The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0022-0418.htm

Search result list evaluation versus document evaluation: similarities and differences

Iris Xie and Edward Benoit III
School of Information Studies, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

Received 29 September 2011
Revised 26 March 2012
Accepted 6 May 2012

Abstract

Purpose – The purpose of this study is to compare the evaluation of search result lists and documents, in particular evaluation criteria, elements, association between criteria and elements, pre/post and evaluation activities, and the time spent on evaluation.

Design/methodology/approach – The study analyzed the data collected from 31 general users through prequestionnaires, think aloud protocols and logs, and post questionnaires. Types of evaluation criteria, elements, associations between criteria and elements, evaluation activities and their associated pre/post activities, and time were analyzed based on open coding.

Findings – The study identifies the similarities and differences of list and document evaluation by analyzing 21 evaluation criteria applied, 13 evaluation elements examined, pre/post and evaluation activities performed and time spent. In addition, the authors also explored the time spent in evaluating lists and documents for different types of tasks.

Research limitations/implications – This study helps researchers understand the nature of list and document evaluation. Additionally, this study connects elements that participants examined to criteria they applied, and further reveals problems associated with the lack of integration between list and document evaluation. The findings of this study suggest more elements, especially at list level, be available to support users applying their evaluation criteria. Integration of list and document evaluation and integration of pre, evaluation and post evaluation activities for the interface design is the absolute solution for effective evaluation.

Originality/value – This study fills a gap in current research in relation to the comparison of list and document evaluation.

Keywords Document evaluation, Search result list evaluation, Comparison, Evaluation criteria, Relevance criteria, Evaluation elements, Evaluation activities, Evaluation time, Information retrieval, Searching

Paper type Research paper

Introduction

Information retrieval requires users' evaluation of both search result lists and documents. Evaluation is one of the key search activities in the information retrieval process. The evaluation process involves a high degree of complexity and cognitive abilities, while representing primary information filtering and value judgments. The terms "lists" and "surrogates" are used interchangeably in LIS literature. Evaluation of

The authors thank the University of Wisconsin-Milwaukee for its Research Growth Initiative program for generously funding the project, and Tim Blomquist and Marilyn Antkowiak for their assistance on data collection and Huan Zhang for her assistance on data analysis. The authors would also like to thank the anonymous reviewers for their constructive comments.

Journal of Documentation
Vol. 69 No. 1, 2013
pp. 49-80

0022-0418
DOI 10.1108/00220411311295324

search result lists, hereafter referred to as lists, is defined as the identification of useful or relevant documents from the retrieved listing of document surrogates and/or the overall list page. Although list evaluation focuses on surrogates, the organization of lists and the overall layout of result webpages occasionally influence the evaluation. Evaluation of lists is broader than evaluation of document surrogates. Document evaluation, on the other hand, assesses the usefulness or relevance of an individual document retrieved or browsed.

Previous studies focused more on relevance criteria than elements or other dimensions of evaluation. Studies comparing both list and document evaluation criteria focus on relevance (Bade, 2007; Borlund, 2003; Saracevic, 2007a,b; Savolainen and Kari, 2006; Vakkari and Hakala, 2000). Search result list evaluation receives significantly less attention than document evaluation. Its research remains divided between coverage aspect of relevance (Park, 1994; Purgailis Parker and Johnson, 1990) and quality aspect of relevance (Rieh, 2002; Su, 1994; Voiskunskii, 1997). The majority of current research on evaluation remains focused on coverage aspect of relevance criteria within document evaluation (Bales and Wang, 2005; Barry, 1994, 1998; Greisdorf, 2003; Schamber *et al.*, 1990; Spink *et al.*, 1998; although credibility (Fogg *et al.*, 2003; Metzger *et al.*, 2003; Xu and Chen, 2006) and other aspects of relevance criteria (Barry and Schamber, 1998; Wang and Soergel, 1998, 1999) occasionally appear.

The elements evaluated attract little attention for both result list evaluation (Aula *et al.*, 2005; Granka *et al.*, 2004; Rele and Duchowski, 2005) and document evaluation (Kelly *et al.*, 2002; Maglaughlin and Sonnenwald, 2002; Tombros *et al.*, 2005; Wang and Soergel, 1998). Fewer studies examine the evaluation activities, with the majority focused on list evaluation (Chi *et al.*, 2001; Joachims *et al.*, 2005; Shneiderman *et al.*, 1997) or evaluation time; either from a list evaluation perspective (Hoeber and Yang, 2009; Jansen and Spink, 2006; Jansen *et al.*, 2009; Silverstein *et al.*, 1999; Spink *et al.*, 2001) or a document evaluation perspective (Hansen and Karlgren, 2005; Jansen and Spink, 2003; Kelly and Cool, 2002).

The limitations of previous research call for further exploration of users' evaluation of both lists and documents. In this study, the authors compare the evaluation of documents and lists in terms of criteria and elements, along with their associated pre/post and evaluation activities, and the time spent during evaluation.

Research problem and research questions

Previous research on evaluation mainly focuses on document evaluation, and less on list evaluation. Few studies compared the similarities and differences between document and list evaluation. Moreover, relevance criteria is the main research topic, but other dimensions of evaluation, such as the elements or components that users examine for evaluation and their evaluation activities with associated pre and post activities, and the amount of time users spend evaluating, are less investigated. Here are the four research questions that are addressed by this study:

(1) In what ways do criteria of evaluating lists and documents match or differ?

(2) In what ways do elements of evaluating lists and documents match or differ?

(3) What types of associations do exist between evaluating criteria and elements?

(4) In what ways do evaluation activities match or differ between list and document evaluation?

(5) Does evaluation time differ between list and document evaluation? What are the variations and associated reasons, in time spent within list and document evaluation?

Literature review

This paper reviews relevant literature on evaluation criteria applied, elements examined, activities performed, and time spent for both list and document evaluation. For each dimension, the literature review starts from the discussion of research on both lists and documents, then transitions to individual discussions of list evaluation and document evaluation.

Criteria

With few exceptions, evaluation criteria research focuses on either list evaluation or document evaluation. Despite the difference, both types of research highlight the role of relevance and credibility in users' evaluation. Document evaluation received more attention, as its research evolved from printed to the electronic documents; however recent research further develops the multidimensional understanding of list evaluation.

Criteria – lists and documents

Few studies explore both list and document evaluation criteria. Vakkari and Hakala (2000) evaluated the changing nature of relevance and relevance criteria during the information search process based on Kuhlthau's model. Through three rounds, the eleven participants evaluated both lists and document relevance for a real world task. The study identified six major categories of relevance criteria: information content of the documents, sources of documents, the document as physical entity, user's situation; user's experience and preference, and information types. Additionally, the analysis found twenty-five sub-categories of relevance criteria. Overall, they discovered relevance criteria varied widely, although the criteria used for list evaluation remained more consistent than those of document evaluation. While Vakkari and Hakala's (2000) study focuses on stage-oriented research in academic setting, Savolainen and Kari (2006) investigated how users select hyperlinks within lists and choose web documents during web-based search from a relevance standpoint. The study found twelve criteria used in relevance judgments regarding both links and web pages. These include: ability to understand, accessibility, affectiveness, clarity, cost, curiosity, currency, familiarity, language, novelty, reliability, security, specificity, time constraints, topicality, usability, validity, and variety. Additionally, the study compared the evaluation criteria used between hyperlinks and web pages, finding specificity and topicality are the highly used criteria for both.

Previous research also highlights the dynamic nature of list and document evaluation within the relevance concept. Spink *et al.* (1998), for example, move beyond exploration of binary relevance toward an understanding of the degrees of relevancy suggesting these may shift over time. Borlund (2003) outlines a "relevance framework," based on the many different conceptions of relevance in previous research including, "classes, types, degrees, criteria, and levels of relevance" (p. 923). Furthermore, Bade (2007) argued against misinterpretations of relevance evaluation arguing for a multidimensional understanding over a binary one. Viewing relevance as either objective or subjective leads to two major flaws: "The first is that any relevance judgment by any human being will always be subjective to some degree, no matter

how objective that person may strive to be. The second is that nothing is relevant to anything for any machine" (p. 840).

As a pioneer of relevance research, Saracevic (2007a, b) updated his 1975 review of relevance research including 30 years of new research on the concept. The review summarizes the different attributes or criteria that users concentrate on while making relevance inferences. Furthermore, he highlighted both list and document evaluation relevance concepts. Through his review, the concepts of relevance criteria remain diverse and disputed with studies indicating a wide range of divergent criteria.

Criteria – lists

Unlike document evaluation, list evaluation criteria research began within the past twenty years. It focuses on relevance judgment and quality aspect of relevance. In one of the earliest list studies, Purgailis Parker and Johnson (1990) found users' judgment of document relevance is not altered based on the organization of surrogates within a list if less than 15 documents are presented in a result list. The study used a three point scale of relevance (relevant, relevance is not known, and document is clearly not relevant), and found it indicated some evidence of bias when over 15 documents displayed, however not enough information existed for a conclusion. In a different study, Park (1994) concluded, "From a user's perspective, relevance is intimately related to thought processes and criteria used in the evaluation of citations retrieved in an IR situation" (p. 136).

