Proceedings of International Conference on Technology and Social Science 2017
Invited Paper

A Browser Application for Keyword Recommendation Based on User Behavior

Chen Kuo Noriaki Yoshiura

chenkuo@fmx.ics.saitama-u.ac.jp yoshiura@fmx.ics.saitama-u.ac.jp

Information System Department, Faculty of Engineering, Saitama University,

255 Shimokubo, Sakura-ku, Saitama-shi, Saitama-ken, Japan

Keywords: Association rule mining, Apriori algorithm, Browser fingerprinting, Keyword recommendation

Abstract. Keywords are the words and phrases that are typed into the search box of a search engine (such as Google) by the Internet users to find out the websites that match what the Internet users are trying to find. Keywords can reflect the personal preferences and needs of a user. Some keywords on web pages have garnered much attention. However, the keywords you are interested in may rarely appear. Although third-party web tracking and Google Analytics can analyze keywords of users, both of them work for commercial advertising and web analytics technologists, and neither of them serves the users. Our goal is to provide some tailor-made keywords for users in the browser. We should help users to easily find keywords that they might be interested in. Thus, this paper presents a browser application collects and analyzes the data of user, and recommends the user some keywords that the user may be interested in.

1. Introduction

The communication between a browser and a web server is realized by HTTP. If a user sends a page request, a web server will respond simply. Then close the connection with the user. When a browser sends a request to the web server, no matter the browser is the first time to visit or not, the server will treat the browser as the first time. Because HTTP is a stateless protocol [1]. A stateless protocol does not require the server to retain information or status about each user for the duration of multiple requests. However, some web applications may have to track the progress of users from page to page. For example, a web server tracks the progress of a user is necessary, when a web server is required to customize the content of a web page for a user. To compensate for the defect of that HTTP is a stateless protocol, Netscape developed an HTTP cookie to save identification information of the user. An HTTP cookie is sent from a website and stored on hardware of the user through browsers. The most common example of using a cookie is to store names, preferences and password remember options of users. It is also one of the ordinary and mostly asked interview questions. Cookies help websites to provide visitors with a better experience.

Now the Internet has been a part of our lives. When a user visits some websites creating a Flash cookie [2,3] from a third-party advertisement company to be stored on computer the user, even though the user did not have a click on an advertisement or Flash video. The Flash cookie is undiscovered by users. Flash cookies are another means of tracking your movement on the Internet and storing lots of information on the user [4]. Flash cookies are not controlled by the browser or shown in the list of cookies manager of the browser. Furthermore, they did not appear in databases or other browser-specific storage locations. However, third-party web tracking [4,5] is a common phenomenon in networks.

In recent years, some advertisement companies would like to deliver advertisements to special users that may be interested in the advertisements. This situation causes most of the websites to use third-party web tracking. Users cannot benefit from third-party web tracking, furthermore, users receive target advertisement. On the other hand, browsing behavior analysis services of a user usually only

provide the analysis results of data of all users by a website. As a result, the user behavior analysis of a website is not a personal service. In other words, the service works for web analytics technologists. Although some websites offer user keyword analysis service, it is limited to browsing behavior within the websites. Thus, it is hard to find what a user is interested in when the user is browsing.

With the fast development of Internet, every day there are a lot of pictures, blogs, and video uploaded to the Internet. As a result, it is becoming more and more difficult for people to find the information they need in the massive data [9]. In this case, the search engine (Google, Bing, Baidu, etc.) has become the best way to quickly find the target information. However, the search engine cannot fully meet the requirements of users for information discovery. Because the results should be consistent with personal preferences in many cases.

On the other hand, the idea and range of knowledge of a user would have limitations. Ideas would be easily confined to their most commonly used vocabulary. However the keyword ever-changing, and the experience of each Internet user is also different. They may not be able to accurately enter a keyword to search engine for the content that they need. To summarize, in these cases, users donot realize their own needs.

Thus, we propose a keyword-based analysis tool for browsers. The aim of the keyword-based analysis tool is to provide users with a tool so that users can record and analyze their keywords without cookies. Then users will get some other keywords that they are interested in. At the same time, this tool consists of the following advantages. First of all, the tool can help users analyze their keywords without cookies and each user has his own analysis service. Second, users do not need to register or log in. In addition, the tool offers an extensive range of analytical service, including many search engine sites. Finally, the tool can help users to find keywords that they have not entered the keywords. At the same time, the recommended keywords conform to the requirements and preferences of users. Our purpose is to make the browser application is like a personal secretary and adviser to users.

