Features of Documents Relevant to Task- and Fact-Oriented Questions

Diane Kelly[a], Vanessa Murdock[b], Xiao-jun Yuan[a], W. Bruce Croft[b] & Nicholas J. Belkin[a]

[a]School of Communication,
Information & Library Studies
Rutgers University
New Brunswick, NJ 08901

[diane, yuanxj, belkin]@scils.rutgers.edu

[b]Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts
Amherst, MA 01003

[vanessa, croft]@cs.umass.edu

ABSTRACT
We describe results from an ongoing project that considers question types and document features and their relationship to retrieval techniques. We examine eight document features from the top 25 documents retrieved from 74 questions and find that lists and FAQs occur in more documents judged relevant to task-oriented questions than those judged relevant to fact-oriented questions.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software – Performance evaluation

General Terms
Measurement, Performance

Keywords
Query Classification, Question Types, Task-Oriented Questions

1. INTRODUCTION
Consider the following two questions, "How long does it take to get a passport?" and "How do I get a passport?". The type of information requested by each question is quite different. Information likely to satisfy the former question will be short and factual, such as "98 months," while information likely to satisfy the latter question will be longer and might consist of a set of instructions, a description of a process or a form.

Quite often the type of information requested by a user of an information retrieval system is about a process, or how to perform a task. Most document retrieval systems treat all requests uniformly, regardless of their grammar, orientation or form. Indeed, it has traditionally been the case that users are required to request information with a query consisting of keywords. Specifying a query as a question is becoming a more and more common mode of input; a growing body of research is dedicated to question-answering systems [c.f. 1]. Allowing users to express information needs as questions rather than keywords provide users with a better opportunity to express the type(s) of information needed to resolve the information problem.

The work presented in this paper is part of an ongoing research project whose aim is to 1) distinguish between different questions according to type of information requested; 2) identify salient features of documents relevant to each question type and 3) design retrieval techniques that consider these differences. The first of these issues is addressed in [4], where we investigated automatic classification techniques for question types using a language modeling approach. The results from this work provide evidence for the distinction between task- and fact-oriented questions. In this paper, we address issue 2), by gathering relevance judgments of documents retrieved for both task- and fact-oriented questions from real users and examining the features of each set of documents.

Section 2 describes our relevance assessment study. Section 3 presents the results from the analysis of document features of relevant and non-relevant documents for task and fact questions. Section 4 presents our conclusions and plans for future work.

2. RELEVANCE ASSESSMENTS
To identify relevant and non-relevant documents for both task and fact questions, we conducted a human relevance assessment study, where we asked participants to evaluate the relevance of documents retrieved for both types of question. Study participants were asked to consider six different questions and judge the relevance of a set of documents to each of these questions.

2.1 Question Collection
Our corpus of questions came from the query logs of the Govbot[1] search engine. Govbot allowed users to access government information on the web and indexed primarily documents from the .gov and .mil domains. The query logs contained over a million queries, both in the form of keywords and questions. While most Govbot queries were one or two words long, there were a number of well-formed questions, approximately 4 in every 1000 queries. We identified queries that started with question words (who, what, where, when, why, how) as question-queries and selected these for further investigation.

The average length of questions was 8 words. After pre-processing, we had 4100 unique questions, with 3700 unique words pertaining to government. From this set, we selected randomly a list of 120 questions for our study. Questions were classi-

---

[1] In operation at the Center for Intelligent Information Retrieval, University of Massachusetts (1995–2001).

fied manually as either fact or task, depending upon the type of information expected. The reliability of this classification was independently validated by four people. The classification of all but 7 of 120 questions was agreed upon by all four people. These 7 questions were eliminated from the set.

2.2 Document Collection

Each of the 113 questions was submitted to GovBot, which uses Inquery [3] to retrieve documents. We followed each returned URL to ensure that all links functioned properly and were accessible. From our efforts, we obtained the top 25 unique documents retrieved by GovBot for each question. We further screened the list of questions for currency and document availability until we had 80 total questions, 40 task and 40 fact, which we then used in our relevance assessment study.

