Evaluation of a Scatter/Gather Interface for Supporting Distinct Health Information Search Tasks

Yan Zhang and Ramona Broussard
School of Information, University of Texas at Austin, 1616 Guadalupe Street, Austin, TX 78701. E-mail: {yanz, ramona}@ischool.utexas.edu

Weimao Ke and Xuemei Gong
College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104. E-mail: {wk, xg45}@drexel.edu

Web search engines are important gateways for users to access health information. This study explored whether a search interface based on the Bing API and enabled by Scatter/Gather, a well-known document-clustering technique, can improve health information searches. Forty participants without medical backgrounds were randomly assigned to two interfaces: a baseline interface that resembles typical web search engines and a Scatter/Gather interface. Both groups performed two lookup and two exploratory health-related tasks. It was found that the baseline group was more likely to rephrase queries and less likely to access general-purpose sites than the Scatter/Gather group when completing exploratory tasks. Otherwise, the two groups did not differ in behavior and task performance, with participants in the Scatter/Gather group largely overlooking the features (key words, clusters, and the recluster function) designed to facilitate the exploration of semantic relationships between information objects, a potentially useful means for users in the rather unfamiliar domain of health. The results suggest a strong effect of users' mental models of search on their use of search interfaces and a high cognitive cost associated with using the Scatter/Gather features. It follows that novel features of a search interface should not only be compatible with users' mental models but also provide sufficient affordance to inform users of how they can be used. Compared with the interface, tasks showed more significant impacts on search behavior. In future studies, more effort should be devoted to identify salient features of health-related information needs.

Introduction

In recent decades, various factors, including rising healthcare costs, the consumerism movement, and the availability of large amounts of health information, have encouraged consumers to become active seekers of health information. According to reports from the Pew Internet & American Life Projects, as of 2011, 80% of Internet users in the United States have searched online for health information, and the information they found had a significant impact on their healthcare decisions (Fox, 2011; Zickuhr, 2010). At the same time, effective access to health information has been promoted as a national priority for achieving population health. In the report *Healthy People 2010*, the U.S. Department of Health and Human Services explicitly outlined how improving consumers' health literacy, that is, their ability to obtain, process, and understand basic health information and services, should become an important component of health communication in any form (U.S. Department of Health and Human Services, 2010).

Consequently, great efforts have been made to make more high-quality health information available online and to provide more readable information to consumers with inadequate literacy skills (Zun, Downey, & Brown, 2011). Despite these efforts, studies continued to suggest that searching for health information remains a difficult task. Many consumers expressed difficulties with the information search process, including articulating needs, assessing relevance, and evaluating quality (Arora et al., 2008; Bundorf, Wagner, Singer, & Baker, 2006). These difficulties, to a large extent, can be attributed to inadequate user support in current search systems (Zhang, 2011). To improve user support, it is necessary to improve the interfaces that mediates user–system interactions.

Defining and designing interfaces that support effective information seeking is a task fundamental to information science (Marchionini & Komlodi, 1998). In the past several years, efforts have been made to improve health information search interfaces. Typical approaches include

Received January 27, 2013; revised April 30, 2013; accepted April 30, 2013

 DOI: 10.1002/asi.23011

query recommendation and expansion based on medical thesauri or user feedback (e.g., Zeng et al., 2006) and faceted display of search results based on document attributes (e.g., Mu, Ryu, & Lu, 2011). However, few have attempted to examine whether document clustering, a technique that automatically groups documents into meaningful semantic clusters and that requires minimal human intervention, could effectively facilitate consumers' exploration and sense-making of the often unfamiliar domain of health information. To explore the potential of this approach, we implemented a search system using Scatter/Gather, one of the best-known document clustering techniques (Hearst, 2009), based on the Bing API. Scatter/Gather was incorporated with Bing because search engines remain the gateway to health information despite the availability of a large number of Web-based health information sources (Fox & Jones, 2009).

This Scatter/Gather-enabled and Bing API-based search system differs from general web search engines in its organization of search results and in its ability to help identify latent associations among documents and topics. Instead of presenting a ranked list of documents, the system organizes search results into topically coherent groups (clusters) and presents descriptive key words for each group. This design offers not only an overview of the result but also a focused view of separate topic clusters. In addition, it allows users to select any number of interesting clusters (gather) and recluster the documents in these clusters (scatter) to explore the selected conceptual space further. In this article we report a between-subjects experimental study to evaluate the effects of this Scatter/Gather system on users' health information search behavior and performance. To prepare for the evaluation, a baseline system that resembles typical web search engines was also created.

Related Literature

Health Information-Searching Behavior and Search Interfaces

A search interface is intended to aid users "in the expression of their information needs, in the formulation of their queries, in the understanding of their search results, and in keeping track of the progress of their information seeking efforts" (Hearst, 2009, p. 1). It is obvious that a search interface serves as a channel through which information searching proceeds, so the design of effective search interfaces hinges on an in-depth understanding of the information search process (Marchionini & Komlodi, 1998).

In studies on consumer health information searching, two major activities involved in the search process—query formulation and access and evaluation of search results—have received much attention. Studies have consistently reported that, similar to general search queries, health-related queries were simple and short, with the majority ranging from one to three terms (Spink et al., 2004; Zeng et al., 2006). Toms and Latter (2008) characterized consumers' query behavior as a trial-and-error process. Zhang, Wang, Heaton, and Winkler (2012) further revealed that, in completing health tasks that explore relationships between different conditions, 55% of participants' query reformulations were intended to make the query conceptually more specific or more general, 30% switched search topics or partially replaced one of the concepts in the previous query with a new concept, and the remaining 15% were query iterations (including re-executing the previous query and replacing it with synonyms).

Examining and evaluating search results are parallel cognitive processes, with the former driving users to perceive certain interface elements and the latter directing users' cognitive resources to make judgments (Zhang, 2012a). When examining search results, users pay attention to titles, URLs, and summaries and search key words (Toms & Latter, 2008) and sometimes terms related to health and medicine, such as MD, doctors, and treatments (Zhang, 2012a). Sillence, Briggs, Fishwick, and Harris (2004) proposed a two-stage model to account for users' evaluation of health information. At the first stage, users quickly reject certain sites based on design factors (e.g., layout and navigational aids); at the second stage, they meticulously select websites based on an appraisal of content factors (e.g., accuracy and readability). These criteria have been identified in several other studies (e.g., Eysenbach & Kohler, 2002).

Studies have also documented users' difficulties with health information searches. One of the most widely recognized difficulties is related to the representation and articulation of information needs. Users often had imprecise representations of their conditions (Keselman, Browne, & Kaufman, 2008; White & Horvitz, 2009) and at the same time lacked proper medical terminologies to describe those conditions. Furthermore, many had difficulty in correctly spelling medical terms (Boden, 2009; Zeng et al., 2006). Because of the idiosyncratic nature of most health conditions, and associated treatments and coping requirements, some users also expressed frustration with finding personally relevant information and with evaluating the quality of information in relation to their specific conditions (Arora et al., 2008).

To accommodate users' health information search processes as well as to alleviate some of the difficulties, researchers have attempted to design novel interfaces and interactive features to support consumer health information searches. For example, Pratt (1997) developed a search system, DynaCat, that organizes results into categories based on subject headings defined in MeSH (Medical Subject Heading) and semantic types defined in UMLS (United Medical Language System). A user study consisting of 15 cancer patients and caregivers indicated that, compared with the traditional document ranking, users found significantly more answers in a fixed amount of time and felt more satisfied. Zeng et al. (2006) developed the Health Information Query Assistant system, which suggests alternative/additional query terms related to a user's initial

DOI: 10.1002/asi

query to make the query more specific based on semantic relationships between terms in medical thesauri, log data, and medical literature. A user study confirmed that the system resulted in higher rates of successful queries but had no impact on users' satisfaction or ability to accomplish search tasks.

Recently, Mu, Ryu, and Lu (2011) implemented and evaluated a faceted search interface for health information retrieval and navigation based on a PubMed data set. They found that the faceted interface significantly reduced the number of clicks users had to perform to find relevant documents. Participants expressed favorable views toward the system's ability in finding relevant documents, in providing related medical terms, and in presenting search results that were better organized.

