Understanding Search Failures in Consumer Health Information Systems

Alexa T. McCray, Tony Tse
National Library of Medicine, Bethesda, MD

ABSTRACT

We examined queries that led to search failures on two National Library of Medicine Web-based consumer health sites, ClincialTrials.gov and MEDLINEplus. The purpose of the study was to analyze and categorize queries resulting that led to no results with the ultimate goal of developing interventions to assist users in recovering from those failures. We first analyzed over 2,700 queries, iteratively developing a coding scheme. We subsequently applied the codes to an additional set of 2,000 queries. We found that most of the queries were in scope, relevant to the system being searched, and did not exhibit so-called consumer language. As the final step, we developed a taxonomy based on whether the search failures were due primarily to content issues, to problems in query formulation, or to limitations of the search system. The results reported here have informed the further development of our own systems, and they may be helpful to others as they seek to improve consumer access to health information.

INTRODUCTION

Online information systems are among the primary resources used by consumers seeking personal healthcare information. Despite the increasing availability of consumer health information systems, consumers often encounter barriers in information seeking [1]. Some of the obstacles to effective retrieval that have been identified include submission of ill formed query strings, mismatch in terminology, and terms that are too broad, too narrow, or out-of-scope [2-3]. Such queries lead to incomplete retrieval at best, and to irrelevant or no results at worst. In this study, we examined queries that led to no results on two National Library of Medicine (NLM) Web-based consumer health sites, *ClincialTrials.gov* and MEDLINEplus. The purpose of the study was to analyze and categorize the queries resulting in search failures, with the ultimate goal of developing interventions to assist users in recovering from those failures.

System designers and intermediaries have long sought to understand users' online information seeking behavior to evaluate system effectiveness and effect improvements. Transaction log analysis facilitates the analysis of large numbers of user actions and corresponding system responses, since log data are gathered unobtrusively and are machine-readable [4]. Brewer has found, however, that, on average, queries submitted to Web search engines consist of no more than 1.7 words [5]. This limited input contributes to the primary problem faced by log analysis methods, namely that they can assess users' information needs only indirectly. In addition, it is not always possible to know whether the source of a query is direct user input or an application acting on behalf of a user. Nonetheless, these methods do provide a range of useful insights about information retrieval systems and the queries that are posed to them. An early study by Tonta investigated search failures in online catalogs, finding that failures could be grouped into four broad categories. Only one of the categories related to query formulation by users, while the others were concerned with system design or performance issues [6]. Jansen and colleagues used automated means to characterize over 51,000 queries submitted to the Excite search engine [7]. They noted that there were a high number of incorrect uses and errors in the queries reviewed. Most of the problems related to the misuse of supported operators, or to the use of unsupported operators.

As consumers increasingly seek information online to become better informed about personal health care issues, they face various barriers to effective information retrieval. Perhaps because medical terms are often difficult to remember and spell, ill-formed queries and term mismatches are particularly common problems. Our earlier analysis of the query strings submitted to the NLM homepage indicated that, although the majority of the terms were medical in nature, many contained misspellings and were otherwise ill-formed [8].

We undertook the work reported here in order to gain a better understanding of the types and frequency of queries that result in no retrieval, thereby informing not only the further development of our own systems, but also informing the research of others as they seek to improve consumer access to health information.

METHODS

We analyzed over 4,700 queries from two NLM consumer health sites by inspecting samples of their log files. All log files used in this study were anonymized before we began work on them. We first analyzed approximately 2,000 queries that had resulted in no retrieval on *ClinicalTrials.gov* during the summer months of 2001. *ClinicalTrials.gov* contains basic information on ongoing and completed clinical trials and is intended for use primarily by patients. In this phase of the analysis we were interested in identifying the phenomena that might account for the search failures. The goal was to develop a coding scheme that we would subsequently validate on new data. For each query, we noted the nature of the phenomena that may have led to the failure and began to build the taxonomy based on the observed data. This process continued until we discovered no new phenomena.

In the refinement phase, we coded an additional 750 queries in four separate rounds. The first round involved coding 300 queries and the subsequent three rounds each involved 150 queries. During each round both authors coded all of the queries independently. We checked for inter-rater reliability by assessing whether there was complete, partial, or no agreement in assigned codes. Queries for which there was only partial or no agreement were discussed until we reached consensus. This phase also resulted in the development of a set of guidelines for use during the final coding process.

