Automated Gathering of Web Information: An In-Depth Examination of Agents Interacting with Search Engines

BERNARD J. JANSEN, TRACY MULLEN
The Pennsylvania State University
AMANDA SPINK
The University of Pittsburgh
and
JAN PEDERSEN
Overture Services, Inc.

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. Some assume 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 first provide a classification for information agent using stages of information gathering, gathering approaches, and agent architecture. We then examine an implementation of one of the resulting classifications in detail, investigating how agents search for information on Web search engines, including the session, query, term, duration and frequency of interactions. For this temporal study, we analyzed three data sets of queries and page views from agents interacting with the Excite and AltaVista search engines from 1997 to 2002, examining approximately 900,000 queries submitted by over 3,000 agents. Findings include: (1) agent sessions are extremely interactive, with sometimes hundreds of interactions per second (2) agent queries are comparable to human searchers, with little use of query operators, (3) Web agents are searching for a relatively limited variety of information, wherein only 18% of the terms used are unique, and (4) the duration of agent-Web search engine interaction typically spans several hours. We discuss the implications for Web information agents and search engines.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

General Terms: Algorithms, Performance,

Additional Key Words and Phrases: Search engines, Web searching, agent searching

Authors' addresses: B.J. Jansen, and T. Mullen, The Pennsylvania University; email: jjansen@hq.acm.org; A Spink, Queensland University of Technology; J. Pedersen, Yahoo Search Marketing Service.

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 [Hölscher and Strube 2000; Jansen and Pooch 2001; Spink et al. 2002], 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 article, we refer to spiders, softbots, metasearch 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. There is a general assumption that these agents represent a more sophisticated method of searching relative to human searchers (see [Aridor et al. 2002, p. 1; Budzik and Hammond 1999, p. 9; Chen and Sycara 1998, p. 1]. 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. More particularly, we investigate (1) agent searching characteristics when using search engines, (2) the frequency and duration of interaction, and (3) the types of information these agent retrieve. An understanding of how Web agents search is an important research area with ramifications for Web search engine design, information architecture, and 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 [Koster 1998]. In Etzioni's information food chain [Etzioni 1996a], Web crawlers and search engines are information herbivores grazing on Web pages and hyperlinks. Under this metaphor, intelligent information agents are information carnivores, autonomously employing intelligence to hunt down information and dine on information resources. Overwhelming users with too much information is a real concern, and these specialized agents can help the user acquire and integrate information, effectively playing the role of a personal assistant [Huhns and Singh 1998]. The Semantic Web vision of agents navigating a machine-understandable Web is gradually moving closer to reality with the increased use of XML and the development of ontology markup languages, such as DARPA Agent Markup Language (DAML), Resource Description Language (RDF), and Ontology Web Language (OWL) [Hendler 2001; Miller 2004].

In Table I, we categorize recent information agent research according to the stage of the information retrieval process and the approach taken towards making Web search agents more customizable and intelligent. We divided

Table I. Information Agent Approaches and Stages

Stages	Information Acquisition	Query Formation	Information Delivery and Management
Approach			
Algorithmics	(Brandman et al. 2000), (Broder, Najork, and Wiener 2003), (Edwards, McCurley, and Tomlin 2001), (Wolf et al. 2002)[Brandman et al. 2000; Broder et al. 2003; Edwards et al. 2001; Wolf et al. 2002]	(Chen et al. 2000), (Madden et al. 2002) [Chen et al. 2000], [Madden et al. 2002]	N.A.
Information filtering	(Chen and Sycara 1998), (Hurley and Wilson 2001), (Lieberman, Fry, and Weitzman 2001), (Doorenbos, Etzioni, and Weld 1997), (Bollacker, Lawrence, and Giles 1998), (Lu and Sterling 2000), (Knoblock et al. 2001), (Menczer et al. 2001), (Pant and Menczer 2002), (Rowe 2002), (Cesarano, d'Acierno, and Picariello 2003), (Chakrabarti, van den Berg, and Dom 1999), (Aridor et al. 2002), (Diligenti et al. 2000), (Joachims, Freitag, and Mitchell 1997), (Rhodes and Maes 2000) .()[Chen and Sycara 1998], [Hurley and Wilson 2001b], [Lieberman et al. 2001], [Doorenbos et al. 1997], [Bollacker et al. 1998], [Lu and Sterling 2000], [Knoblock et al. 2001a], [Menczer et al. 2001], [Pant and Menczer 2002], [Rowe 2002b], .()[Cesarano et al. 2003], [Chakrabarti et al. 1999], [Aridor et al. 2002],[Diligenti et al. 2000], [Joachims et al. 1997], [Rhodes and Maes 2000]	(Chen and Sycara 1998), (Lieberman, Fry, and Weitzman 2001), (Aridor et al. 2002), (Lee et al. 1998), (Yu, Koo, and Liddy 2000), (Pitkow et al. 2002), (Martin and Jose 2003), (Somlo and Howe 2003), (Flake et al. 2002), (Glover et al. 2001), (Ghani, Jones, and Mladenic 2001),(Lin and Knoblock, 2003) [Chen and Sycara 1998], [Lieberman et al. 2001], [Aridor et al. 2002], [Lee et al. 1998], [Yu et al. 2000], [Pitkow et al. 2002], [Martin and Jose 2003], [Somlo and Howe 2003a], [Flake et al. 2002], [Glover et al. 2001], [Ghani et al. 2001b], [Lin and Knoblock 2003]	(Hurley and Wilson 2001), (Aridor et al. 2002), (Yu et al. 2000), (Pitkow et al. 2002), (Thomas and Fischer 1997), (Voss and Kreifelts 1997), (Joshi 2000), (Chen et al. 2001), (Budzik and Hammond 1999) [Hurley and Wilson 2001b], [Aridor et al. 2002], [Yu et al. 2000], [Pitkow et al. 2002], [Thomas and Fischer 1997a], [Voss and Kreifelts 1997a], [Joshi 2000], [Chen et al. 2001a], [Budzik and Hammond 1999]
Collaborative filtering	(Good et al. 1999)[Good et al. 1999]	(Joshi 2000), (Glance 2001), (Fitzpatrick and Dent 1997) [Joshi 2000], [Glance 2001a], [Fitzpatrick and Dent 1997]	(Voss and Kreifelts 1997), (Joshi 2000), (Good et al. 1999), (Mladenic 1999) [Voss and Kreifelts 1997a], [Joshi 2000], [Good et al. 1999], [Mladenic 1999]
Information Integration	(Barish et al. 2000), (Knoblock et al. 2001), (Lu and Sterling 2000))[Barish et al. 2000; Knoblock et al. 2001a; Lu and Sterling 2000]	(Barish et al. 2000), (Knoblock et al. 2001), (Lu and Sterling 2000) [Barish et al. 2000; Knoblock et al. 2001a; Lu and Sterling 2000]	(Barish et al. 2000), (Knoblock et al. 2001), (Lu and Sterling 2000) [Barish et al. 2000; Knoblock et al. 2001a; Lu and Sterling 2000]

Note: We include the author and publication date for ease of reading, along with the citation, for example, (Chen et al. 2000) [13].

the stages into *information acquisition*, *query formulation*, and *information delivery and management*. Information acquisition entails searching and retrieving resources from the Web to be indexed by search engines. Query formulation involves translating the user's information need into one or more lists of query terms to retrieve information from search engines. Information delivery and management requires delivering and filtering query results, as well as allowing the user to save, edit, and otherwise manage the resultant information.

