JIS

Popularity and findability through log analysis of search terms and queries: the case of a multilingual public service website

Gilad Ravid

Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheva, Israel and Annenberg Center for Communication, University of Southern California, Los Angeles, CA, USA

Judit Bar-Ilan and Shifra Baruchson-Arbib

Department of Information Science, Bar-Ilan University, Ramat-Gan, Israel

Sheizaf Rafaeli

Center for the Study of the Information Society, University of Haifa, Haifa, Israel

Received 3 September 2006
Revised 23 December 2006

Abstract.

SHIL on the Web is the website of the Israeli Citizens' Advice Bureau. It provides information about rights, social benefits, government and public services and civil obligations. Activity on the site approaches 10,000 pages visited per day. It has interfaces in four languages: Hebrew, Arabic, Russian and English. Logfile analysis of the SHIL website revealed to our surprise that about 60.7% of the requests reaching SHIL from external sites (excluding requests from robots) are from general search engines (e.g. Google and MSN), and users reach a specific page on the site linked from the search results page. This finding seems to indicate that the site is not known well enough to the public. On the other hand the site is very active, thus it seems to serve Israeli citizens well, even without being a well known brand. In this paper we analyzed the external requests coming from search engines. The analysis is based on the 266,295 queries from search engines that reached SHIL during March–October 2005. Studying queries submitted to search engines is a novel technique for analyzing the access patterns to the site and provides a better understanding of the user needs and intentions than analyzing the distribution of the visited pages only. We are not aware of any previous study that analyzed the relation between

Correspondence to: Judit Bar-Ilan, Department of Information Science, Bar-Ilan University, Ramat Gan, 52900, Israel. Email: barilaj@mail.biu.ac.il

the query submitted to the search engine and the webpage the user clicked on the search results page. Since search engines provide snippets, when the user clicks on a specific page he already has some information on what is to be found on the page and the user makes a conscious decision to click on the specific result. Thus, this type of analysis provides additional information about the users' actual information needs.

Keywords: logfile analysis; information on public and governmental services and entitlements; queries; search engines

1. Introduction

What can an individual do when he needs information on public and governmental services and entitlements, especially when he is not sure what government office will provide an answer to his information need? He may ask friends and relations – strong social ties – or he may approach a Citizens' Advice Bureau (CAB, e.g. [1]) or use a phone hotline, like the 2–1–1 information and referral helpline in the United States and Canada [2]; he can also go to the library or to a nearby information centre. However, today more and more people turn to the web in order to fulfil their information needs. The world wide web has become a major information source in the developed world (e.g. [3]). Suppose that our citizen decided to seek a solution to his information problem through the web. Again he has a number of options: perhaps he is aware of some site that might provide the necessary information – in this case he can type in the URL to the location bar of his browser or retrieve the URL from his bookmark list; for Israeli citizen information, SHIL on the Web (www.shil.info), the website of the Israeli Citizens' Advice Bureau, would be a good choice. Another option is to recall an ad on the TV or on the radio, inviting him to visit the governmental portal – in Israel, the Government portal (www.gov.il) is constantly advertised, or he may choose to browse a directory like Yahoo! (dir.yahoo.com) or The Open Directory (www.dmoz.org), or a local directory, like Walla (www.walla.co.il) in Israel. Of course, he may also type a query relating to his information need into the search box of a search engine and hope that by clicking on one of the results displayed for his query, he will be able to solve his information problem.

It turns out that many web users choose the last option and even if they are aware of some of the other possibilities they prefer to turn to a search engine, most likely to Google [4, 5]. According to the findings published in [4] an increasing number of the search queries are 'navigational queries' [6]. Thus, it seems that users do not bother to remember or to store the URLs of websites useful to them; rather they prefer to look up the addresses of these sites at a search engine.

'Findability' is defined by Morville [7, p. 4] as

(a) the quality of being locatable or navigable;

(b) the degree to which a particular object is easy to discover or locate; and

(c) the degree to which a system or environment supports navigation and retrieval.

The main goal of SHIL on the Web is to enhance findability of information on public and governmental services and entitlements.

In this paper we studied a large log of the SHIL website, focusing on requests from external referrers – called 'external hits' [8]. External hits from search tools, where the referral URL contains a query (i.e. the user reached the site after submitting a query at the referral site) are called 'external queries'. The 'referer' (misspelled in the official HTTP specification [9]) or the 'referring page' is the URL of the previous page from which a link was followed [10]. Note, that here we analyze external requests that originated from search engines, and do not study the distribution of the visitors to the website.

The analyzed log contained 757,697 external hits and covered an eight-month period between March and October 2005. About 330,000 of these external requests originated from crawlers. Out of the remaining 438,289 external hits, 65.8% were external queries. The remaining 34.2% external requests did not contain any information on the source and a small minority came from other sites that link to SHIL. One plausible explanation for the large percentage of external queries could be that users

use search engines to locate the SHIL website. However, the analysis showed that this was not the case: only 0.6% of the queries were looking for the SHIL website. A better explanation is that many users are unaware of SHIL's existence, but still use it extensively to fulfil their information needs related to public and governmental services and entitlements, as we will show in the following sections.

Previous logfile analyses either studied search engine logs or logs of specific sites. Search engine logs were analyzed in order to characterize the submitted queries (popular search terms, query length, number of search pages viewed, number of modifications, length of search session, etc.). Logfile analyses of specific sites, on the other hand usually analyze page visit distributions, user profiling and/or internal navigation patterns. In the current study we analyzed the queries submitted to general search engines that directed users to the SHIL website. This kind of analysis allowed us to learn about the users' information problems which were submitted as queries to the search engine against the site's 'response' – the page that they reached from the search engine. We are not aware of any previous studies that employed such methodology. The user sees snippets for all the results presented by the search engine for his query and makes a conscious decision to click on the specific result – a result that seems relevant to the information problem at hand.

2. The SHIL website

SHIL on the Web operates on a proprietary content management system; it has a directory-like structure (see Figure 1) with top-level categories, listed here in the order of appearance:

- Economics
- Transportation
- Work relations
- Welfare
- National insurance
- Absorption and immigration
- Health
- Environment
- Consumers
- Taxes and fees
- Army and security
- Housing and accommodation
- Education
- Family matters
- Registrars
- Law and justice
- Other
- SHIL offices

Most of the information is in Hebrew; some of it is translated into Arabic and Russian as well. In each category there are a number of articles, explaining topics related to the category. Information sources for the articles include governmental publications and communiqués, as well as popular press articles, suggestions and contributions from the public and constituent organizations. A specific article may belong

