Reformulation of Consumer Health Queries with Standard Medical Vocabulary

Samina Raza Abidi, Krista Elvidge, Hadi Kharrazi, Michael Shepherd, Carolyn Watters, & Jack Duffy

Faculty of Computer Science, Dalhousie University, Halifax, Canada, abidi@cs.dal.ca.

Despite the proliferation of consumer health websites, consumers, for the most part, do not utilize effective search queries when seeking online health information. Previous studies have suggested that reformulating consumer health queries (lay terminology) with standard medical terminology could increase the overall effectiveness of the search query. Five different medical topics were presented in this study and the participants were asked to formulate queries (keywords) for them. These user-formulated queries were then mapped to MeSH (Medical Subject Heading), which was used as the medical terminology source to create reformulated queries. R-precision was calculated for each query (original and reformulated) and then compared using an ANOVA on the first ten Google results. Unexpectedly, the mean R-precision of the original lay terminology queries was significantly higher than the mean R-precision of the reformulated queries. The significantly higher R-precision recorded in this study can probably be explained, to a great extent, by the fact that the reformulated queries returned documents with more medical vocabulary and most participants found it difficult to understand standard medical vocabulary; and therefore, they were unable to identify the relevance of the documents returned from the Google search.

Keywords

Consumer Health, Google, Health Information Retrieval, MeSH, Query Reformulation

1. Introduction

The Internet has become an important resource for health consumers seeking health and medical information [1]. According to Skinner, Biscope, and Poland, millions of Canadians use the Internet to search for health information, and this is the fastest-expanding Internet category with a growth rate of 34% per year [2]. Health information retrieved from credible Internet sources has the potential to empower consumers [3, 4]. Overall, such knowledge leads to better-informed decisions, the adoption of healthy behaviors, the strengthening of patient-physician relationships and an increase in patient compliance and satisfaction, resulting in improved health care outcomes and a more efficient utilization of health care resources [3, 4].

Unfortunately, despite the proliferation of consumer health websites, consumers, in an attempt to manage their health, do not utilize effective search queries which could facilitate their retrieval of credible information [5]. Previous research by Zeng *et al*. [6] has shown that there is often a significant mismatch between the vocabularies used by consumers when retrieving health information and the terminology of the health website, which could impede the effectiveness of health information retrieval. This mismatch of consumer health vocabulary and medical content terminology could be bridged using query reformulation [7].

In fact, Plovnick and Zeng [7] suggest that reformulating consumer health queries (lay terminology) with standard medical terminology could increase the overall effectiveness and efficiency of the search query, thereby improving the consumers' retrieval of relevant health information.

This research explored the following question: Does the reformulation of consumer health queries with standard medical vocabulary improve the results of the search query?

2. Methodology

Our approach for query reformulation involved manually mapping the participants' free-text queries to the concepts in MeSH (Medical Subject Heading), which was used as the source of standard medical terminology. MeSH is the National Library of Medicine's controlled vocabulary thesaurus. Controlled Internet searches using both free-text and reformulated queries were carried out by the participants using the Google search engine. Our null hypothesis stated that the average R-precision of a health information search using original, lay terminology queries would be equal to or less than the R-precision for reformulated queries. In order to evaluate the null hypothesis we performed one-way analyses of variance (ANOVA) on the data obtained from our study. R-precision was chosen as the dependent variable for the statistical analysis. The order of presentation of the search results (original query or reformulated query) was also examined to determine if the order of presentation affected the participants' judgements. The detailed methodology for this study can be decomposed into the following steps.

2.1 Participant Recruitment

Ten computer science students (undergraduate/graduate) were recruited from Dalhousie University for this study. Recruitment methods included class announcements, mass emails, and personal contacts. Students who were affiliated with health or medical disciplines were excluded from participating in this study. Since the selection of the participants was random, no screening was done with respect to the English language profficiency and both native and non-native speakers of English language were recruited. There was no remuneration for participation in this study.

