Web Usage Analysis in Six Arab States and Three Universities

Areej Adel Abd Alrahman AlQwaqenah*, Mohammed Al-Kabi** and Belal Abu Ata***

Received on Feb. 20, 2011 Accepted for publication on Oct. 19, 2011

Abstract

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 research project concerned with discovering the Web usage in six Arab countries (Jordan, Iraq, Lebanon, Palestinian territories, Saudi Arabia and Syria). Also the Web usage in 3 Jordanian universities (Al-Balqa' Applied University, Hashemite University, and Yarmouk University) is discovered through the analysis of the Web log files of their servers. The test results show significant differences in Web usage within the three Jordanian universities under consideration. Also the tests reveal that there are significant differences in Web usage within the six Arab states under consideration. The tests reveal that top Web sites and Web usage are related to each other and not independent.

Keywords: Search trends, Web usage analysis, Web server log files, Transactional log, Web log

Introduction

The last twenty years witnessed the emergence of the Internet and a boom in its usage. Web log files represent an electronic record of the interaction processes between different Web users from one side and the Web servers from the other side. Therefore many studies attempts to analyze and mine the Web log files within different Web servers, Web search engines and Internet Service Providers (ISPs) to discover the Web usage and search trends around the world or within certain cities or states. These studies can be beneficial in many fields such as education and business. Although the literature is full with many studies dedicated with Web usage analysis and mining, but there are very few studies related to the analysis and mining of the Web usage in the Arab world.

The Web is growing rapidly and the internet users are increased in the Middle East and North Africa, so there is a need to discover whether the Internet is used for
entertainment, games, education, business …etc. So there is a need to discover the top sites and top Web searches within these Arab countries.

Analyzing users’ Web log data and extracting their search terms is important and a little bit challenging task for Arabic terms. Web logs are maintained by Web servers and contain information about users accessing the site. 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 1 shows the basic items of a Web log file.

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

Figure 1: Single row from a Web log file

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 table1:

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

To discover the top Web searches there is a need to obtain the Web log files (search engine’s transactional logs) from each well known Web search engines such as Google, Bing, Yahoo, Ask, …etc, or we have to get Web log file from different ISPs. Kindly
note that Google as Web search engine publishes the top Web searches within different
countries through it famous service which is called Google Insights for Search.

Similarly discovering the top sites necessitate the need to obtain the Web log files
from different ISPs or get these top sites from specialized Web sites such as

The analysis of this study is based on a program which designed and implemented
in the computer center of Hashemite University to extract different Web queries whether
they are in Arabic or English from the Web log files. Other Software’s are used to
accomplish the task of analysis of the large amount of information Queries & Top Web
sites) stored within different files.

This study divided into a number of sections, where the next section presents
related studies to this one beginning with those related to the analysis of Web log files,
and then presenting some of the studies that present the mining of Web log files. Goals
and approaches section presents the two approaches used in this study and the goals of
this study. Then we present the experiments and evaluation section which shows the
details of the conducted experiments and their results. Finally we present the conclusion
and future work section.

Related Work

This section discusses the relevant background literature for this research study
highlighting related work done in the field of Web log analysis.

Analyzing Web Server Log

Research for analyzing Web log data has been done by many researchers in the
field of Web usage mining; while Pei et al. [4] 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. [5] 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 an analysis of the site-keyword graph, while Borges et al. [6] modeled users' to
capture Web navigation patterns.

Huiying et al. [7] and Tanasa et al. [8] are working on data preprocessing which
includes users' identification, session identification, path completion, transaction
identification. Koutsoupias et al. [9] analyzed the Web log file using statistical analysis
method and provided a tool and interpretation of the preprocessed statistical results
produced from Web log data.

Murate et al. [10] described a method for clarifying users' interests based on
analysis of site-keyword graphs generated from Web log data. Their method based on
the assumption that sites/keywords regarding a users interests appear in several different
transaction of Web log data.

Web Usage Analysis in Six Arab States and Three Universities
The study of Suneetha et al. [11] 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 improves the performance.

