WEB QUERY STRUCTURE:
IMPLICATIONS FOR IR SYSTEM DESIGN

Bernard J. Jansen
Computer Science Program
University of Maryland (Asian Division)
Seoul, 140-022 Korea
Email: jjansen@acm.org

Amanda Spink
School of Information Sciences and Technology
The Pennsylvania State University
511 Rider I Building,
120 S. Burrowes St.
University Park PA 16801
E-mail: spink@ist.psu.edu

Major Anthony Pfaff
Department of English
United States Military Academy
West Point, New York 10996

Please Cite: Jansen, B. J., Spink, A., & Pfaff, A. 2000. Web Query Structure: Implications for IR System Design. Proceedings of the 4th World Multiconference on Systemics, Cybernetics and Informatics (SCI'2000). Orlando, USA, July 23-26, 2000.

See Other Publications

ABSTRACT

Translating an information need into a form understandable by an information retrieval system typically requires the use of terms and queries. Terms form the queries for information retrieval systems, and queries are a representation of the user’s information needs that information retrieval systems can understand. Therefore, terms and how they are used in queries are the essential components of a user's interaction with any information retrieval system. By modeling terms, term semantics, and query syntax, one could tailor an information retrieval system to confirm to this model. This may provide assistance to the user in finding relevant information. In a step towards this model, we analyzed a transaction log containing over a 50,000 queries posed by over 18,000 users of *Excite*, a major Internet search service. Web queries are usually two terms. We further isolated basic query structure and syntactic patterns. Based on this analysis, we classified queries into five (5) general categories. Web queries are overwhelming noun phrases, usually in the form of a modifying noun followed by the modified noun. We conclude with the implications of this user model on system design of IR systems.

Keywords: web searching, information retrieval, linguistic, queries

INTRODUCTION

Information retrieval (IR) and Web user modeling is a growing area of research as the realization has increased that the user must be considered part of the complete IR system (Brajnik 1987; Saracevic, Spink, and Wu 1997). Saracevic, Spink, and Wu (1997) reviewed the history and state of user modeling research in traditional IR systems. There is also a growing body of literature focusing on IR in the context of the Web (Jansen, Spink, & Saracevic, 2000; Jansen and Pooch (under review), Lawrence & Giles, 1998; Lynch, 1997). However, many Web studies have focused on user characteristics and empirical analysis of users' queries, with little attention to theory development or theory application.

In this study, we investigate the applicability of linguistic analysis of user Web queries for the improvement of IR and especially Web system. Users of such systems are natural language users. Knowing how natural language users structure their queries, in an attempt to model their information need, may reduce the gap between how a computer works and how the "typical user", (i.e., a user with limited knowledge about how an IR system works) thinks the system works. By analyzing the user queries for structure, syntax, and semantics, we may be able to develop strategies that will benefit future IR system design.

In pursuit of this line of investigation, we analyzed a transaction log from the *Excite* search engine, a major Web media company. This paper reports the methods, findings and results from a linguistic analysis of this corpus of queries from users of the *Excite* search engine.

EXCITE DATA CORPUS

Founded in 1994, *Excite*, Inc. is a major Internet media public company that offers free Web searching and a variety of other services. The company and its services are described at its Web site [http://www.excite.com]. Only the search capabilities relevant to our results are summarized in this paper. *Excite* searches are based on the exact terms that a user enters in the query. Capitalization is disregarded, with the exception of logical commands AND, OR, and AND NOT. Stemming is not available. An online thesaurus and concept linking method called Intelligent Concept Extraction (ICE) is used, to find related terms in addition to terms entered. Search results are provided in a ranked relevance order. A number of advanced search features are available.

Focusing on a term level of analysis, we a term as: any unbroken string of characters (i.e. a series of characters with no space between any of the characters). The characters in terms included everything – letters, numbers, and symbols. Terms were words, abbreviations, numbers, symbols, URLs, and any combination thereof. We counted logical operators in capitals as terms, however, in a separate analysis we isolated them as commands, not terms.

Some general statistics about the data corpus are presented in Table 1.

