Cybrarians Journal
ISSN 1687-2215
دورية إلكترونية فصلية محكمة متخصصة في مجال المكتبات والمعلومات

العدد 35، سبتمبر 2014

A Comparison between seven open and Free Search Engines Software

Amin Mubark Alamin Ibrahim
Department of mathematics, Faculty of Education
Shagra University- KSA
aminibrahim2005@gmail.com

Awad Hag Ali
College of Computer Science and Information Technology
Elneelain University, Khartoum, Sudan
awadha@neelain.edu.sd

1-Abstract:
The aim of this paper was to compare between seven free and open search engines: phpdig, sphider, Juggernautsearch, Webglimpse, Webinator, Swish-E and Perlfect, in terms of 30 features. A model design was developed to incorporate these features into four packages according to the impact and efficiency of the search engine functions. Each package was then given a weigh from a scale of 100%: searching mechanisms was given 40%, followed by crawler and indexer, searching features and others representing 26%, 20%, and 14%, respectively. Results revealed that phpdig and sphider were the best engines, there were significantly differences (F-test) over other engines according to all features (p=0.021).

Citation
Ibrahim, Amin Mubark Alamin. A Comparison between seven open and Free Search Engines Software / Awad Hag Ali .- Cybrarians Journal .- No. 35 (September 2014) .- Accessed <State here the date you accessed the article> .- Available at: <Copy here the URL of the current page>

2-Introduction:

Cybrarians Journal
ISSN 1687-2215
دورية إلكترونية فصلية محكمة متخصصة في مجال المكتبات والمعلومات
العدد 35، سبتمبر 2014

The Internet and computer technology has immeasurably increased the availability of information. However, the size of information increases, it becomes difficult for users to retrieve relevant information. Search engines have been developed to facilitate fast information retrieval. There are many software packages for search engine creation on the Internet. Many of them are free or free for noncommercial use. With so many software packages, selecting a suitable search engine software is very difficult than retrieving relevant information efficiently from websites. Free search engine software is not well-documented, which makes it difficult to understand the functions they provide. According, to whoever provides the actual search service, free search tools can be categorized into remote site search service and the server-side search engine (). In the former, the indexer and query engine run on a remote server that stores the index file. When the time of search be, a form on a user's local Web page sends a message to the remote search engine, which then sends the query results back to the user. A server-side search engine is what we usually think of as a search engine. It runs on the user's server, and takes that server's CPU time and disk space. In this paper, the term *search engine* refers only to server-side search engines. According to what is indexed, search engines are classified as *file system search engines* and *website search engines*. File system search engines index only files in the server's local file system. Website search engines can index remote servers by feeding URLs to web crawlers. Most search engines combine the two functions, and can index both local file systems and remote servers. A fully functional website search engine software package should have the following four blocks:

- A *Web Crawler or Spider* that follows HTML links in Web pages to gather documents;
- An *Indexer* that indexes the documents crawled using some indexing rules and saves the indexed results for searching;
- A *Query Engine* that performs the actual search and returns ranked results;
- An *Interface* that allows users to interact with the query engine.

The seven software packages compared are either have all four blocks or allow adding the missing blocks.

Scant work were done to build a model for comparison between the search engines software. Accordingly, this study aimed to develop a model to compare between seven search engines in terms of their impacts and efficiencies. These seven engines are: phpdig, sphider, Juggernautsearch, Webglimpse, Webinator, Swish-E and Perlfect.

3. Basic Information of the seven Search Engine Software Packages:

This shows some basic information about each search engine software package. The information includes licensing, where to find, source code availability, what is written in, platform compatibility, completeness of the package, and who built it. Licensing refers to whether the software is a freeware or is free under some conditions. Source code availability provides the website address to download the source code if it is available. What is written *in* tells what programming language is used in implementing the software. Platform compatibility specifies what operating systems that the software can run on. If the software package is fully functional, i.e. it has a web crawler, an indexer, a query engine, and a query interface, we consider the package to be complete. Who built it tells us the developers of the software. Discussed in [1],[2] and [3]

4-Material and methods

The search engine software packages were compared and contrasted under four packages. A design was developed in which weights are given to each criteria according to the impact and the efficiency the search engine function. These four packages are searching mechanism, Crawler and Indexer, searching feature and others. each of these packages includes different features.

