A study of medical and health queries to web search engines

Amanda Spink*, Yin Yang†, Jim Jansen†, Pirrko Nykanen‡, Daniel P. Lorence§, Seda Ozmutlu¶ & H. Cenk Ozmutlu¶, *School of Information Sciences, University of Pittsburgh, Pittsburgh, †School of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA, ‡VTT, University of Tampere, Tampere, Finland, §Department of Health Policy and Administration, The Pennsylvania State University, University Park, PA, USA and ¶Department of Industrial Engineering, Uludag University, Bursa, Turkey

Abstract

This paper reports findings from an analysis of medical or health queries to different web search engines. We report results: (i) comparing samples of 10 000 web queries taken randomly from 1.2 million query logs from the AlltheWeb.com and Excite.com commercial web search engines in 2001 for medical or health queries, (ii) comparing the 2001 findings from Excite and AlltheWeb.com users with results from a previous analysis of medical and health related queries from the Excite Web search engine for 1997 and 1999, and (iii) medical or health advice-seeking queries beginning with the word 'should'. Findings suggest: (i) a small percentage of web queries are medical or health related, (ii) the top five categories of medical or health queries were: general health, weight issues, reproductive health and puberty, pregnancy/obstetrics, and human relationships, and (iii) over time, the medical and health queries may have declined as a proportion of all web queries, as the use of specialized medical/health websites and e-commerce-related queries has increased. Findings provide insights into medical and health-related web querying and suggests some implications for the use of the general web search engines when seeking medical/health information.

Introduction

Medical and health information is proliferating on the web. A recent study by the American Medical Association[1] reported an increase in the use of the Internet by physicians. Consumers also search the web for information related to medical and health issues. This information may become the basis for shared patient/provider communication and decision-making. In addition to seeking factual or general information on medical and health issues, people may use web search engines to seek advice on personal problems, where related queries may be very detailed and solicit specific medical or health information.

In this growing shared-decision-making medical environment, the quality of web-based information retrieved by the patient becomes of growing importance to the provider.[2] Many studies have examined aspects of consumer-health related web use, including websites and electronic lists.[3–5] Studies have evaluated the accuracy, quality and reliability of medical information on the web[6,7–9]

Correspondence: Amanda Spink, School of Information Sciences, University of Pittsburgh, 610 IS Building, 135 N. Bellefield Avenue, Pittsburgh, PA 15260, USA. E-mail: aspink@sis.pitt.edu

McCray *et al.*[10] studied users' queries on the National Library of Medicine website. Some 94% of terms submitted were medical terms covering a broad range of medical topics. Many medical terms were misspelled and short (less than four words). Many users also asked specific questions that went beyond the NLM website.

Medical website quality is being tested through the URAC accreditation process, 'health on the Net' logo program, and the HI Ethics project. As part of the European Union's eEurope: Health Online actions, Member State representatives and experts have agreed to establish a set of guidelines for quality criteria for health-related Websites. Such criteria are designed to increase user confidence in use of such sites and foster best practices in the development of sites.

Barnas and Kahn[11] surveyed users of the Medical College of Wisconsin HealthLink website. The study revealed a broad range of users' medical topic interests, including women's health, weight control, allergies, and back problems. In a related study, Eysenbach *et al.*[12] discussed MedCERTAIN, a project attempting to label the quality of Internet health information.

Medical and health information seeking constitutes an important use of the web. The Pew Internet Project[13,14] one of the largest national surveys accomplished to date, estimate that 62% of Internet users (some 73 million people living in the United States) search the web for health information. Results from the Pew study indicated that 93% of health information seekers surveyed looked for information about a specific illness or condition, 65% sought information on exercise, nutrition or weight control, 64% for prescription drugs, and 33% for sensitive health information. More than half the respondents reported using the web for health information every few months or less frequently.

Clearly, many Americans are using the web for medical and health information, often as a supplement to seeking help from medical providers. Recently, Pew[15] found that half of American adults surveyed have searched on-line for health information. Some 80% of adult Internet users (93 million people), have searched for at least one of 16 major health topics on-line. The Pew study[15] suggests that the act of looking for health or medical information is a popular on-line activity. Pew[15] did not compare the frequency of health searching with other topics.

