Assessment of the Web using Genetic Programming

Reginald L. Walker*
Computer Science Department
University of California at Los Angeles
Los Angeles, California 90095-1596
EMAIL: rwalker@cs.ucla.edu
TELE: 310/606-3231

Abstract

The generation of a Web page follows distinct sources for the incorporation of information. The early sources for a Web page design were organized displays of known information based on the page designers' interest and/or design parameters. The sources may have been published in books or other printed literature, or disseminated as general information about the page designer. The growth in the number of Web pages has led to the development and refinement of several search engines. The use of the refined search engines still results in an array of diverse information when the same set of keywords are used in a Web search. Some consistency in the search results can be achieved over a period of time using the same search engine. Unfortunately, most initial Web searches are also treated as the final searches on a given topic following some form of refinement of the keywords used in the search process. Search strategies behind the current search engines for the World Wide Web were studied to determine the applicability of Genetic Programming for the current diverse set of Web documents. This assessment will be applied to the incorporation of Web documents that have not yet been developed.

Keywords: Genetic Programming, Information Retrieval, Internet, World Wide Web, Search Engines.

*This work was supported by the Raytheon Fellowship Program.

1 The Early Search Engines

The early Web search engines [8] were explicitly divided into separate components: a search engine index browser, search engine indexers, and the search engine. The early popular search engine and index browser were provided by the Gopher integrated search engine software. The early search engine indexers were Archie, Veronica, and Jughead. The Gopher integrated search engine software provided a useful search engine and index browser, but its indexing mechanism was not adequate. The inefficiency of the indexing mechanism led to the development of Archie. The lack of multiple word searches in the Archie indexer led to the development of Veronica and eventually to the development of Jughead. Each of these indexers only provided an indexing mechanism for the Gopher search engine. The Wide Area Information Service, WAIS, provided an integrated search engine, indexer, and browser. Its indexer browser also provided the user with indexed results that were generated by the Gopher-related indexers. Mosaic[1][11] provided users with an alternative to the Gopher index browser for its indexed results as well as FTP archives of software.

2 Comparison of Current Search Engines

Comparisons of the general characteristics that comprise the current search engine technology [3, 11] showed that Yahoo! [16] has the closest ties to the early search engine strategies and Mosaic, since it was initially developed to track Gopher, FTP, and Telnet sites. This search engine also followed an early biology-based classification system, the result of work

[1] It should be noted that Netscape [12] and Mosaic share design methodologies in conjunction with some of the same software developers.

Table 1: **Comparison of the number of relevant documents for each search engine using "sassafras tea" as the basis.**

	AltaVista	Excite/ NetCenter	HotBot	Infoseek	LookSmart	Yahoo!
"sassafras tea"	336	91	0	126	337	73
+"sassafras tea" +herb	66	21	0	5	66	0
"sassafras tea" NEAR herb	143311	- -	- -	- -	143311	- -
"sassafras tea" OR herb	555226	- -	- -	- -	555226	- -
"sassafras tea" AND herb	679028	- -	- -	- -	679028	- -

Table 2: **Comparison of the number of relevant documents for each search engine using "Sassafras tea" as the basis. Note: the search results of the form** 23/54484 **imply** $Webpage/Website$.

	AltaVista	Excite/ NetCenter	HotBot	Infoseek	LookSmart	Yahoo!
"Sassafras tea"	195	91	0	126	196	73
+"Sassafras tea" +herb	42	21	0	23/54484	42	0
"Sassafras tea" NEAR herb	142324	- -	- -	- -	142324	- -
"Sassafras tea" OR herb	582609	- -	- -	- -	582609	- -
"Sassafras tea" AND herb	679028	- -	- -	- -	679028	- -

performed by the 18th century botanist Linnaeus. AltaVista [1] and Inktomi [6] search engines followed the indexer strategy of WAIS by using no fixed categories for indexing the millions of documents. The inclusion methodologies for all the database entries follow two approaches: via a Web crawler and/or human editors. Each engine provided access to millions of Web pages through their databases and/or through AltaVista's or Inktomi's database. AltaVista provided the most up-to-date database by rebuilding the database every 24 hours. This approach eliminated problems associated with new pages, moved pages, and deleted pages. The updating of other databases occurred every seven to ten days. The Information Retrieval (IR) system [7, 15, 13] associated with these two indexers worked very well. LookSmart [9] used 24,000 categories, which reflected the response time of its IR system. This slow response time was also a major factor for the Lycos search engine. The slower response time indicated the use of distributive databases. Infoseek [5] used both approaches and allowed the requester to choose the IR system with or without categories. The more categories supplied by the indexer, the slower the IR system.