Despite the new developments, there is abundant room for improving health-related search interfaces. In a review of 18 popular web-based health search interfaces, Zhang (2011) found that basic cognitive assistance functions, such as query autocompletion and spelling check, were absent in most systems. There was also a lack of functions and mechanisms, such as search histories, to assist users in learning and making sense of the information found during the search process.

During the past several decades, an improved understanding of users' information needs and search processes, coupled with the fast development of technologies, has not only fostered the development of more effective search algorithms but also catalyzed the emergence of a wide variety of novel search interfaces. For example, to support the exploration of a document collection, efforts have been made to integrate navigation with search (e.g., faceted search and social tagging), to visualize search results (e.g., visualizing search results as clusters or as tree maps), and to utilize text-mining techniques to identify latent relationships between information objects (e.g., visualizing citation relationships and Scatter/Gather techniques; Capra, Marchionini, Oh, Stutzman, & Zhang, 2007; Dunne, Shneiderman, Gove, Klavans, & Dorr, 2012; Gossen, Nitsche, & Nürnberger, 2012; Gwizdka, 2009; Hearst, 2009; Kules, Capra, Banta, & Sierra, 2009; Wilson, 2011). To explore whether new techniques and search mechanisms could aid users in health information searches, a domain in which help is often needed, this study examines whether Scatter/Gather, one of the best-known document clustering techniques, can effectively support users' completion of health-related search tasks. The next section provides a brief introduction to Scatter/Gather.

Scatter/Gather

Scatter/Gather is a highly interactive model for information retrieval (IR) and collection browsing based on text clustering (Cutting, Karger, Pedersen, & Tukey, 1992). It can stand alone as a browsing tool, allowing users to explore latent associations among documents and topics in a collection (Hearst & Pedersen, 1996; Ke, Sugimoto, & Mostafa, 2009). In each Scatter/Gather iteration, the system presents to the user a set of clusters (topical groups of documents) in the collection. The user then picks one or more clusters in which he or she is interested (gather), upon which the system performs the clustering function again to identify new topical groups (scatter). By providing such iterative user selection and interactive text clustering, Scatter/Gather is able to help users not only clarify their needs but also navigate large, complex information spaces.

However, challenges associated with clustering efficiency and scalability have hindered the adoption of Scatter/Gather in IR. In particular, many clustering algorithms are computationally complex. Even efficient classical methods, such as k-means, are not sufficiently fast to support on-the-fly clustering when the number of documents is large. Therefore, the use of Scatter/Gather for web browsing is desirable but practically challenging because of the web's scale and dynamics. Several approaches have been proposed to boost online clustering efficiency through parallel computing and data precomputation (Jensen, Beitzel, Pilotto, Goharian, & Frieder, 2002; Ke, Mostafa, & Liu, 2008).

Scatter/Gather can also be implemented with a search system to facilitate the navigation and exploration of search results (Gong, Ke, & Khare, 2012; Hearst & Pedersen, 1996). In this hybrid approach, clustering is performed on retrieved documents to identify major themes/topics in the search results, from which users can learn about important vocabularies associated with the topics and delve into interesting clusters for further exploration. Usability studies have suggested that this hybrid approach was more effective than the standalone Scatter/Gather (Pirolli, Schank, Hearst, & Diehl, 1996).

The effectiveness of Scatter/Gather also depends on factors associated with the nature and context of an information need. Prior studies have revealed that Scatter/Gather was more effective when users were familiar with the query topic and when the topic was broad rather than specific (Ke et al., 2009). Further analysis is needed to understand how important contextual factors, such as search tasks, may affect the effectiveness of Scatter/Gather.

Health Information Search Tasks

Tasks have significant effects on information searching behaviors. Different task types also impose different requirements on systems and user interfaces (Vakkari, 2003; Woodruff, Rosenholtz, Morrison, Faulring, & Pirolli, 2002). Health-related search tasks may be characterized by the attributes of conditions of interest (e.g., acute vs. chronic, severity, rarity, complexity, and whether the condition is stigmatized) or by situation-related factors (e.g., searching for self or for others and the specificity and types of the information needed Zhang, in press). Little is known about the effects of these attributes on users' search behaviors.

Attempts have been made to address this research question. Zhang et al. (2012) observed people using MedlinePlus to solve exploratory tasks that make sense of controversial medical subjects or that evaluate relationships between

DOI: 10.1002/asi

conditions or treatments. They found that, when the task became more exploratory, as defined by the number of concepts (conditions or treatments) involved, users relied more on the search function, used more types of sources, made more transitions between searching and browsing, and reformulated queries with higher frequencies. Participants expressed difficulties in exploring relationships between multiple health conditions. By analyzing search queries oriented toward exploratory tasks of diagnosing illness from symptoms, Cartright et al. (2011) found that users took either an evidence-based approach by focusing on looking for details and relevance of symptoms, or a hypothesis-directed approach by focusing on examining facets of one or more illnesses (e.g., treatments) and on the differences among different conditions.

In another study, Zhang (2013a) observed people searching for specific factual health information using MedlinePlus. She found that participants typically used the browsing strategy, issued very few queries, checked out fewer than one result per search, and expected to find answers on the search results page or within a couple of clicks from the results page.

Based on these prior studies, two types of tasks are defined in this study: lookup and exploratory. Lookup tasks are tasks oriented toward finding particular health-related facts, whereas exploratory tasks are oriented toward learning, investigating, and making sense of specific health issues. This categorization of tasks also recently attracted interest in general information search studies (Marchionini, 2006; Wildemuth & Freund, 2009). To evaluate the Scatter/Gather interface in supporting consumers searching for health information, the following research questions are addressed:

1. Do search interfaces, specifically, a baseline search interface and a Scatter/Gather-enabled search interface, affect users' behaviors while completing two types of health-related search tasks: lookup and exploratory?
2. Do the two search interfaces affect users' performance on the lookup and exploratory tasks?
3. What are users' experiences with the two different search interfaces?

At the same time, we explore whether tasks affect users' search behaviors and task performance.

Methods

A 2×2 between-subjects experiment with 40 participants was conducted. The participants were randomly assigned to two interface conditions: the baseline interface or the Scatter/Gather interface. All participants performed both lookup and exploratory tasks.

Platform Systems

As mentioned previously, the following two systems were developed: (a) a baseline search system and (b) a Scatter/Gather-enabled search system. The baseline system is based on the Bing search API. After the user enters a search query, the system sends the query to Bing, retrieves the first 200 results, and presents them in the classic relevance-based sequential order. Figure 1 shows the baseline interface.

The Scatter/Gather system is also based on the Bing search API. In addition, it adopts the Weka machine-learning package for text clustering (Hall et al., 2009). After the user enters a search query, the system sends the query to Bing and retrieves the first 200 results. The Scatter/Gather system then tokenizes the documents, removes stop words, and performs vectorization using the TF*IDF weighting scheme (Robertson, 2004). The top 1,000 frequent words are kept as features for document vector representations (DF thresholding). The k-means implementation in Weka is used for the identification of topical clusters in the results (Arthur & Vassilvitskii, 2007; Witten, Frank, & Hall, 2011). The top 10 key words are identified based on their TF*IDF weights within each cluster. Figure 2 shows the Scatter/Gather interface.

By default seven clusters are shown, but the user can adjust the number of clusters by moving the "Desired number of clusters" slider in the upper right corner. Each cluster panel includes information about the number of results, representative key words, and the first two documents in the cluster. The user can click on "More" to see additional results in the cluster and "Less" to hide them. Similar to the baseline interface, related searches and search history are shown on the left side.

To the right of each cluster, there is also a check box with which the user can select or deselect the cluster. After selecting clusters in which he or she is interested, the user may click the "Gather and scatter" button to gather the selected clusters (gather) and perform reclustering (scatter). As a result, all the documents in the selected clusters are classified into the desired number of clusters (the number can be set using the "Desired number of clusters" slider). The user can repeat this process until a relevant subset is reached.