Xue et al. [12] proposed a novel re-ranking method based on user logs within Websites. Each page’s access frequency and the traversal patterns of information finding are extracted from web logs. The proposed method resolves the diversity problem of user's access behavior and discovers general patterns.

Spiliopoulou et al. [13] [14] proposed a "web utilization miner" (WUM) to find interesting navigation patterns. The interestingness criteria for navigation patterns are dynamically specified by the human expert using WUM's mining language which supports the specification of statistical, structural and textual criteria.

Spink et al. [15] analyzed the characteristics of general Web search logs from different perspectives: terms, queries, sessions, and result pages. Rose et al. [16] has works 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. [17] studied the popularity of query topics for geographic queries in general search. Their study indicates that geographically related queries are a significant sub-set of the queries submitted to a search engine and it differ from other queries.

**Web Mining**

Kosala et al. [18] defined a Web mining as the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web.

Dunham et al. [19] categorized Web mining into three active research areas according to what part of Web data is mined: content mining, structure mining, and Web usage mining. While Srivastava et al. [20] classified such data into several types: Content, Structure, Usage, and User Profile.

Several researches centered on content mining and usage mining. Baglioni et al. [21] introduced a project aiming to extract navigation behavior models of a site’s visitors. Davison et al. [22] proposed an approach to preloading Web pages into the local cache for a visitor. While the Liu et al. [23] presented a novel approach to classify user navigation patterns and predict users’ future requests by combining Web usage logs and content mining.

There are several researches in Web structure mining, where some of these presented algorithms that find hubs and authorities are described in the study of Slattery et al. [24] and of Chen et al. [25].
Goals and Approaches

In this study, two approaches were used to discover the Web usage and search interest.

First, discovering 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 of these Universities. These three universities were selected due to the availability of their log files and also all fields of study are also included in these universities.

Second, discovering the differences in search interests and Web usage within six Arab countries (Jordan, Iraq, Lebanon, Palestinian territories, Saudi Arabia and Syria) through gathering of the top sites for these countries, as ordered by Alexa.com. These countries were selected due to their geographical and cultural interconnection. These countries represent the other Arab countries in our opinion, so they represent a good data set sample to work on.

Overview of Methodology

The overall project methodology is graphically presented in figure 2 which consists of two major parts:

1. Web usage within three Jordanian universities
2. Web usage within Six Arab countries.

Each part consists of the following steps:

1. Data collection
2. Data preprocessing
3. Extracting top Web pages
4. Data analysis
5. Results and Conclusion
First Part: Web Usage within Three Jordanian Universities

First we start collecting the Web log files from the search samples from the three Jordanian universities (Al-Balqa' Applied University, Hashemite University, and Yarmouk University). Afterward the Web log files are refined to extract the required data for the analysis phase.

Figure 3 summarizes the following general steps to satisfy the main purposes to compare the Web usage within three Jordanian universities.

- Data collection: Gathering the Web log files from several servers of three Jordanian universities.
- Data preprocessing: This step includes revision, refinement, and cleaning the Web log files.
- 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.
Web Usage Analysis in Six Arab States and Three Universities

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

![Diagram](image)

**Figure 3:** General steps of the first part.

The proposed algorithm for extracting top Web pages is illustrated in figure 4. This algorithm is applied on a Web log files represented by set of URLs as discussed in the previous step in the data preprocessing phase. The idea of using this algorithm is to perform an iterative search over a Web server log files which collected from three Jordanian universities to extract different Web pages.
**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 4*: The proposed algorithm to extract top 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. After extracting all top Web pages from Web log files of the three Jordanian universities by using the ISA tool and application of ranking top Web pages. Data analysis includes comparing the top URLs of the three Jordanian universities to find the percentages of common top Web pages within the three Jordanian universities.

The results of this phase will be used to compare top Web pages of three Jordanian universities that are generated from applying the proposed algorithm discussed in the previous subsection of extracting top Web pages.

