



COVER SHEET

Spink, Amanda and Ozmultu, H.Cenk (2002) Characteristics of question format web queries: an exploratory study. *Information Processing and Management* 38(4):pp. 453-471.

Copyright 2002 Elsevier.

Accessed from: <http://eprints.qut.edu.au/archive/00004756>

Information Processing & Management, 2002, 38(4), 453-471

CHARACTERISTICS OF QUESTION FORMAT WEB QUERIES: AN
EXPLORATORY STUDY

Amanda Spink*
School of Information Sciences and Technology
The Pennsylvania State University
511 Rider I Building, 120 S. Burrowes St
University Park PA 16801
Tel: (814) 865-4454 Fax: (814) 865-5604
E-mail: spink@ist.psu.edu

H. Cenk Ozmultu
Dept. of Industrial Engineering
The Pennsylvania State University
University Park PA 16802
E-mail: hco100@psu.edu

* To whom all correspondence should be addressed.

ABSTRACT

Web queries in question format are becoming a common element of a user's interaction with Web search engines. Web search services such as Ask Jeeves — a publicly accessible question and answer (Q&A) search engine — request users to enter question format queries. This paper provides results from a study examining queries in question format submitted to two different Web search engines – Ask Jeeves that explicitly encourages queries in question format and the Excite search service that does not explicitly encourage queries in question format. We identify the characteristics of queries in question format in two different data sets: (1) 30,000 Ask Jeeves queries and (2) 15,575 Excite queries, including the nature, length, and structure of queries in question format. Findings include: (1) 50% of Ask Jeeves queries and less than 1% of Excite were in question format, (2) most users entered only one query in question format with little query reformulation, (3) limited range of formats for queries in question format — mainly “where”, “what”, or “how” questions, (4) most common question query format was “Where can I find.....” for general information on a topic, and (5) non-question queries may be in request format. Overall, four types of user Web queries were identified: keyword, Boolean, question, and request. These findings provide an initial mapping of the structure and content of queries in question and request format. Implications for Web search services are discussed.

INTRODUCTION

Effective query processing is a major challenge for Web search services. Increasingly, Web querying services, such as Ask Jeeves [<http://www.askjeeves.com>], encourage users to enter queries in question format. An underlying assumption of such Web search engines as Ask Jeeves may be that users find expressing their queries in natural language question queries less difficult to construct than keyword or Boolean queries. Based on this trend, systems designers are working on the development of more effective question query processing techniques (Agichtein, Lawrence & Gravano, 2001; Prager, Brown, Coden & Radev, 2000) for Web and IR systems. Much IR and Web research has focused on the analysis of keyword and Boolean queries (Spink & Saracevic, 1997; Spink, Wolfram, Jansen & Saracevic, 2001). Therefore, studying the characteristics of users queries in other format, such as question format, is an important and growing field for the development of more effective “question and answer” (Q&A) access to the Web.

Unlike Ask Jeeves, general Web search services, such as Excite, are not currently encouraging users to submit queries in question format. However, questions or requests for information by a user are an element within a dialogue-based approach to modeling user-Web/information retrieval (IR) system interaction. This research also falls within the general framework of research that attempts to model the interactive Web search process. Web user modeling research is part of a larger body of user modeling research in artificial intelligence (AI) and cognitive science (Kobsa & Wahlster, 1989). To develop a Web dialogue-based model, a first necessary step is the development of a grammar of Web users' interaction and to identify the elements within this grammar. Queries in questions format can be regarded as one syntactic structure of a grammar of

Web interaction in need of further exploration. Queries in question form a component of a dialogue-based model of Web IR and the potential functioning of intelligent interfaces that assist users with Web IR tasks and building interactive processes the way humans normally conduct them. Web search engines that ask users to “ask questions”, such as Ask Jeeves, are increasing in popularity. In this paper we report findings from a study comparing the prevalence of question format queries to Ask Jeeves - a Web search engine that encourages users to enter question format queries, with users’ question format queries to the Excite Web search engine that does not encourage question-format queries. We first provide an overview of related research, summarize our research objectives and research design, and then report the findings of our analysis.

RELATED STUDIES

A growing number of studies have examined the nature of queries to Web search engines. Large-scale studies (Jansen, Spink & Saracevic, 1999; Silverman, Henzinger, Marais and Moricz, 1999; Spink, Wolfram, Jansen and Saracevic, 2000) show that most Web users enter few queries consisting of few search terms, conduct little query reformulation and have difficulty developing effective keyword or Boolean queries. Web search engines process millions of queries daily. On average Web search engine users enter few queries and few search terms. Many Web users find Boolean logic difficult to master as a form of query construction and create many failed Web queries.