This paper reports findings from an analysis of medical or health-related web queries. To date, few studies have reported results from an analysis of large-scale medical and health-related web queries submitted to commercial web search engines.

Related studies

Medical web searching

Many researchers have studied users' medical searching and information retrieval within complex abstracting systems, such as MEDLINE. Hersh and Hickman[16] and Hersh *et al.*[17] found that medical and nurse practitioner students were moderately successful at answering clinical questions correctly when searching MEDLINE. Eysenbach and Kohler[18] found that for 21 consumer web users, searching for health information is suboptimal, but users nevertheless found health information successfully. Atlas[19] found that first-year medical students reported that general web search engines produced better results than meta-medical websites and medicine-specific search engines. Many researchers have called for a new generation of web-based medical retrieval tools.[20–22]

Recent studies by Spink *et al.*[23] show that users general web search engine sessions are usually short and contain few search terms or queries. Medical and health-related topics generated approximately 9.5% of Excite Web searches in 1997, 7.8% in 1999 and 7.5% in 2001. Spink and Ozmutlu[24] found that 11.5% of Ask Jeeves question queries were medical or health related.

Studies on health information seeking on the web also indicate that lay terminology is only partly successful at locating useful health information and search effectiveness[20] where web searching often yields misleading or unrelated information for the lay health consumer. Berland *et al.*[25] likewise concluded that accessing health information using search engines and simple search terms was not efficient, as high reading levels are required to comprehend web-based health information. Patients further seek medical advice from physicians via e-mail[26,27] and from the web.[28,29]

Macleod[8] summarized the limitations of the web as an electronic consumer resource:

'while immediate access to such information has been of great benefit to health-care professionals and patients, there is growing concern that a substantial proportion of clinical information on the web might be inaccurate, erroneous, misleading, or fraudulent, and thereby pose a threat to public health …'

Often, then, the web-enabled health information seeker must know, within the realm of language, the near-specific location of the knowledge they seek. When an exact clinical term is not known, most laypersons will resort to their only available resource, the popular/lay terminology for the concept, illness, or subject of interest. While the capacity of the patient to gather and collect such information remains a matter for the clinician to assess, it is undeniable that Internet-derived information can serve as a powerful catalyst for seeking health services.[30,31]

In this paper we report findings from a large-scale study of medical and health-related web querying on different commercial web search engines. These results are also compared with previous findings from Spink *et al*.[23] who examined medical and health query trends from 1997 to 2001.

Research questions

The objective of this study was to examine the characteristics of medical and health information queries submitted to the Excite, AlltheWeb.com and Ask Jeeves Web search engines, including:

1 What are the characteristics of medical or health-related web searching?
2 What medical or health-related topics are searched for on web search engines?
3 What is the nature of medical or health advice-seeking web queries on the Ask Jeeves Web search engine?

Research design

Data collection

Excite.com and AlltheWeb.com are major Internet media companies offering web searching and web search engines. We analysed data sets from Excite taken from May 2001 and for AlltheWeb.com submitted on February 6, 2001. The entries are given in the order they arrived at the web search engine. New sessions/users are identified through a user ID and each query is given time stamps in hours, minutes and seconds. The majority of Excite users are from the United States and AlltheWeb.com users are believed by AlltheWeb.com to be from Europe, mostly from Germany.

Each web query log records contained three fields:

- *Identification*: anonymous code assigned by web company server to a user machine.
- *Time of day*: in hours, minutes, and seconds.
- *Query*: user terms as entered.

The data used in this study was obtained from three web search engines—Alltheweb.com, Excite and Ask Jeeves. At the time of this study, investigators were not able to secure data from other web search engines, e.g. Google. There was also no opportunity to sample web queries over time, such as across days or weeks. The authors had no control over the quantity or structure of the data provided by the web search engines.

Web query analysis

Excite query analysis. We quantitatively analysed 10 000 queries randomly sampled from a log of Excite 1.2 million queries from the 2001 data set, for medical or health-related queries.

