Analysis of Information Needs of Users of MEDLINEplus, 2002 – 2003

Alicia Scott-Wright[1,3], Jon Crowell[1], Qing Zeng[1], David W. Bates [2,3], Robert Greenes[1,3]

[1]Decision Systems Group, and, [2]Dept. of Medicine, Brigham & Women's Hospital,
Harvard Medical School, Boston, MA
[3]Dept. of Health Policy and Management, Harvard School of Public Health, Boston, MA

ABSTRACT

We analyzed query logs from use of MEDLINEplus to answer the questions: Are consumers' health information needs stable over time? and To what extent do users' queries change over time? To determine log stability, we assessed an Overlap Rate (OR) defined as the number of unique queries common to two adjacent months divided by the total number of unique queries in those months. All exactly matching queries were considered as one unique query. We measured ORs for the top 10 and 100 unique queries of a month and compared these to ORs for the following month.

Over ten months, users submitted 12,234,737 queries; only 2,179,571 (17.8%) were unique and these had a mean word count of 2.73 (S.D., 0.24); 121 of 137 (88.3%) unique queries each comprised of exactly matching search term(s) used at least 5000 times were of only one word. We could predict with 95% confidence that the monthly OR for the top 100 unique queries would lie between 67% - 87% when compared with the top 100 from the previous month. The mean month-to-month OR for top 10 queries was 62% (S.D., 20%) indicating significant variability; the lowest OR of 33% between the top 10 in Mar. compared to Apr. was likely due to "new" interest in information about SARS pneumonia in Apr. 2003.

Consumers' health information needs are relatively stable and the 100 most common unique queries are about 77% the same from month to month. Website sponsors should provide a broad range of information about a relatively stable number of topics. Analyses of log similarity may identify media-induced, cyclical, or seasonal changes in areas of consumer interest.

Keywords: MEDLINEplus, Consumer Informatics

INTRODUCTION

Surveys of Internet usage show that health information is one of the most cited reasons for searching the internet. Results of data mining commercial search engine query logs indicate that in general web users type in short queries, mostly look at the first 10 results only, seldom modify their initial query, and seldom use Boolean operators to enhance the specificity of their search terms.[1] [2] [3] That consumers may have problems forming effective queries has also been established by the work of McCray et al [4] of the National Library of Medicine. Zeng et al [5] [6] interviewed users to determine what information they were seeking and then analyzed MEDLINEplus query log files to find that consumers submit queries that often result in failure to find the information they desire.

We examined query logs, from use of MEDLINEplus, a consumer health information website. Our objective was to answer the questions: Are consumers' health information needs stable over time? and To what extent do users' queries change over time? Analyzing health information query logs files may identify patterns of consumer usage and, thus, assist website developers to link consumers to the information they seek.

METHODS

MEDLINEplus Web Access Query Log Files

In October 1998, the National Library of Medicine (NLM) announced a new resource, MEDLINEplus intended for consumers of health information.[7] The original files we received from NLM consisted of daily web access logs for all English queries submitted by consumers to (www.MEDLINEplus.org) between 11/21/2003 and 9/30/2003. We developed a JAVA program to parse the complex query log structure into fields including: the actual query search terms, and a month, day, year, hour, minute, and seconds timestamp of a submitted query. Files were imported into SAS v. 9.1[8] for all analyses reported in this paper. To identify patterns in usage of consumer search terms over time, we aggregated log data into daily intervals to present it in time order. When plotted, time series formats can be visually assessed for patterns of periodicity, seasonality, and outliers.

Unique Query Classifications

All exactly matching query search terms were classified as a unique query. Unique queries comprised of 5000 or more instances of the same exact search terms were qualitatively examined and categorized into eight health topics. The goal of these categorizations was not to create an ontology for consumer search terms but to identify themes that reflect how consumers might specify their areas of information need. These categorizations did not constrain unique queries to only one topic. For example, drug names were listed under the health topic, drugs, but also under a topic relevant to the clinical indication for that drug. Unique queries were classified as names of: body parts, chronic health problems, drugs, general interest or miscellaneous topics not easily included in other topics, infections, mental health or neurological disorders, signs and symptoms; and disorders, and states, or procedures related to women's health.

