A Quantitative Analysis of the User Behavior of a Large E-Broker*

Virgílio Almeida † Wagner Meira Jr. † Victor Ribeiro ‡ Nivio Ziviani †‡

† Dept. of Computer Science
Univ. Federal de Minas Gerais, Brazil
{virgilio, meira, nivio}@dcc.ufmg.br

‡ Miner Technology Group
Belo Horizonte, Brazil
{victor}@miner.com.br

Abstract

The Internet and the World Wide Web provide a global virtual marketplace. However, there is little information about the behavior of e-commerce users worldwide. The goal of the paper is twofold. First, we give an overview of the architecture and implementation of the Miner Family of Web Agents for e-commerce. Then, we present a quantitative study of the user behavior of a large e-broker (i.e., the BookMiner). Considering that the e-broker is used by a large number of users that only speak Portuguese and live in Brazil, we discuss the influence of regional and cultural issues on the e-commerce activities. Although the Web opens a company to a global market, our findings clearly indicate that e-commerce is strongly tied to regional issues, such as language, national customs and regulations, currency conversion and logistics. Also, the Internet infrastructure, mainly the intercontinental links, affects the user behavior.

1 Introduction

Web-based electronic markets are adequate for information-based products (e.g., news, software, financial services, ticketing services) and also for order retailing of some non-digital products such as books, CDs, flowers, cars, among others. More and more companies around the world are creating e-commerce sites that support lists of products and/or services, price information, and commercial transactions. As a consequence, the amount of available information and the number of potential customers in the Web is growing very rapidly [13].

*This work has been partially supported by Project SIAM/DCC/UFMG, grant MCT/FINEP/PRONEX number 76.97.1016.00, CNPq grant 520916/94-8 (Nivio Ziviani), CNPq grant 300437/87-0 (Virgílio A.F. Almeida) and CNPq grant 380134/97-7 (Wagner Meira Jr.)

Though useful information may exist somewhere, it is not always easy to find what a user is looking for on the Web. Since the Web is large and growing exponentially, it is impractical to exhaustively browse the Web looking for products and services. Therefore, one of the biggest challenges faced by electronic customers is the information overload, that hampers the growth of the online buying process. Although there are several different models for representing e-customer behavior, there exist some basic steps that are shared by most models [9], such as: need identification, product search, merchant search, negotiation, purchase and delivery, and product service and evaluation. In order to boost e-commerce activities, tools and services are needed to help customers in each of these basic steps. Two classes of services have been particularly useful for Web customers: (i) *search engines* and *directories*, and (ii) electronic brokers (*e-brokers*).

Large-scale *search engines* try to be comprehensive and any search usually returns many of related and unrelated information as a result of a user query. Most search engines are robot-based and index the whole Web as a full-text database. Robots [10, 12, 17] are software programs that traverse the Web collecting new or updated pages and sending them to a server where they are indexed. The index is used to answer queries submitted from anywhere in the Internet. According to a recent study [13], the number of Web pages is estimated from 200 to 320 million. The same study says that the largest search engines in Web coverage are Hotbot, AltaVista, Northern Light, Excite, Infoseek, and Lycos, and the percentage of the Web indexed by those search services varies from 3% (Lycos) to 34% (Hotbot). A different study [6] says that the coverage of the Web by the larger search engines ranges from 28% to 55%. The main problems of search engines are high volume of data, recollection of data because of the dynamic nature of the Web, saturated communication links, and overloaded Web servers.

On the other hand, high quality human maintained *di-*

rectories cover popular topics effectively, being able to focus a search in smaller collections of Web site descriptions. Directories are hierarchical taxonomies of classified human knowledge. The best example of human maintained index is Yahoo. The main advantage of directories is that the answer to a query is useful, in most cases. The main drawbacks stem from the fact that they are not specialized enough, and in many cases they cover a very small portion of the indexable pages on the Web. Directories also are slow to improve and expensive to build and maintain.