AlltheWeb.com query analysis. We quantitatively analysed 10 257 queries randomly sampled from a data log of 1.2 million AlltheWeb.com web search engine queries for medical and health-related queries.

Ask Jeeves query analysis. We also extracted all queries containing the advice-seeking word 'should' from an Ask Jeeves web query-log of 800 000 queries from 20 December 1999 that contained an anonymous user id, time of the day, and the query terms entered by the users. The word 'should' was selected as a word often found in advice-seeking questions. We qualitatively analysed the sample of 1792 Ask Jeeves web queries for types of medical and health-related

queries. Some 332 (24%) 'should' queries were medical or health related—e.g. 'What should I do about haemorrhoids' These queries were further classified into taxonomy of the medical and health advice sought.

Medical or health query classification. Web queries were qualitatively examined for medical and health content by two researchers. The purpose of the analysis was to classify queries as either medical or health-related query. To judge the query intent as medical or health related, strong evidence of such intent had to be present in the query log. If queries did not include explicitly medical or health-related terms, they were not classified.

Inter-coder agreement is defined as the similarity in which each coder in the study decided whether a query was medically or health related. To check coding consistency, each researcher re-coded 10 000 queries previously classified by the other researcher. After exchanging and re-coding the searches, the researchers met again in order to make a final decision about the classifications. The two researchers discussed each disputed query until a classification consensus for that query was reached.

Results

The results extend findings reported in Spink *et al*.[32]

Medical/health queries

Table 1 compares results from previous web query studies with results from the analysis of the 2001 Excite and Alltheweb.com data sets for medical or health queries.

Medical or health issues were a declining proportion of the queries submitted to the Excite web search engines. The decline has progressed from a high of 9.5% in 1997 to 7.5% in 2001. The proportion of medical or health-related queries submitted to the European AlltheWeb.com web search engine was also lower at 3.2%. For both the Excite and AlltheWeb.com data sets, the mean terms per query was 2.3, the mean queries per user session was 2.2, and the mean pages of 10 websites viewed per user session was 1.6. We had no click-through data for these sessions. These figures are fairly equivalent to statistics for general web searching.[23] In other words, medical or health-related web searches were equivalent in length, complexity and lack of query reformulation to non-medical or health searches.

Medical advice-seeking

Table 2 shows the percentage of medical or health-related queries in the 2001 Excite searches and Ask Jeeves advice-seeking question-answering data.

Some 332 (24%) of the 1792 advice-seeking question-answering Ask Jeeves queries were medical or health related.

General medical/health. Many queries were related to general medical or health issues. A query was considered to be searching for General Medical/Health information if it merely named a medical term, such as a disease, a pharmaceutical drug, or a symptom or general health issue. Example of such terms would be 'AIDS', 'dizziness', and 'flu'.

Table 1 Trends in prevalence of medical and health-related web queries.

1997 Excite queries (2414 queries)*	1999 Excite web queries (2539 queries)*	2001 Excite web queries (10 000 queries)*	2001 AlltheWeb.com Web queries (10 257 queries)
9.5% medical or health queries	7.8% medical or health queries	7.5% medical or health	3.2% medical or health
Mean of 2.3 terms per medical or health query	Mean of 2.3 terms per medical or health query	Mean of 2.3 terms per medical or health query	Mean of 2.3 terms per medical or health query
Mean of 2.2 queries per medical or health session	Mean of 2.2 queries per medical or health session	Mean of 2.2 queries per medical or health session	Mean of 2.2 queries per medical or health session

*Spink *et al.* (2002).

Table 2 Excite and Ask Jeeves medical and health advice-seeking topic categories.

