JAMIA PrePrint: Accepted Article. Published January 9, 2007 as doi:10.1197/jamia.M2191

A Day in the Life of PubMed: Analysis of a Typical Day's Query Log

Jorge R. Herskovic[1], MD, MS, Len Y. Tanaka[1, 2], MD, William Hersh[3] MD, Elmer V. Bernstam[1, 4] MD, MS, MSE

[1]The University of Texas School of Health Information Sciences at Houston

[2]Department of Pediatrics, Division of Pediatric Critical Care, The University of Texas School of Medicine at Houston

[3] Department of Medical Informatics & Clinical Epidemiology, Oregon Health and Science University

[4]Department of Internal Medicine, Division of General Internal Medicine, The University of Texas School of Medicine at Houston

Corresponding author: Dr. Elmer V. Bernstam, The University of Texas School of Health Information Sciences at Houston

7000 Fannin St. Suite 600

Houston, TX 77030

Elmer.V.Bernstam@uth.tmc.edu

Abstract

Objective: To characterize PubMed usage over a typical day and compare it to previous studies of user behavior on Web search engines.

Design: We performed a lexical and semantic analysis of 2,689,166 queries issued on PubMed over 24 consecutive hours on a typical day.

Measurements: We measured the number of queries, number of distinct users, queries per user, terms per query, common terms, Boolean operator use, common phrases, result set size, MeSH categories, used semantic measurements to group queries into sessions, and studied the addition and removal of terms from consecutive queries to gauge search strategies.

Results: The size of the result sets from a sample of queries showed a bimodal distribution, with peaks at approximately 3 and 100 results, suggesting that a large group of queries was tightly focused and another was broad. Like Web search engine sessions, most PubMed sessions consisted of a single query. However, PubMed queries contained more terms.

Conclusion: PubMed's usage profile must be considered when educating users, building user interfaces, and developing future biomedical information retrieval systems.

I. INTRODUCTION

PubMed is an interface to MEDLINE, the largest biomedical literature database in the world. The United States National Library of Medicine (NLM) of the National Institutes of Health (NIH) publishes general usage statistics (1), but not detailed query information. Information Retrieval (IR) researchers use log analyses (2-4) to understand user behavior such as typical query length and complexity (5, 6), how many results users look at (7), and use of Boolean operators (8) among others. These data provide insight into system performance, and inform user interface design and user education. The goal of this study was to obtain similar insight into PubMed for the IR community, providers of biomedical search systems, educators, and the general public.

A. Background

Query logs are usually derived from server logs and contain queries issued by users. Queries are traditionally grouped into sessions, which are series of related queries issued by the same user. Analyses of Web search engine query logs form the foundation of what we know about user searching on the Web.

In a 1998 study focused on AltaVista, Silverstein et al. found that most users issued simple queries of three or fewer terms, used operators in approximately 20% of cases, and rarely went beyond the first page of results (2). Similarly, Jansen et al. studied Excite and found that 66% of users issued only one query and those queries were usually short. Users were equally likely to narrow the query by adding terms, or broaden by removing terms, during a single session (4). As in the AltaVista study, few users clicked on results after the first page, although they reviewed some results after the first page (4, 9). Chau et al. analyzed the query log of a Utah government site search engine and found that this special-purpose search engine had a different usage profile than general purpose engines (10). Thus, PubMed may have a different usage profile than general Web search engines.

The NLM estimated in 2002 (the last year for which we could obtain this information) that one third of PubMed's users were members of the general public, while the remaining two thirds were health care professionals and researchers (11). MEDLINE users leveraged its unique features (12) and studies show that experienced MEDLINE users such as medical librarians perform searches with higher recall and precision than novice clinicians or members of the general public (13, 14). PubMed users may employ different search strategies than Web search engine users.

Three kinds of queries have been characterized according to their underlying intent (15). "Informational queries" are intended to satisfy information needs on a topic. For example, a user may search for "myocardial infarction." In contrast, "navigational queries" are intended to retrieve a specific document or set of documents. For example, the query "j am med inform assoc [journal] AND 2006 [dp] AND 96 [pg]" intends to retrieve a specific article. When users issue "transactional queries," they are searching to perform web-mediated activities such as shopping or banking. Transactional queries do not have a direct PubMed equivalent.

The distinction between informational and navigational queries reflects the distinction between IR and database access. Whereas IR focuses on access to relatively unstructured data (e.g., free text), database management systems provide access to highly structured data. Therefore, identifying which records to return is a critical issue in IR. In contrast, compact storage and efficient retrieval are critical database issues. If PubMed users issue primarily navigational queries, then researchers should focus on optimizing database access. However, if informational queries are common, then IR issues must be addressed.

The goal of this study was to understand PubMed usage. Specifically, we were interested in the length of a typical query/session, the size of the result sets, use of Boolean operators, whether queries were informational or navigational, common search topics, and search strategies.

II. METHODS

A. Log file

We obtained a single day's query log from PubMed, anonymized by the NLM to protect user privacy. The file is publicly available at ftp://ftp.ncbi.nlm.nih.gov/toolbox/pubmed/query-logs, dated October 17, 2005, and is described as "a typical day," but the date of collection is not provided for confidentiality reasons. The file includes: user ID (scrambled), timestamp (seconds since midnight EST) and the query string as entered by the user. According to the NLM, the user ID is provided "so that multiple queries from the same user can be matched" (16) and does not rely solely on the IP address or cookie on the user's computer. In contrast to previous log analyses, our data did not include the entire server log. Specifically, we did not have access to results returned in response to the query, user selections (clicks) or page views.

The log file contained 2,996,301 queries issued by 627,455 different users. Each query represents a single "question" submitted by the user to PubMed. Thus, a session consists of one or more queries (Figure 1). The log file included all queries issued over 24 hours, from midnight to midnight. We arbitrarily excluded very prolific users (over 50 queries/24 hours), since they could represent institutional proxies or programmatic searchers ("bots") rather than individuals. This represented 2,941 users (0.47%) who issued a total of 307,135 queries. We analyzed the remaining 2,689,166 queries. From here on, we refer to this as our entire dataset.

We used custom Python (http://www.python.org) and R (http://www.r-project.org) scripts running on Mac OS X version 10.4.6 (Apple Computer Corp., Cupertino, CA) and Gentoo Linux 2005.0 (http://www.gentoo.org). For analyses that required PubMed servers, we used random samples of queries to avoid overloading the servers or violating their terms of use.