No. of users	No. of queries	Non-unique terms	Mean of terms And Range	Unique terms with case sensitive	Unique terms without case sensitive
18,113	51,473	113,776	2.21 0-10	27,459	21,837

Table 1. Numbers of users, queries, and terms

DATA ANALYSIS

Terms

There were 113,793 terms (all terms from all queries). After eliminating duplicate terms, there were 21,862 unique terms that were non-case sensitive (in other words, all upper cases are here reduced to lower case). In this distribution logical operators AND, OR, NOT were also treated as terms, because they were used not only as operators but also as conjunctions.

There were 74 terms (of the 21,837 unique terms) that occurred more than 100 times in all queries. On the other end of the spectrum**,** there were 9,790 terms that occurred only once. The 74 terms used 100 or more times had a frequency of 20,698 appearances as search terms in all queries. They represent 0.34 % of all unique terms, yet they account for 18.2 % of all 113,776 search terms in all queries. If we delete the 11 common terms that do not carry any content by themselves (*and*, *of*, *the*, *in*, *for*, *+*, *on*, *to*, *or*, *&*, *a*) that altogether had 9,121 occurrences, we are left with 63 subject terms that have a frequency of 11,577 occurrences – that is 0.29% of unique subject terms account for 10.3% of all terms in all queries.

On the other end of the distribution, the 9,790 terms that appeared only once amounted to 44.78% of all unique terms and 8.6% of all terms in all queries. The tail end of unique terms is very long and warrants in itself a linguistic investigation. However, we could find no comprehension studies of what terms, the distribution of those terms, the modification of those terms, etc. of Web queries. Out of the complete rank-frequency-table we took the top used terms i.e. those that appeared 100 times or more, as presented in Table 2.

Term	Frequency	Term	Frequency	Term	Frequency
And (incl. ‘AND’, & ‘And’)	4828	&	188	estate	123
of	1266	stories	186	magazine	123
the	791	p****	182	computer	122

sex	763	college	180	news	121
nude	647	naked	180	texas	119
free	610	adult	179	games	118
in	593	state	176	war	117
pictures	457	big	170	john	115
for	340	basketball	166	de	113
new	334	men	163	internet	111
+	330	employment	157	car	110
university	291	school	156	wrestling	110
women	262	jobs	155	high	109
chat	256	american	153	company	108
on	252	real	153	florida	108
gay	234	world	152	business	107
girls	223	black	150	service	106
xxx	222	porn	147	video	105
to	218	photos	142	anal	104
or	213	york	140	erotic	104
music	209	A	132	stock	102
software	204	Young	132	art	101
pics	202	History	131	city	100
ncaa	201	Page	131	porno	100
home	196	Celebrities	129		

Table 2: Listing of Terms Occurring More Than 100 Times (**** = expletive).

LINGUISTIC ANALYSIS

In English, the modifying term almost always precedes the term that it modifies, as in the query "red chair." Another example, the term "beautiful" is an adjective. When one hears it, one expects it to always precede the term it modifies. In fact, it would sound odd if an adjective went after the term it modifies, as in "women beautiful." However, this "odd sounding" phrase was an actual query from the data set.

Sometimes, it is not clear to what lexical category a term belongs. Consider the query "soccer team", which was also an actual query from the data set. Which term modifies which? The answer cannot be determined by looking at the form of the terms (as one could with the term "beautiful"), but only by where the terms are placed in the query. In English syntax, the modifying term precedes the term that is modified, we know that "soccer" modifies "team." When a noun, like "soccer" modifies another noun (in this case "team") it becomes an attributive noun. In short, attributive nouns function like adjectives, but they do not have the form of an adjective. In this way, the syntax of the language projects onto the semantics of the expressions allowed by the syntax. With this simplified linguistic base, we now move to results of the lexical analysis.

LEXICAL ANALYSIS

For the purposes of this preliminary work, we performed a lexical analysis of the first 511 queries from the data set. We examined the lexical patterns for individual queries as well as for entire sessions (i.e., the entire series of queries by a particular searcher). All the queries examined used English terms. While a complete analysis will require the examination of a much larger set, some interesting results emerged from this incipient analysis.

Generally, one can say that users do not apply the normal rules of English syntax in any coherent or consistent manner. This is in line with our expectations following our term analysis. Users rely on a variety of lexical patterns to "explain" (i.e., formulate the query) to the "computer" (i.e., the IR system) the information need, item, or topic they are trying to locate.