Quality aspect of relevance criteria is another important issue for list evaluation. Su (1994) found "quality of references is more important than the quantity." Quality here refers to, "the relevancy of the articles (i.e. the right pieces, the appropriateness of the articles, the directly relevant); and the completeness (all that is there or no missing information)" (p. 211). Voiskunskii (1997), on the other hand, found a "lack of a criterion for comparison of search results which could actually be used in practice" (p. 133). Rieh (2002) explored users' reliance on information quality and cognitive authority as judgment measures for selection and evaluation of information. The study revealed that users' selection of a particular website or document (from a list) is a practice of predictive judgment followed by the evaluative judgment of the document itself. The study found five dimensions of information quality including goodness, accuracy, currency, usefulness, and importance. An additional six areas fell under the cognitive authority concept including trustworthiness, reliability, scholarliness, credibility, officialness, and authoritativeness. The results indicate a more complex and dynamic understanding of evaluative/predictive behaviors beyond content aspect of relevance.

Criteria – documents

As noted above, the concept of relevance engulfs the document evaluation criteria literature. Several studies reflect early calls to unpack the concept of relevance for a better understanding, such as Janes (1994) stated:

> Perhaps what we have called "topicality," "utility," "satisfaction," "pertinence," and a variety of other names are in fact dimensions of a larger, multidimensional, dynamic concept, as discussed by Schamber *et al.* (1990). They called this concept "relevance," but that word too carries baggage in people's minds and may be causing problems of its own. At present, we conceive of this concept in the abstract, encompassing much that is described by previous definitions and instantiations, but more as well [. . .] it may be possible to begin to pick apart the black box of "relevance" and see what wonders reside within (pp. 167-168).

Barry and Schamber (1998) represent those focused attempting to unpack relevance. Through a comparison of previous personal studies, they identified over 20 criteria, such as access, affectiveness, users' beliefs and preferences, geographic proximity, clarity, dynamism, and presentation quality.

Throughout the past 20 years, researchers have investigated relevance criteria from different aspects (Saracevic, 2007a,b). Barry (1994) found three groupings of criteria: tangible characteristics of documents (e.g. the information content of the document, the provision of references to other sources of information), subjective qualities (e.g. agreement with the information provided by the document) and situational factors (e.g. the time constraints under which the user was working). Her later study (1998) concluded document representations might differ in their effectiveness as indicators of potential relevance because different types of document representations vary in their ability to present clues for specific traits and/or qualities. Wang and Soergel (1998, 1999) investigated the use and evaluation of documents during research projects, and found several key criteria. They divided the evaluation criteria into two categories, "the document content information criteria concerning topic, content, depth, style, etc, and the situational criteria relating a document to the user's professional and personal situation" (Wang and Soergel, 1998, p. 122). The highlighted criteria included: topicality, orientation/level, discipline, novelty, quality, recency, reading time, availability, special requisite, authority, and relation/origin.

Greisdorf (2003) identified relevance as a problem-solving and decision-making process, in which topicality, pertinence, and utility of a retrieved item are considered while Xu and Chen (2006) "suggest that topicality and novelty are the two major underlying dimensions of relevance" (p. 969). Bales and Wang's (2005) meta-ethnography of 16 relevance studies found a wide range of definitions for various elements and criteria. Regarding criteria, the study found 14 major criteria including: topicality, recency, accessibility, quality, novelty, affordability, intelligibility, serendipity, and visibility.

Credibility or authority also garnishes significant research attention recently. Through a large study of real users and settings, Fogg *et al.* (2003) concluded website presentation most significantly impacted credibility. Metzger *et al.*'s (2003) compared college students' and nonstudents' evaluation of web-based resources, concluding nonstudents view internet resources in a vastly different way. Although both groups view variance of credibility of internet-based sources, "the nonstudents indicated that they verified online information more than the students did, although both groups reported that they verify online information only rarely to occasionally" (p. 285). Students may use/find the internet sources for convenience, and therefore do not think of the quality of their work. While disconcerting on face value, the authors suggest the real numbers may be more shocking. They state "Student participants reported that they verify online information only 'rarely' to 'occasionally,' and it is quite possible that these results are somewhat inflated due to the social desirability inherent in this measure" (p. 287).

Overall, current research indicates users rely on relevance judgments during the evaluation of both lists and documents; however recent studies discussed more dimensions of criteria for documents than lists. Previous literature relating to list evaluation remains thin and rarely compares users' evaluation criteria between documents and lists.

Elements

Similar to criteria, element research focuses on either list evaluation or document evaluation. While list evaluation studies explore element gaze patterns of the list, document evaluation research identifies more specific elements examined for different types of tasks. Also similar to the criteria literature, fewer studies explore list evaluation than document evaluation. There is little research comparing the elements examined in both list and document evaluation.

Elements – lists

List research on elements is related to gaze-based research, or more specifically eye tracking analysis. A different type of research explored the effects of rank within result lists and interface design on evaluation through eye tracking analysis. Granka *et al.* (2004) stated, "After the second link, fixation time drops off sharply [. . .] Once the user has started scrolling, rank becomes less of an influence for attention. A sharp drop occurs after link 10, as ten results are displayed on a page" (p. 2). Additionally, the study found users scanned lists from top to bottom, and the users who chose the lower ranked documents viewed more abstracts proportionately. Rele and Duchowski (2005) found no significant difference for how users view a tabular or vertical listing of search results, and concluded users typically follow a top to bottom scanning strategy. Aula *et al.*'s (2005) eye tracking analysis suggest experienced users follow an economic evaluation style, requiring fewer elements than novice users who are exhaustive evaluators (pp. 1,060-1,061).

Elements – documents

As noted above, a majority of evaluation element research focuses on document evaluation rather than list evaluation. Wang and Soergel (1998) identified 18 document information elements (DIE). In order of frequency of use, the DIEs were: title, abstract, journal, author, geographic location, publication date, document type, author's affiliation, descriptors, language, publisher, document length, volume and issue, subfile, edition, author's expertise, table contents, and citation status. Similarly, Maglaughlin and Sonnenwald (2002) found six categories of relevance judgment elements:

(1) abstract;
(2) author;
(3) content;
(4) full text;
(5) journal/publisher; and
(6) personal.

Users combined the evaluation of individual elements to create a degree of relevance (i.e. relevant, partially relevant, not relevant). Bales and Wang (2005) further discovered 28 document elements including:

- author;
- title;
- abstract;

- subject index;
- language;
- length; and
- source for document evaluation.

Focusing on elements examination in different tasks, Tombros *et al.* (2005) explored the elements online users assessed during information-seeking tasks. The study assigned each participant a series of three tasks. The first task required general gathering of background information, the second focused on making an informed decision, and the final task required participants to gather a list of information. The study found the elements and their use varied based on the type of task and users' progression through a course of tasks. Specifically, during the decision task, users focused more on "factual features of documents (numbers, pictures) and of scope and depth of the available information" (p. 339). In contrast, participants concentrated on query terms, pictures, and links more for the listing task. Another study compared elements (i.e. tables, forms, files, links, FAQs, etc) within relevant documents for different tasks and fact questions (Kelly *et al.*, 2002). The results indicate variations between task (those relating to a process) and fact questions (those relating to short, factual information). Users rely on various lists in documents, such as definition lists, during evaluation for task question. In contrast, users examine tables more often in the evaluation for fact question.

Existing list evaluation element research offers intriguing insight through eye tracking analysis, but its variation also indicates a need for further research. Current research indicates document evaluation elements vary by task, and additional research is needed for further codification of the elements and their association with evaluation criteria.

Evaluation activities and time

Although limited research focuses on evaluation activities, there are some of the related studies on different approaches applied to investigate evaluation activities. Chi *et al.* (2001), for example, translated use behavior based on the information foraging theory to predict user evaluation activities. The modeling helps design better interfaces assisting user evaluation in relation to predicting their movements and potential pages of interest to them. Joachims *et al.* (2005) analyzed clickthrough data to track evaluation activities and concluded it was not a reliable metric. They concluded that users' clicking decisions are influenced by the relevance of the results. However, they are also affected by the trust they have in the retrieval mechanism and by the quality of the result set. Another study proposes a four-phase framework for searching textual databases based on users' relevance feedback during their evaluation of lists (Shneiderman *et al.*, 1997).

While there are few studies on comparison of time spent on list and document evaluation, previous literature on time research have different foci for list and document evaluation. Instead of investigating the amount of time users spent, previous literature of list evaluation focuses on number of result pages viewed. According to one study, while 28.6 percent of users examined only one page of results, an additional 19 percent looked at two pages only (Spink *et al.*, 2001). Jansen and Spink (2006) found the percentage of users who view only one page of results has dramatically risen from 29 percent in 1997 to 73 percent in 2002. Log analysis also reveals disparity. Jansen *et al.*

(1998), for example, found users viewed 2.21 screens per query, on average. Silverstein *et al.* (1999) found users look at an average of 1.39 result list screens per query, with only 4.3 percent of users going beyond the third screen. After testing a dynamic visualization method, HotMap, Hoeber and Yang (2009) observed the time required to find ten relevant documents for one specific task. The results show that it took just over 3.5 minutes using Google, and reduced by almost a full minute using the HotMap visualization.

Time research on document evaluation, though limited, has investigated the total time spent evaluating documents. Jansen and Spink (2003) found an average evaluation session time of 15 minutes. The average time spent examining one document was 16 minutes and 2 seconds, adjusted for the outlying data. More than 75 percent of the users view the retrieved documents for less than 15 minutes. The authors also noted nearly half of the users (40 percent) viewed documents for less than 3 minutes; therefore users typically viewed retrieved documents between 3 and 15 minutes. Time for document evaluation is affected by different factors. Research shows that domain knowledge affects document evaluation time. Kelly and Cool (2002) found document evaluation time significantly decreases when the user is familiar with the topic. Their study of 36 participants found document evaluation time ranged from 23.46 seconds (least familiar) to 16.57 seconds (most familiar) within the participant's total evaluation limitation of 20 minutes. Language is another factor that influences the time for document evaluation. Hansen and Karlgren (2005) explored the impact of language on document evaluation time, specifically with native Swedish speakers evaluating English documents. They found participants took slightly longer time evaluating the non-native language (27 seconds per document) than the native language (20 seconds per document).