B. Result sets

We first took a sample (referred to as "query sample" from now on) of the log file. We used a random number generator to include each query from the log file with a 0.001 probability, which gave us a sample of 2,708 queries. To retrieve result counts and PubMed's Medical Subject Headings (MeSH) translation for each query in the sample, we submitted them to PubMed via the E-Utilities interface (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html, accessed January 20, 2006). We computed the mean, median, and standard deviation for the number of results per query. We used the MeSH translation to classify queries as informational or navigational. Queries that contained only bibliographic tags (e.g., [pdat], [au]) were deemed navigational, according to the algorithm shown in Figure 2. In other words, queries were considered informational by default, and were counted as navigational only if positively identified as such. We also compared the number of results retrieved by navigational and informational queries. To verify our query-classification algorithm, we classified the same sample manually. We counted queries in which users searched for authors' names exclusively, citation information (like journal name, date of publication, and page numbers) exclusively, or explicit MeSH Terms. If the query was not explicit but the intent was clear (for example, "Smith, AB" is not mapped to an author tag) we classified it according to its intent. We assumed that all other queries were simply textual searches. Thus, navigational queries should be equivalent to the manually classified author and citation queries.

C. Terms

We performed a term analysis to find the most common words and phrases users entered into PubMed. We lowercased each query to eliminate the effect of case. A term was defined as a string separated from others by punctuation, white space, or a string of characters contained within square or curly brackets, or quotes. For example, [MeSH Major Headings] was a single term. We determined the most common terms by counting every occurrence of each distinct term in the query log (excluding single letters and punctuation) and sorted in descending order. We reported both the most common terms and the most common search field tags. We were unable to group equivalent tags (e.g., [author]=[au]) because we were not able to obtain a definitive list of equivalents from the NLM.

We then performed a second order analysis similar to the one described in (2) that detects two-term correlations, regardless of their relative positions in the query (i.e., terms did not have to be adjacent to each other). We computed a correlation coefficient ρ to judge the strength of the relationship between terms in each pair. For Boolean data (occurrence/non-occurrence), ρ is related to χ^2, a statistical measure of deviation from expected frequencies by the formula χ^2=n × ρ^2. For example, if the terms "gastric" and "cancer" appeared frequently in the same query, but not independently, they had a high correlation. This analysis required quadratic storage space. Thus, we arbitrarily considered only pairs of the 25,000 most common terms. The list was filtered using correlation and frequency cutoffs. Correlations that contained stopwords or bibliographic tags were considered uninteresting and were removed manually.

D. Topics

To determine general search topics, we mapped all queries to MeSH (2005 edition) using the NLM's Metamap batch server (http://skr.nlm.nih.gov/) with the default processing options, plus –M "MMI output." We weighted each MeSH term according to the number of mappings for that particular query. The sum of scores for each query was one. For example, if a query was mapped to three MeSH terms, each term was weighted by 1/3. We used the MeSH hierarchy to categorize terms into its top level. We drilled down into the "disease" category (second level) to determine the most popular clinical topics. When a term was classified into multiple categories, we counted its contribution to all categories.

E. Session separation

To gain insight into users' search strategies, we grouped related queries. PubMed's user hash identifies all queries by the same user, but does not consider their time or topic. For example, a users' query history could include queries for "myocardial infarction AND aspirin" at 11:13 AM, “myocardial infarction prevention AND aspirin” at 11:37 AM, and “gastrointestinal stromal tumors” at 12:00 PM. In this hypothetical example, the user made her first query more specific, probably to retrieve fewer, more relevant results, and then switched topics completely. To perform automated analyses, we must be able to group related queries into sessions. This was traditionally done using time thresholds, i.e. if the user waited more than a certain number of minutes, then a new session began. Since there are no prior PubMed log analyses using sessions, and PubMed users might be different from general Web searchers, separating sessions by time may be overly simplistic. Therefore, we performed a semantic analysis over the entire data set to separate users’ queries into sessions.

We relied on detecting a change in topic by evaluating the semantic distance between consecutive queries. Semantic distance reflects difference in the meaning of two concepts For example, “dog” is closer to “cat” (as they are both mammals) than to “pterodactyl,” so the semantic distance between “dog” and “cat” is smaller than between “dog” and “pterodactyl.” By measuring the semantic distance between queries, we expected to group them into sessions better than with arbitrary time thresholds.

We evaluated this claim by performing a small pilot study. We printed a random sample of 2,390 queries issued by 351 individual users (as identified by the NLM-provided user hash). Two of the authors (LYT and JRH) independently identified session boundaries. Sessions were defined as sets of queries in which the user was pursuing the same information need. We compared the results of this exercise to dividing the queries into sessions using a time cutoff (0 to 120 minutes in 1 minute increments) and to our MeSH-based semantic classifier. We found that the semantic classifier had better concordance with human judgment than all time cutoffs. We also used these results to determine the best distance threshold between sessions (3.8), which was used for the rest of the analysis.

We used MeSH mappings for queries and computed semantic distance between consecutive queries. Distance was defined as the shortest path between pairs of concepts on the MeSH tree as shown in **Figure 3**. For this analysis, we only used the highest scoring mapping returned by Metamap for each concept. When we could not map queries to MeSH terms or concepts, we used WordNet (http://wordnet.princeton.edu/). In this case, we walked WordNet's hypernym/hyponym tree to obtain distance measurements directly from the query as entered by the user. We used WordNet 2.0 via the pywordnet Python interface (http://osteele.com/projects/pywordnet/). To simplify implementation, we only used the WordNet noun database.

We assigned weights to each edge according to its depth in the respective tree. We reasoned that "deeper" steps represent less difference than "shallower" ones. For example, in **Figure 3**, "myocardial infarction" is closer to "myocardial ischemia" than the latter is to "heart diseases." The distance score from "heart diseases" to "myocardial ischemia" is thus greater than the one from "myocardial ischemia" to "myocardial infarction." We used one divided by the depth of the topmost node in a pair as a score: the steps in descending order were worth 1, 0.50, 0.33, 0.25, 0.2, etc. points. Our use of a depth-conscious measure has precedents in the literature and, in particular, is similar to the Leacock-Chodorow distance (17). When one or both of the queries in a pair contained more than one term, we paired each term to its closest counterpart in the other query. We then added the individual distances between pairs of terms to obtain a total distance. When we could not compute a distance, we assumed that the queries were part of the same session.

F. Strategies

Once the queries were divided into sessions, we used a smaller random sample (called "strategy sample") that consisted of approximately 1% of users (6,000 users who issued 25,650 queries) to study search strategy. We eliminated sessions with a single query from the strategy sample, and used the E-Utilities to obtain result counts. We determined whether users looked for broader or more specific result sets by comparing the number of results returned by consecutive queries within the same session. For example, if a user's first query in a session retrieved 12,500 articles and her second query retrieved 700, we deduced that she narrowed her query.