Tasks

Four search tasks, two simple lookup and two exploratory, were used in the study. The lookup tasks involve finding a specific factual answer to a well-defined question. The answer is located on one page and minimal cognitive effort is required. On the other hand, the exploratory tasks involve finding information to make sense of a health issue and to support decision making. The "answer" for these tasks is less certain, and successful completion of the tasks requires users to compare and synthesize information from multiple pages. Table 1 shows the four tasks. The two lookup tasks were adapted from Zhang (2009). The first exploratory task was adapted from Zhang (2009), and the second was adapted from Mu et al. (2011).

Participants

Upon receipt of IRB approval, a recruitment e-mail was posted on a campus-wide listserv of a large research

DOI: 10.1002/asi

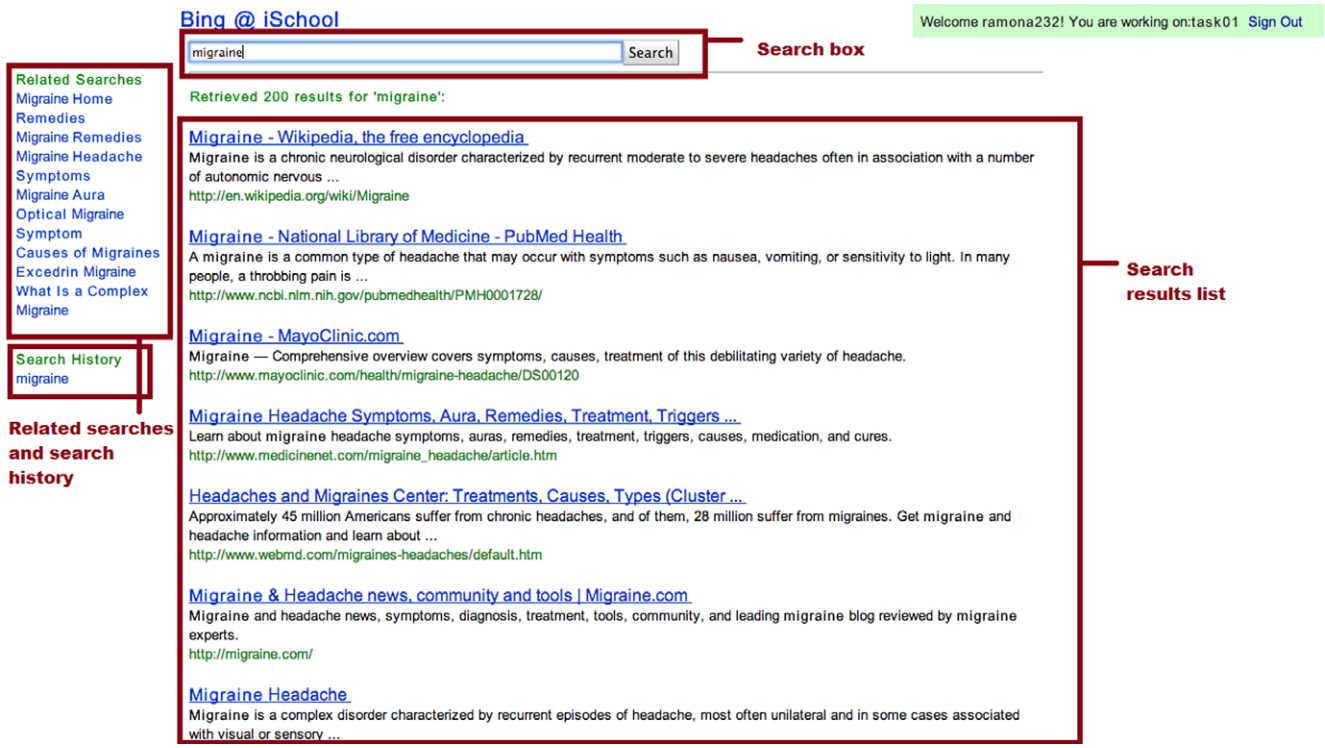

FIG. 1. Baseline search interface. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

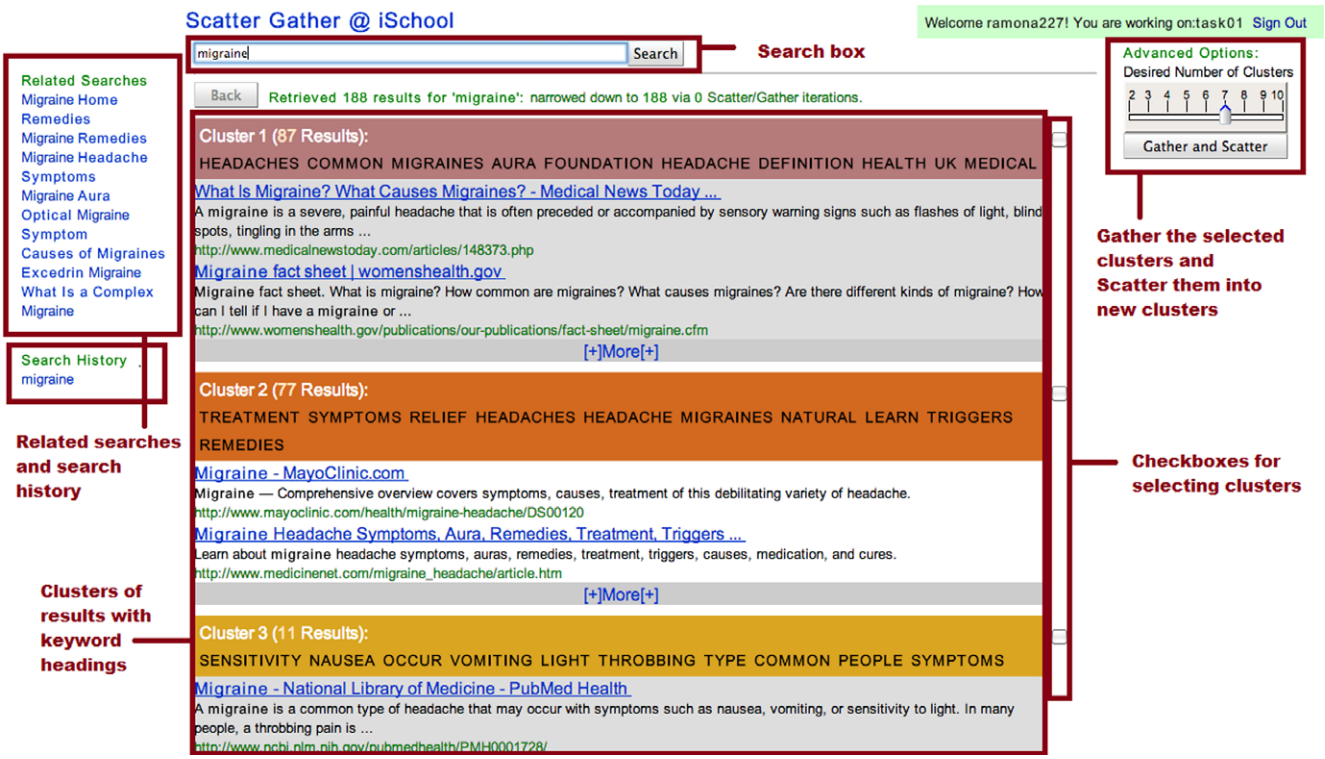

FIG. 2. Scatter/Gather search interface. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DOI: 10.1002/asi

TABLE 1. Search tasks.