The rest of this paper is organized as follows. Section 2 briefly presents the structure of keywords analysis application of browser. Section 3 illustrate experiments and the test results. The concluding remarks are finally made in Section 4.

2. Literature Reference

This section will briefly present general structure of keywords analysis application of browser and the backgrounds of browser fingerprinting.

The design of the tool is based on B/S (Browser, Server) model and three-tier architecture [12] that includes the User Interface layer, Business Logic layer and Data access layer.

2. 1. User Interface Layer

In the User Interface layer, a client will give a presentation in form of a web page. The client is a plug-in [3] of browser that provides the user with the interface. Then the browser is sending requests and required data of the user to server by HTTP protocol. After the server responds to the requests, the client will display the result in a browser. Users do not need to log in to the client. Every user will be created a user ID by browser fingerprinting when the user first time to use the keyword analysis tool. Browser fingerprinting is the systematic collection of information about a remote device, for identification purposes. This is not like a cookie, which is saved information from a site stored on computer of the user. Instead, browser fingerprinting involves digging into settings and configuration information that the browser gives when the page or plugin suggests it. We collect the information of the use includes the user agent string from browser, screen resolution, lists of fonts, the plugins the user have installed, etc. We use the parameters of fingerprinting were described by Table 1.

Passive elements	Active elements
IP address Operating system User agent Language Http accept headers Encoding header	Time zone Screen resolution and its color depth List of plugins Cookies preferences (allowed or not) List of fonts Do Not Track preferences Presence of AdBlock A picture rendered with the HTML Canvas element A picture rendered with WebGL

Table 1. Browser fingerprinting

Then the ID as the unique identification [11] of a user will be stored in a file created by the client. The file stored in installation directory folder of the client. The user ID will be directly read when using the client again. When a user clicks the button of the search engine, the client will get the words and phrases in the box of a search engine by the users. Meanwhile, those words and phrases as well as user ID will be uploaded and stored together in the keywords sheet of the server as the input of user keyword analysis tool. When a user clicks a plug-in icon in the search engine, he or she can see the keywords just searched and another corresponding keyword returned by the server. This layer accepts input of the user, output to the user and offers interface to the user. The User Interface layer flowchart for collecting information a user can be seen in the following Fig. 1:

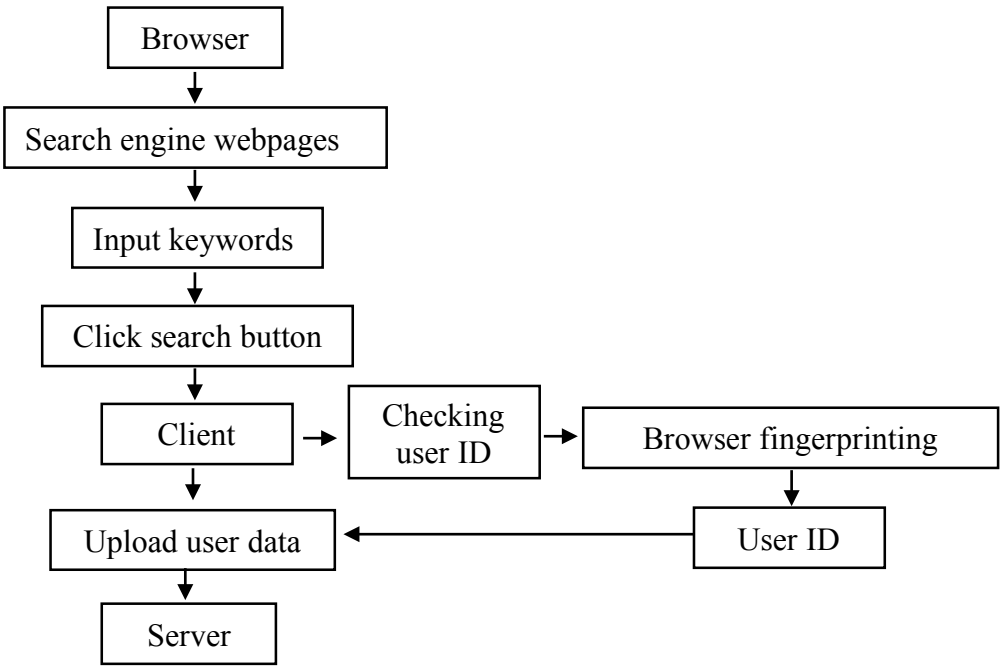

Fig. 1. Collect user information flowchart

2. 2. Business Logic Layer

The Business Logic layer is the application server that is the most important part of the tool. The application server has the functions of transaction processing, database connectivity and messaging. The application server not only accepts requests of the user but also achieves the association analysis of keywords and statistics keywords frequency. The application server will not keep analyzing associated keywords. Because it will make the server response to request of user slow. Thus, we set up a period of time, the application server will conduct an association analysis to keywords of all users. Then we store the results of the analysis into the database. When a user sends a request, the application server will read the recommended keywords directly from the database. The Business Logic layer flowchart can be seen in the following Fig. 2:

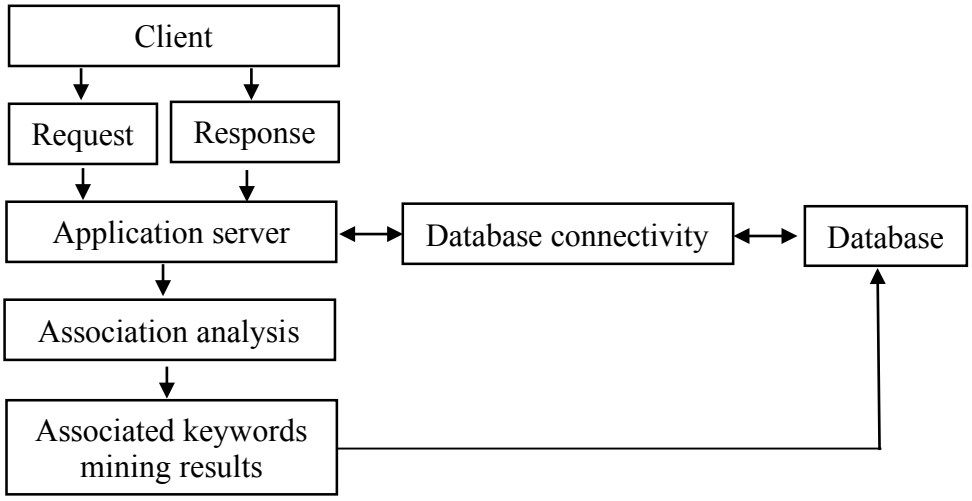

Fig. 2. Business Logic layer flowchart

2. 3. Data Access Layer

Data access layer is SQL (Structured Query Language) Server that responds to requests of the application server and completes operations on the data. The main function of the data access layer is responsible for access to the database. The application server can access database systems, binaries, text documents, and XML documents. A simple statement is to achieve the database table Select, Insert, Update, Delete operation.

3. Tool discussion

3. 1. Application of association rule in research

It is widely known that Apriori algorithm is used to analyze shopping lists of customers of the supermarket. However, Apriori algorithm is used to analyze the keywords of users in our research. On the other hand, the algorithm works for business services, such as help supermarkets to sell more goods. In our study, the algorithm can help users to find some keywords that they may be interested in. How to use association rule in our research can be seen in the following Fig. 3:

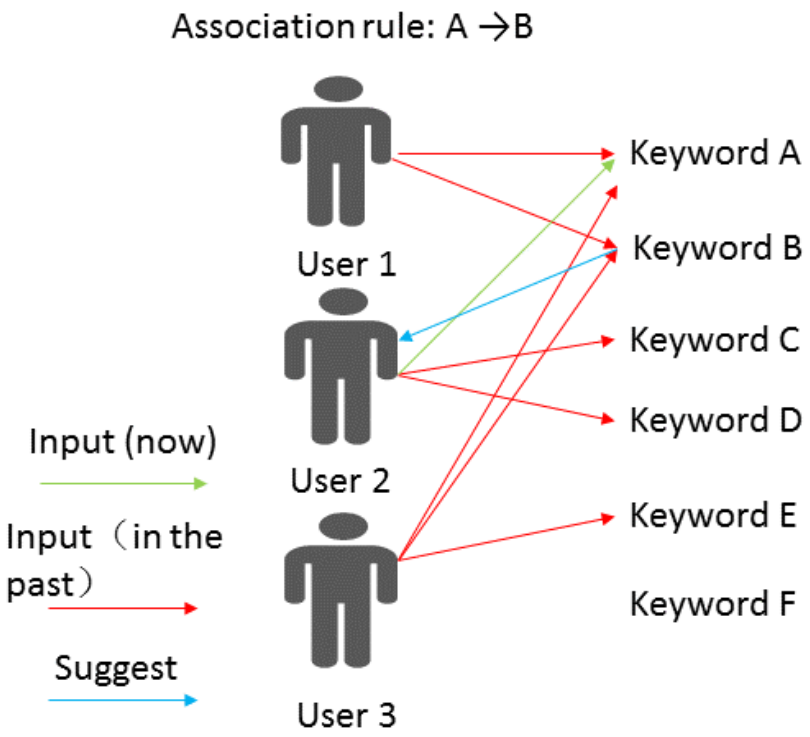

Fig. 3. Keyword recommend method

3. 2. Experimental

In order to test the keyword analysis application recommended keywords to users are helpful or not. We designed two groups of experiments. Two groups of experiments with the same server.