Unique Query Overlap Rate

To determine if consumers' health information needs were stable over time, and to what extent unique queries changed over time, we calculated an Overlap Rate, which has been previously described and used in the analyses of commercial query log files, to measure the similarity between or among query logs for different time periods. [9] [10] Given a series of query log files L_i (i= 1 to n) where L_i represents a log file of unique queries over a specified time period, the Overlap Rate is defined as:

$$OR = (L_1 \cap L_2 \cap L_3 \cap \ldots L_n) / (L_1 \cup L_2 \cup L_3 \cup \ldots L_n)$$

For example, the Overlap Rate for the top 10 unique queries for the months of January and February would be the number of common unique queries in the top 10 from January and February divided by the total number of top 10 unique queries from January or February.

RESULTS

Over the 315 day study period a total query volume of 12,234,737 English queries was submitted to MEDLINEplus. Figure 1 presents the aggregated web access log data in daily intervals in time order. As visually depicted and confirmed by counts, queries were submitted approximately according to a weekly cycle with about 16-17% of the weekly query volume submitted on each of the four days: Monday, Tuesday, Wednesday, and Thursday, about 13 - 14% on Friday, 9% on Saturday, and 10% on Sunday. Figure 1 shows a query volume drop approximately coincident with the year end holiday season celebrated in the United States and many English speaking countries. Analyses of actual query volumes (where all instances of exactly matching queries each contribute one count to the total volume) confirm this: during the week preceding the Christmas holidays, 12/15/02 - 12/21/02, the query volume was 219,489 hits/week; during Christmas week, 12/22/02 - 12/28/02, the volume was 141,453 hits/week, and for the two weeks following Christmas week query volumes were 175,948 and 277,152 hits/week respectively.

