Web Searching Agents, What Are They Doing Out There?

Bernard J. Jansen
School of Information Sciences
and Technology
The Pennsylvania State University
University Park, PA, 16801, USA
jjansen@ist.psu.edu

Amanda S. Spink
School of Information Sciences
University of Pittsburgh
Pittsburgh PA 15260
spink@ist.psu.edu

Jan Pedersen
AltaVista Company
1070 Arastradero Rd.
Palo Alto, CA 94304
jan.pedersen@av.com

ABSTRACT - *The Web has become a worldwide repository of information, which individuals, companies, and organizations utilize to solve or address various information problems. Many of these Web users utilize automated agents to gather this information for them. It is assumed that this approach represents a more sophisticated method of searching. However, there is little research investigating how Web agents search for online information. In this research, we examine how agents search for information on Web search engines, including the session, query, term, duration and frequency of interactions. For this study, we analyzed queries that 2,717 agents submitted to the Alta Vista search engine on 8 September 2002. Findings include: (1) agents interacting with Web search engines use queries comparable to human searchers, (2) Web agents are searching for a relatively limited variety of information, with only 18% of the terms used being unique, and (3) agent – Web search engine interaction typically spans several hours with multiple instances of interaction per second.*

Keywords:
Information retrieval agents, softbots, automated Web searching

1. INTRODUCTION

The Web has become a major source of information that people and organizations utilize to address a variety of information issues. There is a growing body of literature examining how people search the Web [1-3], providing insights into how humans conduct Web searching. However, non-humans now conduct at least a portion of Web searching. These non-humans include agents, automated processes or spiders that search the Web. For this paper, we refer to spiders, softbots, meta-search applications and other automated information gathering processes all as agents.

Web search engines, individuals, commercial corporations, and others use the agents to retrieve information from the Web on their behalf. It is assumed that these agents are a more sophisticated method of searching, relative to human searchers. However, little research has investigated the validity of this assumption. It is this assumption that we challenge in this research.

In this manuscript, we report findings from our analysis that focus on the interactions between Web agents and Web search engines. An understanding of how Web agents search is an important research area with ramifications for Web search engine design, network performance, along with commercial, social and privacy issues.

2. LITERATURE REVIEW

A Web searching agent is a program that automatically traverses the Internet using the Web's hypertext structure. The agent can either retrieve a particular document or use some specified searching algorithm to recursively retrieve all Web documents that are referenced from some beginning base document [4].

Most Web search engines, such as Alta Vista (www.altavista.com) and Google (www.google.com), employ agents as crawlers [5]. In addition to these general-purpose search engines, niche search engines also employ agents. For example, Lawerence, Giles, and Bollacker [6] developed ResearchIndex, which incorporates a software agent to locate computer articles on the Web. By utilizing these agents to gather and organize online data, many of these general and niche search engines have become valuable information resources.

As a result, others retrieve information from these search engines for personal, commercial, and other purposes. Although humans conduct much of this searching, others use agents to retrieve the information. Commercial examples include meta-crawlers search engines such Ithaki (http://www.ithaki.net/indexu.htm) or Dogpile (http://www.dogpile.com).

Unlike standard search engines, meta-crawlers do not crawl the Web themselves to build listings. Instead, they use automated applications to send queries to several search engines simultaneously. The results from all the queried search engines are then blended together into one or more results listing. Other companies also utilize meta-crawler software to locate job information, evaluate page rankings, or locate bargains for certain products or services.

0-7803-7952-7/03/$17.00 © 2003 IEEE.

Research is still on going in the meta-search area. For instance, Chen, Meng, Fowler, and Zhu [7] are developing an intelligent Web meta-indexer for Web searching, which is a stand alone system that utilizes results from other Web search engines.

It is not only corporations and organizations that employ agents. Individuals also utilize agents to gather information. Sample code for Web searching agents is readily available [8], and designing Web agents is now a fairly common student project at many universities [9, 10]. Additionally, there are several inexpensive commercial applications that provide meta-crawler software that runs from a desktop computer [11].

