Unstructured English Queries: How Users Request Information

Carol A. Keene
Department of Computer Science & Engineering
University of Colorado at Denver
P. O. Box 173364 Denver, Colorado 80217-3364
ckeene@carbon.denver.colorado.edu

Abstract

Presented with the task of locating needed information in on-line, full-text documentation, users must express queries in the language of the retrieval system. Many of these query languages are based on Boolean logic or restricted natural language syntax, and users find it difficult to express information needs. Experiments conducted at the University of Colorado asked participants to enter English queries to locate information needed to solve problems ranging from very specific to very general ones. No restrictions were placed upon grammar or vocabulary. The collected queries were very short, telegraphic in style, used few verbs, and contained frequently occurring terms from the stored vocabulary. There were no statistically significant differences in query contents based upon a participant's knowledge of the topic or English communication skills.

1 Introduction

Today's technology is making full-text documents widely accessible to more and more people every day. Collections of articles, complete books, and even encyclopedias are commercially available while the Internet allows one to download documents via phone lines from the convenience of office or home. Software products from operating systems to compilers, word processors, and productivity tools are placing user documentation on-line. This trend towards electronic documentation instead of, or in addition to, paper documents presents a number of problems to novice and expert users alike. The information need must be determined, the source of documents identified, the request expressed in a format appropriate to the retrieval system, and the information extracted from the documents presented. This paper assumes that the problem to be solved and the source of the text documents are known, and focuses on issues relating to expressing the information need.

2 Information Needs

An individual's need for information can occur in a spectrum from the very specific to the very general. Frants and Brush [3] designate the needs as concrete or problem-oriented. Examples of very specific information needs would include determining the syntax of an operating system command, finding the definition of a technical term, or locating the bibliographic citation of an article. This need is satisfied by the retrieval of any occurrence of the target information, whether that information appears in one document or one hundred documents. There is no

need for high levels of recall if the set of retrieved documents includes at least one with the desired answer. The most desirable responses will exhibit high levels of precision so the user does not have to search irrelevant documents. At the other end of the spectrum are the most general types of queries which seek background information or look for every document in a collection that discuss a given topic. These queries require high levels of recall, the higher the better. Users expect to scan multiple documents looking for information and would accept moderate levels of precision. Queries at the center which require a moderate level of specificity of information need relatively high levels of both recall and precision. Retrieving some irrelevant documents will be tolerated if the desired information is found with a minimum of effort.

Document retrieval (DR) systems attempt to satisfy all types of information needs by targeting high levels of both precision and recall for every query. If successful, such systems would certainly provide at least one document containing the answer to a specific request and would have few irrelevant documents to distract users seeking less specific information. Retrieval strategies include full-text scanning, multiple forms of inverted indexing, vector-space models, and probabilistic models augmented by advanced techniques such as automatic thesauri, term weighting, natural language document analysis, or relevance feedback (see [13] and [2]). Many DR systems also include query refinement mechanisms in which users identify the terms or keywords which are most important to the success of the query. Comparisons of DR systems using various benchmark document collections has not identified one retrieval strategy as superior to the others in all situations. The strategy with the best precision and recall performance differs from

collection to collection and query to query [8]. These studies assume that the user writes syntactically and semantically correct queries.

Queries are formatted in a variety of ways and may require knowledge of Boolean logic, a restricted natural language syntax, or knowledge of the set of keywords selected for the documents. The ability to express information needs thus requires knowledge of both the syntax and keyword vocabulary of the DR system being used. These restrictions are accepted in retrieval systems such as Dialog, BRS, SMART, or Medline which access a variety of databases with very large document or citation collections. However, such restrictions are not easily accepted when searching the on-line documentation of a software product or an electronic book. Every computer user has anecdotes detailing frustrations with so-called on-line help systems which are anything but helpful. The result is often that the capabilities of such products remain under-utilized or the product is abandoned in favor of one with a more compatible human-computer interface. Query formatting continues to be a difficult question for DR systems.

3 Formatting Queries

Query formats are as varied as document retrieval strategies and include menu-based and direct-manipulation formatting as well as sublanguages with specialized vocabulary and syntax. Users must learn the formatting procedures of each retrieval system used or rely upon a search intermediary to conduct the search on an unfamiliar system. Menu-based formatting and other direct-manipulation techniques restrict the user to the choices the system allows. This, of course, means the system will receive only queries that it can interpret correctly, but may be frustrating

for users who have difficulty expressing their needs using the required vocabulary and syntax. For large commercial retrieval systems, it is reasonable to require the client to use an intermediary or to learn the query formatting language and procedures. Here the user must learn the vocabulary and syntax from either on-line help or printed manuals. These searches are likely to have very specific requirements for the target documents and need high levels of recall (and precision). Systems using natural language query processing must incorporate syntactic and semantic analysis plus vocabulary and concept recognition and word-sense disambiguation into the retrieval process. Queries may be much less structured, but the user again must learn the required vocabulary and any syntactic rules or restrictions.