4-1 Searching Mechanism:

Searching Mechanism includes both indexing and ranking methods. Pertaining the Indexing method, Most search engines operate on the principle that pre-indexed data is easier and faster to search. The form and quality of the index created from the original documents is of paramount importance to how searches are performed. The commonly used indexing method is the full text inverted index. It takes a large amount of disk space and the indexing process is slow, because it keeps most of the information in a document. Another method is to index only the title, keywords, description, and other parts of a document. In this way, the indexing process can be very fast and the resulted index is relatively small. And some search engines uses

two-level indexing, which we will introduce later. Some free search engine applies Cellular Expansion Algorithm in indexing, which is still kept as a technical secret. This feature was given a weight of 25%, resulting from the sum of the weigh given to each of its components: inverted index (9%), attribute index (5%), keyword index (4%), two level index (6%) and other features (1%). On the other hands, ranking method is referred to the method that decides a document's relevance to a query. Factors such as word frequency in the document, word position in the text, and link popularity are usually considered. Different search engine takes into consideration the different factors. This feature was given a weight of 15% (Gerald Saltan algorithm (3%), word weight (4%,) word weight and frequency (6%) and other operation (2%)). This package was considered as the most important package because it usually determines how many disk space the search engine requires, how fast the indexing process is, and how fast and accurate the search process is. Accordingly, has been given the highest weight (40%)

4-2Crawler and Indexer Features:

Crawler and Indexer Features compare the following functionalities of built-in web. This feature was given a weight of 26%.

- Crawler Retrieval Depth Control. Can the administrator control the maximum depth that a crawler follows in a retrieval process? Given a weight of 3%.
- Duplicate Detection. During the process of crawling and indexing, can duplicated documents be detected and thus not be indexed? Given a weight of 3%
- Robot Exclusion Standard Support. Does the crawler respect the robot exclusion standard that is to not index documents indicated in the robot.txt file? Given a weight of 3%.
- File Format to be Indexed.what files formats can be crawled and indexed by the crawler and indexer? Given a weight of 3%.
- Index Protected Server. Can the crawler retrieve secured pages in password protected sites?
- Auto indexing.it is to repeat the process of indexing pages every certain times, like every seven days so as to combine the old index with new one. Given a weight of 2%.

- index static and dynamic pages. index static page means that index page which depend on language of static design pages like (html). And index dynamic pages that index pages which depend on dynamic language like (php and asp) .Given a weight of 3%.
- Word Forms. Is word stemming supported? Given a weight of 3%.
- another features with weight 1%

4-3Searching Features

Searching features This feature was given a weight of 20%. And it was considered in the following terms:

• Boolean Search. Can the search engine look up pages containing some word and not containing some other word? Does the search engine support the and/or logic among query words? Given a weight of 2%.
• Phrase Matching: Can the search engine match only those documents that contain words in exactly the same sequence as that of the query? Given a weight of 2%.
• Attribute Search: Can search engine perform search within only the body, title, description, keywords, URL, or other parts of documents? Given a weight of 3%.
• Fuzzy Search: Can the search engine match documents that contain words that are similar to requested query? Are search by soundex, metaphone, or substring supported? Given a weight of 2%.
• Wild Card: Is there a wild card character that can be used in search to match any one or more character or symbol? Given a weight of 0.5%.
• Regular Expression. Regular expressions are symbols that users add to their queries to describe complex patterns to match. Is regular expression search supported? Given a weight of 1%.
• Numeric Data Search: Can the search engine deal with numeric queries such as "Quantity > 300"? Given a weight of 0.5%.
• Case Sensitivity: Is the search engine case sensitive, or can it be configured as case sensitive? Given a weight of 1%.
• Nature Language Query: Does the search engine support nature language queries? Given a weight of 1%.
• Nesting capabilities: Does the search engine has the nesting capabilities? Given a weight of 1%. Nesting capabilities point to the order of implementation for Boolean process like (automobile or car) and sales. This statement will return all record which contain (sales) and contain (automobile or car).The result of this query(automobile or car) and sales is not like automobile or car and sales .
• Phonetic similarity: Does the search engine has phonetic similarity? Given a weight of 1%: it contains many algorithms which is used for spell checkers .Such as when we try to enter a word like (nam) suddenly the search engine asked "do you mean name".
• N-grams: Does the search engine support n-grams queries? Given a weight of 1.5%. it is a conditional probable model for n-1 ,a word the background of a prior word of a human text admitted ,then what are the probable words that may follow it, and what is the probability of the sequence of each of those words or letters.
• Field searching.is within specific parts of a record in DB of search engine. Given a weight of 3%.
• Another features with weight 0.5%.

4-4 Other Features:

This includes seven features and was given a weigh of 14%:
• International Language: Can the search engine support languages other than English? Given a weight of 1%.
• Page Limit: How many pages can be indexed for the free version of the software? What is the theoretical or empirical limit? . Given a weight of 2%.

• Customizable Result Formatting: Can the result pages be customized to have a desired look and feel? Given a weight of 1%.
• Program language: Can the search engine support open source languages or close source language? Given a weight of 4%.
• Full license. Can the search engine support full license? Whether the software is a freeware or free under some conditions. Given a weight of 2%.
• Portal features: Can the search engine support portal features? Whether the search engine contains e-mail, news and engines of a special research (pictures ,videos,…). . Given a weight of 2%.
• Output options: Can the search engine support Output options? When a response is transmitted ,so most search engines retrieve the number of records which had been retrieved ,and some engines introduce an alternative coordination options ,which may include a brief coordination that includes an abstract about the page or the initial words of it which are usually ,the (title, and abstract) . Given a weight of 1%.
• Another features with weight of 1%.

F-test and T-test were used to test for the differences between the seven search engine packages.

5-Results and discussion

The results of the Comparison between the seven search engines are presented in Table 5. It is very clear that phpdig, sphider, were the best engines. They got the highest score compared with the others in almost all features.

Table 1: Features weight of the seven Search Engines

	Web-glimpse 2.x	Web-inator	SWISH-E	Perlfect	Juggernaut search 1.0.1	sphider	phpdig
Searching Mechanism							
Indexing Method	6	9	0.5	9	4	9	9
Relevance Ranking	1	1	1	3	4	4	6
crawler and indexer feature							
Robot Exclusion Standard Support	5	5	5	0	5	5	5
Crawler Retrieval Depth Control	3	3	3	0	0	3	3
Duplicate Page Detection	3	3	3	0.05	3	3	3

File Format to be Indexed	1.55	1.75	1.6	1.8	2	1.5	2.5
Index Protected Server	0	0	0	0	0	0	2
Index dynamics pages	0.5	0.5	0.5	0	0	3	3
Word Forms	0	0.5	3	0	0.125	0.125	0.125
Auto indexing	0	0	0	0	0	0	0
Searching Features							
Boolean Search	2	2	2	2	0	2	2
Phrase Matching	0	2	2	0	0	2	2
Attribute Search	0	0	3	0	3	3	3
Fuzzy Search	1	1	1	0	0.125	1	1
Wild Card	0. 5	0. 5	0. 5	0	0.125	0	0
Regular Expression	1	1	0	0	0.125	0.125	1
Numeric Data Search	0	1	0	0	0	1	0
Case Sensitivity	0	0	0	0	0.125	1	1
Natural Language Query	0	1	0	0	0	1	1
nesting capabilities	0.125	0.125	0.125	0.125	0.125	0.125	0.125
Phonetic Similarity	0.125	0.125	0.125	0.125	0.125	0.125	0.125
Field searching	0	0	0	0	0	3	3
N_grams	0.125	0.125	0.125	0.125	0.125	0.125	0.125
other feature							
International Language	1	0	1	1	0	1	0
Page Limit	0. 5	1	0. 5	1	1	1.5	1.75
Customizable Result Formatting	0	1	0. 5	1	0. 5	1	1
program lanquage	0	0	0	0	0	4	4
Free use	0	0	2	2	0	2	2
Output options	0.05	0.05	0.05	0.05	0.05	1	0.5
portal features	0	0	0	0	0	0	0
Total	25.475	34.175	29.025	21.275	23.05	53.625	57.25