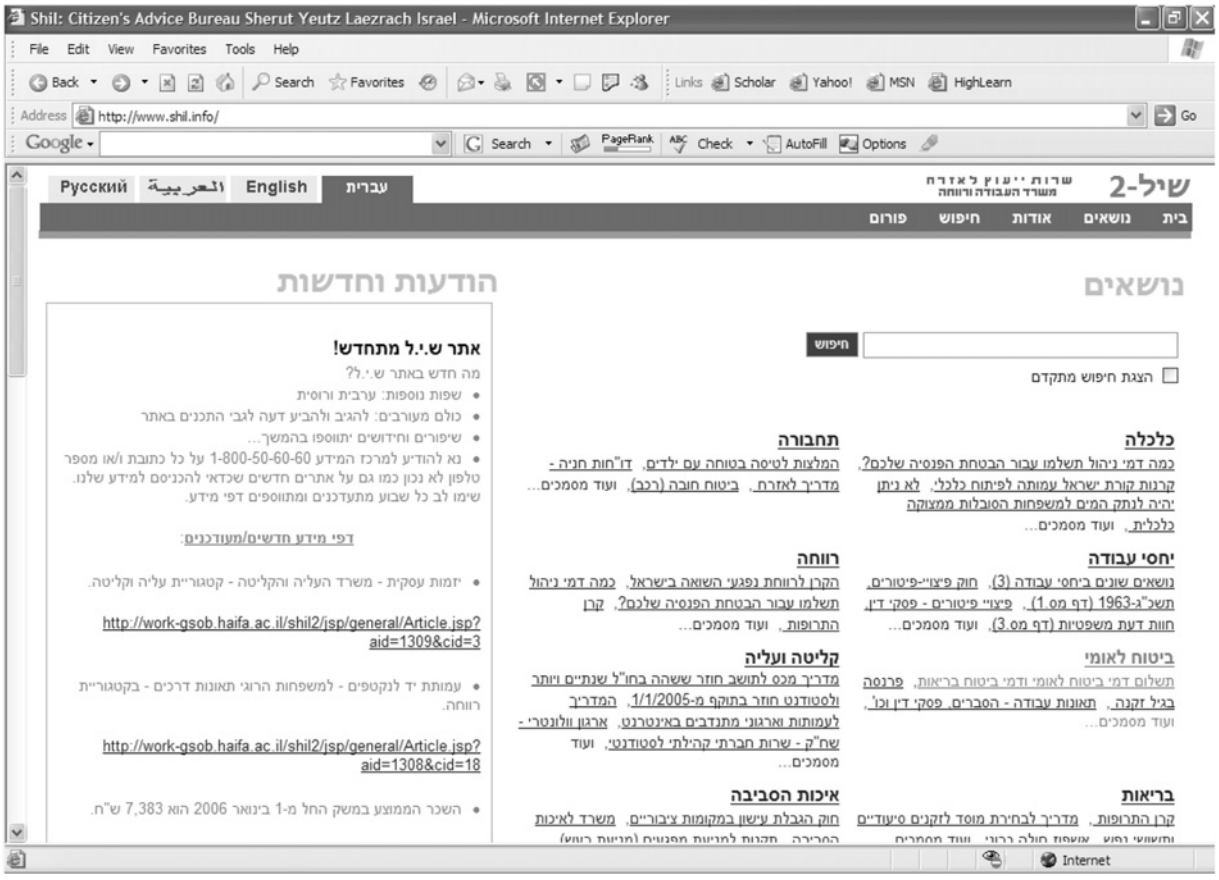

Fig. 1. Hebrew language homepage of SHIL on the Web.

to multiple categories. The site is updated almost daily by SHIL volunteers and staff – the date of last update appears on each article. In many cases there are almost no links in the articles either to other parts of the site or to outside sources. However, there are navigational links to the homepage and to the category or categories the article belongs to. There is a search box on each page that allows the users to search within the site and this feature allows the users to search for further information on their topic. The users are encouraged to provide feedback; they can score the specific article on a scale of 1–5 or comment on it. On the homepage there is a list of new and updated articles. The site operates forums in Hebrew and Russian, where the users ask specific questions and SHIL staff and volunteers answer these questions, often by directing the user to a specific article in the site. The Hebrew language forum receives about 30 questions a day; its Russian language counterpart is less active. An Arabic and a Russian mirror of the entire site, including interactive components, is under development; currently only partial content is available in these languages. Arabic is an official language in Israel, and there is need for information in Russian as well due to the large number of immigrants from the former Soviet Union, who arrived in Israel in the 1990s. There is an English language interface as well.

3. Related studies

There are a number of studies that have analyzed search engine logs. One of the first major studies was based on a set of almost 1,000,000,000 queries presented to AltaVista during a 43-day period in August–September 1998 [11]. Findings of the study included data on the average number of terms

(a single word or a phrase enclosed in quotation marks) per query (2.35). Jansen, Spink and Pedersen [5] compared the results of [11] with a one day log (of 3,000,000 queries) on AltaVista in 2002. The percentage of single-word queries decreased from 25.8 to 20.4%, and the most frequently appearing query changed from 'sex' in 1998 to 'google' in 2002. 'Sex' was at fourth place in 2002. The popularity of the query 'google' shows that web surfers use the search engines as a navigation tool. A subset of the queries of 2002 was categorized, and 'people, places and things' was the most frequently occurring category (over 49%).

Excite's logs were extensively analyzed in a series of papers [12–16] and in a number of additional conference presentations. The results were based on four sets: a set of about 50,000 queries from March 1997, a set of over a million queries from September 1997 and two other sets of similar sizes from December 1999 and May 2001 respectively. The results of the analyses included the number of terms per query (about 2.4 on the average, with an increase to 2.6 in the last set), the percentage of users who used Boolean queries (between 5 and 10%) and data related to search sessions. Spink, Jansen, Wolfram and Saracevic [17] compared the last two datasets, enabling them to identify a shift in the interests of the searchers from entertainment, recreation and sex to e-commerce related topics, like commerce, travel, employment and the economy.

After Excite stopped being an independent search engine, the above-mentioned researchers switched to analyzing logs from the search engine AlltheWeb [5, 15, 16]. The results are comparable, except for some differences in topics searched: less emphasis on e-commerce related issues and more on people and computers, but 'sex' was still the second most popular search term. AlltheWeb users generated slightly more queries per session than Excite users. In the second AlltheWeb study (with data from 2002) there was an increase of single-word queries from 25 to 33%.

Ozmutlu, Spink and Ozmutlu [18] carried out a time-of-day analysis of the search logs of Excite and AlltheWeb based on the local time at the server. For Excite the busiest hour of the day in terms of query arrival was between 9 and 10 a.m., while for AlltheWeb the largest number of queries per hour arrived between 8 and 9 a.m. Hourly traffic of the AOL search engine powered by Google was analyzed by Beitzel et al. [19], and according to their findings the busiest hour of the day was between 9 and 10 p.m. Since AOL users are located in the United States, this means that the morning hours were busiest.

Additional large scale web search engine log analyses were carried out for a Korean search engine [20], where specific techniques were developed to handle language specific problems and also for Vivisimo [21]. Spink and Jansen discuss their search log analysis studies in their book [22] and compare the results of nine large search log analysis studies in a recent paper [23].

Next we review studies analyzing search logs of individual sites. One of the first single website search studies was carried out by Croft, Cook and Wilder [24]. They examined the usage of the THOMAS website, intended to provide government information to the general public on the web. Jones, Cunningham and McNab [25] analyzed more than 32,000 queries that were submitted to the New Zealand Digital Library over a period of more than one year in 1996–7. Cacheda and Vina [26] analyzed the search logs of the Spanish web directory BIWE based on a 16-day log from 2000. Wang, Berry and Yang [27] carried out a four-year longitudinal analysis of queries submitted to the website of the University of Tennessee at Knoxville. Chau, Fang and Sheng [28] analyzed the queries submitted to the Utah state government website during a period of 168 days in 2003. The content provided on the Utah state government website resembles the information available from SHIL on the Web.

