

Popularity and findability through log analysis of search terms and queries: The case of a multilingual public service Web site

Gilad Ravid

Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheva, Israel

Judit Bar-Ilan; 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

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

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

JIS Editorial Process

Received:

Revised:

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 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/she 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

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 in 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

Popularity and findability through log analysis of search terms and queries

users do not bother to remember or to store the URLs of Web sites 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 *referrer* (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 months 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 Web site. However, the analysis showed that this was not the case, only 0.6% of the queries were looking for the SHIL Web site. 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/her 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

Popularity and findability through log analysis of search terms and queries

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 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 neither to other parts of the site nor 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 to 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 to Israel in the 1990's. There is an English language interface as well.

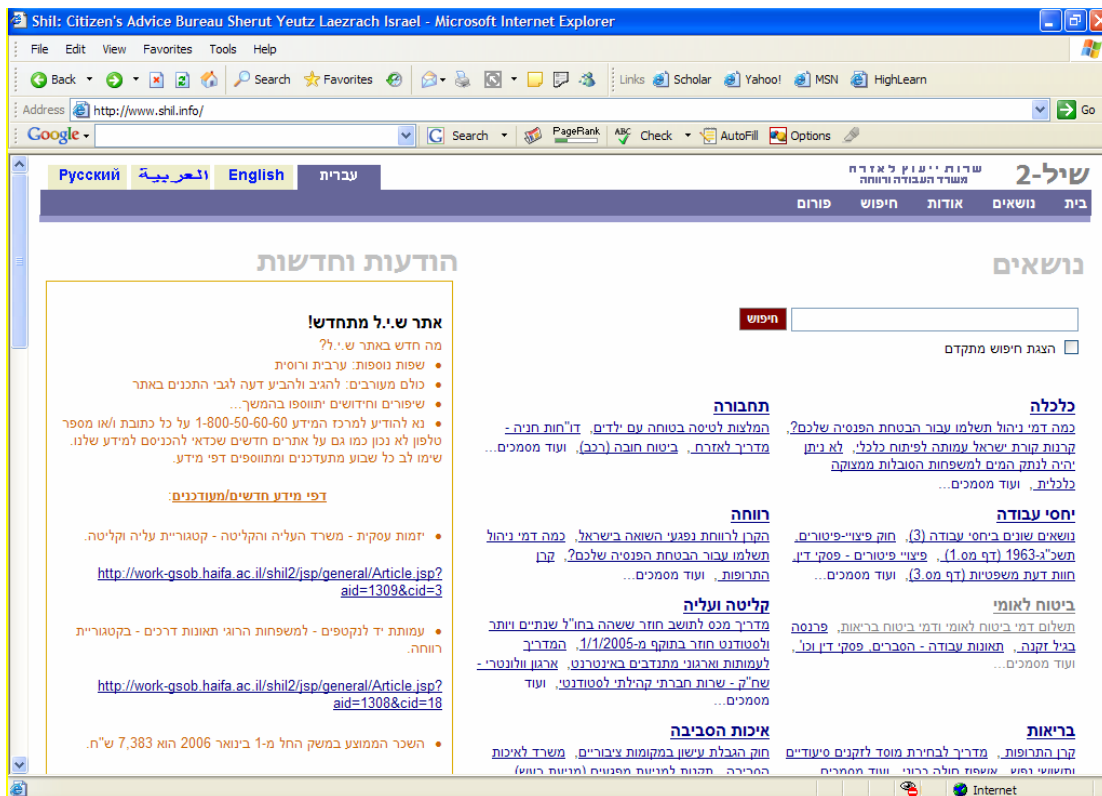


Figure 1. Hebrew language homepage of SHIL on the Web

3. Related studies

There are a number of studies that 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 days 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%).

