# FREQUENT SEARCH KEYWORD IDENTIFICATION FROM THE SEARCH LOG

Nurul Farhana Ismail[1], Wan Hussain Wan Ishak[2] and Faudziah Ahmad[3]

[1] School of Computing, Universiti Utara Malaysia, 06010, Sintok, Kedah
[2] School of Computing, Universiti Utara Malaysia, 06010, Sintok, Kedah. Email: hussain@uum.edu.my
[3] School of Computing, Universiti Utara Malaysia, 06010, Sintok, Kedah. Email: fudz@uum.edu.my

**Abstract**: *Search log is a rich source of information where it contains valuable information such as searcher information, keyword, date and time. Retrieving these information from the search log will provide some insight into searcher searching patterns. The pattern is vital in identifying searcher's interest, thus enable the search system to incorporate a recommendation and support to the searcher. This paper discusses the analysis of the search keywords that are obtained from a university library search log. Through the analysis, a set of frequent search keywords has been identified. The top five most frequent keywords used are management, organization, network, technology, and performance. The analysis also shows that these keywords were used in various databases.*

**Keywords:** Frequent Keyword, Search Log, Keyword Identification, Library.

## Introduction

Search log is an important source of information where it contains information related to the users' activities and behaviour while using the search system (Yamin et al., 2014). Information stored in the search log can be used to study the users' searching pattern (Yamin et al., 2013). This information can be used by the information retrieval system to enhance its operation.

Information on the previous successful searchers activities can be used to assist other searchers to find what they what (Yamin et al., 2015). This approach is known as a recommender system (Adomavicius and Tuzhilin, 2005; Melville and Sindhwani, 2010). Recommender system is a system innovation that assists users by recommending items that are related to the user's interest (Melville and Sindhwani, 2010). As such, the search recommender system assists searchers by providing recommendations based on previous search activities (Grossman, 2010).

Data mining techniques (Han & Kamber, 2006) are potential ways to be used in developing the recommender model (Mobasher et al., 2000; Pierrakos et al., 2003). Data mining can be used to analyze the system log to look for the histories of the item retrieval. The log contains data about the previous searching, database selection and download. Additionally, some log

also stores users' profile such as the computer IPs, sessions, and some time users' ID (Yamin et al., 2013). These data can be utilized to assist other searchers who are looking for similar information.

This paper discusses the findings from the search log analysis. The aim of this study is to identify frequent search keywords that were used by the searchers. Data mining technique, namely web usage mining is applied in the process and the list of frequent search keywords were presented in the findings section.

**Literature Review**

Recommender system is aimed to assist users to find information and creates personalized content to the end users (Mahmood et al., 2009). Typically, the web pages offer three major information: content itself, the structure, and usage patterns. Content data consists of whatever is in a web page; structure data refer to the organization of the content; usage data are the usage patterns of web sites.

There are three main types of recommender systems: collaborative filtering, content-based filtering, and demographic recommender systems (Lopes & Roy, 2015). Collaborative filtering recommender systems recommend items by taking into account the preferences of items to the users. This is with the assumption that users will be interested in items where others might have highly rated. Content-based filtering recommender systems recommend items based on the textual information on an item, under the assumption that users will like similar items to the ones they liked before. Demographic recommender systems categorize users or items based on their personal attribute and make recommendation based on demographic categorizations.

Data mining technique is one of the potential approaches to be used in developing a recommender system. Data mining is the process of finding useful patterns in data is known by different names in different communities (e.g., knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing) (Fayyad et al., 1996). Data mining approaches can be used to analyse the log file for histories of keyword search, ip-address, and session and user id. The data inside log file can be utilized for other users who are looking for the same keyword. The application of the data mining techniques to these different data sets is at the basis of the three different research directions in the field of web mining: web content mining, web structure mining and web usage mining (Zhang & Chang, 2002).

Web content mining is the process of obtaining useful information from web content (Adsod & Chopde, 2014) while web structure mining as the discovery of useful knowledge from the hyperlinks that representing the structure of the web site (Dharmarajan & Dorairangaswamy, 2014). Web usage mining is the process of detecting out what a user's are viewed in the internet. It can be concluded as the observed and survey of user access pattern, while extracting files and data from a web site to identify and enhance the utility of web-based applications (Gupta et al., 2014).