Finally we mention two studies that analyzed the referrer field of website logs (the page visited just before hitting a page on the site). Thelwall [29] analyzed the log of the site of the Wolverhampton University Computer Based Assessment Project for a period of 10 months in 2000 and found that nearly 80% of the external hits were requests from search engines, most of them from Yahoo. He also analyzed query phrasings, but because the targeted site was very small (only five pages) there were only minor variations to the queries.

Davis [30] studied the distributions of a set of referrals to the American Chemical Society's site. The aim of the study was to understand how scientists locate published articles. In this case only 10% of the referrals were from search engines.

4. Theoretical framework

4.1. Information behavior in electronic environments

One of the first models for online searching is Bates's berrypicking model [31]. She described online searching as a process where the user during his search collects pieces of information. These pieces may influence and change the original information problem and at the same time help the user to modify the query, so that the results will better suit the information need. An extensive, user-centered model of information seeking in the electronic environment was introduced by Marchionini [32]. He defined several stages in the information seeking process. These stages may follow one another sequentially, but often the information seeker goes back to a previous stage and changes some of the settings defined at that stage so that he can proceed more successfully. The stages are: recognize and accept an information problem; define and understand the problem; choose a search system; formulate a query; execute the search; examine results; extract information and reflect/iterate/stop.

One of the earliest models of web searching was proposed by Choo, Detlor and Turnball [33]. Their model was based on existing models for information behavior (by Ellis [34]) and scanning (by Aguilar [35]) that were not developed for the web environment. The first model that was developed specifically for searching the web was introduced by Hölscher and Strube [36]. Their model quantified the transitions between different states of the model: information need, direct access, search engine interaction, examining a document and browsing a website. Broder [6] defined a taxonomy of query types. He differentiated between informational, navigational and transactional queries. In the SHIL query log we identified informational and navigational queries (looking for a specific site or page).

4.2. Intermediation in electronic environments

One of the great promises of e-commerce was disintermediation ('displacement of market middlemen who traditionally are intermediaries between producers and consumers by a direct new relationship between manufacturers and content originators with their customers' [37, p. 29]. However, in parallel to disintermediation we also witness re-intermediation. Bailey and Bakos [38] provided empirical findings on the existence of intermediaries in electronic markets. On the other hand, Burt [39] claims that what he calls 'second-hand brokerage' is negligible compared to direct contacts. Our results indicate that this is probably not the case in electronic environments. Sarkar, Butler and Steinfeld [40] claim that intermediation will not disappear and a new form of intermediation, called 'cyberintermediation' will be created. 'Cybermediaries' are a new type of intermediaries who 'perform the mediating tasks in the world of electronic commerce'. Intermediaries in electronic environments have the ability to aggregate search efforts and increase the efficiency of information seeking [41].

One can view the search engines as primary intermediaries – without them we have no efficient access to the huge amounts of information residing on the web. According to [41, 42] one of the roles of intermediaries is to facilitate searching. In our case, without intermediation, users would directly approach government offices to fulfill their information needs. Our results show that the users who reach the SHIL site from search engines (the majority of SHIL users) go through two intermediaries: the search engine and the SHIL website. In many of the cases, in addition to the basic information provided by SHIL, it directs the users to the site of the appropriate ministry or public service.

5. Research questions

Unlike most previous studies of websites, we decided not to analyze simply the logfiles, but to concentrate only on requests that originated from search engines ('external queries'). The external queries convey more information about the users' intentions, information needs and the way they formulate them, than a simple analysis of the distribution of the pages visited on the website. In

combination with the specific pages visited we can gain insights into the users' information problems. In addition we characterized the submitted queries so that the results of the current study can be compared to previous search engine and site log analyses. Differences are expected since SHIL is a multilingual site. There are no previous studies that analyzed query characteristics of Hebrew language queries, thus this study provides a baseline for Hebrew language searches on the web.

Our aim was to understand:

- What information on public and governmental services and entitlements is of interest to users who arrive at SHIL on the Web from search engines?
- How do the queries relate to the titles of the webpages the users reach from the search engine results page?
- What are the characteristics of these queries (query length, query terms, time submitted, etc.)? Are there specific problems because the site is multilingual?
- How do these queries compare with those submitted to general search engines and to queries submitted to local search engines, especially with the queries submitted to the Utah state government website [29] that provides information that is comparable to the information provided by the SHIL site?

6. Research design

6.1. Data collection

The dataset contained all the external hits to the SHIL website for an eight-month period between March 1, 2005 and October 30, 2005. The original log comprised 757,697 external hits. A large number of the requests (319,408) were identified as requests submitted by crawlers. Since our aim is to characterize human information needs, these records were excluded from the dataset. Out of the remaining 438,289 records 306,383 were records with non-empty referrers. Records with empty referrers are requests where either the visitor is a spider or a bot (most of these were excluded in a previous step of the data cleansing), or the user enters the URL manually to the location bar or clicks his Favorites list, disables the referrer or reaches the site through a non-browser link [43]. We have no further information regarding the requests with empty referrer field, and these records were not processed.

Each record contained the IP address of the requester, the exact time and date of the request, the referring site, the referring query for 'external queries' and the requested page on the SHIL site. IP addresses were discarded from the analysis, to avoid privacy issues (like the recent release of the AOL search data [44]).

6.2. Data analysis

All requests with status codes other than 200 (successful) and 304 (cached) were removed from the log, resulting in 297,153 records. This set included all successful external requests with explicit referrers. Out of these only 17,922 did not originate in a query, i.e. the referral site was a portal or some other site with a link to SHIL. Note that only 5.8% of the 'external hits' with explicit referrers did not originate from a search engine. The major source of the 'external queries' was google.co.il (71.6%), followed by google.com (10.6%), search.msn.co.il (6.3%) and walla.co.il (a local portal and search engine, 3.6%). Thus, over 80% of the 'external queries' originated from Google sites.

The huge majority of the queries were in Hebrew, with some Arabic, Russian and Latin characters. The search engines employ various techniques for encoding the non-Latin queries in the referral URL, and it is not always easy or possible to filter out the actual query from this text. For example for queries originating from MSN, the query was passed to the SHIL server and logs as a series of question marks. Some of these problems are caused by the use of several standards: some sites encode Hebrew and Arabic as single byte strings while other sites as multi byte (utf-8) strings. Multi byte queries are unique, but for uni-byte encoded queries where the encoding scheme did not

appear in the referrer URL, ISO8859–8 (Hebrew/English) encoding was assumed. This assumption was not justified in all cases, and sometimes resulted in illegible queries in an unknown language. Additional variations in encoding are caused by the fact that Hebrew and Arabic are written from right to left. Some of the special characters were also problematic. In some cases in Hebrew the quotation mark (“) is used in abbreviations – again posing a problem, both for the search engines and for the interpretation of the queries, especially when in a query there are two abbreviated terms.

After filtering out the majority of illegible queries, e.g. queries with question marks only, null queries, or non-character strings, the dataset including 266,295 ‘external queries’ (60.7% of the total log with requests from crawlers excluded) was loaded into MS Access, a relational database, for further analysis.

For each external query the log file records the page that was visited on the site. All the pages on the SHIL website are organized into categories and articles within the categories, where the top level page in each category has no article ID, and contains a linked list of all the articles in the specific category. Thus, because of the structure of the SHIL site, all pages on the site are categorized. Sometimes an article belongs to several categories.

