

Discovering the Web Usage in three Jordanian Universities

Areej Al-Qwaqenah

CIS Department IT & CS Faculty. Yarmouk University Irbid, Jordan
Areej981@yahoo.com

Belal Abu Ata

CIS Department IT & CS Faculty. Yarmouk University Irbid, Jordan
belalabuata@yu.edu.jo

Mohammed Al-Kabi

CIS Department IT & CS Faculty. Yarmouk University Irbid, Jordan
mohammedk@yu.edu.jo

ABSTRACT—The Analyzing users' Web log data is important and challenging research topic of Web usage mining, which is an important technique to show users' behavior. This paper aims at discovering the Web usage in three Jordanian universities (Al-Balqa' Applied University, Hashemite University, and Yarmouk University) through the analysis of the Web log files of their servers. The results show differences in Web usage between these universities, the tests show that these differences are statistically significant through the use of Chi-square test for independence.

Keywords —*Web usage mining, Web server log files, chi-square test for independence, and level of significance.*

I. INTRODUCTION

Web usage mining is the process of applying data mining techniques to discover usage patterns based on Web log files. Web usage mining consists of the following three main stages [1]: preprocessing, pattern discovery, and pattern analysis. Figure 1 show these stages, where the log file represents an input of these three main stages.

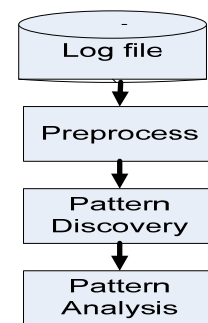


Figure 1. Web log mining structure

Data preprocessing is the process of eliminating irrelevant fields from access log

file. There are two primary tasks in preprocessing:

1. Data cleaning: Data cleaning eliminates unnecessary data (such as button images) from the original Web log file.
2. Transaction identification: The transaction identification process groups the sequences of page requests into logical units, each of which is called a session.

The purpose of data preprocessing is to extract useful data from raw Web log files. Figure 2 shows an overview of data preprocessing.

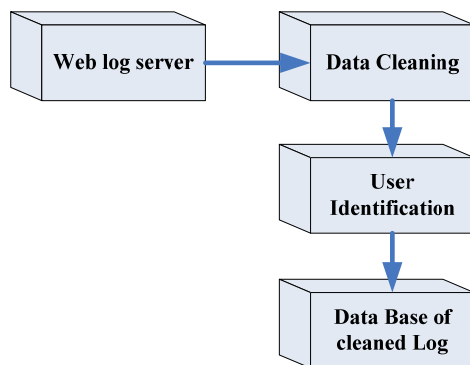


Figure 2. Overview of Data Preprocessing

Analyzing users' Web log data and extracting their search terms is important and a little bit challenging task for Arabic terms. Web logs contain information about users accessing the site and maintained by Web servers. The Web server log files records all client requests (hits) processed by the server. The row of a Web server log file represents a client request (hit) made to a Web server for a single Web page. Each line of a Web server log file specifies a hit to the Web site. Figure 3 shows the basic items of a Web log file, such as the IP address of the visitor, date and time of the visit ...etc.

172.17.1.17	anonymous	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)
2010-03-11	11:15:50	AJLOUN-ISA
10.232.1.77	10.232.1.77	
8080	14954	711 34738 http GET
http://www.google.jo/search?hl=en&source=hp&q=%D8%A7%D9%84%D8%BA%D8%AF&btnG=Google+Search&meta=Upstream		
200		

Figure 3. Web log mining structure

Web log files generated by Web/proxy servers are text files with a row for each HTTP transaction. A typical row contains the following information as shown in Table 1.

Table 1: Fields of the Web log File Row

Field Name	Value
IP address	172.17.1.17
Remote log name	Anonymous
Authenticated user name	AJLOUN-ISA
Timestamp	2010-03-11 11:15:50
Access request	http GET
Result status code	711
Bytes transferred	34738
Referrer URL	http://www.google.jo/search?hl=en&source=hp&q=%D8%A7%D9%84%D8%BA%D8%AF&btnG=Google+Search&meta=
User Agent	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)

In the previous figure, the first entry indicates that an anonymous user with the IP address of 172.17.1.17 at 11:15:50 on 2010-03-11, from a server named AJLOUN-ISA at IP address 10.232.1.77.