In the experiment one, the tables in the database are null before the experiment. We set the minimal support value is 0.6 and the minimal confidence value is 0.5. Then we used five testers to randomly input some keywords. After the server got 100 keywords, we stopped the experiment. The five testers cannot find their suggested keywords. Through this implementation, we found that in order to get the expected keyword of users, this algorithm needs a large number of users and keywords.

In the second experiment, we improved the experimental method. We find 500 names of classical moves that are well known by ourselves, then insert them to frequent keyword set E with support values. Then we make the association rules and set the minimal support value is 0.3 and the minimal confidence value is 0.4. Then we get 100 testers to download and install the plug-in by social networks. This time the tool can create the recommended keyword by inputs of users. We still have 31 active users until now.

4. Conclusion

This paper is presents an overview and method of a keyword-based analysis tool for browser. The keyword analysis browser application is useful in association analysis of keywords and statistics keywords frequency. The web application is an application that records and analyzes keywords of the user of search engine. The keyword analysis indicates the statistics keywords frequency and association analysis keywords of the user.

Meanwhile, we are facing many challenges. It is well known that the association rule algorithm is based on Big Data of users. Thus, we need to prepare a large amount of data as the original data. On the other side, we need to consider improving the efficiency and accuracy of the algorithm.

References

[1] Fielding R, Gettys J, Mogul J, et al. "Hypertext Transfer Protocol -- HTTP/1.1.", ***Computer Science & Communications Dictionary***, Vol. 7, No. 4, pp. 595-599, 1996.

[2] A. Soltani, S. Canty, Q. Mayo, L. Thomas, and C. J. Hoofnagle, "Flash Cookies and Privacy", ***Intelligent Information Privacy Man-agement of AAAI Spring Symposium,*** Vol. 2010, pp.158-163,2010.

[3] M. Ayenson, D. J. Wambach, A. Soltani, N. Good, and C. J. Hoofnagle, J/OL., "Flash cookies and privacy II: Now with HTML5 and ETag respawning", ***Ssrn Electronic Journal***, pp.1-21, 2011.

[4] MAYER, R. Jonathan, MITCHELL, J. John, "Third-party web tracking: Policy and technology", ***Security and Privacy of 2012 IEEE Symposium,*** pp.413-427, 2012.

[5] ROESNER, Franziska, KOHNO, Tadayoshi, WETHERALL and David, "Detecting and defending against third-party tracking on the web", ***Proc. The 9th USENIX on Networked Systems Design and Implementation***, (San Jose, America), pp.12-12, January, 2012.

[6] ECKERSLEY, Peter, "How unique is your web browser?" ***Privacy Enhancing Technologies of Springer Berlin Heidelberg***, pp.1-18, 2010.

[7] XIANG, Jian-chi, Xiang-bin LIU, and Xuan-hua XU, "Research on Data Collection Technology for Web Usage Mining Based on User Behavior", ***Computer and Modernization***, pp.59-62, 2007.

[8] Kuo, Ren Jie, Chie Min Chao, and Y. T. Chiu, "Application of particle swarm optimization to association rule mining", ***Applied Soft Computing***, pp.326-336, 2011.

[9] X. Wu, X. Zhu, G. Q. Wu and W. Ding, "Data mining with big data", ***Transactions on Knowledge and Data Engineering of IEEE***, Vol. 26, No. 1, pp.97-107, 2014.

[10] Liu, Xingtao, S. Bing, and Y. Xie, "An Improved Apriori Algorithm for Mining Association Rules", ***Computer Technology and Development***, Vol. 43, pp.10-12, 2009.

[11] Mulazzani M, Reschl P, Huber M, et al., "Fast and reliable browser identification with javascript engine fingerprinting" ***Proc. Web 2.0 Workshop on Security and Privacy*** **(**San Francisco, American) pp.1-18, May, 2013.

[12] Helal S, Hammer J, Zhang J, et al., "A three-tier architecture for ubiquitous data access", ***Computer Systems and Applications of ACS/IEEE International***, pp.177-180, 2001.