MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Data Mining MEDLINEplus Queries, Circa "9-11"

Alicia O. Scott-Wright[a b], Qing T. Zeng[a]

[a] *Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School, Boston, MA*
[b] *Department of Health Policy and Management, Harvard School of Public Health, Boston, MA*

Abstract

We examined query logs circa "9-11" from use of MEDLINEplus, a consumer health information website. Our objectives were to determine to what extent consumers' health information interests were stable over time, and to identify and describe anthrax-related health information categories of presumed interest to users during critical time periods following the attack on the World Trade Center. Analyzing health information query logs may identify patterns of consumer usage that can assist health information website developers and sponsors to link consumers to the knowledge they seek and thus, in an epidemic crisis situation, facilitate adoption of self-protective behaviors.

Keywords:

consumer health information, MEDLINEplus, anthrax

Introduction

Surveys of Internet usage show that health information is one of the most cited reasons for searching the internet. That consumers may have problems forming effective queries which can often result in failure to find information sought has been established by the work of Zeng and associates.[1 2 3]

Twenty-two cases of anthrax were confirmed during the anthrax outbreak in the United States which followed the tragedy of September 11, 2001 (see Table 1). Unfortunately, the initial case of anthrax, confirmed on October 4, 2001, was erroneously identified as sporadic and due to a waterborne bacterial source. This false information from high ranking government officials may have undermined the credibility of subsequent reports coming from public health agencies. A consumer survey by Blendon et al[4] noted that during the anthrax outbreak most respondents trusted their own physician *most* for health information. However, information on health information websites was trusted *second-most highly* more so than information from public health agencies, and government officials or television.

How did consumers use health information found on the internet? In a survey by Kittler, Hobbs, Volk, Kreps, and Bates,[5] 58% of those searching online for anthrax-related information reported behavior changes. After searching for health information on the web, consumers physically handled mail differently, and 65% reported adopting self-protective behaviors, such as, washing hands more often.

We examined query logs, over the 4 month period 8/1/2001 – 11/30/2001, from use of MEDLINEplus, a consumer health information website. Our objectives were to determine to what extent consumers' health information interests are stable over time, and to identify and describe anthrax-related health information categories of presumed interest to users during critical time periods following the attack on the World Trade Center. Analyzing health information query logs may identify patterns of consumer usage that can assist health information website developers and sponsors to link consumers to the knowledge they seek and thus facilitate adoption of self-protective behaviors.

Materials and methods

MEDLINEplus query log analyses

In October 1998, the National Library of Medicine (NLM) announced a new resource, MedlinePlus, intended for consumers of health information.[6] When a user submits a query to MEDLINEplus, a program writes data about the query into the log file on the server. The original files we received from NLM consisted of log data for all queries submitted by consumers to the website (www.MedlinePlus.org) on the dates August 1, 2001 through 11/30/2001. Personal identifiers were not included in the log data we received.

We developed a JAVA program to parse the complex query log structure into fields which included: the actual query search terms, and a month, day, year, hour, minute, and seconds timestamp of that submitted query. Parsed files were imported into and analyzed using SAS v. 9.1 for all results reported in this paper.

To identify patterns in usage of consumer search terms over time, we aggregated web access 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 search queries

By definition, *One* unique or distinct query represents all queries with exactly matching search terms. For example, the exact search terms "diabetes mellitus" might have been

submitted by consumers numerous times and in differing frequencies per day over a selected time period, but would represent only one unique query.

Unique or distinct search terms were identified over the study period. To determine if consumer queries were stable over time, and to estimate to what extent consumer search queries changed over time, we calculated an *Overlap Rate,* which has been previously described and used in the analyses of query log files,[7 8 9] to measure the similarity between query logs on a month to month basis.

Given a series of query log files $L_1, L_2, ..., Ln,$ the *Overlap Rate* is defined as:

$$OR = (L_1 \cap L_2 \cap L_3 \ldots L_n) / (\cup L_1 \cup L_2 \cup L_3 \ldots L_n) \quad (1)$$

Another example, the Overlap Rate for the top 10 unique queries for the month of January when compared with the top 10 unique queries for the month of February would be the number of *common* unique queries in the top 10 from January and February divided by the total number of *all* unique queries in the top 10 from January or February.

