MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Strategies for Supporting Consumer Health Information Seeking

Alexa T. McCray, Nicholas C. Ide, Russell R. Loane, Tony Tse

National Library of Medicine, Bethesda, Maryland, USA

Abstract

Despite a growing number of available Web-based health information resources, consumers continue to face a variety of barriers as they attempt to access these resources. Developing a system that appropriately responds to user queries poses several challenges. Guided by an earlier study that analyzed a large number of queries submitted to ClinicalTrials.gov, we developed a variety of techniques to assist user information seeking. We tested the efficacy of these techniques by submitting the original user queries to our new search engine to determine if these techniques would result in better system performance. Overall, the number of query failures was reduced, but the largest improvement was found in the system's query suggestion capability. For a subset of query failures, the current system was able to cut the earlier failure rate almost in half, in most cases providing a suggestion rather than directly finding records. The techniques described here provide a new approach for responding to user queries. The techniques are tolerant of certain types of errors and provide feedback to assist users in reformulating their queries.

Keywords

Databases; Information Services; Information Storage and Retrieval; Internet; Terminology

Introduction

Information retrieval systems often place the burden on the user to define and formulate a query to satisfy a particular information need, despite evidence that users frequently do not have sufficient information to accomplish this task [1-3]. Insufficient knowledge about the domain, about the content and nature of the database being searched, and about the idiosyncrasies of the retrieval system, all adversely affect the information seeking process—from knowing *what* to ask and *how* to formulate the question, to evaluating the retrieval set for relevance to the information need.

Formulating a query in a specialized domain such as medicine is particularly challenging [4-7]. Differences between words consumers use to describe medical concepts and technical medical terminology at the lexical (e.g., spelling, inflectional variation), syntactic (e.g., word order) and semantic (e.g., overgeneralization, underspecification) levels may impede effective information retrieval. Furthermore, unfamiliarity with a system's search syntax or the scope of information available further decreases the likelihood of resolving an information need. Even if users are able to correctly identify terms, they are unlikely to adhere carefully to any prescribed syntax.

Consider the situations listed in Table 1. In each case, the query text reflects a user's expression of an information need. The situation is described by a complex interaction between the user's query and the retrieval system.

Table 1: Sample query and situation pairs for a consumer health site.

Query	Situation
dishwashers	The document set does not contain any relevant documents.
nearsightedness, alaska	No documents include both these concepts, but there are documents relevant to either term alone.
stage ii breast cancer herceptin	There are documents including all these concepts and also possibly related documents on herceptin alone.
how do you use interleukin to treat patients with breast cancer?	The query contains non-essential words and there are relevant documents.

We recently conducted a study of search failures on two National Library of Medicine consumer health sites in an effort to determine what types of queries resulted in search failures, with the ultimate goal of developing interventions to assist users in recovering from these failures [8]. The study identified three broad classes of failures, query failures due primarily to content coverage, failures due to user query formulation, and failures due to system functionality. Guided by these results, we subsequently developed a variety of techniques to assist user information seeking. We tested the efficacy of these techniques by submitting a large number of the original failed queries to our new search engine to determine whether these techniques would result in better system performance.

Materials and Methods

For our earlier investigation we collected all the queries that were submitted as basic searches to *ClinicalTrials.gov* during the month of November 2001. We identified all queries that re-

sulted in at least one retrieval result, all those that resulted in no retrieval, and those where we offered spelling assistance. We then fully analyzed 1,000 of the failed queries to determine the cause of the failure. Some of the queries failed because there was no content available at the time. For example, the query term "aggressive fibromatosis" while relevant to *ClinicalTrials.gov*, resulted in no retrieval because the system had no information about it in 2001. Many failures, however, were due to characteristics of the query, such as misspellings, abbreviatory expressions, run-together phrases, and incorrect use of search operators.