Lookup tasks	A friend of yours is an athlete. Now he wants to increase his muscle mass. He has been training without Creatine, but would like to start a regimen. He is seeking your advice on this. You decide to find out what the side effects of taking Creatine are.
	A heart attack is a medical emergency, and prompt treatment increases the chance for survival. According to the American Heart Association, heart attacks cause one of every five deaths. According to the National Institutes of Health (NIH), more than 1.2 million heart attacks occur each year in the United States, and about 460,000 of these are fatal. Approximately 300,000 people die annually from heart attacks before they can receive medical treatment. To be prepared for possible emergencies, you decide to find out what to do when a person around you has a heart attack.
Exploratory tasks	Imagine that one of your close family members has lived with diabetes for years. Recently, he was also diagnosed with hypertension. You decided to do some research on the clinical associations between the two conditions so that you are able to effectively discuss with him about various implications of this diagnosis.
	Imagine that you recently began suffering from migraines. You heard about two possible treatments for migraine headaches, beta-blockers and/or calcium channel blockers, and you decided to do some research about them. At the same time, you want to explore whether there are other options for treating migraines without taking medicines.

university. As a result, 40 participants, including students, staff, and alumni who did not have a medical background, were recruited and participated in the study. Their ages ranged from 18 to 55 years, with 10% younger than 20, 77.5% between 20 and 30, and 12.5% between 30 and 55. Each participant was compensated $15.

Experimental Procedure

Upon arrival, participants were given an overview of the study and asked to read and sign an informed consent document. Next they completed a questionnaire reporting their demographics as well as experience with web search and health information search. Then, participants were randomly assigned to one of the two interface conditions: 20 to the baseline interface and 20 to the Scatter/Gather interface. Each participant completed all four tasks listed in Table 1. The order in which the tasks were presented was randomized to minimize learning effects. Before the search began, participants watched a short video tutorial demonstrating the basic functions of the interface to which they were assigned. The tutorial for the Scatter/Gather interface focuses on explaining the major features (the Scatter/Gather panel, clusters, and key words) of the interface and what these features do to the search results (i.e., the scatter and gather mechanisms). The tutorials for both interfaces were not interactive. Participants' understanding of either interface was not assessed, but they were instructed to ask questions anytime they want.

Participants were given as much time as they needed to complete each task. During the search, whenever they checked out a website from the results list, they were prompted to rank the relevance and usefulness of the site in relation to the task on a 7-point Likert scale (1 [*not relevant*], 7 [*relevant*]; 1 [*not useful*], 7 [*useful*]), which has been identified as the optimal range of numbers for relevance ranking (Tang, Vevea, & Shaw, 1999). The search queries, websites visited, and participants' ratings were logged. As soon as participants completed each task, they were asked to complete a short questionnaire assessing mental effort (how much mental effort they used to complete the task) and satisfaction (how satisfied they were with their performance). The search process was video recorded using Camtasia software.

After completing all four tasks, participants filled out a questionnaire assessing their overall experience with the system that they had used. The questionnaire consisted of a series of statements about users' perceptions of the ease of use and usefulness of the system, understanding of the system's working mechanism, enjoyment, engagement, and intentions to use the system in the future. The rating scale was a 5-point Likert scale (1 [*strongly disagree with the statement*], 5 [*strongly agree*]). Items measuring the ease of use and usefulness were adapted from Capra et al. (2007). The remaining aspects were each measured by a single item. Then, participants were shown a playback of the video of their search behaviors for the last task and asked to comment on their actions, such as selection of key words, reformulation of queries, access of results, and evaluation of websites. After the playback, participants were interviewed about their likes and dislikes about the interface, perceptions of the search tasks, and search experience. The data collection was one-on-one and took place in a private laboratory. All the participants followed the same procedure. Each session lasted approximately 1 to 1.5 hours.

Data Analysis

The main dependent measures were derived from our research questions regarding users' interaction behaviors with the interfaces, task performance, and experience. Interaction behavior was operationalized by three sets of variables: (a) task completion time (measured from the time at which the task was displayed until the participant moved to the questionnaire associated with the task); (b) the number of queries and query terms submitted as well as semantic changes involved in query reformulations; and (c) the number of sites visited, the percentage of sites ranked by the users as relevant and useful (measured on a 7-point Likert scale presented on each page that users viewed from the results list), and the type of sites visited.

The unit of analysis for semantic changes in query reformulations was a query reformulation instance. The coding schema developed by Rieh and Xie (2006) was adopted as an initial framework for the coding, which defined four major

DOI: 10.1002/asi

TABLE 2. Demographics and health information search (HIS) experience of the two groups.

	F	M	Age (years)	Web exp. (years)	HIS exp. (years)	HIS freq. (/month)
BS	13	7	25.9 (9.59)	13.7 (4.57)	3.8 (1.24)	2.7 (.75)
SG	12	8	24.3 (5.84)	12.8 (4.59)	3.6 (1.54)	2.9 (.97)

types of conceptual changes: specification, generalization, parallel movement, and replacement with synonyms. At the same time, open coding was performed to allow new categories to emerge. Two independent coders coded all the query reformulation instances, and intercoder reliability was 98.5%. The discrepancies were resolved by discussion.

Quality is of critical importance for health information. To illuminate the ability of each interface to encourage users to access high-quality information, we categorized the sites that users visited into four types.

1. Evidence-based sites: sites that provide access to empirical medical-related research or sites that belong to reputable health institutes, such as PubMed and Mayo Clinic
2. Health-specific sites, such as WebMD and Healthline
3. General-purpose sites, such as ehow, news sites, and magazines' sites
4. User-generated content sites, such as Wikipedia, Yahoo! Answers, and other online forums

Task performance was measured by two sets of variables: (a) participants' self-reported mental effort for completing each task and satisfaction with their own performance (measured on a 5-point Likert scale presented after each task) and (b) researchers' assessment of participants' performance based on the sites rated by users as useful and relevant. For the lookup tasks, users either found the answer (success) or did not find the answer (failure). For the exploratory tasks, because the nature of the tasks is to explore as many aspects of a particular issue as possible, two factors were taken into account in the assessment: number of sites visited and the quality of the sites (indicated by the four types of sites described previously). The score for each website type (WST) was *evidence-based sites* = 3, *medical-specific sites* = 2, and *general-purpose and user-generated sites* = 1. The performance (P) is then calculated using the formula: $P = \Sigma WST_i$, with i indicating the number of sites visited.

Users' experience with either interface was also measured by two means: (a) the user experience questionnaire that participants completed at the end of the search session (measured on a 5-point Likert scale) and (b) users' descriptions of their perceptions of either interface in the exit interviews. The interviews were transcribed and analyzed by using the qualitative content analysis method. The analysis focused on identifying themes concerning users' experiences.

Results

Demographics

Table 2 summarizes the demographics of the two interface groups: baseline (BS) and Scatter/Gather (SG). Fisher's exact test (Upton, 1992) suggests that the two groups did not differ in gender composition. Indicated by *t*-tests, the two groups also did not differ in age, web experience, health information search experience, or the frequency of searching for health information.

TABLE 3. Mean task completion times in seconds.

	BS (*SD*)	SG (*SD*)
Lookup	446 (155)[a]	409 (184)[b]
Exploratory	578 (205)[a]	530 (177)[b]

Note. a and b denote pairs that differed significantly.

TABLE 4. Average number of queries and terms by task types.

		No. of Q (*SD*)	No. of T (*SD*)
Lookup	BS	2.28 (1.33)[a]	4.78 (1.60)[c]
	SG	2.20 (1.32)[b]	4.88 (1.50)[d]
Exploratory	BS	3.35 (1.61)[a]	3.70 (1.06)[c]
	SG	3.68 (1.57)[b]	3.48 (.83)[d]

Note. a–d denote pairs that differed significantly.

Interaction Behavior

As mentioned in the Data Analysis section, participants' behavior while interacting with the two interfaces was measured in relation to the following three aspects: time to complete tasks, query formulation, and results access.

Task completion time. Table 3 summarizes the mean task completion times for each interface and task type combination. Separate *t*-tests indicate that the two interface groups did not differ in task completion times for either type of task. Paired-sample *t*-tests suggest that task type had an impact on task completion time, with both groups spending significantly more time on the exploratory tasks than on the lookup tasks: BS, $t(19) = 4.02$, $p = .001$; SG, $t(19) = 2.64$, $p = .02$.

Query formulations and reformulations. Query is the main medium through which users interact with a search system (Saracevic, 1997). Thus, users' query formulation and reformulation behavior is an important aspect of user–system interaction. Table 4 shows the average number of queries and query terms submitted to each interface for each task type.

Participants submitted an average of 2.20 to 3.68 queries in completing each type of task. The query length ranged

DOI: 10.1002/asi

TABLE 5. Semantic changes in query reformulations.