Current research on evaluation activities constructs and tests differing metrics. As to evaluation time, research on list evaluation primarily reports varying findings on pages of results per query rather than specific temporal measurements. Document evaluation literature, on the other hand, indicates a wide range of average time from 16.57 seconds to over 15 minutes, and their associated reasons.

Limitations of the existing literature

Existing research on list and document evaluation contains several limitations. Until recently, most of the document evaluation literature focused on relevance criteria, with few exploring the elements examined. List evaluation research also remains confined to criteria, and contains significantly fewer studies than document evaluation. Similarly, few explorations of evaluation activities exist. Alternatively, current studies of evaluation time focus more on documents than lists with the majority of list-based research focusing on average pages evaluated per query than time. While each area discussed contains individual limitations, the overall limitation remains the lack of comparative studies between list and document evaluation.

Methodology

Sampling

In order to recruit participants to represent general public, researchers of this study posted newspaper advertisement (Journal Sentinel) as well as fliers to different community centers and public locations (e.g. local libraries, grocery stores, etc),

listserves (e.g. craigslist, etc). A total of 31 participants participated in the study from the Greater Milwaukee area representing general users of information with different sex, race, ethnic backgrounds; education and literacy levels; computer skills; occupations; and other demographic characteristics. Each participant was paid $75 as an incentive for his/her time and effort spent for the study. Table I presents the demographic characteristics of participants.

Data collection

Data of this study were collected via two pre-questionnaires (general and onsite), think aloud protocols and logs, as well as post-questionnaires. Here are the data collection procedures:

(1) First, participants were asked to fill in general pre-questionnaires in which their demographic information and their experience in searching for information related to their work and search tasks, information resources, search activities and problems were requested before they came to Intelligence & Architecture (IIA) research lab.

(2) Second, participants were invited to come to Information Intelligence & Architecture (IIA) research lab to search for information for one work related and another personal-related task formation. They were instructed to fill

Demographic characteristics	Number	Percentage
Gender		
Male	10	32.3
Female	21	67.7
Age		
18-20	1	3.2
21-30	13	41.9
31-40	5	16.1
41-50	7	22.6
51-60	5	16.1
61 +	0	0.0
Native language		
English	29	93.5
Non-English	2	6.5
Ethnicity		
Caucasian	29	93.5
Non-Caucasian	2	6.5
Computer skills		
Expert	3	9.7
Advanced	21	67.7
Intermediate	7	22.6
Beginner	0	0.0
Occupation		
Administrative assistant, marketing communications, librarian, student, programmer, tutor, school guidance secretary, social worker, portfolio specialist, trust associate, client relationship associate, software developer, self-employed, buyer, nurse, academic advisor, painter, unemployed, etc		

Table I.
Demographic characteristics of participants ($n = 31$)

in onsite pre-questionnaires consisting of information in relation to their work and search tasks, their expectation of the search results as well as their domain and retrieval knowledge about the search tasks before they started their searches.

(3) Third, participants were asked to search for two self-generated tasks. They were instructed to "think aloud" during their search process. Their information search processes including evaluation activities were captured by Morae, a usability testing software that not only records users' movements but also captures their "think aloud," including their thoughts during the search process.

(4) Fourth, participants were asked to fill in the post-questionnaire regarding their experience in their search activities including evaluation activities, their problems, and factors affecting their search activities.

As stated above, each participant was asked to conduct two self-generated tasks instead of assigned tasks. Examples of work tasks included researching for a client, enhancing work related education, and professional writing; examples of personal tasks consist of hobby or travel information, answering household maintenance questions, planning events (such as a wedding), pre-purchasing research, or simply curiosity. However, two out of 62 tasks were not able to be analyzed because of poor quality of the recorded data; thus the total of tasks being analyzed in this study was 60. The recoded data were transcribed. Every movement of evaluation activity for each participant was transcribed, and his or her related verbal protocols were also recorded in the transcription. Table II highlights how data sources were collected and analyzed for each research question.

Data analysis

The unit of analysis is each evaluation activity. As defined in the Introduction, evaluation refers assessment of the usefulness or relevance of both search result lists and documents retrieved or browsed. It starts from a participant's assessment of a query result list, continues to the assessment of individual documents, and ends when the user moves to the next search activity. Types of evaluation criteria, elements, evaluation activities and their associate pre/post activities were analyzed based on open coding which is the process of breaking down, examining, comparing, conceptualizing and categorizing, (Strauss and Corbin, 1990). The open coding process identified the most categories possible since categories emerged from the data rather than applying existing categories. This process, therefore, adhered to a purely inductive analysis for category creation for types of evaluation criteria, elements, and activities rather than a quantitative-based descriptive analysis. This is because the data were mainly derived from think aloud and log data. Think aloud data offer rich and insightful information about what participants think about at that time, but they might not always specify every criterion, element, and activity to show an accurate count of each variable. Frequency of data within specific categories of criteria, elements, and activities offered quantifiable ranking of the associated data and are indicated within the organization of the tables presented and discussed in result presentation. Descriptive analyses were applied to analyze time spent on lists and individual documents. Table III presents the coding scheme of evaluation criteria, elements, evaluation activities and their associated pre/post activities, and time. To

Research question and characteristics of participants	Data collection	Data analysis
Similarities and differences of evaluation criteria applied between list and document evaluation	Think aloud data related to evaluation criteria Post-questionnaire data related to problems in applying evaluation criteria	Open coding, taxonomy of types of criteria
Similarities and differences of evaluation elements examined between list and document evaluation	Log data related to evaluation elements examined Think aloud data related to evaluation elements examined Post-questionnaire data related to problems in examining elements	Open coding, taxonomy of types of elements
Associations between evaluation criteria and elements	Log data related to evaluation elements examined Think aloud data related to evaluation criteria and elements examined	Open coding, taxonomy of types of criteria and associated elements
Similarities and differences of evaluation activities performed during list and document evaluation	Log data related to evaluation activities performed Think aloud data related to evaluation activities performed Post-questionnaire data related to problems in performing evaluation activities	Open coding, taxonomy of types of activities
Differences of evaluation time spent between list and document evaluation and associated reasons	Log data related to evaluation time spent in list and document evaluation Think aloud data related to time spent in list and document evaluation and associated reason Onsite pre-questionnaire data related to search topic and plans Post-questionnaire data related to problems in spending time in list and document evaluation	Descriptive analysis of time spent in list and document evaluation Taxonomies of types of reasons that influence quick and long evaluation a list or a document
Characteristics of participants	General pre-questionnaire	Descriptive analysis of characteristics of participants

Table II.
Data collection and analysis

save space, the detailed discussion of different types of criteria, elements, evaluation activities and their associated pre/post activities, and time with definitions and examples are presented in the Results section.

To test the inter-coder reliability, two researchers independently coded 20 tasks from 10 participants randomly selected from 60 tasks performed by 31 participants. The inter-coder reliability for evaluation of lists and documents in terms of their criteria, elements, and evaluation/pre/post-activities was 0.94/0.90, 0.90/0.95, and 0.93/0.95 respectively, according to Holsti's (1969) reliability formula.

Variables	Definitions	Example (lists)	Example (documents)
Criteria	Participants' judgments applied during their evaluation of a search result list or an individual document	Reputation, "When I scanned it, I saw monster.com and that's a pretty reputable and popular job site, so I will go there and see what they have (S11)."	Scope, "This site was not useful at all it did not provide much info about it (S5)"
Elements	The individual components that participants examined during the evaluation of a result list or an individual document	Rank, "I am going to that one. That's the first one [listed] (S18)"	Text, "Just reading what kind of a plant it is (S18)"
Evaluation/ pre/post activity	Participants' actions taken during, prior to, and following the evaluation of a result list or an individual document	Pre-activity [types www. babyname.com into URL] (S1) [clicks 4th result from Google] (S6) Post-activity "I am going to advanced search (S30)," or [clicks back to Google] (S11)	Evaluation activity "So now [I will] go to each one and compare which gives you the best deals (S7)" "I think I want to look at that [. . .] I am going to Control F it and look for my word, arsenic [using find function] (S19)"
Time	Time that participants spent in evaluating a result list or an individual document from beginning to the end in minutes	1.3 minutes evaluating a Google result list on search query, "nuclear weapons, missiles (S1)"	2.2 minutes evaluating brides.com (S15)

Table III.
Coding scheme

Reliability = 2M/(N1 + N2), where M is the number of coding decisions on which two coders agree, and N1 and N2 refer to the total number of coding decisions by the first and second coder, respectively.

Results

A thorough examination of both similarities and differences of lists and documents answers the proposed research questions in relation to criteria applied, elements examined, associations between criteria and elements, activities performed, and time spent in evaluating lists and documents.

Criteria similarities

Many of the evaluation criteria selected for list and document evaluation are similar. These criteria include: scope, specificity, reputation, depth, credibility, cost, and language. Table IV presents examples of similar evaluation criteria. Upon closer examination, each of these criteria offers interesting differences. Strictly speaking, the differences are not about criteria, instead, they are about the basis for the judgments to be made.