At the time of our earlier investigation we were running a commercial search engine that had been enhanced by our own terminology server. The terminology server handled lexical variation, including providing spelling help, and it allowed for expanding queries with synonyms whenever these were available [9]. Since that time we have developed an in-house XML-based search engine (SE), which works together with the terminology server and a variety of our own retrieval algorithms. SE is currently being used in the production *ClinicalTrials.gov* system as well as in several other of our public systems. Figure 1 below shows the design of the SE search engine.

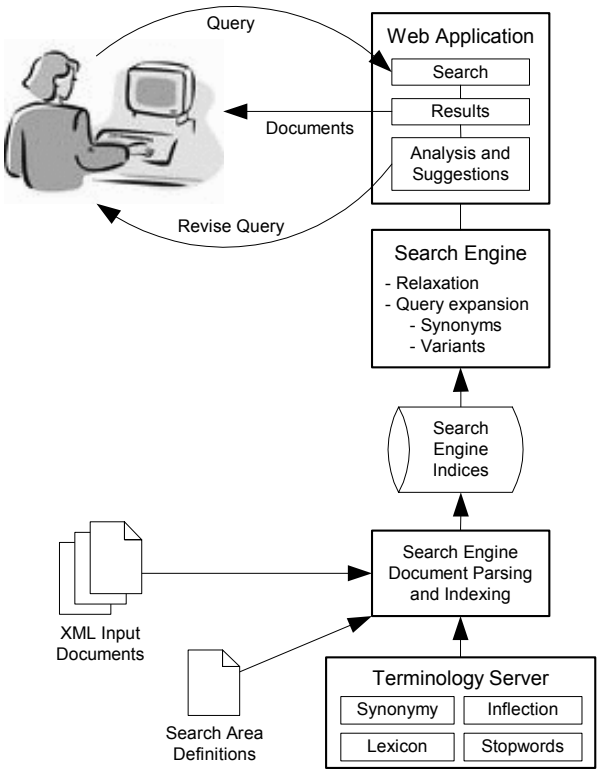

Figure 1 - SE search and retrieval system

The SE search and retrieval system indexes and searches structured documents in eXtensible Markup Language (XML) format. It first parses these documents according to a fine-grained tokenization method, with punctuation characters treated as separate tokens. Thus, the string "non-hodgkin's lymphoma" consists of six tokens: "non, -, hodgkin, ', s, lymphoma". The index process generates a dictionary of all the words in the XML document set each time new data come into the system, and the synonymy set is pruned to the set that overlaps the words in the corpus. At retrieval time this means that any query suggestions are limited by a "closed world assumption". For example, if the user submits the misspelled query "osteoparosis", the system will suggest both "osteoporosis" and "osteopetrosis", but only if there are records for both of these conditions in the current document set.

The terminology server provides synonymy and lexical information from the Unified Medical Language System [10]. The stopword list consists of a small number of words, such as conjunctions and prepositions. Stopwords are retained in the query text, but are treated specially during query analysis.

The tags within the XML documents define names for various parts of the document. SE uses a mapping file to associate a named search area with a set of document parts. Each document part is given a weight indicating its relative importance within the named search area. For example, in *ClinicalTrials.gov* a hit within the brief title document part (with a weighting of .95) is considered to be more relevant than a hit in the location document part (with a weighting of .55).

At search time, the user submits a query to the Web application. The query either retrieves a result set directly, or if nothing is found, the system offers suggestions that may help the user either reformulate the search or try one of the searches suggested by the system.

The search strategy generates multiple interpretations of the user's query, merging the result sets for all interpretations. Interpretations are generated by breaking the text into all possible phrase combinations. The most restrictive interpretation treats the entire text as a single phrase. The least restrictive interpretation treats the text as the conjunction of all the meaningful words. The resulting expression is a "relaxed" alternative to treating the entire text as a phrase.

