The Role of Domain Knowledge in Document Selection From Search Results

Jingjing Liu
School of Library and Information Science, University of South Carolina, 1501 Greene Street, Columbia, SC, 29208. E-mail: jingjing@sc.edu

Xiangmin Zhang*
School of Information Sciences, Wayne State University, 106 Kresge Library, Detroit, MI, 48202. E-mail: ae9101@wayne.edu

It is a frequently seen scenario that when people are not familiar with their search topics, they use a simple keyword search, which leads to a large amount of search results in multiple pages. This makes it difficult for users to pick relevant documents, especially given that they are not knowledgeable of the topics. To explore how systems can better help users find relevant documents from search results, the current research analyzed document selection behaviors of users with different levels of domain knowledge (DK). Data were collected in a laboratory study with 35 participants each searching on four tasks in the genomics domain. The results show that users with high and low DK levels selected different sets of documents to view; those high in DK read more documents and gave higher relevance ratings for the viewed documents than those low in DK did. Users with low DK tended to select documents ranking toward the top of the search result lists, and those with high in DK tended to also select documents ranking down the search result lists. The findings help design search systems that can personalize search results to users with different levels of DK.

Introduction

In various emergent medical situations, such as to diagnose a critical health condition for a patient, or to find ways to stop a fast-spreading disease, a fast response to medical literature search requests is critical. This fast response certainly relies on the system's effective search algorithms, but human factors may likely decide the accuracy and correctness of the results, which is more important than just keywords matching results. In this research, we investigated the effect of a user's prior knowledge on the quality of the system's search results from a big medical science data set, in the domain of genomics.

Users of today's information retrieval (IR) systems have various backgrounds. One goal of designing optimal search systems is to make searching effective for people with different levels of domain knowledge (DK), low or high, on search topics. A good amount of research has been done on user background and also on search result relevance judgment, which enables us to better understand the nature of finding related documents. To add to our knowledge about IR and also to benefit search system designs, it is important to know if users with different levels of knowledge would select the same set or different sets of relevant documents. To achieve this goal, it is desirable to understand how people with different levels of knowledge make decisions to select and evaluate the retrieved documents on the same topic, regardless of the system's searching and ranking techniques.

There has been research effort on the criteria that people use to judge and select a document. For example, Wang and Soergel (1998) studied the user's document selection process and identified the factors that affect such a process. However, their research did not differentiate users with different levels of DK. On the other hand, there has been much research on the differences between domain experts and novices in information seeking. Previous studies have found that users with higher level of DK have different search tactics (namely, querying behaviors), performance (result accuracy, and so on), time spent on task accomplishment and document reading, and so on (for example, White, Dumais, & Teevan, 2009; Wildemuth, 2004; Zhang, Anghelescu, & Yuan, 2005). However, the literature has rarely seen efforts spent on examining how DK affects relevant document selection.

*Corresponding author.

Received June 2, 2018; revised November 8, 2018; accepted December 19, 2018

© 2019 ASIS&T • Published online Month 00, 2018 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.24199

This article aims at exploring this issue, with the following specific research questions (RQs) addressed:

RQ1. Do people with different levels of DK select the same or different set of documents for the same search topic?

As found in previous studies (for example, Hsieh-Yee, 1993; Wildemuth, 2004), DK experts and novices had different search queries or search strategies. Different queries could possibly return different sets of search results. Therefore, our hypothesis is that users with different levels of DK would retrieve and select different sets of documents.

RQ2. For the selected documents on the search result pages (SERPs), are there differences between DK experts and novices regarding the documents' ranking positions?

The rationale behind this RQ is that even if experts and novices could retrieve roughly the same set of documents, they may make different decisions on selecting documents on the SERPs. For instance, some people may mainly pick up top-ranked documents (those deemed more relevant by the system), and some others may (also) view and select documents that are listed down the result list on the SERPs (those deemed less relevant by the system). Understanding the possible different selecting behaviors will be helpful for improving search results rankings.

RQ3. Are there differences between domain experts and novices in assessing the relevance of their selected documents, including the documents that were selected by both groups?

Typically, one would select documents that are at least somehow relevant. However, it is uncertain if there is any difference in the degree or level of relevancy of the documents selected by domain experts and novices. This research question is specifically interested in the relevance judgment scores that domain experts and novices would assign for the documents that they selected.

RQ4. Do the documents selected by domain experts have different features than those selected by domain novices?

Document features to be considered for this RQ include: the number of words in the documents, the number of words in MeSH (Medical Subject Headings),[1] and the specificity of the MeSH terms. The specificity level of the MeSH terms are determined by their hierarchical levels listed in the MeSH tree.

The assumption for this RQ is that domain experts are able to understand the text selected, and therefore would focus on more specific terms. Domain novices, on the other hand, would prefer more general terms in the text because of their insufficient or incomplete understanding of the text due to the lack of knowledge.

By exploring the above-described RQs, this research hopes to contribute to personalizing search results; for example, tailoring the rank and/or presentation of documents on the SERPs, based on users' DK levels. This could help the users more easily find relevant documents that meet their specific need, given their knowledge levels.

Literature Review

Domain Knowledge and Information Search Behaviors

A rich amount of research effort has been spent examining the effects of DK on information search behaviors. Previous studies have mainly focused on three aspects: the overall search process and performance, query strategies (also called search tactics), and dwelling behavior on SERPs and documents. These are introduced below.

Research findings about the general search process revealed that higher levels of knowledge tended to be associated with less effort in preparing for search, such as a shorter time in reading the instructions for the task in an experiment setting (Vibert et al., 2009), but more effort in the search process; for example, longer time spent on a session in a naturalistic web search environment (White et al., 2009), more queries per task session (White et al., 2009; Zhang et al., 2005), longer queries on average (Hembrooke, Granka, & Gay, 2005; White et al., 2009; Zhang et al., 2005), more keywords in their first query (Vibert et al., 2009), and more pages viewed in a session (Vibert et al., 2009; White et al., 2009). However, a few studies had different results from the above findings. Duggan and Payne (2008) found that knowledge level in the music domain did not affect user behaviors, but in the football domain it was negatively correlated with mean query length. Vibert et al. (2009) found that knowledge level in the neuroscience did not affect users' time spent to accomplish tasks. Regarding the search outcome, Duggan and Payne (2008) found that experts in the football domain had a higher accuracy rate of answering questions. Kang, Fu, and Kannampallil (2010) found that experts performed better in finding unique information.

Regarding query strategies, it has been found that higher levels of DK are associated with fewer keywords taken from task instructions (Vibert et al., 2009), less use of a thesaurus for term suggestion (Hsieh-Yee, 1993), but when using a thesaurus, more effectiveness of the use (Sihvonen & Vakkari, 2004), more terms from own knowledge (Hsieh-Yee, 1993), a wider and more specific vocabulary (Vakkari, Pennanen, & Serola, 2003), more efficient selection of concepts (Wildemuth, 2004), less repeated terms (Hembrooke et al., 2005) or synonyms (Hsieh-Yee, 1993), more unique terms (Hembrooke et al., 2005), less use of the same querying structure (Hembrooke et al., 2005), less term combinations (Hsieh-Yee, 1993), less close monitoring of search (Hsieh-Yee, 1993), and so on. Drabenstott (2003) found that domain novices rarely used strategies that are indicators of domain experts, such as citation search, author searching, and the journal run (identifying a journal and then searching its volumn/issues), and that domain novices frequently used subject

[1] MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of descriptors in a hierarchical structure that permits searching at various levels of specificity.

search, in which they used general terms about a subject or general topic.