	Lookup		Exploratory	
	BS (%)	SG (%)	BS (%)	SG (%)
Specification	15 (30.6)	14 (29.2)	27 (28.4)	30 (28.3)
Generalization	3 (6.1)	4 (8.3)	15 (15.8)	17 (16.0)
Parallel move	16 (32.7)	20 (41.7)	21 (22.1)	28 (26.4)
New concept	6 (12.2)[b]	5 (10.4)[c]	20 (21.1)[b]	25 (23.6)[c]
Rephrase	9 (18.4)	5 (10.4)	12 (12.6)[a]	6 (5.7)[a]
Total	49	48	95	106

Note. a–c indicate pairs that had statistically significant differences.

TABLE 6. Number of sites visited and the relevance and usefulness of the sites.

		No. of sites	Relevant (%)	Useful (%)
Lookup	BS	*4.03 (2.12)*	80.59	69.92
	SG	4.07 (2.42)	84.73	77.04
Exploratory	BS	*5.05 (2.32)*	85.88	74.06
	SG	4.80 (1.82)	80.01	68.74

Note. Italic font indicates that the difference between the pair was statistically significant.

from 3.48 to 4.88 terms. Separate *t*-tests indicate that the two groups did not differ in the number of queries they submitted to solve each type of task or in the average number of terms per query.

Paired-sample *t*-tests indicate that both groups submitted significantly more queries to the exploratory tasks than to the lookup tasks: BS, $t(19) = 3.99$, $p = .001$; Scatter/Gather, $t(19) = 4.51$, $p = .001$; but the length of queries for the exploratory tasks was significantly shorter than those for the lookup tasks: BS, $t(19) = 2.53$, $p = .02$; SG, $t(19) = 3.70$, $p = .002$.

When users' understanding of the problem at hand or of the system changes, they change queries. Thus, query reformulation is a reflection of users' dynamic cognitive activities during a search. To examine whether the two groups differ in their cognitive activities in the problem-solving process, we analyzed the semantic changes involved in query reformulations. Five types of semantic changes were found. They were specification (users specify the meaning of the previous query by adding more terms or replacing terms with those that have more specific meaning); generalization (users generalize the meaning of the previous query by deleting terms or replacing terms with those that have more general meaning); parallel movement (users replace one concept in the previous query with a new concept, and the two queries have partial overlap in meaning or deal with different aspects of one concept); switch to a new concept (users change to new concepts and the new query does not overlap with the previous query); and rephrase (users rephrase the previous query by changing the form of the query without changing the meaning, such as correcting misspellings and rephrasing the previous query into a question). Table 5 shows the distribution of the semantic changes for each interface and task type combination.

Suggested by *t*-tests, the two interface groups did not have significant differences in their query reformulation behaviors. There was one exception: When completing exploratory tasks, the baseline group had a significantly higher percentage of query reformulation instances dedicated to rephrasing a query, that is, changing the form of a query without changing the meaning: $t(38) = 2.26$, $p = .001$.

Tasks showed an impact on query reformulations. Separate paired-sample *t*-tests suggest that both groups made a significantly higher percentage of new concept movements in completing the exploratory tasks than in completing the lookup tasks: BS, $t(19) = 3.31$, $p = .004$; SG, $t(19) = 3.32$, $p = .004$.

Results access. Accessing search results is an important, but less well explored, step in information searching (Marchionini, Capra, & Shah, 2008). Here we report participants' behavior for accessing results in relation to the following three aspects: number of sites visited, relevance and usefulness of these sites, and types of sites visited. Table 6 summarizes the first two aspects for each interface and task type combination.

The two groups did not differ significantly in the number of results checked out for either type of task. In the case of the baseline group, the tasks showed an impact. Participants in this group checked out significantly more results when completing exploratory tasks than when completing lookup tasks: $t(19) = 2.43$, $p = .03$.

In terms of the relevance and usefulness of the results, the Scatter/Gather system returned higher percentages of results rated as relevant and as useful for the lookup tasks, whereas the baseline system returned higher percentages of results rated as relevant and as useful for the exploratory tasks; however, the differences were not statistically significant. Paired-sample *t*-tests indicate that tasks did not have a significant impact on participants' ratings of the relevance and usefulness of the results.

As mentioned in the Data Analysis section, four categories of sites were defined: evidence-based, health-specific, general-purpose, and user-generated. Table 7 summarizes the average number of sites visited by category for each interface and task type combination. Both groups, regardless of task, viewed health-specific sites most frequently, followed by general-purpose sites, user-generated sites, and evidence-based sites. *t*-Tests suggest that the baseline group accessed a higher percentage of health-specific sites (e.g., WebMD, everydayhealth.com) in completing the lookup tasks than the Scatter/Gather group, $t(38) = 1.99$, $p = .05$, whereas the Scatter/Gather group accessed a higher percentage of general-purpose sites (e.g., ehow and CNN) in completing the exploratory tasks, $t(38) = 2.02$, $p = .05$. Participants in the Scatter/Gather group accessed higher percentages of evidence-based sites (e.g., PubMD and MedlinePlus) and user-generated sites (e.g., Wikipedia and

DOI: 10.1002/asi

TABLE 7. Average number of sites visited by category.

		Evidence-based (%)	Health-specific (%)	General-purpose (%)	User-generated (%)
Lookup	BS	7 (4.4)	88 (55.0)[a]	41 (25.6)[c]	24 (15.0)
	SG	15 (9.3)	65 (40.4)[a,d]	45 (28.0)[e]	36 (22.4)
Exploratory	BS	13 (6.5)	140 (70.0)	12 (6.0)[b,c]	35 (17.5)
	SG	23 (12.1)	108 (56.8)[d]	25 (13.2)[b,e]	34 (17.9)

Note. a–e indicate pairs that had statistically significant differences.

TABLE 8. Self-reported mental effort and satisfaction with performance, along with performance grading by researchers.

		Mental effort	Satisfaction with performance	Performance grading
Lookup	BS	2.55 (.48)	4.10 (.53)	100%
	SG	*2.28 (.77)*	4.22 (.44)	100%
Exploratory	BS	2.57 (.54)	4.00 (.54)	10.79 (5.48)
	SG	*2.65 (.59)*	4.05 (.65)	8.45 (4.44)

Note. Italic font indicates that the difference between the pair was statistically significant.

Yahoo! Answers) for both task types, but the differences were not statistically significant.

The type of task (simple lookup or exploratory) also had an impact on the access of results. The baseline group accessed a higher percentage of general-purpose sites when completing lookup tasks than when completing the exploratory tasks: $t(19) = 5.78$, $p = .001$. The Scatter/Gather group accessed significantly higher percentages of health-specific sites when completing the exploratory tasks than when completing the lookup tasks, $t(19) = 2.16$, $p = .04$, and accessed more general-purpose sites when completing the lookup tasks, $t(19) = 2.90$, $p = .01$.

Task Performance

Task performance was measured by two means. One was participants' self-report of mental effort and satisfaction with performance; the other was a performance assessment made by the researchers based on the sites visited by the participants. (The rating criteria are explained in the Data Analysis section.) Table 8 summarizes these measurements.

Separate *t*-tests indicate that there were no significant differences between the two groups in their self-reported mental effort and satisfaction with performance for either type of task. Tasks had an impact on participants' mental effort for the Scatter/Gather group: They felt that completing exploratory tasks required significantly greater mental effort than completing the lookup tasks: $t(19) = 2.78$, $p = .01$. Researchers' examination of the results rated by participants as both relevant and useful indicated that both groups were able to find answers to the lookup tasks successfully and that the two groups did not differ in the performance grading for exploratory tasks.

User Experience

After completing all four tasks, participants completed a user experience questionnaire evaluating their experience with both of interfaces. Table 9 summarizes their ratings.

TABLE 9. User experience ratings of the two interfaces: 1 (*strongly disagree*), 5 (*strongly agree*).