Therefore, the key issue continues to be where a piece of information is available in the Web. As a result, *e-brokers* have been developed to help users to find information, products and merchants. A broker is a party which mediates between buyers and sellers in a marketplace [15]. E-brokers can search for products, and retrieve information to help a customer to determine what to buy. E-brokers can also look for merchant-specific information (e.g., price) to help a customer decides whom to buy from. Basically, e-brokers can be viewed as search engines that specializes in specific topics. For example, a bargain broker searches the Web for price and characteristics of products, summarizes the results and presents them to the user. In another example, a broker could search in the e-catalog of many suppliers, which are registered with the broker, and try to match product specification and negotiation requirements.

The Miner Family of Web agents [5] is both a searching utility and a electronic catalog, that also provides brokerage services. The Miner Family were developed mainly for Portuguese language-based services. The search utility services provided by the Miner Family include: (1) MetaMiner, metasearch engine that uses Brazilian and international search engines, (2) DoctorMiner, that searches for information on several sites containing medical and odontological references, (3) SoftMiner, that searches for software in freeware and shareware sites, and just released (4) JavaMiner, that searches for technical information about Java language, and (5) PeopleMiner, that searches for people on the Internet. The search engine service includes (6) NewsMiner, that collects news from Brazilian newspapers, leaving them daily available for the Internet community. Brokerage services include: (7) BookMiner, that searches for books in registered Brazilian and international bookstores to match user's specification and (8) CDMiner, that searches for musical titles in Brazilian and international musicstores to find the user's preferences.

A *portal* is a site that brings together a variety of content and services in one area and attracts a large number of visitors. The idea is to become a single best starting place for as many users as possible. In Brazil, the largest Web site is UOL [7], which is shaping itself as a portal. UOL is a Brazilian site that brings together a variety of content and services in different areas. UOL acts both as a content and service provider offering more than 53 Brazilian magazines, 21 international magazines, 59 Brazilian newspapers and 31 international newspapers. UOL also offer several services, including hundreds of chat rooms that topped 12,000 people, more than 400 product sites, and RadarUOL, a search engine powered by Inktomi [3]. UOL topped 12 million page views in one day, being one of the largest non-English content provider in the world. The Miner Family has a partnership deal with UOL and is one of the services offered at the UOL site. This rich environment provided us part of the data used in this paper.

The goal of the paper is twofold. First, we give an overview of the Miner Family architecture and implementation and point out the differences with existing similar services (e.g. Junglee, Express, and Jango). Then, we present a quantitative study of the user behavior of a large e-broker. Considering that the e-broker is used by a large number of users that only speak Portuguese and live in Brazil, we discuss the influence of regional and cultural issues on e-commerce activities. In order to do that, the paper is organized as follows. Section 2 presents the architecture of a non-English e-broker, named "The Miner Family" and discusses its design rationale and components. Section 3 characterizes the workload of a brokerage service (i.e., BookMiner) of the Miner Family. We present figures that indicate the level of activity of the e-broker and show a model of the customer behavior. To conclude, Section 4 points out some evidences that shows the influence of regional and cultural issues, language in particular, on the quantitative results presented in the paper.

2 Architecture of the Miner Framework

The Miner family of Web agents [5] is a set of tools whose main objective is to help people to find information on the Web. The main idea is to bring multiple search and information sources together in one place. The searching is performed by agents working in parallel, just like metasearchers [16, 11] that use several search engines simultaneously, collecting answers and unifying them. The information may be a price of a book, a new musical release, a freeware or a shareware software, daily news, or any document available on the Web.

The Miner Family was written in Java language and comprises about 23,000 lines of code that run on a Netscape Enterprise Server, and the host platform is a SUN Ultra running Solaris 2.6. The code was implemented emphasizing greater reusability and easier maintenance and is structured into four levels: (1) general library, (2) middleware (e-commerce, search utilities, and search engines), (3) agents, and (4) user's interface. Figure 1 depicts each of these levels, which are explained in detail in the next paragraphs.