Increasingly, new types of Web querying services, such as Ask Jeeves [<http://www.askjeeves.com>], are requesting users to enter queries in the form of questions. An underlying assumption of the Ask Jeeves approach to Web searching is that users find expressing their queries in natural language question queries less difficult to construct than Boolean queries. Few studies have examined the nature of queries in question format to Web search engines. Jansen, Spink, Pfaff and Goodrum (2000)

conducted a linguistic analysis of Excite users queries contained in a 1997 data set and identified less than 1% of queries in elicitation format or requests for information. However, with the emergence of a more question and answer (Q&A) approach to Web querying, the nature of users' queries in question format are becoming important and significant to the development of more effective Web IR systems.

Question asking or elicitation behavior, as requests for information, forms a significant part of day-to-day human communication. Questions are a significant part of interpersonal communication studies examining models of discourse analysis (Cicourel, 1980), models of dialogue (Levinson, 1981); turn taking during conversations (Sacks, Schegloff & Jefferson, 1974), and replies and responses (Goffman, 1976). Kearsley (1976) studied the replies and responses process and identified taxonomy of major question functions, including echoic, expressive, epistemic, and social control. Information retrieval (IR) research and models have increasingly adopted language from communication studies to conceptualize interactive IR as a conversation within a dialogue framework.

With the increasing popularity and associated advertising by Web search services such as Ask Jeeves, we examined the nature and extent of users' queries to the Excite Web search engine in the form of a question. Using data from the Ask Jeeves and Excite Web search engines we can compare the structure of any queries in the form of questions. We sought to examine how many Excite users were entering question queries and the structure of those queries.

RESEARCH QUESTIONS

The purpose of our study was to gain a greater understanding of Web question format queries. We sought to answer the following research questions:

1. What is the prevalence of Ask Jeeves and Excite queries in question format?
2. What is the mean number of terms per question format query?
3. What are the common starting term(s) for question format queries?
4. Are Boolean operators or modifiers used in question format queries?
5. Are question marks used in question format queries?
6. What are the common subjects for question format queries?

Our analysis should reveal data on the nature and extent of Web queries in question format to assist in the design of more effective Web retrieval systems.

RESEARCH DESIGN

Excite Data Set

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 in more detail at its Web site (<http://www.excite.com>). Excite searches are based on the exact terms that a user enters in the query; however, capitalization is disregarded, with the exception of logical commands AND, OR, and AND NOT. There is no stemming. An online thesaurus and concept linking method called Intelligent Concept Extraction (ICE) is used to find related terms in addition to terms entered.

The Excite data set we analyzed consisted of a transaction log of 1.7 million action records from 20 December 1999 containing four fields, that were:

Identification: an anonymous code assigned by the Excite server to a user machine.

Time of Day: measured in hours, minutes, and seconds from midnight of 20 December 1999.

Number of Pages Viewed: the number of pages containing 10 Web sites viewed by the user.

Query: the query terms exactly as entered by the user.

Our analysis focused on the question format queries in the data set. The data analyzed included users sessions, queries, and term level of analysis. A *session* is the entire sequence of queries by a particular user. A *query* is a set of one or more terms entered into the Web IR system during a single search. A *term* is any string of characters bounded by white space.

Question Format Query Classification

To identify the question format queries in the data set of 1.7 million queries, we created a program to automatically identify any query ending with a question mark (?) and any query beginning with one of a list of words commonly associated with human question asking, including: where, what, who, how, when, can, are, is, may, which, has, does, did, will, has, could should, and do. We qualitatively analyzed a subset of the question format queries to determine: (1) the number of terms per question query, (2) the subject of the question format query, the starting term for the question, and (3) classified the question format query as either a query for factual information or a more general query for information. The classification of question query formats is in line with the verificative and topical oriented types of information needs empirically documented by Ingwersen (1996).

Question Query Sessions

A subset of 100 Excite user sessions including at least one question format query were identified and qualitatively examined to analyze the:

1. Subject of the question format query session
2. Number of question format queries per session
3. Number of terms per question format query session
4. Number of pages of 10 web sites viewed by the question format query user
5. Number of question format queries per user session

6. Mean number of terms per question format query
7. Position of the question format queries in the user session
8. Use of Boolean operators or modifiers in question format queries
9. Starting term(s) of question format queries
10. Use of a question mark

Ask Jeeves Data Set

The data set analyzed consisted of a random subset of 30,000 queries selected from a transaction log of 800,000 queries submitted to Ask Jeeves on 20 December 1999. Our statistical and qualitative analysis focused on the 30,000 randomly selected queries, or 3.75% of the data set, including queries and terms. As with the Excite analysis, we created a computer program to automatically identify any query ending with a question mark (?) or beginning with one of the common words associated with human questioning: e.g., where, what, who, how, when, can, ... etc. Similarly to the Excite analysis, we qualitatively analyzed a subset of the question format queries to determine: (1) the number of terms per question format query and (2) the starting term(s) for the question format query.

RESULTS

This paper extends preliminary findings reported in Spink, Milchak, Sollenberger and Hurson (2000).

Number of Queries in Question Format

Table 1 summarizes the results of the question format query analysis for the Excite and Ask Jeeves data sets.