Two interesting questions are (1) Given the opportunity to express information needs in unrestricted English, what are the characteristics of the resulting queries? and (2) What are the implications for processing the queries? The following discussion assumes active users as described by Carroll and Rosson [1] who bring to the current task a history of previous successes and failures and who resist relinquishing previously successful search techniques regardless of their efficacy in the present task.

4 Experimental Results

Two experiments were conducted using Electrical Engineering (EE) and Computer Science (CS) students at the University of Colorado at Boulder to collect unstructured English queries and to test a prototype document retrieval system. (Results of the prototype retrieval system are reported elsewhere [9].) Both experiments used a document collection which consisted of a small number of short

text segments (approximately 1,000 to 2,000 bytes each) selected from on-line materials documenting a VLSI (Very Large Scale Integrated) circuit verification software package called CSIM (the Colorado SIMulator) under development at the University of Colorado.

The first experiment used a subset of 80 document segments (also called nodes). Questions were prepared at three different levels of generality. The most concrete questions requested users to find details of the CSIM structures; the intermediate questions referred to CSIM concepts and procedures; and the least specific questions requested information about general concepts of logic simulation. Seventeen undergraduate students who were experienced users of CSIM were requested to enter a maximum of 3 queries for questions randomly selected from a pool of 25 questions. Because each subject had a limited amount of time, not all subjects were given the same questions or the same number of questions. Only 4 of the 17 subjects consistently used complete sentences with imperative the favored sentence type. Of the 743 queries collected, 85% contained only phrases or groups of words, with an average query length of 4 words.

Since most of the queries just paraphrased the the target question, a second experiment was designed to include a question pool which presented problem situations and allowed the subject to determine the information need based upon the problem presented. The 20 problems represented information needs ranging from the very specific to the very general. The subject pool was expanded to include persons for whom English is not the native language and persons with little or no experience with CSIM in specific or logic simulation in general.

Twenty-three graduate and undergraduate students from the Electrical Engi-

neering and Computer Science Departments participated in the second experiment. Over 900 queries were collected with results similar to the first experiment. Again, 85% of the queries were not complete sentences. The average query length was 3 words. Figure 1 shows one of the problem statements used and a sample of the queries entered. Queries are characterized by lack of grammatical correctness and a telegraphic style which omits verbs and connective words. In addition to the lack of complete sentences, the most common grammatical errors were noun-modifier number disagreement, improper ordering of nouns and modifiers, and subject-verb disagreement.

The subjects of the second experiment were also categorized according to their major department (EE or CS), level of domain knowledge (novice, intermediate, or expert), and native language (English or other). Table 1 shows the information for each subject, and Table 2 summarizes the subject characteristics. There were no statistically significant differences in average query length or use of complete sentences between groups of subjects with different knowledge levels or English communication skills.

Experiment 2 also compared each query against the stored vocabulary for the document collection. The stored vocabulary consisted of the full-text vocabulary with stop-words removed. The stop-word list contained common words of English, certain irrelevant strings contained in coding examples (such as variable names), and a subset of the most frequently-occurring terms. Because the domain includes terms such as *and*, *or*, and *not* as technical terms, these three terms were included in the stored vocabulary. Within the 908 queries entered by the subjects, there were 341 distinct queries. Two queries containing exactly the same terms but in

QUESTION 10

Figure 5 is the behavioral model for an ALU. Where can you find an explanation of the meaning of the indicated (highlighted) statement from IN_LIST?

SUBJECT	QUERY
101	CHDL documentation
102	what is grp
	arguments of grp
103	GRP
	model interface
107	inlist grp
109	INPUT LIST IN CIRCUIT MODEL
	BEHAVIROAL MODEL DESCRIPTION
111	chdl models
112	give me a explanatio of line 3 in IN_LIST
114	help behavioral model IN_LIST
116	grp
117	in_list
	alu
119	interface
120	meaning of GRP
	interface
121	model
	behavioral model
122	grp
	macro
	function grp
123	grp
	indata

Figure 1: Experiment 2: Sample Queries

Table 1: Subject Characteristics

Subject	Department	Knowledge	Language
101	EE	Expert	English
102	EE	Expert	Other
103	EE	Intermediate	Other
104	EE	Expert	English
105	EE	Expert	English
106	EE	Intermediate	Other
107	EE	Intermediate	English
108	EE	Expert	English
109	EE	Intermediate	Other
110	EE	Intermediate	English
111	EE	Expert	English
112	EE	Novice	Other
113	EE	Intermediate	Other
114	EE	Intermediate	English
115	CS	Novice	English
116	CS	Novice	English
117	CS	Novice	English
118	CS	Novice	English
119	CS	Novice	English
120	CS	Novice	English
121	CS	Novice	English
122	CS	Intermediate	English
123	CS	Novice	English

different order were considered identical. Table 3 shows the number of distinct queries and their occurrence frequencies. The frequently occurring queries were entered by more than one subject and, in some cases, by all the subjects for certain questions. Individual subjects entered a single query for more than one question, and different subjects entered a single query for different problems.