Given the amount of information indexed and the number of users, there is a growing body of research examining the use of Web search engines [1, 2, 12, 13]. Jansen and Pooch [1] present an extensive review of the Web searching literature, reporting that Web searchers exhibit different search characteristics than do searchers using other information systems. Jansen and colleagues [12] conduct an in-depth analysis of the user interactions with the Excite search engine, reporting analysis focusing on sessions, query, and terms. Silverstein and fellow researchers [13] conduct a large study on a sample of queries of over a billion queries, also focusing on sessions, and queries. Spink and associates [2] analyze trends in Web searching, reporting that Web searching has remained relatively stable over time, although they note a shift from entertainment to commercial searching.

However, all of these studies examine searching patterns of humans searching for information on Web search engines. None of these articles distinguished between human and agent Web searching.

It has been stated that Web agents offers a more sophisticated method of searching for information on the Web [14]. There is a significant amount of literature on Web agents and their use by Web search engines to gather information [15]. There is also significant research into methods to optimize agent information gathering to avoid unnecessary loads on servers or the network [16-18]. However, researchers have not investigated the actual searching characteristics of these Web agents, even though for some time there have been questions about their effect on information providers [19].

The minimal research investigating agent information gathering characteristics when using search engines is quite surprising. Many researchers have noted the dramatic effect of Web search engines on society [20]. In education for example, research articles online have a higher citation rate relative to those articles not online [21]. In job seeking, niche job boards have dramatically altered the hiring process [22]. In fact, these search engines have become so adept at gathering information and therefore influencing how this information is used that some now consider these search engines security and privacy risks [23]. The entire topic of what information the search engines provides or does not provide may have a dramatic effect on which people or organizations are successful [24].

With individuals, organizations, corporations and others using agents to retrieve information from these search engines, it would seem that an understanding of how these agents interaction with search engines is of great importance. The results of this research have ramifications in terms of system design, e*commerce, network performance, and the societal effects of the Web. These considerations are the drivers for this research.

3. RESEARCH QUESTIONS

More specifically, the research questions driving this study are:

(1) What are the Web searching characteristics exhibited by Web search agents when using search engines?

(2) What is the frequency and duration of the interaction between Web agents and Web search engines?

(3) What types of information are Web agents retrieving?

To address these research questions, we obtained and quantitatively and qualitatively analyzed actual queries submitted to AltaVista, a major U.S. Web search engine, by Web agents.

4. RESEARCH DESIGN

The queries used in this study were submitted to Alta Vista on 8 September 2002 and span a 24-hour period. The queries were recorded in a transaction log and represent a portion of the searches executed on the Web search engine on this particular date. At the time of the data collection, Alta Vista was the ninth most popular search engine on the Web [25].

4.1 Data Collection

The original transaction log contains approximately 3,200,000 records. Each record contains three fields: (1) *Time of Day*: measured in hours, minutes, and seconds from midnight of each day as recorded by the Alta Vista server; (2) *User Identification*: an anonymous user code assigned by the Alta Vista server; and (3) *Query Terms*: terms exactly as entered by the given user.

Using these three fields, we could locate the initial query and recreate the chronological series of actions in a session. In this research, we generally follow the terminology outlined in [1]. Briefly, a term is any series of characters separated by white space. A query is the entire string of terms submitted by a searcher in a given instance. A session is the entire series of queries submitted during one interaction with the Web search engine.

4.2 Data Analysis

The original query transaction log contained searches from both human users and non-human agents. We were interested in only those queries submitted by agents. From the original transaction log, we therefore extracted a sub-set of queries that we deemed were submitted by agents.

To do this, we separated all sessions with greater than 100 queries into an individual transaction log. We chose 100 because it is nearly 50 times greater than the reported mean search session [1] for human Web searchers, and over 70 times greater than the reported standard deviation. We were satisfied that we had retrieved a subset of the transaction log that contained mainly queries submitted by agents or perhaps high volume common user terminals. It is also probably that we are not including some agent requests in our sample; however, this sample certainly represents a substantial portion of agent submissions.