Note: any unknown features in any search engines were given a half of the least weigh of others operation.

5-1Testing the differences between the seven search engines packages

Results of the F-test revealed that there was no significant difference between all search engines with searching mechanism and crawler and indexer features. On the other hand, results of the F-test on the searching feature and other features were found to be highly significant (Table 2 and 3). It is very clear from Table 2 and table 3 that, there were a significant different between each of the search engines sphider and phpdig and other search engines with respect to searching and other features (p>0.10 and p>0.021).This proves that there is a high difference among engines regarding this units. also the results of the F-test between all search engines with all features weight were found to be highly significant (Table 4). It is very clear from Table 4 that, there were a significant different between each of the search engines sphider and phpdig and other search engines with respect to all features (p>0.021).This proves that there is a high difference among engines with all features,also This means that there were differences between phpdig and sphider search engines with other search engines ,and which proves that the search engines sphider and phpdig are the best choice among the other search engines (Figure (1) analysis of all search engines with all features weight)

Table 2: F-test for the seven search engines with respect to searching features

		Mean Difference (I-J)	Std. Error	Sig.
Search Engine (i)	Search Engine (j)			
Juggernaut search 1.0.1	Perlfect	0.12	0.33	0.73
	SWISH-E	-0.38	0.33	0.25
	Web-inator	-0.38	0.33	0.25
	Web-glimpse 2.x	-0.15	0.33	0.64
	sphider	-0.92*	0.33	0.01
	phpdig	-0.89*	0.33	0.01
Perlfect	SWISH-E	-0.5	0.33	0.13
	Web-inator	-0.5	0.33	0.13
	Web-glimpse 2.x	-0.27	0.33	0.42
	sphider	-1.04*	0.33	0.00
	phpdig	-1.00*	0.33	0.00
SWISH-E	Web-inator	0	0.33	1.00
	Web-glimpse	0.23	0.33	0.49

	2.x			
	sphider	-0.54	0.33	0.11
	phpdig	-0.50	0.33	0.13
Web-inator	Web-glimpse 2.x	0.23	0.33	0.49
	sphider	-0.54	0.33	0.11
	phpdig	-0.5	0.33	0.13
Web-glimpse 2.x	sphider	-0.77*	0.33	0.02
	phpdig	-0.73*	0.33	0.03
Sphider	phpdig	0.04	0.33	0.91

ANOVA

	Sum of Squares	df	Mean Square	Sig.
Between Groups	12.874	6	2.146	.010
Within Groups	59.459	84	.708	
Total	72.333	90		

* The mean difference is significant at 95% level.

