The Effects of Topic Familiarity on User search behavior in Question Answering Systems

Azzah Al-Maskari
Dept. of Information Studies
University of Sheffield
Sheffield, S1 4DP, UK
Lip05aaa@shef.ac.uk

Mark Sanderson
Dept. of Information Studies
University of Sheffield
Sheffield, S1 4DP, UK
m.sanderson@shef.ac.uk

Abstract

This paper reports on experiments that attempt to characterize the relationship between users and their knowledge of the search topic in a Question Answering (QA) system. It also investigates user search behavior with respect to the length of answers presented by a QA system. Two lengths of answers were compared; snippets (one to two sentences of text) and exact answers. A user test was conducted, 92 factoid questions were judged by 44 participants, to explore the participants' preferences, feelings and opinions about QA system tasks. The conclusions drawn from the results were that participants preferred and obtained higher accuracy in finding answers from the snippets set. However, accuracy varied according to users' topic familiarity; users were only substantially helped by the wider context of a snippet if they were already familiar with the topic of the question, without such familiarity, users were about as accurate at locating answers from the snippets as they were in exact set.

1 Introduction

Over the past few years, the question answering tracks at the Text Retrieval Conferences (TREC) Vorheese (2002, 2003, 2004) have brought formal and rigorous evaluation methodologies to bear on the question answering task. Although the importance of these evaluations cannot be denied, they measure performance of complex systems that typically involve a combination of information retrieval, information extraction, and natural language processing technologies. Järvelin and Ingwersen (2004) assert that the real issue in information retrieval systems design is not whether recall/precision goes up by a statistically significant percentage, but rather whether it helps the user solve the search task more effectively or efficiently. Therefore, the issue in Information Retrieval (IR) shifts from maximizing the retrieval performance by refining IR techniques and methods to maximizing the understanding of users' behaviors and information need representation during retrieval. Thus there is a great demand to consider the knowledge of users' behavior to be the key solution to successful retrieval.
In QA systems, the users' decision about the correctness and usefulness of an answer mainly depends on the context in which possible answers appear in addition to the users' previous knowledge of the topic of question.

This paper explores if there is a measurable difference in task performance using *exact* or *snippet* sets as an answer sets in a QA system; judges if there is a preference for exact over snippets retrieval or vice versa; and assesses if there is a difference in users' search behavior when they are searching exact or snippet sets (i.e., accuracy, level of confidence and number of answers to be displayed). This paper is divided as follows: in section 2 relevant previous works on answer-on-context is described, section 3 explains the experimental approach followed to conduct this user test and the results found, section 4 draws conclusions and future research.

2 Related Research

Previous studies have examined users' preferences for the length of answers from a QA system. Lin

and Mitamura (2004) explored the roles of context in QA systems in four different options; exact answer, answer-in-sentence, answer-in-paragraph, and answer-in-document. They reported that users like answer-in-paragraph condition best and the exact answer condition the least. López-Ostenero et al. (2004) and Peinado et al. (2005) found that users prefer passage retrieval to document retrieval when searching standard document and passage retrieval engines in an interactive Cross-Language QA system. In contrast, the finding of Figuerola et al. (2004) demonstrated that users appraised document retrieval over passages for QA systems. A more recent study by Navarro et al. (2005) also showed that users considered larger context are better than shorter ones by comparing a passage retrieval system versus a clause retrieval system in an interactive cross language QA system.

Previous research has also examined user topic familiarity. Kelly and Cool (2002) reported on users searching an IR system, showing that as one's familiarity with a topic increases, searching efficiency increases and reading time decreases. In addition, Shiri and Revie (2003) investigated the effects of topic familiarity and topic complexity on cognitive and physical moves in a thesaurus-enhanced online IR system. They defined cognitive moves as user's conceptual analysis of terms or documents; and physical moves – those associated with the use of system features. Their findings indicated that an increased number of cognitive and physical search moves were associated with more complex topics. It was also observed that users searching moderately familiar and very familiar topics used more cognitive and physical moves than users searching for unfamiliar topics, though this difference was not statically significant. It was suggested that contextual factors, such as topic familiarity and task, affected the rate of occurrence of these behaviors. While is has been acknowledged in these studies that topic familiarity is an important factor influencing information seeking, no one appears to have considered QA systems and user topic familiarity. The work in this paper focuses on identifying information search behaviors that might be related to topic familiarity.