When an agent or human searcher submits a query, then views a results page, and then goes to the next results page, the Alta Vista server logs this second interaction with the identical user identification and identical query, but with a new time (i.e., the time of the second visit). This is beneficial information in determining how many of the retrieved results the agent might have visited from the search engine, but unfortunately it also skews the results in analyzing how the agents searched on system.

To address the first research question, we collapsed the data set by combining all identical queries [1] submitted by the same agent. This gave us unique queries in order to analysis sessions, queries and terms.

For the second and third research question, we utilized the complete un-collapsed sessions in order to obtain an accurate measure of the temporal length of sessions and the number of results visited.

5. RESULTS

In this section, we present the results of our study.

Table 1 presents general searching information of the agent – search engine interactions.

Table 1: Aggregate results for general search trends

	Agent Searching Data During Interactions with Alta Vista	
Sessions	2,717	
Queries	896,387	
Terms		
Unique	570,214	17.7%
Total	3,224,840	
Terms per query	mean 3.6	sd 2.8
Terms per query		
1 term	216,105	24%
2 terms	268,076	30%
3+ terms	411,988	46%
Queries per Agent	mean 329.9	sd 1383.9
Agents modifying queries	2,386	88%
Session size		
1 query	331	12%
2 queries	109	4%
3+ queries	2,277	84%
Results Pages Viewed		

Table 1: Aggregate results for general search trends

	Agent Searching Data During Interactions with Alta Vista	
1 page	760,071	85%
2 pages	67,755	8%
3+ pages	68,561	8%
Boolean Queries	177,182	20%
Terms not repeated in data set	411,577	13%
Use of 100 most frequently occurring terms	834,251	26%

At the query level, Web agent queries are comparable to queries submitted by human Web searchers. About 46% of agent queries contained more than 3 terms compared to 45% for human searchers [2]. The standard deviation (2.8) [1] is about twice that of human searchers The use of Boolean operators by agents is about double that of human searchers, but still represents a minimal usage at 20%.

In terms of results pages, over 86% of the Web agents accessed only the first page of results, which is higher than reported in research on human Web searches (approximately 40%) [3, 12, 26].

There are major differences between agent and human searchers at the term and session level of analysis. The percentage of agent sessions with more than three queries (84%), after duplicate queries were removed, was significantly higher than that of human searchers (25%) [2].

The number of unique terms (18%) for was very low compared to human searchers (61%) [2] indicating a tight jargon used by Web agents and a limited subject matter. The use of the 100 most frequently occurring terms (26%) submitted by agents was also high compared to human searchers (usually under 20%) [2].

We further examine agent searching at the term level of analysis in order to get a better understand of what types of information the agents are commonly searching. In Table 2, we present the top term occurrences for the agent data set. Term co-occurrence is useful to assist in determining the specific usage of a term intended by a searcher within the framework of a particular query [27].

Table 2: Top Term Co-occurrences

	bensalem	center	cv	embedded	estate	fax	feasterville	fitness	high	iksar	kunark	luclin	necromancer	neriak	new	number	permafrost	qeynos	real	retail	sale	school	serial	text	title
bensalem	-																								
center		-																							
cv			-																						
embedded				-																					
estate					-																				
fax		1902				-																			
feasterville	512						-																		
fitness		2326				1902		-																	
high									-																
iksar										-															
kunark											-														

Table 2: Top Term Co-occurrences

	bensalem	center	cv	embedded	estate	fax	feasterville	fitness	high	iksar	kunark	luclin	necromancer	neriak	new	number	permafrost	qeynos	real	retail	sale	school	serial	text	title
luclin										515		-													
necromancer										593			-												
neriak										714			539	-											
new															-										
number		528				528		528								-									
permafrost																	-								
qeynos										842	515	521	619	767			551	-							
real																			-						
retail						589														-					
sale					2130														2129		-				
school									515													-			
serial																915							-		
text				509																				-	
title			685																						-
title:resume																									850
title:vitae																									820
trevose	512						978																		
us		1057				1059		1057																	
velious										510								525							
york															1060										

From Table 2, the highest term pairs were *fitness – center* (2,326 occurrences), *estate sale* (2,130), *real sale* (2,129), *fitness fax* (1,902), and *fax – center* (1,902).