7. Results

We report basic characteristics of the external query logs: length of queries, most frequent queries, and the pages visited from the results of these queries.

7.1. Query length and the use of search modifiers

Query length counts the number words in the query, where a word is a string of characters delimited either by a space or by the end of the query. Quotation marks were removed from the queries prior to counting the number of words in the query. The results appear in Table 1.

The mean query length is 2.79, which is comparable with other search engine log analyses [23, 28, 29]. However, the distribution of the query lengths is rather different: in the SHIL log the percentage of single-word queries is surprisingly low as can be seen in Table 1.

As expected, the variability in the four-word queries was higher than for the single-word queries. Altogether 19,205 different four-word queries were identified, out of which 14,122 (73.5%) occurred only once. The number of different single-word queries was 2212, out of which 1432 (64.7%) occurred only once. The most popular single-word query, Superland (amusement park) appeared 3160 times, which is 20% of the single-word queries; while the most popular four-word query, small claims court (in Hebrew) occurred 1392 times, which constitutes only 3.5% of the four-word queries.

Table 1
Query length distribution in absolute numbers and percentages out of the 266,295 queries

Query length in words	No. of occurences	Percentage of queries
1 word	15,817	5.94%
2 words	113,407	42.59%
3 words	77,736	29.19%
4 words	40,234	15.11%
5 words	12,693	4.77%
6 words	4275	1.61%
7 words	1258	0.47%
8 words	491	0.18%
9 words	196	0.07%
10 words	79	0.03%
11 words	32	0.01%
12 words	39	0.01%
More than 12 words	38	0.01%
Total	266,295	100.00%

7.2. Most frequent queries

The query log included 266,295 queries. Of these, 72,799 unique queries were identified. In Table 2 the 25 most frequent queries and their meanings in English are displayed.

All the queries in Table 2 are Hebrew queries; the most frequently occurring query in Arabic was تعليم اللغة العبرية (learning the Hebrew language) which occurred 286 times in the log. The most frequently occurring Russian query, социальное обеспечение (social security), occurred only 22 times. SHIL on the Web is currently being translated into Russian, and at the time the log was collected only a minority of the information in Hebrew was available in Russian as well. In the query log there were 2729 Arabic queries (1.02%), 614 queries in Russian (0.23%) and 1162 queries including Latin characters (0.44%). The 25 most frequently occurring queries comprise 20.95% of the log. The remaining 72,774 queries cover 79.05% of the log.

The queries displayed in Table 2 provide an insight to the users' information needs and intentions. The top queries are for government offices. SHIL is an intermediary for these sites, because quite often the pages the users reach on the SHIL website direct them to the appropriate section of the specific government website.

Figure 2 depicts the rank-frequency distribution of the queries on a log–log scale. There were a few frequently occurring queries, but the majority of the queries (52,065, 71.5% of the unique queries) occurred only once.

We also looked at the frequencies of the query terms. Altogether 742,454 query terms were extracted from the 266,295 queries. The number of unique query terms was 18,845. The most frequently occurring query terms are parts of the most frequently occurring queries. The 25 most popular words cover an even larger percentage of the total words (28.61%) than the 25 most popular queries out of the total number of words (20.95%).

Table 2
Most frequently occurring queries and their meanings, in absolute numbers and in percentages (N = 266,295)

Query in original language	Query in English	Frequency	%	Cumulative %
ביטוח לאומי	National Insurance	7226	2.71%	2.71%
משרד הפנים	Ministry of Internal Affairs	6394	2.40%	5.11%
משרד העבודה	Ministry of Labour	5234	1.97%	7.08%
שעון קיץ	Daylight saving time	5044	1.89%	8.97%
חוזה שכירות	Rental contract	3891	1.46%	10.44%
משרד הרישוי	Licensing Authority	3637	1.37%	11.80%
סופרלנד	Superland (*amusement park*)	3160	1.19%	12.99%
בטוח לאומי	National Insurance (*variant spelling*)	2119	0.80%	13.78%
דמי הבראה	Vacation fees	1933	0.73%	14.51%
פיצויי פיטורין	Dismissal compensation	1728	0.65%	15.16%
משרד הרישוי חולון	Licensing authority Holon (*the central office is in Holon*)	1685	0.63%	15.79%
בית משפט לתביעות קטנות	Small Claims Court	1392	0.52%	16.31%
בית החייל	Soldiers' residence	1383	0.52%	16.83%
היחידה להכוונת חיילים משוחרר	The unit for advising discharged soldiers	1327	0.50%	17.33%
רשם העמותות	Non-profit organizations' registrar	1072	0.40%	17.73%
ש.י.ל.	SHIL	994	0.37%	18.11%
חוק הגנת הצרכן	Consumer protection law	968	0.36%	18.47%
חוק הגנת הדייר	Tenant protection law	941	0.35%	18.82%
מינהל מקרקעי ישראל	Israel Land Administration	867	0.33%	19.15%
חיילים משוחררים	Discharged soldiers	854	0.32%	19.47%
האגודה לתרבות הדיור	Housing Advice Association	848	0.32%	19.79%
קו לחיים	Kav LaHaim (*an organization providing help for sick children*)	830	0.31%	20.10%
משרד רישוי	Licensing Authority (*variant spelling*)	763	0.29%	20.39%
לשכת הבריאות	Ministry of Health	753	0.28%	20.67%
משרד העבודה והרווחה	Ministry of Labour and Social Affairs (*full name of the Ministry*)	750	0.28%	20.95%

Note: text in italics added to clarify the query.

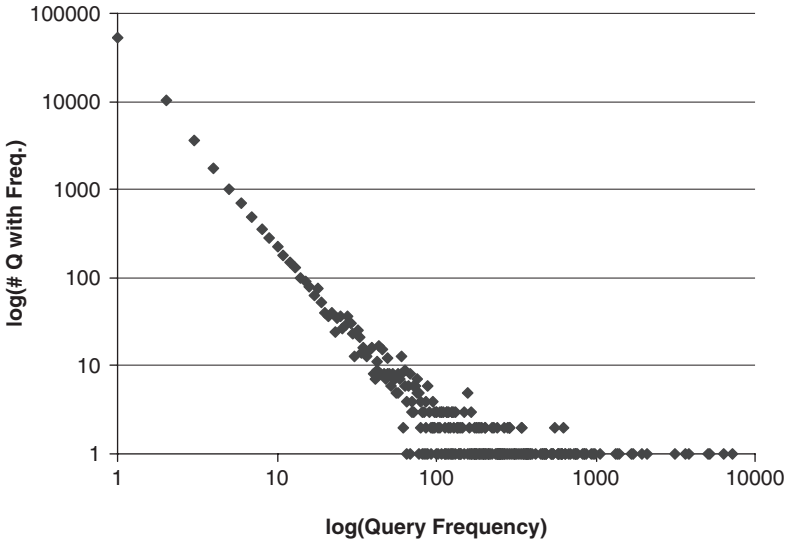

Fig. 2. Rank-frequency distribution of queries.