Anthrax-related time series plots

Our web access log study period spans the time of the events occurring just before and just after terrorist attacks on the World Trade Center in New York City on September 11, 2001 and the anthrax epidemic that followed. We divided our time series data of all queries containing the word (or search term) anthrax into periods based on possible dates of significance. Please note, our time series graphs represent the actual number of times English search terms containing the word anthrax (anthracis) were used on each day of our study period. Thus, every anthrax-related query is counted as a hit in our time series plot as many times as submitted.

Period 1 begins 8/1/2001 and ends 9/10/2001. Period 2 begins 9/11 and extends through and includes 9/17/2001. Period 3 begins 9/18, the day the first two of four letters containing anthrax spores were postmarked and extends through 10/3/2001. Period 4 begins 10/4/2001. The day the first case of anthrax was confirmed and hit the media and extends through 11/30/2003. Other important dates: October 9, 2001, the date two additional letters with anthrax spores were postmarked, and November 16, 2001, the date the final case of anthrax was confirmed, fall within Period 4. Below is a Centers for Disease Control and Prevention CDC chronology of the outbreak.

Table 1 - Chronology of Anthrax Outbreak (Source Centers for Disease Control and Prevention CDC, Atlanta GA.)

Date(s)	Event(s)
Sept 11	Terrorist attack on World Trade Center, New York, New York
Sept 18	Postmark date of 2 envelopes mailed to news media companies
Sept 22-Oct 2	Symptom onset dates for nine cases of anthrax (2 inhalational)
Oct 4	Anthrax first confirmed and made public
Oct 9	Postmark date of 2 envelopes mailed to government leaders
Oct 14-Oct 26	Symptom onset dates for 13 cases of anthrax (7 inhalational)
Nov 15	Symptom onset date for one *final* case of anthrax (inhalational)

Anthrax multi-word query classifications

Anthrax-related multi-word queries contained the word anthrax or anthracis and at least one other word. These multi-word queries were qualitatively examined and classified into the topics: epidemiology, microbiology, symptoms (lung, skin, other), therapy, vaccine, veterinary, warfare, and other complex queries aided by the use of a keyword search program written to place queries into categories according to matches of words in a query with keywords we used to define a category (see Table 2). The goal of these categorizations was to identify themes or topics that might help us better describe areas of presumed consumer interest. These categorizations were mutually exclusive and exhaustive.

Table 2 - Anthrax-related multi-word query categories and keyword list of some of the words or word fragments (*) used to categorize queries

Category	Keywords
Veterinary	cattle, woolsort*, dog, cat
Vaccine	vaccin*, immun*
Sympt:skin	rash, lesion, cutaneo*
Warfare	bioter*, weapon, postal
Sympt:lung	inhal*, lung, chest
Therapy	cipro, floxacin, penicillin
Microbiology	enzyme, mice, bacteria
Epidemiology	pregnan*, water, outbreak
Sympt:other	intestinal, gastr*, mening*

Table 3 - Anthrax-related queries by search strategy (one-word vs. multi-word) and by period with category breakdowns of multiword queries

Query Type	8/1 - 11/30		P1 8/1 - 9/10		P2 9/11 - 9/17		P3 9/18 - 10/3		P4 10/4 - 11/30	
	Count	%	Count	%	Count	%	Count	%	Count	%
Anthx Queries	15,481	100.0	63	100.0	64	100.0	594	100.0	14,760	100.0
One-Word	11,138	71.9	48	76.2	56	87.5	485	81.6	10,549	71.4
Multi-Word	4,343	28.1	15	23.8	8	12.5	109	18.4	4,211	28.5
Sympt-lung	203	1.3	5	7.9	0	0	5	0.8	193	1.8
Sympt-skin	752	4.9	1	1.6	1	1.6	2	0.3	748	7.1
Sympt-other	390	2.5	0	0	0	0	5	0.8	385	3.6
Vaccine	203	1.3	2	3.2	4	6.3	24	4.0	173	1.6
Therapy	418	2.7	0	0	2	3.1	32	5.3	384	3.6
Warfare	220	1.4	0	0	1	1.6	10	1.9	209	2.0
Microbiology	334	2.2	1	1.6	0	0	10	1.9	323	3.1
Epidemiology	737	4.8	1	1.6	0	0	10	1.9	748	7.1
Veterinary	30	0.2	0	0	0	0	1	0.2	29	0.3
Other	1056	6.8	5	7.9	0	0	10	1.7	1019	9.6