2.2 Participant Consent and Training

Prior to commencing the study, the participants were required to sign an informed consent form which outlined the risks and benefits associated with the study, described the study, time commitments, the participants' right to withdraw without consequence, and an assurance of confidentiality and anonymity of personal data. The consent form also included permission to use direct quotations from the transcript in literature resulting from this study (devoid of any information that could be used to identify a participant). Each participant was provided with a unique user ID to ensure confidentiality of responses. Each participant underwent a pre-study training session in order to guarantee that he/she had an understanding of how to use and interact with the system.

2.3 Query Selection and Reformulation

Five different medical topics or medical scenarios were presented in the study. The topics chosen were: 1) high blood sugar and an enlarged heart, 2) nose bleeds in cold weather, 3) sudden bleeding in the brain, 4) infant skin rash, and 5) impaired night vision. The participants selected their own queries (keywords) without any restrictions. The participants were allowed to use as many keywords as they desired in each query, but they could not use any of the special advanced search features in Google.

Once the queries were formulated by the participants, the researcher then mapped these free-text queries to the concepts in MeSH. The participants' queries were reformulated by replacing their original search terms with equivalent medical vocabulary synonyms. For example, for the impaired night vision medical scenario, a participant queried *night vision problems*. In this case the researcher provided the participant with the reformulated MeSH keyword query *nyctalopia*. If the participants' original query terms included any of the standard medical terminology suggested in MeSH, the terms were left unaltered. Given that the researcher was not blinded of the hypothesis, the reformulated MeSH queries were verified by another researcher who was unfamiliar with our study and hypothesis. The queries of this second reformulation were compared with those in the study, and were an exact match.

2.4 Internet Searches using Reformulated and Free-Text User Queries

The participants used both the original queries and the reformulated queries to initiate controlled Internet searches using the Google search engine. Participants were randomly assigned to begin information searches with either their original query or the reformulated query. This randomization was done to prevent an order effect that could have biased the results of this study. A researcher was present with the participants at all times to ensure that the methodology of the study remained uniform for all searches.

2.5 Evaluation of the Searches

The participants were required to evaluate the first ten search results for each query, determine which documents were relevant to the medical topic under investigation, and record their responses on the provided form. After completing this task, participants were asked to complete a separate questionnaire regarding their preferences for the original or reformulated queries. The questionnaire consisted of a modified Likert scale and a comment box.

3. Results

3.1 Quantitative Results

Each of the ten participants selected five free-text queries which were then manually reformulated by the researchers. Next, the queries were randomly assigned to the participants. While ten participants might be a small number by some experimental standards, our repeated measures design yielded one hundred data points which should be more than enough power to detect a true effect.

R-precision was calculated for each of the original and reformulated queries. R-precision can be defined as the precision after R documents are retrieved in response to a query, where R is the number of relevant documents for that query [8]. As Google returns ten items per page, we calculated R-precision for the first ten documents retrieved in response to each query, i.e., the first page returned by Google. Precision is the proportion of these ten documents that were judged by the participant as being relevant to the query [9].

We then computed the average or mean R-precision for both the original (free-text) and reformulated queries for each participant (Table 1). Finally, the micro-average of the individual mean R-precision values of each of the two types of queries was calculated. As shown in Table 2, the average of the individual means and medians for the original queries were higher (with a smaller standard deviation) compared with the reformulated query statistics.

Table 1 Individual Mean Average R-precision Values.

Participant	Original Query Mean R-precision	Reformulated Query Mean R-precision
001	0.794	0.625
002	0.885	0.582
003	0.724	0.414
004	0.742	0.578
005	0.833	0.782
006	0.777	0.757
007	0.769	0.540
008	0.714	0.649
009	0.855	0.819
010	0.745	0.255

Table 2 Micro-Average R-precision Values

	Original Query	Reformulated Query
Mean	0.784	0.600
Standard Deviation	0.058	0.172
Median	0.773	0.604

In order to test our hypothesis we performed several statistical analyses of the data. First we calculated a simple t-test for the data in Table 1. This test indicated that the original query was significantly preferred over the reformulated query ($t(10) = 3.20$, $p = .009$, $d = 1.43$). We also used a 2 x 2 factorial table as our platform for all ten users. We used R-precision as the dependent or response variable for each query. The explanatory variables were: 1) Type (whether a query was original or reformulated) and 2) Order (whether a participant began a search using the original or reformulated query).