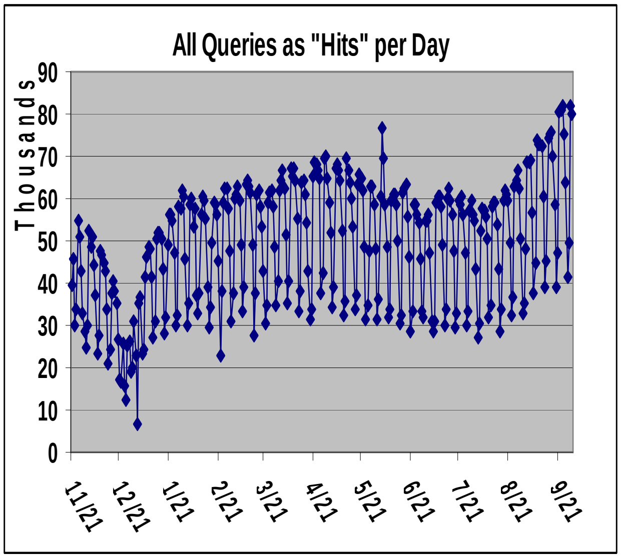

Figure 1: All Queries as No. "Hits" per Day

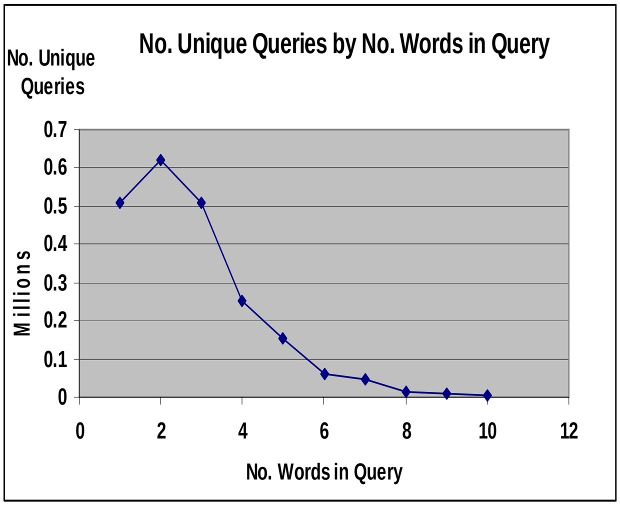

Figure 2: Unique Queries by Number of Words in the Query

Analysis of Unique Query Classifications

Of the total query volume of 12,234,737, only 2,179,571 or 17.8% were unique or distinct queries. Mean number of words used in unique queries was 2.73 S.D., 0.24 see Figure 2. Only 137 unique queries or 0.006% of all unique queries were comprised of 5000 or more instances of exactly matching search terms. When each exactly matching instance was counted as one query, the query volume represented by these 137 unique queries accounted for 1,414,970 of 12,234,737 queries or 11.6% of the total query volume or of the total usage of the service; 121 of 137 or 88.3% of these frequently used queries were of only one word. We categorized the 137 most commonly used queries into eight topics. The eight topics, examples of the most frequent unique queries submitted in that health topic along with the query volume generated (n) by each example are, in alphabetical order names of: **body parts** (gallbladder n = 21,257, thyroid n = 18,821, heart n = 15,044), **chronic health problems** (diabetes n = 45,035, lupus n = 27,025), **drugs** (Lexapro n = 13,129, Zoloft n = 9,114), **general interest or miscellaneous** (cholesterol n = 12,116 triglycerides n = 5,998), **infections** (cellulitis n = 22,702, shingles n = 18,728, herpes n = 15,631), **mental health and neurology** (depression n = 17,736, stress n = 9,328, Zoloft n = 9,114), **signs and symptoms** (hives n = 21,461, edema n = 18,094, cough n = 8,705), and **women's health** (pregnancy n = 23,173, hysterectomy n = 18,253).

In Table 1 the total query volume generated by our categorizations, 1,762,851, exceeds the query volume generated by the 137 unique queries, 1,414,970, because our query categorizations were not constrained to one health theme or topic.

Table 1: Query Volume Generated by 137 Unique Queries Used 5000 or More Times

Topic	Query Volume Generated
Chronic Problems	710,804
Infections	251,744
Symptoms	216,269
Mental Health and Neurology	183,215
Body Part	131,929
Drugs	121,075
Women's Health	86,685
General Interest	61,130

Figure 3 shows the number of unique queries (x-axis) and the cumulative percentage of usage (y-axis).

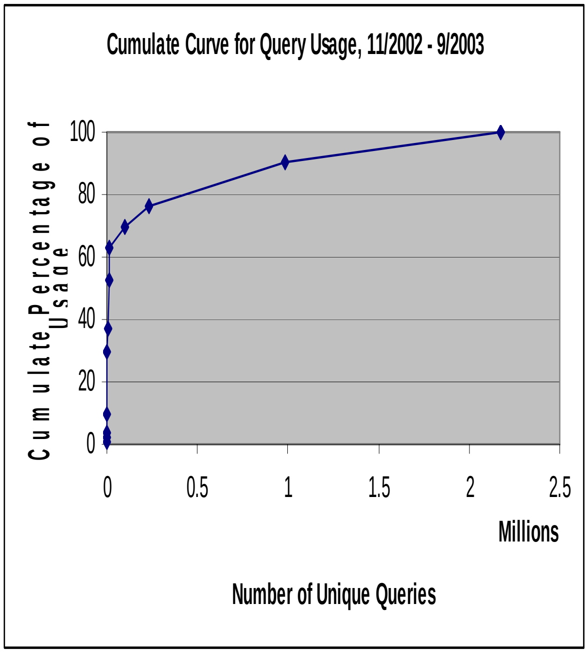

Figure 3: Cumulate Curve for Query Usage, 11/21/2002 - 9/26/2003

The curve is skewed to the left indicating that a relatively small number of unique queries contributed significantly to the total usage of the service. For example, over the study period, 100,910 unique queries or 4.6% of all unique queries accounted for about 70% of the total usage of the service.

Unique Query Overlap Rate

Table 2 shows the month-to-month overlap rates for the top 10 and top 100 unique queries of the two months compared. Also shown is the OR for the first 5 months vs. the last 5 months of our sample.

