Ali Shiri
University of Alberta, Edmonton, Alberta

Preliminary Observations on Health Query Terms in a University OPAC: Transaction Log and Co-occurrence Analyses

Abstract: This paper reports preliminary results of a study that investigated the nature, characteristics and types of health and medical queries submitted to an academic OPAC using transaction log and term frequency analyses. The study found that medical and health related searches submitted tended to be broad and include document genre queries.

Résumé : Cette communication présente les résultats préliminaires d'une étude sur la nature, les caractéristiques et les types de requêtes de santé et médicales posées par l'intermédiaire d'un CIEL. Une analyse de la fréquence des occurrences a été effectuée dans les journaux transactionnels. L'étude conclut que les requêtes de santé et médicales sont de nature générales et font allusion au genre de documents recherchés.

1. Introduction

Examination of information search behaviour of users seeking medical and health information has recently gained particular attention. This is in part due to the importance of understanding users' search behaviours to enable us to design and develop better information retrieval systems. The availability of large user search behaviour data sets there is ample opportunity to study and examine the search and interaction behaviour of a large number of users interacting with search engines, OPACs, commercial databases and digital libraries. The aim of this paper is to report the preliminary results of a research project that investigates the nature, characteristics and types of health and medical queries submitted to an academic OPAC using transaction log and term frequency analyses. This paper reports on the study of the following questions:

- What are the most frequently used terms in health queries?
- What is distribution of subject-based and known-item queries?
- Can co-co-occurrence analysis of OPAC query terms provide additional terms for query formulation and reformulation?

2. Prior research

Several studies have addressed the characteristics of users' health queries in various contexts. Spink et al. (2004) reported a study of health and medical queries submitted to two different search engines. They found that a a) small percentage of web queries were medical or health related, b) the top five categories of medical or health queries were: general health, weight issues, reproductive health and puberty, pregnancy/obstetrics, and human relationships, and c) 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. Zhang et al. (2008) examined the frequently used medical-topic terms in queries submitted to a

Web-based consumer health information system. Using transaction log analysis and multi-dimensional scaling they analyzed users' queries and compared them to the MeSH thesaurus. They found that Consumer health-information-seeking behaviors are better understood in terms of the selected medical subjects. These terms from the user queries focus more on symptoms, medical assistance, causes of a disease, foods and medicines related to disease prevention and treatment, medical test results, and so on. The findings allow doctors and health professionals to better understand and serve their consumers. Poikonen and Vakkari (2009**)** examined the differences between expressions used by lay persons and professionals in nutrition related questions and answers, and to what degree General Finnish Ontology (GFO) and a medical thesaurus (FinMeSH) cover these expressions. They found that The vocabularies of lay persons and professionals were found to be quite similar. All term types expressed in both questions and answers matched better with FinMeSH than with GFT. There were some differences in coverage of the thesauri between patients' terms and physicians' terms. FinMeSH covered 60% of synonyms used by physicians whereas only 38% of synonyms used by patients could be found in that thesaurus. In a transaction log analysis study of search characteristics in differen information retrieval environment, Wolfram (2008) found that users of an OPAC issued more purposeful and longer queries, with more queries per session and longer inter-query times as compared with searches carried out in bibliographic databases, specialized search system and web search engines. Lau and Goh (2006) studied the search patterns on a university OPAC to identify query search failure patterns. They found that 11.8% of all queries involved the use Boolean operators and that The use of keyword searches contributed to 68.9% of all queries while other options such as title, author and subject accounted for 16.5%, 8.2% and 6.4% of all searches respectively. Building upon the above studies, the research reported here has made use of transaction log analysis of a university OPAC query log file to examine the search characteristics of academic users seeking health and medical information. In particular, the nature and characteristic of queries and the advantages of term co-occurrence analysis have been examined.

3. Methodology

This study made use of a combination of data analysis techniques, namely transaction log analysis and co-occurrence analysis. The transaction log file of the University of Alberta OPAC for the time period of May 1 - June 20, 2005 timeframe was acquired to start the process. Data derived from the transaction log file was entered into an Microsoft Excel file for processing. The each record contained the following data elements: query, date, time and the IP address. The data was processed to identify individual queries and sessions. In total 64,836 records were manually reviewed to ensure that all the health and medical queries would be identified. In case of uncertainty about a term or whether it is a health related term, the MeSH thesaurus was used. In total, 5848 Health/Medical-related queries were identified. A *term*, or *keyword*, is defined as a string of one or more alphanumeric characters delimited by non-alphanumeric characters. Queries with Boolean Operators were separated to allow for detailed analysis. All of the 5804 queries were examined in terms of the type of search in order to identify the distribution of subject-based and known item search queries. We used TAPOR (Text Analysis Portal for Research), for term frequency and co-occurrence analyses. The portal provides a wide range of tools including highly frequent terms and co-occurrence and collocation terms.