The general library contains several functionalities that

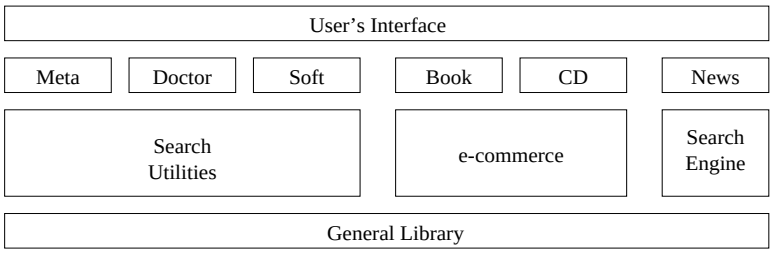

Figure 1. Structure of the Miner Family code

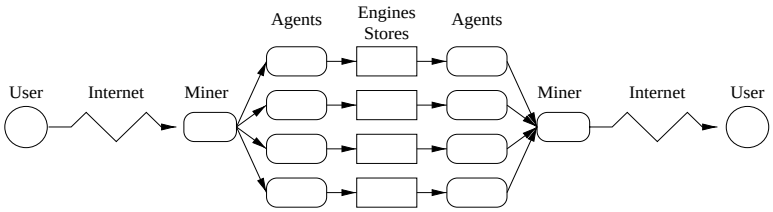

Figure 2. Miner Family functionality

are used by the upper levels, such as handlers (HTTP, cookies, tickets), query caching (for breaking results among pages), data fusion and interface widgets. It corresponds to 25% of the Miner code. The functions and primitives for each of the types of services offered by the Family are implemented in the middleware level, and each of the three services comprises about 2,000 lines of code. The e-commerce code contains classes that abstract goods' characteristics and interface with the stores that sell them. Similarly, the search utilities code contains functions that handle searches in each of the types of sites (software, people, and general) and the respective object classes. The search engines code implements procedures for information management, connection handling , and bots' navigation control. The ethic of bots [12] is also enforced by the procedures of the search engines. The agents for the various sites that are queried comprise 3,000 lines of code total. Among other tasks, these classes store details about site handling, data filtering, and structure of HTML data. Finally, the interface code (7,000 lines) implements all the HTML forms for the queries and the formatting of their results.

By using this structure, the implementation of new Family members becomes trivial. A new search utility, querying ten different sites would require only about 500 new lines of code. As a example, the implementation of the member JavaMiner, which searches for articles on the Java language, cost 16 man-hour and was made available in less than a week after conception.

2.1 Components and Services

All members of the Miner Family work similarly and the main steps to answer a query are depicted in Figure 2. Each query task can be divided into five main steps, as follows: (1) a user submits a query; (2) the Miner server gets the query and dispatch its agents; (3) each agent queries its target engine, store, or site; (4) each agent receives and parses the query results; and (5) the server unifies, formats, and sends the results to the user.

Currently the Miner family has eight members, divided into three groups: (1) search utilities (MetaMiner, DoctorMiner, SoftMiner, JavaMiner, PeopleMiner), (2) search engines (NewsMiner), and (3) e-commerce (CDMiner and BookMiner). Table 1 presents description of each member concerning its target (e.g., search engines, stores, etc.) and the number of registered sites for each member.

Member	Target	#Sites
MetaMiner	search engines	13
DoctorMiner	medical and odontological	17
NewsMiner	newspapers	13
BookMiner	bookstores	16
CDMiner	musicstores	13
SoftMiner	software	10
PeopleMiner	people	7
JavaMiner	Java language	5

Table 1. Members of the Miner Family

2.2 Related Work

There are related works in this area. They are new and change a lot in short periods of time. Excite has a shopping guide to find products and prices on the Web, which is called ProductFinder and is powered by Jango [4]. Junglee has developed a technology which aggregates information and prices for merchandise sold on the Web, enabling consumers to compare and shop for online products. Their technology was used by Yahoo at the time the experiments presented below were done. Junglee also announced its Shopping Guide that can be reached by Web consumers by hitting an e-commerce button found on Presario PCs from Compaq [1]. More recently, Infoseek announced Express [2], which uses many search engines to multiple search for products.