The top 100 queries were on average 77% the same month-to-month with a 95% Confidence Interval (CI) [67% - 87%] which shows statistical significance. Top 10 queries were on average 62% the same month-to-month with a 95% CI [23% - 101%] which includes unity and thus is not statistically significant. The "Overlap Rate Top 10" of 33% as measured for the time periods 11/21 - 12/31/2002 vs. Jan. 2003 and Mar. 2003 vs. Apr. 2003 represents a mismatch of five unique queries between the time periods compared (see Table 3).

Table 2: Month-to-Month and First 5-month-to-Last 5-month Overlap Rates with Standard Deviations (S.D.) for top 10 and top 100 queries

Month-to-Month (m-t-m)	**Overlap Rate Top 10**	**Overlap Rate Top 100**
Nov. 21, 2002 - Dec. 2002 vs. Jan. 2003	0.33	0.73
Jan. 2003 vs. Feb. 2003	0.67	0.79
Feb. 2003 vs. Mar. 2003	0.67	0.81
Mar. 2003 vs. Apr. 2003	0.33	0.69
Apr. 2003 vs. May 2003	0.67	0.73
May 2003 vs. Jun. 2003	0.67	0.82
Jun 2003 vs. Jul. 2003	0.67	0.84
Jul. 2003 vs. Aug. 2003	1.00	0.79
Aug. 03 vs. Sep. 26, 03	0.67	0.76
Mean m-t-m (std dev)	0.62 (0.20)	0.77 (0.05)
First Five Months-to-Last Five Months Nov 21, 02 to Apr. 03 vs. May 03 - Sep. 26, 03	0.67	0.67

Table 3: Top 10 Unique Queries for Months with Low Overlap Rates. Query Search Terms Arranged in Descending Order of Query Volumes.

Nov/Dec. 2002 vs. Jan. 2003		**Mar. 2003 vs. Apr. 2003**	
Diabetes	Diabetes	Diabetes	Diabetes
Cancer	Lupus	Lupus	Sars
Lupus	Hives	Low Blood Pressure	Pneumonia
Hypertension	Low Blood Pressure	Hives	Low Blood Pressure
Hives	Lexapro	Asthma	Pregnancy
Shingles	Pregnancy	Anemia	Lupus
Asthma	Asthma	Hysterectomy	Heart
Depression	Fibromyalgia	Cellulitis	Cancer
Anemia	Cellulitis	Thyroid	Gallbladder
Lexapro	Hysterectomy	Shingles	Shingles

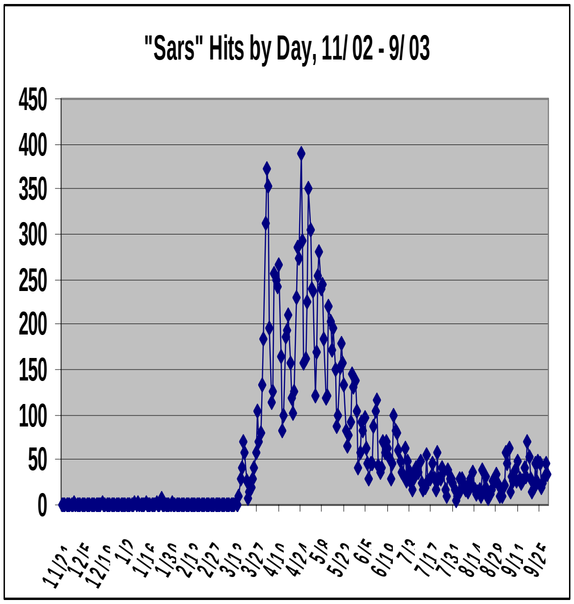

Figure 4 : "SARS" Hits by Day, 11/02 - 9/03

The low overlap rate top 10 between Nov./Dec. and Jan. was a result of "common" terms replacing "common" terms; no new search terms were introduced. The introduction to the top 10 in April 2003 of the never used before one-word search term, SARS, along with the one-word search term pneumonia contributed to the 33% overlap between Mar. and Apr., 2003. The term pneumonia had not previously appeared as a top 10 search term and during the 315 day study period only appeared as a top 10 search term during April, 2003. Figure 3 shows a time series graph for the number of SARS hits by day.

Diabetes was the most common search term in each month and overall. Only the search terms diabetes and lupus were among the top 10 search terms used in every month in the study period.