3 Experimental Approach

To start the experiments, 92 factoid questions were randomly selected from the TREC QA track and issued to an in-house QA system, AnswerFinder[1]. AnswerFinder was first tested by issuing 199 TREC questions; 92 factoid, 51 definition and 56 list. As shown in Table 1, AnswerFinder performed best at answering factoid questions, thus for this experiment only factoid questions were used.

	Factoid	Definition	List
Correct	63%	6%	15%
Not exact	1%	6%	9%
Wrong	22%	15%	13%
No answer	6%	23%	18%

Table 1: AnswerFinder Performance

The answers were manually assessed following an assessment scheme similar to the answer categories in iCLEF 2004:

- *Correct:* answer string is valid and supported by the snippets.
- *Non-exact:* answer string is missing some information, but the full answer is found in the snippets.
- *Wrong:* answer string and the snippets are missing important information or both the answer string and the snippets are wrong compared with the answer key.
- *No answer:* system returns no answer.

Table 1 illustrates AnswerFinder's effectiveness.

Answer Type		
Correct	65	68.5%
Non exact	1	1.1%
Wrong	19	6.5%
No Answer	7	23.9%
Total	**92**	**100.0%**

Table 2: Overall View of AnswerFinder Performance.

Procedure and Design of the User's Tests

To conduct the user test 44 master students participated in judging the TREC questions. These users were of differing nationalities and

[1] A web-based question answering tool Greenwood (2004). http://www.dcs.shef.ac.uk/~mark/phd/software/

backgrounds. Each participant was asked to judge AnswerFinder's answers to 20 different questions; 10 questions of snippets set and 10 of the exact set, each user was allotted 15 minutes to complete their task. Figure1[2] and Figure2[3] typify snippets and exact sets.

When did the royal wedding of Prince Andrew and Sarah take place?

1. 1986
   It was great fun when it did take place and you can have a look by clicking ... to the summer of **1986** and the nuptials of Prince Andrew and Sarah Ferguson.
2. July 23
   Finally, on November 20, 1947, the long-awaited Royal wedding took place . ... Prince Andrew and Sarah were married at Westminster Abbey on **July 23**, ...
3. 1992
   ... the royal marriage at Windsor's Guildhall, the ceremony will now take place on ... The Duchess of York, who was divorced from Prince Andrew in **1992**, ...
4. November 20, 1947
   Finally, on **November 20, 1947**, the long-awaited Royal wedding took place .
5. April 8th
   The wedding will take place on **April 8th** .

Figure 1: Answer Type: Snippet set

When did the royal wedding of Prince Andrew and Sarah take place

1. 1986
2. July 23
3. 1992
4. November 20, 1947
5. April 8th

Figure 2: Answer Type: Exact answer

A Latin-square matrix design was adopted to minimize the effects of user-specific, question specific and order-related factors on the tasks. To prevent any learning effect masking any system effect, every question was introduced to the participants in one set only, half of the participants were given the exact set first while the other half were given the snippets set first. Figure 3 shows a sample presentation order.

User	Search Order (condition: A|B, question 1….100)																			
1	1	18	4	21	17	6	30	44	11	3	49	68	58	78	52	73	79	67	60	55
2	51	82	63	60	80	45	76	54	65	27	2	7	21	29	44	17	30	19	15	5
3	2	5	19	18	38	37	31	28	15	35	62	69	73	87	81	77	80	88	70	59
4	74	52	46	79	65	48	78	83	72	64	10	35	31	20	4	36	29	87	44	8
5	12	41	23	36	2	26	32	24	22	5	70	71	82	73	47	66	87	85	83	62
6	59	75	84	51	47	61	85	82	46	58	9	14	36	32	37	22	33	23	18	7
.																				
.																				
42	69	56	88	65	72	46	79	83	97	50	19	28	23	4	14	7	92	21	41	10
43	93	54	73	85	48	61	66	58	82	92	16	3	42	5	36	26	43	39	13	97
44	11	25	15	8	29	20	40	33	94	99	86	74	65	80	50	70	54	58	63	81

Figure 3: Questions Distribution: the shaded area means the snippets set and the white area means the exact set

The Test Measures

The following measures were used to examine users' search behavior and topic familiarity in QA system; these measures are explained in the succeeding sections:

- Participants' accuracy in identifying the correct answers in each set.
- The effect of question familiarity on participants.
- The effect of answer sets, snippets and exact, on participants.
- Confidence of participants in their judgments in each set.
- Number of answers preferred by the participants in QA systems.