III. RESULTS

Table 1 shows basic descriptive statistics. Figure 4 shows the distribution of queries per user. Of the 2,708 queries in the query sample, 436 (16.10%) had no results. The result sets from the remaining 2,272 queries are described in Table 1 and **Figure 5**. Of this sample, 599 queries (22.11%) were classified as navigational. Manual classification of the same dataset showed that approximately one quarter of queries were navigational. Specifically, 22.90% of the queries contained only authors' names or a PubMed author tag; 2.47% contained citation information, and 0.26% had both author names and other citation information. The remaining three quarters (74.37%) were informational searches, and none used MeSH terms explicitly. Excluding queries with no results, approximately 50% of users issued tightly focused queries that returned 10 results or less. In this sample, navigational queries had a median of 37 results per query (range 1 – 133,100) and informational queries had a median of 100 (range 1-4,845,000). The difference is statistically significant (Mann-Whitney test, $p < 0.001$).

Of 2,689,166 queries, 302,386 used at least one Boolean operator (11.24%). The exact number is difficult to ascertain since, officially, PubMed recognizes only uppercase Boolean terms (18) (Table 2). However, it rewrites lowercase Boolean operators to uppercase internally, apparently trying to match the user's intent. This limitation arises because MeSH terms can contain any word including Boolean operators. For example, "Bone and bones" and "not expressed in choriocarcinoma clone 1, human" are MeSH terms. The query log contained 695,018 unique terms. The most common terms are listed in Table 3 (PubMed stopwords (18) were removed). The 50 most common PubMed tags are listed in Table 4. While term counts varied considerably, the majority of queries had fewer than 10 terms (**Figure 6**) with a median of three terms per query.

The second order analysis identified 2,552,940 highly correlated term-pairs. We filtered the list down to a manageable size by arbitrarily keeping all term pairs with a correlation coefficient ρ greater than 0.6, and over 100 occurrences in the query log. This yielded 26 term pairs (Table 5). Uninteresting pairs, like those involving stopwords, did not have high correlation. Evidence-based medicine-related term pairs figure prominently in this list ("randomized controlled", for example, was the most frequent term pair, although it was present in only 0.13% of queries). Other highly correlated terms may be seasonal, such as phrases related to Lyme disease ("burgdorferii garinii").

MetaMap provided MeSH mappings for 1,495,354 of the queries (55.61%). The most common MeSH categories were "Chemicals and drugs" (24.61%), "Diseases" (20.16%), "Biological sciences" (10.79%), and "Anatomy" (10.27%) (Table 6). The subdivisions of the "Diseases" category are shown in Table 7. The most common disease category was "Pathological conditions, signs and symptoms," with 13.03% of all Diseases.

We performed a semantic distance analysis on all queries to divide them into sessions. The majority of users conducted a single search session during the day (**Figure 7**). The query log contained 740,215 sessions. Most of these sessions were short (62.75% had a single query) which is similar to Silverstein's finding that 63.7% of AltaVista sessions consisted of a single query (2) (**Figure 8**).

Excluding sessions with one query from the strategy sample left 4,997 sessions. Of these, 23.30% had monotonically increasing result counts, while 23.66% had monotonically decreasing result counts. The rest did not have a consistent strategy. Therefore, users broadened and restricted their searches in roughly equivalent numbers.

IV. DISCUSSION

We found that PubMed users issued diverse queries on a broad range of topics without dominant phrases. Like Web users, PubMed users favored short queries and issued few queries per session. When they edited consecutive queries in a single session, they were equally likely to broaden or narrow the search. Approximately one quarter (22-26%) of queries were navigational and three quarters were informational. Advanced MEDLINE features such as MeSH terms were seldom used.

Users issued a median of two queries, although there was large variability in query counts per user. Result set sizes had a bimodal distribution, suggesting that there were two classes of queries. There were focused queries with less than ten results and less focused queries with a mode of approximately 100 results. While we do not know whether users actually clicked on these results, the low numbers suggest that they preferred small result sets. Previous studies showed that experienced users were able to search MEDLINE more effectively (13, 14). The bimodal distribution reflects the distinction between navigational and informational queries; it may also be, in part, due to different usage patterns of professional, highly trained users compared to the general public.

PubMed queries had a median of three terms, higher than reported for Excite and AltaVista. 11.2% of PubMed queries contained operators, which was lower than AltaVista (21.4%) (2) and similar to Excite (10.0%) (9). It is possible that PubMed users intended to issue more Boolean queries, but did not uppercase them properly. If we disregard case, 21.8% of the queries contained at least one Boolean operator (Table 2). Therefore, the intended number of Boolean operators was somewhere between 11.2% and 21.8%. In contrast to the differences in Boolean operator use, sessions were of similar length, with approximately as many users issuing a single query per session on PubMed as on Web search engines. The high frequency of sexual and pornographic terms in Web query logs was not seen on PubMed.

Lack of clickthrough information was one of our major limitations. We can only speculate about the number of results users actually reviewed. Other limiting factors were that we can only report on the typical day's log, and thus we cannot exclude temporal artifacts. For example, one may wonder if an article on Lyme disease appeared in the press on the same day these data were captured, or if this level of interest in Lyme disease is constant.

Our findings regarding search strategy rely on the accuracy of our session separation algorithm. Our technique is better at grouping related queries than using a time cutoff, but less straightforward and requires an arbitrary threshold. While we believe that the technique is generalizable, this has not been empirically demonstrated.

In future work, we would like to strengthen and deepen this analysis by including clickthrough data and information on retrieved results, perhaps by using data from local Web proxy logs. In addition to knowing how users search and what they search for, we could determine whether they are successful. For example, a search where the user downloads the full text of an article via a link from PubMed would be considered more successful than a search where the user did not click on any results. Given clickthrough data, algorithms that learn from usage (implicit feedback) can be adapted to PubMed (19).

On the Web, few users review results beyond the first page (2). If a significant proportion of PubMed users also focus on the first few results, then we need to develop ranking strategies that place the most important results at the top (20). In future work we plan to leverage this information to improve biomedical IR.

V. CONCLUSION

We studied a full day of PubMed queries to characterize user search behavior. We found that PubMed users resemble Web search engine users in some respects, like session length. They issue a wide variety of queries on a large variety of topics without dominant search terms or topics. The majority of queries were informational. Therefore developing effective information retrieval strategies remains important. Our findings suggest that educators and PubMed user interface researchers should not focus on specific topics, but overall efficient use of the system. We also found that result sets come in two sizes, with some very broad queries. We hope that these results inform the design and evaluation of future biomedical information retrieval tools.