Once we were satisfied that the coding scheme was sufficiently well-developed and that it covered all the cases we had seen in our test data set, we applied it to two new sets of failed queries, one set chosen from the November 2001 logs of *ClinicalTrials.gov* and another from the MEDLINEplus logs for the same time period. NLM's MEDLINEplus system is a consumer health information system containing extensive information on hundreds of health topics. Since *ClinicalTrials.gov* is a specialized consumer site, we were particularly interested to see if the nature or frequency of query failures would be significantly different on the more broadly-based MEDLINEplus system.

We filtered each log file to exclude queries from within NLM, and searches resulting in no retrieval were extracted and aggregated into a unique list. From these lists, we selected a random sample of 1,000 failed queries from each system. We assigned codes to the individual components of complex queries. For example, the complex query "parkinson's disease and pshychology" was divided into three components: two terms and one operator ("and"). For our final counts, we recorded only unique instances of codes for a complex query, since our interest was in search failures at the overall query level.

Both authors independently coded all queries and for those cases where there was disagreement, we discussed the codings until we reached consensus. Most cases of coding variations were superficial; it often being the case that one of us inadvertently omitted a single applicable code. Because *ClinicalTrials.gov* offers spelling assistance to users, we subsequently also reviewed all spelling mistakes to see if users took advantage of this capability in the reformulation of their queries.

RESULTS

Table 1 shows the results of the development and refinement phases for the coding scheme.

		Rater Agreement (%)		
Round	**String (#)**	**Full**	**Partial**	**None**
Development Phase				
1	854	N/A	N/A	N/A
2	1,172	N/A	N/A	N/A
Refinement Phase				
1	300	48	31	21
2	150	49	25	26
3	150	75	20	5
4	150	66	25	9

Table 1. Number of *ClinicalTrials.gov* search failures reviewed and degree of inter-rater coding agreement during the development and refinement of the coding scheme.

Since the goal of the development phase was simply to identify the types of phenomena evident in the data set, we did not attempt to reach consensus on our coding assignments during this phase. During the first round of the refinement phase, we reached full agreement on just under half (48%) of the assignments, partial agreement on 31% of the assignments, and we had no agreement on 21% of the codings. For the partial agreement cases, it was often the case that one of us missed a relevant code, but occasionally we disagreed on the applicability of a particular code. When this happened, we clarified the coding guidelines for that code, or, in some cases, our disagreement led to a further refinement of the coding scheme.

Table 2 shows the results, together with multiple examples for each observed phenomenon, of manually coding the final data set.

Code	Description	*Clinical Trials.gov*	MEDLINE plus	Examples
A	**In Scope**			veterinary cancer trials, National Coverage Decision, infant mortality rate, reiter's syndrome
	Error-free	296	274	
	Coded with ≥1 B code	582	491	
	Total	878	765	
B1	**Misspelling**	268	163	apashia, artories, DIABETIES, mavulur degeneration, pulmanary fibrosis, multile sclerosis
B2	**Abbreviation**	8	9	pt, dx, tx, dr., Calif, MA
B3	**Non-alphanumeric**	207	139	+ % / _ " ; : ' ? *
B4	**Run-together Phrase**	31	12	useofenergyhealing, lymesdisease, lookaheadtrial vaccinesandperiodontaldisease
B5	**Word/Phrase Split Inappropriately**	25	11	myco bacterium; bi,polar disorder; cartilage, replacement; arthritis, rheumatoid; Barre-Guillain
B6	**Search Operator**	147	136	and, or, vs, not, NAC+bronchitis, non-hodgkins lympoma/New jersey; NEAR
B7	**Unordered List of Words**	48	63	Topotecan Gemcitabine lung, antegren multiple sclerosis, allergy peanut
B8	**Truncation**	5	8	neuro, sarco, naltrexon*, ENAXO*
B9	**Ellipsis**	44	46	Downs [syndrome]; hoxsey [therapy]; prinzmetals [angina]; [Citrus] aurantium; afasak [study]
B10	**Natural Language Phrase**	92	79	atlas of orthosis; What does beer do to your bones?; accurate assessment of precision errors
B11	**Acronym**	76	47	EPA, hplc, VIN, adh, C225, efc 4584, FR901228, CHARM, imac, PKD, TS-1
B12	**Possible Consumer Term**	36	36	nose bleed, no feeling in legs, crooked spine, implantable minimized telescope, tubes tied
H	**Web Address**	2	3	ucsf.edu/research/trials; w.w.w. washington e.d.u.
L	**Language other than English**	16	39	fase 3; vitamina C; malignes, Pleuramesotheliom, cardiopathies valvulaires
P	**Publication**	0	10	N. ENG. J. mED 2000 343: 16-22, urology october issue, nejm,
U	**Unknown, Uncertain or Out of Scope**	117	191	rastetieds, calciomimeticagew, Carreca, bargar, J4, NuVue, pea poisson, resources allocation, drum

Table 2. Comparison of manual coding results for 1,000 search failures from each of two consumer health information systems. (Since queries are often multiply coded, the total number of codings exceeds the total number of queries.)