	% of 2001 Excite medical or health-related queries (4.5% of 2500 queries = 112)		% of Ask Jeeves advice-seeking medical or health-related queries (24% of 1792 queries = 332)	
Topic categories	Number	%	Number	%
General medical/health	23	21.3%	118	35.6%
Human relationships	40	35.2%	70	21.3%
Weight	17	15.3%	69	20.8%
Reproductive health/puberty	21	18%	45	13.8%
Pregnancy/baby	11	10.2%	28	8.5%
Total	112	100%	332	100%

Human relationships. A query was regarded as falling into the Human Relationship category if it was seeking information about psychological health issues (e.g., dating, marriage, psychological disorders, known psychologists, or psychological tests). Examples of these queries would be 'How I do I get a mental health doctor?' This type of question required a complex and more psychologically based response.

Weight. Many queries in this category requested information on the appropriate weight for a given height or age. For example, 'What should be my weight if my height is 4′10″? Many users just asked—'How much should I weigh?'

Reproductive health. Reproductive health queries often related to sexual issues and infertility or sexual health issues. For example, different users asked, 'How often should men ejaculate?' or 'Is my vagina normal?'

Pregnancy/baby. Many queries related to health issues during pregnancy or the health of a baby. For example, 'When should I take a pregnancy test' and 'Should pregnant women fly in an airplane?' Table 3 provides the number of advice-seeking queries using various starting terms including the term 'should'.

Most people who used the Ask Jeeves question-answering web search engine asked for medical or health advice within a limited question structure. Of interest, most users *seeking advice* about medical or health asked 'What should' or 'How should' questions, rather then 'Where should I' questions. In general, most non-medical or health web users ask more location-related questions, such as 'Where can I find' questions or 'What is' questions.

Table 3 Ask Jeeves advice-seeking medical or health-related query—starting terms (total = 332 queries).

Should query starting terms	Number of queries	%
What should I	82	24.7%
How should	48	14.4%
What should you	46	13.8%
Should	35	10.6%
Should I	30	9.2%
Should we	27	7.8%
Where should	22	6.8%
How should	16	4.8%
Why should	12	3.7%
When should	10	3%
Which should	4	1.2%
Total	332	100%

Personified and opinion queries. Medical or health-related queries were also often advice seeking and personalized, e.g. 'Hey Jeeves, what should I take for the flu?' Medical/health users also sought more human-like answers. A few users actually addressed Ask Jeeves as a human being by saying: 'Hey Jeeves' or 'Jeeves …' or asking for opinion. Some users requested help by saying: 'Help me Jeeves'. Some users were polite and phrased their query as 'May I …' or 'Please …' However, most users were not that polite, particularly those who entered a request as opposed to a question format query, discussed later in this paper.

While not readily seen in individual search observations, it becomes apparent from a large log of Ask Jeeves question queries that many medical and health information seeking searchers don't clearly understand the web search process. They ascribe human abilities to Ask Jeeves that go way beyond its current capabilities. Many users do not understand how a web search engine works in conjunction with their own knowledge base or how it affects information-seeking and searching processes. Users are often frustrated and emotional during their web search engine interactions.

Discussion

Findings from our analysis suggest that a small proportion of searching on commercial web search engines, in the United States and Europe, are medical or health related. Despite the large percentage of Internet users who conduct medical or health searchers reported by The Pew Internet Project[13,14] survey, examination of large-scale commercial web query logs suggests that medical or health web searching may be a small proportion of web searches. As stated previously, Pew[15] does not compare health searching with other topics.

Our findings also suggest an on-going shift in web users' search topics. From 1997 to 2001, queries related to entertainment or recreation and medical or health have declined proportionally, as the queries related to e-commerce, travel, employment or economy and people, places or things have increased.[23] This proportional shift may also reflect a shift by web users to more specialized medical or health websites, such as WebMD, and a shift towards more e-commerce websites and web searching.

Medical or health web queries and sessions are short and equivalent in length to non-medical or health querying. Users also do not reformulate their medical or health searches to a great extent. Thus, they express medical and health issues succinctly. Few people create long medical or health queries that include synonyms or alternate terms. Few users look beyond the first or second page of websites retrieved in response to their queries. Users' may be finding the information they seek in the first 10–20 websites.