All search terms are expanded with plurals, possessives, hyphen variants, compound words, and synonymy, if possible. All possible relaxed alternatives are evaluated, weighted, and combined to produce the final result. The relaxation search is evaluated by:

- Performing phrase searches, which produces a list of hits for each piece and its expansions.
- Merging the hits, which produces a set of scored documents for each piece.
- Applying AND operators to the pieces, which produces a set of scored documents for each relaxed alternative.
- Weighting the alternatives so that the more relaxed alternatives contribute less to the overall scores. A penalty is applied for each inserted AND operator.
- Applying an OR operator to the weighted alternatives, which produces a set of scored documents for the full relaxation search.

A document containing all the words of the query text, adjacent and in the same order as expressed in the query, will receive a higher score than a document containing all the words somewhere, but not adjacent to each other. In general, pieces of a que-

ry that are known terms are expanded with synonyms and find relevant documents, and also contribute to higher document scores. Pieces of a query that are not recognizable have trivial expansions, rarely occur in the data, and do not contribute significantly to the document scoring.

The search strategy algorithm can be illustrated with an example for the query phrase "heart attack in elderly".

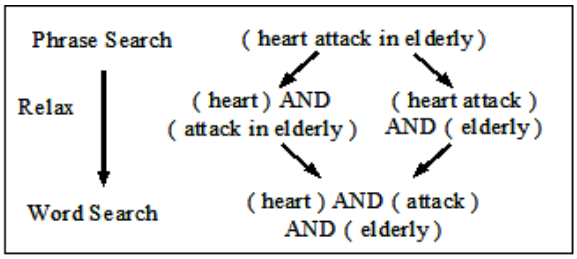

Figure 2 - Relaxation of "heart attack in elderly"

The algorithm involves several steps, as follows.

1) Break the query text into words and identify stopwords.

heart = meaningful word 1
attack = meaningful word 2
in = stopword
elderly = meaningful word 3

2) Create a list of all possible relaxed alternatives by inserting one or more AND operators.

(heart attack in elderly)
(heart) AND (attack in elderly)
(heart attack) AND (elderly)
(heart) AND (attack) AND (elderly)

3) Extract the pieces from the relaxed alternatives.

heart
attack
elderly
heart attack
attack in elderly
heart attack in elderly

4) Expand each piece with its synonyms and inflectional variants.

heart → hearts, cardio
attack → attacks
elderly → older adult, older adults, geriatric, geriatrics, aged person, aged persons
heart attack → heart attacks, heart infarction, heart infarctions, myocardial infarction, myocardial infarctions, …
attack in elderly → attacks in elderly

heart attack in elderly → heart attacks in elderly

5) Weight the alternatives (synonyms and inflectional variants are excluded for expositional clarity).

1.00 * (heart attack in elderly) OR
0.10 * ((heart) AND (attack in elderly)) OR
0.10 * ((heart attack) AND (elderly)) OR
0.01 * ((heart) AND (attack) AND (elderly))

The relaxation approach produces a set of intermediate results that can serve as query suggestions. Since there is exponential growth in calculation with the number of meaningful words in the search text, care must be taken to keep response times reasonable. Since all alternatives and suggestions are AND clauses, when any piece finds zero documents, the entire expression will find nothing and can be skipped. Experience has shown that if the underlying phrase search can be performed in 50 ms, a query with seven meaningful words can be fully evaluated with suggestions in less than a second.

If the result set is not empty, then the results of the search are displayed together with a link to the "query details" page. If the result set is empty, then the query details page is immediately displayed.

Since we intend SE to be an improvement over our earlier search methods, we decided to test its performance by re-evaluating the queries that were posed to *ClinicalTrials.gov* in November 2001. We ran all 155,777 queries through SE and then compared the retrieval results to those of the retrieval system that was in place in 2001. We also ran the smaller set of 1,000 queries (all of which had been selected because they had resulted in query failures) through SE to see whether we were now able to do any better. Because we wanted to control for content issues, we ran SE against the 2001 *ClinicalTrials.gov* data set, rather than against our current content. That is, since we were testing the efficacy of SE, we did not want to count the addition of clinical trial content as a search engine improvement.