Acknowledgements

Supported in part by a training fellowship from the W. M. Keck Foundation to the Gulf Coast Consortia through the Keck Center for Computational and Structural Biology, NLM grant 5K22LM008306 and NCRR grant 1UL1RR024148.

Table 1 – User, query, and result set statistics (excluding users with ≥50 queries/24 hours)

Number of queries	2,689,166
Number of users	624,514
Range of queries per user	1 to 49
Average queries/user	4.31
Standard deviation queries/user	5.88
Median queries/user	2.00
Queries in result set sample	2,272
Range of result set sizes in the result set sample	1 to 4,844,731
Average result set size in the result set sample	14,050
Standard deviation of result set size in the result set sample	145,074
Median result set size in the result set sample	68

Table 2 - Boolean operators in queries

Operator	Number of queries with at least one operator (% of total queries)	Number of queries with at least one operator, regardless of case (% of total queries)
AND	292,286 (10.9%)	572,221 (21.3%)
OR	35,658 (1.3%)	41,928 (1.6%)
NOT	4,932 (0.2%)	6,511 (0.2%)
At least one Boolean	302,386 (11.2%)	586,752 (21.8%)

Table 3 - Common terms

Term	Frequency
[author]	133,492
[au]	56,903
[pmid]	53,605
cancer	46,370
cell	39,687
review	35,272
2005	34,840
[pdat]	34,370
[jour]	28,835
[page]	28,023
[volume]	26,464
[title/abstract]	25,713
disease	21,337
protein	20,417
[auth]	19,466
cells	17,574
human	17,512
[mesh]	17,287
[la]	16,937
receptor	15,837
treatment	15,715
[ti]	15,505
syndrome	15,476
2004	14,837
diabetes	14,788
therapy	14,104
[mh]	13,873
2002	13,519
[ta]	12,641
gene	12,285
trial	11,886
2003	11,791
clinical	11,579
eng	11,263
journal	11,249
kinase	11,217
heart	11,217
with	11,031
[pt]	11,030
brain	10,926

Table 4 - 50 most frequent tags in descending order

Tag	Occurrences	Tag	Occurrences
[author]	133,492	[ptyp]	3,597
[au]	56,903	[mesh terms]	3,566
[pmid]	53,605	[ad]	3,408
[pdat]	34,370	[tw]	3,355
[jour]	28,835	[publication type]	3,272
[page]	28,023	[text word]	3,163
[volume]	26,464	[edat]	3,048
[title/abstract]	25,713	[first author]	2,372
[auth]	19,466	[all]	2,114
[mesh]	17,287	[mh:noexp]	1,994
[la]	16,937	[filter]	1,807
[ti]	15,505	[lang]	1,696
[mh]	13,873	[gr]	1,550
[ta]	12,641	[publication date]	1,435
[pt]	11,030	[corporate author]	1,426
[dp]	10,280	[mesh:noexp]	1,395
[majr]	9,822	[pg]	1,109
[issue]	8,888	[language]	1,069
[sb]	7,048	[word]	1,020
[tiab]	5,496	[author name]	974
[all fields]	5,366	[vi]	895
[text]	4,741	[sh]	881
[journal]	4,567	[majr:noexp]	733
[title]	3,885	[rn]	709
[uid]	3,801	[subheading]	649

Table 5 - Common and highly correlated term pairs

Terms	Frequency	ρ
randomized controlled	3,383	0.670
nitric oxide	2,446	0.785
united states	660	0.667
interferences rnais	494	0.982
sirnas rnais	494	0.948
interferences sirnas	494	0.931
carpal tunnel	349	0.675
rotator cuff	245	0.646
carbonic anhydrase	235	0.797
homo sapiens	222	0.617
obsessive compulsive	207	0.652
guillain barre	203	0.649
sexually transmitted	188	0.699
burgdorferi garinii	157	0.777
burgdorferi afzelii	157	0.777
vena cava	156	0.653
garinii afzelii	155	0.975
foramen ovale	142	0.739
ixodes garinii	132	0.673
ixodes afzelii	132	0.673
spina bifida	130	0.727
puerto rico	125	0.709
epidermolysis bullosa	120	0.751
prader willi	108	0.707
mycosis fungoides	104	0.609
circular dichroism	102	0.637

Table 6 – Queries by category according to MeSH mappings

MeSH Category	Percentage
Chemicals and Drugs	24.61%
Diseases	20.16%
Biological Sciences	10.79%
Anatomy	10.27%
Organisms	9.73%
Analytical, Diagnostic and Therapeutic Techniques and Equipment	7.42%
Psychiatry and Psychology	3.82%
Physical Sciences	3.20%
Health Care	2.44%
Persons	2.18%
Information Science	2.08%
Anthropology, Education, Sociology and Social Phenomena	1.33%
Geographic Locations	0.85%
Technology and Food and Beverages	0.77%
Humanities	0.35%
Total	100.00%

Table 7 - Queries by disease type (according to MeSH categories)

MeSH Disease Category	Percentage
Pathological conditions, signs and symptoms	13.03%
Nervous system diseases	8.79%
Neoplasms	7.99%
Cardiovascular diseases	7.35%
Immune system diseases	6.93%
Bacterial infections and mycoses	5.30%
Nutritional and metabolic diseases	5.19%
Skin and connective tissue diseases	4.80%
Musculoskeletal diseases	4.77%
Virus diseases	4.27%
Digestive system diseases	4.08%
Congenital, hereditary, and neonatal diseases and abnormalities	3.98%
Endocrine system diseases	3.77%
Respiratory tract diseases	3.40%
Disorders of environmental origin	3.24%
Eye diseases	3.03%
Hemic and lymphatic diseases	2.67%
Stomatognathic diseases	2.22%
Urologic and male genital diseases	2.11%
Female genital diseases and pregnancy complications	1.49%
Parasitic diseases	0.90%
Otorhinolaryngologic diseases	0.44%
Animal diseases	0.21%
Total	100.00%

