PERGAMON

Information Processing and Management 38 (2002) 401–426

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

A user-centered approach to evaluating human interaction with Web search engines: an exploratory study

Amanda Spink *

School of Information Sciences and Technology, The Pennsylvania State University, 511 Rider I Building, 120 S. Burrowes St., University Park, PA 16801, USA

Received 2 November 2000; accepted 10 May 2001

Abstract

A growing body of studies is developing approaches to evaluating human interaction with Web search engines, including the usability and effectiveness of Web search tools. This study explores a user-centered approach to the evaluation of the Web search engine Inquirus – a Web meta-search tool developed by researchers from the NEC Research Institute. The goal of the study reported in this paper was to develop a user-centered approach to the evaluation including: (1) *effectiveness*: based on the impact of users' interactions on their information problem and information seeking stage, and (2) *usability*: including screen layout and system capabilities for users. Twenty-two volunteers searched Inquirus on their own personal information topics. Data analyzed included: (1) user pre- and post-search questionnaires and (2) Inquirus search transaction logs. Key findings include: (1) Inquirus was rated highly by users on various usability measures, (2) all users experienced some level of shift/change in their information problem, information seeking, and personal knowledge due to their Inquirus interaction, (3) different users experienced different levels of change/shift, and (4) the search measure precision did not correlate with other user-based measures. Some users experienced major changes/shifts in various user-based variables, such as information problem or information seeking stage with a search of low precision and vice versa. Implications for the development of user-centered approaches to the evaluation of Web and information retrieval (IR) systems and further research are discussed.

1. Introduction

The effective performance of Web search tools is an important challenge for Web designers and a significant growing area of study. How to improve the effectiveness of Web search tools

* Tel.: +1-814-865-4454; fax: +1-814-865-5604.
E-mail address: spink@ist.psu.edu (A. Spink).

0306-4573/02/$ - see front matter
PII: S0306-4573(01)00036-X

and how to measure their effectiveness is a crucial area of research. The evaluation of information retrieval (IR) systems has been a major area of study for more than 40 years (Saracevic, 1995; Sparck Jones & Willett, 1997). Web and IR systems evaluation is also important for users. How are users to evaluate their own interactions with Web/IR systems? Most approaches to Web/IR evaluation are for researchers, not users. Precision and recall are measures largely designed and used by researchers. These measures have limitations when used to measure IR system effectiveness (Hersh et al., 2000; Saracevic, 1995). New user-centered evaluation measures are needed for users and also designers of Web technologies. Meta-search tools enable users to enter a query that is processed concurrently against a number of different commercial Web search engines, such as Excite, Google, Alta Vista, etc. Such tools as WebCrawler and Dogpile are becoming popular for Web searching, each offering different features and services. Web meta-search tools are becoming a fundamental and important part of seeking information on the Web.

This paper reports results from an exploratory study evaluating a user-centered approach to Web/IR systems using the Inquirus Web meta-search tool developed by researchers at the NEC Research Institute (Lawrence & Giles, 1998a,b). The evaluation approach explored in this study is based on a user-centered approach discussed by Spink and Wilson (1999), who proposed that search engine evaluation should focus on measuring the impact of users' interactions on their information problem and their moves through the different stages of their information seeking process. In real life, users evaluate Web tools in the context of their information seeking and retrieving behaviors beyond precision and usability measures (Spink & Wilson, 1999). The goal of the study reported in this paper is to develop a user-centered approach to explore the evaluation of Inquirus *usability* and *effectiveness*, including changes users experience in their information problem and information seeking stages as a result of their interaction with Inquirus, and attempt to measure those changes. Users' interactions with Inquirus were evaluated using a range of standard usability measures, precision measure, and measures based on impact of Inquirus interaction on users' shifts and changes in their information problem and information seeking stages. This evaluation approach is based on the theoretical framework and model that conceptualizes Web searching within a user's information seeking and retrieving context outlined in the next section of this paper.