Results

We have implemented the retrieval techniques described above, and SE is currently being used in the production *ClinicalTrials.gov* system as well as in several other of our public systems. For all query results, by clicking the "query details" button, the user is able to see the precise search executed by the system, including how the query was parsed by the system and any synonyms that were searched. In the case of a query failure, the system offers immediate assistance to the user in the form of query suggestions.

Figure 3 shows the query details page on *ClinicalTrials.gov* for the sample query, "stage ii breast cancer herceptin olympus California".

The query details page offers users both the ability to modify their queries and to receive feedback. Included are suggested query variants and a table showing terms and words from the user's query and how often they appear in the document set. The query suggestions are displayed with an explicit syntax that includes Boolean operators. The intent is to guide the user in

the differences between similar queries. A suggestion may be executed through the TryIt! link. The Individual Terms report lists recognized terms and the number of times they were found in the data set. The report lists the compound terms first followed by the individual words. If a term (or any of its synonyms) is not found in the data set, then this is marked as "none", as is the case for the term "olympus" in this example. Available synonyms for each term and word are shown under "Also searched'.

Query Suggestions	
(stage ii breast cancer AND herceptin AND california)	TryIt!
(stage ii breast cancer AND california)	TryIt!
(stage ii breast cancer AND herceptin)	TryIt!
(breast cancer AND herceptin AND california)	TryIt!
stage ii breast cancer	TryIt!

Individual Terms	Count
"stage ii breast cancer herceptin olympus california"	None
"stage ii breast cancer" Also searched: breast cancer stage ii	72
"breast cancer" Also searched: cancer of the breast breast tumor malignant malignant breast neoplasm malignant tumor of breast	290
"stage ii" Also searched: stage two	419
"california"	1886
"olympus"	None
"herceptin" Also searched: trastuzumab	39
"cancer" Also searched: neoplasms tumor malignancies neoplastic growth	2729
"breast" Also searched: mammary gland	330
"ii"	3004
"stage"	1095

Figure 3 - Partial screen shot of Query Details page in ClinicalTrials.gov for the query: "stage ii breast cancer herceptin olympus California"

To test the efficacy of these new methods, we ran the full set of *ClinicalTrials.gov* November 2001 queries against the earlier and current retrieval systems, using the 2001document set. Table 2 shows the results.

Table 2: Results of searching all November 2001 ClinicalTrials.gov queries on two different retrieval systems

	2001 System	2003 System (SE)
Query failure	30,822 (20%)	24,787 (16%)
Records found	105,126 (67%)	94,168 (60%)
Suggestion offered	19,829 (13%)	36,722 (24%)
Total queries	155,777	155,777

In the 2001 system, 20% of the queries failed completely, while in the current system a somewhat smaller percentage (16%) failed. There was a decrease in the number of records directly found in the 2003 system. Since query suggestions are an important feature of SE, we report those cases where the system is able to offer suggestions. The terminology server feature of the original 2001 system also included a suggestion capability, but these were exclusively spelling suggestions. Thus, all "suggestions offered" for the 2001 system are examples of spelling correction, while the suggestions offered for the 2003 system include not only spelling, but also suggestions of the type shown in Figure 3 above.

Table 3 shows the results for the subset of 1,000 failed queries from November 2001.

Table 3: Results of searching 1,000 November 2001 ClinicalTrials.gov "failed" queries on two different retrieval system.

	2001 System	2003 System (SE)
Query failure	844 (84%)	411 (41%)
Records found	0 (0%)	21 (2%)
Suggestion offered	156 (16%)	568 (57%)
Total queries	1,000	1,000

Since the 1,000 queries were chosen because they had failed, by definition, the 2001 system has zero records found. The current system cut the earlier failure rate almost in half (411/844), in most cases providing a suggestion (568) rather than directly finding records (21). The 2001 system was able to make a suggestion in only 16% of the cases. The 2003 SE system, on the other hand, was able to do so for 57% of the queries.