Criteria	List examples	Document examples
Scope	"Hmm I have never seen this site [. . .] from all different sites wow over 2,000 [results] [. . .] this is a good site, my-it-career (S17)"	"So this looks like it might be a good resource, this resource compares several different treatments, psychotic drugs other treatment drugs, shock therapy, vocational rehab [. . .] I like this study because it is very clear cut ABC (S7)"
Specificity	"There is a UNESCO website that I think might be reliable but other than that these sites are not really specific (S16)"	"Qualifications [. . .] 18 years old by August 1, 2008 [. . .] application deadline [. . .] this has basically answered all my questions right off the bat (S31)"
Depth	"There are a lot of dictionary meanings but I think those are too simple [since] I am looking for something more complicated (S11)"	"It went into a lot of detail, it gave personal accounts this man's experience having a dog with it and it gave a lot of detail how he dealt with it what kind of symptoms and treatments that the other sites did not offer and overall this one gave me the most info (S5)"
Reputation	"I was also going to use Wikipedia even though it is frowned upon (S20)"	"He is pretty well known as a budget traveler (S29)"
Credibility	"I am not going to utilize sites being written by somebody. I want it to be a news article or an Irish history page or something like that, that seems more credible (S13)"	"This looks like people are just writing posts it's a personal thing, not reputable (S12)"
Layout	"I am pretty drawn usually to the stuff on top here, the sponsored links, just because they [. . .] hold such a prominent space on the page [. . .] I don't usually look at the stuff on the right here [pointing to column of sponsored links] (S9)"	"I actually like the other website better because it straight out listed things (S25)"

Table IV.
Examples of similar evaluation criteria

Within the area related to content coverage, scope and specificity are the most typical criteria. Scope relates to the extent to which information is covered. Lists provide limited data to explore scope, therefore allowing only quick decisions based on brief summaries of items or their own knowledge of the resources. Unlike lists, documents provide ample information for participants to use scope as a criterion. In the example provided in Table IV, participant 7 selected the document because it compares several different treatments, and these treatments are the ones that she was interested.

Same applies to specificity and depth criteria. Specificity refers to the extent to which information covered by the document is focused to match the user needs. Depth refers to the extent to which detailed information is provided by the document. As noted above, participants relied on the short descriptions within lists to make their judgments. Therefore, lists forced quick decisions based more on past experience with similar sites and instinct rather than information. In the examples provided in Table IV, participants made decisions based on their perceptions of Google and UNESCO. By

contrast, documents provided participants with specific and detailed information to help them make their decisions.

In the criteria apply related to coverage, reputation and credibility are the dominant ones. Reputation refers to the extent to which the source of a document or information is well known or reputable. Credibility refers to the extent to which information provided is reliable. For both reputation and credibility, participants based their decisions on the owner(s) of the website (such as Amazon.com) or the author(s) of the document. Although similar, credibility criteria judgments for lists remain snap decisions, while document evaluation occasionally requires a longer analysis of credibility judgments. For example, participant 16 spent two minutes evaluating a Google result list based on credibility, while the same participant required two times as much time for the same criteria evaluation of a document.

In addition to the evaluation criteria related to document surrogates in lists, two criteria are associated with structure of lists and list pages. While layout specifies the similarities in criteria between lists and documents, organization (discussed in Criteria differences below) signifies the unique criterion for list evaluation only. Layout refers to the display of documents or information. Interestingly, participants cared about the layout of lists and specific documents with different preferences. In general, participants do not like that sponsored links hold prominent space on the list page. Different from their preferences of layout of lists, participants preferred documents with clear display of information.

Overall cost and language do not yield differences for the evaluation of lists and documents. Participants examined the costs associated with accessing an individual document through a website. Language provided significant frustrations, as users found potentially relevant items, however they could not evaluate them based on the language barrier. Participant 16, for example, stated, "It's all in French that's not really helpful [. . .] this website has a list of actual books, research guides that could help me in my search but it is all in French and I don't speak French."

Criteria differences

Despite their similarities, the differences between list and document evaluation are represented by unique criteria applied for the evaluation of lists and documents. Participants applied organization in evaluating lists but not documents. Organization refers to the arrangement of query results in the list. Participant 12, for example, stated, "I am scanning the first two hits here [. . .] [Assuming the top hits are more relevant]."

During the evaluation of documents, participants applied twelve unique criteria beyond the ones shared with list evaluation including: unique information, currency, accuracy, intended use, picture, number, ease-of-use, availability, speed, item type, and item length. Lack of information provided for the lists contributed to not applying these criteria in the evaluation process. Participants in this study complained the lack of data needed for their list evaluation judgments, and they had to click the documents in order to make assessments. Related problems reported by participants, such as the unavailability of documents, extended loading time, and language barrier issues.

Three criteria emerge relating to quality (unique information, currency and accuracy) of documents. Although similar, these three criteria address slight differences of quality (see Table V). Unique information refers to the extent to which distinctive viewpoints or ideas are provided. This criterion offers an important aspect

Criteria	Examples
Unique information	"They kind of reiterate what I found in CNN. . .one of the big differences from what the White House says and what the industry says are quite different and this gives actual percentages (S2)" "This is new info so I think this is a good website and this is a bib[liography] at the end (S16)"
Currency	"Also the info comes out probably between three and five days earlier than it does actually having to wait for the print journal scrolling looking at some of the topics [. . .] it almost reads more and more like a newspaper what's current hot news (S2)" "I don't want that [. . .] this is not a current project (S19)"
Accuracy	"This does not seem right, 23 percent lower wow [. . .] this is not right [. . .] I don't know [. . .] oh my I can't believe it (S17)" "$63,000 [. . .] that's a little unrealistic [and] I am going back to Google (S21)"

Table V.
Examples of quality criteria for document evaluation

of quality, since participants applied it as a quality measure. Additionally, unique information was applied more often than other quality-based criteria. Currency refers to the extent which information is timely, recent or up-to-date. Currency is used to both accept and reject a document. Accuracy, however, specifically examines the extent to which information is considered as correct or valid. Unlike currency, usually participants only mention this criterion as a rejection of specific documents. Both currency and accuracy could be presented at the list level to save users time for the initial screening although most of the lists do not contain the information. Participant 2, for example, tried to explain the number of documents on a list based on the topic currency, stating, "This has only been brought out Thursday this week, not even two days ago, obviously there are going to be a lot of attention on this." In this instance, the ability to view document dates could assist list evaluation.

Another unique criterion is related to intended audience of documents, intended use refers to the targeted audience of the document. Participants needed to decide if the document is intended for their use. Here is an example that a participant quickly recognized the intended use of a document, "This site is specific to classroom management for professors (S28)." An additional criterion stems from within the division of design including ease-of-use. Participants appreciated documents (mostly website), which were easy to navigate and use. Participant 7 stated, "It looks like a good source, but then when you look at it [. . .] I don't know how to use this website [and] I am not going to bother with it."

The two accessibility related criteria (cost and language) that apply to both list and document evaluation are discussed above. Another group of criteria in accessibility category unique to document evaluation includes: availability and speed. Availability refers to the extent to which effort is required to obtain information. Unavailability of documents often obstructs users' access to documents (either in part or completely). Participant 14 voiced concerns stating, "I am unable to search the library catalog because I don't have a barcode or pin number from the University of Chicago." Another participant stated, "Oh man [have to login], I thought this was going to be a direct link [closes page] (S24)." Similar frustrations occur in relation to speed when participants

had to wait for a long time to access documents (although occurring in limited numbers), as participant 6 comments, "Here's a site from Germany, probably has a server running in his basement [give up]."

The final group of criteria unique to document evaluation focuses on characteristics of the items including:

- item type;
- item length;
- picture; and
- number (see Table VI).

Item type and item length are closely related, with the former criteria specific to format and the latter one to the length of a document. Item type, however, is applied far more frequently than item length. Participants preferred different types of documents for different types of tasks. Some participants preferred documents containing pictures or numbers, and also used them as elements for evaluation.

Element similarities

The elements participants examined within lists and documents also show similarities. Table VII presents examples of similar evaluation elements. While author/source, time/date, and item type are used for both list and document evaluation, few differences among their use exist. Participants looked at the author/source in both lists and documents for authority judgments. The time/date element remains a consistent evaluation tool for the currency criterion in both list and document evaluation. Unlike the item type criterion, the element version relates to how participants used different item types as tools for evaluation.