2. Related studies

2.1. IR system evaluation studies

2.1.1. Systems approaches

Information retrieval research and evaluation has been largely based on variations of precision and recall measures. The TREC conferences use various precision and recall measures as the basis of comparing the performance of IR systems (Sparck Jones, 1995, 1999). Many researchers have discussed the limitations of precision and recall measures, and called for the development of new IR evaluation measures (Saracevic, 1995). Hersh et al. (2000) examined task-centered approaches to IR system evaluation and documented the disconnection between precision/recall and user success. The importance of evaluation in IR has also been the focus of much attention (Borlund &

Ingwersen, 1997; Harter & Hert, 1997; Rees, 1966; Saracevic, 1995; Tague & Schultz, 1989). The nature, manifestations and effects of human evaluation behavior are both challenging and elusive. Saracevic (1995) suggested that evaluation was an integral part of IR, and stated, "the issue and challenge for any and all IR evaluations are the broadening of approaches and getting out of the isolation and blind spots of single level, narrow evaluations. How can interaction be ignored in IR evaluation at any level?"

2.1.2. Task-centered approaches

Task-centered approaches to IR evaluation are contributing to a better understanding of the user/IR system interaction process. By focusing on the user's task resolution, some studies have proposed new IR evaluation measures, including the *value of the search results as a whole* measure of Su (1998), and Tague and Schultz's (1989) *informativeness* measure. Reid (2000) highlights task-centered approaches to IR evaluation. Greisdorf and Spink (2001) propose a *median measure* to supplement precision and recall. They found that the median point of relevance distributions (on an interval scale) correlates with the point where relevant and partially relevant items begin to be retrieved. IR evaluation approaches have also been adopted in studies evaluating Web search engines.

Task-centered approaches have largely not taken into account the user's information seeking processes that provide a context for their IR interaction. Users' tasks can be variable, but all users are moving through an information seeking process during which their information problem may evolve or change.

2.2. Web search engine evaluation

A growing number of studies have developed approaches to evaluating Web search engines. Most studies are limited to small queries and search numbers, dichotomous relevance judgments, and precision and recall measures, and in general do not use real user relevance judgments (Leighton & Srivastava, 1999; Losee & Paris, 1999). Recent studies have produced valuable insights into Web search engine performance. In a large-scale study Lawrence and Giles (1998b) found that individual Web search engines generally do not cover a majority of Web sites. A recent study by Gordon and Pathak (1999) identifies two forms of search engine evaluations – testimonials or industry assessments, and shootouts in laboratory settings, and provides a valuable comparison of previous search engine evaluation studies. They also found: (1) fairly low absolute retrieval effectiveness, (2) differences in Web search engine retrieval and precision, and (3) a lack of overlap in retrieval by Web search engines.

Few studies have developed user-based approaches to Web evaluation that attempt to measure the impact of the Web interaction on users' information seeking processes. The goal of the study reported in this paper was to evaluate Inquirus for: (1) *effectiveness*: based on the impact of users' interactions on their information problem and information seeking stage, and (2) *usability*: including screen layout and system capabilities for users. This approach to Web evaluation is based on a theoretical framework and model derived from previous human information seeking and retrieving research, discussed in the next section of the paper.

3. Theoretical framework

Fig. 1 presents a theoretical framework for an evaluation approach based on an integrated model of information seeking and retrieving that includes relevance judgments (on the scale highly relevant, partially relevant, partially not relevant, and not relevant) made within a set of situated actions by information seekers within interactive search sessions with Web systems over a period of time.

This model extends and integrates a model of relevance level, region and time developed by Spink (1998) and Spink, Greisdorf, and Bateman (1998), and a model of human information seeking developed by Wilson (1997). The model has various elements:

- *Time* is represented by movements or shifts during interactive search episodes, including tactics, information problem, strategies, terms, feedback, goal states, or uncertainty, and between searches.
- *Interactive search episodes* are represented by interactive IR models, including those of Belkin, Cool, Stein, and Theil (1995), Ingwersen (1992, 1996), and Saracevic (1996b, 1997).
- The *set of situated actions* includes actions, decisions and judgments during an interactive search episode, e.g., relevance, magnitude or strategy feedback, tactics, search strategies, or search terms within a search episode. Sets of situated actions occur during interactions.

