International Journal of Information Science and Management

Evaluating Function of Persian Search Engines on the Web Using Correspondence Analysis

M. A. Erfanmanesh, Ph. D. Student
UM University, Malaysia
email: Maerfan@perdana.um.edu.my

F. Didegah, M. S.
Shiraz University, I. R. of Iran
Corresponding Author: fdidgah@gmail.com

Abstract

This study aimed to evaluate Persian Search Engines based on the criteria obtained from Alexa databank using correspondence analysis. Through searching the web, 23 search engines were found, which due to their cease and no coverage by Alexa, this number was reduced to 16. Data analysis revealed that Ghatreh Search Engine occupied a high rank in most of attributes like traffic rank, time spent on site and number of pages viewed per user. Iranmania Search Engine attained the largest number of links among other search engines. Jostejoogar Search Engine attracted the largest number of foreign users whilst Begardim Search Engine had no foreign users. Correspondence analysis classified the 16 search engines into three groupings which were related on the basis of certain attributes.

Keywords: Persian Search Engines, Alexa Criteria, Correspondence Analysis.

Introduction

There are a whole host of websites on the Internet. No one can give an exact number for the size of the Web. Many surveys and studies have been conducted regarding the dramatic growth of the Internet. One study indicated that the number of hosts has been roughly doubling every year (Kobayashi & Takeda, 2000).

Considering the large amount of organized and unorganized information on the internet, some ways should be applied to improve and facilitate users' information exploration in the cyberspace. The most frequently used and powerful way of finding information on the internet is using search engines. A web search engine is an information retrieval system which is used to locate the web pages relevant to user queries. It is not surprising that the internet users are increasingly utilizing search engines to satisfy their information needs. About 85% of the Web users surveyed claimed to be using search engines to find specific information of interest (Kobayashi & Takeda, 2000).

Although general popular web search engines such as Google and Altavista are getting easier to use, sometimes locating relevant information, especially local information, is still difficult. A growing number of countries are beginning to develop their own search engines

to facilitate the search for local content. The present study applies the correspondence analysis method to evaluate Persian Search Engines based on data obtained from Alexa Website.

Criteria for the Evaluation of Search Engines Sites based on Alexa Databank

Alexa Internet was first established in 1996 by Brewster Kahle and Bruce Gilliat (Alexa Internet, 2009). It collects constantly all types of information from websites and offers free-of-charge evaluation services. Alexa computes traffic rankings by analyzing the Web usage of millions of Alexa Toolbar users and data obtained from other diverse traffic data sources. The traffic rank is a measure of a website's popularity. The rank is calculated using a combination of average daily visitors and page views over the past three months. The site with the highest combination of visitors and page views is ranked first. The websites with less than thousand monthly visits are not considered as a statistical sample. Generally, the traffic ranking after 100'000 is not reliable due to a shortage of data and so is of no statistical significance.

Average time on site is another factor which is a measure of user attention and includes the average minutes a user spent per day on the site over the past three months.

Number of pages visited per user is also calculated by Alexa which shows quality and variety of information on the site.

Another measure is the number of links which a website receives from other websites and shows its reputation. In addition to these factors, Alexa Internet offers information about percentage of people who visit a website (National and International visitors). As it is inferred from Alexa indexes' definitions, these indexes are mostly related to users' preference which means if a website has a proper function based on these indexes, it is more popular and favored by web users.

Research Objectives

This study seeks to examine the web performance and function of Persian search engines based on six criteria, i.e., traffic rank, average number of page views by users, spent time on the site per user, number of links received from other websites and percentage of Iranian and foreign visitors. The search engines will be grouped based on their correlation, and their strong and weak points will be reviewed.

Research Questions

The current study is concerned with answering the following two research questions:

1. Based on the six factors, which search engines function better than the others?
2. Using correspondence analysis, how many groups could Persian search engines be categorized into, and what are the strong and weak points of each group?