Table 2 presents the main characteristics of the three technologies mentioned above and the Miner Family. The first row shows the number of bookstores used by Yahoo.Junglee, Infoseek.Express and BookMiner. In the case of Yahoo.Junglee the number was estimated from the queries submitted as they do not list the actual bookstores. In the case of Infoseek.Express they do not search all five bookstores or musicstores in parallel, but each one at a time, and so we could not include them in our experiment whose results are shown in Table 3. From the 16 bookstores listed in BookMiner 8 are Brazilian. The second row presents the number of musicstores provided by Yahoo.Junglee and CDMiner, and the value for Yahoo.Junglee is again estimated from the queries because they do not list them. From the 13 musicstores listed in CDMiner 5 are Brazilian. The third row presents the number of engines to search for software (freeware and software). Again, in the case of Infoseek.Express they search all 5 software sites one at a time. The fourth row presents the number of search engines and directories used by Infoseek.Express and MetaMiner. From the 13 engines used by MetaMiner 5 are Brazilian. The fifth row shows that only Infoseek.Express searching tools do not perform requests in parallel. Finally, the last row shows the tools that allow users to choose the sites that are to be queried.

Table 3 presents seven queries submitted to Yahoo.Junglee, BookMiner and CDMiner. The first five queries search for books, the first two being titles published in US. The following three are authors of books: one American (i.e., the writer Tom Wolfe), one Portuguese (i.e., the poet Fernando Pessoa), and one Brazilian (i.e., the writer Jorge Amado). The following two queries search for CDs from one American (i.e., the jazz singer Ella Fitzgerald) and one Brazilian (i.e., the bossa nova singer João Gilberto) artist. The last query searches for the sound track of the 1985 movie Subway, which was found only in one Brazilian musicstore at this time. The aforementioned table shows the query results. The first two columns present the answers returned by Junglee and Miner, respectively. The last column (Common) presents the number of answers that appeared in the results returned by both tools. The large number of documents returned by the Miner Family comes from the larger number of registered sites. For queries involving Brazilian and Portuguese names the differences are even larger because of the language influence.

3 Workload Characterization and Analysis

This section presents a workload characterization of the Miner Family. We show a two-level characterization process. The higher level quantifies the overall load for services, in terms of the distributions and characteristics of service requests. The lower level focuses on the behavior of the BookMiner, that is one of the brokerage services of the family.

3.1 Overall Workload

We start out the analysis by partitioning the overall workload according to the services provided by the Miner Family. Table 4 shows the data extracted from logs of a four-week period of usage of the Miner services. The daily average number of requests was 22,086. We divided the data into three categories: (1) request frequency, (2) request characteristics, and (3) hourly distribution. Request frequency represents the percentage of requests addressed to each service. We note that MetaMiner is the most popular service, receiving almost 90% of the total requests. Three other metrics were defined to further characterize the request workload: (1) words per query, (2) match ratio, and (3) answers per query. Words per query quantifies the complexity of the request, which is around 2 words on the average. For instance, 95% of the requests to CDMiner have less than four words.

The match ratio represents the number of requests that returned at least one URL. In this case, we can observe that a high match ratio can result from two different scenarios. The first one is related to services that have broad coverage (i.e., the MetaMiner) and provide answers for most of the queries (although we cannot quantify how meaningful the answers are). The second scenario involve services that are so specialized that the queries are very constrained (i.e., SoftMiner and DoctorMiner). Similar conclusions arise when we look at the average number of answers per query.