Therefore, sets of situated actions may occur during each interactive search episode that takes place over a period of time. This integrated model provides a framework for the development of empirical research to integrate interactive IR research and develop IR evaluation measures within information seeking contexts, and explore their interactive search episodes within their changing

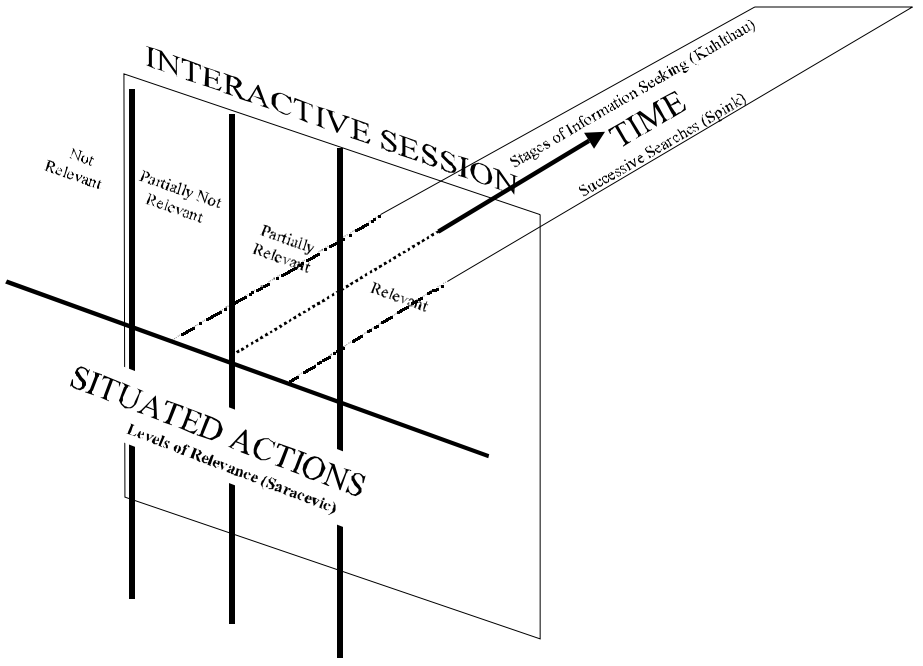

Fig. 1. Model of situated actions, interactive session and time.

information seeking contexts. Effective IR evaluation measures must account for IR interactions taking place within the context of information seeking behaviors. Each facet of the model is briefly discussed to develop a framework for an integrated view of the interactive search processes within changing information seeking contexts.

3.1. Time

An IR evaluation measure should account for the element of time in information seeking behavior. Such a measure includes consideration of time and accounts for the effect of the changes and shifts that occur at the IR interaction level (Robins, 2000; Xie, 2000) that affect the shifts at the information problem level. The set of situated actions during IR interactions occurs over a period of time, such as judgments during an evolving information seeking process or during successive search episodes. Each set of situated actions may be plotted within four attributes: (1) interaction time, (2) successive searching time, (3) information seeking time, and (4) problem solving time.

1. *Problem solving processes* are represented in Wilson's (1997) problem solving model of information seeking behavior in which interactive search episodes provide the information inputs to the problem solving process through which the information seeker's uncertainty level is reduced;
2. *Information seeking stages* are represented in the model by the Kuhlthau (1991) information search process model;
3. *Successive searches over time* relate to the same or evolving information problem (Spink, 1996).