Review of Literature

Though many studies have been conducted on search engines and also several local search engines in Iran have emerged, not much research has been done on Persian web search engines. All the comparative and evaluative studies conducted and published to date mainly involved the already well-known search engines such as Google, Altavista, Infoseek and Excite. Davis (1996) provides an extensive review on the comparison of seven search engines: Altavista, Hotbot, Infoseek, Excite, Lycos, Open text and WebCrawler. The comparisons were based on the search engines features and characteristics.

Chu and Rosenthal (1996) compared and evaluated three web search engines, namely Altavista, Excite and Lycos. The comparison and evaluation were in terms of their search capabilities and retrieval performance. At the end of the study, the authors reasoned out the superiority of Altavista, and proposed a methodology for evaluating other web search engines.

C Net, a company specialized in evaluating online products and services, published the findings of a comparative study of 15 web search engines. The search engines were tested on their accuracy of results, ease of use and provision of advanced options using 15 queries specifically composed for the evaluation. Most of the queries resembled reference questions asked in public libraries. According to the two feature tables generated by the evaluation, Altavista proved to be the best choice among individual search engines (Leonard, 1996).

Schlichting and Nilsen (1997) examined Altavista, Excite, Infoseek and Lycos. They conducted a small empirical study (with five participants) and used signal detection analysis to analyze the data.

The study reported by Boltuk (2000) compared six major web search engines namely Altavista, Excite, Go, Google, Hotbot, and Lycos. The author's focus was on features of these six search engines. The evaluation criteria used in this study were search restrictors, result display, subject directory and other search features.

Sutachun (2000) in his thesis compared five search engines: Altavista, Excite, Hotbot, Lycos and Infoseek in terms of search features and retrieval performance. He also applied the method in evaluating the effectiveness of each engine.

Evaluating Persian Search Engines, Shakeri (2008) investigated Recall and Precision of ten Persian search tools in retrieving Library and Information Science information. The results of her study showed that Webgah, Dahio and Persian Google had the highest rate of Recall and Precision among others.

The use of correspondence analysis in webometric studies and also investigating websites based on Alexa Internet data are recent developments. In two studies, Berthon et al. (1997) and Berthon, Pitt, Ewing, Ramaseshan and Jayaratna (2001) evaluated some industrial and telecom websites respectively using correspondence analysis and claimed that this statistical method could be used in webometric studies.

Using correspondence analysis, Shen, Li and Shen (2006) conducted a study to evaluate fifteen university library websites. They applied six indexes as library website evaluation criteria: traffic, visits, connectivity, speed, page views, and freshness. Using this method, they classified websites into three groupings with their specific attributes. The study has identified the relations among these websites, their strong and weak points and has offered suggestions as how to construct university library websites.

Method

The present study has been employed by using webometrics methods. Also, using correspondence analysis, we have analyzed the attributes of the websites. Persian search engines were chosen as population of the present research. In order to identify these search engines and extract their URLs, some widely known search engines such as Search Engine Colossus and Search Engine Guide, were visited. In this search, 23 Persian search engines were found which due to their cease and no coverage by Alexa, this number was reduced to 16.

The list of the surveyed search engines' websites is provided in Table 1. Six criteria including traffic, links, page views, spent time on site per user and Iranian and foreign visitors were selected in this survey to be analyzed by correspondence analysis method.

The counts of criteria per search engine were extracted from Alexa Website (www.alexa.com) in the first week of April 2009 and conducted within the same time (April) in order to decrease errors commonly associated with frequent website updates.

For this purpose, each search engine URL was searched in Alexa databank and the proper pages were downloaded. Then the necessary data were extracted from downloaded pages and entered into Microsoft Excel software.

Correspondence analysis was used to analyze data. Correspondence analysis is primarily a technique for representing the rows and columns of a two way contingency table in a joint plot. The technique is particularly useful for tables with large numbers of levels where deriving useful information from the table can be difficult (JMP, 2005). Greenacre (2007) believes that correspondence analysis is a generalization of a simple Graphical concept with which we are all familiar namely, the scatterplot.

The software used in this evaluation is JMP package of SAS (Statistical Analysis System). This computer program dynamically links statistics with graphics to interactively explore, understand, and visualize data.