Web usage mining can be used to understand the user's preferences and behaviour (Mobasher et al., 2000). Srivastava et al (2000) describes that web usage mining is the function of data mining approach to identify usage patterns from Web data, in order to meet the requirement of Web-based applications. Aghabozorgi & Wah (2009) defined that in a web site there are three inputs have to be handled which is content, structure and log data. Web usage mining is applied on usage data and it is being used in large scale by the organization to study the behaviour of their web user (Singh et al., 2013). Most of the data needed for web log analysis reside on web servers, proxy servers, enterprise logs, web clients, and etc. The processes in web usage mining begin with data pre-processing in order to prepare the data for the analysis with the data mining algorithm (Dhandi & Chakrawarti, 2016).

## Methodology

In this study the data were obtained from Perpustakaan Sultanah Bahiyah (PSB) of Universiti Utara Malaysia (UUM). The data are the search log of the online databases subscribed by the library. As for the case study, data from the first six months of the year 2012 were taken. This selection is due to the large amount data which required extensive processing time and computing resources. The pre-processing procedure has been conducted on the raw data to remove un-relevant information. Figure 1 shows a sample of the raw data.

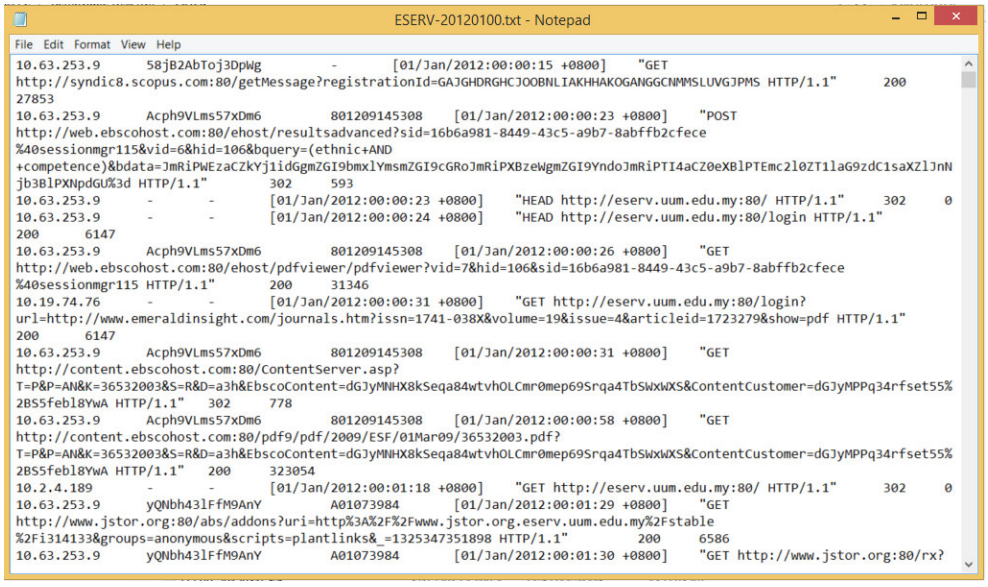


Figure 1: Sample of the raw data

Once the data have been pre-processed, the search keywords used by the users are detected. Typically, the keywords are embedded in the URL that ends with the word "Query", "Keyword" and etc. The search keywords are kept in a separate file according to its database name. Figure 2 shows sample keyword in log file. Finally, the keywords are analysed and the frequency is calculated.

```
GET http://www.emeraldinsight.com:80/search.htm?st1=sukuk+structure&ct=all&ec=1&bf=1 HTTP/1.1
GET http://proquest.umi.com:80/pqdweb?RQT=302&COPT=REJTPUcyODcrM2IxMCZJTlQ9MCZWRVI9Mg==&clientId=28929&cfc=1 H
GET http://www.emeraldinsight.com:80/journals.htm?issn=1753-8394&volume=2&issue=4&articleid=1826941&show=pdf HTTP/1.1
GET http://www.emeraldinsight.com:80/journals.htm?issn=1759-0817&volume=1&issue=1&articleid=1858421&show=pdf HTTP/1.1
GET http://proquest.umi.com:80/pqdweb?SQ=continuous+education+and+retention&DBId=G647&date=ALL&onDate=&beforeDate=&af
GET http://www.emeraldinsight.com:80/search.htm?st1=sukuk+structure&ec=1&bf=1&ct=jnl HTTP/1.1
GET http://proquest.umi.com:80/pqdweb?SQ=continuous+education+factors&DBId=G647&date=ALL&onDate=&beforeDate=&afterDate
GET http://eserv.uum.edu.my:80/connect?session=s5kK6uj4JmxDFOgY&url=menu HTTP/1.1
GET http://eserv.uum.edu.my:80/menu HTTP/1.1