The analysis of advice-seeking and personified web queries suggests that when seeking medical and health information, most consumers fail to understand the limitations of the web search process. Many also ascribe human and advice-seeking abilities to web search engines, such as Ask Jeeves, that go beyond the system's current capabilities. Many users do not really understand how a web search engine works in conjunction with their own information-seeking and searching processes. Users are often frustrated and emotional during their Web search engine interactions. They wish to engage in an advice-seeking interaction, but may be frustrated by the inability of the system to respond to their personal medical and health needs and concerns.

Commercial web search engines were designed to help people use natural language expressions and formulate searches as readable and understandable queries. When seeking medical information, such technologies have the potential to make medical language, terms, and expressions understandable in searching for help and asking opinions from the search engine. This technology could offer more human-like communication, and it is often marketed and promoted as a health information service. Such context, however, is problematic when the use of precise clinical terms, descriptions, or concepts is called for. Current search technologies are becoming more akin to natural language communication, and are being used more frequently by users for health-information retrieval, health services, or the sharing of health experiences. Most users may still lack the specialized vocabulary needed to effectively retrieve the information relevant to their condition.

Most studies to date have presupposed that people are able to reach (arguably) credible clinical or health information through commercial web search engines. However, for many people, the operation and outcomes of the commercial web search engine still pose a significant barrier to usable health information. Research is needed to explore the limitations of web search engines and websites for medical or health information seeking.

Large-scale studies of web queries, as outlined in this paper, have strengths and weaknesses. Such studies using real data from web search engines can show large-scale patterns and trends. Frequently,

however, they lack demographic data on individual users and their web search effectiveness.

Conclusions and further research

In an era of health consumerism, evidence-based medicine, and growing shared-decision making between patients and providers, it becomes imperative to provide consistent, reliable, health information via the Web. People expect and require information they can trust, delivered in a format that is understandable and usable. Further research is underway that examines the characteristics of medical or health related queries and sessions using query data from other web search engines.

Acknowledgement

The authors thank the anonymous reviewers for their helpful comments.

References

1 American Medical Association. *Study on Physicians' Use of the World Wide Web*. Chicago: AMA Press, 2002.
2 Impicciatore, P., Pandolfini, C., Casella, N. & Bonati, M. Reliability of health information for the public on the World Wide Web: a systematic survey of advice and managing fever in children at home. *BMJ* 1997, **314**, 1875–81.
3 Jones, R. B., Balfour, F., Gillies, M., Stobo, D., Cawsey, A. J. & Donaldson, K. The accessibility of computer-based health information for patients: Kiosks and the Web. *Medical Information* 2001, **10**, 1469–73.
4 McCray, A. T., Dorfman, E., Ripple, A., Ide, N. C., Jha, M., Katz, D. G., Loane, R. F. & Tse, T. Usability issues in developing a web-based consumer health site. AMIA'00. *Proceedings of the Annual Fall Symposium of the American Medical Informatics Association*, 2000: 556–61.
5 Rozic-Hristovski, A., Hristovski, D. & Todorovski, L. Users' information seeking behavior on a medical library website. *Journal of the Medical Library Association* 2000, **90**, 210–7.
6 European Commission Information Society. Guidelines for quality criteria. *Workshop on Quality Criteria for Health Related Websites (Workshop Report), September 2001*, 2001.
7 Fallis, D. & Fricke, M. Indicators of accuracy of consumer health information on the Internet: a study of indicators relating to information for managing fever in children in the home. *Journal of the American Medical Informatics Association* 2002, **9**, 73–9.
8 McLeod, S. D. The quality of medical information on the Internet: a new public health concern. *Archives of Ophthalmology* 1998, **116**, 1663.
9 Pandolfini, C. Follow up of quality of public oriented health information on the World Wide Web: systematic re-evaluation. *BMJ* 2002, **324**, 582–3.
10 McCray, A. T., Loane, R. F., Browne, A. C. & Bangalore, A. K. Terminology issues in user access to Web-based medical information. *AMIA'98: Proceedings of the American Medical Informatics Association*, 1998.
11 Barnas, G. P. & Kahn, C. E. Assessing consumers' interest in Internet-based health information. *AMIA'99: American Medical Informatics Association*, 1999.
12 Eysenbach, G., Yihune, G., Lampe, K., Cross, P. & Brickley, D. Quality management, certification and rating of health information on the Net with MedCERTAIN: using a medPICS/RDF/XML metadata structure for implementing eHealth ethics and creating trust globally. *Journal of Medical Internet Research* 2000, **2**(Suppl.): 2E1.
13 Pew Internet & American Life. Search engines. *A Pew Internet Project Data Memorandum* June 2002.
14 Pew Internet & American Life. Vital decisions: how Internet users decide what information to trust when they or their loved ones are sick. *Pew Internet & American Life Project Report* 2002.
15 Pew Internet & American Life. Internet health resources: health searches and e-mail have become more commonplace, but there is room for improvement in searches and overall Internet access. *Pew Internet & American Life Project Report* 2003.
16 Hersh, W. R. & Hickam, D. H. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. *Journal of the American Medical Association* 1998, **280**, 1347.
17 Hersh, W. R., Crabtree, M. K., Hickman, D. H., Sacherek, L., Friedman, C. P., Tidmarsh, P., Mosbaek, C. & Kraemer, D. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. *Journal of the American Medical Informatics Association* 2002, **9**, 283–93.
18 Eysenbach, G. & Kohler, C. How do consumers search for and appraise health information on the World Wide Web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ* 2002, **24**, 573–7.
19 Atlas, M. C. First-year students' impressions of the Internet. *Medical Reference Services Quarterly* 2001, **20**, 11–25.
20 Bin, L. & Li, K. C. The retrieval effectiveness of medical information on the Web. *International Journal of Medical Information* 2001, **62**, 155–63.
21 Rogers, R. P. Searching for biomedical information on the World Wide Web. *Journal of Medical Practice Management* 2001, **15**, 306–13.
22 Suarez, H. H., Hao, X., Chang, I. F. & Masys, D. R. Searching for information on the Internet using IMLS and medical world search. *AMIA'97: Proceedings of the Annual Fall Symposium of the American Medical Informatics Association*. Philadelphia, PA: Hanley and Belfus, 1997: 824–828.