Of the 341 distinct queries, 77 contained no matching terms in the stored vocabulary, and 14 of the 77 occurred with frequencies greater than 1. A closer examination revealed that more than two-thirds of the unmatched queries contained at least one of the frequently-occurring terms that were deliberately omitted from the stored vocabulary.

5 Discussion

The two experiments discussed above found linguistic characteristics similar to those reported by Marchionini [11], Landauer et. al. [10], Furnas et. al. [5] [4], Guindon [6], and Guindon et. al. [7]. When permitted to express information needs in unrestricted English, users entered short queries consisting primarily of ungrammatical sentence fragments or lists of words. Common grammatical errors included lack of subject-verb or verb-object agreement, misspelled words, improperly placed modifiers, and omission of connective words.

The average query length from the experiments reported here is somewhat less than the average length found by other researchers. The difference may be explained by the fact that all the subjects of both experiments were regular computer users and were accustomed to the terse commands required by operating systems. The small percentage of complete sentences among the queries is consistent with

Table 2: Summary of Subject Characteristics

Characteristic		Students
Department		
	EE	14
	CS	9
Domain Knowledge		
	Novice	9
	Intermediate	8
	Expert	6
Language		
	English	17
	Other	6

Table 3: Experiment 2: Occurrence Frequency of Distinct Queries

Frequency of Occurrence	Number of Distinct Queries	Frequency of Occurrence	Number of Distinct Queries
1	188	10	1
2	31	12	2
3	7	16	1
4	9	18	1
5	5	19	1
6	5	20	1
7	4	21	1
8	3	22	1
9	2	48	1

that found by the studies listed above and, thus, cannot be attributed solely to frequency of computer use.

There were some queries used by most subjects for a variety of problems. These queries usually contained at least one of the most frequently-occurring terms in the stored vocabulary. Further investigation in a less constrained experimental environment is needed to determine if this behavior reflects a query vocabulary preference or is caused by the selection of the problem situations presented to the subjects. Monitoring usage of a document retrieval system over an extended period of time with no pre-assigned problems to solve will permit a more accurate characterization of the frequency of vocabulary usage and the rate of inclusion of the most highly-frequently occurring terms in queries.

Although the current experiments did not reveal any significant differences based upon language skills or prior domain knowledge, these results need to be verified. If a document retrieval system is to be used by persons from varied cultural backgrounds, the entire human-computer interaction needs to be examined to insure semantic clarity for all users. Russo and Boor [12] present guidelines for items to examine.

6 Conclusions

Given the ability to enter unrestricted, unstructured English queries to retrieve full-text on-line documents, users prefer short, ungrammatical queries based upon the most highly-frequently occurring terms in the domain vocabulary. Designers of help systems and document retrieval systems need to consider accommodating these preferences as a way of providing maximum functionality with minimum

user effort.

Though achieving the highest levels of precision and recall may not be required for all types of queries, the experimental results presented here indicate that there is no easy way to determine the intent of the query when users enter the same query for both very specific and very general information needs.

References

[1] John M. Carroll and Mary Beth Rosson. Paradox of the active user. In John M. Carroll, editor, *Interfacing Thought*, pages 80–111. MIT Press, Cambridge, MA, 1987.

[2] William B. Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval: Data Structures & Algorithms*. Prentice Hall, Englewood Cliffs, NJ, 1992.

[3] Valery I. Frants and Craig B. Brush. The need for information and some aspects of information retrieval systems construction. *Journal of the Americal Society for Information Science*, 39(2):86–91, March 1988.

[4] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. Statistical semantics: Analysis of the potential performance of keyword information systems. *The Bell System Technical Journal*, 62(6):1753–806, July 1983.

[5] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–71, November 1987.

[6] Raymonde Guindon. How to interface to advisory systems? users request help with a very simple language. In Elliot Soloway, Douglas Frye, and Sylvia B. Sheppard, editors, *Human Factors in Computing Systems: CHI '88 Conference Proceedings*, pages 191–96, 1988.

[7] Raymonde Guindon, Kelly Shuldberg, and Joyce Conner. Grammatical and ungrammatical structures in user-adviser dialogues: Evidence for sufficiency of restricted languages in natural language interfaces to advisory systems. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, July 1987.

[8] Donna Harmon. Overview of the first trec conference. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 36–47, 1993.

[9] Carol Keene. Using unrestricted queries to retrieve on-line documentation. In Preparation, 1994.

[10] Thomas K. Landauer, Susan T. Dumais, Louis M. Gomez, and George W. Furnas. Human factors in data access. *The Bell System Technical Journal*, 61(9):2487–509, November 1982.

[11] Gary Marchionini. Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science*, 40(1):54–66, 1989.

[12] Patricia Russo and Stephen Boor. How fluent is your interface? designing for international users. In *INTERCHI '93 Conference Proceedings: Bridges Between Worlds*, pages 342–47, April 1993.

[13] Gerard Salton. *Automatic Text Processing*. Addison-Wesley, Reading, MA, 1989.