The null hypothesis was tested using a one-way ANOVA on these variables. As previously mentioned, the null hypothesis was that the average R-precision of the searches using the original queries is equal to or less than that of the reformulated queries. The value of alpha chosen for all tests was .05. The assumption of homogeneity of variances was evident in all ANOVA tests, and the data distributions were reasonably normal. In addition, all ANOVA results were tested and confirmed with non-parametric tests.

For the query variable Type, the results rejected the null hypothesis and indicated that the original (free-text) queries were significantly more effective at retrieving results that the participants considered pertinent compared with the reformulated queries, *F*(1,98) = 7.27, *p* = .008, η_p^2 = .07. Although the mean of the variable Type (original query) was significantly higher than that of Type (reformulated query), the effect size estimate indicated that the difference was not substantial. The factor Type of the query accounted for only 7% of overall variance (effect + error), as calculated by using partial eta-squared (η_p^2 = .07). Partial eta-squared for an experimental factor is defined as the proportion of total variation attributable to the factor, partialling out (excluding) other factors from the total non-error variation [10]. The R^2 (coefficient of determination, which measures the proportion of variance in one variable that is explained by the other variable) also echoed similar results. R^2 was 7 % and adjusted R^2 (which takes into account the number of variables in the model) was 6%. These results indicated a small to modest effect size. Therefore, although the mean R-precision for the original queries was significantly higher than the mean R-precision of the reformulated queries, this effect was only modest.

A second ANOVA was used to evaluate any effect the order of the query might have had on search results. In this case, our null hypothesis (that the order of the query would not have any significant effect on the R-precision of the search results) could not be rejected, *F*(1,98) = 3.28, *p* = .073, R^2 adjusted = .023. Since in this study we used p-value of 0.05 as conventional reference point to describe the statistical significance of our results, a p-value of 0.073 indicated that the order of the query did not have a significant effect upon the relevance of the search results.

The box plot of Figure 1 displays the comparative relationships that the variables Type and Order had upon the response variable (R-precision). The left cluster of box plots represents R-precision distribution for the Type original (free-text) queries. The left shaded box plot shows the range of R-precision values when the original queries were performed first (Order 0) and the unshaded box plot shows the range when the original queries were performed second (Order 1). The right cluster represents the Type reformulated queries. The right shaded box plot shows the range of R-precision values when the reformulated queries were performed first (Order 0) and the unshaded box plot shows the range when the reformulated queries were performed second (Order 1).

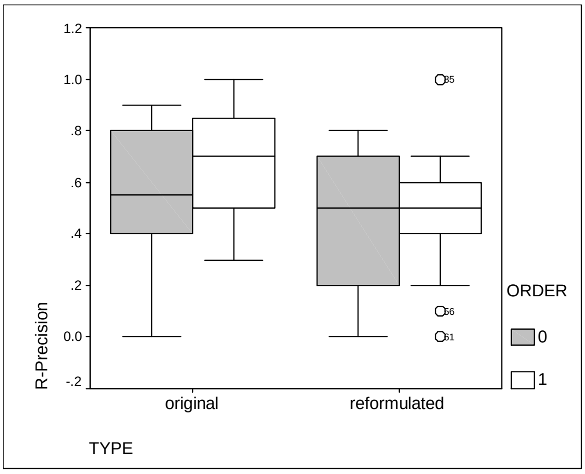

Figure 1 Boxplot for R-precision vs Type and Order

Furthermore, to investigate the impact of Topic on the results, we performed an ANOVA on the five different medical topics (explanatory variable). The null hypothesis, that the different medical scenario topics would not have a significant effect upon R-precision, could not be rejected, $F(4,95) = 1.63$, $p = .172$, R^2 adjusted = .025. This result clearly shows that the effect of the type of query on R-precision was independent of the medical topic selected for the searches.

3.2 Qualitative Results

The participants' were asked to complete a short Likert questionnaire to determine their preference for using either the original or reformulated queries. The first question asked if the results from the reformulated queries were easier to read and understand compared with the results from the original (free-text) queries. Seven of ten participants disagreed with this statement, and three reported a neutral response. In accordance with this result, seven participants preferred using their original queries to retrieve information compared with the reformulated queries.