Time may be plotted from the initiation of an information seeker's information problem, including the measures associated with the attributes of searches and judgments, in a visual model. This study initially uses the Saracevic (1996a) stratified model of IR interaction within our integrated model of information seeking and searching. The model views the interaction as a dialogue between participants, *user* and *computer* (system) through an interface at a *surface* level. *Interaction is the interplay between various levels*. On the user side elements involve at least these levels: *cognitive*, *affective*, and *situational*. The model depicts some elements from information seeking models and interactive IR models that describe the phenomena of successive and related searches of digital environments by humans during an information seeking process.

3.2. Interactive search sessions

IR interactions related to the single search episode can be represented in the model by different theoretical interactive IR models – such as Ingwersen's (1992, 1996) cognitive model of IR interaction, Belkin et al.'s (1995) episodic interaction model, or Saracevic's (1996a, 1997) stratified model of IR interaction, or a combination of elements of all interactive IR models. Therefore, as interactive search sessions occur they exist within the context of time facets such as successive searches, information seeking process and information problem solving. To extend the model, the next level within the facet of time and the interactive search session is the set of situated actions.

3.3. Set of situated actions

The set of situated actions includes actions, decisions and judgments during an interactive search episode, e.g., relevance, magnitude or strategy feedback, tactics, search strategies, or search

terms within a search episode. Situated actions occur and form part of interactive IR episodes that occur within information seeking and then problem solving time. A complete model would include all situated actions during an interactive search episode. In the model shown, we explore a specific set of situated actions related to relevance judgments (Spink et al., 1998). Some specific situated actions, displayed in Fig. 1, include relevance judgments.

3.4. Relevance judgments

The degrees of users' relevance judgments are situated within one of four relevance regions in Fig. 1 – highly relevant, partially relevant, partially not relevant, and not relevant. Therefore, the region of an information seeker's relevance judgment can be situated according to relevance level and relevance degree. For example, an information seeker may judge a retrieved item highly relevant based on the relevance level of topicality. The ability to plot these cognitive relations by inference is an attribute of the second dimension in the set of situated actions, the information seeker's region of relevance attributed to these relations or non-relations. This second attribute also contains positive and negative aspects that can be labeled and depicted graphically.

4. Research objectives

The objectives were to conduct a study to evaluate:

1. The usability of the Inquirus Web meta-search tool;
2. The impact of searching Inquirus on users' information problems and information seeking processes.

5. Research design

5.1. Data collection

Data were collected from 22 volunteer users who were faculty, students or administrators at the University of North Texas during March–April 1999 (Table 1).

Users responded to a call for study participation, seeking those interesting in using a new Web meta-search tool, sent out through the University of North Texas email system. The average age

Table 1
Basic data

Number of users	22
Mean age of users	44.5 years (range: 24–72)
Number of males	13
Number of females	9
Mean search terms per query	2.9
Mean queries per user	8.6
Mean Web site viewed per query	2.3
Mean pages (10 Web sites) viewed	25.7

of the users was 44.5 years (range = 24–72); nine females and 13 males were included. Users searched Inquirus on their own information problem. Before searching Inquirus each user was first briefed by a research assistant on the basic features of Inquirus.

5.1.1. Questionnaires

Before accessing Inquirus, each study participant completed a *consent form*, a *demographic form* and a *pre-search questionnaire*. After their Inquirus interaction each user completed a *post-search questionnaire*. The aim of the pre- and post questionnaires was to capture the state of each user in a number of areas before and after their Inquirus interaction. This allowed the measurement of changes or shifts by users resulting from their interaction with Inquirus. The questionnaires were based on questionnaires used in two major studies of online searching by Saracevic, Kantor, Chamis, and Trivison (1988), and Spink, Wilson, Ellis, and Ford (1998).

Users were asked to give their perceptions on a number of issues related to issues represented in Fig. 2 – a general model of information seeking and searching. This model enhances a similar model presented in Saracevic et al. (1988).