**Second Part: Web Usage within Six Arab Countries**

In the second part of our methodology, data is collected from Alexa.com and from Google Insights for search. Afterward the data is refined to be ready for the analysis. *Figure 5* summarizes the general steps to satisfy the main purposes to compare the Web usage within six Arab countries.
Data Collection

Data collection is divided into two steps. In the first step we collected on a daily bases the top searches for Jordan and five other Arab countries for October-2010. In the second step, the top sites for Jordan and the other five Arab countries were gathered as ordered by Alexa.com.

Google Insights for Search (IFS) [26] is a service by Google, it provide insights into the search terms people have been entering into the Google search engine. Google IFS reveals the top searches categorized by week, month, and year. Popular searches of other states can be obtained from IFS.

We collected data for six Arab countries (Jordan, Iraq, Lebanon, Palestinian territories, Saudi Arabia and Syria) on a daily bases during October – 2010.

Alexa top sites web service provides access to the lists of web sites ordered by Alexa traffic rank. Web sites are ranked in descending order according to the requests received from a certain country to each Web site. The maximum number of top Web sites displayed by Alexa is 100, and could be less depending on the traffic. Alexa; owned by Amazon.com; is a very powerful tool used to rank web site traffic. Alexa Rank is a measure of website popularity.

Data was collected during October 2010 by the authors from Alexa.com for six Arab states. Although the data by Alexa does not change on daily bases, but on monthly bases.
Data Preprocessing and Data Analysis

This step includes revision of the data by eliminating the words that are less important and not meaningfully. First of all we start manipulating the data collected from Google insights for search, these data needs revision through reweighting all top words, finding a rate of repetition of top words in each country and deleting the repetition of words. Afterward these words are compared to determine the percentage of similarity and difference.

The manipulation of data collected from Alexa.com is started after the manipulation of data collected from Google. The manipulation includes a preprocessing of the data by eliminating the description and the hyperlink of Web sites that are less important, in order to find rates of similarities and differences between six countries.

After getting top Web sites and top searches for six Arab countries, analysis of data is started by comparing the top Web sites and top searches to find the percentages of similarities and differences within these countries.

Experiments and Evaluation

Introduction

This section presents the experiments needed to accomplish the main objective of this study and analyzing the data collected from different sources. The data has to be preprocessed before the analysis process, then the results has to be displayed and interpreted in a logic way.

Evaluation Measure

This section focus on use of the chi-square for independence test [27] as statistical significance test. Chi-square test is applied when we have two categorical variables from a single population. It is used to determine whether there is a significant association (relationship) between the two variables of a sample. This test consists of four steps:

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

1- State the Hypotheses

Suppose that a variable A has r levels, and a Variable B has c levels. The null hypothesis states that knowing the level of Variable A does not help you predict the level of Variable B. That is, the variables are independent.

$H_0$ (Null hypothesis): Variable A and Variable B are independent.

$H_1$ (Alternative hypothesis): Variable A and Variable B are dependent.
The alternative hypothesis is that knowing the level of Variable A can help you predict the level of Variable B.

2- Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis ($H_0$). To discover whether there is a relationship between any two variables the following assumptions has to be followed:

- **Significance level.** Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- **Test method.** Use the chi-square test for independence to determine whether there is a significant relationship between two categorical variables.

3-Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the $P$-value associated with the test statistic.

- **Degrees of freedom.** The degrees of freedom ($DF$) is equal to:

$$DF = (r-1) \times (c-1)$$

Where $r$ is the number of levels for one categorical variable, and $c$ is the number of levels for the other categorical variable.

- **Expected frequencies.** The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute $r \times c$ expected frequencies, according to the following formula.

$$E_{r,c} = \frac{(nr \times nc)}{n}$$

where $E_{r,c}$ is the expected frequency count for level $r$ of Variable A and level $c$ of Variable B, $nr$ is the total number of sample observations at level $r$ of Variable A, $nc$ is the total number of sample observations at level $c$ of Variable B, and $n$ is the total sample size.

- **Test statistic.** The test statistic is a chi-square random variable ($\chi^2$) defined by the following equation.
\[ \chi^2 = \sum \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}} \]  

\[ \sum \frac{\chi^2}{E_{r,c}} \]  

Where \( O_{r,c} \) is the observed frequency count at level \( r \) of Variable \( A \) and level \( c \) of Variable \( B \), and \( E_{r,c} \) is the expected frequency count at level \( r \) of Variable \( A \) and level \( c \) of Variable \( B \).