When coding the data, we first assessed whether a query was in scope (A) and then assessed its other characteristics (B1-B12). If the query was either not in scope (U) or fell into one of the other major categories (H, L, or P), we coded it accordingly.

We used the following guidelines when coding the final data set:

1. To determine whether the query is in scope, search Google; for *ClinicalTrials.gov*, use <TERM>, "clinical trial"; for MEDLINEplus, use <TERM>, "health".
2. B codes always co-occur with A. Select all applicable B codes.
3. Codes H, L, P, U stand alone.
4. Apply multiple codes to a query as necessary.

The first guideline indicates that before assigning codes to a query, we checked Google, doing minimal repair of the query terms as necessary (e.g., fixing spelling mistakes). If Google returned relevant results, then this gave us some independent verification of the relevance of the query to the particular NLM site being searched. In other words, if we found that a query from the *ClinicalTrials.gov* data was found in a clinical trials context on another web site, then it was potentially in scope for *ClinicalTrials.gov*. Since MEDLINEplus deals with a broad range of health topics, we searched for these queries with the additional term "health" to determine whether it was potentially in scope.

The second guideline indicates that for those terms that were in scope, we first chose A and then added one or more applicable codes from B. Often this would be only one B code, indicating, for example, a

spelling error, but in many cases multiple B codes were assigned. Some examples are "multiplesclrosis", which was coded as misspelled (B1), and also as a run-together phrase (B4); "obstetrics and gyno", which was coded as having a search operator (B6), and exhibiting truncation (B8); and "congestive, heart, failure AND prognosis AND erderly", coded as misspelling (B1), non-alphanumeric (B3), phrase split inappropriately (B5), and containing a search operator (B6).

Guideline three indicates that for URL's (H), terms formulated in languages other than English (L), citations to the literature (P), and terms that were out of scope or about which we were uncertain (U), we assigned no additional codes.

Guideline four indicates that when multiple different phenomena were evident in a complex query, all phenomena were coded. For example, the query "Hypothyroid it'd it'd" was coded as A for "Hypothyroid", B3 for the apostrophes, and U because we were uncertain what "it'd" meant.

Table 2 indicates that for both systems, most queries were in scope, yet they resulted in no retrieval. Of the 878 A codes assigned to the *ClinicalTrials.gov* data, 296 did not have any B codes assigned. Of the 765 A codes assigned to the MEDLINEplus data, 274 did not have any B codes assigned. This means that between 34% and 36% of the queries that resulted in no retrieval were in scope and error-free, but there happened to be no matching data available on the respective systems. Examples of such queries posed to *ClinicalTrials.gov* are "menstrual synchrony" and "neural stem cells", and examples from MEDLINEplus are "Jimson Weed" and "aberdeen low back pain scale".

Both *ClinicalTrials.gov* and MEDLINEplus offer spelling assistance in the form of a list of possible alternatives. For example, if the user types in "alzhimer", the systems will offer "Alzheimer" as a possibility. We inspected the spelling mistakes (B1) in the *ClinicalTrials.gov* data to determine whether if the system offered an alternative, the user would choose it. For over 60% of the misspellings, the system did, in fact, offer one or more alternatives. Although these were generally, though not always, appropriate, users only chose an alternative in about 45% of the cases in which one was offered.

Because we were particularly interested in whether our data would reveal large numbers of so-called "consumer" terms, we coded these separately (B12). Only a relatively small number of queries (coincidentally 36 for both systems) consisted of what might be considered consumer language.

The phenomena listed in Table 2 do not always signal an error. They do, however, all represent characteristics that may cause difficulties for (or, as in the cases reported here, no retrieval from) a particular information system. Because our primary concern was to understand the nature of search failures so that we might effect improvements in our search systems, we further categorized each observed phenomenon according to whether it reflected primarily content issues, problems in query formulation, or limitations of the search system. Table 3 shows the resulting taxonomy.