7.3. Queries in Arabic

The 2729 queries in Arabic were analyzed separately. Table 3 lists the 12 most popular queries in Arabic. The average length of the Arab language queries was slightly below the average in the log, 2.52 vs 2.79, and the longest query was only seven words long. One has to take into account that not all the information in Hebrew is available in Arabic as well. It seems that the Arab language queries concentrate on learning Hebrew and on topics related to driving. Rather interestingly, the query lesbianism (مساحقات) appeared 10 times (0.37%), while variations of the word lesbian in Hebrew appeared in 56 queries (0.02%). One possible reason for this could be that SHIL is one of the top ranking sources on the topic in Arabic, but not in Hebrew. The specific article in SHIL speaks about single-parent families and the adoption rights of homosexual or lesbian couples in Israel.

7.4. Most frequently requested categories and articles

The five most popular categories appear in Table 4. It is interesting to note that popularity does not correspond exactly to the order in which these categories appear on the home page of SHIL. In Table 5 the five most frequently listed articles can be viewed. We analyzed the most frequently visited pages in

Table 3
Most frequently occurring queries in Arabic and their meanings, in absolute numbers and in percentages ($N = 2729$)

Original query	Translated query	No. of occurrences	Percentage of queries in Arabic	Cumulative %
تعليم اللغة العبرية	Learning the Hebrew language	286	10.24%	10.24%
تعلم السياقة	Driving lessons	98	3.51%	13.75%
التدبير المنزلي	Home economics	70	2.51%	16.26%
اللغة العبرية	The Hebrew language	69	2.47%	18.73%
السياقة	Driving	56	2.01%	20.74%
قانون عمل النساء	Law of working women	52	1.86%	22.60%
تعليم العبرية	Learning Hebrew	50	1.79%	24.39%
رخصة السياقة	Driving license	50	1.79%	26.18%
التامين الوطني	National Insurance	49	1.76%	27.94%
"النساء قانون عمل"	Law of working women	38	1.36%	29.30%
خدمات وزارة الداخلية	Ministry of Interior Affairs	33	1.18%	30.48%
تأهيل المعاقين	Training disabled	24	0.86%	31.34%

Table 4
Most frequently requested categories in absolute numbers and percentages ($N = 266,295$)

Category	Times requested	Percentage requested	Cumulative %	Placement on homepage
Work relations	41,719	15.67%	15.67%	3
National Insurance	25,445	9.56%	25.22%	5
Consumers	23,866	8.96%	34.18%	8
Housing and accommodation	20,184	7.58%	41.76%	10
Registrars	16,148	6.06%	47.83%	13

Table 5
Most frequently requested articles in absolute numbers and percentages ($N = 266,295$)

Article	Belongs to category	Times requested	Percentage requested	Cumulative %
Ministry of Interior Affairs – Services available at Post Offices	Registrars	10,957	4.11%	4.11%
Ministry of Transport – Driving licenses	Other	9507	3.57%	7.68%
National Insurance	National Insurance	8145	3.06%	10.74%
Professional training – Ministry of Social Affairs	Education	7607	2.86%	13.60%
Dismissal compensation – Early notice	Work Relations	6806	2.56%	16.16%

order to understand the relation between the most frequent queries and the most frequently visited pages. In Table 6, we present the most frequently visited page for each of the 10 most frequent queries (from Table 2). This analysis enabled us to relate the users' information needs as expressed by their queries with the answers provided by the website. We are not aware of any previous study that employed this type of analysis.

Let us take a closer look at the article that users reach most often when submitting queries to search engines and choose a result in SHIL. The article is: 'Ministry of Interior Affairs – Services available at Post Offices'. In most of the cases (58.1%, 6366 queries) for which the users are directed to this page the submitted query was 'Ministry of Internal Affairs' (in Hebrew). As of the beginning of February 2006, when submitting this query to google.co.il, which was the major source of the external queries, the SHIL page is the third result, immediately after two pages from the Ministry's site. The title of the page clearly states that the page is about services available at Post Offices, and not about the Ministry or the services it provides in general. Thus, even though the query was very general, the users that clicked on the SHIL page were probably interested in this narrower aspect, or were unable to fulfil their information need at the Ministry's website.

8. Discussion

8.1. What information on public and governmental services and entitlements is of interest to users who arrive at SHIL on the Web from search engines?

The most frequently occurring queries appear in Table 2 for Hebrew and in Table 3 for Arabic. The queries in Hebrew are mainly general and/or navigational queries. The pages the users reach for these queries often contain links to the specific government sites. Thus, SHIL acts as an intermediary in these cases. Especially interesting is the case of the Ministry of Labour, which due to the

Table 6
The most frequent queries and the most frequently retrieved article for the query

Query in English	Query frequency	Most frequently retrieved article	Times retrieved for query	Percentage out of query frequency
National Insurance	7226	Category page for the National Insurance, with links to all the articles in this category	6839	94.6%
Ministry of Internal Affairs	6394	Ministry of Interior Affairs – Services available at Post Offices	6366	99.5%
Ministry of Labour	5234	Professional training – Ministry of Industry, Trade and Employment	4869	93.0%
Daylight saving time	5044	Top level page – daylight saving time 2005, links to a single article in this category	5015	99.4%
Rental contract	3891	Rental contract template	3809	97.9%
Licensing Authority	3637	Ministry of Transportation – driver and vehicle licenses	2884	79.3%
Superland (*amusement park*)	3160	Superland – address, opening hours, prices, description	2994	94.7%
National Insurance (*variant spelling*)	2119	National Insurance – disability	1432	67.6%
Vacation fees	1933	Vacation fees – extensive discussion of rights	1582	81.8%
Dismissal compensation	1728	Dismissal compensation – early notice. First part of a two-part extensive article on rights and obligations	1723	99.7%

change in its name is quite unreachable without the help of an intermediary. The queries in Arabic, unlike the Hebrew queries, are information queries. It will be interesting to further investigate the differences between the Hebrew and Arabic queries. Next, we discuss four top queries in detail.

Let us consider the most frequently occurring query, National Insurance. The users were looking for the National Insurance Institute of Israel (www.btl.gov.il/English/eng_index.asp). The query appeared in several variants, two of them appearing among the top 25 queries. By inspecting the frequently occurring queries, we came across a few more variations; altogether SHIL was queried for the National Insurance Institute at least 10,326 times (not counting queries where the users were looking for specific offices) – 3.9% of the queries. When submitting the query National Insurance (in Hebrew) to Google, the top result, of course, is the National Institute's site. The SHIL result, as of January 17, 2006, comes out number four for the first variant (with the Hebrew letter YUD) and number two for the second variant. We have no information how many users chose the National Insurance site, but a considerable number of users preferred to look at the information provided by SHIL on the topic. A possible reason for this could be that the snippet Google displays for the SHIL page on the National Insurance Institute summarizes the page content in a much more appealing fashion. The SHIL snippet states: 'National Insurance – all the information about your rights …'; while the National Insurance Institute's snippet says: 'Notice to the employees about changes in the insurance fees …'.