DISCUSSION

We analyzed MEDLINEplus query log data and found that only about 1 in 5 queries were unique or distinct. Consumer queries tended to be short (mean of 2.73 words), and the most commonly used search terms were of only one word.

Based on these findings, we can answer the questions posed at the beginning of the paper: Are consumer health information needs stable over time? and To what extent do users' queries change over time? For the top 100 user search terms, queries were relatively stable over time. For any month we could predict with 95% confidence that queries would be the same as the previous month's queries between 67% and 87% of the time.

However, the monthly top 10 user queries demonstrated significant variability. The effects of media coverage on the "new" infection coined SARS, Severe Acute Respiratory Syndrome, (which affected people principally in parts of China and Canada and reached epidemic proportions) were dramatically demonstrated in the time series graph of SARS hits per day. This "spike" in consumer interest in SARS and pneumonia was discernible from a low overlap rate between March and April, 2003 and analysis of the mismatch in search terms between these months.

This study has limitations. We analyzed submitted queries. We could not examine the actual questions users had that led them to submit a query. Also, because of the unsophisticated one word search strategy used by some consumers, a wide range of information might have been sought which could not be discerned from just that one term; for example, a user inputting the search term diabetes may have in mind a myriad number of questions.

In conclusion, this study is a descriptive analysis of MEDLINEplus query log data over a 315-day period. If consumers are to find the information they seek, sponsors of consumer health information websites should provide a broad range of information about a relatively stable number of topics. Periodic or real-time analyses of the similarity between logs of adjacent time periods may identify mismatches in logs due to influence of consumer health information interests by media coverage. Seasonal or cyclical consumer health-related interests may also influence overlap rates.

Our future work will include analyzing and categorizing multi-word queries -- queries of two or more words. Although less frequent than one word queries, multi-word queries, specifically those containing the most popular one word queries (diabetes, lupus, etc.) may yield more specific clues about the content of information consumers sought. For example, the query diabetes and depression is a combination of two very frequent one word search terms and is markedly more specific than either term. While it may be likely that users who input the more sophisticated multi-word queries have different information needs than users inputting one word queries, examination of multi-word queries containing the more common one word queries, using overlap rates and time series analyses of "hits" may allow website sponsors to better index the broad knowledge bases now required in response to less specific one word consumer requests.

REFERENCES

[1] Spink A, Wolfram D, Jansen BJ, Saracevic T. Searching the web: The public and their queries. J A, Soc Inf Tec 2001;52 (3):226-234

[2] Silverstein C, Henzinger M, Marais H, Moricz M. Analysis of a very large web search engine query log. SIGIR Forum 1999;33(3):6-12.

[3] Eysenbach G, Kohler C. Health-related searches on the Internet. JAMA. 2004 Jun 23;291(24):2946.

[4] McCray AT, Tse T. Understanding search failures in consumer health information systems. AMIA Annu Symp Proc. 2003;:430-4.

[5] Zeng QT, Kogan S, Plovnick RM, Crowell J, Lacroix E-M, Greenes RA. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. International J of Medical informatics 2004;73: 45-55.

[6] Zeng QT, Kogan S, Ngo L, Greenes RA. Relationship among different subjective measurements of consumer health information retrieval performance. Medinfo 2004.

[7] Miller N, Lacroix E-M, Backus JE, MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service, Bull. Med. Libr. Assoc. 2000 88;1:11-17.

[8] SAS Institute, Cary, North Carolina

[9] Chowdhury A, Grossman D, Frieder O, McCabe C. Analyses of multiple evidence combinations for retrieval strategies. Proceedings of the 24th Int Congress on Information Retrieval, ACM-SIGR, 9/ 2001.

[10] Badue C, Baeza-Yates B, Ribeiro-Neto B, et al. Distributed query processing using partitioned inverted files. Proceedings of SPIRE 2001, IEEE CS Press, Laguna San Rafael, Chile, pp.10-20, 11/ 2001

Acknowledgements

This research was partially supported by Training Grant LM07092 and R01 Grant LM07222 both from the National Library of Medicine.

Corresponding Author

Alicia Scott-Wright
●aswright@rics.bwh.harvard.edu