In pattern discovery methods, data mining techniques are used in order to extract patterns of usage from the Web data. These discovery methods may include association

rules, clustering, classification, Dependency modeling or mining for sequential patterns. Pattern analysis is the final stage of Web usage mining, the aim of this process is to extract the interesting rules or patterns from the output of pattern discovery process.

The initial trigger to writing this paper is the literature lacks comparative studies related to the Web usage within Jordanian universities. In this paper, we will try to discover the Web usage in three Jordanian universities (Al-Balqa' Applied University, Hashemite University, and Yarmouk University).

This paper will be organized as follows: the second section will introduce related work to this study, the third section presents the goals and approaches of this study, the fourth section presents the experiments and evaluation, section five presents the Evaluation Results, and section six presents the conclusion.

II. RELATED WORK

Research for analyzing Web log data has been done by many researchers in the field of Web usage mining; while Pei *et al.* [2] have successfully used the log data from Web logs to discover frequent patterns, they proposed an algorithm called (WAP) Web access pattern tree for efficient mining of access patterns from pieces of logs, Murate *et al.* [3] highlights the importance of analyzing users web log data and extracting their interests of web-watching behaviors and describes a method for clarifying users interests based on the analysis of the site-keyword graph, while Borges *et al.* [4] modeled users' to capture Web navigation patterns.

The study of Suneetha *et al.* [5] concerned with the in-depth analysis of Web log data of NASA Web site to help system administrator and Web designer to improve their system to arrange their system by determining occurred system errors, corrupted and broken links. They proposed model to form well focused

data of interested users and then frequent pattern mining algorithm is applied on this group instead of considering overall entries, which intern improves the performance. The proposed model consists of two phases. In first phase, the web server log data is preprocessed. The purpose of data preprocessing is to extract useful data from raw web log. In the second phase data is classified using enhanced version of decision tree algorithm C4.5. NASA web server data is used for experimental purpose, which results in less execution time and reduced memory utilization with high accuracy.

Spink *et al.* [6] analyzed characteristics of general Web search logs from different perspectives: terms, queries, sessions, and result pages. They showed top users short queries, a small number of search terms were used with high frequencies, few queries were modified, few result pages per query were be visited, and the popularities of query topics. Rose *et al.* [7] has been done on analyzing user goals and categorized general search queries according to navigational, informational, and resource queries, and they counted the proportion of queries in each category. While the Sanderson *et al.* [8] studied the popularity of query topics for geographic queries in general search, they indicate that geographically related queries are a significant sub-set of the queries submitted to a search engine and it differ from other queries.

III. GOALS AND APPROACHES

This paper presents an approach used to discover the differences in search interests and Web usages on the academic level within Jordan by comparing the users' navigational behavior and interests between three of Jordanian universities (Al-Balqa' Applied University, Hashemite University, and Yarmouk University) by using an application which designed and implemented in this study to extract the top Web pages from Web log

files from these Universities, this approach is graphically presented in Figure 4.

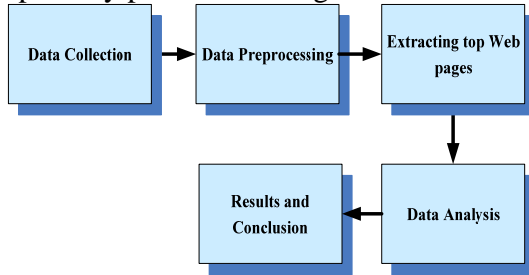


Figure 4. The methodology of the study

Figure 5 summarizes the following general steps to satisfy the main purposes to compare the Web usage within the three Jordanian universities.