Since traffic rank and links numbers were very large, in order to give a clear visualized Graph, data provided in Table 2 were grouped based on the principles in Table 3. Afterwards, the statistical data of Table 4 were entered into the JMP data analysis tool and analyzed.

Table 1

Surveyed Persian Search Engines

Row	Search engine	URL
1	webgah	www.webgah.com
2	jamasp	www.jamasp.com
3	irpars	www.irpars.com
4	parseek	www.parseek.com
5	jostejoogar	www.jostejoogar.com
6	today	www.today.ir
7	rismoon	www.rismoon.com
8	dahio	www.dahio.com
9	iranmehr	www.iranmehr.com
10	iranmania	www.iranmania.com
11	persiangoogle	www.persiangoogle.com
12	maibosearch	www.maibosearch.com
13	ghatreh	www.ghatreh.com
14	begardim	www.begardim.com
15	jasjoo	www.jasjoo.com
16	behjoo	www. behjoo.com

Table 2

Original Counts per criteria obtained from Alexa

Search engine	Traffic Rank	Average time on site (min/day)	Sites linking in	Users from Iran (%)	Users from overseas (%)	Pageview /User
webgah	1,605,810	1	29	72.6	27.4	3
jamasp	109,224	2.3	233	95.9	4.1	3.6
irpars	97,933	2.6	76	86.9	13.1	5.2
parseek	5,400	3.4	587	91.9	8.1	3.99
jostejoogar	489,548	6.5	30	46.5	53.5	5.7
today	486,664	2.5	34	94.3	5.7	3.4
rismoon	59,744	3	111	97.9	2.1	2.03
dahio	181,099	2.2	896	88.8	11.2	2.7
iranmehr	100,933	2	99	85.9	14.1	1.5
iranmania	13,012	3.1	1009	87.9	12.1	3.09
persiangoogle	1,203,681	0.7	14	82.0	18.0	1.1
maibosearch	32,366	2.2	262	91.5	8.5	2.08
ghatreh	4,939	6.3	288	93.1	6.9	5.02
begardim	1,657,122	1.3	6	100.0	0.0	1.7
jasjoo	24,519	2.9	39	95.3	4.7	4.2
behjoo	407,201	2	17	60.4	39.6	1.1

Table 3

Grouping Criteria into Quantitative Data

criteria	Grouping (Quantitative Data)
Traffic Rank	First divided by 1000 below 250= 5 250-500=4 500-750=3 750-1000=2 More than 1000=1
Page views	Rounded off
Time on site	Rounded off
Links	First divided by 10 and then rounded off
Iranian Users	First divided by 10 and then rounded off
Foreign Users	First divided by 10 and then rounded off

Table 4

Data obtained based on grouping

Search engine	Traffic Rank	Average time on site (min/day)	Sites linking in	Users from Iran (%)	Users from overseas (%)	Pageview /User
webgah	1	1	3	7	3	3
jamasp	5	2	23	10	0	4
irpars	5	3	8	9	1	5
parseek	5	3	59	9	1	4
jostejoogar	4	7	3	5	5	6
today	4	3	3	9	1	3
rismoon	5	3	11	10	0	2
dahio	5	2	90	9	1	3
iranmehr	5	2	10	9	1	2
iranmania	5	3	101	9	1	3
persiangoogle	1	1	1	8	2	1
maibosearch	5	2	26	9	1	2
ghatreh	5	6	29	9	1	5
begardim	1	1	1	10	0	2
jasjoo	5	3	4	10	0	4
behjoo	4	2	2	6	4	1

Results and Discussion

The first research question was:

1. Based on the six criteria, which search engines have function better than the others?

In order to answer this question, data provided in Table 2 were analyzed. Based on

data offered in this table, Ghatreh and Parseek Search Engines ranked respectively first and second in traffic factor, while Begardim, Webgah and Persiangoogle Search Engines have the lowest traffic rank among others.