Different IR systems used advanced database categorization [7] schemes. AltaVista used a database IR system based on Dynamic Categorization Technology [1], also known as COW9. This technology does not make any *a priori* or *post priori* assumptions about individual search patterns. HotBot [4] and Lycos[10] used a similar technology through their partnership with Inktomi. Excite[2]/NetCenter[2][12] used an *a priori* approach in its IR system by means of its Intelligent Concept Extraction (ICE) Technology [2]. The ICE technology allows the indexer to grow in terms of search query relationships executed by previous users of its IR system. A stored list of search queries allows the system to stabilize after the ICE system has been exposed to variations of a query word combination. Infoseek used a Context Classification Engine (CCE) Taxonomy [5] that applies technology similar to Knowledge Discovery in Databases (KDD) [14] for its IR system. Yahoo! and LookSmart currently make use of the database query technology provided by AltaVista. These technologies use either Genetic Algorithms (GA) or Genetic Programming (GP) methodologies as a basis.

[2]NetCenter was produced by the Netscape Communication Corporation.

The inclusion methodologies either rely upon human editors and/or Web crawler software. The shortcoming of human editors [11] was the bias that may result from their knowledge about certain subject matters. Some bias also resulted from the guidelines provided by the classification methodologies chosen for individual search engines. Yahoo! provided the Web designer with the initial freedom of categorizing her/his Web page by providing a submission form. A Yahoo! staff editor reviewed the submittal form for the appropriateness of the chosen category. The use of a Web crawler for database inclusion eliminated any bias that was not incorporated into the methodology for the inclusion mechanism. Excite used a human editor for the homepage inclusion and a Web crawler to traverse the links from the homepage. The human editors reviewed some of the Web page sites and judged the contents. The contents of the site were judged in conjunction with the site's design as well as the overall appeal. This approach may add some bias towards an appealing, well designed site.

3 Study of Searches using the Current Search Engines

3.1 Overview

The growth and stability of Web pages were studied using two approaches to track information. The first approach looked at the different character patterns used by the search engines for Web page classification. This study used the term *sassafras tea* as the basis for the search comparisons (see Tables 1 - 3). The choice of the term *sassafras tea* underscores the diversity that exists for Web searches using the same and/or different search engines. Fifteen distinct search patterns generated the results for the search engines in this study. The second approach looked at the generation of new Web pages following a natural disaster. The natural disaster chosen for this search was *Hurricane Mitch*. The initial date for the first Web search occurred November 5, 1998. This data collection effort focused on the impact of *Hurricane Mitch* on *Florida* over an eight-day period. The eight-day period covered the time it takes most search engines to update their databases. Three distinct search patterns generated the results for the search engines in this study.

3.2 Results for Search Patterns in Study 1

The search for the three major strings of "sassafras tea," "Sassafras tea," and "Sassafras Tea" showed some of the current search engines' indexers as case sensitive for different string patterns. Supplying an additional keyword narrowed the scope of the search engines. This keyword was "herb" (see Tables 1 - 3). The use of a sequence of words inside double quotes indicated the searched-for phrase. The use of the + symbol indicates that the string pattern MUST occur within the document. Only the AltaVista and Lycos indexers used the NEAR reserved word. AltaVista uses a ± 10 word range and Lycos uses a ± 25 word range. The use of NEAR, AND, and OR reserved words in conjunction with the strings yielded unexpected results. The search pattern

> "sassafras tea" NEAR herb

yielded

> *hits*(sassafras NEAR herb) + *hits*(tea NEAR herb)

which is equivalent to the following two independent searches

> *hits*(+sassafras NEAR +herb) + *hits*(+tea NEAR +herb).

The search pattern should have been

> +"sassafras tea" NEAR +herb.

Similar search problems occurred using the AND and OR reserved words. There was a dramatic increase in the number of relevant documents (*hits*) when these keywords were part of the search string pattern and used incorrectly. LookSmart used the AltaVista indexer and database to generate its *hits*. This search phrase was not a typical topic searched on a regular basis. HotBot return 0 *hits* for all of the tested search patterns in this study.

3.3 Results for Search Patterns in Study 2

The search results for AltaVista show the most variations between the first and second days (see Table 4). These differences show the instability in the initial search. The results for Days 2 through 5 show little or no variations. A variation only occurred from a decrease in the number of hits. The transition from Day 5 to Day 6 showed an increase in the number of pages for each of the search patterns. Day 6 and Day 8 showed the results as stable and consistent.

The search results for Excite/NetCenter remained stable for Days 1 through 4 (see Table 4). The transition from Day 4 to Day 5 resulted in a 152-fold decrease in

Table 3: **Comparison of the number of relevant documents for each search engine using "Sassafras Tea" as the basis. Note: the search results of the form 23/54483 imply** $Webpage/Website$.