Event	Class of Variables
Information seeker has an information problem to resolve	Information seeker pre-search characteristics Cognitive style Problem statement Knowledge level Information seeking stage Uncertainty level
Information seeking process related to information problem	Information seeking behaviors
Information seeker formulates their information problem into a question	Question statement Question analysis
Presearch interaction with a search intermediary	Intermediary characteristics
Formulation of the search strategy (terms and tactics)	Pre-search characteristics: information seeker
Searching activity and interactions	Search strategy Search characteristics Search processes
Delivery of responses to the information seeker	Items retrieved Forms delivered
Evaluation of output	Relevance Utility
Information seeker evaluation of impact of search	Information seeker post-search characteristics Problem statement Knowledge level Information seeking stage Uncertainty level
Usability	Satisfaction Learning time Amount of information provided Screen arrangement and layout

Fig. 2. A general model of information seeking and searching.

Data were gathered on each element of Fig. 2. Questions and scales for each element of the data collection were adapted from previous studies by Saracevic et al. (1988) and Spink, Wilson, Ford, Foster and Ellis (forthcoming).

Appendix A details the questionnaires used in the study.

5.1.2. Transaction logs

Each user was audiotaped during their Inquirus searching interaction. The audiotapes were professionally transcribed and then qualitatively analyzed to identify users' comments on their Inquirus searching and usability.

5.1.3. Relevance judgments

Users recorded relevance judgments on a worksheet for the first 20 Web sites they retrieved.

Item #	Relevance (place vertical line indicating how relevant this item is)	Judgments (check one box only)				Levels of relevance (check box(es) most important to your judgment)										Describe
		NR	PNR	PR	R	S	T	P	U	M	NS	NT	NP	NU	NM	
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														
	NR |————| R														

This worksheet was developed and used during studies of relevance judgments by Spink and Greisdorf (2001), Spink (1998), and Spink et al. (in press).

5.2. Data analysis

Quantitative and qualitative analysis methods were used. Quantitative analysis concentrated on statistical analysis of the data from questionnaire Likert scales. Standard statistical tools from the Excel package were employed. Part of the quantitative analysis will be a search for and test of statistical models appropriate for these types of events, starting with correlation analysis. However, qualitative methods predominated. The reason for this is that the data involved, from search terms selected in queries to answers to questions as to reasons, interactions, or results, were largely textual. Qualitative methods are based on grounded theory (Strauss & Corbin, 1990). The

qualitative methods included: content analysis, structuring of taxonomies depicting structure and relations of various types of actions and specific variables, derivation of various diagrams and structures to describe shifts, and principles and criteria derived from grounded theory research.

6. Results

6.1. Users' information problem description

The 22 topics searched for by each participant during the study are shown in Table 2.

Users information problems covered a broad range of topics, including the arts, social sciences, physical sciences and education. On average, users reported experience with six Web search engines at the time of the Inquirus interaction. All but three users had conducted a previous Web search for information on their topic and 10 users reported that their previous Web search contributed to the current Inquirus search.

6.2. Search data

Table 3 shows the basic search data from the study.

Inquirus users' interactions were not typical of general Web users. Specifically, Inquirus users' mean of 2.9 terms per query and 8.6 queries per search session was larger than general Web users'

Table 2
Users' information problems

User number	Information problem
1	Controlled vocabulary
2	Electronic books
3	Business
4	Art
5	Puerto Rican statehood
6	Gender differences in newspaper preferences
7	Digital watermarks
8	Information retrieval
9	Music history
10	Digital imaging
11	Public administration
12	Vijayan Pillai
13	Early childhood development
14	Early childhood development
15	Speech communication
16	Adult learning
17	Mental health
18	Decision making by couples
19	Cultural color preference
20	Historical study
21	Bill Clinton
22	Husband and wife communication

Table 3
Basic search data

Total user queries	191
Total user terms	570
Total pages accessed	448
Mean terms per query	2.9
Mean queries per user	8.6
Mean pages viewed per query	2.3

queries of 2.4 terms and lack of query modification by general users as identified by Spink, Wolfram, Jansen and Saracevic (2001).

6.3. Usability measures

How users rated Inquirus on various usability measures is shown in Table 4.