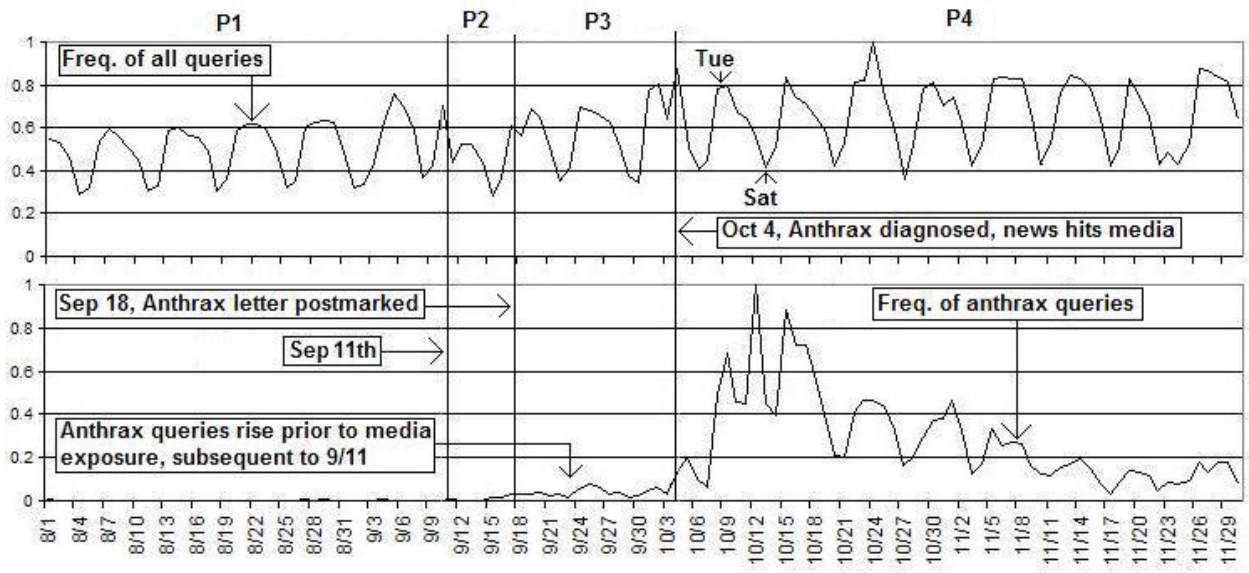

Figure 1 - Top - Normalized graph of all queries showing approximate weekly variation. Bottom - Normalized graph of anthrax-related queries over same time. P1: 8/1-9/10, P2: 9/11-9/17, P3: 9/18-10/3 P4: 10/4-11/30, 2001. Important dates: 9/18 and 10/9 - 2 letters containing anthrax mailed on each date. Nine people become symptomatic between 9/18 and 10/3; thirteen people become symptomatic after 10/4. First case diagnosed 10/4 and hits news the same day. Last anthrax case has symptom onset date 11/16 (CDC)

Results

Over the 121 day study period a *total query volume* of 2,179,571 English queries was submitted to MEDLINEplus; 1,012,542 (46.5%) of these query submissions were of only one word. [Range min: 8,853 queries submitted on 8/4/2001 max: 30,349 on 10/25/2001.] Figure 1 Top presents the normalized 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.

Unique search queries

All submitted queries (regardless of the number of times submitted) or the *total query volume* was 2,179,571; only 634,669 or 29.1% were unique or distinct queries. Mean number of words used in these distinct search terms was 2.23 S.D. 0.14.

Table 4 gives the list of unique queries used 2000 times or more over the 121 day study period and the actual number of times the query was used.

Table 4 - Most commonly used search terms and number of times used

Unique Search Term(s)	Number of Times Used
anthrax	10,958
diabetes	6,093
fibromyalgia	3,549
lupus	3,515
asthma	3,049
multiple sclerosis	2,879
breast cancer	2,516
depression	2,486
cancer	2,392
osteoporosis	2,376
shingles	2,335
hypertension	2,260
alzheimer	2,203
cholesterol	2,138
anemia	2,062
stroke	2,047
arthritis	2,039

Table 5 shows the month-to-month overlap rates for the top 10 and top 100 unique queries of the two adjacent months compared.

Top 10 queries were on average 54% the same month-to-month with a statistically significant 95% Confidence Interval (CI) [34% - 74%].

The top 100 queries were on average 69% the same month-to-month also with a statistically significant 95% Confidence Interval (CI) [53% - 85%].