	AltaVista	Excite/ NetCenter	HotBot	Infoseek	LookSmart	Yahoo!
"Sassafras Tea"	159	91	0	126	160	73
+"Sassafras Tea" +herb	26	21	0	23/54483	26	0
"Sassafras Tea" NEAR herb	142072	- -	- -	- -	142072	- -
"Sassafras Tea" OR herb	583773	- -	- -	- -	583773	- -
"Sassafras Tea" AND herb	677866	- -	- -	- -	677866	- -

Table 4: **Comparison of the number of relevant documents for the AltaVista and Excite/Netscape search engines using "hurricane" as the basis. Note: the * means the search pattern was +"hurricane mitch"+florida.**

	AltaVista				Excite/Netscape		
	hurricane	+hurricane +mitch	+hurricane +mitch +florida		hurricane	+hurricane +mitch	+hurricane +mitch +florida
Day 1	197465	7785	3136	Day 1	74757	440	138
Day 2	517700	94	31*	Day 2	74757	440	138
Day 3	517700	94	31*	Day 3	74757	440	138
Day 4	483951	94	31*	Day 4	74757	440	138
Day 5	517700	94	31*	Day 5	492	4653	55
Day 6	518240	160	46*	Day 6	85674	4653	1230
Day 8	518240	160	46*	Day 8	85674	4653	1230
Mean	467285	1212	479	Mean	67267	2246	438
Standard Deviation	110776	2684	1085	Standard Deviation	27674	2085	502

the "hurricane" search pattern, a 10-fold increase in the "hurricane mitch" search pattern, and a two-fold decrease in the last search pattern. The transition from Day 5 to Day 6 resulted in a 174-fold increase in the "hurricane" search pattern. The "hurricane mitch" search pattern produced stabilized results with the final search pattern producing a 22-fold increase.

The search results for HotBot showed similar results for the first two days of this study (see Table 5). The number of hits during the transition from Day 2 to Day 3 increased 11-fold for the "hurricane" search pattern, 3-fold for "hurricane mitch" search pattern, and 3-fold for the "hurricane mitch florida" search pattern. Days 3 to 8 remained consistent, with minor fluctuations in the number of hits.

The search results for Infoseek remained stable for the "hurricane" search pattern with a two-fold increase Day 8 (see Table 5). The "hurricane mitch" search pattern showed a 6-fold decrease from Day 1 to Day 2. The final search pattern showed two-fold decreases for transitions from Day 1 to Day 2 and from Day 6 to Day 8.

The search results for LookSmart proved identical to the search results for AltaVista, with the exception of the first day of the search (see Table 6). This discrepancy in the results for the first day for these two search engines indicated that the LookSmart results were based on a refinement of the initial results of the AltaVista search. The results for Day 2 through 8 compared the same as the results for AltaVista.

The search results for Yahoo! showed the instability of the initial search with a 175-fold decrease from Day 1 to Day 2 for the "hurricane" search pattern (see Table 6). This decrease was followed by a 250-fold increase from Day 2 to Day 3. There was a 281-fold decrease from Day 1 to Day 2 for the "hurricane mitch" search pattern. The number of hits for Day 2 totaled approximately the same for each search pattern. The results showed stability for Days 3 through 8. The final search pattern showed a 32-fold decrease/increase from Day

Table 5: **Comparison of the number of relevant documents for the HotBot and Infoseek search engines using "hurricane" as the basis. Note: the * means the search pattern was +"hurricane mitch"+florida.**

	HotBot				Infoseek		
	hurricane	+hurricane +mitch	+hurricane +mitch +florida		hurricane	+hurricane +mitch	+hurricane +mitch +florida
Day 1	23571	439	176	Day 1	192999	195	18
Day 2	23609	441	178	Day 2	192999	30	9*
Day 3	271514	1542	560	Day 3	192999	30	9*
Day 4	271514	1542	560	Day 4	192999	30	9*
Day 5	267465	1526	560	Day 5	192999	30	9*
Day 6	271514	1542	560	Day 6	192999	30	9*
Day 8	271514	1542	560	Day 8	366140	27	5*
Mean	200100	1225	451	Mean	217733	53	10*
Standard Deviation	111643	496	173	Standard Deviation	60587	58	4*

Table 6: **Comparison of the number of relevant documents for the LookSmart and Yahoo! search engines using "hurricane" as the basis. Note: the * means the search pattern was +"hurricane mitch"+florida. Note: the search results of the form 0/0 imply** $Webpage/Website$.