Users' rating included judgments on the amount of information provided on the screen, the screen arrangement and layout more than other aspects of the system.

6.3.1. Presentation of Inquirus search results

Overall, many users rated the presentation of results as desirable. One user commented that Inquirus was "very helpful, because of the way it broke things down and ranked them", as a "Search engine was easy enough" and "I found it useful". Alternatively, some users pointed to systems problems: "The blurb didn't help tell me what the web page contained. It was necessary to

Table 4
Usability measures

Usability criteria	Mean user ratings	Range – user ratings
Frustration to satisfaction	5.5	1–9
Difficult to easy	6.3	3–9
Dull to stimulating	5.5	1–9
Time to learn		
Lengthy to easy		
Speed	6.7	1–9
Response time	6.9	1–9
Ease of searching	7	4–9
Amount of information provided		
Inadequate to adequate	7.5	2–9
Screen arrangement		
Logical to illogical	7.5	3–9
Screen layout		
Inadequate to adequate	7.1	4–9
Screen terminology		
Not helpful to helpful	6.5	2–9
Messages	6.2	2–9
Overall reaction to Inquirus	5.9	2–9

click through to make relevance judgments", "I don't like it that a second window opens when I click on something I find interesting", "Confusing codes (e.g., M-4K)", "The format with the arrows was somewhat confusing, as well as having to jump to a different window. I always forgot to click on the separate window" and "Some of the places to execute the search (like find) function were not what I'm used to".

6.3.2. Comparison with other search engines

A complete list of users' positive and negative comments on the differences between Inquirus and other Web searching tools is provided in Fig. 3.

6.3.3. User suggestions

A complete list of users' suggestions to improve Inquirus functionality and search features is provided in Fig. 4.

Users' suggestions were passed directly to the Inquirus developers to help them refine the systems' capabilities.

Complete List of Users Comparative Comments
Superior to Yahoo and Infoseek by far.
Looking forward to having access to such a search engine
GOOD --> keep this coming. Better than other available websites
More intuitive
A positive thing for people who don't have the time to search one engine at a time
I like the combination search of some many search engines. Made it faster to refine and eliminate search topics
OK. But not as good as Spider
Did not find anything on my selected search. HotBot by itself has retrieved pertinent information using similar search terms
I like Excite's top 10 matches and then the ability to pick which match is best
Didn't notice the capability to narrow search by directing a follow on search from within found pages
I've used a different engine in the past that allowed me to do that (Alta Vista???)

Fig. 3. Complete list of users' comparative comments.

Complete List of User Suggestions

I would like to have more capabilities to restrict material to be returned from search (e.g. no .com; .edu only) .The in-context text seemed choppy. Some started with last 2 words of previous sentence. A lot of the key material returned in far down the page, some people will never get to it. They will do something new before they scroll down to it.

Need hyperlinks within page at top.

Needs better Help.

Who cares about stuff that can be downloaded.

What's the no search terms section?

No orientation to the authors & authorities section.

Why not put relevance ranked items before unranked?

I would like it to also access online journals and periodicals.

Fig. 4. Complete list of user suggestions.

6.4. Inquirus search effectiveness

6.4.1. Relevance judgments

Users recorded relevance judgments on a worksheet for the first 20 Web sites they retrieved. Results are summarized in Table 5.

The mean precision per search was 27.7%. Precision was calculated by dividing the number of relevant and partially relevant items retrieved by the number of items retrieved per search.

6.4.2. Overall effectiveness of the search

A complete list of users' comments on the effectiveness of their Inquirus search is provided in Fig. 5.

7. User changes during Inquirus interaction

An aim of this study was to examine the impact of Inquirus interaction on the users at various levels. Questions on the pre- and post-search questionnaires collected data on changes that users experienced due to their Inquirus interaction, including changes in their information problem stage, personal knowledge, information seeking stages, uncertainty level, understanding of their information problem, and resolution of their information problem (Table 6).

