The 4th International Conference on Internet Computing. Las Vegas, Nevada, p. 620 –626. 23 – 26 June 2003.

MONSTERS AT THE GATE: WHEN SOFTBOTS VISIT WEB SEARCH ENGINES

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

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

Abstract

There has been little research investigating how Web agents search for information. We focused on three research questions: (1) How do agents interact with Web search engines? (2). What type of information are Web agents looking for and in what quantity?, and (3) What is the duration and frequency of Web agent interactions The queries examined for this study were submitted to Alta Vista on 8 September 2002, spanning a 24-hour period. Findings include: (1) when agents interact with Web search engines they use simpler queries than those submitted by human searchers, (2) Web agents are searching for a wide variety of information, with 60% of the terms used being unique, and (3) agent – Web search engine interaction is typically over several hours with multiple instances of interaction per second. Implications for Web agents and search engines are discussed.

Keywords: softbots, agents, Web searching

Introduction

There is a growing body of literature examining how people search the Web [1-3]. This research provides some insight into how humans search for information on the Internet and provides a broad framework for examining how agents search. Jansen and Pooch [1] review the Web searching literature and suggest that Web searchers exhibit different search techniques than do searchers on other information systems. Hölscher and Strube [3] examined European searchers and note that experts exhibit different searching patterns than novices. Spink, et al.,[2] show that Web searching characteristics have remained relatively stable over time, although search subjects have shifted from entertainment to commercial searching.

There is little research investigating the effects of agent Web searching, the use of automated processes by Web search engines [4] to locate information, or the use of spiders [5] both by search engines and individuals to retrieve information on the Web. We refer to spiders, softbots, meta-search applications and other automated information gathering as agents. An understanding of agent search has ramifications for Web search engine design, network performance, and commercial, social and privacy issues and, as such, is an important research area.

Research Questions

We focused on three research questions:

1. How do agents interact with Web search engines?
2. What type of information are Web agents looking for and in what quantity?
3. What is the duration and frequency of Web agent interactions?

Research Design

Data Collection

The queries examined for this study were submitted to Alta Vista[1] 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. 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.

Data Analysis

With these three fields, we located initial query and recreated the chronological series of actions in a session. We generally follow the terminology outlined in [1]. 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 by

[1] http://www.altavista.com

a user during one interaction with the web search engine.

The transaction log contained searches from both human users and agents. We were interested in only those queries submitted by agents. From the transaction log, we therefore culled a sub-set of queries that we deemed were submitted by agents. To do this, we separated all sessions with greater than 10,000 queries into an individual transaction log. We chose 10,000 because it is almost 5,000 times greater than the reported mean search session [6] for human Web searchers. We were satisfied that we had retrieved a subset of the transaction log that contained solely queries submitted by agents or other automated process. It is more likely that we are not including some agent requests in our sample; however, this sample certainly represents the largest of the agent submissions (i.e., the "monsters" of the group).

Web Agents

When an agent or human searcher submits a query, then views a document, and returns to the search engine, the Alta Vista server logs this second visit with the identical user identification and 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 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.

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:General Agents Searching Characteristics

	Number	Percentage
Sessions	22	
Queries	219,718	
Terms *Unique*	277,902	60%
Total	459,537	
Mean terms per query	2.3	
Terms per query		
1 term	78,997	36%
2 terms	102,474	47%
3+ terms	38,247	17%
Pages Viewed Per Query		
1 page	18,8747	86%
2 pages	17,155	8%
3+ pages	13,816	6%
Mean queries per user	9,987	
Agents modifying queries	20	91%
Session size		
1 query	-	0%
2 queries	1	5%
3+ queries	21	95%
Boolean Queries	1,737	1%
Terms not repeated in data set	224,561	49%
Use of 100 most frequently occurring terms	49,492	11%

The Web agent queries are simpler than queries submitted by human Web searcher. Only 17% of the agent queries contained more than 3 terms. Some 45% of Web queries submitted by humans are greater than 3 terms. The agent queries are also simpler in terms of structure, with only 1% of the agent queries containing Boolean operators, compared to about 10% for human Web users. Agents also exhibit the similar characteristic as human Web searchers with a very low tolerance for wading through a lot of results. In fact, Web agents appear to have an even lower tolerance for viewing a large number of results. For 86% of the agent's queries, only the first set of results were viewed, which is 30% higher than human Web searchers.