- \( P \)-value. The \( P \)-value is the probability (assuming the null hypothesis is true) of observing a sample statistic as extreme as the test statistic.

4- Interpret Results

We will reject the null hypothesis. Typically, this involves comparing the \( P \)-value to the significance level, and rejecting the null hypothesis when the \( P \)-value is less than the significance level.

Data Preparation for Evaluation

The data of this study was collected mainly from three sources; Alexa.com (http://www.alex.com), Google Insights for Search (http://www.google.com/insights/search/#), and from three Jordanian universities namely: Al-Balqa' Applied University, Hashemite University, and Yarmouk University [28, 29, 30].

The collected data are used into two ways:

1. Web log files for three Jordanian universities (Al-Balqa' Applied University, Hashemite University, and Yarmouk University) was used for discovering the differences in search interests and Web usages within three Jordanian universities.
2. Collected top Web pages for six Arab countries (Jordan, Iraq, Lebanon, Palestinian territories, Saudi Arabia and Syria) and top searches for these countries were used for discovering the differences in search interests and Web usage within these Arab countries.

Discussion of Comparison Results

The top 50 URLs of three Jordanian universities were compared in order to find common top Web pages and their percentages within these universities.

Figure 6 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%, while the percentage of differences is 81%.
The pie chart in figure 7 showed that percentage of overlapped URLs in Al-Balqa' Applied University and Yarmouk University which is equal to 20%, while the percentage of the differences is 80%. This indicates that the mutual interests between the two universities restricted only to 20%, while difference is around 80%.

Figure 8 showed that the percentage of overlapped top visited URLs equals to 19%, while the percentage of different top visited URLs equals to 81%.
Figure 9: Similarities and differences among top URLs in three Universities

Figure 9 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%.

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

1. University size: The number of students in Al-Balqa university is approximately 45,000 students, distributed over 12 colleges distributed in different regions of Jordan, 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.

2. Major Variation: Most of the majors at the Hashemite university are scientific, so the interests of students enrolled in these majors definitely different from interests of students enrolled at Yarmouk university in majors related to arts, literature, sport, economics, …etc. Although there are three only scientific faculties (Science, Hijjawi for Engineering Technology and Information Technology) at Yarmouk university, the other nine faculties (Arts, Economics and Administrative Sciences, Shari’a and Islamic Studies, Education, Physical Education, Law, Fine Arts, Archaeology And Anthropology, and Mass Communication) are artistic. Therefore the majority of students at Yarmouk university are enrolled within artistic majors.

3. Internet Speed: The speed of the internet within these three universities varying from one to another. The speed of the internet at Yarmouk university is 155 Megabits per second, while it is 100 Megabits per second at Al-Balqa university, distributed over 12 colleges in various provinces of Jordan.

4. Search interests: The student search interests of different academic levels are widely varied. The post graduate students used the 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.

5. Gender: 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.

Figure 10 shows the similarities and differences among top URLs in the three universities and top Web sites in Jordan. As shown in this figure the percentage of common top visited Web pages within Jordan and the three universities is equal to 14%, while the percentage of the peculiarities is 86%.

**Figure 10:** Similarities and differences among top visited Web sites between Jordan and three Universities

**Top Searches Comparison within Six Arab Countries**

Figure 11 shows the similarities and differences among top searches in Iraq and Jordan.

**Figure 11:** Similarities and differences among top searches in Iraq and Jordan

**Figure 12:** Similarities and differences among top searches in Lebanon and Jordan
As shown in figure 11, the percentage of common top searches is equal to 3%, while the percentage of differences is 97%. Figure 12 shows the similarities and differences among top searches in Lebanon and Jordan. As shown in this figure, the percentage of common top searches is equal to 13%, while the percentage of differences is 87%.