Ghatreh Search Engine has functioned well in two other factors, time spent on site and number of pages viewed per user. Also, this search engine has received adequate number of links occupying rank four. Meanwhile, Iranmania Search Engine has received the largest number of links from other websites (1009) standing first. Regarding this factor, Dahio and Parseek Search Engines with 896 and 587 links received stand second and third, respectively (Table 2).

A close look at two columns, traffic rank and links counts, reveals a quite significant correlation between them. The result of Spearman correlation between these two columns is the proof (Table 5). Considering the fact that the best websites in traffic are those with the lowest numbered rank, a negative correlation between these two indexes is expected. As it is clear, search engines with a large number of links have ranked top in traffic which shows the importance of linking in cyberspace.

Table 5

Spearman Correlation Coefficient between Traffic Rank and Link number

			Traffic Rank	Link
Spearman's rho	Traffic Rank	Correlation Coefficient	1.000	-.788**
		Sig. (2-tailed)	.	.000
		N	16	16
	Link	Correlation Coefficient	-.788**	1.000
		Sig. (2-tailed)	.000	.
		N	16	16

**. Correlation is significant at the 0.01 level (2-tailed).

Considering number of visitors from overseas and from home, the results show that about half of Jostejoogar Search Engine's visitors came from Austria. After that, Behjoo and Webgah Search Engines have the largest number of foreign visitors (about 40 percent). Of the surveyed search engines, Begardim Search Engine has attracted no foreign visitors and is just visited in Iran (Table 2).

Another attribute for evaluating search engines is average number of pages viewed by users. Offering relevant information and variety of services persuades users to continue searching and browsing the links indexed by a search engine. Hence, larger number of pages viewed by users shows quality of services provided by a search engine. Among

surveyed search engines, Jostejoogar has the largest number of viewed pages. After that, Irpars and Ghatreh Search Engines come second and third, respectively (Table 2).

On the whole, none of the search engines have ranked top in all factors. Although Ghatreh Search Engine has shown a high rank in most of attributes, it has not functioned well regarding foreign visitors. Among all search engines, Begardim and Persiangoogle have shown weaker functions in most criteria.

The second research question was:

2. Using correspondence analysis, how many groups could Persian search engines be categorized into and what are the strong and weak points of each?

The results of Correspondence Analysis appear in Table 6 through 8 and Figure 1. The Chi-square test in Table 6 is highly significant, which means the surveyed factors and variables are not independent and the results are accurate. Tables 7 and 8 also are representing factors and variables coordinate.

As shown in Figure 1, the 16 search engines are classified into three groups; the first grouping consists of 6 search engines, including Iranmania, Dahio, Parseek, Ghatreh, Maibosearch, Jamasp whose strong points lie in linking. As it is obvious, based on data in Table 2, Iranmania, Dahio and Parseek Search Engines have the highest received links and that is why they are nearer to the link attribute in the figure. The other three come in ranks four, five and six based on link factor, and have functioned quite well in some other factors like traffic rank and page views. That is why they are located a bit far from link factor and have a tendency to the second group. On the whole, search engines of this group have functioned fairly well on the web.

The second group includes Rismoon, Iranmehr, Irpars, Jasjoo, and Today. These search engines have functioned similarly in four factors, traffic rank, page views, spent time on site and Iranian users. Hence, they are grouped together located nearer to the center of the plot which shows that overall performance of these search engines is in a moderate condition. Therefore, these search engines are in need of improvements to occupy their proper position on the web.

The third group includes 3 search engines, Webgah, Behjoo and Jostejoogar. Their strength lies in foreign visitors.

Besides these three groups, there are two more websites, Begardim and Persiangoogle Search Engines. As it was implied before, these search engines have functioned weakly in most factors. Hence, due to lower numbers and counts in all factors, they are not associated to any factor and not grouped with other search engines.