Nine participants added a written comment concerning their preference of layman term queries compared with reformulated medical terminology. In essence, these respondents stated that they preferred to use lay terminology queries, and they preferred the results generated from this type of query, as demonstrated by the following participant comment: "The relevant results given from Google were much closer to the title or topic or subject I was searching".

4. Discussion and Future Work

Our results indicated that the original (lay) queries were significantly more effective at retrieving results relevant for these users than the reformulated queries, as measured by participant-rated R-precision. The significantly higher R-precision for original queries recorded in this study can probably be explained, to a great extent, by the fact that the reformulated queries tended to retrieve documents with more medical jargon than did the original queries and most participants found it difficult to understand standard medical vocabulary, and therefore were unable to identify the relevance of the documents returned from the Google search.

There are some limits to generalizing our findings. The participants of this study were computer science students who may not have been familiar with the medical topics used in this study; therefore, the results of this study cannot be generalized to all health consumers. It is likely that health consumers who live with specific health ailments would have a greater knowledge base concerning their condition, including familiarity with medical terminology specific to their health, and for this reason, a similar study which investigates the usefulness of medical terminology queries for these populations may yield different results.

This study required participants to perform original and reformulated queries exclusively. Combining original and reformulated query terms might produce interesting results, and perhaps a future study which uses a blended form of query terms might provide the opportunity to explore whether the participants undergo a learning effect which may affect their interpretation and comprehension of query results. Furthermore, this study was conducted using a Google search engine, and since document sets may differ depending upon the search engine, it might be interesting to conduct a similar study using a variety of search engines to compare query results produced in various search environments.

Finally, this research applied the MeSH vocabulary to map the participants' original queries to their medical equivalent. Future studies, using other medical terminology systems, may yield different results.

5. Conclusions

In this study we investigated the effectiveness of the search query, according to user preference of the reformulation of consumer health queries using standard medical vocabulary over the free-text queries submitted by the participants (original query).The effects of the order in which the queries were conducted, as well as the types of medical scenario topics, were found to be non-significant. Our results indicate that the intuitive belief that reformulated queries will improve search results is not supported by our results. In fact we found just the opposite is true. Further research to uncover the extent of generalizability of our result is needed.

References

1 Bensley, RJ, Brookins-Fisher J. Community health education methods: A practical guide. 2nd ed. Mississauga, ON: Jones and Bartlett; 2003.

2 Skinner H, Biscope S, Poland B, Goldberg E. How adolescents use technology for health information: Implications for health professionals from focus group studies. Journal of Medical Internet Research [serial online] 2003 [cited 2005 Jan 5]; 5(4). Available from: URL: http://www.jmir.org/2003/4/e32/.

3 Borzekowski D, Rickert V. Adolescent cybersurfing for health information: A new resource that crosses barriers. Archives of Pediatrics & Adolescent Medicine 2001; 155: 813-817.

4 Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Education Research 2001; 16(6): 671-692.

5 Spink A, Wolfram D, Jansen BJ, Saracevic T. Searching the Web: the public and their queries. Journal of the American Society for Information Science and Technology 2001; 52(3): 226-234.

6 Zeng Q, Kogan S, Ash N, Greens R, Boxwala A. Characteristics of consumer terminology for health information retrieval. Methods of Information in Medicine 2002; 41(4): 289- 98.

7 Plonvick RM, Zeng QT. Reformulation of Consumer Health Queries with Professional Terminology: A Pilot Study. Journal of Medical Internet Research 2004; 6(3): 1-12.

8 Tague-Sutcliffe J, Blustein J. A Statistical Analysis of the TREC-3 Data. Proceedings of the 3rd Text Retrieval Conference TREC; 1994 November; Gaithersburg, MD, 385-398.

9 Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. New York: Addison Wesley & ACM Multimedia; 1999.

10 Cohen J. Eta-squared and partial eta-squared in fixed factor ANOVA designs. Educational and Psychological Measurement 1973; 33: 107-112.