Approaches taken by the various information agents include *algorithmics*, *information filtering*, *collaborative filtering*, and *information integration*. Algorithmics approaches generally consist of optimizing algorithms and data structures to achieve faster and more accurate results. Information filtering approaches help select relevant information for a given user at a given time and in a given context. This can be done by either filtering out irrelevant information or by actively recommending useful information. Collaborative filtering focuses on identifying users with similar preferences and using their opinions to provide recommendations for information and information sources. Finally, information integration is a process whereby an agent tries to achieve its goals by combining information from multiple heterogeneous information sources. For example, a user who wants to put together a travel package might use an agent to recommend combinations of accommodations, flights, and sightseeing.

We also categorized information agent approaches by architectural design, as shown in Table II. Some approaches consist of a single centralized agent that performs all aspects of the task. Other agents are built as multiple interacting agents or distributed multiagent systems, where different agents specialize in different tasks. The system as a whole uses inter-agent communication protocols to coordinate agent activities. Currently, the majority of agent development focuses on the single agent paradigm, but the number of multiagent systems appears to be growing. As individual agents become more robust and intelligent, multiagent system approaches that coordinate their activities for a more end-to-end approach to information-seeking may very well become more prominent.

While the breakout of categories is helpful, and many agents do fall into exactly one category, there are both approaches and agent systems whose capabilities span two or three of the categories. Thus a common approach towards implementing personalization is creating, managing, and exploiting user profiles. This includes learning user profiles either explicitly through relevance feedback or tacitly through observation [Chen et al. 2001a; Lieberman et al. 2001; Yu et al. 2000]. A key concern is that the learning process should not place an undue burden on the user. A standard means of representing such user profiles is with the *term frequency x inverse document frequency* (TF-IDF) algorithm [Somlo and Howe 2003b], which captures each document as a vector representation of the weighted frequency of terms. A term that occurs frequently in the document, but rarely in the rest of the collection, has a high weight. The user profile can be a single TF-IDF vector or multiple topic-specific ones. When a new document's vector is sufficiently similar to vectors in the user's profile, then it is deemed relevant [Somlo and Howe 2001].

Table II. Information Agent Architectures and Stages

Stages	Information Acquisition	Query Formation	Information Delivery and Management
Arch			
Single Agent	(Brandman et al., 2000), (Broder, Najork, and Wiener, 2003), (Edwards, McCurley, and Tomlin, 2001), (Wolf et al., 2002), (Chen and Sycara 1998), (Hurley and Wilson 2001), (Lieberman, Fry, and Weitzman 2001), (Doorenbos, Etzioni, and Weld, 1997), (Bollacker, Lawrence, and Giles, 1998), (Knoblock et al., 2001), (Menczer et al., 2001), (Pant and Menczer, 2002), (Rowe, 2002), (Chakrabarti, van den Berg, and Dom, 1999), (Aridor et al., 2002), (Diligenti et al., 2000), (Joachims, Freitag, and Mitchell, 1997), (Rhodes and Maes, 2000) [Brandman et al. 2000; Broder et al. 2003; Edwards et al. 2001; Wolf et al. 2002], [Chen and Sycara 1998], [Hurley and Wilson 2001b], [Lieberman et al. 2001] , [Doorenbos et al. 1997], [Bollacker et al. 1998], [Knoblock et al. 2001a], [Menczer et al. 2001], [Pant and Menczer 2002], [Rowe 2002b], [Chakrabarti et al. 1999], [Aridor et al. 2002], [Diligenti et al. 2000], [Joachims et al. 1997], [Rhodes and Maes 2000]	(Chen et al., 2000), (Madden et al., 2002), (Chen and Sycara, 1998), (Lieberman, Fry, and Weitzman, 2001), (Pitkow et al., 2002), (Martin and Jose, 2003), (Somlo and Howe, 2003), (Flake et al., 2002), (Glover et al., 2001), (Ghani, Jones, and Mladenic, 2001), (Lin and Knoblock, 2003), (Joshi, 2000), (Glance, 2001), (Fitzpatrick and Dent, 1997) [Chen et al. 2000], [Madden et al. 2002], [Chen and Sycara 1998], [Lieberman et al. 2001], [Pitkow et al. 2002], [Martin and Jose 2003], [Somlo and Howe 2003a], [Flake et al. 2002], [Glover et al. 2001], [Ghani et al. 2001b], [Lin and Knoblock 2003], [Joshi 2000], [Glance 2001a], [Fitzpatrick and Dent 1997]	(Hurley and Wilson, 2001), (Pitkow et al., 2002), (Thomas and Fischer, 1997), (Joshi, 2000), (Chen, Chen, and Sun, 2001), (Mladenic, 1999), (Budzik and Hammond, 1999) [Hurley and Wilson 2001b], [Pitkow et al. 2002], [Thomas and Fischer 1997a], [Joshi 2000], [Chen et al. 2001a], [Mladenic 1999], [Budzik and Hammond 1999]
Multiple Interacting Agents	(Lu and Sterling, 2000), (Knoblock et al., 2001), (Pant and Menczer, 2002), (Cesarano, d'Acierno, and Picariello, 2003), (Good et al., 1999), (Barish et al., 2000) [Lu and Sterling 2000], [Knoblock et al. 2001a], [Pant and Menczer 2002], [Cesarano et al. 2003], [Good et al. 1999], [Barish et al. 2000]	(Lu and Sterling, 2000), (Knoblock et al., 2001), (Aridor et al., 2002), (Lee et al., 1998), (Yu, Koo, and Liddy, 2000), (Barish et al., 2000) [Lu and Sterling 2000], [Knoblock et al. 2001a], [Aridor et al. 2002], [Lee et al. 1998], [Yu et al. 2000], [Barish et al. 2000]	(Lu and Sterling, 2000), (Knoblock et al., 2001), (Aridor et al., 2002), (Yu, Koo, and Liddy, 2000), (Voss and Kreifelts, 1997), (Good et al., 1999), (Barish et al., 2000) [Lu and Sterling 2000], [Knoblock et al. 2001a], [Aridor et al. 2002], [Yu et al. 2000], [Voss and Kreifelts 1997a], [Good et al. 1999], [Barish et al. 2000]

In the sections that follow, we describe each of the information retrieval stages and the application approaches for that stage in more detail. For agent systems that cross one or more categories, we describe them in the most relevant category.

2.1 Information Acquisition

Information acquisition can be as straightforward as Web robots that crawl the Web periodically and bring back pages to be indexed for future user queries. In this realm, agent research tends to concentrate on improving algorithmics using techniques such as network flow theory [Wolf et al. 2002], uniform resource locator (URL) caching [Broder et al. 2003], or maintaining data on page change rates [Edwards et al. 2001]. One crawler automatically categorizes pages based on ontologies and semantic networks [Cesarano et al. 2003]. [Brandman et al. 2000] turn the problem around and focus on creating more crawler-friendly Web servers. They suggest several Web server strategies (e.g., provide the crawler with information about recently changed material) that could lower the load on both crawlers and servers alike.

Agents can also carry out *focused search*. The focus can be topic-specific [Chakrabarti et al. 1999; Menczer et al. 2001], or oriented towards application and hardware constraints [Aridor et al. 2002], such as those required for the small storage space and low battery resources on most pervasive devices. Since these types of agents only search a portion of the Web, they require fewer resources to keep results current. One key issue in this area is assigning credit across crawled links so that local gains in Web page relevance do not overshadow less immediately promising, but eventually valuable paths [Diligenti et al. 2000]. Given the restricted focus, focused search also has the potential to ease query formulation and filtering tasks. The MySpiders system [Pant and Menczer 2002] deploys adaptive crawlers whose population learns and evolves to better fit the search task. Citeseer [Bollacker et al. 1998] searches for information about research publications, then parses and links them based on their structure format (e.g., abstract, introduction, citations), and finally manages and makes them searchable through an augmented search interface that suggests related papers to the user. The Marie-4 crawler [Rowe 2002a] uses an expert system approach to do caption-based retrieval of images only.

