A Query Analysis of Consumer Health Information Retrieval

Yi Hong, MLIS; Norberto de la Cruz, PhD; Gary Barnas, MD; Eileen Early, RN; Rick Gillis, MD

Office of Clinical Informatics, Medical College of Wisconsin, Milwaukee, Wisconsin

Abstract

The log files of MCW HealthLink web site were analyzed to study users' needs for consumer health information and get a better understanding of the health topics users are searching for, the paths users usually take to find consumer health information and the way to improve search effectiveness.

Introduction

HealthLink (http://healthlink.mcw.edu), a consumer health information resource includes several features to help users find articles of interest. First, a search feature allows retrieving those articles that match the keyword or phrase specified by users. Second, the HealthLink documents include Dublin Core metadata to assist users to locate needed health information from HealthLink web pages through external search engines. Third, most HealthLink articles have been submitted to DMOZ Open Directory (http://dmoz.org/about.html) to assist people to find HealthLink articles through web directories. To learn more about the users' information needs and make the consumer health information more accessible, an exploratory analysis of search queries submitted to HealthLink was performed.

Methodology

Data from HealthLink log files (December 2001) were imported into a Microsoft Access database. The records contain a subset of queries directly submitted to HealthLink, and queries submitted through external search engines. The external query submissions were analyzed to determine what are the main paths that users take to access HealthLink and what search engines are most popular in searching for consumer health information. The internal query submissions were analyzed to determine the health topics of interest and the characteristics of HealthLink query strings. To ascertain subject categories of HealthLink searches, queries that were submitted more than 10 times were categorized into a set of health topics, which are based on HealthLink topics and Medical Subject Headings (MeSH) tree structures. The following data were collected and analyzed: search engine use frequency; query frequency distribution; subject categorization of queries; top health topics that received most queries; query construction and pattern; and average number of words per query.

Results

Most users found HealthLink articles through external search engines rather than searching directly through the HealthLink Web site. The top 10 external search engines used for HealthLink article retrieval in December 2001 were: Google, MSN, Yahoo, AOL, Lycos, Overture, AltaVista, Ask, Excite, and Dogpile.

The top 10 highest frequency search queries occurring in December 2001 were: symptoms 671, carotid 389, shingles 327, low back pain 279, kidney infection 255, calcifications 202, neck 144, urinary tract infection 134, back pain 131, microcalcification 129. The most frequently searched health topics were: infections, nutrition/herbs, skin diseases, kidney diseases, wellness, digestive diseases, neurological disorders, musculoskeletal disorders, back problems, heart diseases.

Only a few search queries were used repeatedly and most search queries were submitted only once. Queries submitted to HealthLink web site were usually simple and short. On the average, a HealthLink query contained 2.1 words. About 74% of all queries contained one or two words. 37.15% of the queries only had one word. Fewer than 5% of the queries had more than 5 words.

Conclusions

The query analysis reveals users' interests and search behaviors. The top lists of search queries and topics show what users look for and their top health concerns. It seems that the more common conditions and diseases along with general health information receive the most searches. Accessing consumer health information could be enhanced using MeSH or other controlled vocabularies to generate indexing terms, organize search results, facilitate retrieval of articles, and improve retrieval effectiveness based on information retrieval feedback.

References