Dwelling behaviors on SERPs and documents has not gained as much attention as query strategies, but the literature is seeing an increase in researchers' effort along this line. Related studies have found that experts used less time to find their first relevant document (Downing, Moore, & Brown, 2005), and had a tendency of spending less time on documents (Duggan & Payne, 2008; Kelly & Cool, 2002). Liu and Belkin (2010) found that only on very useful pages did people with higher knowledge spend less time than those with lower knowledge, and for not useful pages and somewhat useful pages, the first uninterrupted dwell time (excluding revisiting to the same page) on the webpages did not show differences between users with higher and lower levels of knowledge.

The above-introduced studies have clearly shown that knowledge is a significant factor influencing information search. However, none of them examined the selection behavior on SERPs regarding what documents are selected, and at what positions.

Domain Knowledge and Relevance Judgments

In IR, document selection from SERPs is closely related to relevance judgment, because the documents selected most of the time are considered "relevant" to the search topic. The concept of "relevance" has been studied intensively. On the practical side, researchers have found that there are many facets or criteria of "relevance." Research on relevance judgment has identified many factors or criteria that searchers use to make decisions. For example, Taylor, Zhang, and Amadio (2009) and Taylor (2012) emphasize the dynamic nature of the relevance judgment process and provided strong statistical evidence of associations between the information search process and the choices of relevance criteria over the course of a search. Bailey et al. (2008) discussed the agreement between relevance judgments from different types of searchers: "gold standard" judges who are experts in a particular information-seeking task and originated the topic; "silver standard" judges who are task experts but did not create topics; and "bronze standard" judges who are not experts in the task. Analysis shows low levels of agreement in relevance judgments between these three groups. A more recent work by Palotti, Hanbury, Muller, and Kahn (2016) and Palotti, Zuccon, Bernhardt, Hanbury, and Goeuriot (2016) found assessment disagreements between lay people and experts in topical relevance judgment, indicating that relevance judgment relying on lay people for health-related system evaluation could lead to bias and unreliable results. However, they also found that assessment of understandability showed high correlation instead of disagreements between lay people and experts.

While the research on relevance so far has been helpful for understanding how people in general cognitively evaluate search results, and thus select retrieved documents, little evidence has been produced that shows how people with different levels of domain expertise make relevance judgments on documents, and no investigation has been done to reveal the document-viewing behaviors by different types of users.

Relevant Document Selection Behaviors

Document selection is based on a user's relevance judgment of the retrieved documents. While relevance judgment research is important in understanding the users, relevance judgment itself is basically a mental process, which cannot be observed by the search system. From a system design point of view, it may be more important to know the results of relevance judgments, that is, the selected documents, because they are observable by the system so that the system could use the information to create adaptive strategies to help users. Therefore, we focus on document selection issues in the current research.

Regardless of a user's background, the document selection process and decision making in information seeking have been studied. One early notable work is Wang and Soergel (1998). Based on their observations of 25 voluntary faculty and graduate students in agricultural economics, who participated in the research selecting documents from search results in a database (DIALOG), the researchers proposed a general cognitive model of document selection, which follows the human decision-making process. The model includes four major components, each with a set of values: Document Information Elements (such as title, author, and so on), Criteria (such as topicality, orientation, quality, and so on), Values (such as epistemic, functional, and so on), and Decision (acceptance, maybe, and rejection). The researchers note that the decision rules govern the overall document selection process, and knowledge plays an important role in processing information and applying criteria.

As a follow-up study to Wang and Soergel (1998), Wang and White (1999) interviewed 15 out of the 25 participants for their subsequent decisions as they progress through the project on those documents previously selected as relevant. They found that topic knowledge was one of the important factors that affect the user's decision making. More evidence has been reported in recent years. Users with greater knowledge have been found to find more relevant articles (Downing et al., 2005; Nguyen & Santos Jr., 2007), save more information (Kang et al., 2010), and have a higher ratio of relevant documents saved out of all documents viewed (Kelly & Cool, 2002). Meanwhile, some studies had different findings. For example, Zhang et al. (2005) found that people with higher and lower knowledge in the engineering domain did not differ in the numbers of relevant documents found. In addition, domain experts were found to evaluate search results more thoroughly and clicked more often on relevant search results (Cole, Zhang, Liu, Belkin, & Gwizdka, 2011), have more page visits and be more persistent with longer evaluation sessions

(Pallotti et al., 2016), and detect unfruitful search paths faster than nonexperts (Duggan & Payne, 2008). One possible explanation is that higher DK enables a user to comprehend the search results and the content better, which, in turn, enables the user to make informed decisions regarding document or search results evaluations.

Knowledge and Personalization Approaches

Kumaran, Jones, and Madani (2005) differentiated documents into the introductory and the advanced that match different levels of topic knowledge. A classifier was built to classify the documents according to different features (for example, stop-word, line-length) that could be predictive of assumed topic knowledge. An experiment to rerank search results for people with lower topic familiarity showed that the classifier was effective: the portion of introductory pages at the top 5 and top 10 result sets using this method were significantly higher than those in the baseline run using a default search engine ranking. Paukkeri, Ollikainen, and Honkela (2013) recognize that different users have different levels of DK, and documents are written differently for people with such differences. Documents intended for professionals may not be understandable at all by a novice user and documents for novice users may not contain all the detailed, specific, and in-depth information needed by an expert. The authors addressed the issue of how to estimate a document's knowledge (difficulty) level to match that of the user's, so that the user's selection of relevant documents can be personalized, to meet the user's need. The authors proposed a method of matching a document with a user's level of DK, which is based on the comparison of terms appearing in a document and terms known by the user. Using the method, they examined medical documents and found that the method is able to distinguish between documents written for novices and the documents written for experts. The implication of this research is that, documents are written for people with different levels of DK, who should be served with different documents by a system. However, this research did not investigate how people with different levels of DK actually select documents, and whether or not the selected documents by people with different levels of DK would be the same or different.

Method

The current research used a data set collected through a laboratory user experiment. In the experiment, 35 participants each performed four out of five search tasks using a customized experimental search system and following a systematically rotated task order. A detailed description of the research design can be found in Zhang, Liu, Cole, and Belkin (2015). In order to clearly present the current article, the experiment procedure is depicted in Figure 1, and the main components of the methodology are introduced in the subsections below.