Agents can also look for information proactively, based either on standing requests ("I'm interested in new information about information agents") or based on what the agent anticipates the user may need next. Rhodes and Maes [2005] survey these agents and call them *just-in-time information retrieval* (JITIR) agents. One of their agents, the Remembrance Agent, displays a list of documents that are highly correlated to the document the user is currently working on the Agents discussed range across our information acquisition stage (i.e., How can JITIR agents automatically provide users with relevant information given their current environment?), to the information delivery stage (i.e., How should a JITIR agent present information in the most useful nonintrusive fashion?), as well tot he study of the psychological aspects (i.e., How do

JITIR agents affect user information seeking behavior?). Watson [Budzik and Hammond 1999] is a similar JITIR agent that models user behavior and directs search towards tasks in which the user is currently engaged. Letzia [Lieberman et al. 2001] automatically constructs a user profile from Web pages the user has recently visited. It serves as a reconnaissance agent that scouts ahead, following links in the neighborhood of the current Web page and comparing them to the user's current profile. Other similar work in this area includes WebWatcher [Joachims et al. 1997] and WebMate [Chen and Sycara 1998].

2.2 Query Formulation

Human query refinement typically consists of typing a query to a search engine, and then iteratively refining it until the results are satisfactory or the user no longer wishes to continue the process. Agents can support and help automate this process in several ways. The NLI agent [Lee et al. 1998] allows users to ask natural language queries, which include anaphoric terms and elided expressions, and translate them into standard Boolean queries. A Web search agent keeps track of the search history, allowing users to refine the original natural language queries.

At the information filtering level, advances in retrieving and querying over structured data, typically in XML-like format, allows agents to reason about the data using inference and other formal methods. Under this research, queries generally pull data upon request. Alternatively queries periodically push new information to the user. Research into continuous queries allows users and agents to specify queries that monitor for significant changes in updated and streaming data sources [Chen et al. 2000; Madden et al. 2002].

Agents also use automatic query expansion techniques which augment the query based on a query *context* [Fitzpatrick and Dent 1997] by capturing previous user queries and searches [Martin and Jose 2003; Pitkow et al. 2002], external sources such as Web browser histories and user profiles [Somlo and Howe 2003b], or other indications of the user's current context. Thus, if a user is looking for pages on weather, and searches for "forecast", the system might add the terms "weather" to return weather forecasts and not stock forecasts. Automated learning methods can be used to bias queries toward specific subtopics using features extracted from positive and negative examples of the subtopic [Flake et al. 2002; Glover et al. 2001]. Automated query expansion is used by CorpusBuilder [Ghani et al. 2001a] to build a corpus of documents in a specified language (e.g., Slovenian). An inverse-geocoder agent [Lin and Knoblock 2003] uses a complex query strategy to provide inverse lookup services that can, for example, allow the user to lookup restaurant information by zip code even when it is stored by the restaurant name.

Augmented queries can also be constructed as part of a collaborative searching process where earlier related queries within the community of users can be used to augment the current one [Fitzpatrick and Dent 1997]. The community search assistant enables users to tap into other users queries by building a query graph that measures relatedness based on the documents returned by the query and not the query terms themselves [Glance 2001b].

2.3 Information Delivery and Management

Information delivery and management is what happens after the user has typed in a query. Query results may be processed (e.g., filtered, integrated, etc.) before presenting them to the user and the subsequent information organized and managed for the user's future use. User profiles often filter the query results. The BASAR agent system [Thomas and Fischer 1997b] stores a search query, results, and subsequent user actions in a user's profile to use to filter future queries. For example, it may notice that a user prefers information from *edu* sites and filters search results accordingly. DubLet [Hurley and Wilson 2001a] recommends rental properties culled from online advertisements. It relies on the intrinsic structure of such advertisements to extract various features such as monthly rent, location, and number of bedrooms, into a common representation. Users generate profiles by checking off features that matter most to them. One can dynamically adjust user profiles by requesting that recommended properties be, for example, cheaper or nicer than the current advertisement. The system then filters the result based on this dynamic user profile.

Collaborative filtering agent approaches are more common for filtering movies, music, or news rather than general text searches [Mladenic 1999]. However, in the SOAP agent system [Voss and Kreifelts 1997b], task agents use recommender agents to find out about similar past user search results and recommendations on a per-task basis. Users rate results returned by the task agent, and these ratings are in turn passed back to the recommender agent. The W^{3}IQ agent [Joshi 2000] acts as a proxy and supports information-seeking on mobile, low bandwidth devices by using collaborative filtering to better ensure that the information sent is truly relevant. In Good et al. [1999], collaborative filtering involves not only communities of users, but also communities of users and recommender agents.

Ariadne [Knoblock et al. 2001b] is an information integration agent system that allows users to easily create information agents that can access and integrate information across multiple sites. Given a domain model (e.g., locating theatres and restaurants for an evening out), Ariadne creates a plan for querying the appropriate sources and determines how to integrate the data into a meaningful result. Thus, if a user wants to see a certain film and wants to find a nearby restaurant, an Ariadne-based agent might plan to first find theatres showing that film and then use those locations to find nearby restaurants. The SportsAgent [Lu and Sterling 2000] system answers sports queries by combining and integrating the services of several specialized information agents. SportsAgent uses a mediator agent to determine how to combine services to answer queries and to coordinate the activities of the information agents.

Once the query results have been processed, a key component of information management is managing trade-offs between the value of each piece of information and the human attention costs of keeping more and more information. As such, it falls under the more general rubric of personal information management. One of the key problems is information fragmentation, caused in part by the abundance of inconsistent, uncoordinated organizational tools [Jones 2004]. Agents have the potential to help by either talking a common language or by using wrappers to translate from one schema to another [Michalowski et al. 2004].

2.4 Agent Searching on Web Search Engines

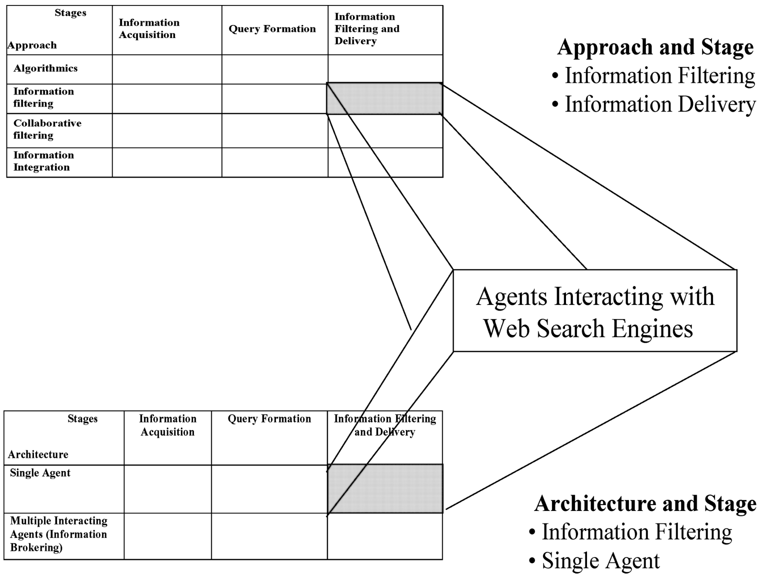
The use of search engines by agents is an example of information filtering and information delivery. Most Web search engines, such as Alta Vista (www.altavista.com) and Google (www.google.com), employ agents as crawlers [Sullivan 2002]. In addition to these general-purpose search engines, niche search engines also employ agents. For example, Lawrence [1999] developed CiteSeer, which incorporates a software agent to locate computer articles on the Web. By utilizing these agents to gather and organize online data, many of the general and niche search engines have become valuable information resources.