[Place Table 1 Here]

The Excite data set of 1.7 million queries contained 15,575 (or approximately 1%) queries in question format. Interestingly, only half the Ask Jeeves queries were queries in question format. The Excite data set contained some 1% of queries in question format that was not requested by the search engine. However, queries in question format are still a small part of Excite querying. Overall, our analysis also revealed four query types: keyword, Boolean, question, and request. Some 9430 Ask Jeeves non-question queries (64%) were request queries, in the form of requests for information, commands, or orders. Request queries are examined in more detail later in the paper.

Terms Per Question Format Query

We first examined the number of terms used per Excite and Ask Jeeves question format query (Figures 1 & 2).

[Place Figure 1 Here]

The mean number of terms per Excite question format query was 7.8. This query length is much longer than the 2.4 terms for non-question format queries or a mean of 14 terms for mediated online searching (Spink & Saracevic, 1997; Spink, Wolfram, Jansen & Saracevic, 2001). Question format queries also contain common function words such as “a”, “the” and “are” normally occurring in question phrases. Non-question format queries are usually constructed as a adjective/noun (nominal) phrases (Jansen, Pfaff, Spink & Goodrum, 2000). For Ask Jeeves, we examined the number of terms per query, including non-question format queries (Figures 2).

[Place Figure 2 Here]

For Ask Jeeves queries, the distribution of the number of terms for all queries has two peaks and appears to be a combination of two different distributions - non-question format queries with a mean of 3 and the mean number of terms per question

format query was 7. This compares to the mean number of terms per Excite question format query of 7.8. This query length is much longer than the 2.4 terms per query previously reported for non-question format Web queries (Jansen, Spink & Saracevic, 2000).

Use of Boolean Operators/Modifiers

Previous research shows that few Excite users submit queries with Boolean operators or modifiers (Jansen, Spink & Saracevic, 2000; Spink, Wolfram, Jansen & Saracevic, 2001) (Table 2).

[Place Table 2 Here]

Table 3 shows the number of Ask Jeeves question format queries that included Boolean operators/modifiers.

[Place Table 3 Here]

Few Ask Jeeves users included Boolean operators in their queries in question format. The most common inclusions were full stops at the end of a question format query, use of “And” and the use of quotation marks “ “ in some queries in question format around a person's name. Since the Ask Jeeves search engine allows natural language entries, “and” and “or” is mostly used as in their original meaning not as Boolean operators. Relevance feedback use was also low at 0.02% when compared to 9.7% in the percentage of relevance feedback by Excite users (Jansen, Spink & Saracevic, 2000). Further research could compare the presentation and explanation of relevance feedback by both search engines.

Question Format Query Starting Term

How did Ask Jeeves and Excite users structure their question format queries? We compared the starting terms for question format queries (Table 4).

[Place Table 4 Here]

In both data sets, the most common term used to begin a question format query was either “where”, “what” and “how’: Approximately,

- 1 in 2 question format queries began with the term “where”
- 1 in 4 question format queries began with the term “what”
- 1 in 6 question format queries began with the term “how”

Overall, users enter a limited range of question query formats.

Questions Format Query Starting Terms

Table 5 shows the starting terms of Excite queries in question format for 100 sessions and Table 6 shows the starting terms for Ask Jeeves queries in question format for 11,743 sessions. Figure 3 shows some commons question formats.

[Place Table 5 Here]

[Place Table 6 Here]

[Place Figure 3 Here]

Where Queries

Most commonly users phrased the beginning of their question queries as “Where can/would I find/get/buy...../”. An example query is “Where can I find information on cats?” A “where can I find” question implies a desire to find a place or physical location for the information desired. However, few users asked “which” web site contained the information they sought. Questions beginning with the term “where” may be similar to the opening words for many questions to reference librarians. Most users framed their questions as requesting the location of information on a particular topic. Users may be conceptualizing “where” in terms of web site location.

What Queries

“What” was the second most popular form of question starting term and the terms “what is” was the second most popular form of question. Using the word “what” was usually associated with a request for a specific or factual piece of information. An example query is “What is the name of the leader of Ghana?”

How Queries

Some users began questions with the term “how”. They were generally asking how to find something. An example query is “How do I find information on gardening?” These users seemed to be asking for instructions to assist them in their information seeking. The nature of the opening phrase for a question query can change the nature of the question and the response to the question. For example the query: “Where can I find information on peanut growing?” requires a different response than to the question “How do I grow peanuts?” or “What is a peanut” that seeks more specific information. A question such as “Do you have the new book on peanut growing by Tim Hancock” requires a simpler yes or no answer followed by the location of a publication.

Opinion Queries

Some users requested an opinion from Ask Jeeves on their personal problems. For example, one questioner asked: “Should I talk during sex?” This type of question requires a complex and more psychologically based response. Another questioner asked: “Was lincoln a great leader?” Many questions reflected a human need for opinion, communication or help with some personal information issue or problem. One user made a passionate plea for help from Ask Jeeves: “My kid won’t stop biting people?”