However, there are substantial differences. The number of unique terms is high at 60% of the total term occurrences, and the session lengths are long. Over 95% of the agents had sessions greater than 3 queries, after duplicate queries were removed. This is four times what human Web searchers submit [7]. Almost 50% of terms were not repeated in the data, indicating a very disperse language used and diverse information needs. However, the use of the 100 most frequently occurring terms was similar to that reported by human Web users, indicating that query terms at the high end of the rank – frequent spectrum may adhere to some common distribution.

Table 2: Session Lengths

Agent	Queries	Agent	Queries
A	38,294	L	10,150
B	26,044	M	7,933
C	21,343	N	6,042
D	14,516	O	3,051
E	13,123	P	2,817
F	13,102	Q	2,660
G	12,633	R	1,681
H	12,471	S	1,053
I	11,645	T	111
J	10,558	U	6
K	10,483	V	2

Sessions and Queries

The agent sessions are exponentially longer than the normal Web users with a mean session length of 9,987 queries and a very high standard deviation of 9,420. As a comparison, the mean Web human session is about 2 queries [7]. Again, these results are for the collapsed transaction logs, so some of the session lengths are less than the original 10,000 indicating that there were several duplicate queries with in the session. For example, we see that agent V had only 2 unique queries, although this agent's un-collapsed session contained 10,092 queries.

Table 3 presents information on query lengths.

Table 3: Number of Occurrences of Query Lengths

Query Length	Occurrences	Percentage	Query Length	Occurrences	Percentage
1	79,000	35.96%	11	58	0.03%
2	102,474	46.64%	12	44	0.02%
3	10,021	4.56%	13	26	0.01%
4	5,029	2.29%	14	10	< 0.01%
5	18,065	8.22%	15	6	< 0.01%
6	2,931	1.33%	16	3	< 0.01%
7	1,224	0.56%	17	3	< 0.01%
8	425	0.19%	19	1	< 0.01%
9	307	0.14%	21	1	< 0.01%
10	90	0.04%			

As for query length, the typical agent queries are similar to those of typical Web users, generally about 2 terms per query. However, only 17% of the agent queries contained more than 3 terms, which is substantially lower than the percentage of normal Web users.

Terms

Table 4 presents the most frequently occurring terms.

Table 4: Most Frequently Occurring Terms

Term	Occurrences	Term	Occurrences
center	2,442	south	322
fitness	2,326	hotel	294
real	2,265	virginia	283
estate	2,256	supplies	278
sale	2,183	dakota	274
fax	1,905	facsimile	269
us	1,071	fotos	247
manufacturers	615	online	240
new	615	california	235
colombia	557	gratis	232
wholesale	552	mexico	229
number	533	historia	229
manufacturing	432	service	216
equipment	358	york	213
north	350	gift	205
para	332	lasik	203
carolina	326	home	200

From a review of these terms relative to top terms reported in other studies [8], the most notable difference is the absence of sexual terms and the lack of popular entertainers or celebrities. Terms in agent queries are more commercially or location focused (e.g., *real*, *estate*, *columbia*, *carolina*, *virginia*, *california*, and *mexico*). Some 37% of the top terms refer to location.

We also analyzed the queries using term co-occurrence analysis, that looks for the simultaneous occurrence of terms within queries [9].

Table 5 presents the most frequently occurring term co-occurrences. The four most frequently occurring pairs were (1) *fitness* and *center* (2,312 co-occurrences), (2) *real* and *estate* (2,252), and (3) *estate* and *sale* (2,082), and (4) *real* and *sale* (2,082).