- **Data collection:** Gathering the Web log files from several Web servers within the three Jordanian universities. This step includes accessing the server of these universities daily during October – 2010 and collecting the Web log files which are text files containing information about the users behavioral on the web. Some of this information is beyond the scope of this project.
- **Data preprocessing:** This step includes revision, refinement, and cleaning the Web log files, such as images, multimedia files, and page style files, for example, requests for graphical page content (.jpg and .gif) are removed.
- **Extracting top Web pages:** Applying proposed algorithm on the preprocessed data in the previous step to extract the top Web pages during specific period of time. This algorithm is illustrated in Figure 6, and it is applied on a Web log files represented by set of preprocessed URLs. The idea of using this algorithm is to perform an iterative search over a Web server log file that collected from the three Jordanian universities to extract Web pages. The outcome of this algorithm is the File_of_URLs consisting of all Web pages. After that we determine the category of each Web page whether it

is low visited Web page, medium visited Web page, and high visited Web page. Highly visited Web pages will be analyzed and compared with highly visited Web pages collected from other sources. This implemented algorithm is called Ranking of top Web pages.

- **Data analysis:** In this step we will map the top Web pages to find similar and different ones within the three Jordanian universities.

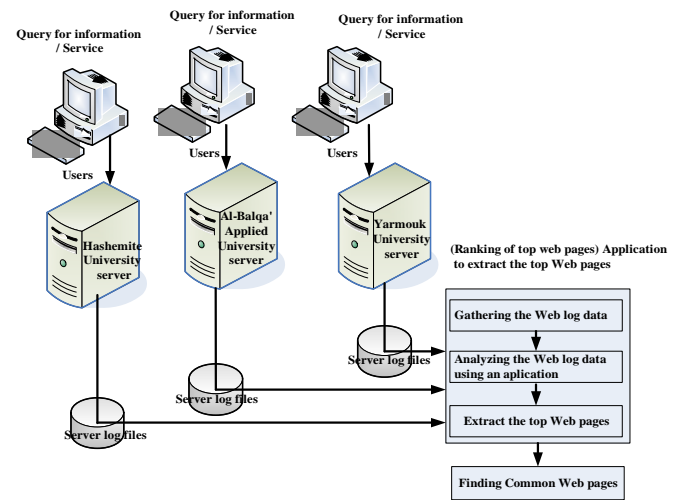


Figure 5. General steps of the approach

Algorithm: Extracting Top Web pages from Web log files

Input: Web Server Log File

Output: File_of_URLs

Step1: Read LogRecord from Web Server Log File

Step2: If ((LogRecord.url-stem (http://))
then

Insert LogRecord into a File_of_URLs
// ("URL") and ("URLFrequency").

Step3: Repeat the above two steps until end of file
(EOF) (Web Server Log File)

// that indicates that all records have been read from
the log files

Step 4: Stop the process.

Figure 6. The proposed algorithm to extract top Web pages

IV. EXPERIMENTS AND EVALUATION

The top URLs of three Jordanian universities were compared in order to find the common top Web pages and their percentages within these universities. Figure 7 presents a similarities and differences among top URLs in Al-Balqa' Applied and Hashemite Universities. As shown in this figure, the percentage of common top visited URLs equals to 19% of the 100 Web pages, while the percentage of differences is 81%.

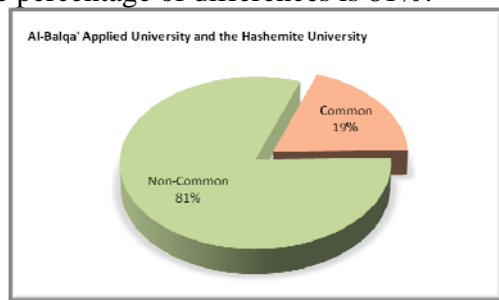


Figure 7. Similarities and differences among top URLs in Al-Balqa' Applied and Hashemite Universities

Similarities and differences among top URLs in Al-Balqa' Applied University and Yarmouk University are shown in Figure 8.