Personified Queries

A few users actually addressed Ask Jeeves as a human being by saying: “Hey Jeeves” or “Jeeves....”. Some users requested help by saying: “Help me Jeeves”. Some users were polite and phrased their query as “May I....” or “Please...”. However, most

users were not that polite, particularly those who entered a request as apposed to a question format query, discussed later in this paper. When you analyze a large log of Ask Jeeves question queries, you realize that many people don't really understand the Web search process. They ascribe human abilities to Ask Jeeves that go way beyond its current capabilities. Many users do not really understand how a Web search engine works and their own information seeking and searching processes. Users are often frustrated and emotional during their Web search engine interactions.

Top 75 Query Terms

Spink, Wolfram, Jansen and Spink (2001) analyzed the top 75 query terms for the large 1999 Excite data set (Table 7).

[Place Table 7 Here]

To compare the top 75 terms by frequency appearing in question format queries we analyzed the Ask Jeeves data set (Table 8).

[Place Table 8 Here]

There were 19,775 unique terms in the Ask Jeeves data set among the total number of 160,498 terms. Since natural language is common even in the non-question format queries, it may be expected that terms such as "I", "can", "the", "of", "a", "is", "in", "for", "on", "to", etc. be in the list of top occurring terms. Those terms are essential for any kind of natural language query. As expected, in contrast to the top 75 Excite terms, most of the top 75 Ask Jeeves terms are the words needed to generate meaningful sentences such as "I", "can", "where", etc. that form part of questions or requests. The terms "I", "find", "get", "me" and "need" were also part of request queries.

Question Format Query Subject Terms

Ask Jeeves Subject Terms

Table 9 shows the top 30 question format query subject terms.

[Place Table 9 Here]

Ask Jeeves users were often asking for pictures and free products or services. Also, due to the pre-Christmas nature of the data set (December 20, 1999), it is not unexpected that the term Christmas appeared with high frequency. Also, the terms sex and nude are not unusual as high frequency terms (Jansen, Spink & Saracevic, 2000).

Excite Subject Terms

We also examined the subject categories for Excite users' question format queries. These are listed in Table 10.

[Place Table 10 Here]

In general, the subject of the Excite question format queries were scattered across a broad range of subjects. Approximately,

- 1 in 5 users were seeking information on people, places or things.
- 1 in 6 users were looking for information related to commerce, employment, travel or economic issues.
- Less than 1 in 10 questions related to sex, pornography or preferences.

Question Format Query Reformulation

We could not judge the success of particular queries beyond the number of pages (10 Web sites per page) viewed. However, to determine the level of query reformulation we analyzed a subset of 100 Excite user sessions containing question format queries. A comparable analysis could not be conducted on the Ask Jeeves data set, as the Ask Jeeves queries could not be grouped in sessions. The analysis at the Excite user sessions that included at least one question query gave us a different picture of question format query usage (Table 11).

[Place Table 11 Here]

The sample of 1000 user sessions that included question format queries, were fairly short and included only one question format query. Question format queries were usually the first and only query submitted during the user sessions. There was little query modification during these question format query sessions. Even though users submitted queries in question format, few reformulated the question format query into a keyword or Boolean query, or even reformulate their initial query based on the results of the retrieval. Some Excite users entered a Boolean or keyword query and subsequently reformulated it into a question format query – this behavior is difficult to explain. Also, the 6.4 terms per query was somewhat shorter in this sample than the 7.8 terms per query for the larger Excite data set.

Use of Question Marks

In both the Excite and Ask Jeeves data sets, only half the users included a question mark at the end of their question format queries. The inclusion of a question mark is a normal pattern in the writing of a question. However, it was interesting to see that nearly 50% of users did not include a question mark at the end of their question. Maybe, they realized that the search engine would not process the question mark?

REQUEST FORMAT QUERIES

One in three Ask Jeeves non-question queries were in request, command, or order format. We did not analyze the request format queries in the Excite data. As request format queries appeared to be a major component of Ask Jeeves non-question queries, Table 12 shows the starting term for those Ask Jeeves queries in request format.

[Place Table 12 Here]

Instead of the common question format query “Where can I find”, requests are generally in the form “Find me....”, “I need”, “I want”, “Get me”, “Give me”, “Show me”, or “I am looking”. Some users make fairly interesting requests, such: (1) “Want to hear farts” that could require quite a bit of query interpretation and audio response, and (2) “Take me to Mexico City” that implies possible transportation capabilities.

The difference between question and request format query users is an area for further research. Differences may relate to users perceptions of and expectations of the search engine or how users generally interact with other people when looking for information, their interaction style or gender. Some users may prefer to use a question format and some a request format. Request queries were shorter and more concise than question format queries. Some users may feel more comfortable using a request rather than a question format. One could speculate that males may be more likely to enter a request/command or order query than females, given their often more instrumental rather than relational patterns of speech. Females may be more likely to ask questions in a more relational style. As we have no data on the gender of each Ask Jeeves query, this is an area for further research.