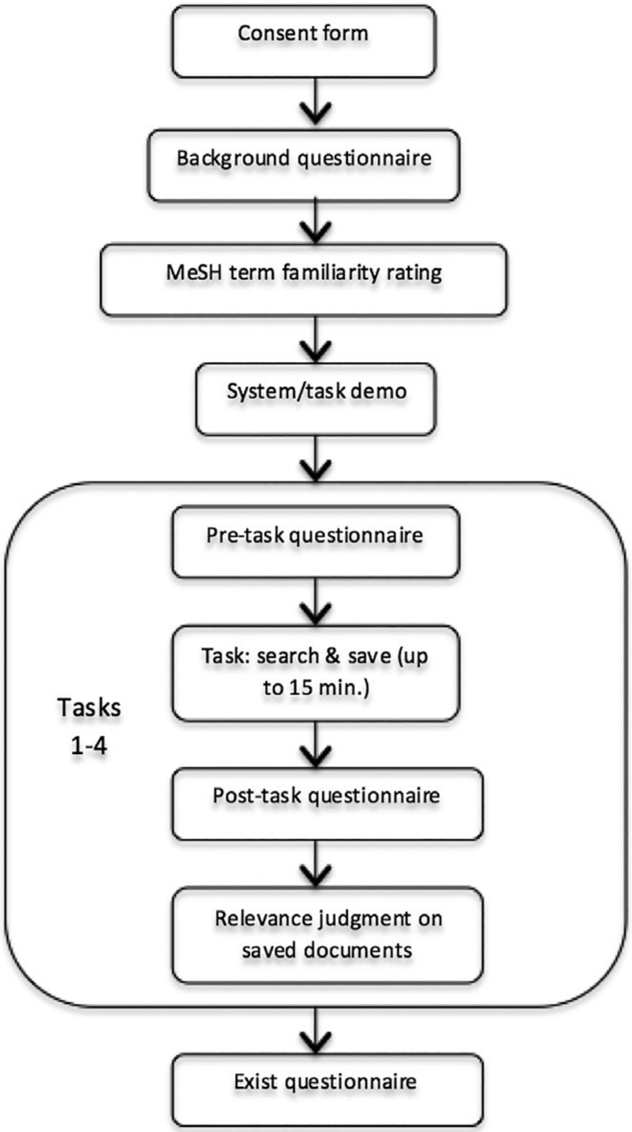


FIG. 1. Experiment procedure.

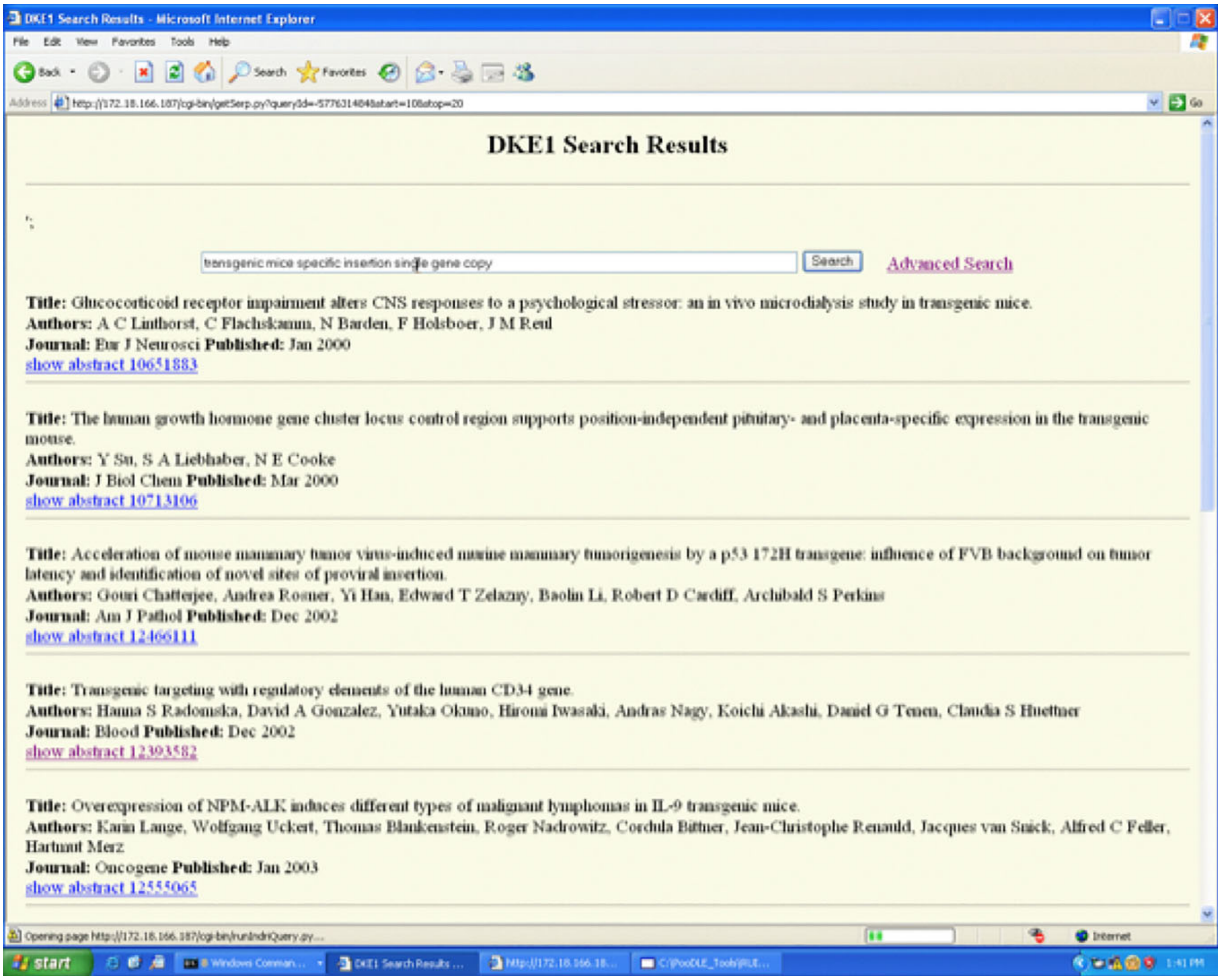
Experimental System and Data Set

The study designed an experimental search system using the Indri search engine from the Lemur toolkit.[2] The system's underlying data set used a subset of the MEDLINE bibliographic database (Hersh & Voorhees, 2009) in the Text Retrieval Conference's (TREC[3]) 2004 Genomics track data collection. This subset was for the period of year 2000–2004 (n = 1.85 million), the size of which was large enough to allow reasonable retrieval efficiency.

Users could choose to use either simple or advanced searches in the system. They could formulate any queries they liked. The system's sample SERP is displayed in Figure 2. Retrieved documents on the SERPs

[2] http://lemurproject.org

[3] http://trec.nist.gov/



http://172.18.166.187/cgi-bin/getSerp.py?queryId=-577631484&start=10&stop=20

DKE1 Search Results

transgenic mice specific insertion single gene copy

Title: Glucocorticoid receptor impairment alters CNS responses to a psychological stressor: an in vivo microdialysis study in transgenic mice.
Authors: A C Linthorst, C Flachskamm, N Barden, F Holsboer, J M Reul
Journal: Eur J Neurosci **Published:** Jan 2000
show abstract 10651883

Title: The human growth hormone gene cluster locus control region supports position-independent pituitary- and placenta-specific expression in the transgenic mouse.
Authors: Y Su, S A Liebhaber, N E Cooke
Journal: J Biol Chem **Published:** Mar 2000
show abstract 10713106

Title: Acceleration of mouse mammary tumor virus-induced murine mammary tumorigenesis by a p53 172H transgene: influence of FVB background on tumor latency and identification of novel sites of proviral insertion.
Authors: Gouri Chatterjee, Andrea Rosner, Yi Han, Edward T Zelazny, Baolin Li, Robert D Cardiff, Archibald S Perkins
Journal: Am J Pathol **Published:** Dec 2002
show abstract 12466111

Title: Transgenic targeting with regulatory elements of the human CD34 gene.
Authors: Hanna S Radomska, David A Gonzalez, Yutaka Okuno, Hiromi Iwasaki, Andras Nagy, Koichi Akashi, Daniel G Tenen, Claudia S Huettner
Journal: Blood **Published:** Dec 2002
show abstract 12393582

Title: Overexpression of NPM-ALK induces different types of malignant lymphomas in IL-9 transgenic mice.
Authors: Karin Lange, Wolfgang Uckert, Thomas Blankenstein, Roger Nadrowitz, Cordula Bittner, Jean-Christophe Renauld, Jacques van Snick, Alfred C Feller, Hartmut Merz
Journal: Oncogene **Published:** Jan 2003
show abstract 12555065

Opening page http://172.18.166.187/cgi-bin/runIndriQuery.py...