As a result of these large information repositories, others retrieve information from the search engines for personal, commercial, and other purposes. Although humans conduct much of this searching, some use agents to retrieve the information. Commercial examples include metacrawlers search engines, such as Ithaki (http://www.ithaki.net/indexu.htm) or the more popular Dogpile (http://www.dogpile.com). Unlike standard search engines, metacrawlers do not crawl the Web themselves to build listings. Instead, they use automated applications to send queries to several search engines simultaneously. The metacrawler than blend the results from all the queried search engines together onto one results listing. Other companies utilize metacrawler software to locate job information, evaluate page rankings, or locate bargains for certain products or services. Research is still ongoing in the metasearch area. For instance, Chen [2001b] are developing an intelligent Web metaindexer for Web searching which is a stand-alone system that utilizes results from existing 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 [SearchTools.com 2001], and designing Web agents is now a fairly common student project in many university courses [Berry and Browne 1999; Youngblood 1999]. Additionally, there are several inexpensive commercial applications that provide metacrawler software that runs from a desktop computer [Sullivan 2003].

Figure 1 shows the classification of these agents that interact with Web search engines. Referring to our approach and stages taxonomy, these types of agents programmatically handle information filtering and delivery and are usually single agent, although their results may be aggregated with results from other agents. It is this classification of agent that we examine further in this article.

It has been stated that these Web agents offer a more sophisticated method of searching for information on the Web [Etzioni 1996b]. There is a significant amount of literature on Web agents and their use by Web search engines to gather information [Arasu et al. 2001]. There is also significant research into methods to optimize agent information gathering to avoid unnecessary loads on servers or the network [Shkapenyuk and Suel 2002; Talim et al. 2001; Xiaohui



Stages Approach	Information Acquisition	Query Formation	Information Filtering and Delivery
Algorithmics			
Information filtering			▨
Collaborative filtering			
Information Integration			

Stages Architecture	Information Acquisition	Query Formation	Information Filtering and Delivery
Single Agent			▨
Multiple Interacting Agents (Information Brokering)			

Fig. 1. Approach, stage, and architecture of search engine querying agents.

et al. 2001]. However, the actual searching characteristics of these Web agents has not been investigated, even though for some time there have been questions about their effect on information providers [Selberg and Etzioni 1995].

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 [Brody 2000]. In education, for example, research articles online have a higher citation rate relative to those articles not online [Lawrence 2001]. In job seeking, niche job boards have dramatically altered the hiring process [Cappelli 2001]. 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 [Hernandez et al. 2001]. The entire topic of what information the search engines provide or do not provide may have a dramatic effect on which people or organizations are successful [Introna and Nissenbaum 2000].

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 interact with search engines is of great importance. The results of this research have ramifications in terms of system design, ecommerce, network performance, and the societal effects of the Web. These considerations are the drivers for this research.

3. RESEARCH QUESTIONS

Specifically, the research questions driving this study are the following

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

What types of information are Web agents retrieving?

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

To address these research questions, we obtained and quantitatively and qualitatively analyzed actual queries and result page requests submitted to Excite and AltaVista, two major. Web search engines, by Web agents.

4. RESEARCH DESIGN

We collected the data used in this study at two Web search engines. Two of the transaction logs are from Excite (http://www.excite.com), a major Web search engine during the time of the data collection. Each transaction log holds a large and varied set of queries. The transaction logs spanned several hours of user searching on the following dates: 16 September 1997 (Tuesday, midnight to 8 a.m.) and 30 April 2001 (Monday, midnight to midnight). Excite was the second most popular Web site in 1997 [Munarriz 1997], and was the fifth most popular in 2001 as measured by the number of unique visitors [Cyber Atlas 1999; 2001]. In 2002, Excite changed its business model from a search engine to an information portal.

The third transaction log is from the AltaVista search engine. The information in the transaction log was collected on 8 September 2002 and spans a 24-hour period. The queries recorded in the transaction log 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 [CyberAtlas 2002].

4.1 Data Collection

The original transaction logs contain a varied but substantial amount of queries, approximately 1,200,000 to 3,200,000 records. Each record contains at least three fields: (1) time of day measured in hours, minutes, and seconds from midnight of each day as recorded by the Excite and Alta Vista servers; (2) user identification an anonymous user code assigned by the search engine 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 [Jansen and Pooch 2001]. Briefly, a term is any series of characters separated by white space. A query is the entire string of terms submitted 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 logs contain searches from both human users and non-human agents. We were interested in only those queries submitted by agents. From the original transaction logs, we therefore extracted a subset of queries that we deemed were submitted by agents. Unfortunately, there is no definitive way to tell if an interaction with a search engine is from a human or an agent [Silverstein et al. 1999]. So, one must use heuristic methods.

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 [Jansen and Pooch 2001] 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 probable 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 document and returns to the search engine, the search engine server logs this second visit 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 result pages the agent visited from the search engine, but, unfortunately, it also skews the results in analyzing how the agents searched on the system relative to the number of queries and frequency of terms.

To address the first research question, we collapsed the data set by combining all identical queries [Jansen and Pooch 2001] submitted by the same agent. This gave us unique queries in order to analyze sessions, queries, and terms.

For the second and third research question, we utilized the complete uncollapsed 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 analysis. Earlier results from an analysis of agent searching on the AltaVista search engine appear in Jansen et al. [2003a; 2003b].

5.1 Agent Searching Behavior

Table III presents the general searching information of the agent-search engine interactions.

At the term level of analysis, the number of unique terms (17% to 37%) was very low compared to human searchers (57% to 61%) [Spink et al. 2002], indicating a tight jargon used by Web agents and limited subject matter. The use of the 100 most frequently occurring terms (21% to 28%) submitted by agents was also high compared to human searchers (usually well under 20%) [Spink et al. 2002].

Examining the query level, Web agent queries are comparable to queries submitted by human Web searchers. About 40% to 46% of agent queries contained 3 or more terms compared to about 45% for human searchers [Spink et al. 2002]. The standard deviation (1.7 to 2.6) is about twice that of human searchers, however [Jansen and Pooch 2001]. The use of Boolean operators by agents on the Excite search engine is nearly the same as human searchers [Spink et al. 2002], but the percentage is about double that of human searchers on the AltaVista search engine (approximately 10%) [Jansen et al. 2005], although it still represents a minimal usage at 20%.

Table III. Aggregate Results for Agent General Search Trends

	Excite 1997	Excite 2001	AltaVista 2002
Sessions	145	170	2,717
Queries	5,756	6,024	896,387
Terms			
Unique	5,534 (37%)	5,779 (34%)	570,214 (17%)
Total	14,760	16,825	3,224,840
Terms per query	mean = 2.56 (sd = 1.68)	mean = 2.79 (sd = 1.84)	mean = 3.6 (sd = 2.6)
Terms per query			
1 term	1,320 (23%)	1,462 (24%)	216,105 (26%)
2 terms	1,964 (34%)	1,749 (29%)	268,076 (30%)
3+ terms	2,358 (41%)	2,813 (47%)	411,988 (46%)

Queries per Agent	mean = 39.7
	sd = 45.3
		mean = 35.4
		sd = 102.5
			mean = 329.9
			sd = 1883.9

Agents modifying queries	145 (100%)	158 (93%)	2,386 (88%)
Session size			
1 query	0 (0%)	12 (7%)	331 (12%)
2 queries	3 (2%)	6 (4%)	109) (4%)
3+ queries	142 (98%)	152 (89%)	2,277 (84%)
Results Pages Viewed			
1 page	3,287 (57%)	2,419 (40%)	760,071 (85%)
2 pages	972 (17%)	430 (7%)	67,755 (8%)
3+ pages	1,483 (26%)	3,715 (53%)	68,561 (8%)
Boolean Queries	184 (3%)	173 (3%)	177,182 (20%)
Terms not repeated in data set	3,126 (21%)	3,685 (22%)	411,577 (13%)
Use of 100 most frequently occurring terms	3,085 (21%)	16,320 (28%)	834,251 (26%)

In terms of results pages, nearly 85% of the Web agents access only the first page of results on AltaVista, which is higher than human searchers (72%) [Jansen et al. 2005]. On the Excite search engine, the variance is similar with 40% to 57% of agents viewing only the first page versus human searchers at 29% to 42% [Spink et al. 2002].

