Do users with different domain knowledge select different sets of documents?

Xiangmin Zhang[1] Jingjing Liu[2], Xiaojun Yuan[3], Michael Cole[4], Nicholas Belkin[4], Chang Liu[4],
xiangminz@gmail.com,jliujingjing@gmail.com,xiaojunyuan@gmail.com,m.cole@rutgers.edu,
belkin@rutgers.edu,changl@eden.rutgers.edu
[1]Wayne State University, Detroit, MI 48202 [2]University of South Carolina, Columbia, SC 29208
[3]SUNY at Albany, Albany, NY 12222 [4]Rutgers University, new Brunswick, NJ 08901

ABSTRACT

In this paper, we report findings on an examination of the document selecting behaviors of people with different levels of domain knowledge (DK). We found that people with high and low DK levels generally select different sets of documents to view; those high in DK read more documents and gave higher relevance ratings for the documents they view than those low in DK. The reasons of the findings, as well as implications on systems design, are discussed.

Keywords

Domain knowledge, user behaviors, relevance judgment

INTRODUCTION

As searching has become a social activity, involving people from different ages and different levels of knowledge, one goal of designing highly effective search systems is to make searching by those who are less knowledgeable as effective as those who are more knowledgeable. To achieve this goal, it is desirable to understand how people with different levels of knowledge actually find and value the retrieved documents on the same topic.

There has been much research on the difference between novice and experts in terms of information seeking, and previous studies have found that users with higher level of domain knowledge have different search tactics (querying behaviors), performance (result accuracy, etc.), and time spent on task accomplishment and document reading, and so on (e.g., Wildemuth, 2004; White, Dumais, & Teevan, 2009; Zhang, Anghelescu, & Yuan, 2005). However, little research has examined users' document selecting and viewing activities. In this paper, we address the following research questions:

1) Do people with different levels of domain knowledge select the same set of documents for the same search topic?
2) What would be the differences in selecting behaviors between the two groups of people?
3) If they have selected the same set of documents, do they have similar relevance judgments on the same document (related to the same search topic)?

Answers to these questions will provide a foundation for personalization of search methods based on users' domain knowledge levels.

METHOD

Experiment Design

A controlled lab experiment was conducted, using an experimental search system based on Indri, in the genomics domain, using TREC dataset and topics. A total of 40 participants (the log data of two were missing) were invited individually to an interaction lab, working on four out of five topics in assigned task orders. Users were asked to rate their familiarity levels of the selected MeSH terms that were in the topic areas. Questionnaires were used before and after each task to elicit users' self-judged knowledge ratings. Users' interaction with the search system was logged.

Knowledge Assessment

The participant's DK level was calculated as follows. We first calculated the average of the MeSH term ratings for each participant. Then correspondingly the average of the pre-task topic familiarity and expertise ratings were calculated. Because the MeSH ratings and pre-task questionnaire ratings used different scales, the average ratings for MeSH and pre-task questions were standardized as Z scores, which are standard deviations from the mean (Kachigan, 1991). These two Z scores for each participant were consequently averaged as DK for each participant:

$$DK_{user} = \frac{\left(Z\left(meshK_{user}\right) + Z\left(\frac{\left(familiarity_{user} + expertise_{user}\right)}{(2)} \right) \right)}{(2)}$$

Proceedings of HCIR 2012, Cambridge MA, Oct 4-5, 2012

Based on the knowledge scores, users were divided into two groups. Those whose scores were above the median (value was -.155) were placed in a high DK group and those below were in a low DK group.

Document Selecting and Relevance Assessment

For each topic, participants were asked to save the documents that they thought relevant during the searching process. At the end of the search for the topic, they would evaluate the saved documents, making relevance judgments.

RESULTS

General Viewing Behaviors of the Two DK Groups

Non-parametric tests were used in this study. Specifically, Mann-Whitney test for the comparison between two groups and Kruskal-Wallis test for comparisons among three and more groups were used.

The total number of documents retrieved and viewed by all participants was 507. Among the 507 documents, approximately 29% (149) were viewed only by the low DK level participants, and approximately 47% (236) were viewed by the high DK level participants only. The remaining 24% (122) were viewed by both groups of participants. This distribution is listed below:

- Set 1 (N=149, 29%): docs viewed only by low DK users
- Set 2 (N=236, 47%): docs viewed only by high DK users
- Set 3 (N=122,24%): docs viewed by both groups

In general, higher DK people viewed significantly more documents than low DK people (*Z*=2.019, p=.043).