Table 3: F-test for the seven search engines with respect to other features

		Mean Difference (I-J)	Std. Error	Sig.
(I) S_ENGINE	(J) S_ENGINE			
Juggernaut search 1.0.1	Perlfect	-0.58	0.48	0.23
	SWISH-E	-0.42	0.48	0.39
	Web-inator	-0.08	0.48	0.86
	Web-glimpse 2.x	0	0.48	1
	sphider	-1.41*	0.48	0.01
	phpdig	-1.20 *	0.48	0.02
Perlfect	SWISH-E	0.17	0.48	0.73
	Web-inator	0.5	0.48	0.31
	Web-glimpse 2.x	0.58	0.48	0.23
	sphider	-.83 *	0.48	0.09
	phpdig	0.17	0.48	0.73
SWISH-E	Web-inator	0.33	0.48	0.49
	Web-glimpse 2.x	0.42	0.48	0.39
	sphider	-1.00*	0.48	0.05

	phpdig	-0.79	0.48	0.11
Web-inator	Web-glimpse 2.x	0.08	0.48	0.86
	sphider	-1.33*	0.48	0.01
	phpdig	-1.12*	0.48	0.03
Web-glimpse 2.x	sphider	-1.41*	0.48	0.01
	phpdig	-1.20*	0.48	0.02
Sphider	phpdig	0.20	0.48	0.67
ANOVA				
	Sum of Squares	Df	Mean Square	Sig.
Between Groups	12.140	6	2.023	.021
Within Groups	24.385	35	.697	
Total	36.525	41		

The mean difference is significant at the .90 level.

Table4: F-test for the seven search engines with respect to all features weight

		Mean Difference (I-J)	Std. Error	Sig.
(I) S_ENGINE	(J) S_ENGINE			
Juggernaut search 1.0.1	Perlfect	0.04	0.44	0.92
	SWISH-E	-0.25	0.44	0.57
	Web-inator	-0.40	0.44	0.36
	Web-glimpse 2.x	-0.16	0.44	0.71
	sphider	-1.06*	0.44	0.02
	phpdig	-1.17*	0.44	0.01
Perlfect	SWISH-E	-0.29	0.44	0.50
	Web-inator	-0.45	0.44	0.31
	Web-glimpse 2.x	-0.21	0.44	0.64
	sphider	-1.10*	0.44	0.01
	phpdig	-1.21*	0.44	0.01
SWISH-E	Web-inator	-0.15	0.44	0.73
	Web-glimpse 2.x	0.09	0.44	0.84
	sphider	-.81*	0.44	0.06
	phpdig	-.92*	0.44	0.04

Web-inator	Web-glimpse 2.x	0.24	0.44	0.59
	sphider	-0.66	0.44	0.13
	phpdig	-.77*	0.44	0.08
Web-glimpse 2.x	sphider	-.90*	0.44	0.04
	phpdig	-1.01*	0.44	0.02
Sphider	phpdig	-0.11	0.44	0.81

ANOVA

	Sum of Squares	Df	Mean Square	Sig.
Between Groups	44.159	6	7.360	.021
Within Groups	584.847	203	2.881	
Total	629.006	209		

* The mean difference is significant at the .10 level.

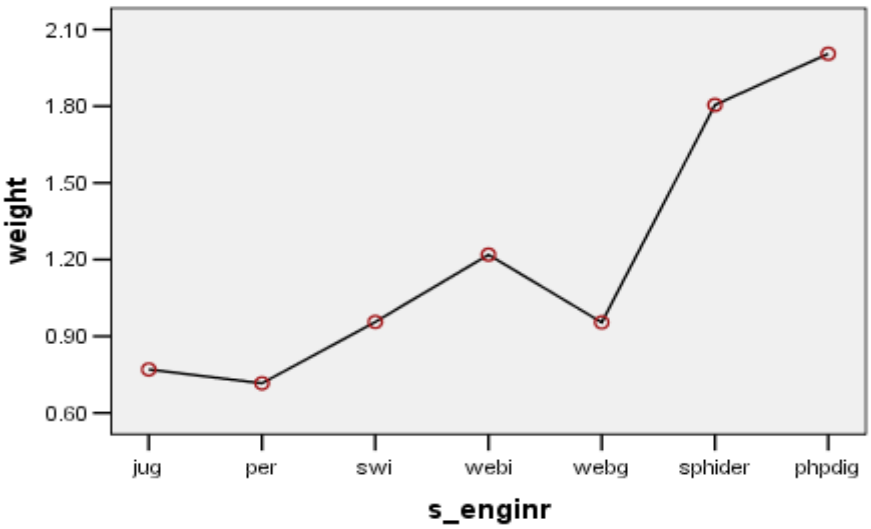

Figure (1)