FIG. 2. A search results page as displayed in the search system. [Color figure can be viewed at wileyonlinelibrary.com]

are hyperlinked to their abstracts.[4] The displayed abstract is depicted in Figure 3.

Search Topics

For this controlled laboratory user study, five search topics were selected from the total 50 search topics used in the TREC Genomics Track. The topics were selected based on the consideration of balancing (i) MeSH categories, (ii) the specificity levels of the topics, and (iii) search difficulty levels. For MeSH categories, the topics had one from Category I (Genetic Processes), two from Category II (Genetic Phenomena), and two from Category III (Genetic Structure). Task specificity was determined by the level of the task topic keywords in the MeSH tree, specifically, the path length to the root in the MeSH category tree, as judged by an external expert in the biomedical area hired in the study. A topic having a MeSH hierarchy level of higher than three was considered as a general, and otherwise, specific. Task difficulty was determined by the ease of finding relevant documents using the topic's title keywords as queries in the experimental system. The topics and their corresponding MeSH categories are listed in Table 1.

The topics were presented unchanged from the TREC Genomics Track descriptions. The following lists their MeSH category and the topic description.

Category I: Genetic processes

7. DNA repair and oxidative stress

Need: Find correlation between DNA repair pathways and oxidative stress.

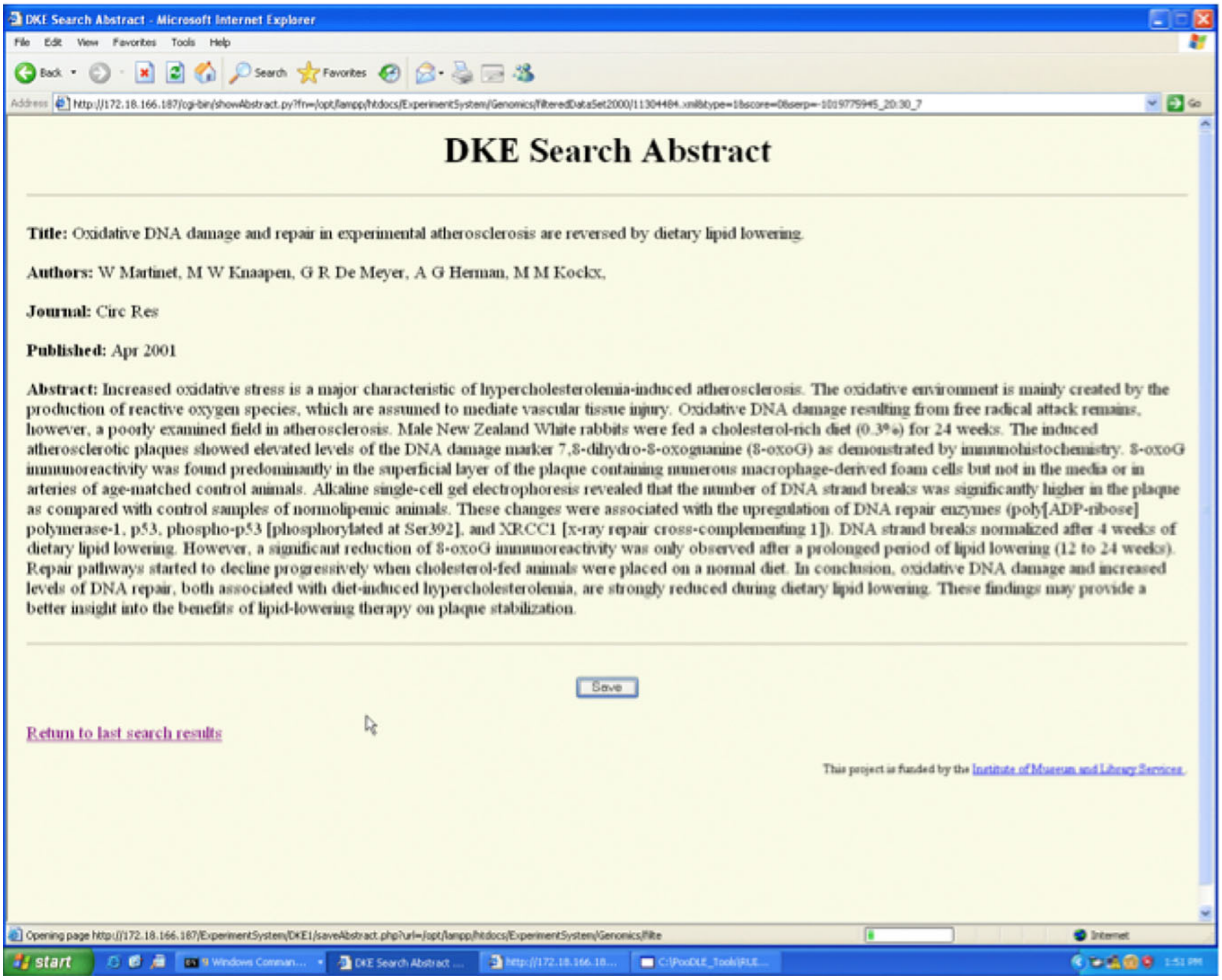
Context: Researcher is interested in how oxidative stress affects DNA repair.

Category II. Genetic phenomena

42. Genes altered by chromosome translocations

Need: What genes show altered behavior due to chromosomal rearrangements?

[4] In this experiment, only the abstracts of articles are shown. The term "document" used in the current article therefore refers to the "abstract" in the experiment system.


DKE Search Abstract - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://172.18.166.187/cgi-bin/showAbstract.py?fn=/opt/lampp/htdocs/ExperimentSystem/Genomics/filteredDataSet2000/11304484.xml&type=1&score=0&serp=-1019775945_20:30_7
Go
Save
Opening page http://172.18.166.187/ExperimentSystem/DKE1/saveAbstract.php?url=/opt/lampp/htdocs/ExperimentSystem/Genomics/filte
Internet
start
9 Windows Comman...
DKE Search Abstract ...
http://172.18.166.18...
C:\PooDLE_Tools\RUE...
1:51 PM

DKE Search Abstract

Title: Oxidative DNA damage and repair in experimental atherosclerosis are reversed by dietary lipid lowering.