Figure 1 - Log file sample

The PubMed log had three columns. The first column contains a PubMed-generated user hash, the second a timestamp (in seconds since midnight at the server's location), and the third is the actual query as submitted. These are a few consecutive lines from the raw log.

```
96NLg4IOFuEAAFsWRTEAAAAF|626|below the knee amputation
C9taYIOFkQAADnfeMkAAAAB|626|ige anti drug detection
Qdbx-4IOFkoAAB64Yrs|626|ht1080 seattle
V9794IOFlwAADdOVHcAAAAF|626|immunoassay blank matrix
faRKGoIOFkMAAHCTgQAAAAAG|626|"electrophysiological characterization"
oAAAAG|626|systems biology neuroscience
so77pYIOFpIAADI2hL8AAAAN|626|basal ganglia memory
swAAAAC|626|roach g, fletcher a, dawson d
swAAAAC|626|roach g, fletcher a, dawson d
-Q09RIIOFkIAAEqRjiEAAAAB|627|Nellgard B anesthetics 2000
-Qv1VYIOFlwAADb2Wq0AAAAP|627|microRNA and neuron
-QzAjoIOFloAABxIo70AAAAM|627|15764753
-RGnboIOFlwAAECtUPYAAAAI|627|C2C12
5ZWgIIOFj0AAD83nhAAAAAD|627|electrostimulation for stroke rehabilitation
9HMgMoIOFpIAAB5rgV0AAAAR|627|"karasuyama.h"[au]
9y4aiYIOFloAAByyn7MAAAAQ|627|death and neuron
BHqvGIIOFkUAAHjljFIAAAAI|627|brummelkamp bernards barcode
CN4YIOFj0AADtmwM8AAAAM|627|chaperon*
QZESZYIOFkgAAGxtz0YAAAAH|627|MCAD 2002
e47uoIOFkEAACF3jWgAAAAJ|627|Hirsch M, O'Donnell J, Olsson A
oulksoIOFpIAAEYyBpwAAAAI|627|((ADHD[ALL])) AND (2005/9/27[Entrez
Date]:2005/10/5[Entrez Date])
sAAAAJ|627|Reh TA retina
xwH11YIOFpEAAHY9epsAAAAD|627|clenbuterol
```

Figure 2 - Algorithm to classify queries as informational vs. navigational

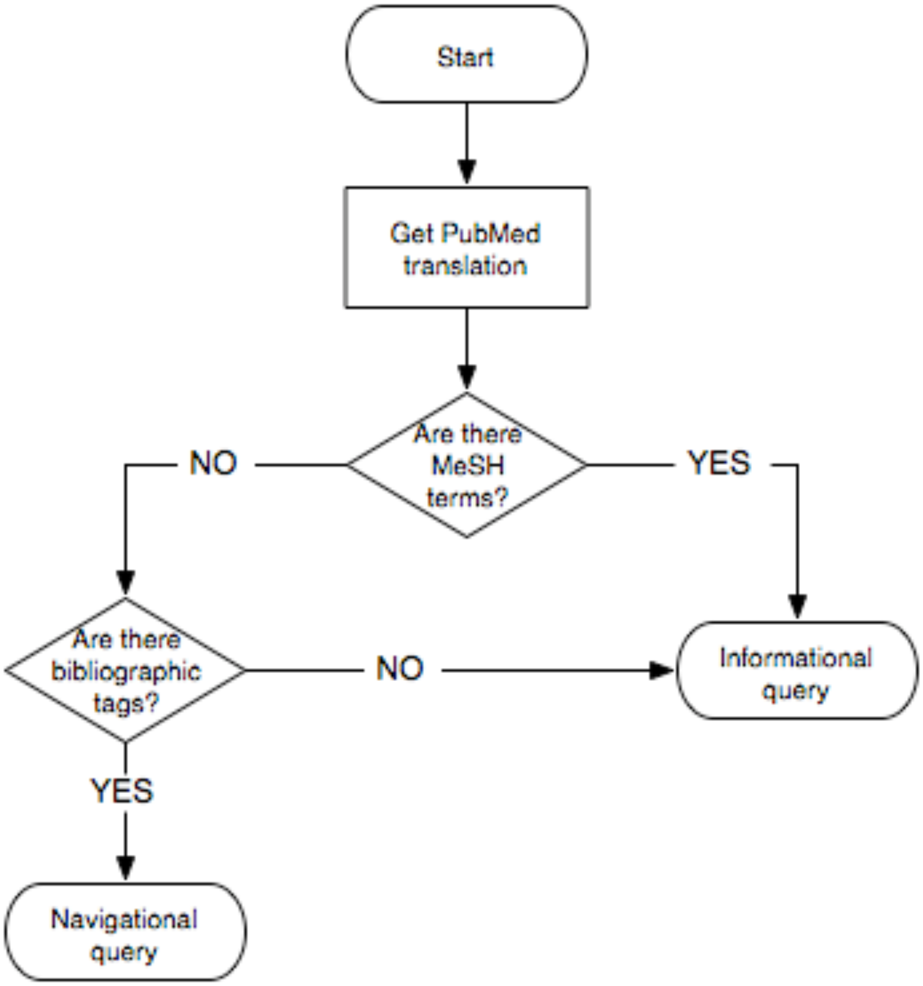
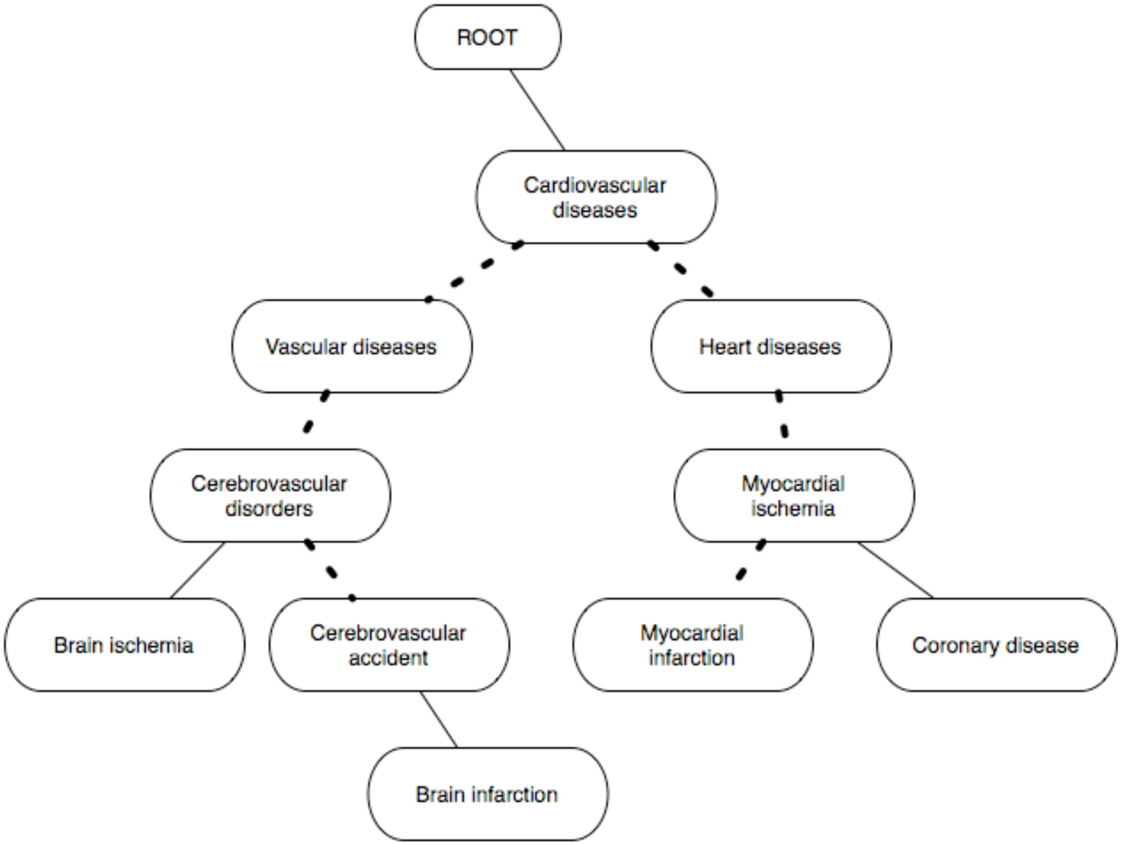