The lines represents that, the average of phpdig and sphider is greater than the other search engines

5-2Testing the differences between the phpdig and sphider search engines packages

Results of the T-test between sphider and phpdig search engines with all features weight were found to be there was no significant (Table 5). It is very clear from Table 5 that, there was no significant different between each of the search engines sphider and phpdig (p=0.83).

Table5: T-test for the phpdig and sphider search engines with respect to all features

Cybrarians Journal
ISSN 1687-2215
دورية إلكترونية فصلية محكمة متخصصة في مجال المكتبات والمعلومات

العدد 35، سبتمبر 2014

Group Statistics

	s_enginr	N	Mean	Std. Deviation	Std. Error Mean
f_weight	sphider	30	1.8500	1.90632	.34804
	phpdig	30	1.9583	2.03269	.37112

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
f_weight	Equal variances assumed	.062	.804	-.213	58	0.83	-.10	.50
	Equal variances not assumed			-.213	57.763	.832	-.10	.50

5. Conclusion:

1- Differences were tested between seven of search engines in the searching Features unit, Tests showed that there is statistically significant differences between the search engines, where the value of sig is (0.010).As shown in Table (Table2)
2- Differences were tested between seven of search engines in the other Features unit, Tests showed that there is statistically significant differences between the search engines, where the value of sig is (0.021).As shown in Table (Table3)
3- Differences were tested between all search engines with all features weight, Tests showed that there is statistically significant differences between the search engines, where the value of sig is (0.021),as shown in Table (Table4), ,so as to know the significant between any two search engine we used the properties of multi comparisons, This pointer to differences between phpdig and sphider search engines with other search engines ,and which proves that the search engines phpdig and sphider are the best choice among the other search engines,also as shown in figure(1)
4- Differences were tested between phpdig and sphider search with all features weight, Tests showed that there is no statistically significant differences between the search engines, where the value of sig is (0.832).As shown in Table (Table5).

6.References:

[1]The original version of this paper was finished in 2002 as a project paper for the course, *Information Sciences and Technology 511: Information Management – Information and Technology*, taught by Dr. Lee Giles at the College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA. The author was a graduate student at the College of Information Sciences and Technology, The Pennsylvania State University at that time. This paper removed all obsolete contents of the original version.

http://www.searchtools.com/analysis/free-search-engine-comparison.html

[2]Anyone pajolot 2001 - 2003 who develop the php search engine

www.phpdig.net

[3] Andro spass 2005-2007 who develop sphider search engine

www.sphider.eu

[4]Lawrence Philips which he published in the June, 2000 issue

of C/C++ Users Journal

http://swoodbridge.com/DoubleMetaPhone.

[5] Donald T. Kasper, Juggernautsearch Internet Search Engine 1.0.1 Technical Responses and Comparison to HTDIG (HT://DIG), May 2001. http://juggernautsearch.com/htdig.htm.

[6] http://www.perlfect.com/freescripts/search/development.shtml

[7] http://perlfect.com/freescripts/search/

[8] http://swish-e.org/docs/index.html

[9] Webinator WWW Site Indexer Version 5.0. http://www.thunderstone.com/site/webinator5man/webinator5.pdf

[10] Manber, U.; Wu, S., "GLIMPSE: A Tool to Search Through Entire File System". TR 93-34, Department of Computer Science, University of Arizona, Tucson, Arizona, 1993.

[11] Udi Manber, Mike Smith,, Burra Gopal, "WebGlimpse Combining Browsing and Searching" , Proceedings of 1997 Usenix Technical Conference, Jan 6-10, 1997

[12] http://www.webglimpse.net/features.html