Table 5: Term Co-Occurrence

Term	Term	Occurrence	Term	Term	Occurrence
fitness	center	2,312	south	carolina	159
real	estate	2,252	north	carolina	157
estate	sale	2,082	new	jersey	150
real	sale	2,082	west	virginia	147
fax	center	1,902	north	dakota	147
fax	fitness	1,902	zfacsimile	center	143
fax	us	1,057	zfacsimile	fitness	143
us	center	1,057	lasik	surgery	128
us	fitness	1,057	south	dakota	124
fax	number	528	north	real	117
number	center	528	north	estate	117
number	fitness	528	equipment	supplies	109
facsimile	center	267	north	sale	108
facsimile	fitness	267	carolina	estate	106
new	york	215	carolina	real	106
new	estate	174	new	mexico	101
new	real	174	hilton	head	101

From the analysis of term co-occurrence, the trend identified in the term analysis continues with a lack of sexual or celebrity pairing, and the location searching becomes more apparent with at least 18 (51%) of the pairs referring to locations (e.g., *real estate*, *new york*, *south carolina*, *north carolina*, *new jersey*).

Surprisingly, of the more 217,000 queries submit, there were no query that was submitted by more than one agent, although there were some common terms among queries. However, several agents submitted queries multiple times. As stated previously, whenever an agent viewed a document and then returned to the search engines, this subsequent visit was recorded with a matching query. This allowed us to analyze the number of results visited by the agents based on the number of duplicate queries.

Table 6 presents the top number of documents visited ranked by query.

Table 6: Top Occurring Queries (Number of Documents Visited)

Query	Occurrences	Query	Occurrences
link:www.dimpleart.com	35,304	like:http://www.usms.org/ comp/calendar.htm	415
hocking hills wedding	16,915	sexo	406
britney spears	16,502	juegos	306
link:www.balancedliving.com	9,273	gambling	300
link:www.releasetechnique.com	7,476	musica	205
link:www.drproactive.com	7,453	web casino	200
link:www.microsoft.com/office/ word/default.asp	7,070	internet casino	200
link:www.dare2believe.com	4,829	online casino	200
link:www.thaifooddb.com	4,330	video poker	200
dogs	4,305	online gambling	200
link:http://www.nurseshift.com	3,962	gamble	200
link:www.workathomewith healthnutrition.com	3,960	like:http://www.nal.usda.gov/fnic/ pubs/bibs/gen/freelow.html	198
link:www.harrisdigitalpublishing.com	3,854	Unfinished Wholesale Furniture	195
link:www.ultimatesuccesstips.com	3,722	like:http://www.ascx.com/gymco.htm	155
link:www.innergear.com	3,651	horoscope	153
link:theinspirationplace.com	3,423	entertainment	145
hocking hills	962	dog	145
sony dvd player	868	mp3	135
halibut	615	POEMAS	112
chat	494	like:http://www.usa-gymnastics.org /suppliers/	111
postales	426	britney+spears	107

From Table 6, it is apparent that some of the agents are very Web results driven, visiting hundreds and even thousands of Web documents. At this level of analysis, we see for the first time the appearance of celebrity queries (e.g., *britney spears*) and popular searching topics (e.g., *mp3*, *online gambling*). The agent with the largest session (99,595 queries) was an agent interested in tracing hyperlinks to particular web sites. There are several companies employing agents for such services (e.g., Web Position Gold, a common Web master ranking software).

Time and Frequency of Interaction

Table 7 presents the overall statistics for the session time and number of interactions. These results are derived from the un-collapsed sessions, so the duplicate queries appear in the calculations.

Table 7: Statistics for Session Time, Number of Queries, and Queries Per Second

	Session Time				Queries	
	Hours	Minutes	Seconds		Number of Queries	Queries Per Second
Average	10	4	12		20,809.09	1.62
St Dev	8	57	2		18,738.56	2.44
Max	23	39	48		99,595	11.78
Min	0	24	19		10,093	0.15

Typical user sessions are a few minutes [10] while our results indicates that agent sessions are several hours. Not only was the duration of the interaction lengthy, but also the frequency of the interaction was on very intense with the agents submitting 1.6 queries per second on average.