At the session level of analysis, the percentage of agent sessions with more than three queries (84% to 98%), after duplicate queries were removed was significantly higher than that of human searchers (approximately 25%) [Jansen and Spink 2005; Jansen et al. 2000, 2005; Spink et al. 2002].

Table IV. Information Agent Approaches and Stages

	Excite 1997		Excite 2001		AltaVista 2002	
Query Operators	Human	Agent	Human	Agent	Human	Agent
Total Queries In Data Set	545,206	5,756	1,025,910	6,024	1,073,388	896,387
Total Boolean (AND, OR, NOT)	30,602 (2.9%)	184 (3.2%)	72,250 (7.0%)	173 (2.9%)	50,584 (4.7%)	177,182 (19.8%)
Total Operators (“, +, −)	66,271 (6.4%)	270 (4.7%)	44,378 (4.3%)	904 (15.0%)	140,587 (22.5%)	94,610 (10.6%)
Total	96,873 (9.3%)	545 (7.9%)	116,628 (11.3%)	1,077 (17.9%)	191,171 (27.3%)	271,792 (30.4%)

5.2 Query Syntax Usage

We examine agent searching at the query level of analysis in order to get a better understanding of how agents structure their queries. In Table IV, we present the top term occurrences for the agent data set. Percentages and numbers for human query operators are from Spink and Jansen [2004].

Table IV shows that Boolean and Web query operator usage for agents is comparable to that of human searchers from each data set. The usage of operators, especially with the Excite search engine, is relative low compared to usage on more traditional IR systems. The low usage of Web query operators places even greater emphasis on the proper selection of Web query terms. Certainly, the agent searching does not appear to be more sophisticated or complex than that of human searching human searching.

5.3 Term Level Analysis

We further examine agent searching at the term level of analysis in order to get a better understanding of what types of information the agents are commonly searching. In Table V, we present the top term occurrences for the agent data set.

The bolded cells represent the most frequently occurring term from that data set. NA means that the term did not appear in the top terms for that data set. The percentages from AltaVista are comparable to human searchers on that search engine. For example, the most frequently occurring term from human searchers was *free* (18,404 occurrences representing 0.6% of all terms) followed by *sex* (7,771 occurrences representing 0.2%) [Spink and Jansen 2004]. For the Excite search engine, the percentages for agents are higher than terms used by human searchers. For 1997, the most frequently occurring term (*of*) represented 1.49% of all term occurrences compared to *free* (the most common human used term) at 0.9%. In 1999, the most commonly used term by agents was *return* (2.23%) compared to *free* which was used 1.1% by human searchers.

In order to gain additional insight into the information needs of these agents, we qualitatively analyzed a random sample of approximately 2,600 queries from each of the three data set, into 12 general topic categories developed by [Spink et al. 2002]. Two independent evaluators manually classified each of the queries independently. The evaluators then met and resolved discrepancies. Table VI displays the evaluation results.

We see from Table VI that agents are predominantly searching for information on *People, Places, and Things* with approximately 35% of all agent queries

Table V. Aggregate Terms—Level for Agent General Search Trends

Terms	Excite 1997	Percentage	Excite 2001	Percentage	AltaVista 2002	Percentage
center	25	0.17%	19	0.11%	**2,442**	**0.08%**
fitness	19	0.13%	20	0.12%	2,326	0.07%
real	21	0.14%	38	0.23%	2,265	0.07%
estate	19	0.13%	33	0.20%	2,256	0.07%
sale	15	0.10%	376	2.23%	2,183	0.07%
find	19	0.13%	20	0.12%	387	0.01%
fax	NA	NA	NA	NA	1,905	0.06%
of	**220**	**1.49%**	114	0.68%	533	0.02%
maps	22	0.15%	55	0.33%	432	0.01%
volleyball	72	0.49%	131	0.78%	283	0.01%
in	125	0.85%	79	0.47%	248	0.01%
can	20	0.14%	27	0.16%	235	0.01%
you	26	0.18%	49	0.29%	213	0.01%
football	28	0.19%	178	1.06%	190	0.01%
the	156	1.06%	66	0.39%	173	0.01%
free	83	0.56%	102	0.61%	169	0.01%
return	21	0.14%	**376**	**2.23%**	168	0.01%
basketball	19	0.13%	150	0.89%	138	0.00%
on	40	0.27%	21	0.12%	134	0.00%
and	201	1.36%	296	1.76%	102	0.00%

appearing in this category. There has been a noticeable drop in *Sex or pornography* searching among agents, reflecting a trend noted in that of human searching on search engines as well [Spink et al. 2002].

Table VII shows the corresponding percentages in each topic from an analysis of human searching from each data set. Percentages in Table VII are from Spink and Jansen [2004].

Both Table VI and VII are organized using the 2002 data set in descending order. We see that the topic *People, places or things* is the top category in 2002 for both agent and human searching (36% for agent and 49% for human searchers). However, the second category for agents is *Unknown*. The vast majority of these queries were of a standard structure, 5 terms OR'd together. These are most likely queries from agents of Web optimization companies to gather ranking results. Example queries of this type are: *abaciscus OR intradural OR swinehead OR washman OR unparcel* and *adirondacks OR defensive OR groomed OR badgers OR distortion*. The other topical domains between agent and human searchers generally follow a similar ranking, although the percentages for agent searching are lower.

5.4 Session Duration

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

An interaction is either a query submission or a request to view a results page. These interactions occur over a session. Referring to Table VIII, the **Duration of Interaction** shows the average, standard deviation, maximum, and

Table VI. Distribution of General Topic Categories for Agent Searching

Rank	Categories	Excite 1997 (2,094 queries)	Excite 2001 (2,700 queries)	AltaVista (2,717 Queries)
1	People, places or things	**742** (35.4%)	**945** (35%)	**986** (36.3%)
2	Unknown	20 (1.0%)	756 (28%)	822 (30.3%)
3	Computers or Internet (or technology stuff)	207 (9.9%)	270 (10%)	236 (8.7%)
4	Commerce, travel, employment or economy	239 (11.4%)	162 (6%)	165 (6.1%)
5	Entertainment or recreation (music, TV, sports)	163 (7.8%)	81 (3%)	135 (5.0%)
6	Health or sciences (physics, math)	126 (6.0%)	54 (2%)	90 (3.3%)
7	Sex or pornography	352 (16.8%)	108 (4%)	79 (2.9%)
8	Education or humanities	83 (4.0%)	81 (3%)	64 (2.4%)
9	Society, culture, ethnicity or religion	96 (4.6%)	81 (3%)	60 (2.2%)
10	Other	12 (0.6%)	54 (2%)	43 (1.6%)
11	Performing or fine arts (i.e., ballets, plays, etc)	9 (0.4%)	54 (2%)	19 (0.7%)
12	Government (or military)	45 (2.1%)	54 (2%)	18 (0.7%)
		2,094 (100.0)%	2700 (100%)	2,717 (100.0%)

minimum duration of agent interaction. The **Interactions During Period** shows the same statistics for the number of interaction during the data collection periods. The **Interactions Per Second** show the statistics for number of interactions occurring per second.