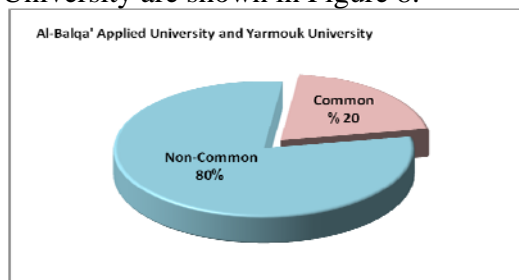


Figure 8. Similarities and differences among top URLs in Al-Balqa' Applied University and Yarmouk University

The pie chart in the above figure showed that percentage of overlapped URLs in Al-Balqa' Applied University and Yarmouk University equals 20% of the 100 Web pages, while the percentage of the differences is 80%. This indicates that the mutual interests

between the two university consists only 1/5, while differences consists 4/5.

Figure 9 shows the similarities and differences among top URLs in The Hashemite University and Yarmouk University.

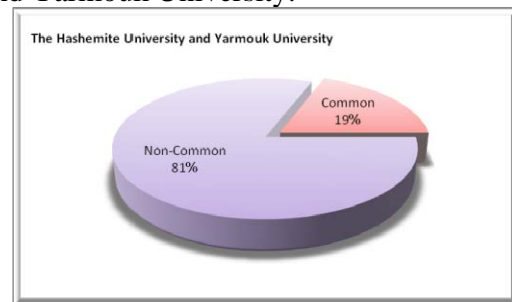


Figure 9. Similarities and differences among top URLs in The Hashemite University and Yarmouk University

Figure 9 showed that the percentage of overlapped top visited URLs equals to 19% of 100 Web pages, while the percentage of different top visited URLs equals to 81%.

Figure 10 shows the overall results of matching top visited URLs within the three universities within the same time duration is equal to 11% of top visited URLs were common between the three universities, while 71% of top visited URLs are not common and represents uniqueness of the URLs within these universities, and this indicate the peculiarities is larger than common interests. And the percentage of mutual interests between any two universities is 18%.

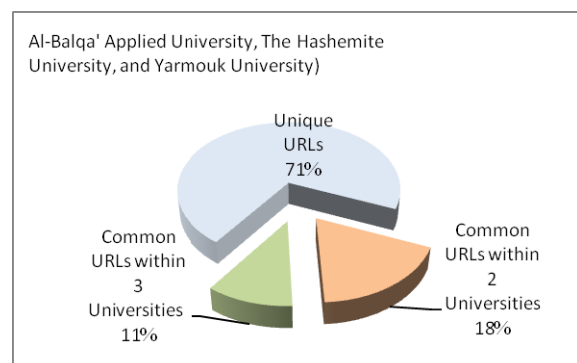


Figure 10. Similarities and differences among top URLs in three Universities

The reasons behind the results of the similarity and differences among top URLs in three Universities may be:

- The number of students in Al-Balqa university is approximately 45,000 students, distributed over 12 colleges nationwide, and the number of students in Yarmouk university is approximately 32,000 and the number of students in the Hashemite university is approximately 20,000 students.
- Most of the majors at the Hashemite university are scientific, so the interest of these majors definitely different from interest of majors related to arts, literature, sport, economics, ...etc., where the majority of students at Yamouk university affiliated within these majors.
- There is also variation in the speeds of the internet, which would make the search on the Web faster for students of Yarmouk that have speed of 155 Mega bits per second in contrast to the internet connection speed of 100 MB per second at Al-Balqa university, distributed over 12 college in various provinces of the Kingdom.
- The student search interests of different academic levels are widely varied. The post graduate students used to internet to look for subjects related to their courses and research. Also the age factor affects the search interests, and usually the average age of post graduate students is higher than average age of their bachelor's counterpart.
- Some colleges of Al-Balqa University are restricted to girls only like Ajloun and Irbid. The gender factor in this case affects the search interests.

V. EVALUATION RESULTS

Chi-square test for independence was applied in this study to examine statistically whether these differences are significant or not.

Problem

Are there any differences in search interests and Web usage within the three Jordanian universities?

Are the top Web pages differ significantly within the three Jordanian universities?

We used a 0.05 level of significance

Solution

It is important to keep in mind that the chi-square test for independence only tests whether two variables are independent or not, it cannot address questions of which is greater or less.