As shown in figure 13, the percentage of common top searches between Palestine and Jordan equals to 23%, while the percentage of differences is 77%. Figure 14 shows the similarities and differences among top searches in Saudi Arabia and Jordan. As shown in figure 14, the percentage of common top searches is equal to 16%, while the percentage of differences is 84%.

As shown in figure 15, the percentage of common top searches among top searches in Six Arab countries.
Figure 15 shows the similarities and differences among top searches in Syria and Jordan. As shown in this figure the percentage of common top searches is equal to 20%, while the percentage of differences is 80%.

Figure 16 shows the similarities and differences among top searches in six Arab countries. As shown in figure 16, the percentage of common top searches within six Arab countries is equal to 1%. While 62% of top searches are not common within these countries, and this indicate the peculiarities is larger than common interests. Also the percentage of mutual interests between any two countries is 20%, while the percentage of mutual interests between any three countries is 9%, and while the percentage of mutual interests between any four countries is 5%, and the percentage of mutual interests between any five countries is 3%.

The test results show the similarity of top interests between Iraq and Jordan is 3%, while the similarity of top interests between Jordan and Lebanon is 13%, and with Saudi Arabia is16%, and with Syria is 20%. The greatest common similarity of top Web searches is 23% between Jordan and Palestinian territories.

**Top Web Sites Comparison within Six Arab Countries**

Figure 17 shows the similarities and differences among top Web sites in Jordan and Iraq. As shown in this figure, the percentage of common top Web sites equals to 20%, while the percentage of differences is 80%. Figure 18 shows the similarities and differences among top Web sites in Jordan and Lebanon.
As shown in figure 18, the percentage of common top Web sites between Jordanian internet users and Lebanese internet users is equal to 16%, while the percentage of differences is equal to 84%.

Figure 19: Similarities and differences among top Web sites in Jordan and Palestine

Figure 20: Similarities and differences among top Web sites in Jordan and Saudi Arabia

Figure 19 shows the similarities and differences among top Web sites between Jordanian internet users and Palestinian internet users. As shown in this figure, the percentage of common top Web sites is equal to 20%, while the percentage of differences is 80%. Figure 20 shows the similarities and differences among top Web sites between Jordanian internet users and Saudis internet users. As shown in this figure, the percentage of common top Web sites is equal to 17%, while the percentage of differences is 83%.

Figure 21: Similarities and differences among top Web sites in Jordan and Syria

Figure 22: Similarities and differences among top Web sites between Six Arab countries
Figure 21 shows the similarities and differences of top Web sites among Jordanian internet users and Syrian internet users. As shown in figure 21, the percentage of common top Web sites equals to 18%, while the percentage of differences is 82%. Figure 22 shows the similarities and differences among top Web sites between Six Arab countries, where the percentage of common top Web sites is equal to 7%, while the percentage of differences is 93%.

**Results Evaluation**

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

The next subsection presents evaluation results of three Jordanian universities, and shows how we applied Chi-square test for independence on three Jordanian universities. Evaluation results of six Arab countries subsection presents how we applied Chi-square test for independence on six Arab countries to examine statistically whether these differences are significant or not.

**Evaluation Results of 3 Jordanian Universities**

Problem 1

- 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 four steps mentioned previously in the "Evaluation Measure" subsection are followed to solve this problem.

Table 2 shows how the top Web pages of the three Jordanian universities are related. The values in this table are the observed values to be used by Chi-square test for independence on Web usage within the three Jordanian universities under consideration. This table presents the total number of top Web pages is equal to 150 top Web pages, which distributed in 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

<table>
<thead>
<tr>
<th></th>
<th>Three Jordanian universities</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BAU (Observed)</td>
<td>HU (Observed)</td>
<td>YU (Observed)</td>
</tr>
<tr>
<td>BAU</td>
<td>12</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>HU</td>
<td>16</td>
<td>9</td>
<td>26</td>
</tr>
<tr>
<td>YU</td>
<td>8</td>
<td>26</td>
<td>21</td>
</tr>
<tr>
<td>Column total</td>
<td>36</td>
<td>51</td>
<td>63</td>
</tr>
</tbody>
</table>

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 dependent (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 used a Minitab statistical analysis software to 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.