Regarding hourly distribution, we consider three characteristics of the workload: peak period, peak hour, and peak/average ratio. Peak period represents the hours during which the number of requests is higher than the daily average. As we can see in Table 4, this information uncovers an interesting characteristic of Miner users, who usually

Characteristics	Technologies			
	Yahoo.Junglee	Excite.Jango	Infoseek.Express	Miner Family
Bookstores	6†	–	5	16 (8 Brazilian)
Musicstores	4†	–	5	13 (5 Brazilian)
Software	–	10	5	10
Metasearch engines	–	–	7	13 (5 Brazilian)
Parallel search	yes	yes	no	yes
Where to search option	no	no	yes	yes

† Estimated

Table 2. Characteristics of the search tools

Queries	Answers		
	Junglee	Miner	Common
Sphere (by title)	75	261	65
Jurassic Park (by title)	71	106	58
Tom Wolfe (by author)	77	46	40
Fernando Pessoa (by author)	30	160	27
Jorge Amado (by author)	39	225	35
Ella Fitzgerald (by artist)	42	161	20
Joao Gilberto (by artist)	28	76	11
Subway (by title)	0	1	0

Table 3. Different types of queries submitted to Yahoo.Junglee and the equivalent Miner tools (BookMiner and CDMiner)

query information during work time, probably using a non-modem connection. The peak hour is the time slot when the maximum number of requests was observed. In all cases but two, we noticed the peak hour is right after lunch time in Brazil. This behavior can be explained by the profile of the Brazilian Internet user, where the majority of the users (i.e., 70%) access the Internet from work. One of the exceptions occurs for the NewsMiner service, whose peak is around 7:00am, when users log to get the daily and breaking news. DoctorMiner peak hour is around 8:00pm, when home users look for medicine news. Finally, peak ratio measures the request rate at the peak hour over the average rate [14]. Specific services such as BookMiner and NewsMiner are more bursty than generic search service like MetaMiner. Their peaks are 7 and 13 times higher than their average, respectively, while the peak ratio of the MetaMiner is only 2.29.

3.2 Workload of a E-Broker

This section focuses on the workload of the BookMiner. As described in the previous section, BookMiner is a brokerage service. The goal is to study the workload generated by Brazilian customers searching for books on global electronic bookstores as well as on national bookshops. Our analysis is based on logs of a four-week period. At first, we can note from Table 4 that almost one fourth (24.35%) of the requests did not match any query in the registered bookstores and were discarded. The characterization is based on data collected from two logs. The first log traces overall results of the BookMiner activities, while the second one provides per-bookstore information. IP addresses were masked in order to protect users' privacy. We merged the two logs based on time, date, and masked IP address. As a result, the merged log provides the following information: date and time of the request, search keyword(s), keyword type (title or author), elapsed request response time, overall number of titles returned to the user, response time for each bookstore, and number of titles returned by each bookstore.

Figure 3 shows the BookMiner customer behavior graph. For each registered bookstore, we measured the click-through frequency, given a BookMiner response. The click-through determines which bookstore was chosen by the user. The percentage associated with each path of the BookMiner graph represents the click-through frequency. We note that 76% of the Brazilian customers prefer Brazilian bookstores. Among the global bookstores, Amazon.com was chosen by most of the users (50% of the users), followed by Barnes & Noble and BookStacks. Siciliano, a Brazilian bookshop, is responsible for one fourth of the click-throughs among the Brazilian bookstores, followed by Cultura, Booknet and Loyola. An interesting observation is