4. Results

4.1 Highly frequent query terms

A list of all the terms in the query logs was created in order of frequency. The terms were closely examined to choose a sample of the most frequently occurring terms. The top 40 most frequently used terms appearing in health queries appear in Table 1

Term	**Frequency**
nursing	403
health	356
care	193
journal	176
medicine	172
therapy	116
medical	114
clinical	110
psychology	108
practice	91
general	90
guide	88
disease	81
disorders	80
surgery	78
human	76
pubmed	73
manual	70
canadian	69
anatomy	67
physiology	66
drug	66
physical	65
assessment	63
mental	62
handbook	62
american	61
Research	60
nutrition	60
development	59
brain	53
canada	52
personality	51
nurses	50
occupational	47
e-therapeutics	46
patient	45
child	45
history	45
communication	45

Table 1. Forty most frequently used query terms

A review of the most frequently used terms in the above table shows that medical and health related queries submitted to OPACs tend to have a low level of specificity. In other words the query terms are generally broad. An interesting observation is related to the genre related query terms. As can be seen in the above table terms such as journal, handbook, or manual have been used quite extensively. This finding reflects the mental model of searchers looking for health and medical information on an academic OPAC.

4.2 Query types

Unlike many web search log analysis studies, OPAC searches contain a significant number of queries associated with journal or book titles. They contain a large number of known item searches such as title and author searches. Table 1 shows the distribution of queries across search types. As can be seen a significant number of queries (41%) are associated with keywords-in-title queries. This is particularly interesting when is compared with the finding in Lau and Goh (2006) who found that title searches accounted for 16.5% of the searches. It should be noted that title queries include book as well as journal titles. Known Item Searches (KIS) are the searches for specific titles, authors or publishers, which constitute less than 2% of the queries.

Query types	No of queries	% of total queries	% of KIS queries
Subject queries	3303	57	N/A
Title queries	2404	41	96
Author/Publisher queries	96	2	4
Unknown	2	0	N/A
Total	5805	100%	100%

Table 2. Distribution of known item searches

4.3 Term co-occurrence analysis

We choose two terms for co-occurrence analysis to see the extent to which these co-occurring terms would be useful to support users’ query formulation and reformulation. The selected terms were ‘health’ and ‘care’. The top twenty co-occurring terms were examined to see if there are any kinds of relationships between them. Table 3 shows the terms that have co-occurred with the term ‘health’ in the data set. One of the main findings of in relation to the terms co-occurring with the term ‘health’ is that these terms can assist searchers in making their queries more specific. For instance, we see ‘public’, ‘community’, ‘physical’, and ‘mental’ that when they are individually combined with the term ‘health’ provide a more specific contextual information and allow users to formulate more well-defined queries.

Co-occurring term	**frequency**
Care	98
Nursing	80
Community	34
Journal	34
Mental	33
Canada	33
Public	31
Medicine	27
Research	27
Policy	22
General	22
Information	22
Physical	22
Psychology	20
Assessment	19
Child	18
Practice	17
Law	17
Management	16
Qualitative	16
Technology	15
Clinical	15
Evidence	15

Table 3. Terms co-occurring with the term ‘health’

These terms may not necessarily fall under any hierarchical relationships that are present in thesauri. However, they do support users’ query formulation. Table 4 shows the terms that co-occur with the term ‘care’ in the dataset.

Co-occurring term	**Frequency**
Health	99
Nursing	82
Critical	23
Patient	20
Plan	19
Intensive	16
Palliative	15
Clinical	12
Physiology	11
Primary	11
Surgery	11
Medicine	11
Medical	11
Cardiac	9
Psychiatric	7
Practice	7
Death	7
Child	7
Chronic	7
Wound	6

Table 4. Terms co-occurring with the term ‘care’

As can be seen from Table 4 terms co-occurring with the term ‘care’ provide an interesting variety of related terms that may be used for query formulation or modification. To show this graphically, a visual representation of the co-occurring terms is provided in Figure 1. The user can browse these terms and may choose to narrow down his/her search query using the examples of various types of care.