Specific Fact Versus More General Information Queries

We also examined whether the Excite question or request was for a specific fact or for more general information (Table 13).

[Place Table 13 Here]

Nearly 40% of users were looking for factual answers. An example factual query is: “How much is first class postage to England?” Many Excite users were looking for more general information on a topic. An example query is: “Where can I learn more about peasants/serfs in the middle ages?” Answers to this type of question would be

more involved and complex than simple factual answers. We did not conduct this the factual versus general information analysis on the Ask Jeeves data set.

DISCUSSION

Web search engine users generally enter four types of queries: keyword, Boolean, question, and request. Overall, most Web question format queries are about 7 terms in length, and non-question/request queries are less than 5 terms long, and contain few Boolean operators or modifiers. The form in which information seekers express themselves is fairly limited when using either question or request query format. When users expressed themselves in the form of questions they generally asked either “where”, “what”, or “how” questions. The most common form of question format query begins with the words “Where can I find.....” for general information on a topic. Much less frequently do they ask “which”, “when”, or “does” questions. They are sometimes likely to ask for subjective opinion and more likely to request directions to information. The most common form of request format query was “Find me information on.....”. One could speculate that the use of a question or request format queries may relate to gender differences in male and female personal communication or interaction styles.

The appropriate analysis and response to a user's question or request query is a fairly complex task beyond the stripping down to a simple adjective/noun (nominal) phrases for processing against the database of Web sites. We are quite a way off from replicating a reference interview via Web question and answer (Q&A) interaction, but increasingly such a Q&A is becoming the interaction mode of choice for Web users. Effective question and request query processing is a major challenge for Web system designers. Not only are question and request queries longer than keyword or Boolean queries, but also they are more complex than just an adjective or noun phrase. The nouns and adjectives in a query are important. However the first few words used to begin a question or request query are also important as they set the initial context and

expectations of the user in relation to the processing of the noun and adjective phrases making up the rest of the query.

Despite the current low use of question format queries by Excite users, it was interesting to see the nature of question format queries presented to the search engine. There was little query reformulation by the users despite the fact that Excite does not process question format queries as questions. A few users entered a question format query and then seemed to realize they should reformulate into a Boolean-type query. Another small group of users began with a Boolean-type query structure and then decided to reformulate their query into a question. However, most users entered only one question format query and then examined the results. The low query reformulation rate during question format query sessions was consistent with other studies of non-question Web querying (Spink, Jansen & Ozmutlu, 2000). Users seemed to take what they were given in Web sites in response to a question. In a traditional reference interview, a reference librarian generally answers a question with another question. Therefore, Excite users are engaging in the first steps of a reference interview without the appropriate "librarian-type" response from the Web system.

Despite the frequently broad nature of the information requests, the form in which information seekers express their requests are fairly limited to a request for a location. When information seekers expressed their information requests in the form of questions they generally asked either "where", "what", or "how" questions. They are seeking to find a location, but do not use the terminology "Find me a website on....". Much less frequently did they ask "which", "when", or "does" questions - they were less likely to ask for subjective opinion and more likely to request directions to information as they may in a library.

Many information seekers were looking for more general information beyond a simple factual answer. The Web system's response to a more general information

request needs to be more complex than for a factual question. In the context of a reference interview, a more general information request would normally engender an interchange between information seeker and reference librarian to refine the information seeker's request. As we can see from the studies by Spink, Goodrum and Robins (1998), a reference interview preceding a mediated online database search includes a lengthy conversation to clarify the information seeker's real needs. However, the appropriate analysis and response to a user's question query is a fairly complex task beyond the stripping down to a simple adjective or noun phrase for processing against the database of Web sites. We are quite a way off from replicating a reference interview via Web question and answer (Q&A) interaction, but increasingly such a Q&A is becoming the interaction mode of choice for Web users.

CONCLUSIONS AND FURTHER RESEARCH

Overall, there seems to be some common patterns of Web question format query structure. These limited patterns of question and request query structures need to be tested further in other data sets. An interesting point for further research is whether question format queries or non-question format queries contain more significant terms? How can the structure of question or request format queries be improved for more effective retrieval?

In addition, further research is needed to relate question and request query construction to users' gender, communication style, or interaction style. A common element of database, IR systems, and Web search engine research is the human querying process. Further research is also needed to highlight the similarities and differences that occur when users interact with different types of structured and unstructured data over different types of retrieval systems, such as online public access catalogs (OPACs), Web search engines, IR systems, and databases.

ACKNOWLEDGMENTS

The authors would like to thank Jack Xu from Excite, Inc for access to the data set examined in this study. We also acknowledge the financial support from the Penn State WISER program (Women in Science and Engineering Research) that funded our study. The authors would also like to thank the reviewers of this paper for their suggestions that improved the paper.