	Content	**User**	**System**
In Scope	√		
Misspelling		√	
Abbreviation		√	
Non-alphanumeric			√
Run-together Phrase			√
Incorrect Split		√	
Search Operator			√
Unordered Words			√
Truncation		√	
Ellipsis		√	
Natural Language			√
Acronym		√	
Consumer Term		√	
Web Address			√
Non-English			√
Publication			√
Out of Scope	√		

Table 3. Taxonomy of search failures. Query failures that are due primarily to **content** coverage, to **user** query formulation, or to **system** functionality.

The first column of Table 3 indicates that queries that are in scope (and error free) but still result in no retrieval are problems of content coverage. That is, the system being queried may have no data about the requested topic. Queries that are out of scope are also problems of content coverage, because they indicate that there is a mismatch between users' expectations of the content of the information system, and what it, in fact, has to offer.

Queries that seem to fail primarily due to the user's formulation of the query are indicated in the second column. For example, abbreviations and acronyms while not incorrect, do present particular challenges to search systems because of their underspecified and

ambiguous nature [9]. Consumer terms, if not recognized by the system, present similar problems.

Finally, the third column indicates those queries for which the failure may be due to limitations of the retrieval system itself. For example, non-alphanumeric characters in a string, such as an underscore between words may cause a failure if the system does not treat it appropriately at search time. Run together phrases most likely do not come directly from the user, but, rather, from some intermediary referral site which has failed to maintain the original spaces between words. These phrases, as well as those that are split inappropriately, are especially difficult to handle, since it is not immediately clear where to insert word breaks or how to reconstruct the intended phrase. Natural language-like phrases are not ill-formed from a user perspective, and the challenge for the search system in these cases is to parse the query, returning information about individual concepts whenever possible.

DISCUSSION

The results reported here indicate that for both of the systems we investigated a large number of queries were in scope, yet they resulted in no retrieval. For some queries, the information, though relevant, was not available on the particular system being searched, and in other cases, certain characteristics of the query, such as misspellings, lead to the search failure. In addition to content coverage issues and problems in query formulation, we found that some search failures may be due to the limitations of the search system itself.

Each of these phenomena is potentially amenable to an intervention that will improve consumer access to the information system. Queries that are in scope and whose topics are not covered, particularly if they are asked with some frequency, might signal possible additions to the information system itself. Queries that are out of scope might indicate that the information system needs to be clearer about its coverage. Queries that fail due to problems in query formulation, may well lend themselves to interactive intervention at search time, and those queries that fail because of a limitation in system functionality indicate clear areas for system improvement.

CONCLUSIONS

In this study we investigated the nature of search failures on two consumer health systems. We found no significant differences between the two systems, although there was some variance. Only a relatively small number of queries consisted of what might be considered consumer language. It, therefore, does not appear that, at least for the majority of the data we considered here, consumers are using terminology that is significantly at odds with the terms in the systems being searched. Instead, our taxonomy of search failures indicates that for those queries that resulted in no retrieval there were three primary classes of phenomena that may have accounted for the failures. These were issues in content coverage, user query formulation, and system functionality.

In this study we investigated only those queries that resulted in no retrieval. It is clear that this is only the first step in gaining an understanding of whether consumers are actually finding the information they need and in a form that is accessible to them. The results reported here have, however, informed the further development of our own systems, and they may be helpful to others as they seek to improve consumer access to health information.

REFERENCES

1. Eysenbach G, Jadad AR. Evidence-based patient choice and consumer health informatics in the Internet age. *J Med Internet Res* 2001 Apr-Jun;3(2):E19.

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

3. Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. *Methods Inf Med*. 2002;41(4):289-9.

4. Peters, TA. The history and development of transaction log analysis. *Library Hi Tech*; 1993;42(11:2):41-66.

5. Brewer EA. The consumer side of search. *Communications of the ACM* 2001;45(9):41.

6. Tonta YA. *An analysis of search failures in online library catalogs*. 1992. Doctoral Dissertation. University of California, Berkeley, 318 pp.

7. 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;36:207-227.

8. McCray AT, Loane RF, Browne AC, Bangalore AK. Terminology issues in user access to Web-based medical information. *Proc AMIA Symp* 1999;:107-11.

9. Fredriuk CS. The effect of abbreviations on MEDLINE searching. *Acad Emerg Med* 1999 Apr;6(4):292.