	LookSmart				Yahoo!		
	hurricane	+hurricane +mitch	+hurricane +mitch +florida		hurricane	+hurricane +mitch	+hurricane +mitch +florida
Day 1	483985	81	1	Day 1	64382	0/103046	0/0
Day 2	517700	94	31*	Day 2	367	0/367	0/357
Day 3	517700	94	31*	Day 3	91826	2/368	0/11
Day 4	483951	94	31*	Day 4	91826	2/366	0/361
Day 5	517700	94	31*	Day 5	91826	2/369	0/364
Day 6	518240	160	46*	Day 6	91826	2/378	0/373
Day 8	518240	160	46*	Day 8	91826	2/380	0/375
Mean	508217	111	31*	Mean	74840	1/15039	0/263
Standard Deviation	15338	31	14*	Standard Deviation	31844	1/35929	0/163

2 to Day 3 and Day 3 to Day 4 respectively.

All the search engines displayed instability during the first three days, with an exception of Excite. The Excite instability occurred during the three-day period starting with Day 4. All of the search engines showed an increase in the number of hits except Infoseek. Infoseek showed a decrease in the number of relevant hits when compared by time. The means and standard deviations associated with each search pattern registered the inconsistencies among each individual search engine. These values also displayed the variations among the distinct search engines as a group.

4 Conclusion

The results of searching the Web over a period of eight days showed that the databases for the distinct search engines stabilized after the completion of the initial search. Searches repeated over a series of days showed the refinement of the database indexers for the chosen search engines. The user can optimize the results initially produced by repeating the search over a period of time.

These results also proved that the most popular search engines did not produce accurate results for the initial search. The initial search may contain some inherent errors that are not apparent or documented. Literature on search strategies, as well as the actual search engine Web pages, inform the user that the results may vary among search engines. However, this information does not mention that the results produced by each search engine may vary over a period of days due to the undocumented refinements for the same keywords.

The search engines used in this study were AltaVista, Excite, HotBot, Infoseek, LookSmart, and Yahoo!. These search engines were chosen because their respective indexers return numeric values for the relevance rating for the chosen search patterns. The Lycos search engine did not receive consideration in these studies because its index browsers returned the first relevant 25 Web documents for a given search pattern.

5 Acknowledgements

The author wishes to thank the reviewers for their helpful comments. Also, Walter Karplus, Zhen-Su She, and Lixia Zhang for their direction and suggestions.

References

[1] AltaVista. AltaVista Web Page. Digital Equipment Corporation, Maynard, MA, November 1998.

[2] Excite. Excite Web Page. Excite Inc. Mountain View, CA, November 1998.

[3] A. Glossbrenner and E. Glossbrenner. *Search Engines for the World Wide Web.* Peachpit Press, Berkeley, 1998.

[4] HotBot. HotBot Web Page. HotBot Inc. San Francisco, CA, November 1998.

[5] Infoseek. Infoseek Web Page. Infoseek Corporation. Santa Clara, CA, November 1998.

[6] Inktomi. Inktomi Web Page. Inktomi Corporation, San Mateo, CA, November 1998.

[7] D.H. Kraft, F.E. Petry, B.P. Buckles, and T. Sadasivan. The use of genetic programming to build queries for information retrieval. In *Proceedings of the First IEEE Conference on Evolutionary Computation*, volume 1, pages 468–73, June 1994.

[8] E. Krol. *The Whole Internet User's Guide and Catalog.* O'Reilly and Associates, Inc., Sebastopol, CA, 1994.

[9] LookSmart. LookSmart Web Page. LookSmart Ltd. San Francisco, CA, November 1998.

[10] Lycos. Lycos Web Page. Lycos Inc. Framingham, MA, November 1998.

[11] C. Musciano and B. Kennedy. *HTML: The Definitive Guide.* O'Reilly and Associates, Inc., Sebastopol, CA, 1997.

[12] Netscape. Netscape Web Page. Netscape Communications Corporation. Mountain View, CA, November 1998.

[13] M.L. Raymer, W.F. Punch, E.D. Goodman, and L.A. Kuhn. Genetic Programming for Improved Data Mining - Application to the Biochemistry of Protein Interactions. In *Proceedings of the First Annual Conference on Genetic Programming*, pages 375–80, July 1996.

[14] T.-W. Ryu and C.F. Eick. MASSON: Discovering Commonalities in Collection of Objects using Genetic Programming. In *Proceedings of the First Annual Conference on Genetic Programming*, pages 200–8, 1996.

[15] M. Stillger and M. Spiliopoulou. Genetic Programming in Database Query Optimization. In *Proceedings of the Genetic Programming 1996 Conference*, pages 388–93, July 1996.

[16] Yahoo! Yahoo! Web Page. Yahoo! Inc. Santa Clara, CA, November 1998.