REFERENCES

- Agichtein, E., Lawrence, S., & Gravano, L. (2001). Learning search engine specific query transformations for question answering. *Proceedings of the 10th World Wide Web Conference, May 1-5, 2001, Hong Kong.*
- Cicourel, A. V. (1980). Three models of discourse analysis: The role of social structure. *Discourse Processes, 3*, 101-132.
- Goffman, E. (1976). Replies and responses. *Language in Society, 5*(3), 257-313.
- Ingwersen, P. (1996). Cognitive analysis and the role of the intermediary in information retrieval. In: Davis, R., ed. *Intelligent Information Systems*. Chichester, West Sussex, England: Horwood (pp. 206-237).
- Jansen, B. J., A. Spink, A. Pfaff, & Goodrum, A. (2000). Web query structure: Implications for design. *SCI 2000: Systematics, Cybernetics & Informatics, July 2000, Orlando - Florida.*
- Jansen, B. J., Spink, A., & Saracevic T. (2000). Real life, real users and real needs: A study and analysis of users queries on the Web. *Information Processing and Management, 36*(2), 207-227.
- Kearsley, G. P. (1976). Elicitations and elicitation asking in verbal discourse: A cross-disciplinary review. *Journal of Psycholinguistic Research, 5*(4), 355-375.

Kobsa, A., & Wahlster, W. (eds.), *User Models in Dialog Systems*. New York: Springer (1989).

Levinson, S. C. (1981). Some observations on the modeling of dialogue. *Discourse Processes*, 4, 93-116.

Prager, J., Brown, E., Coden, A., & Radev, D. (2000). Question-answering by predictive annotation. *Proceedings of ACM SIGIR, July 2000, Athens, Greece* (pp. 184-198).

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. *Language*, 50, 696-735.

Saracevic, T., Spink, A., & Wu, M. M. (1997). Users and intermediaries in information retrieval: What are they talking about? *Proceedings of the 6th International Conference on User Modeling, Chia Laguna, Sardinia, Italy – June 2-5, 1997* (pp. 44-54).

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, 33, 3.

Spink, A., Goodrum, A., & Robins, D. (1998). Elicitation behavior during mediated information retrieval. *Information Processing and Management*, 34(2/3), 257-274.

Spink, A., Jansen, B. J., & Ozmutlu, H. C. (2000). Use of query reformulation and relevance feedback by Web users. *Internet Research: Networking Applications and Policy*, 10(4), 317-328.

Spink, A., Milchak, S., Sollenberger, M., & Hurson, A. R. (2000). Elicitation queries to the Excite Web search engine. *CIKM 2000: 9th International Conference on Information and Knowledge Management. Washington DC. November 2000* (pp. 134-140).

Spink, A., & Saracevic, T. (1997). Interaction in information retrieval: Sources and uses of search terms. *Journal of the American Society for Information Science*, 48(8), 728-740.

Spink, A., D. Wolfram, B. J. Jansen, & T. Saracevic, (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science*, 52(2), 226-234.

Table 1: Summary of the question query data.

	Excite Data	Ask Jeeves Data
Total data set	1.7 million queries	800,000 (30,000 random sample)
No. of question queries	15,575 (1%)	15,431 (51.4%)
No. of non-question queries	99%	14,569 (48.6%) 9,430 request queries (64% of non-question queries)

Figure 1. Terms per Excite question format query.

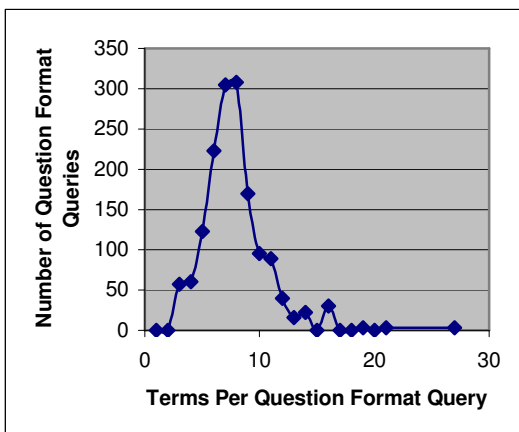


Figure 2. Terms per Ask Jeeves query (including non-question format queries).