Authors: W Martinet, M W Knaapen, G R De Meyer, A G Herman, M M Kockx,

Journal: Circ Res

Published: Apr 2001

Abstract: Increased oxidative stress is a major characteristic of hypercholesterolemia-induced atherosclerosis. The oxidative environment is mainly created by the production of reactive oxygen species, which are assumed to mediate vascular tissue injury. Oxidative DNA damage resulting from free radical attack remains, however, a poorly examined field in atherosclerosis. Male New Zealand White rabbits were fed a cholesterol-rich diet (0.3%) for 24 weeks. The induced atherosclerotic plaques showed elevated levels of the DNA damage marker 7,8-dihydro-8-oxoguanine (8-oxoG) as demonstrated by immunohistochemistry. 8-oxoG immunoreactivity was found predominantly in the superficial layer of the plaque containing numerous macrophage-derived foam cells but not in the media or in arteries of age-matched control animals. Alkaline single-cell gel electrophoresis revealed that the number of DNA strand breaks was significantly higher in the plaque as compared with control samples of normolipemic animals. These changes were associated with the upregulation of DNA repair enzymes (poly[ADP-ribose] polymerase-1, p53, phospho-p53 [phosphorylated at Ser392], and XRCC1 [x-ray repair cross-complementing 1]). DNA strand breaks normalized after 4 weeks of dietary lipid lowering. However, a significant reduction of 8-oxoG immunoreactivity was only observed after a prolonged period of lipid lowering (12 to 24 weeks). Repair pathways started to decline progressively when cholesterol-fed animals were placed on a normal diet. In conclusion, oxidative DNA damage and increased levels of DNA repair, both associated with diet-induced hypercholesterolemia, are strongly reduced during dietary lipid lowering. These findings may provide a better insight into the benefits of lipid-lowering therapy on plaque stabilization.

Return to last search results

This project is funded by the Institute of Museum and Library Services.

FIG. 3. A document's abstract as displayed in the search system. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 1. Search topics.

TREC topic id	Topic title keywords	MeSH category	Specificity (hierarchical level in MeSH)	Difficulty level
2	Generating transgenic mice	Genetic structure	Specific (4)	Hard
7	DNA repair and oxidative stress	Genetic processes	General (1)	Easy
42	Genes altered by chromosome translocations	Genetic phenomena	Specific (4)	Easy
45	Mental Health Wellness-1	Genetic phenomena	General (1)	Hard
49	Glyphosate tolerance gene sequence	Genetic structure	General (1)	Hard

Context: Information is required on the disruption of functions from genomic DNA rearrangements.

45. Mental health wellness-1

Need: What genetic loci, such as Mental Health Wellness 1 (MWH1) are implicated in mental health?

Context: Want to identify genes involved in mental disorders.

Category III: Genetic structure

2. Generating transgenic mice

Need: Find protocols for generating transgenic mice.

Context: Determine protocols to generate transgenic mice having a single copy of the gene of interest at a specific location.

49. Glyphosate tolerance gene sequence

Need: Find reports and glyphosate tolerance gene sequences in the literature.

Context: A DNA sequence isolated in the laboratory is often sequenced only partially, until enough sequence is generated to identify the gene. In these situations, the rest of the sequence is inferred from matching clones in the public domain. When there is difficulty in the laboratory manipulating the DNA segment using sequence-dependent methods, the laboratory isolate must be reexamined.

Tasks

The participants' task for each topic was to search in the experiment system for the "Need" in the topic's description, and save the documents that they thought relevant. They were given up to 15 minutes for each task. Their search process was recorded by logging software Morae.[5] At the end of the search for each topic, participants evaluated the viewed documents recorded by the logger, and made relevance judgments on a 5-point Likert scale, in which 1 was for *not relevant*, 3 for *somewhat relevant*, and 5 for *highly relevant*. Compared with the TREC gold standard judgment (see also the Results subsection: Examination of Relevance Judgments) that used a 3-point scale, 0 for *not relevant*, 1 for *somewhat relevant*, and 2 for *highly relevant*; the 5-point scale provided more granularity.

Participants

A total of 35 students (9 male, 26 female) from Rutgers University participated in the study. The participants' ages ranged from 18–32 years. To ensure that the task topic matches the participants' background, participants were students and postdoctoral researchers from biology-related schools and departments, including biology, pharmacy, animal science, biochemistry, and public health. The number of graduate and postgraduate participants and the number of undergraduate participants were roughly balanced. All participants considered themselves expert web searchers (rated 6 or 7 on a 7-point Likert scale). Each participant was paid $25 upon completion of the experiment session.

Assessment of the Participants' Domain Knowledge Level

A participant's DK level was determined by two factors: (i) his/her familiarity rating for the MeSH terms, and (ii) his/her search topic familiarity and expertise ratings in the pretask questionnaires. Familiarity with MeSH terms was assessed in the MeSH term rating step (see Figure 1), in which participants rated their understanding of the MeSH concepts[6] in three MeSH concept trees related to the topics: genetic processes (G05), genetic phenomena (G13), and genetic structures (G14). A 5-point scale was used, and each point was associated with a textual explanation: 1 for *no knowledge*, 2 for *vague idea*, 3 for *some knowledge*, 4 for *high knowledge*, and 5 for *can explain to others*. The 5-point scale was used since it was easier to attach textual meaning for each point and was easier for the participants to better understand and respond.

Participants were also asked to assess their familiarity and expertise with each particular search topic/task after they had read the topic description and before they started working on the task. A 7-point scale was used, with 1 standing for: *not at all*, 4 for *somewhat*, and 7 for *extremely*. The use of a 7-point scale for these questions was consistent with the scale use of other questions/statements in all questionnaires (including background, pretask, and posttask questionnaires) in the experiment.

For each participant, the average score of the MeSH term ratings and that of the pretask topic familiarity and expertise were calculated individually. Since the MeSH ratings and pretask questionnaire ratings used different scales, their average ratings were standardized by Z scores (Kreyszig, 2011). These two Z scores for each participant were consequently averaged as the DK level for each participant, as shown in the following equation:

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

Based on the DK scores, users were divided into two groups based on the median value, which was −0.155. Those scored above the median were put in a high DK group and those below were in a low DK group.

Results

An exploration of the data was first conducted, which revealed that the variables were generally not normally distributed. Therefore, nonparametric tests were used in the analysis whenever appropriate. Specifically, for comparison of two groups, the Mann–Whitney *U*-test was used, and for comparison of more than two groups, the Kruskal–Wallis test was used (Corder & Foreman, 2014).

An Overview of the Viewed Documents

Documents viewed by the two DK level user groups. First examined was an overview of the documents viewed by the users with different levels of DK. In general, users viewed 1,143 documents (including repeatedly reviewed documents). Among them, about 40.7% (n = 477) were viewed by low DK users, and 59.3% (n = 666) were viewed by high DK users. A chi-square test detected that high DK users viewed significantly more times than low DK users, $\chi2 = 31.25$, $p = .000$.

The three sets of documents. Some documents were viewed by multiple users, and the total number of unique documents that were viewed by users in this experiment was 506. Among them, about 30% (n = 151) were viewed only by the low DK level participants (labeled as Set L); about 47% (n = 238) were viewed by the high DK level participants only (labeled as Set H); and about 23%

[5] https://www.techsmith.com/morae.html

[6] http://www.nlm.nih.gov/MeSH/trees.html

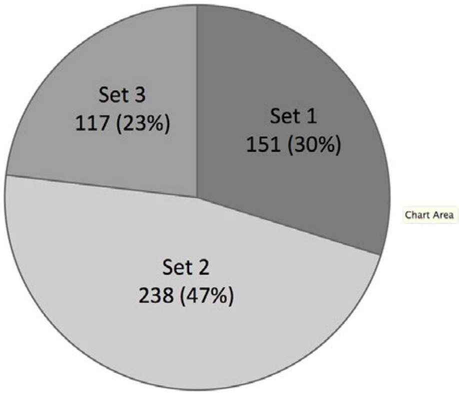


FIG. 4. Distribution of the documents viewed by low and high DK users. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2. Comparison between the numbers of different sets of documents.