The duration and frequency of agent-search engine interactions is substantially different than that of human searchers. The mean agent session (approximately 6-½ to 9-1/2 hours) is approximately 38 times the mean human session of 15 minutes [Jansen and Spink 2003]. However, the standard deviation was relatively high at just over 6 to 8 hours. The maximum session duration was the full temporal span of the data sampling periods. The minimum duration was 1 second.

The mean number of interactions per session (150 to 615) is 75 to 300 times that of human searchers (just over 2) [Jansen et al. 1998]. Again, the standard deviations were quite large, indicating that different agents or types of agents adhere to different interaction patterns.

Table VII. Distribution of General Topic Categories for Human Searching

Rank	**Categories**	**1997** **(2,414 queries)**	**2001** **(2,453queries)**	**2002** **(2,603 Queries)**
1	People, places, or things	162 (6.70%)	483 (19.70%)	**1,282** (49.27%)
2	Commerce, travel, employment, or economy	321 (13.30%)	**606 (24.70%)**	326 (12.52%)
3	Computers or Internet	302 (12.50%)	235 (9.60%)	232 (12.40%)
4	Health or sciences	229 (9.50%)	184 (7.50%)	195 (7.49%)
5	Education or humanities	135 (5.60%)	110 (4.50%)	132 (5.07%)
6	Entertainment or recreation	**435 (19.90%)**	162 (6.60%)	119 (4.57%)
7	Sex and pornography	406 (16.80%)	209(8.50%)	85 (3.26%)
8	Society, culture, ethnicity, or religion	138 (5.70%)	96 (3.90%)	81 (3.11%)
9	Government (or military)	82 (3.40%)	49 (2.00%)	41 (1.57%)
10	Performing or fine arts	130 (5.40%)	27 (1.10%)	18 (0.69%)
11	Non-English or unknown	75 (4.10%)	227 (11.30%)	0%)
12	Other	0%)	15 (<.05%)	1 (∼0%)

Using the 2002 data, the average agent submits a query or views a results page about every 2 seconds with a standard deviation of approximately 4 interactions. The maximum session frequency was just less than 100,000 queries in the 24-hour span and the maximum queries per second were 137.

6. DISCUSSION

We examined data from three data sets of recorded interactions by agents with Web search engines, which is in the information filtering/information delivery of our approach and stage taxonomy, and in the single agent/information filtering category of our architecture taxonomy. We analyzed patterns of information filtering and information delivery agents interacting with two major Web search engines. Naturally, the results reported here are general patterns of interaction. Just as with human searchers, agents are (may be) different with possible different patterns of interaction.

We now return to our research questions, which are as follows.

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

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

Table VIII. Time, Interactions, and Interactions Per Second

Duration of Interaction	Excite 1997	Excite 2001	AltaVista 2002
Hours: Minutes: Seconds			
Average	9:27:49	6:22:49	9:27:30
St Dev	6:03:56	6:45:09	8:05:49
Max	23:39:23	23:59:56	23:59:57
Min	0:00:01	0:09:21	0:00:02
Interactions During Period			
Average	150	319	615
St Dev	66	1,819	2,609
Max	482	23,666	99,595
Min	100	100	101
Interactions Per Second			
Average	0.04	0.03	0.43
St Dev	0.25	0.04	4.17
Max	2.55	0.27	137
Min	<0.00	<0.00	<0.00

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

6.1 Agent Searching Characteristics

Agents interacting with Web search engines use queries similar to those submitted by human searchers. Agents submit very short and generally simple queries, but they are persistent in submitting queries with over 84% to 98% of agents submitting more than 3 queries, and the mean interactions ranging from 39 to 329 queries per agents. As with humans, most agents are not interested in viewing a lot of results, although some agents did exhibit this behavior by viewing hundreds of results pages. Due to the nature of transaction logging, we cannot classify the agents into any type of grouping. However, using the averages and standard deviations, we see that there is, at times, large variation in agent interaction behavior.

6.2 Frequency and Duration of Interaction

The agent-search engine interaction typically takes place over several hours with multiple instances of interaction every few seconds. This is substantially longer than human searchers, but this does not necessary mean the agents are more sophisticated in their searching approach. Although the mean duration was about nine and a half hours, several agent interactions continued for the entire 24-hour period. Using the 2002 data, the maximum frequency of interaction was over 600 interactions per second, 137 of these as page requests. This means the agent was viewing, and possibly downloading, over 1,370 Web documents a second, assuming a standard ten-results page view. The mean interaction was about one query every 2 seconds. The lack of an external economic incentive may be contributing to the inefficient but high-volume searching employ by these agents.

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

6.3 Agent Information Seeking

Agents are searching for a fairly limited variety of information with less than 37% to 18% of the terms used unique. This small number of terms indicates that the agents are searching for targeted sets of topics. This is was also evident in the percentage of terms not repeated in the data set (13% to 22%). Similar to human searchers, agents are interested in information about people, places, and things.

7. IMPACT AND CONCLUSION

As one of the first studies of how agents are searching on the Web, this study contributes to our understanding of Web searching in several important ways. First, we provide a classification method along stages, approaches, and architecture axes. With this classification methodology, we provide an extensive literature review of the current state of the field. Second, the data in our analysis 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 outside of the lab. Third, the sample is quite large, with approximately 900,000 queries submitted by over 3,000 agents. Fourth, we collected the data from two very popular search engines as measured by both document collection and number of unique visitors to ensure that our results were generalizable. Fifth, it is one of the few trend comparisons of agent versus human Web searching available, or human, providing valuable insight into the changing patterns of Web searching interaction. Finally, concerning methods, the study illustrates that transaction log analysis is a viable method for analyzing real agents interacting with real systems in the complex environment of the Web. This complex environment is difficult to recreate in a laboratory setting [Dumais 2002].

The study also has limitations. First, there is no accurate way to distinguish an agent query from a human query. This is certainly an area for possible future research and system development. Second, the sample data comes from only two major Web search engines and three sampling periods, introducing the possibility that the queries do not represent the queries submitted by the broader Web agent population. However, Jansen and Pooch [2001] suggest that characteristics of human searchers are very consistent across search engines. We can hypothesis that this may hold for agents also. However, prior work has also shown that particular dates do affect term usage, so one could expect some variation in this regard.

Overall, the results reported in this study provide a useful characterization of agent information searching and gives insight into the queries, terms and terms pairs that are most frequently used. The study shows that agent searching is similar to human searching with the exceptions of the duration and speed of interaction. Agent searching is certainly not substantially more sophisticated than that exhibited by human searchers. Equipped with this information, search engines developer and other Web information providers can design their Web sites to accommodate or hinder these automated information gathers. Further research should continue to examine the changing trends in automated searching and begin to explore more directly the manner in which agents use search engines to locate information. From this analysis, one can then draw inferences concerning the economic, technical, and social effects of agent Web searching.

ACKNOWLEDGMENTS

Our thanks to Excite and AltaVista for providing access to the data without which we could not have conducted this research.

REFERENCES

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A. and Raghavan, S. 2001. Searching the Web. *ACM Trans. Internet Techn.* 1,1, 2–43.

Aridor, Y., Carmel, D., Maarek, Y. S. and Soffer, A. 2002. Knowledge encapsulation for focused search from pervasive devices. *ACM Trans. Inform. Sys.* 20,1, 25–46.