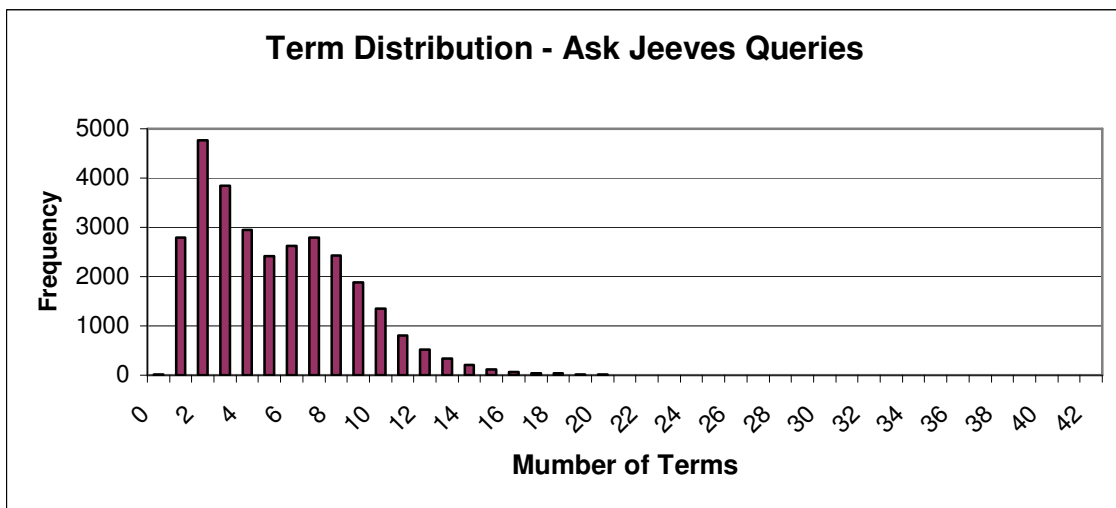


Table 2. Use of advanced search features in Excite queries. (Number of queries = 1,025,910)

Feature	Number of Queries	% of Queries
AND/and/And	29,146	3%
OR/or/Or	1,149	1%
NOT/AND NOT	307	0.0003%
+ plus (correct)	17,028	2%
+ plus (incorrect)	27, 292	3%
+ plus total	44,320	5%
- minus (correct)	1,656	0.001%
- minus (incorrect)	20,295	2%
- minus total	21,951	2%
“ “ (quotations)	52,354	5%
'!' (periods)	51,804	5%
':' (colons)	1,459	1%
&	3,342	3%

Relevance feedback	99,033	9.7%
--------------------	--------	------

Table 3: Use of Boolean operators/modifiers in Ask Jeeves queries.

Boolean Modifier	Boolean Operator	Frequency	%
. period		1127	3.75
And		803	2.6
incorrect -		598	1.99
"" quotation		336	1.12
&		131	0.43
Or		138	0.4
incorrect +		115	0.38
correct +		56	0.18
: colon		47	0.15
correct -		31	0.103
	AND	31	0.103
	NOT	22	0.07
	OR	6	0.02
relevance feedback		6	0.02
Total		3447	11.3

Table 4: Comparison of starting term for Excite and Ask Jeeves question format queries.

Starting Term	Number of Excite Question Format Queries	%	Number of Ask Jeeves Question Format Queries	%
Where	825	53%	8603	48.3%
What	360	23%	3144	17.6%
How	192	12%	2204	12.3%
Who	43	2%	708	3.9%
Can	23	1%	949	5.3%
Is	22	1%	688	3.8%
Why	20	2%	258	1.4%
When	17	1%	234	1.3%
Are	10	0.5%	219	1.2%
Which	9	0.5%	77	0.4%
Do	7	0.5%	411	2.3%
Does	6	0.5%	148	.08%
Did	4	0.5%	20	.01%
Will	3	0.5%	17	.09%
Has	2	0.5%	18	.01%
Was	2	0.5%	53	.02%
Could	1	0.5%	4	.002%
Should	1	0.5%	24	0.01%
Total	1547	100%	17,779	100%

Table 5. Starting terms for Excite question format query sessions.

Starting Terms	Number of Sessions	% of Sessions
Where can/would I find/get/buy	46	46%
What is	25	25%
How do I/you	4	4%
Who was	3	3%
What are the	3	3%
Are	3	3%
Where is/does	2	2%
Who are the	2	2%
When did	2	2%
What kind	1	1%
May I	1	1%
Is there a	1	1%
Do you have	1	1%
Why are	1	1%
Which	1	1%
Will	1	1%
Has	1	1%
What does	1	1%
What causes	1	1%
Total	100	100%

Table 6. Starting terms of Ask Jeeves question query sessions

Starting Terms	Number of Sessions	% of Sessions
Where can I find	4945	42.1%
What is	1562	13.3%
Where can I find information	981	8.3%
Where can I find information about	481	4%
Where can I buy	462	3.9%
What are	414	3.5%
How do I	344	3%
Where is	333	2.9%
Where can I get	298	2.5%
Who is	231	2%
What is a	227	2%
How to	225	2%
Who was	195	1.6%
What does	165	1.4%
How do you	156	1.3%
Where can I see	132	1.2%
Where do I find	123	1%
How many	121	1%
Is	121	1%
Are	114	1%
I need	113	1%
Total	11743	100%

Figure 3. Common Ask Jeeves questions (based on first four words).

