Failure Analysis in Query Construction: Data and Analysis from A Large Sample of Web Queries

Major Bernard J. Jansen
Department of Electrical Engineering & Computer Science
United States Military Academy
West Point, New York 10996
dj9395@exmail.usma.army.mil

Dr. Amanda Spink
School of Library and Information Sciences
University of North Texas
Denton, TX 75203 USA
spink@lis.admin.unt.edu

Dr. Tefko Saracevic
School of Communication, Information and Library Studies
Rutgers University
New Brunswick, NJ 08903
tefko@scils.rutgers.edu

ABSTRACT
This paper reports results from a failure analysis (i.e., incorrect query construction) of 51,473 queries from 18,113 users of Excite, a major Web search engine. Given that many digital libraries are accessed via the Web, this analysis points to the need for redesign of the traditional search engine interfaces.

KEYWORDS: Web queries, query construction

INTRODUCTION
Many digital libraries are or will be connected to the World Wide Web (Web). It is therefore important to understand how users are currently searching the Web because they may use the same methods when searching in digital libraries. This research is relevant to design of search engines and interfaces, improving document storage and site design, identifying mega-tag data, and thesaurus development. It also increases understanding of how people are using the Web, and therefore, how they may use digital libraries.

EXCITE
Founded in 1994, Excite, Inc. is a major Internet media company which offers free Web searching and a variety of other services. According to an independent study, "during a 28-day period from Sept. 29, 1997 to Oct. 26, 1997, there were a total of 11,793,000 unique visitors to the Excite Network" (Excite press release, November 17, 1997). While this includes all the visits, in addition to searches, it is safe to assume that the overwhelming number of Excite visits are searches. This provides a picture of the huge size of the traffic on Excite.

We provide only a brief description of Excite search capabilities – more details are on its Web site. Excite searches are based on the exact terms that a user enters in the query, however, capitalization is disregarded.

Digital Libraries 98 Pittsburgh PA USA

Stemming is not available. Multiple term searches return pages with any or all terms in any order somewhere in a page. Results are provided in a ranked relevance order. Searching and ranking of results are done using proprietary algorithms and procedures. A number of advanced search features are available. Those that pertain to our results are described here:

- Boolean operators AND, OR, AND NOT, but these operators must appear in ALL CAPS and with a space on each side. Parentheses can be used for nested Boolean operators.
- A set of terms enclosed in quotation marks (no space between quotation marks and terms) returns answers with the terms as a phrase in exact order.
- A + (plus) sign before a term (no space) requires that the term must be in a document (i.e., a positive infinite weight for that term).
- A – (minus) sign before a term (no space) requires that the term must NOT be in a document (i.e., a negative infinite weight for that term).

The queries examined are a random subset of Excite searches on 10 March 1997. Each transaction record contained three fields: Time of day, User identification, and Query terms. The first field is time of day measured in hours, minutes, and seconds that a user accessed the Excite server measured from midnight of 9 March 1997. The next field is an anonymous user identification assigned by the Excite server, and the third field is the actual query.

BOOLEAN OPERATORS
We examined how many queries explicitly utilized Boolean operators, including nesting, as presented in Table 1. Boolean operators must be upper case. Additionally, to receive the correct result the Boolean operator NOT must be used with AND.

Boolean operators were used very sparingly. Only 5,323 queries, *or about one in every 9 queries contained a Boolean operator,* and in those AND was used by far the most. A minuscule percentage of queries used OR or AND NOT. Only 273 of the total number of 5,323 queries with operators used nested logic – i.e. *only one in*

about nineteen Boolean searches placed some of the terms with operators in parentheses.

Operator	Number of queries	Percent of all queries
AND	4798	8.68
OR	132	0.26
AND NOT	120	0.23
( )	273	0.53

TABLE 1: Use of Boolean Operators in Queries.

We then examined how many of these Boolean queries were incorrect. These results are presented in Table 2. The column Incorrect displays the number of queries containing a specific Boolean operator that were constructed incorrectly. The last column is the percentage of queries containing Boolean operators that were incorrectly constructed.

Of those queries that used the Boolean operators, 1262 or a whooping 26% of uses were incorrect. *About one in every four queries that used Boolean operators or parentheses was not entered as required by Excite.* The very small use of Boolean operators and the very large percentage of mistakes when they are used shows that the Web searchers are not up to Boolean. System or interface redesign seems to be in order.

Operator	Incorrect	Percent incorrect
AND	1262	26.30
OR	46	34.85
AND NOT	79	65.83
( )	88	32.23

TABLE 2: Incorrect Use of Boolean Operators.

From teaching of information retrieval (IR), we know that people have difficulty in making a distinction between the Boolean AND and *'and'* as a conjunction. For example, in the query: *"I am interested in stock trends on the exchanges in New York, London and Tokyo"* the conjunction *'and'* translates into a Boolean OR – that is why we use Venn diagrams to make it clear.

MODIFYING OPERATORS

Excite also permits three modifiers to query terms: '+' (plus), '–' (minus) and quotation marks as defined above. Table 3 shows the occurrence of these modifiers in queries.

Modifiers	Number of queries	Percent of all queries
+	3009	5.85
-	2573	5.00
" "	2507	4.87

TABLE 3: Use of Query Term Modifiers.

The '+' and '-' modifiers were used more than Boolean operators. *Together '+' and '-' were used in 5,570 queries, or in about one in every nine queries.* But a majority of uses were mistakes: 75% of use of these operators was incorrect.

The modifier mistakes are presented in Table 4. The column Incorrect represents the number of incorrect queries with modifiers, and Percent is the percent of queries with modifiers that were incorrect. There were several common errors, including placing blank spaces after the modifiers and not including blank space between another term and before the modifiers.

In particular, the '-' modifier had a high number of mistakes. The same symbol is used in terms that users hyphenated such as *on-line,* so this incorrect percentage may include some hyphenated terms. The ability to create phrases (terms enclosed by quotation marks) was also seldom used – *only one in every twenty queries requested a phrase*, but when phrases were used the mistakes amounted only to some 8%.

Modifiers	Incorrect	Percent incorrect
+	1684	55.97
-	2495	97.42
" "	200	7.98

TABLE 4: Incorrect Use of Query Term Modifiers.

MISCELLANEOUS OBSERVATIONS

Drawing from their experiences with other search engines, some users also used commands not allowed by Excite, such as: NEAR, a proximity operator, was used 19 times, and '*' and '?' as stemming operators were also used a fair number of times, all resulting in mistakes. It seems that users and modifiers do not get well together. Spaces are the biggest source of incorrect entries. For Boolean operators space is required and for modifiers space cannot be used. This probably causes confusion. Also, the modifiers '+' and '-' are analogous to the addition (+) and subtraction (-) symbols in mathematics. Many users placed these modifiers and terms in queries like mathematical formulas (e.g., *digital+libraries-conferences* versus *+digital +libraries -conferences*). This is another feature ripe for redesign.

CONCLUSIONS

The results of the failure analysis emphasize the need to approach design of digital libraries or any system accessed via the Web in a significantly different way than the design of traditional IR systems. The low use of advanced searching techniques would seem to support the continued research into new types of user interfaces, intelligent user interfaces, or the use of software agents to aid users in a much simplified and transparent manner.