Table 6

Chi-Square Test

Test	Chi-Square	Prob>ChiSq
Likelihood Ratio	273.970	0.000
Pearson	293.196	0.000

Table 7

Criteria Coordinates

criteria	c1	c2	c3
Fusers	1.1165	1.1446	-0.489
Iusers	0.5972	-0.3178	-0.1723
Links	-0.4923	-0.4923	-0.0217
Traffic	0.4399	-0.0988	0.1729
pageviews	0.6055	0.0216	0.2705
time/min	0.5072	0.3261	0.4836

Table 8

Websites Coordinates

Search Engines	c1	c2	c3
begardim	0.9254	-0.7272	-0.7272
Behjoo	0.9991	0.6018	-0.3122
Dahio	-0.5417	0.0474	-0.0658
ghatreh	0.0160	0.0537	0.2394
iranmania	-0.5616	0.0637	-0.0485
iranmehr	0.3561	-0.2061	-0.0704
Irpars	0.5328	-0.151	0.1740
Jamasp	0.0117	-0.1897	0.0846
Jasjoo	0.7007	-0.4257	0.2373
jostejoogar	0.9732	0.7746	0.3419
maibosearch	-0.0761	-0.0578	-0.0275
parseek	-0.3683	0.0344	0.0052
persiangoogle	0.9180	-0.0722	-0.7509
rismoon	0.3218	-0.2814	0.1551
Today	0.7874	-0.2934	-0.0301
webgah	0.8902	0.3366	-0.3791