Comparisons of Ranking Positions of the Selected Documents by Different Groups of Users

We further examined the ranking positions of documents returned by the search system. Categorizing the viewed documents into two groups: viewed by high DK and those viewed by low DK, no significant difference was found between the two groups in terms of average ranking position and highest ranking position. (Table 1):

	Mean (SD)		Mann-Whitney Z (p)
	Low DK	High DK	
Average rank of viewed docs	23.57 (26.54)	26.86 (29.18)	.83 (.41)
Highest rank of viewed docs	9.18 (8.92)	11.20 (11.49)	1.46 (.15)

Table 1. Rankings of documents viewed by two groups of people

However, if the viewed documents are separated into three sets as outlined earlier, with those viewed by both groups of people as a third set (Set 3), results showed that in terms of the average rank, documents in Set 2 (viewed only by high DK) were lower than those in set 3, while documents in Set 1 (viewed only by low DK) did not have significant differences with the other two sets of documents. This means that the documents viewed only by high DK people had a lower rank on average than the documents viewed by both groups of people. It seems that high DK people still tended to view documents even though their rankings were low in the search results.

	Mean (SD)			Kruskal-Wallis H (p)
	Set 1	Set 2	Set 3	
Average Rank	17.58 (21.70)	22.57 (27.19)	12.79 (12.45)	**6.45 (.040)**

Table 2. Average ranks of three sets of documents

Documents Viewed on Different SERPs

For the Set 3 documents, we also looked at the average ranks of the documents that were viewed by different groups of people. A paired t-test found no significant difference: the average rank position viewed by low DK people was 12.10, and that by high DK people was 13.15 (t(121)=-.617, p=.539).

		Low DK	High DK	Mann-Whitney Z(p)
SERP 1	Count	360	444	--
	Rank	3.82 (2.59)	4.44 (2.62)	**3.45 (.001)**
	Frequency	8.11 (6.45)	8.35 (7.26)	-.25 (.80)
SERP 2	Count	66	120	--
	Rank	14.42 (2.74)	14.47 (2.65)	.19 (.85)
	Frequency	5.94 (7.21)	6.63 (7.39)	.86 (.39)
SERP 3 and above	Count	78	145	--
	Rank	41.92 (22.93)	47.80 (26.27)	**1.97 (.049)**
	Frequency	2.51 (2.72)	3.27 (3.76)	1.43 (.15)

Table 3. Documents viewed by two groups of people per different SERPs

Another way to look at the document ranking positions is to check their positions on search result pages (SERPs). We also examined the differences

between low and high DK people in their viewing behaviors by looking at the SERPs that the viewed documents were on. As Table 3 shows, on all SERPs, high DK people viewed more documents than low DK people. Nevertheless, on all SERP levels, there was no difference in the frequency of the documents viewed by the two groups of people. With regard to the rank, on the first SERP, high DK people viewed lower ranked documents than low DK people; on the second SERP, there was no difference in the average rank of the viewed documents by the two groups of people; on the third and latter SERPs, high DK people again viewed lower ranked documents than low DK people.

Document Relevance Judgments

Comparison by DK levels

The examination on documents' relevance judgments (Table 4) found that high DK people had higher scores than low DK people (Z(1212)=6.763, p=.000). We also looked at the TREC assessors' judgments, which was on a 3-point scale, where 0 was for not relevant, 1 for partially relevant, and 2 for very relevant. Again, the documents viewed by high DK people were more relevant than those by low DK people (Z(1212)=2.486, p=.013).

	Mean (SD)		Mann-Whitney Z (p)
	Low DK	High DK	
Self-judged relevance (raw scores)	2.02 (2.34)	2.90 (2.14)	**6.76 (.000)**
Self-judged relevance (grouped scores)	0.80 (0.69)	1.06 (0.69)	**6.54 (.000)**
Gold-standard relevance	0.48 (0.71)	0.59 (0.76)	**2.49 (.013)**

Table 4. Relevance judgments comparison of two groups of people

We also transformed our users' judgments into 3-levels, following TREC assessors' scales. Based on our 5-point scale meaning, we treated the original score 1 as "0", for not relevant, scores 2 to 4 as "1", for partially relevant, and score 5 as "2", for very relevant. Using this transformed data, the analysis again showed that high DK people had higher rating scores than low DK people (Z(1212)=6.543, p=.000).

Comparison by different sets of document

When looking at the relevance judgments by different sets of documents, results (Table 5) showed that the self-rated relevance raw scores of those documents viewed by low DK people were lower than those viewed by high DK and also lower than those viewed by both groups of people (H(1211)=50.393, p=.000). The same pattern was found for the grouped self-rated relevance scores (H(1211)=55.965, p=.000).

Comparison with the TREC Gold-Standard Results

When looking at the gold-standard relevance judgments, it was found that documents viewed by both low and high DK people had higher scores than those documents viewed by either group of people only (H(1211)=100.049, p=.000).