Figure 3 - Semantic distance determination



In this example, we walk through the tree to determine the semantic distance between "Myocardial infarction" and "Cerebrovascular accident," which is six steps long (only one of the possible paths is shown)

Figure 4 – Histogram of queries issued per user, for all 2,689,166 queries

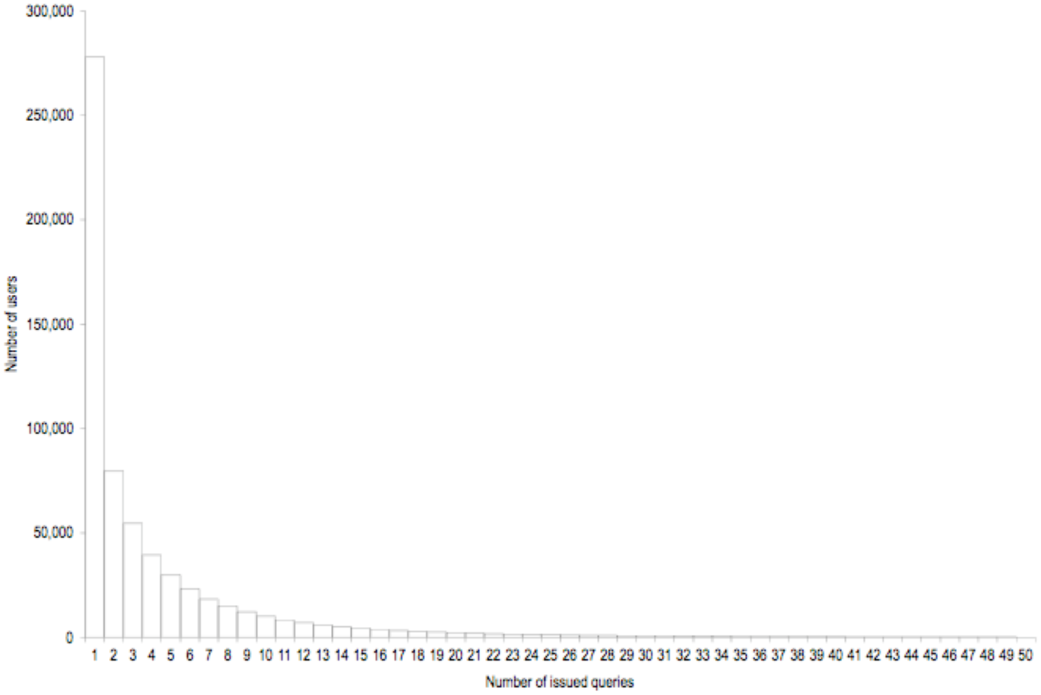


Figure 5 – Logarithm of the size of the result set for a sample of 2,272 queries