Participants looked at both the title/subtitle and abstract/snippet for evaluation, but these elements played different roles in the evaluation process. As to the title/subtitle, participants primarily used this element to assist them selecting which specific documents to evaluate from lists. Similarly, some titles may also indicate the volume of the information as specified in the example about 109 job interview questions. In

Criteria	Examples
Item type	"I am not going to watch a video right now (S23)" "I am trying to figure out what this is about because we have not done any [. . .] always looking for ways to mix up lesson and make it less boring, if I had a video it might be helpful [teaching the class] (S12)"
Item length	"The second site was useful but not as detailed as the first [. . .] this 6th site I feel was the best out of all of them [as it] gave a lot of detail [observer's note: Participant 5 equates the word detail with length] (S5)" "I got a one paragraph explanation of it (S11)"
Picture	"There are pictures of the hotels so I can get an idea of what they look like [. . .] the Ramada looks good (S28)"
Number	"This one catches my eye because it looks like it has some statistics that I am interested (S12)"

Table VI.
Examples of item characteristic criteria for document evaluation

Elements	List examples	Document examples
Title/subtitle	"The first thing I do is I review the titles to see [. . .] This one does not have an abstract but the title is too good to pass up so I will keep it (S4)" "The first one [result] says 109 typical job interview questions that seems too big for me (S12)"	"It's a little bit harder if the title does not jump out at you from the titles other forum topics then you are hoping you will come across something that will help you (S26)"
Abstract/snippet	"Oh now that is interesting I am going to have to read the abstract that's an interesting title [. . .] and if I look at the abstract and it does look good or interesting, I am going to click on the item; If it does not look interesting I am not going to select the item (S4)" "I picked it [website] based on the short description it's called the mathmuseum.org website (S26)"	"They have an abstract here I can get it [. . .] (S6)"
Author/source	"Some of these [results] are from Stanford and other medical schools (S3)"	"It's [written by] the International Osteoporosis Foundation. That's probably one of the better ones (S25)"
Time/date	"They do have an article from yesterday [clicks an article link] (S2)."	"This is an article from a magazine called RFID update. . .this article is 2 years old [clicks back to Google] (S6)."
Item type	"Right away at the top is a map and I am not going to bother with the map right now (S23)"	"I will watch this video for just a second [. . .] it is pretty much a report of people who have already been there (S12)"

Table VII.
Examples of similar evaluation elements

general, title was less examined during document evaluation. In a few cases of title/subtitles being used to evaluate documents, it is often through selecting a part of a document to analyze based on a subheading or subtitle.

Likewise, participants relied on abstracts or snippets within lists more than those found within documents. Similar to the use of titles, abstracts or snippets are less used for document evaluation since most participants read an abstract or snippet prior to selecting the document for evaluation in the first place and they now had the full-text documents to find relevant or useful information. In some instances, however, the abstracts may not be available within the lists or the documents could provide a more detailed abstract than previously viewed. In these cases, the abstracts retain their powerful role within the document evaluation process.

Element differences

Similar to criteria, the study found significant differences between the elements participants evaluated within lists and documents. Unlike criteria, however, both lists and documents in current IR systems provide additional elements. Specifically, when evaluating lists, participants examined unique elements, such as the URL, the number of results, rank, and keywords. For document evaluation, however, participants

examined other elements, such as the full text, picture(s), link(s), number(s), and specific features that are only offered in documents. The differences in elements for both list and document evaluation are discussed separately below.

Participants encountered unique elements in the evaluation of lists, mainly relating to the information provided as well as the organization of the lists, most often the *URL*. Although similar to a document title, a URL represents the domain of the document rather than its specific title. Participants applied criteria, such as credibility, to the URL for evaluation. Another element is the number of results. Participants evaluated this element for indication of the success of their searching; a type of Goldilocks' effect, in which participants often went from too few results to too many, settling in the end on the right number. Perhaps more important for most participants is the element of rank, referring to the order of a list in a query result page. Participants often chose items higher ranked initially, before moving to other elements. Finally, participants occasionally scanned lists for keywords, often as an evaluative method for reducing the number of results (Table VIII).

In evaluating documents, participants examined full-text of documents, in particular pictures, embedded links, numbers, and specific features of the documents. According to the participants, these elements could provide quick and key information for them to make judgments of documents rather than going through the whole document. Table IX presents these elements with specific examples.

Perhaps the most complex elements surround specific features. These elements refer to the features of a given document display used for evaluation. Although these elements were evaluated less often than others, they offer distinctive information for the participants for their decision making. Table X presents the types of features participants mentioned with examples.

Criteria/element relationship

Participants examined specific elements in order to apply their own criteria to evaluate lists and documents. Table XI illustrates the relationships between criteria and the elements examined. Additionally, the table identifies which type of evaluation (list,

Elements	Examples
URL	"A lot of times when I am looking at Google I look at the URL for the link to go into [. . .] there is one that says tripadvisor.com and chances are they are trying to sell you a vacation package and I don't want that I just want to look for myself (S30)"
	"The first one is the food network [www.foodnetwork.com]which I love the food network website so I usually go to that (S21)"
No. of results	"There are too many results so let me go back and change my query (S30)"
	"Ok, so my initial search came up with 98 hits put 25 in the maybe category now we are down to 21 good results that I will print out and send to the boss see what she says if they are really good results to get articles for (S4)"
Rank	"I am scanning my first two hits here and see which one might be [the] most credible (S12)"
Keywords	"If it [the result] has silicosis I am going to keep it (S4)"

Table VIII.
Examples of elements for list evaluation

Elements	Examples
Full text	"Again I see some good info to write down what the difference type of missile that was used in Kuwait and it also says it deals with the patriot system upgrades (S1)" "I am a big Civil War buff so this gets my interest right away [. . .] I am not getting a lot of info but I am just basically looking for tidbits so this gives me enough [. . .] less info on this page than I was hoping (S23)"
Picture	"This is city data [. . .] these are photos of Columbus Georgia (S23)" "I skipped over the first result because it looked like it had younger kids on the front cover but this looks geared towards older students (S28)"
Link	"At the bottom another thing I like is that they have a lot of links to other articles in the database so for instance interview questions and answers, illegal interview questions [links], really good (S12)"
Number	"I see some statistics I am quickly going to scan and see (S12)"
Specific features	See Table X

Table IX.
Examples of elements for document evaluation

Feature types	Document examples
Search function	"They have a search function on their website which I am going to use (S21)"
References	"I am scanning [same webpage] because I know this website usually has some good references in their subheadings they have related articles that some would be relevant to this topic too [to this article] so I am going to choose one of them (S2)"
Indexes and tabs	"I have used that site before and it has a good variety of indexes and tabs to follow that I am sure something this huge would definitely be front and center (S2)"
Zoom	"[clicks 2nd result: Lake Michigan territory 1778 and zooms in and out of map by clicking the zoom function] (S4)"
Customer review	"It also gives a link with costumer reviews so you get more insight on how other people feel about each unit and that might give a better idea of how good the product is actually based on other people's experiences (S5)"
Comparison	"In Tripadvisor if you put in your dates [. . .] they give you different discounts for comparison [. . .] so now go to each one and you can compare which gives you the best deals (S7)"
Classification/category	"This shows a nice category listing not only by frequency but by type [. . .] looks like a little information hierarchy [. . .] nice classifications (S6)" "Ok this looks a good website [. . .] I am going to select a category [browsing subcategories] (S17)"
Help/tips/FAQ	"I am also looking to see if they have a help menu and Frequently Asked Questions by category so if I give this website to patrons I will be able to guide them thru the experience (S8)"

Table X.
Examples of specific features used as evaluation elements

Criteria	List and/or document	Element(s) examined
Credibility	L & D	Abstract/Snippet, Author/Source, Full text, Keywords, Number(s), Title/Subtitle, & URL
Depth	L & D	Abstract/Snippet, Full text, Link(s), Number(s), Picture(s), Specific features, & Title/Subtitle
Language	L & D	Abstract/Snippet, Full text, Keywords, Link(s), Specific features, & Title/Subtitle
Coverage	D	Abstract/Snippet, Full text, Number(s), Specific features, & Title/Subtitle
Reputation	L & D	Abstract/Snippet, Author/Source, Link(s), Title/Subtitle, & URL
Specificity	L & D	Abstract/Snippet, Full text, Number(s), Specific features, & Title/Subtitle
Scope	L & D	Abstract/Snippet, Full text, Specific features, & Title/Subtitle
Intended use	D	Abstract/Snippet, Full text, & Title/Subtitle
Item type	D	Full text, Link(s), & Picture(s)
Speed	D	Full text, Link(s), & Picture(s)
Unique information	D	Full text, Number(s), & Specific features
Accuracy	D	Full text, & Number(s)
Cost	L & D	Full text, & URL
Currency	D	Full text, & Time/Date
Ease-of-use	D	Full text, & Picture(s)
Item length	D	Abstract/Snippet, & Full text
Layout	L & D	Full text, & Picture(s)
Organization	L	No. of results, & Rank
Picture	D	Picture(s)
Number	D	Number(s)
Availability	D	Full text

Table XI.
Relationships between criteria and elements

document, or both) a criterion is applied. The evaluation criteria are not equally supported with elements offered by IR systems. Based on the data, the credibility, depth, and language criteria supported with the most number of elements with six or more elements supporting each. A second tier of supported criteria includes the coverage, reputation, specificity, and scope with four or five supporting elements. Finally, the least supported criteria, in descending order of supporting elements, are: intended use, item type, speed, unique information, accuracy, cost, currency, ease-of-use, item length, layout, organization, picture, number, and availability.

Pre/post and evaluation activities

Unlike criteria and elements, participants followed different paths of pre/post and evaluation activities in evaluating lists and documents. Figures 1 and 2 present the pre/post and evaluation activities of list and document evaluation. Among the three types of activities, the most similar one is the evaluation activities, as both list and document evaluation include examine information and find keywords. Comparatively speaking, evaluation activities in relation to documents are more dynamic and complicated than list. The limited activities of list evaluation reflect the limited information provided within a result list; similarly, the highly dynamic nature of document evaluation activities echoes the increased options and information offered within documents.