The following four steps are followed to solve this problem:

- (1) State the hypotheses
- (2) Formulate an analysis plan
- (3) Analyze sample data
- (4) Interpret results.

Table 2 shows how the top Web pages of the three Jordanian universities are related. Then we used the values in this table as observed values to apply Chi-square test for independence on Web usage in the three Jordanian universities. As shown in this table we see a total number of top Web pages is equal to 150 top Web pages, which distributed equally on the three universities, where 16 represents the number of common top web pages within the three Jordanian universities.

Table 2 Numbers of top Web pages within the three Jordanian universities.

	Three Jordanian universities			Row total
	BAU (Observed)	HU (Observed)	YU (Observed)	
BAU	12	16	16	44
HU	16	9	26	51
YU	8	26	21	55
Column total	36	51	63	150

Where BAU stands for Al-Balqa' Applied University, HU stands for The Hashemite University, and YU stands for Yarmouk University.

Now we have to apply the above four steps:

▪ **State the hypotheses.**

The first step is to state the null hypothesis (H_0) and an alternative hypothesis (H_1).

H_0 : Top Web pages and Web usage are independent.

H_1 : Top Web pages and Web usage are not independent (are related).

▪ **Formulate an analysis plan**

The level of significance being used is 0.05. Using a sample data, we will conduct a chi-square test for independence.

▪ **Analyze sample data.**

We calculate:

- Degrees of freedom: which is equal to the number of columns in the table minus one multiplied by the number of rows in the table minus one.

$$\begin{aligned}\text{Degrees of Freedom} &= (c - 1) \times (r - 1) \\ &= 2 \times 2 = 4\end{aligned}$$

- Expected frequency counts: the expected values have been calculated for each cell. The way to calculate the expected cell frequency is to multiply the column total for that cell, by the row total for that cell, and divide the product by the total number of observations for the whole table.

• **Chi-square test statistic**

Based on the chi-square statistic and the degrees of freedom, we determine the P-value as shown in Figure 11.

Chi-Square Test: C1; C2; C3

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

	C1	C2	C3	Total
1	12 10.56 0.196	16 14.96 0.072	16 18.48 0.333	44
2	16 12.24 1.155	9 17.34 4.011	26 21.42 0.979	51
3	8 13.20 2.048	26 18.70 2.850	21 23.10 0.191	55
Total	36	51	63	150

Chi-Sq = 11.836; DF = 4; P-Value = 0.019

Figure 11. Determine the P-value

• **Results Interpretation.**

Since the P-value (0.019) is less than the significance level (0.05), we cannot accept the null hypothesis (H_0). Thus, we conclude that there is a relationship between top Web pages and Web usage, and they are not independent.

VI. CONCLUSION

This paper focused on the discovering the Web usage within the 3 Jordanian universities (Al-Balqa' Applied University, Hashemite University, and Yarmouk University).

The results show differences in Web usage between these universities. Tests show that these differences are statistically significant

through the use of Chi-square test for independence.

REFERENCES

- [1] Mobasher, R. Colley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining", ACM, volume 48, 2000.
- [2] J. Pei, J. Han, B. Mortazavi-Asl, H. Zhu, "Mining access patterns the efficiently from web logs" in PADKK '00: Proceedings of the 4 Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. London, UK: Springer-Verlag, pp. 396-407, 2000.
- [3] T. Murata and K. Saito, "Extracting Users Interests from Web Log Data", Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, China, pp.343-346, 2006.
- [4] J. Borges, M. Levene, "A Fine Grained Heuristic to Capture Web Navigation Patterns," ACM SIGKDD Explorations, Vol.2, No.1, pp.40-50, 2000.
- [5] K. R. Suneetha, R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.
- [6] A. Spink and B.J. Jansen, "Web search: Public searching on the Web", Dordrecht: Kluwer Academic, 2004.
- [7] D. Rose, D. Levinson, "Understanding user goals in web search", ACM, Press, 13-19, 2004.
- [8] M. Sanderson, J. Kohler, "Analyzing geographic queries", in proceedings of the SIGIR workshop on geographic, 2004.