Popularity and findability through log analysis of search terms and queries

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 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 sessions 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 9:00 and 10:00 AM, while for AlltheWeb the largest number of queries per hour arrived between 8:00 and 9:00 AM. Hourly traffic of the AOL search engine powered by Google was analyzed by Beitzel et al. [19], according to their findings the busiest hour of the day was between 9:00 and 10:00 PM. 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 Web site search studies was carried out by Croft, Cook and Wilder [24]. They examined the usage of the THOMAS Web site, 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 referer field of Web site 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 ten 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 behaviour in electronic environments

One of the first models for online searching is Bates' berrypicking model [31]. She described online searching as a process where the user during his/her 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-centred 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/she can proceed more successfully. The stages are: recognize and accept an information problem; define and understand the problem;

Popularity and findability through log analysis of search terms and queries

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 Choo, Detlor and Turnball [33]. Their model was based on existing models for information behaviour (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 Holscher and Straube [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 Web site. 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 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 [40, 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 need and the way they formulate them, than a simply analysis of the distribution of the pages visited on the Website. In combination with the specific pages visited we can gain insights to 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 to SHIL on the Web from search engines?
- How do the queries relate to the titles of the Web pages 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 multi-lingual?
- How do these queries compare to those submitted to general search engines and to queries submitted to local search engines, especially to the queries submitted to the Utah state government Website [29] that provides information that is comparable to the information provided by the SHIL site?

Popularity and findability through log analysis of search terms and queries

6. Research design

6.1. Data collection

The dataset contained all the external hits to the SHIL Website for an eight months 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 *referers* 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 her 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 referer 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 referers. 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 referers 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, 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 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.

Popularity and findability through log analysis of search terms and queries

The mean number of query length is 2.9, 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 2,212, out of which 1,432 (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. Occurrences	% 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	4,275	1.61%
7 words	1,258	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 twenty-five 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 to 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 2,729 Arabic queries (1.02%), 614 queries in Russian (0.23%) and 1162 queries included Latin characters (0.44%). The twenty-five 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.

Popularity and findability through log analysis of search terms and queries

Table 2. Most frequently occurring queries and their meanings, in absolute numbers and in percentages (N=266,295). Text in *italics* was added to clarify the query

query in original language	query in English	Freq.	%	cum %
ביטוח לאומי	National Insurance	7,226	2.71%	2.71%
משרד הפנים	Ministry of Internal Affairs	6,394	2.40%	5.11%
משרד העבודה	Ministry of Labour	5,234	1.97%	7.08%
שעון קיץ	daylight saving time	5,044	1.89%	8.97%
חוזה שכירות	rental contract	3,891	1.46%	10.44%
משרד הרישוי	Licensing Authority	3,637	1.37%	11.80%
סופרלנד	Superland (<i>amusement park</i>)	3,160	1.19%	12.99%
בטוח לאומי	National Insurance (<i>variant spelling</i>)	2,119	0.80%	13.78%
דמי הבראה	vacation fees	1,933	0.73%	14.51%
פיצויי פיטורין	dismissal compensation	1,728	0.65%	15.16%
משרד הרישוי חולון	licensing authority Holon (<i>the central office is in Holon</i>)	1,685	0.63%	15.79%
בית משפט לתביעות קטנות	Small claims court	1,392	0.52%	16.31%
בית החייל	Soldiers' residence	1,383	0.52%	16.83%
היחידה להכוונת חיילים משוחררים	The unit for advising discharged soldiers	1,327	0.50%	17.33%
רשם העמותות ש.י.ל.	Non-profit organizations' registrar SHIL	1,072	0.40%	17.73%
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 (<i>an organization providing help for sick children</i>)	830	0.31%	20.10%
משרד רישוי	Licencing Authority (<i>variant spelling</i>)	763	0.29%	20.39%
לשכת הבריאות	Ministry of Health	753	0.28%	20.67%
משרד העבודה והרווחה	Ministry of Labour and Social Affairs (<i>full name of the Ministry</i>)	750	0.28%	20.95%

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.