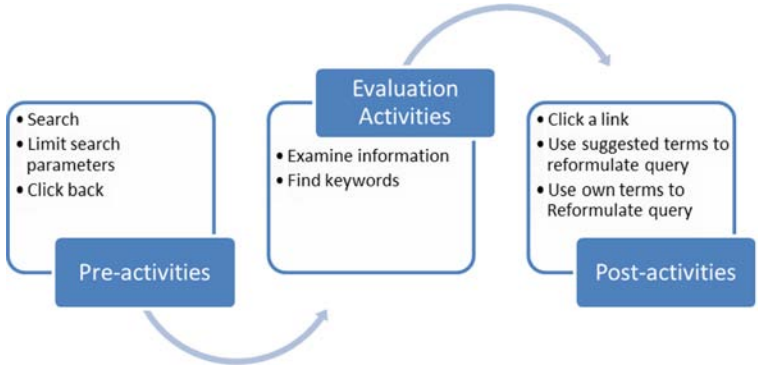

Figure 1.
Pre/post and evaluation activities for result list evaluation

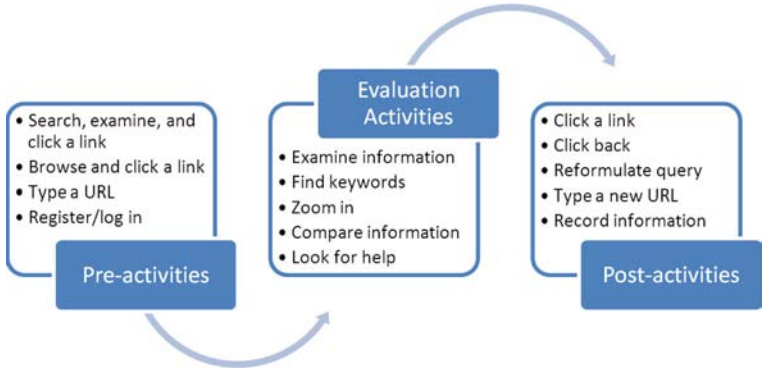

Figure 2.
Pre/post and evaluation activities for document evaluation

Table XII presents pre/post and evaluation activities in relation to list and document evaluation with examples. For pre-activities, while search results are the products of search, documents could be the products of searching, browsing and directly accessing. For evaluation activities, in addition to examining information and find keywords, participants also engaged in several activities that are required to access and examine relevant information in document evaluation. Some of them could be offered in the list, which can help participants make a decision before they access the document, for example, whether they need to register or log in, in order to access the documents. Comparing information is also a typical tactic that participants had to apply in order to evaluate and find useful information. Theoretically, comparing information could also be a tactic for post activities while users used the selected document/information for comparison although we did not find incidences in this study.

For post activities, the differences lie at the differences of outcomes. While lists offered relevant documents, participants would click a link. If they did not contain relevant documents, then participants had to reformulate queries. Post activities for documents are much more complicated because they are related to not only the access and use of a document but also going back to the list and starting another round of search. Because of the design of IR systems, participants always had to click back to the list in order to access the next relevant document. Although both list and document

Table XII.
Pre/post and evaluation activities with examples

Evaluation activities	List examples	Document examples
Pre-activity		
Search	"I am going to type in 'cook county government Illinois elections history' and I have over 500,000 search results (S14)"	
Limit search parameters	"OK, we want to limit [clicks limits option]. Searched for published in the last, I will do 180 days because that is an easy limit (S4)"	
Click back	"I am going back to my Google search [clicks back once to Google] (S16)"	
Search, examine, and click a link		Uses Google.com to search for "Bruce Schneier," and clicks 2nd result link (S6)
Browse and click a link		Browsing SpaceWar.com and clicks the link "biomedical news," to go to "Hospital and Medical News (S1)"
Type a URL		[types URL http://www.foodnetwork.com] (S9)
Register/log in		"You need to login. I think I have been here before [. . .] I think I will register [fills registration form] (S3)"
Evaluation activity		
Examine information	[looking at links] (S19)	"OK over viewing the site, it does not look like it is too involved, which is nice. Nothing jumps out at me, just looking at the directions for cooking. It looks simple enough (S9)"
Find keywords	"Basically I am judging everything on the keywords that pop up (S29)"	"I am going to use the 'find function' to search for the word 'election' on this homepage [gengateway.com] and I see politics of Cook County elections campaigns but it is not covering the year that I need (S14)"
Zoom in		clicks a Google map result: Columbus, GA; zooming in on map (S23)
Compare information		"In Tripadvisor if you put in your dates [. . .] they give you different discounts for comparison (S7)"

(*continued*)

Evaluation activities	List examples	Document examples
Look for help		"I am also looking to see if they have a help menu so if I give this website to patrons I will be able to guide them thru the experience [clicks link: help] (S8)"
Post-activity		
Click a link	"There is a Columbus Georgia homepage I am going to click on the second entry [clicks 1st result: Columbus Georgia home page] (S23)"	"There is a link to 'This document' and I click on that it brings me to US Department of State website [travel.state.gov] (S29)"
Click back		Clicks back button to the result list in Google.com (S25)
Reformulate query	"[query: Best Vendors LLC] [. . .] and when I type in just the name which is 'Best Vendors LLC' I get nothing [. . .] so I am going to type in 'Best Vendors Vending Machine Service' [. . .] and found it (S22)"	"Oh so soft baby blanket [clicks link] it's kind of pretty I kind of like the pattern on that one [. . .] so what I am going to do is put, red heart soft baby [copies this as a Google query] put that info into Google search (S8)"
Use suggested terms to reformulate query	[query in basic search: rfid security][clicks a link from suggested topics: radio frequency identification AND computer security](S6)	
Type a new URL		Types URL http://www.cnn.com (S2)
Record information		"It kind of just gives five different topics, bouquets, best blooms, favorite colors so I am going to write down a few of the different flower colors and things you can add to it [writing info on paper] (S15)"
		"I am going to bookmark this page and come back to that (S6)"
		"This looks good to me so I am going to email it to myself (S21)"
		"I think that is the first class I would be interested in [. . .] ok so I am going to print all course details (S24)"

Table XII.

evaluation include query reformulation, reformulation following document evaluation incorporates information derived from the document evaluation process.

Time

The study found participants spent 246 percent more time evaluating individual documents than individual result lists. The comprehensive information provided by documents lead to the more time spent on the evaluation of documents. While the average time participants spent on evaluating a list is 0.90 minutes with standard deviation 1.13, the average time participants spent on evaluating an individual document is 2.23 minutes with standard deviation 3.29. Although the time spent on list evaluation is similar, time spent on document evaluation is diverse (see Table XIII).

A majority of participants spent less than one minute evaluating a list (64.3 percent), with an additional 25.7 percent spending between one to two minutes. On the long extreme, participant 4 spent 13.25 minutes evaluating a list. This variation is due to the vast amounts of information provided in the list by the digital collection she searched. A list evaluation time exceeding three minutes only occurred six times. Document evaluation times, however, contained more variability with 68.2 percent of instances requiring less than two minutes for evaluation and 6 percent requiring between 7 and 25.5 minutes. Since lists typically provide limited elements, the users make quick decisions, whereas documents offer ample elements requiring additional evaluation time. For example, participant 7 was searching for information regarding the relative effectiveness of treatments for schizophrenia. During her search, she spent 1.61 minutes on average on list evaluation, while using 4.22 minutes on average for individual document evaluation. More telling, she spent 25.5 minutes on an individual document, during which she spent the majority of the time evaluating the multi-page article in-depth.

The reasons that determine quick evaluation time mainly associated with the nature of documents retrieved and reviewed. Irrelevant information, requiring extra efforts, advertisement, duplication, non-authoritative, out-of-date, and disorganized

Table XIII.
Frequency distribution of time spent evaluating lists and documents

Time (minutes)	List ($n = 269$)	Document ($n = 465$)
<1	173	180
1-2	69	138
2-3	19	52
3-4	3	30
4-5	2	12
5-6	1	14
6-7	1	11
7-8	0	3
8-9	0	3
9-10	0	8
10-11	0	3
11-12	0	1
12-13	0	0
13-14	1	0
14-15	0	1
>15	0	9

information as well as foreign language were the main reasons that direct to quick evaluation. Here is one typical example illustrates how irrelevant information led to quick evaluation. "That's not what I want I found a type of disclaimer that's not what I want; I want a definition," according to participant 11.

Compared to quick evaluation, it is more complicated in terms of the types of reasons that lead to participants engaging more time for the evaluation. In general, search topics have impact on time spent for individual documents. When participants have to read the documents more carefully to understand the information, the evaluation process is longer. Simultaneously, research topics also prolong the evaluation process, such as "peer-review literature on silicosis or exposure to silica in the last 4 months (S4)," etc. In general, personal tasks in relation to everyday life require less time for the evaluation. However, personal interests and the characteristics of the document co- influence the evaluation time. For instance, participant 31 needed information about new running shoes, and he found a document that "it is 7 pages long there is a lot of information here." He spent 14.2 minutes to assess the document. It is interesting to find that the majority of the documents that participants spent more than 10 minutes to evaluate are from authoritative and reputable sources in a specific area. For example, participant 3 spent 18 minutes on aamc.org, "I would like get to more info on MCAT in general about the MCAT, when it is taken, scores, and the best ways to practice. I am on the medical college admission test official website right now [aamc.org]."

Discussion

While previous research mainly focuses on document evaluation of relevance criteria, the major contribution of this study is its comparison of list evaluation and document evaluation in terms of its criteria applied, elements examined, pre/post and evaluation activities performed, time spent as well as the association between criteria and elements. The theoretical and empirical implications of the findings of this study can be addressed from the following aspects by:

- comparing the similarities and differences between list and document evaluation;
- revealing multidimensionality of evaluation criteria;
- associating evaluation criteria and associated evaluation elements to identify what types of support users still need in applying different types of evaluation criteria;
- integrating list and document evaluation;
- integrating pre/post and evaluation activities; and
- reducing evaluation time for different types of tasks.