23 Spink, A., Jansen, B. J., Wolfram, D. & Saracevic, T. From e-sex to e-commerce: Web-searching changes. *IEEE Computer* 2002, **35**, 107–9.
24 Spink, A. & Ozmutlu, H. C. Characteristics of question format Web queries: an exploratory study. *Information Processing and Management* 2002, **38**, 453–71.
25 Berland, G. K., Elliot, M. N., Morales, L. S., Algazy, J. I., Kravitz, R. L. & Broder, M. S. Health information on the Internet: accessibility, quality, and readability in English and Spanish. *Journal of the American Medical Association* 2002, **285**, 2612–21.
26 Borowitz, S. M. & Wyatt, J. C. The origin, content, and workload of e-mail: consultations. *Journal of the American Medical Association* 1998, **280**, 1321–4.
27 Spielberg, A. R. Sociohistorical, legal and ethical implications of e-mail: for patient-physician relationship. *Journal of the American Medical Association* 1998, **280**, 1353–9.
28 Ferguson, T. Digital doctoring: opportunities and challenges in electronic patient-physician communication. *Journal of the American Medical Association* 1998, **280**(15), 1361–2.
29 O'Connor, J. B. & Johanson, J. F. Use of the Web for medical information by a gastroentology clinic population. *Journal of the American Medical Association* 2000, **284**, 1902–4.
30 Muhlhauser, I. & Berger, M. Evidence-based patient information in diabetes. *Diabetic Medicine* 2000, **17**, 823–9.
31 Rivera, S., Kim, D., Garone, S., Morgenstern, L. & Mohsenifar, Z. Motivating factors in futile clinical interventions. *Chest* 2001, **119**, 1944–7.
32 Spink, A., Nykanen, P. & Pomeroy, M. Seeking medical and health information, or advice, on the Web. *Proceedings of Internet Research 2: International Conference of the Association of Internet Researchers, October 10–14, 2001*. Minneapolis, MN: University of Minnesota, 2001.