GET http://www.emeraldinsight.com:80/journals.htm?issn=0828-8666&volume=26&issue=1&articleid=1846113&show=pdf HTTP/1.1
GET http://www.emeraldinsight.com:80/search.htm?st1=sukuk+structure&ec=1&bf=1&ct=jnl&nolog=167626&page=2 HTTP/1.1
GET http://eserv.uum.edu.my:80/public/rightmenu.htm HTTP/1.1
GET http://www.emeraldinsight.com:80/search.htm?st1=sukuk+structure&ec=1&bf=1&ct=jnl&nolog=167626&page=2 HTTP/1.1
GET http://www.emeraldinsight.com:80/journals.htm?issn=1753-8394&volume=2&issue=2&articleid=1795358&show=pdf HTTP/1.1
GET http://www.emeraldinsight.com:80/journals.htm?issn=1759-0817&volume=1&issue=1&articleid=1858422&show=pdf HTTP/1.1
GET http://www.emeraldinsight.com:80/search.htm?st1=sukuk+structure&ec=1&bf=1&ct=jnl&nolog=167626&page=3 HTTP/1.1
GET http://www.emeraldinsight.com:80/search.htm?st1=attitude&ec=1&bf=1&ct=jnl&nolog=814577&page=2 HTTP/1.1
GET http://www.emeraldinsight.com:80/search.htm?st1=changing+attitude&ct=all&ec=1&bf=1 HTTP/1.1
GET http://eserv.uum.edu.my:80/connect?session=sjZEm3Xln7jzCwY3&url=menu HTTP/1.1
```

Figure 2: Keywords detection

## Findings

In this study a total of 19146 keywords have been extracted from the dataset. These keywords are associated with their databases, where 11 databases were identified from the log file. Table 1 summarizes the number of keywords obtained from this study. As shown in Table 1, the largest number of keywords was obtained from the IEEE database with a total of 4862 keywords. From of this figure, 3418 keywords are unique while 808 keywords are frequent. The second largest database is SerialSolution with 3043 keywords. The rest of databases recorded less than 2000 total keywords. This figure shows that IEEE and Serial Solution are the most popular database with the searching activity.

Table 1: Summary of Keywords by Database

| NO | DATABASE | NUMBER OF KEYWORDS | FREQUENT KEYWORDS | SUM OF KEYWORDS |
|---|---|---|---|---|
| 1. | ebscohost | 1552 | 217 | 1835 |
| 2. | Emerald | 912 | 74 | 1011 |
| 3. | Eric | 462 | 341 | 1081 |
| 4. | IEEE | 3418 | 808 | 4862 |
| 5. | JournalCambridge | 1080 | 67 | 1148 |
| 6. | JSTOR | 1321 | 150 | 1503 |
| 7. | OnlineLibrary | 956 | 80 | 1044 |
| 8. | ScienceDirect | 1561 | 255 | 1906 |
| 9. | SCOPUS | 1226 | 124 | 1379 |
| 10. | SerialSolution | 2408 | 363 | 3043 |
| 11. | XREF | 307 | 18 | 334 |
| | **TOTAL** | **6143** | **2497** | **19146** |

Table 2 shows the analysis of the frequent keywords by databases. The table shows that most of the keywords repeat two and three times in the databases. Some databases such as IEEE, SerialSolution, ScienceDirect, Ebscohost, and Eric contains keywords that repeat more than

three times. IEEE database has 9 keywords that repeat more than 10 times, while SerialSolution has 5 keywords that repeat more than 10 times.