First, this study fills in the gap on comparison of list and document evaluation. Participants applied more evaluation criteria and examined more associated elements in document evaluation than list evaluation because of the information provided by IR systems. By comparing list and document evaluation, the authors are able to identify not only the similarities of but also the differences of the two types of evaluation in terms of evaluation criteria applied, elements examined, pre/post and evaluation activities performed and evaluation time spent. It is interesting that even though participants did apply several similar criteria for both list and document evaluation,

there are differences in the two types of evaluation. First, participants made quick and cursive evaluation at the list level, and then go to documents to make more thorough judgments. Second, list evaluation in general focus on examining one element while document evaluation involves viewing multiple elements. Here is one example from document evaluation. Participant 7 stated, "When choosing which source to use, I first look at the credibility of the source (i.e. is it from Wikipedia or Mayo Clinic?). I also look for things that are most relevant to what I am seeking. Finally, I look for articles that I can understand easily; I would stay away from sources that use highly technical or scientific language that I don't understand."

At the same time, list and document evaluations are interrelated to each other. Some of the post-activities (i.e. click a link) of list evaluation are pre-activities of document evaluation; some of post-activities (i.e. click back) of document activities are the pre-activities of list evaluation. This study also reveals that participants spent similar time in evaluating lists but diverse time in evaluating documents. The main reasons for quick evaluation are related to the nature of the documents and accessibility, such as irrelevant information, requiring extra efforts, advertisement, duplication, non-authoritative, out-of-date and disorganized information as well as foreign language while both search topics and the characteristics of the document lead to longer document evaluation (Xie *et al.*, 2010). The problems of a lack of support for list evaluation and a lack of integration from list evaluation to document evaluation call for the need to enhance the design of interfaces discussed in detail below.

Second, this study identifies multi-dimentionality of relevance and several new dimensions that help researchers understand nature of list and document evaluation. The examination of criteria reiterated the importance of relevance judgments as previous research has emphasized (Bade, 2007; Barry, 1994 and, 1998; Borlund, 2003; Fitzgerald and Galloway, 2001; Purgailis Parker and Johnson, 1990; Park, 1994; Vakkari and Hakala, 2000, Saracevic, 2007a,b). Most importantly, this study found that multidimensionality of relevance criteria can be extended to content coverage, content quality, design criteria, accessibility, item characteristics as users apply relevance criteria to meet their personal understanding of relevance. In particular, quality is further characterized by reputation, currency, unique information, credibility and accuracy in this study. This study indicates it is not enough just to provide comprehensive and high quality content. Participants also care about design, accessibility, and item characteristic criteria in relation to layout, organization, ease-of-use, language, cost, availability, speed, item type and item length. Twenty-one evaluation criteria highlight the complexity and dynamic nature of list and document evaluation.

Third, findings of this study associate elements with criteria that participants applied. Although evaluation activities are not fully supported at both list and document evaluation levels, less support is at list level. Since current lists do not offer enough elements that users need for evaluation, participants had to spend extra time to go through document evaluation. Moreover, some of the documents that participants selected from the list were considered not useful at all because some of the key elements were not presented. This study shows that current IR systems do not provide enough elements to support quality and document characteristic evaluation at the list level. The design of IR systems need to incorporate the following elements to support quality related elements by presenting information in relation to reputation (reputation ranking, source, etc), unique information (highlighting unique information), credibility (reviews on

credibility), and accuracy (reviews on accuracy, comparison of similar information, etc). Additionally, adding more system features, such as customer reviews, document descriptions will also help the evaluation. For effective evaluation, it is also important to incorporate currency related elements such as date, and document related elements, such as length, whether there are pictures, and videos in the lists.

Fourth, this study reveals the problems of lacking of integration between list and document evaluation. Participants had to go through two levels of evaluation in order to find the information they need. This not only leads to more time spent on the evaluation but also confusion and losing tracking of which documents have been evaluated. Two options can be taken to integrate list and document evaluation. One option is to combine list and documents into one level, lists can be presented on the main page while documents can be opened in a pop up window next to it. By doing that, users do not need to click each document and click back to look for the next one. Google offers previews of screenshots of each document on the list next to the list, but it is hard to see its content because of the small size of screenshots. The other option is to allow users to select all the relevant documents altogether, and then open each documents one by one.

Fifth, this study presents the complete picture of evaluation activities, which also include pre and post evaluation activities. Participants performed evaluation activities going beyond the basic examination of information. Comparatively speaking, more evaluation activities were performed during document evaluation than list evaluation. Evaluation activities, to some extent, is a mini version of the search process which includes searching, browsing, examining, comparing, looking for help and registering to access fulltext, etc. System design needs to support quick list evaluation to identify relevant documents as well as thorough document evaluation to extract useful information for users as suggested at the third aspect. While the pre-activities appear straight forward, the post-activities of both list and document evaluation indicate a complex variety of user behaviors depending mainly on the outcomes of the evaluation. System design needs to support the integration between list and document evaluation as suggested at the fourth aspect. Moreover, the evaluation activities highlight a non-linear information search process. It is also crucial to design IR systems to facilitate pre, evaluation, and post activities. One design suggestion is to make evaluation path available to guide users to easily move from pre-evaluation activity to evaluation activity and from evaluation activity to its post activity. Finally, system design needs to go beyond highlighting the key words in the document in supporting evaluation activities. It is important to offer search function, explicit and implicit help in relation to how to find specific information as well as different features (e.g. zoom in) to identify specific information (tables, pictures, numbers, etc). It is also helpful to make it easy for users to access previous documents for comparison purpose.

Sixth, this study presents the time spent on list and document evaluation, in particular highlighting the brief and extended evaluation for both lists and documents. This study not only explains why users spend short period in evaluating documents as identified by Jansen and Spink (2003) but also reveals that uselessness and inaccessibility of documents are the key factors leading to quick evaluation while the factors behind longer evaluation are more related to users' interest in the topic, their tasks as well as the documents themselves. Since users take longer time in evaluating documents than lists, IR systems need to play a more active role in facilitating users effectively evaluating individual documents. For lengthy documents, IR systems need

to offer table of contents, best passages, abstracts, key facts, etc for successful evaluation. This study indicates that participants spent more time in document evaluation when performing research related search topics than they did when performing personal search topics. As users in general work on their scholarly topics repeatedly on the same area, it is useful for system design to allow users to create their profiles and alerts that contain their evaluation criteria for these topics. These targeted services will generate more relevant documents and help users reduce time for document evaluation. Personal tasks are more easily categorized, such as shopping, travel, medical information, entertainment, do-it-yourself, and news. Designing searching, browsing, and evaluation mechanisms for specific types of personal tasks that enable users to search, browse, view, and compare specific information/format can save more time for evaluation activities.

Conclusion

Evaluation is an essential activity that users perform in their information retrieval process. Evaluation represents a mini information retrieval process itself. Both users and IR systems need to collaborate together in order to effectively accomplish evaluation activities. Users play a more active role in evaluation because judgments have to be made by themselves. At the same time, users do need IR systems to support them by offering, identifying, highlighting or extracting key information needed for evaluation.

The main contributions of this study lie in its theoretical and practical implications of the findings. Theoretically, the findings of this study help researchers understand the nature of evaluation by examining multiple dimensions of evaluation ranging from criteria, elements, activities, to time. More important, the associations between dimensions, in particular, evaluation criteria and elements are analyzed to identify criteria that are less supported by current IR systems. By comparing list and document evaluation and their associated dimensions, these two types of evaluation are no longer considered as separate activities; instead, they are interrelated and can be transformed and integrated. Practically, the findings of this study suggest more elements, especially at list level to be available to support users applying their evaluation criteria. Integration of list and document evaluation and integration of pre, evaluation and post evaluation activities for the interface design is the absolute solution for effective evaluation.

Of course, this study also has its limitations. First, despite the wide variety of participants, the sample of 31 participants limits the generalizability of the findings. Second, while think aloud protocols and log data provide what, how, and why participants evaluate list and documents in their retrieval process, think aloud protocols cannot capture all their thinking in the evaluation process, and log data do not offer what specific elements they view and the duration of their examination time for each element.

In order to generalize the results, further research needs to involve more participants with a variety of tasks. Specifically, more in-depth examinations of evaluation activities need to be performed, such as using eye tracking devices to analyze in what ways and to what extent users evaluate every element of list and documents. Moreover, further research needs to test different prototypes of interfaces that are designed based on the findings of this study and other related studies.

References

Aula, A., Majaranta, P. and Räihä, K.-J. (2005), "Eye-tracking reveals the personal styles for search result evaluation", in Costabile, M.F. and Paternò, F. (Eds), *INTERACT 2005 IFIP TC13 International Conference, Lecture Notes in Computer Science*, Vol. 3585, pp. 1058-61.

Bade, D. (2007), "Relevance ranking is not relevance ranking or, when the user is not the user, the search results are not search results", *Online Information Review*, Vol. 31 No. 6, pp. 831-44.

Bales, S. and Wang, P. (2005), "Consolidating user relevance criteria: a meta-ethnography of empirical studies", poster paper in *Proceedings of the 68th ASIS&T Annual Meeting*, Vol. 42 No. 1, available at: http://onlinelibrary.wiley.com/doi/10.1002/meet.14504201277/abstract

Barry, C.L. (1994), "User-defined relevance criteria: an exploratory study", *Journal of the American Society for Information Science*, Vol. 45 No. 3, pp. 150-9.