Member	MetaMiner	BookMiner	CDMiner	SoftMiner	NewsMiner	DoctorMiner
Queries(%)	89.15	2.60	2.65	2.34	1.89	1.37
Words/Query	1.98	2.05	1.87	1.55	1.66	1.69
Match Ratio(%)	93.64	75.65	77.63	88.00	55.60	95.81
Answers/Query	53.97	42.40	41.06	63.74	11.05	47.78
Peak Period	7am–9pm	7am–9pm	11am–7pm	8am–11pm	5am–5pm	8am–10pm
Peak Hour	1pm	1pm	2pm	1pm	7am	8pm
Peak Ratio	2.29	7.52	6.41	7.50	13.12	9.37

Table 4. Overall Workload Statistics

Bookstore	Availability (% of requests)	Response Time(sec.)	Book Hit Ratio
Barnes & Noble	95.55	25.4	18.5%
Bookstacks	84.75	8.1	22.0%
BookPool	99.50	10.4	4.7%
McGraw Hill	99.20	28.0	4.3%
O'Reilly	100.00	12.7	4.6%
Prentice Hall	100.00	7.1	7.2%
iBS	100.00	17.1	13.9%
Amazon	99.23	13.0	19.1%
Booknet	98.27	12.5	49.3%
Campus	100.00	2.0	7.2%
Cultura	100.00	14.3	33.6%
Siciliano	76.12	24.8	69.4%
Sodiler	100.00	11.4	38.4%
Tempo Real	100.00	12.5	11.5%
Loyola	100.00	8.9	56.0%
artepaubrasil	100.00	8.9	55.7%
BookMiner	100.0	48.5	

Table 5. BookMiner Performance Results

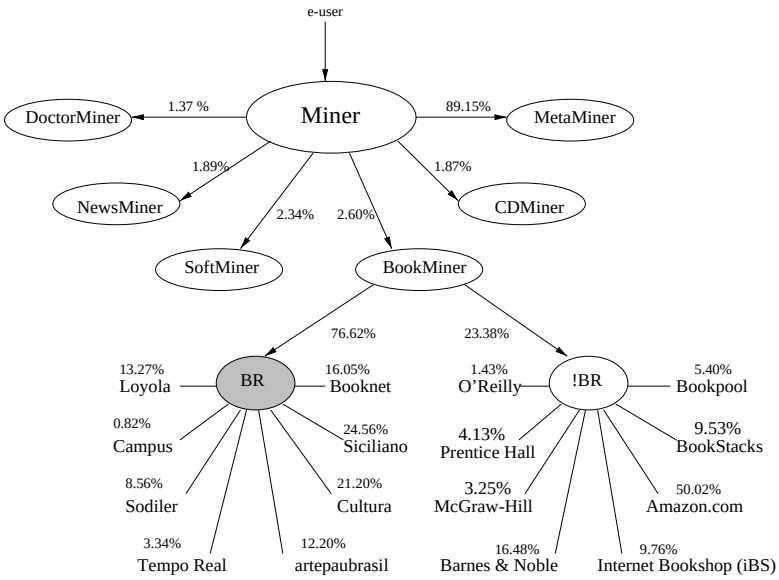

Figure 3. Customer Behavior Graph

that Siciliano does not exhibit a good service level indicator. It has the lowest availability (23.88% of the queries timed out as shown in Table 5) among the Brazilian bookstores. However, Siciliano is a well established company, having many bookstores in the main cities of Brazil, which somehow makes the company familiar to customers, even in the Internet.

E-commerce service levels are usually assessed through response time and availability. For the purpose of our analysis, a server is considered available when it answers the book request within the user-defined timeout (i.e., 60 seconds by default). Elapsed request response time is the interval of time needed for receiving a response from the server. Table 5 shows availability and elapsed response time of the registered bookstores that are queried by the BookMiner. We note that almost all bookstores exhibit a good level of availability no matter they are Brazilian or global. On the other hand, the same table shows a high variance for response times. Average elapsed response time of national bookstores is lower than international bookstores. We conjecture that this phenomenon is a consequence of the heavy traffic on the international links between Brazil and US.