Even in those sessions where users perform multiple queries, the query patterns often vary widely and seldom conform to the rules of English syntax. From a linguistic point of view, there is no "language" to Web queries. A language must have rules of syntax that permit one to distinguish a well formed from an ill-formed query. There does not appear to be any such syntax with web queries.

While there did not seem to be any grammatical consistency to the queries, the syntax of the queries did fall into five categories. The five categories are listed below, followed by a discussion of each.

- Adjective and noun queries (where one term was modified and the others terms were doing the modifying).
- Complete and grammatically correct English sentences.
- Queries comprised of verbs or verbals.
- Random strings of terms of a variety of lexical categories but which seem to belong to the same category.
- Miscellaneous (i.e., URLs and email addresses).

Adjective and Noun Phrase

This first category was by far the most represented, 458 of the 511 queries. Most of the queries in this category conformed to normal English syntax where the modified term (usually a noun is the last term in the query and the modifying term/s (usually an adjective) are to the left. Additionally, the least restrictive term was usually closest to the modified term and the most restrictive term modifier was farthest away.

For example, in the query "brazillian soccer teams" (sic), the terms "brazillian" and "soccer" modify the term "teams". The term "brazillian" is the more restrictive relative to the modifier "soccer." When a noun, like "soccer" modifies another noun (in this case "team") it becomes an attributive noun. In short, attributive nouns function like adjectives, but they do not have the form of an adjective.

In some cases, the term being modified came first, as in the query "women beautiful." In this case, the user begins with the broadest category and then seeks to modify it into a more specific category. This situation is analogous to a person shopping in a department store. The person goes to the shoe department, then to the running shoes, then to a particular brand of running shoe and so on.

Grammatically Correct

In regards to the second category (14 of 511 queries), almost all queries of this type took the form of a question. Further, almost all took the form of a Wh-phrase. A Wh-phrase is an interrogative phrase that begins with words like what, where, when, how, why, which, and whose. A typical query of this type is: ‘what is empty space in the universe composed of?’ In nearly all of these sentences, the verb almost always had a two-place argument structure, which were usually theta marked as agent and theme or agent and location. This theta-marking pattern is also true of those few phrases that contained a verb.

Theta-marking is a way of delineating what kinds of words can be used as arguments for a particular verb. For instance, the verb *kill* has a two-place argument structure (e.g. The boy killed the deer). This is usually formally represented as Kbb, where K represents the predicate *kill* and the b represents the boy and the d represents the deer. But not just anything can go in those places.

For the verb *kill* one of the arguments must be something that can kill and the other something that can be killed. We can call the first the **agent** and the latter the **patient.** The thematic category limits the lexical category of possible responses. For example, in the case of an agent, it will almost always be a noun phrase such as, "The boy." This means that in the event a word can have more than one lexical category (for example, "play," it can be a verb as well as a noun). Knowing the theta-marking of a particular verb will determine which lexical category the word falls in.

Theta-marking also imparts some semantic information about the word. For example, an agent is almost always a noun phrase, and it also has to be something capable of causing an effect (in this example, death). Additionally, the patient must be something capable of receiving an effect (again, in this case, death).

Verbal Phrase

This category (11 of 511 queries) was queries that contained verbs or verbals, which were not complete, grammatically correct English sentences. Verbals are nouns that have "ing" added to them. Verbals function as participles and/or gerunds. Queries containing verbs were extremely underrepresented giving their abundance occurrence in normal English. The queries containing verbals outnumbered the queries containing verbs six (6) to five (5). In many cases, the verbals stood alone, making it impossible to determine if they were meant as gerunds or participles, (e.g. as with the query ‘hunting’).

Where it was possible to determine, we discovered that most verbals were gerunds. In this category more of the verbs (including the root verbs the verbals were created from) had a two-place argument structure, most of which were theta marked for agent and theme or agent and location. The ones that had only a one-place argument structure were theta marked as agent. A typical example of a verb query was "boy and wolf cried", and an example of a verbal phrase query was "flood plains flooding."

Random Category

The fourth category (13 of 511 queries) contains those expressions that contained a series of words of varying lexical categories and which defied syntactical categorization. The query "‘alicia silverstone’ cutest crush batgirl babysitter clueless" serves as a good, and one of the few non-x-rated, examples of this particular pattern.