	χ2	Adjusted *p* values
Set L vs. Set H	19.46	**.000**
Set L vs. Set B	4.31	.15
Set H vs. Set B	41.24	**.000**
(Set L + Set B) vs. (Set H + Set B)	12.15	**.002**

(*n* = 117) were viewed by both groups of participants (labeled as Set B), as shown in Figure 4.

The chi-square test was conducted to examine if the differences between the numbers among the three sets of documents were significant. The *p* values for the pairwise comparisons were adjusted using Bonferroni correction (Dunn, 1961). The results in Table 2 show that the number of documents viewed by the higher DK people only (Set H) were significantly more than the number of documents viewed by the low DK people only (Set L), χ2 = 19.46, *p* = .000; the number of documents viewed by the higher DK people only (Set H) were significantly more than the number of documents viewed by both groups of users (Set B), χ2 = 41.24, *p* = .000; and there was no difference between the number of documents viewed by the low DK people only (Set L) and the number of documents viewed by both groups of users (Set B), χ2 = 4.31, *p* = .15.

Meanwhile, among the 506 documents, the total number of documents viewed by low DK level participants was 268 (Set L + Set B), for a percentage of 53.0% (268/506); the total number of documents viewed by high DK level participants was 355 (Set H + Set B), for a percentage of 70.2% (355/506). A chi-square test found that, again, the total number of documents viewed by the higher DK people (*n* = 355) were significantly more than the total number of documents viewed by the low DK people (*n* = 268), χ2 = 12.15, *p* = .002.

Document Ranking

Ranking position of the documents viewed by two DK level user groups. Following the number of documents viewed, further examined was the users' more specific document-viewing decisions, starting from the ranking position of the documents retrieved and selected to view. The ranking position of a document on the search result list goes from 1 as the top one on the list to lower position on the list with higher ranking position numbers. In this section, the ranks are continuously numbered, namely, the top position on SERP 1 is rank #1, the top position on SERP 2 is rank #11, and the top position on SERP 3 is rank #21, and so forth.

A Mann–Whitney *U*-test found that the rank position number of the documents viewed by low DK users were significantly smaller than that of the documents viewed by high DK users, Z = 4.07, *p* = .000 (Table 3). This means that, in general, low DK users tended to view documents that were listed toward the top of the result lists.

TABLE 3. Ranks of documents viewed by different DK level users.

	Low DK Mean (SD)	High DK Mean (SD)	Mann–Whitney Z (*p*)
Rank position	11.06 (16.42)	14.78 (20.98)	**4.07 (.000)**

Ranking position of the viewed documents in the three document sets. We also examined the ranks of the three different document sets. The Kruskal–Wallis test found a significant difference existing in the rank of the three sets of documents (H = 113.03, *p* = .000) (Table 4). The pairwise comparison, with Bonferroni-adjusted *p* values for repeated comparison, further detected that the differences were between sets L versus B, and sets H versus B (Table 5). Specifically, the average rank of the documents viewed by either DK user group was higher than that of the documents viewed by both groups of users. This means that documents that were viewed by both high and low DK people tended to be the low-ranked ones, and the average number was below 10, meaning that they were on the first SERP page on average. Descriptively, the average rank of Set L documents (average on the second SERP) was lower than that of Set H documents (average on the third SERP).

Ranking position of the viewed documents on different SERPs. Another way to examine the ranking positions of the viewed documents is to check their locations on different SERPs. We examined SERP #1, SERP #2, and SERP #3 and beyond, and the results of the comparison are listed in Table 6.

The Mann–Whitney test found that on SERP #1, high DK people viewed documents with higher rank numbers than low DK people (Z = 2.88, *p* = .004) viewed; on SERP #2, there was no difference in the average rank of the viewed documents by the two groups of people (Z = 0, *p* = 1.00); on SERP 3 and beyond, high DK people again viewed documents with higher rank numbers than low DK people (Z = 2.02, *p* = .043) did. The results indicate that the rank difference of the documents viewed by low and high DK people were mainly on SERP #1, and on SERP #3 and beyond, but not on SERP #2.

Examination of Relevance Judgments

Relevance judgments by users with different DK levels. A Mann–Whitney test examining users' judgments on document relevance found that high DK people had higher rating scores than low DK people, Z = 3.39, *p* = .001 (Table 7). The TREC assessors' judgments, called the gold-standard assessment scores, on the viewed documents by the participants were also examined. The gold-standard assessment was based on a 3-point scale, where 0 was for *not relevant*, 1 for *partially relevant*, and 2 for *very relevant*. The Mann–Whitney test of the gold-standard scores found that documents viewed by high DK people received higher ratings than those by low DK people, Z = 2.66, *p* = .008.

The participants' judgments on the 5-point scale were transformed into 3-level, following TREC assessors' scales. Based on the 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 these transformed data, the analysis again showed that high DK people had higher rating scores than low DK people, Z = 3.31, *p* = .001.

TABLE 4. Ranks of the three sets of documents.

	Set L Mean (SD)	Set H Mean (SD)	Set B Mean (SD)	Kruskal–Wallis H (*p*)
Rank	16.96 (21.44)	22.43 (27.39)	9.10 (12.97)	**113.03 (.000)**

TABLE 5. Pairwise comparison of the average ranks of the three sets of documents.

	Standard test statistic H	Adjusted *p* value
Set L-Set H	−1.68	.281
Set L-Set B	6.24	**.000**
Set H-Set B	9.77	**.000**

Relevance scores for the three sets of documents. When looking at the relevance judgments by different sets of documents, the results (Tables 8 and 9) show that the user-rated relevance raw and transformed scores of those documents viewed by low DK people only (Set L) was lower than those documents viewed by high DK only (Set H). However, for gold-standard relevance judgment, documents viewed by either group only (Set L and Set H) had lower scores than documents viewed by both groups (Set B).

Relevance judgments by different DK level users on the shared documents. For those documents viewed by both high and low DK people (that is, Set B documents), the user-rated relevance score of high DK people, both the raw (Z = 5.39, *p* = .000) and the transformed (Z = 5.40, *p* = .000), were higher than that of the low DK people (Table 10). This indicated that even though both groups of people viewed the same documents, high DK people rated the documents as more relevant than the low DK people.

Comparison between user-rated and gold-standard relevance scores. 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 transformed) with the gold-standard scores. Table 11 shows the paired *t*-test results for all documents viewed, documents viewed by each DK user group, and by both DK user groups, as well as the three sets of documents. Please note that this comparison examined all document view times instead of unique documents, meaning that repeated views of the same documents were all included in the analysis. The results show that self-judged scores were greater than gold-standard scores. This may be due to the differences in assessors' background (many users in the current study vs. two assessors in TREC), or the different ways of how the scores were assigned

TABLE 6. Documents viewed by two groups of people on different SERPs.

		Low DK	High DK	Mann–Whitney Z (*p*)
SERP #1	Number of docs viewed	340	421	—
	Rank Mean (SD)	3.85 (2.59)	4.38 (2.61)	**2.88 (.004)**
SERP #2	Number of docs viewed	64	113	—
	Rank Mean (SD)	14.48 (2.74)	14.44 (2.64)	.000 (1.00)
SERP #3 and beyond	Number of docs viewed	73	132	—
	Rank Mean (SD)	41.64 (23.24)	48.27 (26.99)	**2.02 (.043)**

TABLE 7. Relevance judgments comparison of the two groups.