The Accuracy of the Participants in Identifying the Correct Answer

Users' effectiveness was measured by the "correct answer" identified by each participant that was checked against an answer sheet (created by the evaluator). Table 3 illustrates the participants' overall accuracy in judging the exact and the snippets sets; accuracy was measured as the ratio of correct answers identified to the total number of questions. According to t-test (significant test), there is no significant difference between judging the snippets and the exact sets.

	Correct	Wrong
Exact Set	53.3%	46.7%

[2] This shows how AnswerFinder originally produces the answers

[3] This is the modified exact answer set taken from Figure 1.

Snippets Set	59.3%	40.7%

Table 3: Users' accuracy in identifying "correct answer"

An examination of failure cases was made: in some cases, the answers were clearly presented in the snippets but the participants did not choose the right answer this could be due to their understanding or lack of knowledge of the topic. Therefore, judgments for some questions have entailed human errors and there exists legitimate differences of opinions and perceptions among the participants. In addition, in some cases the QA system provided an incomplete or incoherent piece of text which did not help the users to deduce accurate answers. Thus, the low accuracies are due to both the system and the participants.

The Effect of Questions' Familiarity on the Performance of Participants

Participants completed a post search questionnaire, which determined their familiarity with each question on three scales:

1) Personal knowledge: if users know the answer to the question before looking at the given choices.
2) No idea: if they have no idea about the answer.
3) Topic is familiar but they don't remember the answer.

Tables 4, 5 and 6 depict how the participants' knowledge affected their choices of answers. The general results showed that accuracy varied with respect to topic familiarity; when users were more familiar with a topic their accuracy in identifying answers was high. Perhaps unsurprisingly when users knew the answer, their accuracy in identifying it correctly was similar in the exact and snippet. Table 5 shows users were better able to locate correct answers in the snippets if they had previous knowledge of the topic. Table 6 illustrates similarity in users' accuracy in both sets when users lacked knowledge about the topics. This was surprising as it was presumed such a question was one where users needed the most help.

Set	Correct	Wrong
Exact	75.2%	24.8%
Snippet	73.6%	26.4%

Table 4: User knows answer

Set	Correct	Wrong
Exact	53.3%	46.7%
Snippet	62.7%	37.3%

Table 5: User is familiar with topic

Set	Correct	Wrong
Exact	44.9%	55.1%
Snippet	47.7%	52.3%

Table 6: User lack the familiarity and knowledge

The Effect of the Snippets Set in Identifying the Correct Answers

Participants completed a post search questionnaire, which assessed if the snippets had helped them in identifying the correct answer. More than half of them agreed that the snippets frequently assisted them in identifying the correct answers. However 22.2% asserted it did not help and 20% thought that it only helped occasionally.

Participants were also asked about their preference of the exact and the snippets sets; 24.4% of them favor the exact set while 75.6% prefer the snippets set. The participants justified the reason for their choices; the group who chose the snippets explained it gave more detail and insight about the answers, which resulted in users feeling more confident. On the other hand, the group who chose the exact said the snippets did not help them to pick an answer, but it rather confused them for the same reason identified earlier. They further confirm their contentment with the exact answers for factoid questions because factoid questions do not require explanation. To summarize, the snippets set helped participants to answer the questions in many cases albeit in some cases it failed.

Apparently this shows different preferences among participants as Schamber (1994) states that for text retrieval, different people have different opinions about whether or not a given document should be retrieved for a query.

Confidence of the Participants on their Judgments

Participants' confidence was considered an important facet of their assessment; different judgments indicated different confidence. Participants were required to rate their confidence after judging each question. The users' confidence ranged from 1 (not at all confident) to 5

(completely confident). Table 7 shows users' confidence with snippets was higher than with exact. The majority of participants complained about the difficulty of choosing an answer from the exact set because they lacked the knowledge about the questions' topics.