Table 2: Keyword Frequencies by Database

| NO. | DATABASE | KEYWORD FREQUENCIES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | >10 |
| 1. | ebscohost | 1335 | 168 | 40 | 6 | 1 | 1 | 0 | 0 | 1 | 0 |
| 2. | Emerald | 838 | 54 | 16 | 3 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3. | Eric | 121 | 121 | 194 | 7 | 8 | 10 | 0 | 1 | 0 | 0 |
| 4. | IEEE | 2610 | 516 | 151 | 60 | 39 | 18 | 9 | 3 | 3 | 9 |
| 5. | JournalCambridge | 1013 | 66 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6. | JSTOR | 1171 | 130 | 11 | 6 | 3 | 0 | 0 | 0 | 0 | 0 |
| 7. | OnlineLibrary | 876 | 75 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 8. | ScienceDirect | 1306 | 196 | 41 | 11 | 3 | 2 | 2 | 0 | 0 | 0 |
| 9. | SCOPUS | 1102 | 107 | 8 | 6 | 3 | 0 | 0 | 0 | 0 | 0 |
| 10. | SerialSolution | 2045 | 240 | 59 | 36 | 9 | 9 | 3 | 1 | 1 | 5 |
| 11. | XREF | 289 | 12 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| | **TOTAL** | **12706** | **1685** | **527** | **139** | **68** | **40** | **14** | **5** | **5** | **14** |

Table 3 shows the 30 highest frequent keywords. Keyword *management, performance, entrepreneur, system, service, business, theory, and relationship* were found in all databases. Keyword management is the highest frequency in all databases with the total of 68 times and average 6.18. The lowest is *methodology* that only available in 7 databases with average 3.43.

Table 3: Statistics of the 30 highest frequent keywords

| NO | KEYWORD | NUMBER OF SOURCE DATABASES | KEYWORD FREQUENCY | | | | |
|---|---|---|---|---|---|---|---|
| | | | TOTAL | MIN | MAX | AVERAGE | MEDIAN |
| 1. | management | 11 | 68 | 1 | 20 | 6.18 | 4 |
| 2. | organization | 9 | 44 | 2 | 12 | 4.89 | 5 |
| 3. | network | 10 | 39 | 1 | 15 | 3.90 | 2 |
| 4. | technology | 10 | 37 | 1 | 10 | 3.70 | 2 |
| 5. | performance | 11 | 36 | 1 | 7 | 3.27 | 3 |
| 6. | entrepreneur | 11 | 35 | 1 | 10 | 3.18 | 2 |
| 7. | communication | 10 | 34 | 1 | 14 | 3.40 | 2 |
| 8. | organizational | 10 | 34 | 1 | 6 | 3.40 | 3 |
| 9. | information | 10 | 32 | 1 | 9 | 3.20 | 2.5 |
| 10. | system | 11 | 32 | 1 | 11 | 2.91 | 2 |
| 11. | service | 11 | 31 | 1 | 7 | 2.82 | 3 |
| 12. | business | 11 | 30 | 1 | 8 | 2.73 | 2 |
| 13. | company | 10 | 30 | 1 | 10 | 3.00 | 2 |
| 14. | application | 10 | 29 | 1 | 10 | 2.90 | 2.5 |
| 15. | entrepreneurship | 10 | 29 | 1 | 9 | 2.90 | 1.5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 16. | attitude | 10 | 28 | 2 | 4 | 2.80 | 2.5 |
| 17. | characteristic | 9 | 28 | 1 | 4 | 3.11 | 3 |
| 18. | strategy | 10 | 28 | 1 | 6 | 2.80 | 2 |
| 19. | theory | 11 | 28 | 1 | 6 | 2.55 | 2 |
| 20. | journal | 10 | 27 | 1 | 6 | 2.70 | 2.5 |
| 21. | Knowledge | 9 | 27 | 1 | 6 | 3.00 | 2 |
| 22. | Malaysia | 9 | 27 | 1 | 10 | 3.00 | 2 |
| 23. | student | 10 | 26 | 1 | 6 | 2.60 | 2 |
| 24. | difference | 10 | 25 | 1 | 5 | 2.50 | 2 |
| 25. | effect | 10 | 25 | 1 | 6 | 2.50 | 2 |
| 26. | employee | 10 | 25 | 1 | 5 | 2.50 | 2 |
| 27. | impact | 10 | 25 | 1 | 5 | 2.50 | 2 |
| 28. | relationship | 11 | 25 | 1 | 4 | 2.27 | 2 |
| 29. | development | 10 | 24 | 1 | 6 | 2.40 | 2 |
| 30. | methodology | 7 | 24 | 1 | 11 | 3.43 | 2 |

## Conclusion

The findings of this study show the searchers' searching pattern which includes the keyword they used and the database that they are accessing. This study has identified more than fifteen thousand keywords. From this figure, the most frequent keyword is management followed by organization, network, technology, performance, entrepreneur and etc. These keywords were found in almost all databases. The most popular database is IEEE database. This choice is probably due to the coverage and features offered by IEEE.