2.3 Participants and Procedures

We solicited participants from a graduate course in Library and Information Science to judge the relevance of documents for each question. Each participant was required to evaluate 25 documents for each of six different questions. In total, each participant evaluated 150 documents. Of the six questions, three were task and three were fact. Participants completed the study online at the location of their choice. Participants were not required to complete the entire study within one session, but were required to complete evaluations of all 25 documents for any given question during a single session. There was no time limit for making the evaluations. Participants were presented with four choices for relevance, 1) *Relevant*: the information on this web page (not a linked page) satisfies the query; 2) *Partially Relevant*: the information found on this web page satisfies the query only in part; 3) *Not Relevant*: the information on this web page does not satisfy the query; and 4) *Unsure*: unable to determine the relevance based upon the information on this web page.

3. RESULTS

For practical reasons, not all 80 questions were judged. Each question was judged by one to three participants. The results that we present in this section are based on evaluations of 74 questions (37 fact and 37 task) made by 23 participants. Documents were scored as relevant if the majority of judgments were relevant. Documents were scored as non-relevant if the majority of judgments were non-relevant. A document that was rated as relevant by one participant and non-relevant by another was not included in the analysis. A document that was rated as partially relevant by one participant and relevant by another was included in the partially relevant set. Thus, the partially relevant set includes documents judged as relevant as well as partially relevant.

We examined the distributions of term frequencies, and tf.idf weights for the relevant and non-relevant documents for task and fact questions to obtain a baseline for comparison. Figures 1 and 2 show the distributions for relevant and non-relevant documents. As expected, the distributions of the tf.idf scores are similar.

We further examined the relevant and non-relevant documents for task and fact questions for the presence of eight document features: lists, tables, FAQs, forms, downloadable files, question terms present in special markup, links, and length. Table 1 shows the average occurrence of these features in relevant, partially relevant and non-relevant documents for each question type; the results for each feature are discussed in greater detail below.
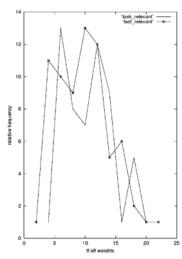

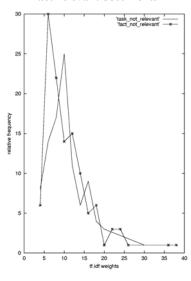**Figure 1. Distributions of tf.idf scores for task-relevant and fact-relevant documents**

Figure 2. Distributions of tf.idf scores for non-relevant documents

3.1 Lists and Tables

We considered lists to be ordered lists, unordered lists, and definition lists. We also included tables in this feature category because many people use tables to format their pages in place of list tags. We acknowledge that tables are used for other purposes as well, but make no effort to distinguish between these uses. To identify each feature we detected the following HTML tags: <OL>, <UL>, <DL>, and <TABLE>.

Documents relevant to task questions have more ordered and unordered lists in them (8.15) than those relevant to fact questions (3.19). Documents not relevant to task questions have fewer (4.27) ordered and unordered lists than documents not relevant to fact questions (2.76). Definition lists are more prevalent in pages relevant to fact questions (.59) than in those relevant to task questions (.32). There are no differences for the number of table tags in pages relevant to either question type.

3.2 FAQs

We considered a document to be an FAQ if the number of questions (sentences ending with a question mark) appearing on the document reached a minimum number (N=5). We were unable to detect FAQs where question marks were omitted and we did not count documents that contained the word "FAQ" or pointed to documents containing FAQs. The appearance of FAQs in task and fact relevant documents is similar (.16 and .10, respectively).