For a more detailed examination, Table 8 presents the session lengths, number of queries, and queries per second for the top three agents measured by number of queries and the top three agents as measured by queries per second. For identification purposes, we labeled each other the agents A to V.

Table 8: Agent Session Period, Number of Queries, and Queries Per Second

Agent	Session Time			Queries (by Highest Number of Queries)	
	Hours	Minutes	Seconds	Number of Queries	Queries Per Second
A	23	2	7	99,595	1.20
B	4	58	27	38,294	2.14
C	23	0	31	26,088	0.31
Agent	Session Time			Queries (by Highest Number of Queries Per Second)	
	Hours	Minutes	Seconds	Number of Queries	Queries Per Second
J	0	24	19	17,188	11.78
G	2	1	19	20,418	2.81
O	1	19	32	13,269	2.78

Discussion

Agent's interacting with Web search engines use simpler queries than those submitted by human searchers. Agent queries are very short with almost no advanced searching operators and are substantially shorter in general than Web queries submitted by human searchers. Agents are also persistent in submitting queries, with over 95% of agents submitting more than 3 queries, with the mean being just under 10,000 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 wide variety of information, with 60% of the terms used being unique. There was no query submitted by more than one agent, which is surprising given the large number of queries in the sample. Agents do not appear to be searching for sexual or other popular terms, but are more focused on commercial information, especially location-related information. Three of the four most frequently occurring terms pairs were location-related. It was only at the individual query level of analysis that popular search topics appeared in the results. Although on average agents are not interested in a large number of results, with 86% of the agent only viewing no more than the top ten documents for a particular query, some agents were extremely result driven, viewing hundreds or thousands of results.

The agent – Web search engine interaction is typically over several hours with multiple instances of interaction per second. Although the mean duration was about 10 hours, several agent interactions continued for the entire 24-hour period. The maximum frequency of interaction was over 11 queries per second including duplicate queries. This means the agent was viewing, and possibly downloading, over 11 Web documents a second. The mean interaction was over 2 instances per second. The long duration and high frequent raises concerns about what information is being collected for and the effect of these interactions on Web as a public resource. The lack of an external economic incentive may be a contributing factor to the inefficient searching techniques employment by these agents.

This study contributes to our understanding of Web searching in several important ways. First, the data comes from real agents or automated processes 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. Second, our sample is quite large, with over 200,000 Web queries, permitting us to examine and report results from a variety of perspectives. Finally, we obtained the data from a popular and established Web search engine.

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, Jansen and Pooch (2001) have suggested that characteristics of Web sessions, queries, and terms are very consistent across search engines. We also do not have information about the systemic characteristics of the Web agents who submitted queries or their designers. So, we must infer their intentions from terms and co-occurrence analysis.

Conclusion

This study provides a useful characterization of Web agent information searching and gives insight into the queries, terms and term pairs that are most frequently used. Armed with this information, search engines and other Web information providers can design their Web sites to accommodate or exclude these automated information gathers. Further 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.

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] 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.

[5] M. Chau, D. Zeng, and H. Chen, "Personalized Spiders of Web Search and Analysis," presented at Joint Conference on Digital Libraries, Roanoke, VA., 2001.

[6] 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.

[7] A. Spink, D. Wolfram, B. Jansen, and T. Saracevic, "The Public and their Queries," *Journal of the American Society for Information Science*, vol. 52, pp. 226-234, 2001.

[8] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic, "Searching of the Web: The Public and Their Queries," *Journal of the American Society of Information Science and Technology*, vol. 52, pp. 226-234, 2001.

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

[10] D. He, A. Göker, and D. J. Harper, "Combining Evidence for Automatic Web Session Identification," *Information Processing & Management*, vol. 38, pp. 727 - 742, 2002.