Confidence Level	Snippets	Exact
completely confident	28.4%	15.8%
Very confident	15.6%	8.4%
fairly confident	15.1%	16.4%
Slightly confident	14.4%	20.0%
Not Confident	26.4%	39.3%

Table 7: participants' confidence in both sets

Number of Answers Preferred by the Participants

In the final test, participants were asked about the number of answers they deemed suitable for QA systems. According to Table 8, five answers are reasonably enough; fewer answers are also acceptable by some users.

No. of Answers	
1	6.7%
2	6.7%
3	20%
4	20%
5	35.6%
more than 5	8.9%
10	2.2%

Table 8: No. of answers preferred by the participants

Participants who chose one or two answers claimed that for factoid questions only one or two answers were sufficient and more answers may lead to confusion. The users who preferred more than five answers wanted a chance to compare and verify the results especially when they did not know the topic. Thus, the less knowledgeable they are about the topic, the more answers to be displayed.

4 Conclusion

A user test was conducted to investigate a QA system. It was concluded that participants preferred and obtained higher accuracy in finding answers from the snippets than from exact, although there was no statistically significant difference in accuracy between the two. The general trend suggests that the longer context, the better the users' accuracy. Some information search behaviors were also discussed such as accuracy and confidence which vary with respect to topic familiarity. Accuracy was found to increase with topic familiarity; the more familiar participants were with a topic, the more accurate their answers. Participants' familiarity with the topic was boosted by the context provided by the snippets. Nevertheless, with no previous familiarity of the topic, users' ability to locate correct answers was similarly poor for both exact and snippet sets.

QA systems are built to fulfill the goal of fact-finding by providing users with short answers quickly and concisely. However, according to this study and the previous studies, users are not always satisfied with short answers and they often prefer longer context to verify the accuracy of the answers. Thus, it can be concluded that it is difficult to establish a fixed context for QA systems.

For future research, it is recommended to build an interface with an answer space that enables users to navigate information clusters i.e., to view the exact answer, the passage or the full document. This is to assess how far each user searches in the information clusters to detect how much context users really want in inclusive of their knowledge level.

A further study could combine information context with cultural and physical contexts by taking a broader viewpoint, not only the analysis and evaluation of the system performance, but also include contextual and situational factors such as users and their knowledge levels, information needs; work-tasks, characteristics and types.

References

FIGUEROLA, C. G., ANGEL F. ZAZO, BERROCAL, J. L. A. & ALDANA, E. R. V. D. (2004) REINA at iCLEF 2004. *iCLEF*.

JÄRVELIN, K. & INGWERSEN, P. (2004) Information seeking research needs extension towards tasks and technology. *Information Research*, 10(1).

KELLY, D. & COOL, C. (2002) The Effects of Topic Familiarity on Information Search Behavior.

Joint Conference on Digital Libraries (JCDL), 74-77.

LIN, F. & MITAMURA, T. (2004) Keyword Translation from English to Chinese for Multilingual QA. *Association for Machine Translation in the Americas, AMTA* Washington, USA.

LÓPEZ-OSTENERO, F., GONZALO, J., PEINADO, V. & VERDEJO, F. (2004) Interactive Cross-Language Question Answering: Searching Passages versus Searching Documents. *iCLEF*.

NAVARRO, B., MORENO-MONTEAGUDO, L., NOGUERA, E., VÁZQUEZ, S., LLOPIS, F. & MONTOYO, A. (2005) "How much context do you need?" An Experiment about Context Size in Interactive Cross-language Question Answering. . *iCLEF*.

PEINADO, V., LÓPEZ-OSTENERO, F., GONZALO, J. & VERDEJO, F. (2005) UNED at iCLEF 2005: Automatic Highlighting of Potential Answers. *iCLEF*.

SCHAMBER, L. (1994) Relevance and information behavior. *ARIST*, 3-48.

SHIRI, A. A. & REVIE, C. (2003) The effects of topic complexity and familiarity on cognitive and physical moves in a thesaurus-enhanced search environment. *Journal of Information Science,* 29**,** 517.