These findings show that searchers are actively using an online database to search and access information especially academic articles. The searching pattern shows that searchers are not just focusing on one database. The searchers seem to use more than one database in order to find what they want. This is evident from the 30 most frequent keywords presented in Table 3, where these keywords were used in almost all 11 databases.

## Acknowledgements

The authors wish to thank the Universiti Utara Malaysia for funding this study under University Research Grant and UUM Sultanah Bahiyah Library for supplying the data

## References

Adomavicius, G. and Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transaction on Knowledge and Data Engineering*, *17(6)*, 734-749.

Adsod, A. A. and Chopde, N. R. (2014). A Review on: Web Mining Techniques. *Int. J. Eng. Trends Technol., 10(3),*108–113

Aghabozorgi, S.R., and Wah, T.Y. (2009). Recommender systems: incremental clustering on web log data. *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, pp. 812-818

Dhandi, M., and Chakrawarti, R. K. (2016). A comprehensive study of web usage mining. *Symposium on Colossal Data Analysis and Networking (CDAN)*, pp. 1-5. DOI: 10.1109/CDAN.2016.7570889

Dharmarajan, K., and Dorairangaswamy, M.A. (2014). Current Literature Review - Web Mining, *Elysium Journal, 1(1)*, 38–42

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Mag.*, 17(3), 37–54

Grossman, L. (2010). How Computers Know What We Want — Before We Do, *Time* (Thursday, May 27,2010). Url: http://www.time.com/time/printout/0,8816,1992403,00.html

Gupta, A., Arora, R., Sikarwar, R., and Saxena, N. (2014). Web usage mining using improved Frequent Pattern Tree algorithms. *International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, pp. 573-578. DOI: 10.1109/ICICICT.2014.6781344

Han, J. & Kamber, M. (2006). *Data Mining: Concepts and Techniques (2nd)*. Morgan Kaufmann Publishers: US.

Lopes, P., and Roy, B. (2015). Dynamic Recommendation System Using Web Usage Mining For E-Commerce Users, *Procedia Computer Science, 45*, 60–69.

Mahmood, T., Ricci, F., and Venturini, A. (2009). Learning adaptive recommendation strategies for online travel planning, *Inf. Commun. Technol. Tour. 2009*, pp. 149–160

Melville, P. and Sindhwani, V. (2010). Recommender Systems. In C. Sammut, G. Webb (eds.), *Encyclopedia of Machine Learning, Springer-Verlag Berlin Heidelberg*, pp: 1-9.

Mobasher, B., Cooley, R. & Srivastava, J. (2000) Automatic Personalization Based on Web Usage Mining. *Communications of the ACM*, 43(8), pp: 142-151.

Pierrakos, D. Paliouras, G. O., Papatheodorou, C. & Spyropoulos, C. D. (2003). Web Usage Mining as a Tool for Personalization:A Survey. *User Modeling and User-Adapted Interaction*, *13*, pp: 311-372

Singh, N., Jain, A., and Raw, R.S. (2013). Comparison Analysis of Web Usage Mining Using Pattern Recognition Techniques. *International Journal of Data Mining & Knowledge Management Process (IJDKP), 3(4)*, 137-147

Srivastava, J., Cooley, R., Deshpande, M., and Tan, P-N. (2000). Web usage mining: discovery and applications of usage patterns from Web data. *ACM SIGKDD Explorations Newsletter, 1(2),* 12-23

Yamin, F.M., Ramayah, T., and Ishak, W.H.W. (2013). Search Interface to Capture Searchers Behaviour. *International Journal of Computing Academic Research, 2(2),* pp: 67-74.

Yamin, F.M., Ramayah, T., and Ishak, W.H.W. (2015). Does User Search Behaviour Mediate User Knowledge and Search Satisfaction?, *International Journal of Economics and Financial Issues (IJEFI), 5*, pp. 34-39

Yamin, F.M., Ramayah, T., Ishak, W.H.I., & Othman, S.N. (2014). Search Log Analysis Method to Uncover User Search Behaviour on Web Searching Environment. *Proceedings of International Conference on Research Methods in Management and Social Sciences (ICRMMS-2014)*, 139-151.

Zhang, F., and Chang, H-Y. (2002). Research and development in web usage mining system-key issues and proposed solutions: a survey. *Proceedings International Conference on Machine Learning and Cybernetics*, 2, pp. 986-990. DOI: 10.1109/ICMLC.2002.1174531