Barish, G., Chen, Y., Knoblock, C. A., Minton, S., Philpot, A. G. and Shahabi, C. 2000. The TheatreLoc virtual application. In *Proceedings of 12th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI'00).* 980–987.

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

Bollacker, K. D., Lawrence, S. and Giles, C. L. 1998. CiteSeer: An autonomous Web agent for automatic retrieval and identification of interesting publications. In *Proceedings of the 2nd International ACM Conference on Autonomous Agents.* 116–123.

Brandman, O., Cho, J., Garcia-Molina, H. and Shivakumar, N. 2000. Crawler-friendly Web servers. In *Proceedings of the Workshop on Performance and Architecture of Web Servers (PAWS).* Santa Clara, California.

Broder, A. Z., Najork, M. and Wiener, J. L. 2003. Efficient URL caching for world wide web crawling. In *Proceedings of the 12th International Conference on World Wide Web (WWW).* Budapest, Hungary, 680–689.

Brody, R. 2000. Illusions of plenty: The role of search engines in the structure and suppression of knowledge. In *Proceedings of the IEEE International Symposium on Technology and Society.* Rome, Italy, 157–161.

Budzik, J. and Hammond, K. 1999. Watson: Anticipating and Contextualizing Information Needs. In *Proceedings of the 60nd Annual Meeting of the American Society for Information Science.* 727–740.

Cappelli, P. 2001. Making the most of online recruiting. *Harvard Bus. Rev.* 79,3, 139–146.

Cesarano, C., D'acierno, A. and Picariello, A. 2003. An Intelligent Search Agent System for Semantic Information Retrieval on the Internet. In *Proceedings of the 5th ACM International Workshop on Web Information and Data Management (WIDM'03).* New Orleans, LA. 111—117.

Chakrabarti, S., Van Den Berg, M. and Dom, B. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Comput. Netw.* 31,11–16, 1623–1640.

Chen, C. C., Chen, M. C. and Sun, Y. 2001a. PVA: A self-adaptive personal view agent system. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (SIGKDD'01).* San Francisco, CA, 257–262.

Chen, J., Dewitt, D. J., Tian, F. and Wang, Y. 2000. Niagara CQ: A scalable continuous query system for internet databases. In *Proceedings of SIGMOD.* 379–390.

Chen, L. and Sycara, K. 1998. WebMate: A personal agent for browsing and searching. In *Proceedings of the 2nd International Conference on Autonomous Agents and Multi Agent Systems (AGENTS '98).* 132–139.

CHEN, Z. X., MENG, X. N., FOWLER, R. H. AND ZHU, B. 2001b. Features: Real-time adaptive feature and document learning for Web search. *J. Amer. Soc. Inform. Science.* 52,8, 655–665.

CYBER ATLAS. 1999. U.S. top 50 internet properties, Dec. 1999, at home/work combined. 1 (July 2000).

CYBER ATLAS. 2001. U.S. top 50 internet properties, (May 2001) at home/work combined. (July 2000).

CYBER ATLAS. 2002. (Nov. 2002) internet usage stats. (Jan. 2002).

DILIGENTI, M., COETZEE, F. M., LAWRENCE, S., GILES, C. L. AND GORI, M. 2000. Focused crawling using context graphs. In *Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000).* 527–534.

DOORENBOS, B., ETZIONI, O. AND WELD, D. 1997. A scalable comparison-shopping agent for the World Wide Web. In *Proceedings of the 1st International Conference of Autonomous Agents (AGENTS-97).* Marina Del Ray, CA. 39–48.

DUMAIS, S. T. 2002. Web experiments and test collections. *The 11th International World Wide Web Conference.* 2003 (April),

EDWARDS, J., MCCURLEY, K. AND TOMLIN, J. 2001. An adaptive model for optimizing performance of an incremental web crawler. In *Proceedings of the World Wide Web 10 Conference (WWW10).* Hong Kong, 106–113.

ETZIONI, O. 1996a. Moving Up the information food chain: Deploying softbots on the World Wide Web. In *Proceedings of the 13th National Conference on Artificial Intelligence and the 8th Innovative Applications of Artificial Intelligence Conference.* 1322–1326.

FITZPATRICK, L. AND DENT, M. 1997. Automatic feedback using past queries: social searching? In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR97)* Philadelphia, PA, 306–313.

FLAKE, G. W., GLOVER, E. J., LAWRENCE, S. AND GILES, C. L. 2002. Extracting query modifications from nonlinear SVMs. In *Proceedings of the 11th International World Wide Web Conference (WWW'02).* Honolulu, HI, 317–324.

GHANI, R., JONES, R. AND MLADENIC, D. 2001a. Online learning for query generation: Finding documents matching a minority concept on the web. In *Proceedings of Asia-Pacific Conference on Web Intelligence.* 508–513.

GLANCE, N. S. 2001a. Community search assistant. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI'01).* Sante Fe, NM, 91–96.

GLOVER, E. J., FLAKE, G. W., LAWRENCE, S., BIRMINGHAM, W., KRUGER, A., GILES, C. L. AND PENNOCK, D. 2001. Improving category specific Web search by learning query modifications. In *Proceedings of IEEE Symposium on Application and the Internet (SAINT).* 23–31.

GOOD, N. G., SCHAFER, J. B., KONSTAN, J. A., BORCHERS, A., SARWAR, B., HERLOCKER, J. AND RIEDL, J. 1999. Combining collaborative filtering with personal agents for better recommendations. In *Proceedings of the 1999 Conference of the American Association of Artificial Intelligence (AAA-99).* 439–446.

HENDLER, J. 2001. Agents and the semantic Web. *IEEE Intelligent Syst. 16*,2, 30–37.

HERNANDEZ, J. C., SIERRA, J. M., RIBAGORDA, A. AND RAMOS, B. 2001. Search engines as a security threat. *Comput.* 34,10, 25–30.

HÖLSCHER, C. AND STRUBE, G. 2000. Web search behavior of internet experts and newbies. *Int. J. Comput. Telecomm. Networ.* 33,1-6, 337–346.

HUHNS, M. N. AND SINGH, M. P. 1998. Personal assistants. *IEEE Internet Comput.* 2,5, 90–92.

HURLEY, G. AND WILSON, D. C. 2001a. DubLet: An online CBR system. In *Proceedings of the 4th International Conference on Case-Based Reasoning, ICCBR'01.* Vancouver, BC, Canada.

INTRONA, L. AND NISSENBAUM, H. 2000. Defining the Web: The politics of search engines. *Comput.* 33,1, 54–62.

JANSEN, B. J. AND POOCH, U. 2001. Web user studies: A review and framework for future work. *J. Amer. Soc. Inform. Science Techn.* 52,3, 235–246.

JANSEN, B. J. AND SPINK, A. 2003. An analysis of Web information seeking and use: Documents retrieved versus documents viewed. In *Proceedings of the 4th International Conference on Internet Computing.* Las Vegas, NV, 65–69.

JANSEN, B. J. AND SPINK, A. 2005. An analysis of Web searching by European Alltheweb.com users. *Inform. Process. Manag. 42*, 1, 248–263.

JANSEN, B. J., SPINK, A., BATEMAN, J. AND SARACEVIC, T. 1998. Real life information retrieval: A study of user queries on the Web. *SIGIR Forum.* 32,1, 5–17.

JANSEN, B. J., SPINK, A. AND PEDERSON, J. 2003a. Monsters at the gates: When Softbots visit web search engines. In *Proceedings of the 4th International Conference on Internet Computing.* Las Vegas, NV, 620–626.