    Degrees of Freedom = $(c - 1)(r - 1) = 2 \times 2 = 4$

  - 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. Figure 23 shows the results of Minitab program.
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 related to each other.

Evaluation Results of Six Arab Countries

Problem 2

Are there any differences in search interests and Web usage within six Arab countries?

Are the top Web sites differ significantly within the within six Arab countries?

We used a 0.05 level of significance

Solution

Table 3 presents how the top Web sites of six Arab countries are related. The values in this table are the observed values to be used by Chi-square test for independence on Web usage within the six Arab countries.
Al-Qwaqenah, Al-Kabi and Abu Ata

Table 3: Numbers of top Web sites within six Arab countries

<table>
<thead>
<tr>
<th></th>
<th>Jordan (Observed)</th>
<th>Iraq (Observed)</th>
<th>Lebanon (Observed)</th>
<th>Palestinian (Observed)</th>
<th>Saudi (Observed)</th>
<th>Syrian (Observed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jordan</td>
<td>12</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>Iraq</td>
<td>23</td>
<td>43</td>
<td>34</td>
<td>55</td>
<td>49</td>
<td>50</td>
</tr>
<tr>
<td>Lebanon</td>
<td>23</td>
<td>34</td>
<td>35</td>
<td>34</td>
<td>45</td>
<td>34</td>
</tr>
<tr>
<td>Palestinian</td>
<td>23</td>
<td>55</td>
<td>34</td>
<td>30</td>
<td>20</td>
<td>43</td>
</tr>
<tr>
<td>Saudi</td>
<td>23</td>
<td>49</td>
<td>45</td>
<td>20</td>
<td>30</td>
<td>39</td>
</tr>
<tr>
<td>Syrian</td>
<td>23</td>
<td>50</td>
<td>30</td>
<td>43</td>
<td>39</td>
<td>33</td>
</tr>
</tbody>
</table>

The four steps mentioned previously in the "Evaluation Measure" subsection are followed to solve this problem.

- **State the hypotheses.** The first step is to state the null hypothesis ($H_0$) and an alternative hypothesis ($H_1$).
  
  $H_0$: Top Web sites and Web usage are independent.
  
  $H_1$: Top Web sites and Web usage are dependent (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 used Minitab statistical analysis software to 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.
  
  - Expected frequency counts: the expected values have been calculated for each cell. 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. Figure 24 shows the results of Minitab program.
Chi-Square Test: Jordan; Iraq; Lebanon; Palestinian; Saudi Arabia; Syrian

<table>
<thead>
<tr>
<th></th>
<th>Jordan</th>
<th>Iraq</th>
<th>Lebanon</th>
<th>Palestinian</th>
<th>Arabia</th>
<th>Syrian</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>127</td>
</tr>
<tr>
<td></td>
<td>0.122</td>
<td>0.475</td>
<td>0.188</td>
<td>0.115</td>
<td>0.100</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>43</td>
<td>34</td>
<td>55</td>
<td>49</td>
<td>50</td>
<td>254</td>
</tr>
<tr>
<td></td>
<td>26.55</td>
<td>53.10</td>
<td>42.02</td>
<td>42.86</td>
<td>43.07</td>
<td>46.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.475</td>
<td>1.921</td>
<td>1.531</td>
<td>3.441</td>
<td>0.818</td>
<td>0.278</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>34</td>
<td>35</td>
<td>34</td>
<td>45</td>
<td>34</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>21.43</td>
<td>42.86</td>
<td>33.91</td>
<td>34.59</td>
<td>34.76</td>
<td>37.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.115</td>
<td>1.830</td>
<td>0.035</td>
<td>0.010</td>
<td>3.019</td>
<td>0.519</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>55</td>
<td>34</td>
<td>30</td>
<td>20</td>
<td>43</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>21.43</td>
<td>42.86</td>
<td>33.91</td>
<td>34.59</td>
<td>34.76</td>
<td>37.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.115</td>
<td>3.441</td>
<td>0.000</td>
<td>0.010</td>
<td>6.266</td>
<td>0.820</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>49</td>
<td>45</td>
<td>20</td>
<td>30</td>
<td>39</td>
<td>206</td>
</tr>
<tr>
<td></td>
<td>21.53</td>
<td>43.07</td>
<td>34.99</td>
<td>34.76</td>
<td>34.58</td>
<td>37.64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.100</td>
<td>0.818</td>
<td>3.500</td>
<td>6.266</td>
<td>0.695</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>50</td>
<td>30</td>
<td>43</td>
<td>39</td>
<td>33</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td>22.79</td>
<td>45.57</td>
<td>36.96</td>
<td>36.78</td>
<td>36.96</td>
<td>39.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.092</td>
<td>0.430</td>
<td>1.020</td>
<td>1.051</td>
<td>0.112</td>
<td>1.172</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>127</td>
<td>254</td>
<td>201</td>
<td>205</td>
<td>206</td>
<td>222</td>
<td>1215</td>
</tr>
</tbody>
</table>