<i>Where can I find</i>	Example: "Where can I find the lyrics to Feliz Navidad?"
<i>Where can I buy</i>	Example: "Where can I buy a world globe?"
<i>Where can I get the Internet?"</i>	Example: "Where can I get a free fax program on
<i>Where can I see</i>	Example: "Where can I see six little monkeys?"
<i>Where can I learn</i>	Example: "where can I learn the history of my last name?"
<i>Where do I find make a</i>	Example: "Where do I find information on how to barrier reef aquarium?"
<i>Where can I download zip files"</i>	Example: "where can I download unzip program for
<i>Where can I locate</i>	Example: "Where can I locate Boobzilla?"
<i>What is drop?"</i>	Example: "What is the date of the first atomic bomb
<i>How can I find</i>	Example: "How can I find the spread of a stock?"
<i>How do I find my</i>	Example: "How do I find someone with e-mail address in hometown?"
<i>Can you find me</i>	Example: "can you find me info on v-tec engines"

Table 7. Listing of 75 most frequently occurring terms within 531,416 unique queries (Excite treats everything as lower case).(p**** = expletive).

Term	Frequency	Term	Frequency	Term	Frequency
And	21385	naked	1968	web	1366
Of	12731	american	1961	history	1359
Sex	10757	stories	1958	video	1356
Free	9710	software	1908	sports	1351
The	8013	games	1904	california	1345
nude	7047	diana	1885	men	1327
pictures	5939	p****	1876	national	1306
In	5196	black	1823	big	1290
university	4383	on	1813	york	1277
Pics	3815	photos	1799	texas	1276
Chat	3515	jobs	1735	porno	1263
For	3431	world	1734	maps	1256
adult	3385	a	1711	employment	1234
women	3211	magazine	1690	city	1222
New	3109	nudes	1690	canada	1204
Xxx	3010	news	1687	playboy	1197
Girls	2732	football	1627	car	1195
music	2490	page	1591	erotic	1189
porn	2400	computer	1533	weather	1184
To	2265	princess	1461	map	1159
Gay	2187	airlines	1409	internet	1156
school	2176	download	1381	international	1113
home	2150	real	1381	high	1113
college	2043	education	1376	star	1110
state	2010	art	1374	asian	1110

Table 8: Top 75 Ask Jeeves query terms.

Term	Frequency	Term	Frequency	Term	Frequency
I	8548	get	587	internet	195
Can	7629	my	477	pics	193
Where	7427	does	458	make	188
Find	5951	online	440	see	188
The	5122	from	423	some	186
Of	3299	site	417	address	185
A	3236	web	417	naked	184
What	3007	new	392	games	179
Is	2885	an	376	york	179
In	2385	me	345	map	176
For	2303	have	340	list	170
How	2025	info	302	it	169
On	1996	with	292	game	166
To	1519	was	272	women	164
Do	1359	nude	268	at	162
information	1271	sex	260	computer	162
You	1155	website	258	city	160
About	981	when	241	did	159
And	872	out	240	car	157
Are	861	that	224	codes	157
Pictures	789	best	214	name	155
Free	722	there	208	By	153
Who	656	why	207	check	153
Buy	625	music	198	were	152
christmas	591	download	196	need	151

Table 9: Top 30 Ask Jeeves question format query subject terms.

Term	Frequency	Term	Frequency
pictures	789	women	164
Free	722	computer	162
christmas	591	city	160
online	440	car	157
nude	268	codes	157
Sex	260	name	155
music	198	picture	149
Pics	193	phone	148
address	185	people	145
naked	184	state	144
games	179	world	142
york	179	history	141
Map	176	home	137
List	170	recipe	137
game	166	number	135

Table 10. Subject categories of Excite question format queries.

Subject Category	Number of Question Format Queries	% of Question Format Queries
People, Places, Things	355	22.9%
Commerce, Travel, Employment, Economy	212	13.7%
Health, Science	178	11.5%
Computers, Internet, Technology	175	11.3%
Entertainment, Recreation	167	10.7%
Education, Humanities	118	7.6%
Sexual, Pornography	113	7.8%
Government	55	3.5%
Society, Culture, Ethnicity, Religion	51	3.2%
Performing & Fine Arts	47	3%
Crafts, Cooking, Do-It-Yourself	36	2.3%
Indiscernible	40	2.5%
Total	1547	100%

Table 11. Structure of Excite question format in 1000 query sessions.

Mean number of queries	2.1
Mean number of Excite pages viewed	2.8
Mean number of question format queries per session	1
Mean number of terms	6.4

Table 12: Starting terms: Ask Jeeves request format queries.

Term	Frequency	Term	Frequency
Find	5951	use	72
Buy	625	know	69
Get	587	take	67
download	196	looking	65
Make	188	purchase	63
See	188	give	56
Check	153	shop	45
Need	151	sell	43
Tell	110	write	43
Want	97	say	42
Show	91	build	40
Help	89	happened	39
Look	81	start	39
Go	80	wrote	38
Search	74	order	36

Table 13. Excite: Fact or information question.

Type of Question Format Query	Number of Question Format Queries	% of Question Format Queries
Fact	595	38.4%
Information	952	61.6%
Total	1547	100%