	Low DK Mean (SD)	High DK Mean (SD)	Mann–Whitney Z (*p*)
Self-judged relevance (raw scores)	3.63 (1.01)	3.86 (1.02)	**3.39 (.001)**
Self-judged relevance (transformed scores)	1.51 (0.53)	1.63 (0.51)	**3.31 (.001)**
Gold-standard relevance	0.48 (0.71)	0.59 (0.76)	**2.66 (.008)**

(transformed vs. original using 3-point scale), but the results may explain the different patterns in Table 11.

Document Features

Another aspect of interest to us was: Are there differences between the documents that users with different DK levels clicked and viewed? We examined the following document features: (i) document length, that is, the number of terms in the abstract; (ii) the number of MeSH index terms in the abstract; (iii) the number of general MeSH terms (as judged by our assessor) in the abstract; and (iv) the number of specific MeSH index terms (as judged by our assessor) in the abstract. Since our data set sizes are small, we used simple keyword matching method to identify the MeSH terms in our data sets, rather than using the relatively complex MetaMap tool (Aronson & Lang, 2010).

TABLE 8. Relevance judgments between different sets of documents.

	Set L Mean (SD)	Set H Mean (SD)	Set B Mean (SD)	Kruskal–Wallis H (*p*)
Self-rated relevance (raw scores)	3.52 (1.04)	3.89 (.99)	3.77 (1.02)	**7.92 (.019)**
Self-rated relevance (transformed scores)	1.47 (0.57)	1.64 (0.52)	1.58 (0.51)	**6.95 (.031)**
Gold-standard relevance	0.21 (0.56)	0.26 (0.61)	0.54 (0.76)	**24.95 (.000)**

TABLE 9. Standard test statistics (adjusted *p* values) of pairwise comparisons.

	Self-rated relevance (raw scores)	Self-rated relevance (transformed scores)	Gold-standard relevance
Set L vs. Set H	**−2.81 (.015)**	**−2.60 (.028)**	−1.00 (.95)
Set L vs. Set B	−2.04 (.13)	−1.73 (.25)	**−4.71 (.000)**
Set H vs. Set B	1.53 (.38)	1.64 (.31)	**−4.21 (.000)**

TABLE 10. Set B documents relevance judgment between two DK user groups.

	Low DK Mean (SD)	High DK Mean (SD)	Mann–Whitney Z (*p*)
Self-judged relevance (raw scores)	3.67 (1.00)	3.84 (1.03)	**2.17 (.030)**
Self-judged relevance (transformed scores)	1.52 (.52)	1.62 (0.50)	**2.23 (.026)**

Documents viewed by two DK level user groups. The results in Table 12 show that documents viewed by low DK users had shorter document length, smaller number of MeSH terms, and smaller number of general MeSH terms than those viewed by high DK users. This means that in general, high DK users tended to view longer documents, which had more MeSH terms and general MeSH terms than their counterparts.

The three sets of documents. The results in Tables 13 and 14 show that Set H documents had a higher number of MeSH terms than Sets L and B; Set L documents had a smaller number of general MeSH terms than Sets H & B; Sets L & H documents had more specific MeSH terms than Set B. This indicates a pattern that, in general, Set H documents (viewed by high DK only) tended to have more MeSH terms.

Discussion

The results in this study answered our research questions. RQ1 asked if people with different levels of DK select the same or different set of documents for the same search topic. Our results showed that, while there were some shared documents selected by both groups of users, there were also different documents selected by either group of users with different levels of DK. RQ2 asked about the ranking positions for the selected documents on the SERPs, and the results showed that domain novices tended to view documents ranked on the top of the SERPs (that is, those deemed more relevant by the systems). RQ3 asked if there would be differences between domain experts and novices in assessing the relevance of their selected documents. The answer was yes: domain experts were found to have higher self-assessed relevance scores, and their selected documents also had higher relevance scores assessed by the TREC gold-standard evaluation. For RQ4 addressing the document features, it was found that domain experts selected documents that were longer, had more MeSH terms, and more general MeSH terms.

These results revealed several interesting points. First, while it is important to enable a fast response system to

TABLE 11. Comparison between user-rated and gold-standard relevance judgments.

	N	Valid *N*[a]	User-rated Mean (SD)	Gold-standard Mean (SD)	Paired *t* (*p*)
All	1,143	854	1.58 (0.52)	0.63 (0.76)	**30.92 (.000)**
Low DK viewed	504	315	1.51 (0.53)	0.56 (0.71)	**19.12 (.000)**
High DK viewed	709	539	1.62 (0.51)	0.66 (0.78)	**24.30 (.000)**
Set L documents	184	86	1.47 (0.57)	0.26 (0.60)	**12.71 (.000)**
Set H documents	292	198	1.64 (0.52)	0.34 (0.64)	**22.03 (.000)**
Set B documents	737	570	1.58 (0.51)	0.78 (0.77)	**21.40 (.000)**

[a] Excluding docs missing user relevance scores.

TABLE 12. Document features comparison between two DK user groups.

	Low DK Mean (SD)	High DK Mean (SD)	Mann–Whitney Z (*p*)
Document length	15.62 (5.87)	16.58 (5.96)	**2.75 (.006)**
Number of MeSH terms	2.01 (2.13)	2.60 (2.13)	**5.46 (.000)**
Number of general MeSH terms	1.62 (1.61)	2.10 (1.60)	**5.33 (.000)**
Number of specific MeSH terms	0.40 (.87)	0.50 (1.08)	1.13 (.26)

TABLE 13. Document features comparison among three document sets.

	Set L Mean (SD)	Set H Mean (SD)	Set B Mean (SD)	Kruskal– Wallis H (*p*)
Document length	15.51 (5.65)	16.25 (5.68)	16.3 (6.10)	.92 (.63)
Number of MeSH terms	2.24 (2.53)	**2.84 (2.54)**	2.2 (1.86)	**11.84 (.003)**
Number of general MeSH terms	**1.59 (1.82)**	2.11 (1.84)	1.89 (1.47)	**12.46 (.002)**
Number of specific MeSH terms	0.65 (1.06)	0.73 (1.33)	**0.31 (0.80)**	**53.28 (.000)**

TABLE 14. Standard test statistics (adjusted *p* values) of pairwise comparison.

	Number of MeSH terms	Number of general MeSH terms	Number of specific MeSH terms
Set L vs. Set H	**−3.21 (.004)**	**−3.29 (.003)**	.36 (1.00)
Set L vs. Set B	−1.38 (.502)	**−3.24 (.004)**	**5.41 (.000)**
Set H vs. Set B	**2.77 (.017)**	.66 (1.00)	**5.97 (.000)**

quickly, in a limited amount of time, return search results to the user, whether or not the user is knowledgeable about the search request makes it quite different in terms of the system's effectiveness. While the high DK users could handle significantly more search results, the low DK users could not. Returning more search results to low DK users could potentially slow down the document selection time for the low DK users.

Second, returning highly accurate results, in this case, highly relevant documents, should be one of the basic requirements for a fast response system. Our results demonstrated that, compared with the low DK level people, those high in DK were able to retrieve more relevant documents (had higher relevance rating scores on the selected documents). This may indicate that higher DK people may comprehend the documents more thoroughly and can recognize relevant documents, although they may not appear to be relevant from a surface reading. Further research is needed to confirm the reasons.