The third query, Ministry of Labour is also an interesting query. Until February 2003 [45], the ministry in charge of labor affairs was the Ministry of Labour and Social Welfare, but since then the ministry in charge is called the Ministry of Industry, Trade and Employment (the official name in English is Ministry of Industry, Trade and Labour, however in Hebrew the word employment is used) [46]. The site of the Ministry of Social Affairs has no information related to labor issues; there is not even a link on its homepage to the site of the Ministry of Industry, Trade and Employment. In

spite of this, when searching Google in August 2006 for 'Ministry of Labor' (in Hebrew), the top two results are from the Ministry of Social Affairs, while the next two results are from the SHIL site with links to the Ministry of Industry, Trade and Employment. The second result is especially relevant, since it is a list of addresses and telephones of the offices of the 'former Ministry of Labour', now 'Ministry of Industry, Trade and Employment'. Only the fifth result is a page from the site of the Ministry, Trade and Employment (see Figure 3). Thus, the popularity of SHIL for the query 'Ministry of Labour' is not surprising.

The SHIL page is the top result for 'daylight saving time' (in Hebrew) on Google as of January 17, 2006. This is an example of a query where it is rather unclear in advance which site can provide an answer. We presume the users were interested in the dates Israel switched to and from daylight saving time. In Israel, the settings of daylight saving time change frequently, as it is the topic of political haggling. Currently the SHIL page on Superland is only the seventh result on Google (it could have been much higher during the time the data was collected). Seemingly the Superland amusement park did not have a homepage of its own at the time of the data collection, thus people looking for information on the park or on a raging controversy regarding its admission policies for disabled children had to turn to other sites.

Previous studies [4, 5] report the frequent use of search engines as a navigational tool. In this context we examined how many queries contained the term SHIL – in English or in Hebrew with variant spellings (SHIL, SH.I.L or SHI'L). We found 2735 queries (1.03%) that contained a variant of the word SHIL. Users who submitted queries containing the word SHIL were obviously aware of the existence of the SHIL project and used the search engine for navigational purposes.

8.2. How do the queries relate to the titles of the webpages the users reach from the search engine results page?

Table 6 provides details about the webpage the users most often reach for the most frequent queries. Some of the pages are link pages (with links to all pages in the given category), but most of them are content pages, with extensive and updated information. It is interesting to note that the search engines direct the users to different webpages for variant spellings of National Insurance. The reason for this is that National Insurance is spelled differently on the category page and on the page for the disabled. The search engines very consistently direct the users to the same webpages. Since before clicking on a search engine result, users see the title, the URL and the snippet of the page, we assume that the users clicked on the SHIL pages because they were somewhat confident that they would find an answer to their information problem.

8.3. What are the characteristics of these queries (query length, query terms, time submitted, etc.)? Are there specific problems because the site is multilingual?

We have not experienced specific problems because of the multilingualism of the site, except perhaps that we discarded some external queries because they were not encoded properly.

In the SHIL on the Web log, 86.9% of the queries are between two and four words long, whereas for the other logs the percentage ranges between 59.7% and 67.8% (see Table 7). This finding is rather surprising, especially since Hebrew is a 'compact' language. Most of the prepositions, the definite article and some of the conjunctions are prefixed to the word and are not stand-alone words, for example The Ministry of Finance is only a two-word phrase in Hebrew – MISRAD HAOTZAR. Thus, we expected that in the SHIL log short queries would be more prevalent than in the mainly English logs. The SHIL log is the only log for which the percentage of four-word queries is higher than the percentage of single-word queries. Still, if we consider the most frequently occurring queries (see Table 2), we see that only two queries in the set are single-word queries.

It would be of great interest to analyze the Hebrew language queries submitted to general search engines with the SHIL logs in order to find out whether the Hebrew language searcher behaves differently from American or European searchers in terms of query phrasing, i.e. whether the Hebrew language searcher submits longer queries in general or only when looking for information on public and governmental services and entitlements.

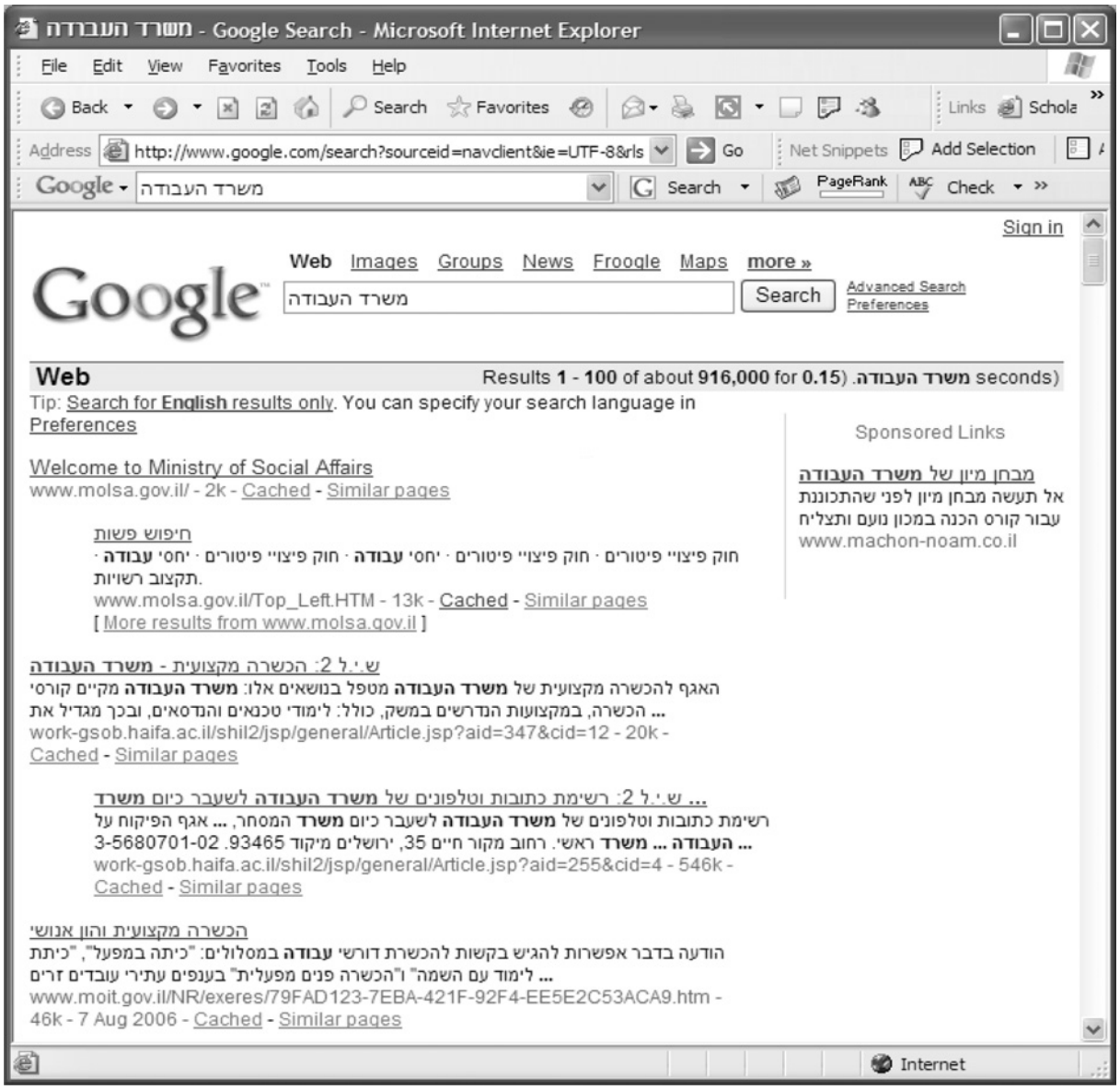

Fig. 3. Results of the query Ministry of Labour (in Hebrew) as of August 9, 2006.