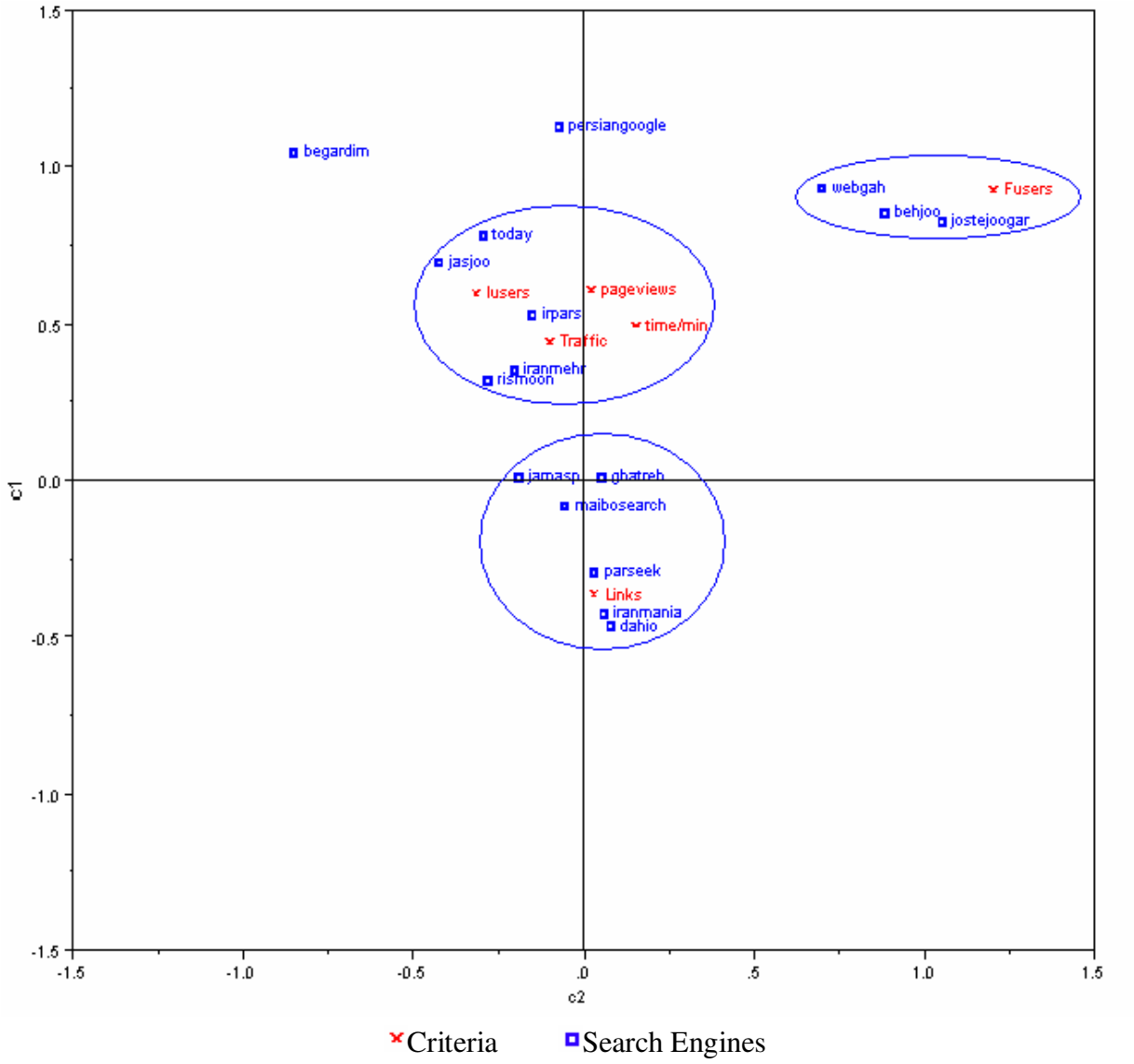

Figure 1. Coordinates.

Conclusion

The present research findings provide an evaluation of Persian Search Engines status in terms of their functions and performance on the web. Results show that most of the surveyed Search Engines do not act successfully on the web and require improvement. The investigated search engines had a weak function in traffic rank and most of them were ranked top of 100'000. Also, the short time each user spent on searching via these tools and the small number of pages viewed reveals that surveyed search engines were not successful enough in attracting users.

Considering the importance of local search engines as proper tools to browse and find local information within a country, the authorities in charge and search engine designers should pay adequate attention to and take due care for the betterment of their services in

order to remove the problems and gain satisfaction of the users.

References

Alexa Internet (2009). *Alexa History*. Retrieved February 3, 2009, from www.alexa.com.

Berthon, P. et al. (1997). Mapping the marketspace: Evaluating industry web sites using correspondence analysis. *Journal of Strategic Marketing*, *5*(4), 233-242.

Berthon, P., Pitt, L., Ewing, M., Ramaseshan, B., & Jayaratna, N. (2001). Positioning in cyberspace: Evaluating telecom web sites using correspondence Analysis. *Information Resources Management Journal*, *14* (1), 13 –21.

Boltuk, D. (2000). *Update to search engines compared*. Washington: Catholic University of America.

Chu, H. & Rosenthal, M. (1996). *Search engines for the world wide web: A comparative and evaluation methodology*. ASIS 96 annual conference proceedings. Retrieved February 3, 2009, from http://www.asis.org/annual-96/electronic procedings/chu.html.

Davis, E. T. (1996). *A comparison of seven search engines*. Retrieved February 3, 2009, from http://www.iwaynet.net/~lsci/search/paper-only.html.

Greenacre, M. (2007). *Correspondence analysis in practice*. Boca Raton: Taylor & Francis Group.

JMP (2005). *JMP statistics and graphics guide*. Release 6. Carey, NC: SAS.

Kobayashi, M. & Takeda, K. (2000). Information retrieval on the web. *ACM Computing Survey*, *32*(2), 144-173.

Leonard, A. J. (1996). *Where to find anything on the net?* Retrieved February 3, 2009, from http://www.cnet.com/ Content/Reviews/Search.

Schlichting, C. & Nilsen, E. (1997). *Signal detection analysis of www search engines*. Proceedings of the second human factors on the web conference. Retrieved February 3, 2009, from http://www.microsoft.com/usability/webconf/schlichting/schlichting .htm.

Shakeri, S. (2008). Rate of recall and precision of Persian web search tools in retrieving LIS informtion. *Faslname-Ketab*, *73*, 177-200.

Shen, X., Li, D. & Shen C. (2006). Evaluating china's university library web sites using correspondence analysis. *Journal of the American Society for Information Science and Technology*, *57*(4), 493-500.

Sutachun, T. (2000). *A comparative study of internet search engines*. Retrieved February 3, 2009, from http://websis.kku.sc.th/abstract/thesis/mart/lis/2543/lis430009t. html.