Table 1. Average number of features for relevant and non-relevant documents, for task and fact questions

Feature	Relevant		Partially Relevant		Non-relevant	
	Task	Fact	Task	Fact	Task	Fact
Unordered Lists	7.65	3.12	6.27	3.01	3.89	2.56
Ordered Lists	.50	.07	.47	.07	.38	.20
Definition Lists	.32	.59	.25	.62	.14	.29
Tables	3.42	5.04	3.75	4.85	2.29	4.42
FAQs	.16	.10	.17	.14	.10	.05
Forms	.04	.10	.07	.11	.13	.15
Downloads	.24	.27	.22	.55	.78	.26
Special Markup	2.97	2.55	2.52	2.33	.89	1.40
Links	32.26	43.72	36.26	44.78	29.26	30.6
Length (bytes)	38251	26522	35252	27653	24842	23495

3.3 Forms and Downloadable Files

Forms were counted by first identifying the <FORM> tag and then screening for those that were application forms or government forms. In many cases, the <FORM> tag is used in search fields and we wanted to exclude these elements. It is often the case that forms appear as downloadable files. We attempted to count forms appearing in this manner by averaging the number of downloadable files per document. We found that downloadable files are more common in pages not relevant to task questions (.78) than those not relevant to fact questions (.26).

3.4 Question Terms in Special Markup

We examined the occurrence of task and fact question terms in numerous HTML tags used for special markup including those used for creating headings, emphasizing text, hyper-linking text and formatting text. We also examined the contents of the <title> tag. The occurrence of question terms in special markup are present more often in documents relevant to both task and fact questions (2.97 and 2.55) than those documents not relevant (.89 and 1.40) to either type.

3.5 Links

We found that the average number of links per page for task-relevant pages (32.26) was less than for fact-relevant pages (43.72). Overall, non-relevant documents for both task and fact questions had fewer links per page (29.26 and 30.6, respectively) than their relevant counterparts.

3.6 Document Length

We measured document length as the complete text of the document, including HTML tags. Task-relevant documents were longer (38251) than fact-relevant documents (26522). Of course, this does not necessarily mean there is more relevant information in the task-relevant documents; rather, it may just be that it took more text, or more markup tags, to express the content of the document.

4. CONCLUSIONS AND FUTURE WORK

We have shown in our previous work [4] that there is a measurable difference between task questions and fact questions. We were able to classify questions as fact or task in three ways: 1) based on question words, 2) using grammatical structure, and 3) training language models. Our current work demonstrates a difference between the features of relevant pages for task and fact questions. We found that lists occur more often in documents relevant to task questions, FAQs are more common in task questions, and links are more common in documents relevant to fact questions. In addition, documents relevant to task questions are longer, on average, than other documents. There were no differences in the tf.idf weights, downloadable files, question terms appearing in special text, or presence of forms or tables of documents relevant to task and fact questions.

Our work provides evidence that the information requested by a question can differ depending on task or fact-orientation. This suggests that retrieval techniques specific to each type of question should be considered. If documents relevant to task questions share a common set of features, it may be possible to exploit this feature set to improve retrieval. It may also be the case that the most relevant part of the document is found within the feature. For instance, with lists it may be possible to focus passage retrieval to the list elements to aid in question answering; we leave this to future work.

In other future work, we would like to test the techniques presented here on many more questions, and further refine the classification described in [4] by training on larger data sets. Further, we would like to investigate whether page ranking algorithms as in [2] are appropriate for task questions.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] Abney, M. C. S., and Singhal, A. Answer extraction. In *Proceedings of ANLP '00*.

[2] Amento, B., Terveen, L, and Hill, W. Does "authority" mean quality? Predicting expert quality rating of web documents. In *Proceedings of SIGIR '00*, 296-303.

[3] Callan, J., Croft, W. B., and Harding, S. The INQUERY retrieval system. In *Proceedings of the International Conference on Database and Expert Systems Applications '92*, 78-83.

[4] Murdock, V. and Croft, W. B. Task orientation in question answering. In *Proceedings of SIGIR '02*, 355-356.