8.4. Comparison with previous studies

The results of our analysis are compared with the data reported for the Utah state government website [29] and with the results of the analyses of the logs of general search engines [23]. The Utah state government website provides comparable information to the information provided by SHIL. Comparison with logs of general search engines is mainly relevant for query length.

It is possible that only queries submitted to SHIL are relatively long, because users that reach the SHIL website have rather complex problems, and the phrasing of such problems requires longer queries. If we accept this explanation, we should expect a similar finding for the Utah government website. However, the query length distribution for the Utah site does not differ from the distribution for the general search engines. The longest query for the Utah government website was a 40-word query, for SHIL the longest query contained 56 words (due to some automatic transformation by the search engine); the longest 'real' query was 40 words long for SHIL on the Web as well.

Table 7
Reported query lengths in several studies, cumulative percentages in parentheses

Query length	SHIL 2005	Utah 2003 [28]	AlltheWeb 2002 [22, p. 84]	Alta Vista 2002 [5]	AlltheWeb 2001 [22, p. 84]	Excite 2001 [16; 22, p. 93]	Knoxville 1997–2001 [27]
1	5.9% (5.9%)	30.7% (30.7%)	33.1% (33.4%)	20.4% (20.4%)	25.1% (25.1%)	25.0% (25.0%)	38.8% (38.8%)
2	42.6% (48.5%)	37.0% (67.7%)	32.6% (66.0%)	30.8% (51.2%)	35.8% (35.8%)	31.7% (56.7%)	41.5% (80.3%)
3	29.2% (77.7%)	19.2% (86.9%)	18.9% (84.9%)	22.8% (74.0%)	22.4% (83.3%)	22.8% (79.5%)	13.4% (93.7%)
4	15.1% (92.8%)	7.6% (94.5%)	8.2% (93.1%)	12.0% (86.0%)	9.6% (92.9%)	8.8% (88.3%)	6.2% (99.9%)
Mean term length	2.79	2.25	2.30	2.92	2.40	2.60	2.0

The most frequently used query modifier was the quotation marks for phrase search (7617 queries, 2.9% of the queries). Wolfram et al. [14] in their analysis of an Excite log of about 50,000 queries found that Boolean operators and query modifiers (+/–) were used in 24.7% of the queries, and were misused in about 34.9% of the cases. In our query log, the occurrence of Boolean operators and search modifiers was only 3.6%. A possible reason could be that it is not straightforward to use some of these operators with Hebrew or Arabic; the use of the quotation marks for phrase searches is especially difficult. The query looks OK when searching only for a phrase, but when trying to search for a phrase and some standalone word(s), the quotation marks move unexpectedly and it is difficult to tell whether the search engine is able to interpret the query correctly.

Comparing the top queries in the SHIL log to the top queries in the Utah state website [28], we see some similarities. Frequently occurring queries in the Utah log included: employment, unemployment, jobs (queries similar to 'dismissal compensation', 'Ministry of Labour'), real estate (somewhat similar to 'tenant protection law', 'Housing Advice Association'), drivers license (similar to 'Licensing Authority') and Medicaid ('Ministry of Health').

9. Limitations

The logfiles of a single site during a specific time period were analyzed. The findings are not necessarily generalizable. However, the method of pairing queries with the visited webpages can be applied in other settings as well.

10. Conclusions

SHIL on the Web is a large content site that provides information on public and governmental services and entitlements. It seems that awareness to this site among Israeli internet users was not very high as a considerable number of external hits were initiated through search engines. SHIL serves the public well even without brand awareness; it achieves its mission by enhancing findability on the web. 'Findability is the biggest story on the Web today, and its reach will only grow as the tidal waves of channel convergence and ubiquitous computing wash over our shores' [6, p. 13]. In addition to providing information, SHIL is an intermediary, directing users to the site of the appropriate public service.

In this study we employed a novel technique of pairing the external queries with the webpages the user was shown when clicking on the search result. The query together with the visited webpage provide a better insight than other log analysis methods to the information problem the user faced when searching the web. Users who reached the SHIL pages as a result of their searches saw the search snippet first and consciously chose to view the SHIL page.

Rather surprisingly longer queries were more prevalent in the SHIL log than in the other logs analyzed in the literature, even though Hebrew is a 'compact' language, with articles and prepositions prefixed to the nouns. The differences between the most frequent Hebrew and Arab queries should be further investigated. In addition, in the future we plan to incorporate the log analysis with user studies in order to gain a better understanding of users' information needs on public and governmental services and entitlements.

References

[1] R. Marcella and G. Baxter, The information needs and information seeking behaviour of a national sample of the population in the United Kingdom, with special reference to needs related to citizenship, *Journal of Documentation* 55(2) (1999) 159–83.

[2] K.E. Fisher, M. Saxton, C. Naumer and C. Pusateri, *WIN 2–1–1: Performance Evaluation and Cost-Benefit Analysis of 2–1–1 I&R systems* (2005). Available at: http://ibec.ischool.washington.edu/win211.pdf (accessed 7 August 2006).

[3] L. Rainie and J. Horrigan, A decade of adoption: how the internet has woven itself into American life. *PEW Internet and American Life Project* (2005) Available at: www.pewinternet.org/pdfs/Internet_Status_2005.pdf (accessed 7 August 2006).

[4] Nielsen/Netratings, Top search terms reveal that users rely on search engines to navigate their way to common Web sites, according to Nielsen/Netratings. *Nielsen/Netratings Press Release*. (2006). Available at: www.nielsen-netratings.com/pr/pr_060118.pdf (accessed 7 August 2006).

[5] B.J. Jansen, A. Spink and J. Pedersen, A temporal comparison of AltaVista web searching, *Journal of the American Society for Information Science and Technology* 56(6) (2005) 559–70.

[6] A. Broder, A taxonomy of Web search, *ACM SIGIR Forum* (2002). Available at: www.sigir.org/forum/F2002/broder.pdf (accessed 7 August 2006).

[7] P. Morville, *Ambient Findability* (O'Reilly Media, 2005).

[8] T.P. Novak and D.L. Hoffman, New metrics for new media, *W3C Journal* 2 (1997). Available at: www.w3journal.com/5/s3.novak.html (accessed 7 August 2006).

[9] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Mastiner and T. Berners-Lee, *RFC 2616 – Hypertext Transfer Protocol – HTTP/1.1*. (1999). Available at: www.faqs.org/rfcs/rfc2616.html (accessed 7 August 2006).

[10] Wikipedia, *Referer* (2006). Available at: http://en.wikipedia.org/wiki/Referer (accessed 7 August 2006).

[11] C. Silverstein, M. Henzinger, H. Marais and M. Moricz, Analysis of a very large Web search engine query log, *ACM SIGIR Forum* 33(1) (1999) 6–12. Available at: www.acm.org/sigir/forum/F99/Silverstein.pdf (accessed 7 August 2006).

[12] N.C.M. Ross and D. Wolfram, End user searching on the Internet: an analysis of term pair topics submitted to the Excite search engine, *Journal of the American Society for Information Science* 51(10) (2000) 949–58.

