4] HotBot, Available: http://www.hotbot.com/.

[5] Northern Light, Available: http://www.northernlight.com/.

[6] Excite, Available: http://www.excite.com/.

[7] Yahoo!, Available: http://www.yahoo.com/.

[8] LookSmart, Available: http://www.looksmart.com/.

[9] eBLAST, Available: http://www.eblast.com/.

[10] NewHoo, Available: http://www.newhoo.com/.

[11] O. Etzioni, “The World Wide Web: quagmire or gold mine?” *Communications of the ACM*, Vol. 39, No. 11, pp. 65-68, Nov. 1996.

[12] B. Cooley, B. Mobasher, and J. Srivastava, “Web mining: information and pattern discovery on the World Wide Web,” *IEEE International Conference on Tools with Artificial Intelligence*, Newport Beach, CA, USA, Nov. 3-8, 1997.

[13] R. Kosala and H. Blockeel, “Web mining research: a survey,” *SIGKDD Explorations*, Vol. 2, No. 1, pp. 1-15, Jun. 2000.

[14] R. Cooley, *Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data*, Ph.D. Thesis, Department of Computer Science, University of Minnesota, May 2000.

[15] S. Chakrabarti, “Data mining for hypertext: a tutorial survey,” *SIGKDD*, Vol. 1, No. 2, pp. 1-11, Jan. 2000.

[16] S. Madria, S. S. Bhowmick, W. K. Ng, and E. P. Lim, “Research issues in web data mining,” *International Conference on Data Warehousing and Knowledge Discovery*, pp. 303-312, 1999.

[17] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large database,” *SIGMOD*, pp. 207-216, May 1993.

[18] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large database,” *International Conference on Very Large Data Bases*, pp. 478-499, Sep. 1994.

[19] J. Han and Y. Fu, “Discovery of multiple-level association rules from large database,” *International Conference on Very Large Data Bases*, pp. 420-431, Sep. 1995.

[20] J. S. Park, M. S. Chen, and P. S. Yu, “An effective hash based algorithm for mining association rules,” *SIGMOD*, pp. 175-186, May 1995.

[21] J. S. Park, M. S. Chen, and P. S. Yu, “Efficient parallel data mining for association rules,” *International Conference on Information and Knowledge Management*, Nov. 29-Dec. 3, 1995.

[22] R. Cooley, B. Mobasher, and J. Srivastava, “Data preparation for mining World Wide Web browsing patterns,” *Knowledge and Information Systems*, Vol. 1, No. 1, pp. 5-32, 1999.

[23] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer, and A. Swami, “An interval classifier for database mining applications,” *International Conference on Very Large Data Bases*, pp. 560-573, Aug. 1992.

[24] T. M. Anwar, H. W. Beck, and S. B. Navathe, "Knowledge mining by imprecise querying: a classification-based approach," *International Conference on Data Engineering*, pp. 622-630, Feb. 1992.

[25] J. Han, Y. Cai, and n. Cercone, "Knowledge discovery in database: an attribute-oriented approach," *International Conference on Very Large Data Bases*, pp. 547-559, Aug. 1992.

[26] G. Piatetsky-Shpario, "Discovery, analysis and presentation of strong rules," *Knowledge Discovery in Databases*, pp. 229-248, 1991.

[27] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, Vol. 1, pp. 81-106, 1986.

[28] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," *International Conference on Foundations of Data Organization and Algorithms*, Oct. 1993.

[29] J. T. L. Wang, G. W. Chirn, T. G. Marr, B. Shapiro, D. Shasha, and K. Zhang, "Combinatorial pattern discovery for scientific data: some preliminary results," *SIGMOD*, Minneapolis, MN, pp. 115-125, May 1994.

[30] R. Agrawal and R. Srikant, "Mining sequential patterns," *International Conference on Data Engineering*, pp. 3-14, Mar. 1995.

[31] M. S. Chen, J. S. Park, and P. S. Yu, "Data mining for path traversal patterns in a web environment," *International Conference on Distributed Computing Systems*, pp. 385-392, 1996.

[32] M. S. Chen, J. S. Park, and P. S. Yu, "Efficient data mining for path traversal patterns," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 10, No. 2, pp. 209-221, Mar./Apr. 1998.

[33] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic personalization based on web usage mining," *Communications of the ACM*, Vol. 43, No. 8, pp. 142-151, 2000.

[34] A. G. Buchner and M. D. Mulvenna, "Discovering internet marketing intelligence through online analytical web usage mining," *SIGMOD Record*, Vol. 27, No. 4, pp. 54-61, 1998.

[35] D. W. Cheung, B. Kao, and J. Lee, "Discovering user access patterns on the World Wide Web," *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 23-24, Singapore, Feb. 1997.

[36] M. Perkowitz and O. Etzioni, "Adaptive sites: automatically learning from user access patterns," *Technique Report TR-97-03-01*, Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA, 1997.

[37] Yahoo! Kimo, Available: http://tw.yahoo.com/.