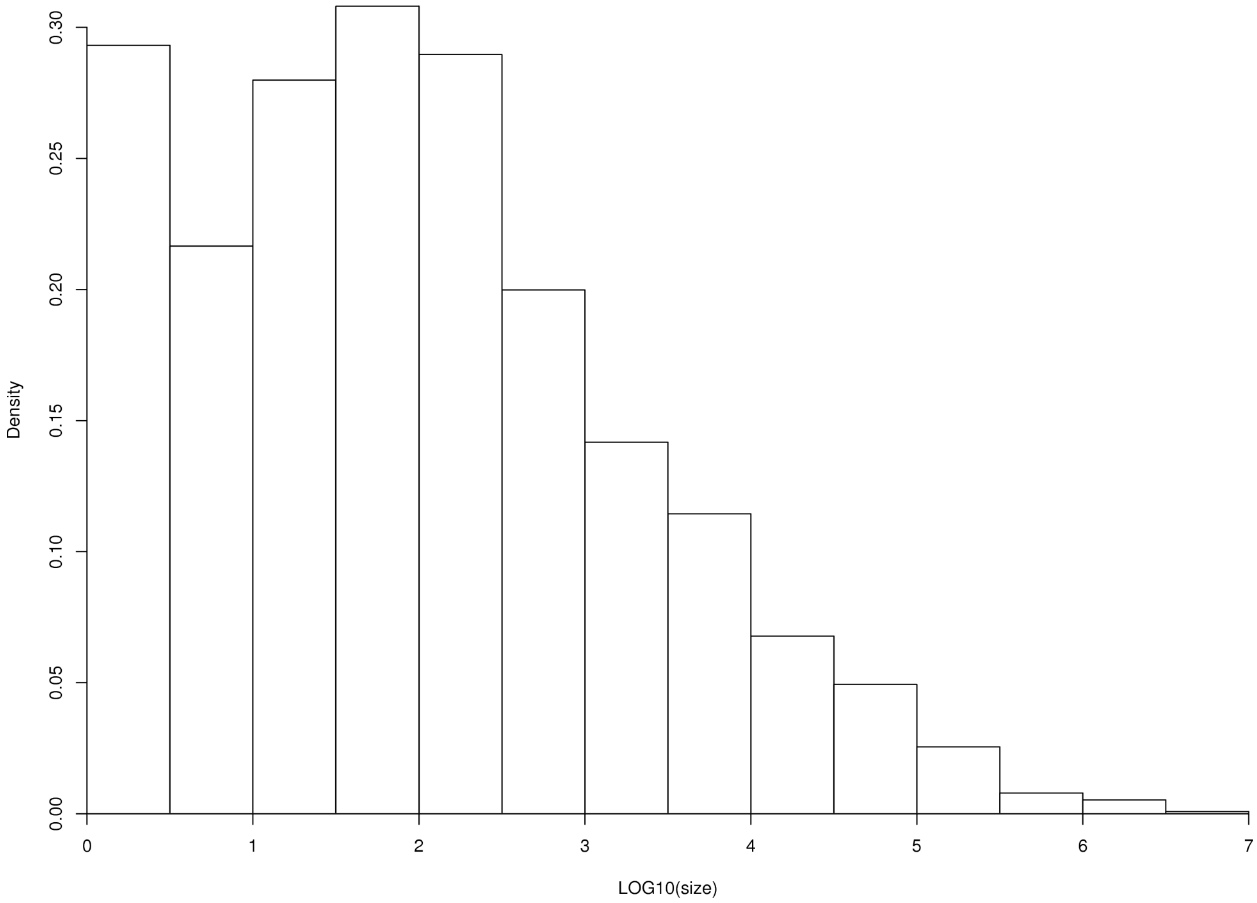


Figure 6 – Relative frequency of term counts for 2,689,166 PubMed queries issued during a single day (graph truncated at 20 terms)

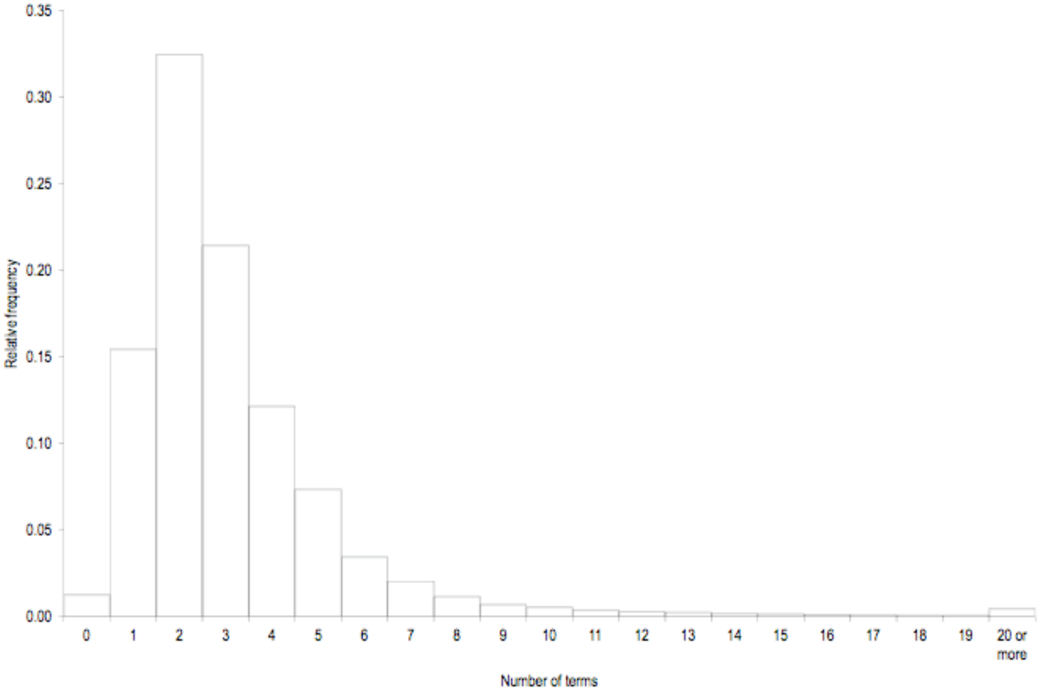


Figure 7 - Number of sessions per user for 2,689,166 queries issued on a single day.

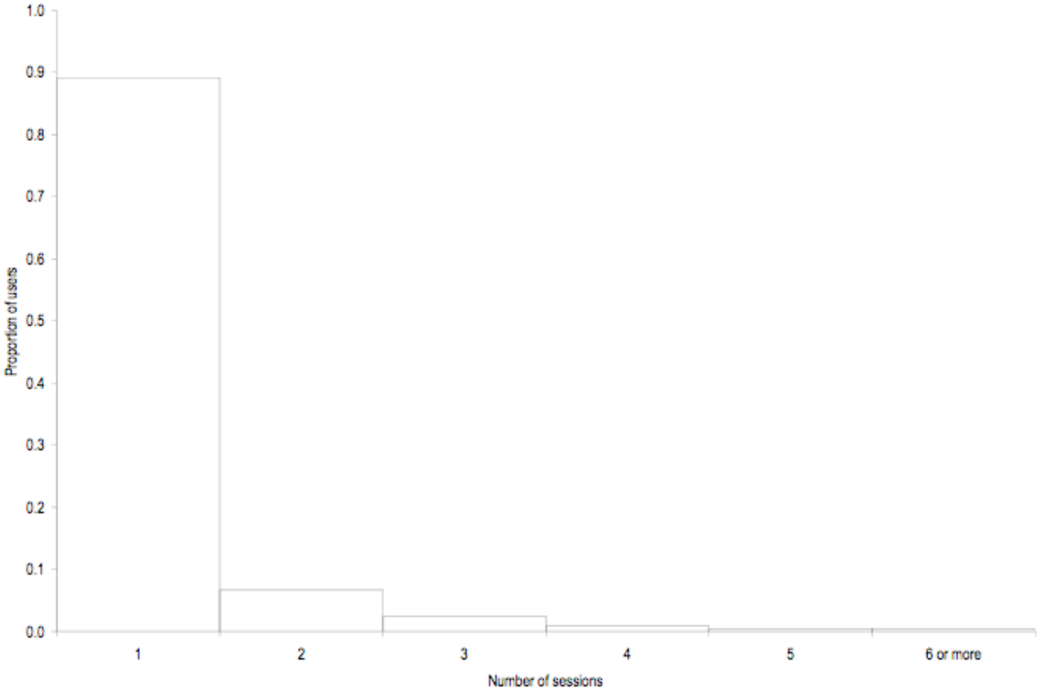
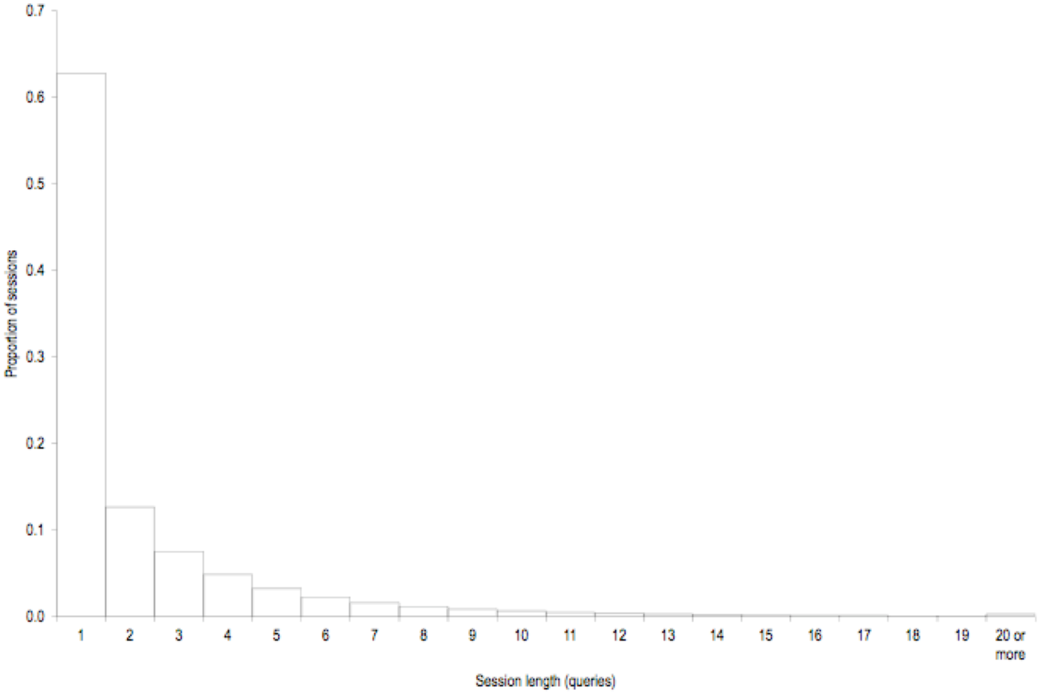