[13] A. Spink, D. Wolfram, B.J. Jansen and T. Saracevic, Searching the Web: the public and their queries, *Journal of the American Society for Information Science and Technology* 52(3) (2001) 226–34.

[14] D. Wolfram, A. Spink, B.J. Jansen and T. Saracevic, Vox populi: the public searching of the Web, *Journal of the American Society for Information Science and Technology* 52(12) (2001) 1073–4.

[15] A. Spink, S. Ozmutlu, H.C. Ozmutlu and B.J. Jansen, U.S. versus European Web searching trends, *SIGIR Forum* Fall (2002) 32–8. Available at: www.acm.org/sigir/forum/F2002/spink.pdf (accessed 7 August 2006).

[16] B.J. Jansen and A. Spink, An analysis of Web searching by European AlltheWeb.com users, *Information Processing and Management*. 41(2) (2005) 361–81.

[17] A. Spink, B.J. Jansen, D. Wolfram and T. Saracevic, From e-sex to e-commerce: Web search changes, *IEEE Computer* 35(3) (2002), 107–9.

[18] S. Ozmutlu, A. Spink and H.C. Ozmutlu, A day in the life of Web searching: an exploratory study, *Information Processing and Management* 40(2) (2004) 319–45.

[19] S.M. Beitzel, E.C. Jensen, A. Chowdhury, D. Grossman and O. Frieder, Hourly analysis of a very large topically categorized Web query log. In: K. Jarvelin et al. (eds), *Proceedings of the 27th Annual International ACM Conference on Research and Development in Informational Retrieval, SIGIR '04, Sheffield, 25–29 July* (ACM, 2004) 321–28.

[20] S. Park, J.H. Lee, and H.J. Bae, End user searching: a Web log analysis of NAVER, a Korean Web search engine. *Library and Information Science Research* 27(2) (2005) 223–31.

[21] Y. Xie and D. O'Hallaron,. Locality in search engine queries and its implications for caching. In: P. Kermani et al. (eds), *Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'02) 23–27 June 2002, New York* (IEEE, 2002) 307–17.
[22] A. Spink and B.J. Jansen, *Web Search: Public Searching of the Web* (Kluwer Academic, Dordrecht, 2004).
[23] B.J. Jansen and A. Spink, How are we searching the World Wide Web? A comparison between nine search engine transaction logs, *Information Processing and Management* 42(1) (2006), 248–63.
[24] W. Croft, R. Cook and D. Wilder, *Providing government information on the Internet: experiences with THOMAS*. Paper presented at the *Digital Libraries Conference, Austin, Texas* (1995). Available at: http://thomas.loc.gov/home/dlpaper.html (accessed 7 August 2006).
[25] S. Jones, S.J. Cunningham and R. McNab, Usage analysis of a digital library. In: I. Witten et al. (eds), *Proceedings of the Third ACM Conference on Digital Libraries, 23–26 June 1998, Pittsburgh* (ACM, 1998) 293–4.
[26] F. Cacheda and A. Vina, Experiences retrieving information in the World Wide Web. In: *Proceedings of the 6th IEEE Symposium on Computers and Communications, 3–5 July 2001, Hammamet, Tunisia (ISCC'01)* (IEEE Compuer Society, Los Alamitos, 2001) 72–9.
[27] P. Wang, M.W. Berry and Y. Yang, Mining longitudinal Web queries: trends and patterns, *Journal of the American Society for Information Science and Technology* 54(8) (2003) 743–58.
[28] M. Chau, X. Fang and O.R.L. Sheng, Analysis of the query logs of a Web site search engine, *Journal of the American Society for Information Science and Technology* 56(13) (2005), 1363–76.
[29] M. Thelwall, Web log file analysis: backlinks and queries, *Aslib Proceedings* 53(6) (2001) 217–23.
[30] P.M. Davis, Information-seeking behavior of chemists: a transaction log analysis of referral URLs, *Journal of the American Society for Information Science and Technology* 55(4) (2004) 326–32.
[31] M.J. Bates, The design of browsing and berrypicking techniques for the online search interface, *Online Review* 13 (1989) 407–24.
[32] G.M. Marchionini, *Information Seeking in Electronic Environments* (Cambridge University Press, Cambridge, 1995).
[33] C.W. Choo, B. Detlor and D. Turnball, Information seeking on the Web – an integrated model of browsing and searching. In: *Proceedings of the 1999 ASIS Annual Meeting* (1999). Available at: http://choo.fis.utoronto.ca/fis/respub/asis99/ (accessed 2 December 2006).
[34] D. Ellis, A behavioural model for information retrieval system design, *Journal of Information Science* 15(4–5) (1989) 237–47.
[35] F. Aguilar, *Scanning the Business Environment* (Macmillan, New York, 1967).
[36] C. Hölscher and G. Strube, Web search behavior of Internet experts and newbies. In: *Proceedings of WWW9* (1999). Available at: www9.org/w9cdrom/81/81.html (accessed 2 December 2006)
[37] K.C. Laudon and C.G. Traver, *E-commerce: Business, Technology and Society* (Prentice Hall, Upper Saddle River, NJ, 2006)
[38] J.P. Bailey and Y. Bakos, An exploratory study of the emergence of electronic intermediaries, *International Journal of Electronic Commerce* 1(3) (1997) 7–20.
[39] R.S. Burt, Second-hand brokerage: evidence on the importance of local structure for managers, bankers, and analysts, *Academy of Management Journal* (forthcoming).
[40] M.B. Sarkar, B. Butler and C. Steinfeld, Intermediaries and cybermediaries: A continuing role for mediating players in the electronic marketplace, *Journal of Computer Mediated Communication* 1 (1995). Available at: http://jcmc.indiana.edu/vol1/issue3/sarkar.html (accessed 7 August 2006)
[41] C. Vishik and A. B. Whinston, Knowledge sharing, quality and intermediation. In: D. Georgakopoulos et al. (eds), *Proceedings of the International Joint Conference on Work Activities Coordination and Collaboraton (WASS '99) 22–25 February 1999, San Francisco*, (ACM, 1999) 157–66.
[42] G.M. Giaglis, S. Klein and M. O'Keefe, Disintermediation, reintermediation, or cyberintermediation? The future of intermediaries in electronic marketplaces. In: S. Klein et al. (eds), *Proceedings of the 12th International Bled Electronic Commerce Conference, 7–9 June 1999, Bled, Slovenia* (1999) 389–407. Available at: http://citeseer.ist.psu.edu/252518.html (accessed 6 February 2007).
[43] B. Graham, *A New Way of Tracking Blank Referrals* (2005). Available at: www.webmasterworld.com/forum39/3213.htm (accessed 7 August 2006).
[44] *AOL Apologizes for Release of User Search Data*. (2006). Available at: http://news.com.com/AOL+apologizes+for+release+of+user+search+data/2100–1030_3–6102793.html (accessed 9 August 2006).
[45] Government of Israel, *All Ministers in the Ministry of Labor and Social Welfare* (2006). Available at: www.knesset.gov.il/govt/eng/GovtByMinistry_eng.asp?ministry=15 (accessed 7 August 2006).
[46] Government of Israel, *All Ministers in the Ministry of Trade and Industry* (2006). Available at: www.knesset.gov.il/govt/eng/GovtByMinistry_eng.asp?ministry=6 (accessed 7 August 2006).