Barry, C.L. (1998), "Document representations and clues to document relevance", *Journal of the American Society for Information Science*, Vol. 49 No. 14, pp. 1293-303.

Barry, C.L. and Schamber, L. (1998), "Users' criteria for relevance evaluation: a cross-situational comparison", *Information Processing & Management*, Vol. 34 Nos 2/3, pp. 219-36.

Borlund, P. (2003), "The concept of relevance in IR", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 10, pp. 913-25.

Chi, E.H., Pirolli, P., Chen, K. and Pitkow, J. (2001), "Using information scent to model user information needs and actions on the web", *SIGCHI '01*, Vol. 3 No. 1, pp. 490-7.

Fitzgerald, M.A. and Galloway, C. (2001), "Relevance judging, evaluation, and decision making in virtual libraries: a descriptive study", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 12, pp. 989-1010.

Fogg, B.J., Soohoo, C., Danielson, D.R., Marable, L., Stanford, J. and Trauber, E.R. (2003), "How do users evaluate the credibility of web sites? A study with over 2,500 participants", *Proceedings of the Conference on Designing for User Experiences, San Francisco*, available at: http://portal.acm.org/ citation.cfm?doid_997078.997097

Granka, L., Joachims, T. and Gay, G. (2004), "Eye-tracking analysis of user behavior in www search", in Sanderson, M., Järvelin, K., Allan, J. and Bruza, P. (Eds), *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004,* ACM, New York, NY, pp. 478-9.

Greisdorf, H. (2003), "Relevance thresholds: a multi-stage predictive model of how users evaluate information", *Information Processing and Management*, Vol. 39 No. 3, pp. 403-23.

Hansen, P. and Karlgren, J. (2005), "Effects of foreign language and task scenario on relevance assessment", *Journal of Documentation*, Vol. 61 No. 5, pp. 623-39.

Hoeber, O. and Yang, X.D. (2009), "HotMap: supporting visual explorations of web search results", *Journal of the American Society for Information Science and Technology*, Vol. 60 No. 1, pp. 90-110.

Holsti, O.R. (1969), *Content Analysis for the Social Sciences and Humanities*, Addison-Wesley, Reading, MA.

Janes, J.W. (1994), "Other people's judgments: a comparison of users' and others' judgments of document relevance, topicality, and utility", *Journal of the American Society for Information Science*, Vol. 45 No. 3, pp. 160-71.

Jansen, B.J., Spink, A., Bateman, J. and Saracevic, T. (1998), "Real life information retrieval: a study of user queries on the Web", *SIGIR Forum*, Vol. 32 No. 1, pp. 5-17.

Jansen, B. and Spink, A. (2003), "An analysis of web information seeking and use: documents retrieved versus documents viewed", *Proceedings of the 4th International Conference on Internet Computing, Las Vegas, NV*, pp. 65-9.

Jansen, B. and Spink, A. (2006), "How are we searching the world wide web? A comparison of nine search engine transaction logs", *Information Processing & Management*, Vol. 42 No. 1, pp. 248-63.

Jansen, B., Zhang, M. and Schultz, C. (2009), "Brand and its effect on user perception of search engine performance", *Journal of the American Society for Information Science & Technology*, Vol. 60 No. 8, pp. 1572-95.

Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G. (2005), "Accurately interpreting clickthrough data as implicit feedback", *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 15-19.

Kelly, D. and Cool, C. (2002), "The effects of topic familiarity on information search behavior", *Proceedings of the Second ACM/IEEE Joint Conference on Digital Libraries, ACM, New York, NY*, pp. 74-5.

Kelly, D., Murdock, V., Yuan, X.J., Croft, W.B. and Belkin, N.J. (2002), "Features of documents relevant to task- and fact- oriented questions", *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM '02), ACM, New York, NY*, pp. 645-7.

Maglaughlin, K.L. and Sonnenwald, D.H. (2002), "User perspectives on relevance criteria: a comparison among relevant, partially relevant, and not-relevant judgments", *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 5, pp. 327-42.

Metzger, M., Flanagin, A. and Zwarun, L. (2003), "College student web use, perceptions of information credibility, and verification behavior", *Computers & Education*, Vol. 41 No. 3, pp. 271-90.

Park, T.K. (1994), "Toward a theory of user-based relevance: a call for a new paradigm of inquiry", *Journal of the American Society for Information Science*, Vol. 45 No. 3, pp. 135-41.

Purgailis Parker, L.M. and Johnson, R.E. (1990), "Does order of presentation affect users' judgment of documents?", *Journal of the American Society for Information Science*, Vol. 41 No. 7, pp. 493-4.

Rele, R.S. and Duchowski, A.T. (2005), "Using eye tracking to evaluate alternative search results interfaces", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 49 No. 15, pp. 1459-63.

Rieh, S.Y. (2002), "Judgment of information quality and cognitive authority in the web", *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 2, pp. 145-61.

Saracevic, T. (2007a), "Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 13, pp. 1915-33.

Saracevic, T. (2007b), "A relevance: a review of the literature and a framework for thinking on the notion in information science. Part III: behavior and effects of relevance", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 13, pp. 2126-44.

Savolainen, R. and Kari, J. (2006), "User-defined relevance criteria in web searching", *Journal of Documentation*, Vol. 62 No. 6, pp. 685-707.

Schamber, L., Eisenberg, M.B. and Nilan, M.S. (1990), "A re-examination of relevance: toward a dynamic, situational definition", *Information Processing & Management*, Vol. 26 No. 6, pp. 321-43.

Shneiderman, B., Byrd, D. and Croft, W.B. (1997), "Clarifying search: a user-interface framework for text searches", *D-Lib Magazine*, Vol. 3 No. 1, available at: www.dlib.org/dlib/january97/retrieval/01shneiderman.html

Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. (1999), "Analysis of a very large web search engine query log", *ACM SIGIR Forum*, Vol. 33 No. 1, pp. 6-12.

Spink, A., Greisdorf, R. and Bateman, J. (1998), "From highly relevant to non-relevant: examining different regions of relevance", *Information Processing & Management*, Vol. 34 No. 5, pp. 599-621.

Spink, A., Wolfram, D., Jansen, M.B.J. and Saracevic, T. (2001), "Searching the web: the public and their queries", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 3, pp. 226-34.

Strauss, A. and Corbin, J. (1990), *Basics of Qualitative Research*, Sage, Newbury Park, CA.

Su, L.T. (1994), "The relevance of recall and precision in user evaluation", *Journal of the American Society for Information Science*, Vol. 45 No. 3, pp. 204-17.

Tombros, A., Ruthven, I. and Jose, J.M. (2005), "How users assess web pages for information seeking", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 4, pp. 327-44.

Vakkari, P. and Hakala, N. (2000), "Changes in relevance criteria and problem stages in task performance", *Journal of Documentation*, Vol. 56 No. 5, pp. 540-62.

Voiskunskii, V.G. (1997), "Evaluation of search results: a new approach", *Journal of the American Society for Information Science*, Vol. 48 No. 2, pp. 133-42.

Wang, P. and 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*, Vol. 49 No. 2, pp. 115-33.

Wang, P. and Soergel, D. (1999), "A cognitive model of document use during a research project. Study II: decision at the reading and citing stages", *Journal of the American Society for Information Science*, Vol. 50 No. 2, pp. 98-114.

Xie, I., Benoit, E. III and Zhang, H. (2010), "How do users evaluate individual documents? An analysis of dimensions of evaluation activities", *Information Research*, Vol. 15 No. 4, available at: http://informationr.net/ir/15-4/colis723.html

Xu, Y.C. and Chen, Z. (2006), "Relevance judgments: what do information users consider beyond topicality", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 7, pp. 961-73.

Further reading

Balatsoukas, P., Morris, A. and O'Brien, A. (2009), "An evaluation framework of user interaction with metadata surrogates", *Journal of Information Science*, Vol. 35 No. 3, pp. 321-39.

Cool, C., Belkin, N.J., Frieder, O. and Kantor, P. (1993), "Characteristics of texts affecting relevance judgments", in Williams, M.E. (Ed.), *Proceedings of the 14th National Online Meeting, Learned Information, Oxford*, pp. 77-84.

Park, T.K. (1993), "The nature of relevance in information retrieval: an empirical study", *Library Quarterly*, Vol. 63 No. 3, pp. 318-51.

Saracevic, T. (1969), "Comparative effects of titles, abstracts and full text on relevance judgments", *Proceedings of the American Society for Information Science*, Vol. 6, pp. 293-9.

Schamber, L. (1991), "Users' criteria for evaluation in a multimedia environment", *Proceedings of the 54th ASIS Annual Meeting*, Vol. 28, pp. 126-33.

About the authors

Iris Xie is a Professor in the School of Information Studies at the University of Wisconsin-Milwaukee. Her research interests and expertise focus on information seeking and retrieving, in particular interactive information retrieval between users and IR systems and its implications for the design and evaluation of a variety of IR systems in the digital age. She is the principal investigator on many research grants awarded by different agencies, which include the Institute of Museum and Library Services (IMLS), Online Computer Library Center (OCLC), ALISE, etc. She has a strong record in publishing refereed articles and presenting at international conferences in the field of library and information science. Iris Xie is the corresponding author and can be contacted at: hiris@uwm.edu

Edward Benoit III is a Doctoral Student in Information Studies at the University of Wisconsin-Milwaukee. His dissertation project focuses on the use of social tagging within digital collections. In addition, his research areas include digital archives, digital libraries, information retrieval, and multimedia preservation. He received his BA in History, MA in History, and MLIS from the University of Wisconsin-Milwaukee.