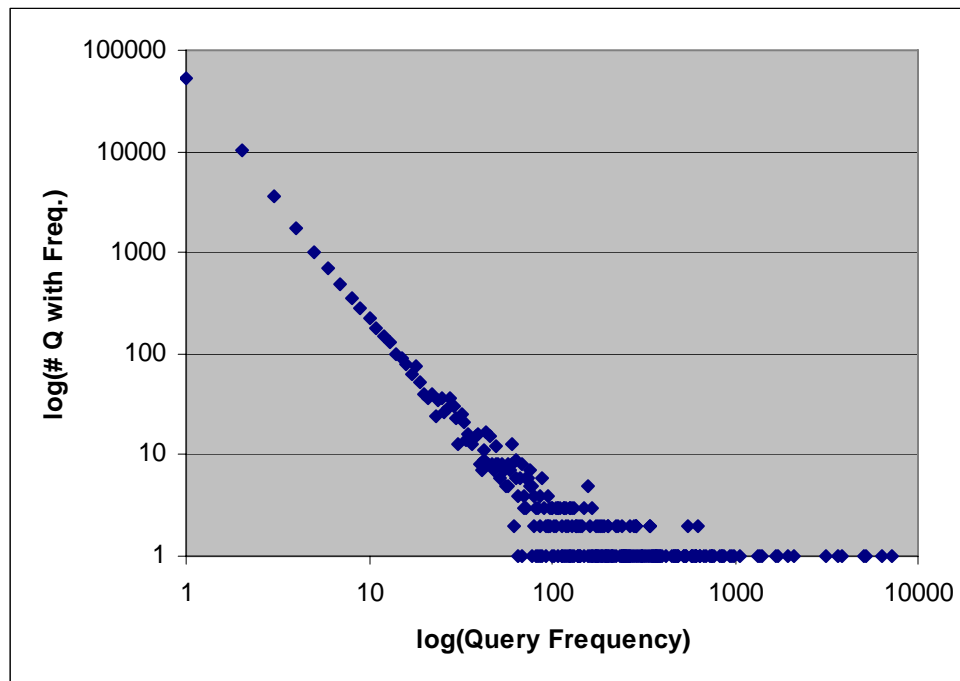


Figure 2. Rank-frequency distribution of queries

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 twenty five most popular words cover an even larger percentage of the total words (28.61%) than the twenty five most popular queries out of the total number of words (20.95%).

7.3. *Queries in Arabic*

The 2,729 queries in Arabic were analyzed separately. Table 3 lists the twelve 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

Popularity and findability through log analysis of search terms and queries

related to driving. Rather interestingly, the query lesbianism (مساحقات) appeared ten 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.

Table 3. Most frequently occurring queries in Arabic and their meanings, in absolute numbers and in percentages (N=2,729)

original query	translated query	no. occurrences	% of queries in Arabic	cum.%
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	the 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%
التأمين 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%

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 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 ten 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 Website. We are not aware of any previous study that employed this type of analysis.

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

Category	times requested	% requested	cum. %	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	% requested	cum. %
Ministry of Interior Affairs – Services available at Post Offices	Registrars	10,957	4.11%	4.11%
Ministry of Transport – Driving licenses	Other	9,507	3.57%	7.68%
National Insurance	National Insurance	8,145	3.06%	10.74%
Professional training – Ministry of Social Affairs	Education	7,607	2.86%	13.60%
Dismissal compensation – Early notice	Work relations	6,806	2.56%	16.16%

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 Web site.

Popularity and findability through log analysis of search terms and queries

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

Query in English	Query freq.	Most frequently retrieved article	Times retrieved for query	% out of query freq.
National Insurance	7,226	Category page for the National Insurance, with links to all the articles in this category	6,839	94.6%
Ministry of Internal Affairs	6,394	Ministry of Interior Affairs – Services available at Post Offices	6,366	99.5%
Ministry of Labour	5,234	Professional training – Ministry of Industry, Trade and Employment	4,869	93.0%
daylight saving time	5,044	Top level page – daylight saving time 2005, links to a single article in this category	5,015	99.4%
rental contract	3,891	Rental contract template	3,809	97.9%
Licensing Authority	3,637	Ministry of Transportation – driver and vehicle licenses	2,884	79.3%
Superland (<i>amusement park</i>)	3,160	Superland – address, opening hours, prices, description	2,994	94.7%
National Insurance (<i>variant spelling</i>)	2,119	National Insurance - disability	1,432	67.6%
vacation fees	1,933	Vacation fees – extensive discussion of rights	1,582	81.8%
dismissal compensation	1,728	Dismissal compensation – early notice. First part of a two parts extensive article on rights and obligations	1,723	99.7%

8. Discussion

8.1. What information on public and governmental services and entitlements is of interest to users who arrive to 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 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 (http://www.btl.gov.il/English/eng_index.asp). The query appeared in several variants, two of them appear among the top twenty-five queries. By inspecting the frequently occurring queries, we came across a few more variation, 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 you 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 charged is called Ministry of Industry, Trade and Employment (the official name in English is Ministry of Industry, Trade and Labour, however in Hebrew the word emploment is used) [46]. The site of the Ministry of Social Affairs has no information related to labour 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 telephone 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

Popularity and findability through log analysis of search terms and queries

users were interested about the dates Israel switched to and from the daylight saving time. In Israel, the settings of daylight saving time changed frequently, as it was the topic of political haggling. Currently the SHIL page on the Superland is only the seventh result on Google (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 have to turn to other sites.