Other than these highest occurring pairings, two other trends appear. First, there is a high number of job related terms (*title cv title: resume title:vitae*). Cappelli (22) has reported that corporations and job boards employ measures to seek out passive job seekers by searching for online resumes online. They search for filename and document titles containing the term "resumes", for examples.

For the other trend, there is a high occurrence of fantasy gaming terms (*luclin necromancer neriak qeynos*). All of these terms, and many others, refer to the various game popular fantasy games.

In addition to analyzing the queries and terms, we also examined the duration and frequency of the agent interactions with the search engine. In Table 3, we report the results of this analysis.

The duration and frequency of agent – search engine interaction is substantially different than that of human searchers. The mean agent session (approximately 9 1/2 hours) is 38 times the mean human session of 15 minutes [1]. However, the standard deviation was relatively high at just over 8 hours. The maximum sessions duration was the full temporal span of the data sampling period. The minimum duration was 2 seconds.

The mean number of queries per session (615) is 300 times that of human searchers (just over 2) [28]. The average agent submits a query about every 2 seconds, with a standard deviation of approximately 4 queries. The maximum session frequency was just less than 100,000 queries in the 24-hour span, and the maximum queries per second was 137.

Table 3. Time, Queries, and Queries Per Second

Duration of Interaction	Hours: Minutes: Seconds	Queries	Queries Per Second
Average	9:27:30	615	0.43
St Dev	8:05:49	2,609	4.17
Max	23:59:57	99,595	137
Min	0:00:02	101	< 0.00

6. IMPACT AND CONCLUSION

Agents interacting with Web search engines use queries similar to those submitted by human searchers. Agent submit very short and generally simple queries, but they are persistent in submitting queries, with over 84% of agents submitting more than 3 queries, with the mean being just more than 600 queries.

Further investigation is needed to determine if there is a relationship between these simple queries and long sessions. Perhaps, if the queries were more sophisticated, the sessions may not need to be so lengthy. This has implications for Web search engine performance during peak usage periods and for network bandwidth usage.

Agents are searching for a fairly limited variety of information, with less than 18% of the terms used being unique. This small number of terms indicates that the agents are searching for a fairly limited subject matter. From the term co-occurrence analysis, it appears that commerce and entertainment topics are of the most interest to those using agents to gather information.

The agent – search engine interaction is typically over several hours with multiple instances of interaction every few second. Although the mean duration was about nine and a half hours, several agent interactions continued for the entire 24-hour period. The maximum frequency of interaction was over 137 requests per second. This means the agent was viewing, and possibly downloading, over 137 Web documents a second. The mean interaction was about a query every 2 seconds. The lack of an external economic incentive may be contributing to the inefficient but high volume searching employment by these agents.

This study contributes to our understanding of Web searching in several important ways. First, the data comes from real agents, deployed by real users, submitting real queries and looking for real information. Accordingly, it provides a realistic glimpse into how Web agents search, without the self-selection issues or altered behavior that can occur with lab studies or survey data.

The study also has limitations. The sample data comes from only one major Web search engine, introducing the possibility that the queries do not represent the queries submitted by the broader Web agent population. However, [1] suggests that characteristics of human searchers are very consistent across search engines. We can hypothesis that this may hold for agents also.

sFurther research is continuing to examine the changing trends in automated searching and explore more directly the manners by which agents use Web search engines to locate information.

7. ACKNOWLEDGMENTS

Our thanks to Alta Vista for providing access to the data, without which this research could not have been conducted.

8. REFERENCES

[1] B. J. Jansen and U. Pooch, "Web User Studies: A Review and Framework for Future Work," *Journal of the American Society of Information Science and Technology*, vol. 52, pp. 235-246, 2001.

[2] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic, "From E-sex to E-commerce: Web Search Changes," *IEEE Computer*, vol. 35, pp. 107-111, 2002.