Chi-Sq = 41.259; DF = 25; P-Value = 0.022

Figure 24: Minitab program to solve problem 2

- Results Interpretation.

Since the $P$-value (0.022) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between top Web pages and Web usage, and they are related to each other.

Conclusion and Future Work

Conclusion

This study divided in two parts. In the first part of this study an exploration of the Web usage within three Jordanian universities (Al-Balqa' Applied University, Hashemite University, and Yarmouk University) is presented. The results of this exploration reveal a significant statistical differences in Web usage among the academic communities of these universities. Second part of this study an exploration of Web usage within six Arab countries (Jordan, Iraq, Lebanon, Palestinian territories, Saudi Arabia and Syria) is presented. The test results of this exploration reveal significant statistical differences in Web usage throughout these six states.

Also the researchers of this study conducted a test which reveals that top Web sites and Web usage are related to each other and not independent.
Future Work:

Many issues regarding this study may be addressed for the future, such as:

- Discovering the Web usage in all Arab countries
- Discovering the Web usage in all Arab and Jordanian Universities

Also future studies could consider navigational behavior, user interests, and other recommendations that if followed would discover the Web usage of all Arab countries and all Jordanian Universities.

تحليل استخدام الإنترنت في ستة دول عربية وثلاث من الجامعات

أريج قواقنة ومحمد الكببي وِلال أبوالعطا

ملخص

تعتبر عملية تحليل البيانات الموجودة ضمن ملفات حواسيب الخادم المعروفة باسم files مهمة. ويتطلب لها من قبل الباحثين على أنها تمثل تحديثاً ضمن حقل التنقيب عن استخدام الإنترنت. تعتبر عملية تحليل هذا النوع من الملفات من التقنيات المهمة للكشف عن سلوك واهتمامات مستخدمي شبكة الإنترنت. تتناول هذا البحث في الجزء الأول من عملية التحليل والكشف عن استخدام شبكة الإنترنت في ثلاث من الجامعات الأردنية (جامعة البلقاء التطبيقية، الجامعة الهاشمية، وجامعة اليرموك). أما الجزء الثاني من هذا البحث فتناول تحليل البيانات الخاصة باستخدام شبكة الإنترنت في الأردن والدول العربية الأربعة (فلسطين، العراق، سوريا، والمملكة العربية السعودية). المجاورة له إضافة إلى لبنان. يهدف التعرف على سلوك واهتمامات مستخدمي شبكة الإنترنت في هذه الدول الست. بينت النتائج التي خلص إليها هذا البحث وجود فروق معنوية من الناحية الإحصائية بين استخدامات الإنترنت في الدول العربية الست موضوع البحث، كما تبين وجود فروق معنوية من الناحية الإحصائية بين استخدام شبكة الإنترنت ضمن الجامعات الأردنية الثلاث موضوع البحث. وبناء الإختبارات هنالك علاقة وثيقة ما بين صفحات الإنترنت المفضلة لدى مستخدمي الإنترنت واستخدامات الإنترنت في الدول العربية الست التي تناولتها هذه الدراسة.
Web Usage Analysis in Six Arab States and Three Universities

References