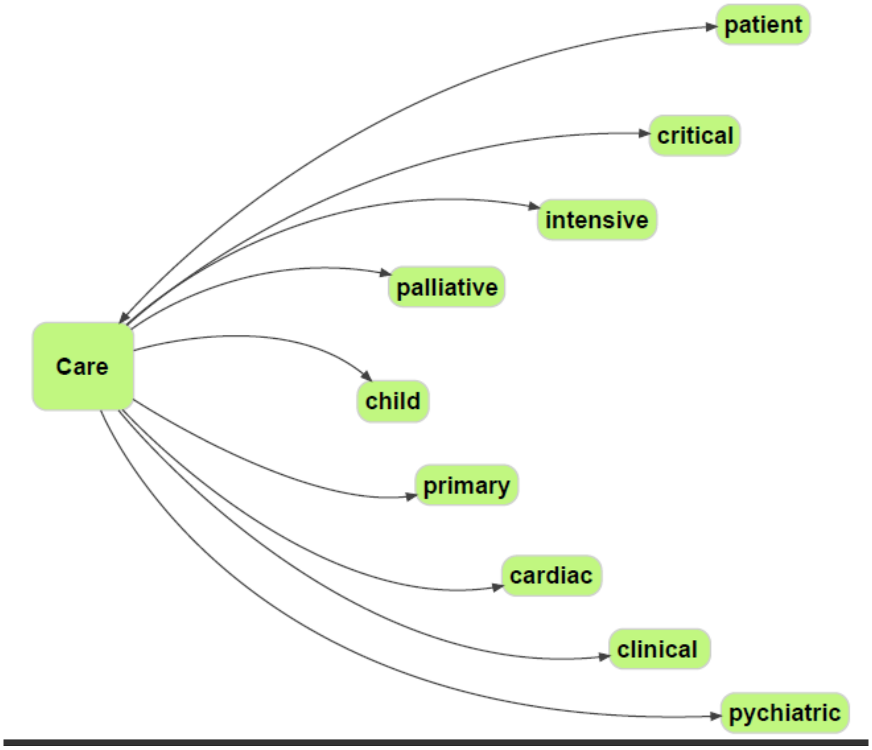

Figure 1. Visual representation of a select number of co-occurring terms with ‘care’

5. Conclusion

Preliminary results from the above study suggest that transaction log analysis can help us understand how medical and health information searchers look for information on an academic OPAC. An analysis of the query terms used by health information searchers shows that they come to the OPAC with a particular mental model, thinking of an information retrieval system that can assist them in finding mainly book-type information sources. With the recent developments in the expansion of the OPAC as an all-encompassing campus-wide information retrieval system that includes all types of documents and document genres, this mental model may not successfully support searchers for specific queries. Providing concept and visual maps of co-occurring terms, derived from transaction logs, will allow users to form more specific and well-defined queries and to form a more inclusive perspective of the content of an OPAC. Also, it is claimed that such techniques as transaction log analysis, query term co-occurrence analysis and visualization are powerful tools that offer a number of ways in which users query data can be effectively utilized to provide information-rich user interfaces for IR systems.

Acknowledgement: The author would like to thank Maria Tan for her assistance in data processing and preparation.

References

Lau, E. P., & Goh, D. H. 2006. In search of query patterns: A case study of a university OPAC. *Information Processing and Management*, *42*(5), 1316-1329.

Poikonen, T. & Vakkari, P. 2009. Lay persons' and professionals' nutrition-related vocabularies and their matching to a general and a specific thesaurus. Journal of Information Science 35(2), 232-243.

Spink, A., Yang, Y., and Jansen, J. et al. 2004. A Study of Medical and Health Queries to Web Search Engines. Health Information and Libraries Journal 21(1): 44-51.

Wolfram, D. 2008. Search characteristics in different types of Web-based IR environments: Are they the same? *Information Processing & Management*, 44 (3), 1279-1292.

Zhang, J., Wolfram, D., Wang, P., Hong, Y., & Gillis, R. 2008. Visualization of health subject analysis based on query term co-occurrences. *Journal of the American Society for Information Science and Technology*, 59(12), 1933-1947.