Table 5 - Month to month overlap rates for top 10 and top 100 unique queries in a month

Months compared	Overlap Rate (OR) Top 10	Overlap Rate (OR) Top 100
Aug. 01 vs. Sep. 01	0.43	0.64
Sep. 01 vs. Oct. 01	0.67	0.70
Oct. 01 vs. Nov. 01	0.54	0.74
Mean OR (S.D.)	0.54 (0.10)	0.69 (0.04)

Anthrax-related time series plots

Over the 121 day study period, 15,481 queries were submitted containing the word anthrax or anthracis; 11,138 of these were just the one-word anthrax or anthracis. Multi-word anthrax queries totaled 4,343 separate (non-unique) query submissions. [Range min: 0 queries submitted on various, but not all, dates between 8/1 and 9/10/2001 max: 885 on 10/13/2001.] Figure 1 Bottom shows a normalized graph of anthrax queries coincident with all queries (including all anthrax queries) on the same timeline.

Anthrax-related unique query classification

Table shows the multi-word anthrax-related query classifications overall and over previously defined periods (see Table 3). Note again, categories are mutually exclusive and that 1056 or 24% of anthrax-related multi-word queries were complex queries on wide-ranging topics and not assignable to the categories selected. Virtually all unassigned queries were submitted during period 4.

Discussion

We analyzed query logs from use of a consumer health information website, MEDLINEplus, over the period August 1, 2001 through November 30, 2001. We found that (for all queries submitted over this time) only about 1 in 3 were unique or distinct. Also, consumers used simple, unsophisticated search strategies; that is, about one-half (46.5%) of consumer queries were of only one-word.

Over our short study period, we determined that consumer health information needs were generally stable and predictable from month to month. With 95% confidence we can say that the top 10 queries would be between 34% - 74% the same month to month and the top 100 queries would be the same between 53% - 85% month to month.

Consumers responded to the events of 9-11 with a presumed interest in information about anthrax using the simple one-word search term for about three quarters of their anthrax-related searches. However, only 127 anthrax-related queries were submitted between 8/1 and 9/18. Once the first case was diagnosed and made public on October 4, 2001, consumers responded with a deluge of "anthrax" queries.

However, between 9/19 and 10/4 before an epidemic was announced, anthrax queries increased to 594. Mounting interest in anthrax, a possible biological warfare agent, following 9-11 might explain this increase, but information on other biological agents such as smallpox was not sought frequently (see Table 4). The predominant interest in anthrax in period 3 seemed to be about therapeutic agents and vaccines or immunization. While in period 4 (after media announcement of first case), the predominant interest was (other) not easily assignable to our selected topics and about cutaneous anthrax or the epidemiology of anthrax.

This study has several limitations. We analyzed queries which are only surrogate measures of actual questions. We could not examine the actual questions users had that led them to submit a query. Also, because of the relatively unsophisticated one word search strategy used for most searches a wide range of information might have been sought which could not be discerned from just that one term.

In conclusion, this study is a descriptive analysis of MEDLINEplus query log data over a 121-day period circa 9-11. In general, consumers used unsophisticated search strategies to access information on a relatively stable body of topics. Sponsors of consumer health information web-sites may need to provide a broad range of information about a relatively stable number of themes to meet consumer needs.

Also, we conclude that the events of 9-11 sparked an interest in one potential biological warfare agent and this interest was significantly increased once an actual case of anthrax was identified. What else would account for the increase in interest in anthrax between 9/18 and 10/4?

References

[1] 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.

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

[3] 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.

[4] Blendon RJ, Benson JM, DesRoches CM, Herrmann MJ. Harvard School of Public Health, Robert Wood Johnson Foundation Survey Project on Americans' Response to Biological Terrorism. International Communications Research. October 24-28, 2001

[5] Kittler AF, Hobbs J, Volk LA, Kreps GL, Bates DW. The internet as a vehicle to communicate health information during a public health emergency: a survey analysis involving the anthrax scare of 2001. J Med Internet Res 2004;(9):E8.

[6] 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

[7] 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.

[8] 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

[9] Scott-Wright A, Crowell J, Zeng Q, Bates DW, Greenes RA. Analysis of information needs of users of MEDLINEplus, 2002 – 2003. Paper presented AMIA November, 2006 Washington, D.C..

Acknowledgements

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

Address for correspondence

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