	BS	SG
Ease of use	4.35 (.48)	4.11 (.53)
Usefulness	3.93 (.65)	3.88 (.85)
Understand system	3.78 (.73)	3.63 (.72)
Enjoyment	3.75 (.85)	3.65 (.88)
Engagement	3.10 (.91)	3.45 (1.00)
Future use	3.43 (.94)	3.65 (1.03)

The baseline interface received higher ratings on the majority of measurements, including ease of use, usefulness, understanding of the system, and enjoyment; however, *t*-tests suggest that the differences were not statistically significant. The Scatter/Gather interface received a higher rating on engagement and future use, but the differences were also not statistically significant.

The analysis of the exit interviews revealed that participants expressed difficulties with medical terms, but, overall, they spoke positively about the interfaces, regardless of which one they used. With regard to the baseline interface, participants expressed a feeling of comfort as a result of its familiarity, for example, one participant commented about the baseline interface:

> I liked it because it's in similar format to Google or Bing, something that I would already use, but other than that . . . I didn't dislike anything really. It was fairly straightforward.

Participants expressed mixed feelings with regard to the Scatter/Gather interface. On one hand, some liked the color-coding of different clusters, suggesting that it helped "remember which one was that you are going for." Some

DOI: 10.1002/asi

reported favoring the fact that the interface grouped similar results into clusters, with a few further commenting that the clusters or "bundles" demonstrated correlations between results or groups of results. However, no participant was able to specify the substance of the correlations. On the other hand, some participants noted difficulties in interacting with the Scatter/Gather interface, which were associated mainly with the key words and the clusters. For key words, participants pointed out that the key words appearing at the top of each cluster were easy to recognize, but they were positioned out of context, making it hard to understand what they represented. Correspondingly, the difficulty with clusters lay in understanding why results were grouped in the way presented. One participant's comment illustrated this struggle:

> It took me a second to figure out what the bold words were in each of the colored sections. It looked to me they were sort of key words, but it looked like a sentence or a phrase. There were a lot of words in there, so it was not easy to quickly identify what the category was or what the grouping was. So I didn't find those particularly helpful, just because it was not clear right away how they were being grouped.

In addition, the color-coding of results was found by some participants to be perceptually confusing and misleading, as one of them commented:

> I didn't like the different colors [associated with] the clusters, because, in some ways, it intrigued me to a certain color. . . . I would always look at the green, and I would never look at the yellow and orange. . . .

The other difficulty expressed by the participants was associated with the visual aspect, or the "look and feel," of the interface. Several participants noted that the color schema of the interface was not appealing and that the results pages were cluttered and overwhelming.

One of the features distinguishing Scatter/Gather from most clustering methods is its ability to support reclustering of a selected subset of clusters, allowing users to drill down to reach a more refined conceptual space. However, only one participant in the study used the Scatter/Gather function, reflected as two interface elements: the check box associated with each cluster that allows users to select the cluster (gather) and the cluster slider (on the right side of the interface) that allows users to specify how many clusters they want to generate based on the selected subset of clusters (scatter). Most participants did not even notice the Scatter/Gather function. Several participants noticed it but believed that it was not necessary to use it, as one commented:

> I didn't use these little check boxes on the side because I didn't really feel there is anything I wanted to change about that. I didn't even think to bother with it.

Another commented:

> I guess I didn't think [recluster] was necessary. My understanding was it will give me more or less these clumped options, I didn't feel [that the results] were overwhelming for me. Also, I didn't scroll all the way down so I didn't feel I needed more options. I felt there really wasn't anything that I needed to change about them.

Discussion

In this study, participants' search behavior in both the baseline and the Scatter/Gather search interfaces was examined from three aspects: time taken to complete the tasks, queries and query reformulations, and results access. Task performance was measured by self-reported mental effort and satisfaction as well as researchers' assessments based on the search results visited. The results indicate that the two interface groups spent about equal amounts of time, submitted a similar number of queries, showed similar patterns in query reformulations, visited an equal number of sites, and rated similar percentages of results visited as relevant and useful. Unsurprisingly, both interface groups' performance ratings (on both task types) and their experience with the corresponding interface were similar as well. An observation of the video-recorded search sessions further revealed that, although one of the intended uses of Scatter/Gather is helping users navigate and explore a complex information space (Hearst & Pedersen, 1996), participants used the results clusters in the Scatter/Gather interface in the same way as they were using search results in regular search engines: they checked out results ranked at the top and rarely used the Scatter/Gather function (we observed only one participant using the reclustering function).

These results may be an illustration of mental models at work. The mental model theory posits that users have a set of predefined assumptions (mental model) about how a type of system should work (Carroll & Olson, 1987; Johnson-Laird, 1983; Zhang, 2010). When users use a system of a particular type, this mental model will be activated, operated, and revised as a result of the interplay of the system, tasks, and time (Zhang, 2013b). Prior research has consistently suggested that users have a simple mental model of search, which consists of a search box, a search button, and a list of ranked results (Nielsen, 2005; Zhang, 2008). Compared with typical search systems, the hybrid Scatter/Gather system implemented in this study not only allows users to search by key words but also organizes results into clusters based on topic similarities, provides key words to represent each cluster, and allows users to explore semantic relationships between results and between topics through a gather and scatter process (Hearst, 2009). The similarities of participants' behaviors in the two interfaces, their equivalent performance, and the nonuse of the novel features indicate that the participants may have failed to build a proper mental model of the Scatter/Gather system.

Postsearch session interviews with the participants provided further insight into this failure. Participants acknowledged the difficulty in understanding two important elements

DOI: 10.1002/asi

of the Scatter/Gather interface, key words and clusters, particularly what they are, where they come from, and what they can do. With regard to the clusters, most participants overlooked them. Several participants noticed them but thought that they were ranked by relevance (as in a generic Web search engine), when the clusters were actually ranked by size (the number of hits in a cluster). For example, one participant commented that "The [clusters] at the top [of the result list] were more relevant I think. So that was good." This misconception suggests that the current Scatter/Gather interface failed to help users understand the system. A redesign should provide explicit explanations about what the key words and the clusters represent and how they are generated. Furthermore, it should display key words and clusters in a way that clarifies to users what they can do with the features; in other words, improve the features' affordances (Norman, 1988).

As mentioned, mental model construction is affected by tasks at hand (Zhang, 2012b, 2013b). Thus, participants' difficulties in understanding key words, clusters, and the Scatter/Gather function may also be attributed to their unfamiliarity with the subject domain of health (as they acknowledged in the postsearch session interviews). Numerous prior studies have suggested that lay users have difficulties in understanding medical terminologies and relationships between them (e.g., Boden, 2009); they tend not to know that unfamiliar terms contained in multiple documents discuss the same information and tend not to be able to discern how documents are related to one another (Mu et al., 2011). It is possible that users might have been less confused with the key words, clusters, and the Scatter/Gather function if the tasks were from familiar domains (e.g., shopping and entertainment).

Participants' failure to construct a proper mental model of the Scatter/Gather system, particularly as manifested by the underuse of the Scatter/Gather function, may also result from the cognitive cost associated with making sense of and using the function. Scatter and gather are two complex conceptual processes involving dynamic defining and refining of conceptual spaces. However, the current system provides only check boxes for gathering and parameters (number of clusters) for scattering. Both were at an abstract level, failing to demonstrate explicitly what changes each selection would entail. The current placement of the Scatter/Gather function to the far left on the search results page might also have contributed to the cognitive load; prior research has indicated that users' focus of attention is the results list during a search, and a display of separated facets or hierarchy could be distracting and overwhelming to users (Kules & Shneiderman, 2008). In future research, we will identify the kinds of cognitive load that the gathering and scattering processes may entail and explore ways to reduce the load. A visual interface that clearly delineates each step of the operation and presents conceptual changes in vivo will be one promising direction (Ke et al., 2009).

It is worth noting that training might also play a role in participants' failure to construct a mental model of and adapt to the Scatter/Gather interface. Although participants viewed a short video tutorial explaining the Scatter and Gather features before they approached the tasks, it is possible that the tutorial, in its current form, was not sufficient in helping them learn how the system works. In future studies, it would be worthwhile to assess users' knowledge of the Scatter/Gather interface before they proceed to tasks.