Figure 8 - Number of queries per session for 2,689,166 queries issued in a single day, as a proportion of sessions with the specified number of queries. Figure truncated at 20 queries.



References

1. United States National Library of Medicine. Resource statistics. [Web page] c2005 [cited 2006 January 17]; Available from: http://www.ncbi.nih.gov/About/tools/restable_stat_pubmeddata.htm

2. Silverstein C, Henzinger M, Marais H, Moricz M. Analysis of a very large AltaVista query log. Technical Note: Digital Equipment Corporation; 1998 October 26. Report No.: SRC Technical Note 1998-014.

3. Spink A, Wolfram D, Jansen B, Saracevic T. Searching the web: the public and their queries. J Am Soc Inf Sci Technol. 2001;52(3):226-34.

4. Jansen BJ, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the web. Inf Process Manage. 2000;36(2):207-27.

5. Eiron N, McCurley KS. Untangling compound documents on the web. In: Ashman H, editor. Conference on Hypertext and Hypermedia; 2003 August 26-30; Nottingham, England: ACM Press; 2003. p. 85-94.

6. Manfred L, Jeroen A. Beyond information searching and browsing: acquiring knowledge from digital libraries. The Netherlands: Tilburg University; 2001 February. Report No.: ITRS-008.

7. Moreno P, Van Thong J, Logan B, Fidler B, Maffey K. SpeechBot: a content-based search index for multimedia on the web. First IEEE Pacific-Rim Conference on Multimedia; 2000 December 13-15; Sydney, Australia; 2000.

8. Luo Q, Naughton JF. Form-based proxy caching for database-backed web sites. In: Apers PMG, Atzeni P, Ceri S, Paraboschi S, Ramamohanarao K, Snodgrass RT, editors. Proceedings of the 27th International Conference on Very Large Data Bases; 2001 September 11-14: Morgan Kaufmann Publishers Inc.; 2001. p. 191-200.

9. Spink A, Jansen B, Wolfram D, Saracevic T. From e-sex to e-commerce: web search changes. Computer. 2002;35(3):107-9.

10. Chau M, Fang X, Sheng ORL. Analysis of the query logs of a web site search engine. J Am Soc Inf Sci Technol. 2005;56(13):1363-76.

11. Lacroix E-M, Mehnert R. The US National Library of Medicine in the 21st century: expanding collections, nontraditional formats, new audiences. Health Info Libr J. 2002;19(3):126-32.

12. Hersh WR, Crabtree MK, Hickam DH, Sacherek L, Friedman CP, Tidmarsh P, et al. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. J Am Med Inform Assoc. 2002;9(3):283-93.

13. Bernstam EV, Kamvar SD, Meric F, Dugan JM, Chizek SC, Stave C, et al. Oncology patient interface to MEDLINE. Proc Am Soc Clin Oncol. 2001;20:244a (Abstract 974).

14. Haynes RB, McKibbon KA, Walker CJ, Ryan N, Fitzgerald D, Ramsden MF. Online access to MEDLINE in clinical settings. A study of use and usefulness. Ann Intern Med. 1990 Jan 1;112(1):78-84.

15. Broder A. A taxonomy of web search. SIGIR Forum. 2002;36(2):3-10.

16. United States National Library of Medicine. README file. [FTP] 2005 [updated 2005 October 17; cited 2005 October 17]; Available from: ftp://ftp.ncbi.nlm.nih.gov/toolbox/pubmed/query-logs/README

17. Budanitsky A, Hirst G. Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. In: Knight K, editor. Language Technologies 2001: The Second Meeting of the North American Chapter of the Association for Computational Linguistics; 2001 June 2-7; Pittsburgh, PA: Information Sciences Institute; 2001.
18. United States National Library of Medicine. Searching PubMed. PubMed Help [Web page] c2004 [updated 2006 August 8; cited 2006 September 25]; Available from: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.section.pubmedhelp.Searching_PubMed#pubmedhelp.Combining_search_ter
19. Joachims T. Evaluating retrieval performance using clickthrough data. Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval; 2002 August 12-15; Tampere, Finland; 2002.
20. Bernstam EV, Herskovic JR, Aphinyanaphongs Y, Aliferis CF, Sriram MG, Hersh WR. Using citation data to improve retrieval from MEDLINE. J Am Med Inform Assoc. 2006 Jan-Feb;13(1):96-105.