JANSEN, B. J., SPINK, A. AND PEDERSON, J. 2003b. Web searching agents: What are they doing out there? In *Proceedings of the 2003 IEEE International Conference on Systems, Man and Cybernetics.* Washington, DC, 10–16.

JANSEN, B. J., SPINK, A. AND PEDERSON, J. 2005. Trend analysis of altaVista Web searching. *J. Amer. Soc. Inform. Science Techn. 56*, 6, 559–570.

JANSEN, B. J., SPINK, A. AND SARACEVIC, T. 2000. Real life, real users, and real needs: A study and analysis of user queries on the Web. *Inform. Process. Manag.* 36,2, 207–227.

JOACHIMS, T., FREITAG, D. AND MITCHELL, T. 1997. WebWatcher: A tour guide for the World Wide Web. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI 97).* 770–775.

JONES, W. 2004. Finders, keepers? The present and future perfect in support of personal information management. *First Monday.* 9, 3.

JOSHI, A. 2000. On proxy agents, mobility, and web access. *Mobile Netw. Appl.* 5, 233–241.

KNOBLOCK, C. A., MINTON, S., AMBITE, J. L., ASHISH, N., MUSLEA, I., PHILPOT, A. G. AND TEJADA, S. 2001a. The Ariadne approach to Web-based information integration. *Int. J. Coopera. Inform. Syst. (IJCIS).* 10,12, 145–169.

KOSTER, M. 1998. The Web robots FAQ. www.robotstxt.org/wc/faq.html 15 (March 2002).

LAWRENCE, S. 2001. Online or invisible? *Nature.* 411,6837, 521.

LAWRENCE, S., GILES, C. L. AND BOLLACKER, K. 1999. Digital libraries and autonomous citation indexing. *IEEE Comput.* 32,6, 67–71.

LEE, G., LEE, J.-H., RHO, H., PARK, Y.-T., CHOI, J. AND SEO, J. 1998. Interactive NLI agent for multiagent Web search model. In *Proceedings of the International Workshop on Intelligent Agents on the Internet and Web, in 4th World Congress on Expert Systems.* Mexico City, Mexico, 67–74.

LIEBERMAN, H., FRY, C. AND WEITZMAN, L. 2001. Exploring the Web with reconnaissance agents. *Comm. ACM.* 44, 8, 69–75.

LIN, S.-D. AND KNOBLOCK, C. A. 2003. Exploiting a search engine to develop more flexible Web agents. In *Proceedings of IEEE/WIC International Conference on Web Intelligence.* 54–60.

LU, H. AND STERLING, L. 2000. Interoperability and semi-structured data in an open Web-based agent information system. In *Proceedings of Proceedings of the Workshop on Information Systems Engineering (WISE00).* Hong Kong, 80–86.

MADDEN, S., SHAH, M., HELLERSTEIN, J. M. AND RAMAN, V. 2002. Continuously adaptive continuous queries over streams. In *Proceedings of the ACM SIGMOD International Conference on Management of Data.* Madison, WI, 49–60.

MARTIN, I. AND JOSE, J. M. 2003. A personalized information retrieval tool. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03).* Toronto, Canada, 423–424.

MENCZER, F., PANT, G., SRINIVASAN, P. AND RUIZ, M. E. 2001. Evaluating topic-driven Web crawlers. In *Proceedings of the 24th Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'01).* New Orleans, LA. 241–249.

MICHALOWSKI, M., AMBITE, J. L., KNOBLOCK, C. A., MINTON, S., THAKKAR, S. AND TUCHINDA, R. 2004. Retrieving and semantically integrating heterogeneous data from the Web. *IEEE Intelligent Syst.* 19, 3, 72–79.

MILLER, E. 2004. The W3C's Semantic Web activity: An update. *IEEE Intelligent Syst.* 19, 95–97.

MLADENIC, D. 1999. Text-learning and related intelligent agents: A survey. *IEEE Intelligent Syst.* 14, 4, 44–54.

MUNARRIZ, R. A. 1997. How did it double? www.tool.com/ddouble/1997/ddouble 970812 html/. 10 November,

PANT, G. AND MENCZER, F. 2002. MySpiders: Evolve your own intelligent web crawlers. *Autonom. Agents Multi-Agent Syst.* 5, 221–229.

PITKOW, J., SCHUTZE, H., CASS, T., COOLEY, R., TURNBULL, D., EDMONDS, A., ADAR, E. AND BREUEL, T. 2002. Personalized search. *Commu. ACM.* 45, 9, 50–55.

Rhodes, B. J. and Maes, P. 2000. Just-in-time information retrieval agents. *IBM Syst J.* 39, 3 & 4, 685–704.

Rowe, N. C. 2002a. Marie-4: A high-recall, self-improving web crawler that finds images using captions. *IEEE Intelligent Syst.*

Rowe, N. C. 2002b. Marie-4: A high-recall, self-improving web crawler that finds images using captions. *IEEE Intelligent Syst.* 17, 4, 8–15.

searchtools.com. 2001. Source Code for Web Robot Spiders.

Selberg, E. and Etzioni, O. 1995. Multi-service search and comparison using the metacrawler. In *Proceedings of the 4th International World-Wide Web Conference.* Boston, MA.

Shkapenyuk, V. and Suel, T. 2002. Design and implementation of a high-performance distributed Web crawler. In *Proceedings of the 8th International Conference on Data Engineering.* San Jose, CA, 357–368.

Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. 1999. Analysis of a very large Web search engine query log. *SIGIR Forum.* 33, 1, 6–12.

Somlo, G. and Howe, A. E. 2001. Adaptive lightweight text filtering. In *Proceedings of Intelligent Data Analysis (IDA'01).* Lisbon, Portugal.

Somlo, G. and Howe, A. E. 2003a. Using Web helper agent profiles in query generation. In *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'03).* Melbourne, Australia, 812–818.

Spink, A. and Jansen, B. J. 2004. *Web Search: Public Searching of the Web.* Kluwer, New York, NY.

Spink, A., Jansen, B. J., Wolfram, D. and Saracevic, T. 2002. From E-sex to E-commerce: Web search changes. *IEEE Comput.* 35, 3, 107–111.

Sullivan, D. 2002. Search Engine Math. www.searchenginewatch.com/showPage.html 11 April,

Sullivan, D. 2003. Search Utilities. www.searchenginewatch.com 16 (March 2002).

Talim, J., Liu, Z., Nain, P. and Coffman, E. G. 2001. Controlling the robots of Web search engines. In *Proceedings of ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems.* Cambridge, MA, 236–244.

Thomas, C. G. and Fischer, G. 1997a. Using agents to personalize the Web. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI 97).* Orlando, FL, 53–60.

Voss, A. and Kreifelts, T. 1997a. SOAP: Social agents providing people with useful information. In *Proceedings of the international ACM SIGGROUP Conference on Supporting Group Work (Group97).* Phoenix, AZ, 291–298.

Wolf, J. L., Squillante, M. S., Yu, P. S., Sethuraman, J. and Ozsen, L. 2002. Optimal crawling strategies for Web search engines. In *Proceedings of WWW 2002.* Honolulu, HI, 136–147.

Xiaohui, Z., Huayong, W., Guiran, C. and Hong, Z. 2001. An autonomous system-based distribution system for Web search. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics.* Tucson, AZ, 435–440.

Youngblood, G. M. 1999. Web hunting: Design of a simple intelligent Web search agent. *ACM Crossroads Magazine.* 5, 4, 1–4.

Yu, E. S., Koo, P. C. and Liddy, E. D. 2000. Evolving intelligent text-based agents. In *Proceedings of the 4th International Conference on Autonomous Agents (Agents00).* Barcelona, Spain, 388–395.

Received August 2004; revised December 2004, February 2005; accepted August 2005