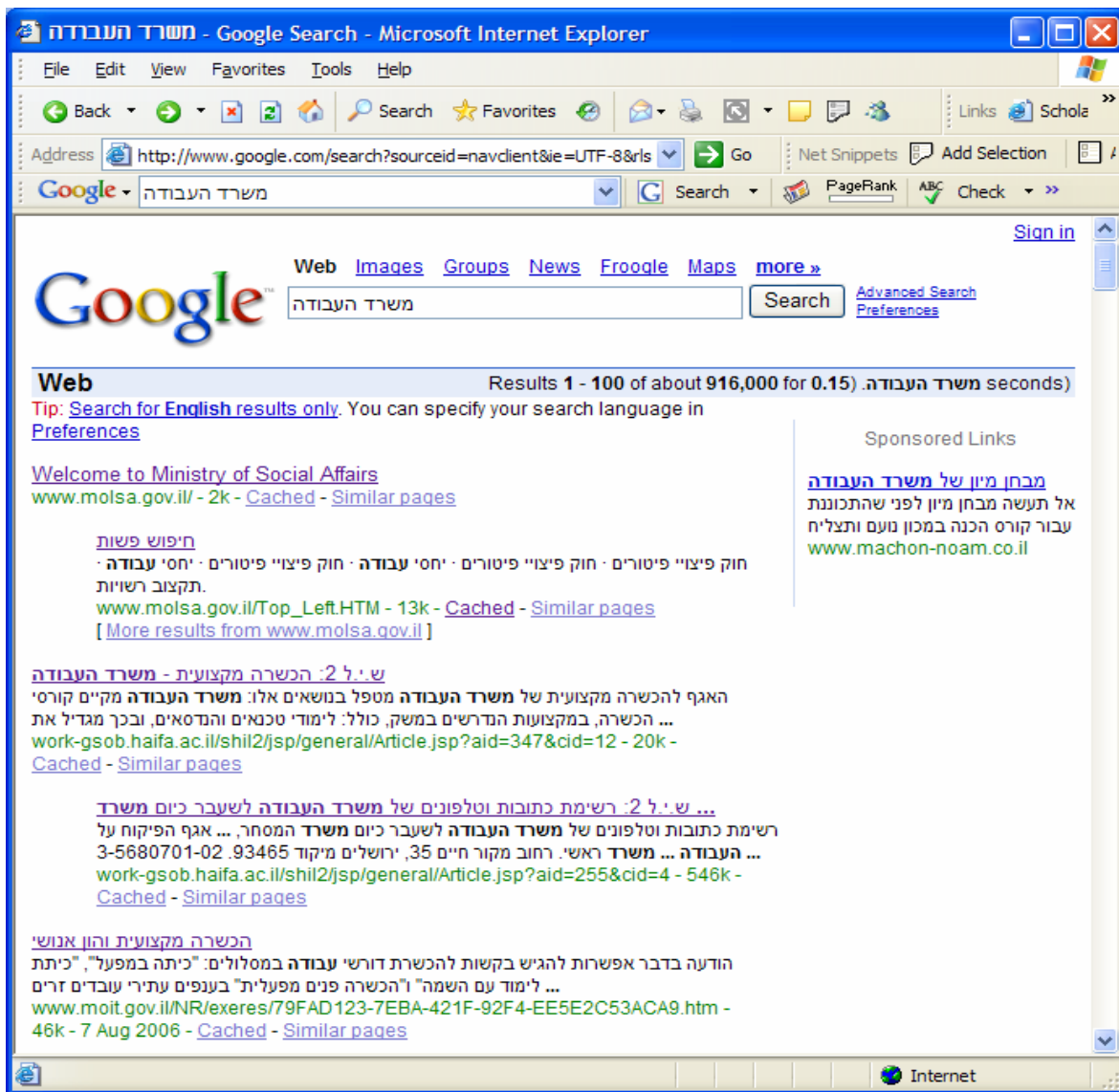


Figure 3: Results of the query Ministry of Labour (in Hebrew) as of August 9, 2006