Discussion

The 2003 system has a smaller number of failed queries than the earlier system. The difference is a modest 4%, yet it reflects some important underlying changes to our search methodology, including "query relaxation", improved tokenization, and synonym enhancement. For example, the query "mini-transplants" previously did not match the document term "mini-transplant" because there was no known variant for that term. SE now tokenizes the term into "mini, -, transplants," generates the variant "transplant," recombines the tokens ("mini-transplant"), and retrieves five documents. Several of the earlier query failures now provide results using SE because additional synonyms have been added to the terminology server either by the UMLS, or through our own synonym enhancement activities. As an example, the tradename "Xeloda" is now expanded to "capecitabine". Perhaps more interesting is the case where query relaxation interacts with phrase level synonymy. For example, the earlier failed que-

ry "multiple system atrophy with postural hypotension" now retrieves a document because the phrase "postural hypotension" finds the synonym "orthostatic hypotension".

The 2001 system found 7% more records than the 2003 system. This can be accounted for by the fact that the 2001 search algorithm allowed, as its least restrictive search, any word in the query to retrieve a document. That is, if nothing was otherwise found, the query was reduced to an OR expression of its constituent words and a retrieval set was presented to the user. In the current system, the search is not executed until the user interacts with the system. This query suggestion capability is perhaps the largest overall improvement. The earlier system provided only spelling suggestions, while the current system provides suggestions based on our query relaxation methods. For example, the query "high dosage of chemotherapy on breast cancer cells" which failed in the earlier system, now offers as one of its suggestions "dosage AND chemotherapy AND breast cancer", which retrieves one document. Similarly, the failed query "nutrition oncology cancer diet", now suggests "nutrition AND oncology", which retrieves two documents.

Not all query suggestions will reflect what the user intended. However, providing feedback to users on what is available allows them to either reformulate or abandon their search. Given the difficulty and ambiguity of query interpretation, our system attempts to convey to users how their queries were interpreted and how they relate to the underlying document set.

Conclusions

We have developed a variety of techniques to assist user information seeking. We tested the efficacy of these techniques by submitting a large number of user queries to our new search engine to determine whether the system we had developed would result in better system performance than an earlier retrieval system that was in use for *ClinicalTrials.gov*. Overall, the number of query failures was reduced, but the largest improvement was found in the system's query suggestion capability. For a subset of query failures, the current system was able to cut the earlier failure rate almost in half, in most cases providing a suggestion rather than directly finding records. The techniques described here provide a new approach for responding to user queries. The techniques are tolerant of certain types of errors and, importantly, provide feedback to help users in reformulating their queries. We continue to investigate the feedback that we are giving with the goal of improving the assistance we offer to users as they seek information on our consumer health information systems.

References

[1] Belkin NJ, Oddy RN, Brooks HM. ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*. 1982;38(2):61-71.

[2] Kuhlthau CC. Inside the search process: information seeking from the user's perspective. *Journal of the American Society for Information Science*. 1991;42(5):361-71.

[3] Jansen BJ, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*. 2000;36:207-27.

[4] Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. *Methods Inf Med* 2002;41(4):289-98.

[5] Smith CA, Stavri PZ, Chapman WW. In their own words? A terminological analysis of e-mail to a cancer information service. *Proc AMIA Symp*. 2002;:697-701.

[6] Patrick TB, Monga HK, Sievert ME, Houston Hall J, Longo DR. Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. *J Med Internet Res*. 2001 Jul-Sep;3(3):E24.

[7] Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. *Health Educ Res*. 2001 Dec;16(6):671-92.

[8] McCray AT, Tse T. Understanding search failures in consumer health information systems. *Proc AMIA Symp*. 2003:;430-4.

[9] McCray AT, Ide NC. Design and Implementation of a national clinical trials registry. *J Am Med Inform Assoc*. 2000 May-Jun;7(3):313-23.

[10]Unified Medical Language System. National Library of Medicine. http://umlsinfo.nlm.nih.gov/.

Address for correspondence

Alexa T. McCray
National Library of Medicine
8600 Rockville Pike
Bethesda, Maryland 20894 USA
mccray@nlm.nih.gov