Despite the overall similarity, the two interface groups did show significant differences in two behavioral measurements. One is query reformulation: In completing exploratory tasks, the baseline group had a higher percentage of query reformulation instances belonging to rephrasing queries (query reformulation instances that do not change the conceptual meaning of the queries) than the Scatter/Gather group (the difference in completing the lookup tasks was in the same direction but not statistically significant). This may be because users are more likely to focus on a preconceived idea and more reluctant to consider alternative possibilities in a rather familiar interface environment (Baron, 2000). Similar behavior (i.e., users keep trying the search strategy that failed them in the first place) has been observed in users' use of online catalog systems (Dickson, 1984).

The other difference between the groups was the sites visited. In completing the lookup tasks, a significantly higher percentage of the sites visited by the baseline group were health-specific sites (e.g., WebMD and Livestrong), and, in completing the exploratory tasks, a significantly higher percentage of the sites visited by the Scatter/Gather group were general-purpose sites. Such differences may be accounted for by the working mechanism of the Scatter/Gather system. As mentioned, in the Scatter/Gather system, results were arranged into clusters, and the clusters were sorted by size, with the largest cluster appearing at the top. It is possible that health-specific sites used very different terms in the content than other types of sites, so that they formed smaller clusters and appeared farther down in the results list. General-purpose sites (e.g., Wikipedia and ehow) tended to appear in larger clusters and were more likely to be at the top of the results. This result suggests that more thought should go into result ranking in designing clustering-based search systems. Specific to health information search, one possible strategy is to rank the results in increasing order of cluster size, as opposed to what we did in the study, so that those "unique" or "special" health-specific sites will appear first.

Compared with interfaces, tasks appeared to have a greater impact on participants' behaviors. In both interface conditions, participants spent significantly more time, submitted more queries, visited more sites, and exerted greater mental effort to complete the exploratory tasks. These results are consistent with those of Zhang et al. (2012). The query length in this study ranged from 3.48 to 4.88 terms, which is slightly longer than lengths reported from previous studies on health queries (e.g., Spink et al., 2004). It is worth noting that queries for the lookup tasks were significantly longer than those for the exploratory tasks. An examination revealed that about 20% of the queries for the

DOI: 10.1002/asi

lookup tasks were phrased as questions (e.g., how to respond to a heart attack), which contributes to longer queries.

Tasks also impacted query reformulations, with the exploratory tasks eliciting a significantly higher percentage of the new concept movement in both interfaces. This result suggests that participants' representation of the exploratory tasks encompassed a more complex network of concepts. In addition, tasks impacted the results visited. Both groups accessed a significantly higher percentage of general-purpose sites when completing the lookup tasks and accessed a higher percentage of health-specific sites when completing the exploratory tasks. These results indicate that participants tended to rely more on general-purpose sites (e.g., Wikipedia and ehow) when looking for specific answers and rely more on health-specific sites when exploring a medical problem or trying to make a decision.

It is also worth noting that, without prompting, participants also commented on the look and feel of the two interfaces, particularly the color and the density of information. Look and feel have been greatly emphasized in the design of many different types of websites, including health sites, because they directly impact a site's perceived credibility (Eysenbach & Kohler, 2002; Fogg et al., 2001). This study adds to this recognition by demonstrating that look and feel are also important for a search interface.

The purpose of the study is to examine whether a Scatter/Gather-enabled search system would be more effective than a baseline system in assisting users in completing health-related tasks. Thus, a controlled experiment is an appropriate method. Nevertheless, it should be recognized that the study has a few limitations. First, the participants were a group of computer-literate users, with the majority being young and well educated. Thus the behaviors demonstrated should not be generalized to other health consumers. Second, because the interface was a between-subjects factor, participants did not use both the Scatter/Gather and the baseline systems and thus could not provide more direct comments on the pros and cons of each interface. The third limitation lies in the researchers' evaluation of participants' performance based on the type of site visited, particularly the value ordering of different types of sites. This evaluation is based on simple heuristics that one type of site provides better quality health information than other types of sites. In future studies, to improve the objectivity of the evaluation, quality of information should be accessed on a case-by-case basis by domain experts.

Conclusions

Web search engines are gateways for consumers to access health information. Today's most typical search interface, characterized by type-keywords-in-an-entry-form and view-results-in-a-vertical-list, provides limited cognitive assistance to users searching for health information, a domain in which they need help with both terms and concepts. In this study, we explored whether Scatter/Gather, a well-known document-clustering technique that identifies major topic clusters in a document collection, can improve health information access. We found that, overall, the Scatter/Gather implementation in the study was not superior to the baseline system. Participants used the two systems in a very similar way and achieved equivalent performance. The results help to shed light on the perceptual and cognitive processes underlying users' interactions with search interfaces by suggesting that users have a simple and sticky mental model of how search engines work and how to use a search engine. This mental model is at work when they search for health information. This understanding implies that, when designing novel interfaces for health information search, designers should try to make the new features compatible with users' existing mental models of search to ensure easy adoption. If a new feature tends to break out of the paradigm of the simple mental model of search, such as clustering or visualized display of results (e.g., a tree map view), it should be able to reduce the cognitive demand by effectively conveying its meaning to users and by providing enough affordance so that users are clear about what they can do with the feature.

Compared with the interfaces, task type (lookup vs. exploratory tasks) had a more significant impact on users' health information search behavior. It follows that efforts should be made in future studies to identify salient features of consumers' health information needs and explore the relationships between the features and search behaviors. Such knowledge is necessary for the design of more effective health information search interfaces.

Acknowledgments

We thank all of our participants for their valuable time and input. This work was partially supported by the Alumni Fellowship from the School of Information at the University of Texas at Austin.

References

Arora, N., Hesse, B., Rimer, B., Viswanath, K., Clayman, M., & Croyle, R. (2008). Frustrated and confused: The American public rates its cancer-related information-seeking experiences. Journal of General Internal Medicine, 23(3), 223–228.

Arthur, D., & Vassilvitskii, S. (2007). k-Means++: The advantages of careful seeding. In Proceedings of the SIAM '07, pp. 1027–1035.

Baron, J. (2000). Thinking and deciding. New York: Cambridge University Press.

Boden, C. (2009). Overcoming the linguistic divide: A barrier to consumer health information. Journal Canadian Health Libraries Association, 30(3), 75–80.

Bundorf, M.K., Wagner, T.H., Singer, S.J., & Baker, L.C. (2006). Who searches the Internet for health information? Health Services Research, 41, 819–836.

Capra, R., Marchionini, G., Oh, J.S., Stutzman, F., & Zhang, Y. (2007). Effects of structure and interaction style on distinct search tasks. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 442–451.

Carroll, J.M., & Olson, J.R. (1987). Mental models in human–computer interaction: Research issues about what the user of software knows. Committee on Human Factors, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