[3] C. Hölscher and G. Strube, "Web Search Behavior of Internet Experts and Newbies," *International Journal of Computer and Telecommunications Networking*, vol. 33, pp. 337-346, 2000.

[4] M. Koster, "The Web Robots FAQ," vol. 2002, 1998.

[5] D. Sullivan, "Search Engine Features For Webmasters," vol. 2002: SearchEngineWatch.com, 2002.

[6] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital Libraries and Autonomous Citation Indexing," *IEEE Computer*, vol. 32, pp. 67-71, 1999.

[7] Z. X. Chen, X. N. Meng, R. H. Fowler, and B. Zhu, "Features: Real-time Adaptive Feature and Document Learning for Web Search," *Journal of the American Society for Information Science*, vol. 52, pp. 655-665, 2001.

[8] SearchTools.com, "Source Code for Web Robot Spiders," SearchTools.com, 2001.

[9] M. Berry and M. Browne, *Understanding Search Engines: Mathematical Modeling and Text Retrieval.* Philadelphia: SIAM, 1999.

[10] G. M. Youngblood, "Web Hunting: Design of a Simple Intelligent Web Search Agent," *ACM Crossroads Student Magazine*, vol. 5, pp. 1-4, 1999.

[11] D. Sullivan, "Search Utilities," vol. 2002: SearchEngineWatch.com, 2003.

[12] B. J. Jansen, A. Spink, and T. Saracevic, "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web," *Information Processing and Management*, vol. 36, pp. 207-227, 2000.

[13] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, "Analysis of a Very Large Web Search Engine Query Log," *SIGIR Forum*, vol. 33, pp. 6-12, 1999.

[14] O. Etzioni, "Moving U p the Information Food Chain:Deploying Softbots on the World Wide," vol. 2002, 1996.

[15] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan, "Searching the Web," *ACM Transactions on Internet Technology*, vol. 1, pp. 2 - 43, 2001.

[16] J. Talim, Z. Liu, P. Nain, and E. G. Coffman, "Controlling the robots of Web search engines," presented at 2001 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, Cambridge, Massachusetts, 2001.

[17] Z. Xiaohui, W. Huayong, C. Guiran, and Z. Hong, "An Autonomous System-based Distribution System for Web Search," presented at 2001 IEEE International Conference on Systems, Man, and Cybernetics, Tucson, AZ, 2001.

[18] V. Shkapenyuk and T. Suel, "Design and implementation of a high-performance distributed Web crawler," presented at 8th International Conference on Data Engineering, San Jose, CA, 2002.

[19] E. Selberg and O. Etzioni, "Multi-Service Search and Comparison Using the MetaCrawler," presented at 4th International World-Wide Web Conference, Boston, Massachusetts, USA, 1995.

[20] R. Brody, "Illusions of plenty: the role of search engines in the structure and suppression of knowledge," presented at IEEE International Symposium on Technology and Society, Rome, Italy, 2000.

[21] S. Lawrence, "Online or Invisible?," *Nature*, vol. 411, pp. 521, 2001.

[22] P. Cappelli, "Making the most of on-line recruiting," *Harvard Business Review*, vol. 79, pp. 139-146, 2001.

[23] J. C. Hernandez, J. M. Sierra, A. Ribagorda, and B. Ramos, "Search engines as a security threat," *Computer*, vol. 34, pp. 25-30, 2001.

[24] L. Introna and H. Nissenbaum, "Defining the Web: the politics of search engines," *Computer*, vol. 33, pp. 54-62, 2000.

[25] CyberAtlas, "November 2002 Internet Usage Stats," vol. 2002: Nielsen//NetRatings Inc., 2002.

[26] F. Cacheda and Á. Viña, "Experiences retrieving information in the World Wide Web," presented at The 6th IEEE Symposium on Computers and Communications, Hammamet, Tunisia, 2001.

[27] L. Leydesdorff, "Words and co-words as indicators of intellectual organization," *Research Policy*, vol. 18, pp. 209-223, 1989.

[28] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real Life Information Retrieval: A Study of User Queries on the Web," *SIGIR Forum*, vol. 32, pp. 5-17, 1998.