In these cases, it is not clear at all that the words are serving the syntactic capacity that one would expect from their position in the query. This query pattern does not conform to a standard, grammatically correct English sentence or phrase nor does it seem to conform to the first query pattern analyzed where one term is modified and the other terms do the modifying. So, while we can pick out the lexical categories of most of the words, that does not help make sense of the expression.

It is also significant that one cannot pick out the lexical category of all the words, for example: "crush." Since the expression does not conform to a standard English syntactical pattern one can not tell if the term is a noun (as in "I have a crush on her") or a verb (as in "I will crush you").

While there does not seem to be a syntactic account for the meaning of this query, there is a semantic one. The terms all seem to relate to a particular movie actress. A human, with the appropriate background, can identify this semantic relationship this because each one of the terms has something to do with the actress Alicia Silverstone, the movies she has made, or the roles that she has played.

Miscellaneous

We have included in the miscellaneous category (15 of 511 queries) any query pattern represented less than ten (10) times. The most prevalent of these are queries concerning URLs, email addresses, and grammatically incorrect English phrases, most being proper names. Since this category is of little interest to a linguistic analysis, we will not include them in the discussion section.

IMPLICATIONS FOR SYSTEMS DESIGN

Several aspects the findings have implications for system design in Web and possibly information retrieval in general. from the above discussion, at least three strategies for system design emerge for addressing the lack of syntax.

Web and IR Systems could "recognize" certain syntactical patterns like those described above. For example, let us look at the *Adjective and Noun Phrases*, where the modified word is last in the series and the modifying words precede it. While this is a simple pattern, it is rich in information. Just by its form, one knows which word contains the category of information the user is seeking, that is the last word in the query. One also knows, of the modifying words, which is most and which is the least restrictive, the first term. A computer can perform this simple evaluation and apply term weighting or suggest general indices of subjects,

In instances where there is a verb, the *Verbal Phrase* category, if the IR system can detect the theta-structure of the verb, it will "know" what kind of item to look for, even if the system cannot tell to what category the item belongs. This is case, the first term of the query could be given the most weight in a term weighting scheme.

For the Random Category, a thesaurus of terms based on some stored dictionary or perhaps collaborative thesaurus based on previous searches could suggest categories to the system. For example, if queries from previous users contained terms such as: "batgirl babysitter clueless" along with "alicia silverstone", the IR system could categorize these terms. In fact, this is similar to how the *Excite* on-line thesaurus works, except *Excite* uses these as terms to suggest to the users. Excite also selects the terms to offer based on the queries of other users.

CONCLUSION

Web and IR systems currently model the user’s information need via the query. However, most Web and traditional search IR engines follow a statistically query term and document term comparison. The premise of this analysis is that if one can correctly model the query, it would be a major step forward in correctly modeling a user’s information need. Previous IR modeling has focused on the user – system discourse, not on the query. Is there a linguist component to IR research? Is there a linguistic identification for query structure? It appears that there is some basic syntactic structure to queries. User modeling should also into account the syntax and semantic of the query. Syntax can provide information on the meaning of query terms.

REFERENCES

Brajnik, G., Guida, G., & Tasso, C. (1987). User Modeling in Intelligent Information Retrieval. *Information Processing and Management 23*, 305-320.

Croft, W. B, Cook, R., & Wilder, D. (1995). Providing Government Information on the Internet: Experiences with THOMAS. *Proceedings of Digital Libraries '95 Conference* (pp. 19-24).

Jansen, B. J. and Pooch, U. (Under Review) Web user studies: A review and framework for future work. Submitted to the *Journal of the American Society of Information Science*.

Jansen, B. J., Spink, A., & Saracevic, T. (2000) Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*. *36*(2), 207-227.

Lawrence, S. , & Giles, C.L. (1998). Searching the World Wide Web. *Science, 280*(5360), 98-100.

Lynch, C. (1997). Searching the Internet. *Scientific American, 276,* 50-56.

Saracevic, T., Spink, A., & Wu, M. M. (1997). Users and Intermediaries in Information Retrieval: What are they talking about? *Proceedings of the Sixth International Conference on User Modeling* (pp. 43 – 51).

Spink, Wolfram, Jansen, Saracevic (Under Review). Searching the Web: the public and their queries. Submitted to Journal of the American Society of Information Science.