DOI: 10.1002/asi

Cartright, M., White, R.W., & Horvitz, E. (2011). Intentions and attention in exploratory health search. In Proceedings of the ACM SIGIR, pp. 65–74.
Cutting, D., Karger, D., Pedersen, J., & Tukey, J. (1992). Scatter/Gather: A cluster-based approach to browsing large document collections. In Proceedings of the ACM SIGIR, pp. 318–329.
Dickson, J. (1984). An analysis of user errors in searching an online catalog. Cataloging and Classification Quarterly, 4(3), 19–38.
Dunne, C., Shneiderman, B., Gove, R., Klavans, J., & Dorr, B. (2012). Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. Journal of the American Society for Information Science and Technology, 63(12), 2351–2369.
Eysenbach, G., & Kohler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. British Medical Journal 324, 573–577.
Fogg, B.J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., . . . Treinen, M. (2001). What makes web sites credible? A report on a large quantitative study. In Proceedings of the Conference on Human Factors in Computing Systems, pp. 61–68.
Fox, S. (2011). The social life of health information. Pew Internet and American Life Project. Retrieved from http://www.pewinternet.org/~/media/Files/Reports/2011/PIP_Social_Life_of_Health_Info.pdf
Fox, S., & Jones, S. (2009). The social life of health information. Pew Internet and American Life Project. Retrieved from http://www.pewinternet.org/~/media//Files/Reports/2009/PIP_Health_2009.pdf
Gong, X., Ke, W., & Khare, R. (2012). Studying Scatter/Gather browsing for web search. In Proceedings of the American Society for Information Science and Technology.
Gossen, T., Nitsche, M., & Nürnberger, A. (2012). Knowledge journey: A web search interface for young users. In Proceedings of the Symposium on Human–Computer Interaction and Information Retrieval 2012. Article No. 1.
Gwizdka, J. (2009). What a difference a tag cloud makes: Effects of tasks and cognitive abilities on search results interface use. Information Research, 14(4). Retrieved from http://informationr.net/ir/14-4/paper414.html
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009). The weka data mining software: An update. SIGKDD Exploration Newsletter, 11(1), 10–18.
Hearst, M.A. (2009). Search user interfaces. New York: Cambridge University Press.
Hearst, M., & Pedersen, J. (1996). Reexamining the cluster hypothesis. In Proceedings of the SIGIR 1996, pp. 76–84.
Jensen, E.C., Beitzel, S.M., Pilotto, A.J., Goharian, N., & Frieder, O. (2002). Parallelizing the Buckshot algorithm for efficient document clustering. In Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM), pp. 684–686.
Johnson-Laird, P.N. (1983). Mental models. Cambridge, MA: Harvard University Press.
Ke, W., Mostafa, J., & Liu, Y. (2008). Toward responsive visualization services for scatter/gather browsing. In Proceedings of the American Society for Information Science and Technology, 45(1), 1–10.
Ke, W., Sugimoto, C.R., & Mostafa, J. (2009). Dynamicity vs. effectiveness: Studying online clustering for scatter/gather. In Proceedings of the ACM SIGIR '09, pp. 19–26.
Keselman, A., Browne, A.C., & Kaufman, D.R. (2008). Consumer health information seeking as hypothesis testing. Journal of the American Medical Informatics Association, 15(4), 484–495.
Kules, B., & Shneiderman, B. (2008). Users can change their web search tactics: Design guidelines for categorized overviews. Information Processing and Management, 44(2), 463–484.
Kules, B., Capra, R., Banta, M., & Sierra, T. (2009). What do exploratory searchers look at in a faceted search interface? Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 313–322.
Marchionini, G. (2006). Exploratory search. Communications of the ACM, 49(4), 41–46.
Marchionini, G., & Komlodi, A. (1998). Design of interfaces for information seeking. Annual Review of Information Science and Technology, 33, 89–120.
Marchionini, G., Capra, R., & Shah, C. (2008). Focus on results: personal and group information seeking over time. In Proceedings of the Workshop on Human–Computer Interaction and Information Retrieval (HCIR) 2008.
Mu, X., Ryu, H., & Lu, K. (2011). Supporting effective health and biomedical information retrieval and navigation: A novel facet view interface evaluation. Journal of Biomedical Informatics, 44(4), 576–586.
Nielsen, J. (2005). Mental models for search are getting firmer. Retrieved from http://www.useit.com/articles/mental-models-for-search/
Norman, D. (1988). The design of everyday things. New York: Basic Books.
Pirolli, P., Schank, P., Hearst, M.A., & Diehl, C. (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. In the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 1996, pp. 213–220.
Pratt, W. (1997). Dynamic organization of search results using the UMLS. Journal of the American Medical Informatics Association, Suppl. S, 480–484.
Rieh, S.Y., & Xie, H. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. Information Processing and Management, 42(3), 751–768.
Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for idf. Journal of Documentation, 60, 503–520.
Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. Proceedings of the ASIS Annual Meeting, 34, 313–327.
Sillence, E., Briggs, P., Fishwick, L., & Harris, P. (2004). Trust and mistrust of online health sites. In Proceedings of the ACM CHI 2004, pp. 663–670.
Spink, A., Yang, Y., Jansen, J., Nykanen, P., Lorence, D.P., Ozmutlu, S., & Ozmutlu, H.C. (2004). A study of medical and health queries to web search engines. Health Information and Libraries Journal, 21, 44–51.
Tang, R., Vevea, J.L., & Shaw, W.M. (1999). Towards the identification of the optimal number of relevance categories. Journal of the American Society for Information Science, 50(3), 254–264.
Toms, E., & Latter, C. (2008). How consumers search for health information. Health Informatics Journal, 13(3), 223–235.
U.S. Department of Health and Human Services. (2010). Retrieved from the Department of Health and Human Services: CDC website on March 17, 2010 from http://www.cdc.gov/nchs/data/hpdata2010/hp2010_final_review.pdf
Upton, G.J. (1992). Fisher's exact test. Journal of the Royal Statistical Society. Series A (Statistics in Society), pp. 395–402.
Vakkari, P. (2003). Task-based information searching. Annual Review of Information Science and Technology, 37, 413–464.
White, R.W., & Horvotiz, E. (2009). Cyberchondria: Studies of the escalation of medical concerns in web search. ACM Transactions on Information Systems, 27(4), Article 23.
Wildemuth, B., & Freund, L. (2009). Search tasks and their role in studies of search behaviors. In Proceedings of the 3rd Workshop on Human–Computer Interaction and Information Retrieval (HCIR).
Wilson, M.L. (2011). Search user interface design. Synthesis Lectures on Information Concepts, Retrieval, and Services, 3(3), 1–143.
Witten, I.H., Frank, E., & Hall, M. (2011). Data mining: Practical machine learning tools and techniques, 3rd ed. San Francisco: Morgan Kaufmann.
Woodruff, A., Rosenholtz, R., Morrison, J.B., Faulring, A., & Pirolli, P. (2002). A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for web search tasks. Journal of the American Society for Information Science and Technology, 53(2), 172–185.
Zeng, Q.T., Crowell, J., Plovinick, R.M., Kim, E., Ngo, L., & Dibble, E. (2006). Assisting consumer health information retrieval with query recommendations. Journal of American Medical Informatics Association, 13(1): 80–90.
Zhang, Y. (2008). Undergraduate students' mental models of the web as an information retrieval system. Journal of the American Society for Information Science and Technology, 59(13), 2087–2098.
Zhang, Y. (2009). The construction of mental models of information-rich web spaces: The development process and the impact of task complexity. PhD dissertation, University of North Carolina at Chapel Hill.

Zhang, Y. (2010). Dimensions and elements of people's mental models of an information-rich web space. Journal of the American Society for Information Science and Technology, 61(11), 2206–2218.

Zhang, Y. (2011). A review of search interfaces in consumer health websites. In Proceedings of the Workshop on Human–Computer Interaction and Information Retrieval (HCIR) 2011. Mountain View, CA.

Zhang, Y. (2012a, October). Consumer health information searching process in real life settings. In Proceedings of the 75th Annual Conference of the American Society for Information Science & Technology (ASIST '12). Baltimore, MD.

Zhang, Y. (2012b). The impact of task complexity on people's mental models of MedlinePlus. Information Processing and Management, 48(1), 107–119.

Zhang, Y. (in press). Beyond quality and accessibility: Source selections in consumer health information searching. Journal of the American Society for Information Science and Technology.

Zhang, Y. (2013a). Searching for specific health-related information in MedlinePlus: Behavioral patterns and user experience. Journal of the American Society for Information Science and Technology (in press).

Zhang, Y. (2013b). The development of users' mental models of MedlinePlus in information searching. Library and Information Science Research, 35, 159–170.

Zhang, Y., Wang, P., Heaton, A., & Winkler, H. (2012). Health information searching behavior in MedlinePlus and the impact of tasks. In Proceedings of the ACM International Health Informatics (IHI) Symposium 2012, Miami, FL, pp. 641–650.

Zickuhr, K. (2010). Generations 2010. Pew Internet and American Life Project (2010). Retrieved from http://pewinternet.org/~/media//Files/Reports/2010/PIP_Generations_and_Tech10.pdf

Zun, L.S., Downey, L., & Brown, S. (2011). Completeness and accuracy of emergency medical information on the web: Update 2008. Western Journal of Emergency Medicine, 12(4), 448–454.

DOI: 10.1002/asi