We define another metric called "book hit ratio" (BHR) that represents the number of times that a bookstore suggests at least one title in response to a customer request over the total number of requests sent to the bookstore. Looking at Table 5, it is evident that Brazilian bookstores are more effective in finding in their selection the books requested by Brazilian customers. The BHR of Brazilian bookstores is higher than the BHR of the global bookstores. This fact stems from cultural factors such as English proficiency and local interests. Around 50% of Brazilian Internet users do not speak English. Also, Brazilian bookstores have much larger selection of books on topics that are part of the Brazilian culture [8] than global bookstores.

4 Concluding Remarks

This paper shows an overview of the Miner Family architecture and implementation. Then, we present a quantitative study of the user behavior of a large e-broker. We characterize the workload of the Miner Family and focus on the behavior of the BookMiner, one of its brokerage services.

Based on the statistics shown in the paper we found interesting observations about e-commerce. First, we note that 76% of the Brazilian customers prefer Brazilian bookstores. The reasons are multiple. First, network infrastructure affects customer behavior. Average elapsed response time of national bookstores is lower than international bookstores. We conjecture that this phenomenon is a consequence of the heavy traffic on the international links between Brazil and US. Language, culture and social aspects play a major role in the behavior of e-commerce customers. Brazilian bookstores are more effective to attract customers, not because they offer more titles, but because they offer Brazilian books written in Portuguese, which are far more popular. Customs, currency conversion and delivery logistics also help local bookstores.

Although the Web opens a company to a global market, our findings clearly indicate that e-commerce is strongly tied to regional issues, such as language, culture, national customs, currency conversion and logistics. The Internet infrastructure, mainly the intercontinental links, hampers a

consistent performance and affects user behavior. Our future work will focus on the quantitative analysis of the behavior of e-commerce users to come up with models to describe workloads of e-commerce components, such as portals, brokers and merchants. Moreover, we would like to answer questions such as what regional features should be present in a portal site considering cultural and language characteristics.

References

[1] Compaq Corp.: Junglee Shopping Guide. `http://www.compaq.junglee/compaq/ top.html`.

[2] Express:. `http://www.express.infoseek.com`.

[3] Inktomi:. `http://www.inktomi.berkeley.edu`.

[4] Jango:. `http://www.jango.com`.

[5] Miner family:. `http://miner.com.br`.

[6] Search engine watch:. `http://www.searchenginewatch.com`.

[7] Universo online:. `http://www.uol.com.br`.

[8] V. Almeida, M. Cesário, R.Fonseca, W. Meira Jr., and C. Murta. The influence of geographical and cultural issues on the cache proxy server workload. *Computer Networks and ISDN Systems*, 30:601–603, 1998.

[9] Y. Bakos. Reducing buyer searching costs: Implications for electronic marketplaces. *Management Science*, 43(12), 1997.

[10] F. Cheong. *Internet Agents Spiders, Wanderers, Brokers and Bots*. New Riders, 1996.

[11] D. Dreilinger. Savvysearch:. `http://guaraldi.cs.colostate.edu:2000/`.

[12] M. Koster. Guidelines for robot writers. `http://web.nexor.co.uk/mak/doc/robots/guidelines.html`.

[13] S. Lawrence and C. Giles. Searching the world-wide web. *Science*, 280(5360):98, April 3, 1998.

[14] D. A. Menascé and V. A. F. Almeida. *Capacity Planning for Web Performance – Metrics, Models, & Methods*. Prentice Hall, PTR, 1998.

[15] A. Segev, D. Wan, and C. Beam. Designing eletronic catalogs for business value: Results from the commercenet pilot. Technical Report Working Paper CITM-WP-1005, Fischer Center for Information Technology – U.C. Berkeley, October 1995.

[16] E. Selberg and O. Etzioni. Multi-service search and comparison using the metacrawler. In *Proc. of the Fourth International World-Wide Web Conference*, 1995.

[17] J. Williams. *Bots and Other Internet Beasts*. Sams Net, 1996.