Third, we found 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 was viewed by both groups of users. They selected relevant ones from different sets of documents. We also found that the documents viewed by high DK users ranked lower than those viewed by low DK users. High DK people appeared to go further down to the search result list, to the bottom of an SERP, as well as more SERPs, to look for relevant documents. While further research is needed to find out the reasons for why they did so, one possible explanation could be that high DK people understood what the document was about and they did not mind reading the documents that were determined not as relevant by the system (that is, the system assigned them the higher rank numbers), which may not seem as relevant at first glance. Or possibly, the high DK people were less affected by the system ranking than the low DK users, who may have understood the documents to a lower degree.

Further, we found that both the low and the high DK participants rated higher average relevance scores than the gold-standard. We thought several reasons could possibly 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 from theirs. Third, our experiment tasks were 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.

Finally, document features were found to show differences between those viewed by different DK user groups, and among different document sets. The pattern that compared them with domain novices, domain experts tended to read documents that were longer and had more MeSH terms (for the documents that they viewed differently) could be reasonably due to their higher knowledge levels, which may have helped them better comprehend MeSH terms and longer documents. Document length and number of MeSH terms could be indicators of the document being read by domain experts and novices, and systems could personalize their search result lists for experts and novices based on these document features. Specifically, systems may consider ranking shorter documents and/or documents with fewer MeSH terms on the top ranks for domain novices, and vice versa.

As any research, the current study had limitations. The study focused on the medical domain, and care needs to be taken when generalizing the results to other domains. Also, the participants were students, which may not represent some types of domain experts such as medical scientists and researchers. Future research would consider recruiting domain experts of these types. It would also be interesting to see if there are differences in the document selection behaviors between medical students and scientists/researchers.

Conclusion

Through a controlled lab experiment using the TREC genomics track data, the current study examined the document selection decisions made by experts and novices in an interactive search process in the genomics domain. The results show that, in general, domain experts and domain novices selected different sets of documents to view, and domain experts read more documents and gave higher relevance ratings for the viewed documents than domain novices did. While the domain novices tended to select the documents that ranked on top of the search result lists, domain experts tended to also select documents that ranked toward the bottom of the search result lists. Documents viewed by domain experts and novices were found to have different features, which may contribute to why experts and novices viewed different sets of documents. The findings have implications for designing personalized search systems that can help domain novice users.

Acknowledgment

This research was partially supported by IMLS Grant #LG 06-07-0105-07. We thank the PooDLE project group members for their assistance with data collection and data sharing.

References

Aronson, A.R., & Lang, F.M. (2010). An overview of MetaMap: Historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3), 229–236.

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A.P., & Yilmaz, E. (2008). Relevance assessment: Are judges exchangeable and does it matter? In Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 667–674). New York: ACM.

Cole, M.J., Zhang, X., Liu, C., Belkin, N.J., & Gwizdka, J. (2011). Knowledge effects on document selection in search results pages. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1219–1220). New York: ACM.

Corder, G.W., & Foreman, D.I. (2014). Nonparametric statistics: A step-by-step approach. Hoboken, NJ: Wiley.

Downing, R.E., Moore, J.L., & Brown, S.W. (2005). The effects and interaction of spatial visualization and domain expertise on information seeking. Computers in Human Behavior, 21, 195–209.

Drabenstott, K.M. (2003). Do nondomain experts enlist the strategies of domain experts? Journal of the American Society for Information Science and Technology, 54(9), 836–854.

Duggan G.B. & Payne, S.J. (2008). Knowledge in the head and on the web: Using topic expertise to aid search. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 39–48). New York: ACM.

Dunn, O.J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56(293), 52–64.

Hembrooke, H., Granka, L., & Gay, G. (2005). The effects of expertise and feedback on search term selection and subsequent learning. Journal of the American Society for Information Science and Technology, 56 (8), 861–871.

Hersh, W.R., & Voorhees, E.M. (2009). TREC genomics special issue overview. Information Retrieval, 12(1), 1–15.

Hsieh-Yee, I. (1993). Effects on search experience and subject knowledge on the search tactics of novice and experience searchers. Journal of the American Society for Information Science, 44(3), 161–174.

Kang, R., Fu, W.-T., & Kannampallil, T. (2010). Exploiting knowledge-in-the-head and knowledge-in-the-social-web: Effects of domain expertise on exploratory search in individual and social search environments. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 393–402). New York: ACM.

Kelly, D. & Cool, C. (2002). The effects of topic familiarity on information search behavior. In G. Marchionini (Ed.), Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries (ACM JCDL '02) (pp. 74–75). New York: ACM.

Kreyszig, E. (2011). Advanced engineering mathematics (10th ed.). Hoboken, NJ: Wiley.

Kumaran, G, Jones, R., & Madani, O. (2005). Biasing web search results for topic familiarity. In H-J. Schek, N. Fuhr, A. Chowdhury, & W. Teiken (Eds.), Proceedings of the 14th International Conference on Information and Knowledge Management (ACM CIKM '05) (pp. 271–271). New York: ACM.

Liu, J. & Belkin, N. J. (2010). Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. Proceedings of the 33rd Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR '10), July 19–23, 2010, Geneva, Switzerland (pp. 26–33).

Nguyen, H., & Santos, E., Jr. (2007). Effects of prior knowledge on the effectiveness of a hybrid user model for information retrieval. Proceedings of SPIE, 6538, 65380V.

Palotti, J., Hanbury, A., Muller, H., & Kahn, C.E.J. (2016). How users search and what they search for in the medical domain. Information Retrieval Journal, 19(1), 189–224.

Palotti, J., Zuccon, G., Bernhardt, J., Hanbury, A., & Goeuriot, L. (2016) Assessors agreement: A case study across assessor type, payment levels, query variations and relevance dimensions. In Proceedings of CLEF 2016 (pp. 40–53).

Paukkeri, M.-S., Ollikainen, M., & Honkela, T. (2013). Assessing user-specific difficulty of documents. Information Processing and Management, 49, 198–212.

Sihvonen, A., & Vakkari, P. (2004). Subject knowledge improves interactive query expansion assisted by a thesaurus. Journal of Documentation, 60(6), 673–690.

Taylor, A. (2012). User relevance criteria choices and the information search process. Information Processing and Management, 48, 136–153.

Taylor, A., Zhang, X., & Amadio, W.J. (2009). Examination of relevance criteria choices and the information search process. Journal of Documentation, 65(5), 719–744.

Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. Information Processing & Management, 39(3), 445–463.

Vibert, N., Ros, C., Bigot, L.L., Ramond, M., Gatefin, J., & Rouet, J.-F. (2009). Effects of domain knowledge on reference search with the PubMed database: An experimental study. Journal of the American Society for Information Science and Technology, 60(7), 1423–1447.

Wang, P., & Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. Journal of the American Society for Information Science, 49(2), 115–133.

Wang, P., & White, M.D. (1999). A cognitive model of document use during a research project. Study II. Decisions at the reading and citing stages. Journal of the American Society for Information Science, 50(2), 98–114.

White, R., Dumais, S.T., & Teevan, J. (2009). Characterizing the influence of domain expertise on Web search behavior. In Proceedings of the Second ACM International Conference on Web Search and Data Mining (pp 132–141). New York: ACM.

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.

Zhang, X., Anghelescu, H.G.B., & Yuan, X. (2005). Domain knowledge, search behavior, and search effectiveness of engineering and science students. Information Research, 10(2), 217.

Zhang, X., Liu, J., Cole, M., & Belkin, N.J. (2015). Predicting users' domain knowledge in information retrieval using multiple regression analysis of search behaviors. Journal of the Association for Information Science and Technology, 66(5), 980–1000.

DOI: 10.1002/asi