	Mean (SD)			Kruskal-Wallis H (p)
	Set 1	Set 2	Set 3	
Self-rated relevance (raw scores)	1.39 (2.36)	2.75 (2.26)	2.74 (2.17)	**50.39 (.000)**
Self-rated relevance (grouped scores)	0.60 (0.65)	1.01 (0.72)	1.02 (0.68)	**55.97 (.000)**
Gold-standard relevance	0.28 (0.63)	0.32 (0.63)	0.70 (0.77)	**100.49 (.000)**

Table 5. Relevance judgments between different sets of documents

Relevance judgments by different groups on the shared documents

For those documents viewed by both high and low DK people, the self-judged relevance scores of high DK people, both the raw (Z(735)=3.859, p=.000) and the grouped (Z(735)=3.727, p=.000), were higher than the low DK people (Table 6). This indicated that even though both groups of people viewed the same documents, high DK people thought the documents were more relevant than the low DK people.

	Mean (SD)		Mann-Whitney Z (p)
	Low DK	High DK	
Self-judged relevance (raw scores)	2.39 (2.26)	3.01 (2.06)	3.86 (.000)
Self-judged relevance (grouped scores)	0.91 (0.68)	1.10 (0.67)	3.73 (.000)

Table 6. Set 3 documents relevance judgment between two groups of people

Self-judged vs. gold-standard

To further help understand the difference between people with different levels of DK in their relevance

judgment, we also compared the self-judged relevance scores (the grouped) with the gold-standard scores. Table 7 showed that self-judged scores were greater than gold-standard for all documents, those viewed by low DK people only, viewed by high DK people only, as well as by both groups.

	N	Mean (SD)		Paired t(p)
		Self-judged	Gold-standard	
General	1213	0.95 (0.70)	0.54 (0.74)	**15.38 (.000)**
Low DK	504	0.80 (0.69)	0.48 (0.71)	**8.00 (.000)**
High DK	709	1.06 (0.69)	0.59 (0.76)	**13.35 (.000)**
Set 1 documents	184	0.60 (0.65)	0.28 (0.63)	**4.80 (.000)**
Set 2 documents	292	1.02 (0.63)	0.32 (0.72)	**12.87 (.000)**
Set 3 documents	737	1.02 (0.77)	0.70 (0.68)	**9.46 (.000)**

Table 7. Comparison between self-judged and gold-standard relevance judgments

DISCUSSION & CONCLUSIONS

In terms of document selections, the results from this study are similar to the findings from an earlier study by the first author, in which 26 participants used Google search engine for searching on five topics. The participants were categorized into low knowledge (n=8) and high knowledge (n=18) groups. In total, 1164 documents were retrieved, with 255 (22%) viewed only by low DK users, 724 (62%) viewed only by high DK users, and 185 (16%) by both. The high DK users judged the documents they retrieved more relevant than the low DK users did.

Our results revealed several interesting points. First, it is demonstrated that, given the same search topic, low DK and high DK users retrieved and viewed primarily separate sets of documents. Only a small fraction (less than 1/4) of the documents were viewed by both groups of users. They selected relevant documents from different sets of documents.

We also found that high DK users viewed significantly more documents than low DK users, and the documents viewed by high DKs ranked lower than those viewed by low DK users. High DK people appeared to go further to the bottom of a SERP as well as later SERPs to look for relevant documents. While figuring out probable reasons requires further research, one possible explanation could be that high DK people do not mind reading the documents that were determined not as relevant by the system (i.e., the system put them in the lower ranks), which may not seem so applicable at first glance.

Another finding is that compared with the low DK level people, those high in DK had higher relevance scores on the saved documents. This may indicate that for these recall-based tasks, higher DK people may comprehend the documents more and can recognize relevant documents although this might not be recognizable immediately. Again, further research is needed to confirm the reasons.

Further, we found that both the low and the high DK participants rated higher average relevance scores than the gold-standard. We think several reasons could lead to this. First, TREC assessors were two biology-majored students: one PhD and one undergraduate student. It is possible that the participants in our study, especially the high DK people, had higher levels of DK than TREC assessors. Second, the gold-standard assessments were for the documents retrieved by the TREC participating groups, and from our observations, some documents retrieved by our system were different than theirs. Third, our experiment task was recall-based which asked the participants to try to find as many relevant documents as possible, but this was not the case in the TREC assessment process.

The findings of this research have implications on developing personalization techniques for IR systems. Since the users with low level DK tended to view different documents from the higher level DK users, and tended to have an inferior search performance, the system may learn from the higher level DK users' behaviors, and help the low level DK users adopt these behaviors to achieve a better performance.

ACKNOWLEDGMENTS

This research was partially supported by IMLS Grant #LG 06-07-0105-07.

REFERENCES

1. Kachigan, S. K. (1991). Multivariate Statistical Analysis. Radius Press, New York.
2. White, R., Dumais, S.T., & Teevan, J. (2009). Characterizing the influence of domain expertise on Web search behavior. Proceedings of WSDM 2009.
3. Wildemuth, B. (2004). The effects of domain knowledge on search tactic formulation. Journal of the American Society for Information Science and Technology, 55(3), 246-258.
4. Zhang, X., Anghelescu, H. G. B. &Yuan, X. (2005). Domain knowledge, search behavior, and search effectiveness of engineering and science students. Inf. Res., 10(2), 217.