Individual users experienced different changes and reactions to their Inquirus interaction. For example, the search of highest precision was User 11 with 63%. However, User 11 experienced no change in information problem stage as a result of Inquirus interaction, being at Stage 2 before and after the search. User 11 did report a shift of one information seeking stage as a result of the

Table 5
User relevance judgments (first 20 sites retrieved)

User no.	Relevance judgments						Clicked through to make judgments		
	NR	PNR	PR	R	Total	Precision (%)	Yes	No	%
1	14	3	2	1	20	15	10	10	50
2	6	3	7	4	20	55	20	0	100
3	14	2	1	3	19	21	10	9	55
4	3	4	2	0	9	22	3	6	33
5	16	2	1	1	20	10	6	14	30
6	16	1	1	2	20	15	10	10	50
7	14	3	1	2	20	15	14	6	70
8	11	1	3	5	20	15	1	19	5
9	10	4	0	4	18	22	9	9	50
10	18	0	1	1	20	10	0	20	0
11	4	1	2	12	19	63	13	6	72
12	19	0	1	0	20	5	7	13	35
13	1	1	2	1	5	60	2	3	40
14	20	0	0	0	20	0	2	18	10
15	7	1	6	6	20	60	3	17	15
16	13	0	2	5	20	35	20	0	100
17	6	2	3	6	17	52	4	13	24
18	12	3	1	4	20	25	7	13	37
19	14	2	1	1	18	11	15	3	88
20	18	0	1	1	20	10	14	6	70
21	3	6	5	6	20	55	20	0	100
22	10	3	4	3	20	35	16	4	80
Total	236	33	38	59	366	Mean = 27.7%	164	192	46

Inquirus interaction from Stage 3 to Stage 4. Alternatively, User 8 retrieved few relevant Web sites (precision 15%), but did report a one-stage shift in information problem and a two-stage shift in information seeking stage from Stage 2 to Stage 4.

Overall, different users experienced different levels of change on various criteria.

7.1. Changes related to the user's information problem

7.1.1. Change in information problem stage

As Table 6 shows, different users experienced different levels of change in their *information problem stage* due to their Inquirus interaction.

- 5 (31%) users shifted one information problem stage.
- 13 (50%) users stayed in the *same* information problem stage.
- 4 (19%) users shifted to a *previous* information problem stage.

Interestingly, half the study participants remained in the same information problem stage – measured before and after their Inquirus interaction – and felt the Inquirus interaction had not affected a change. Nearly one in six users shifted to a previous information problem stage. These users may have over-estimated their information problem stage in the pre-search form or felt the

Complete List of User Inquirus Effectiveness Comments

It was worth trying, as I learned from my mistakes regarding terminology. I feel I got a few relevant documents

I am glad that I had a chance to try out a new and innovative searcher for information. I just wish it could have found the info I needed

Seems like a good product. However, it may just not adequately cover my research topic. (I seemed to only find personal web pages that mention the topic. No educational or research sites resulted.)

I have such limited ability on the Internet I am not sure that I can competently judge the system due to my uncertainty about the limits of the information that I am seeking. The concept seems good but has not greatly increased the information that I am seeking.

Did not find anything on my selected search. HotBot by itself has retrieved pertinent information using similar search terms

The problem was the information that I wanted may not be on the Web (...) the amount of info retrieved did not seem to be extensive.

There were many separate pages within a larger site (don't know the terminology here) that made it seem repetitive. One recurring instance was the English First (EF) page. I kept coming to different areas within the page.

(...) it may just not adequately cover my research topic. (I seemed to only find personal web pages that mention the topic. No educational or research sites resulted.)

Fig. 5. Complete list of user Inquirus effectiveness comments.

interaction gave them information that convinced them they were actually at an earlier stage than they thought.

7.1.2. Change in information seeking stage

As Table 6 shows, different users experienced different levels of change in their *information seeking stage* on their topic due to their Inquirus interaction.

- 11 (45%) users shifted at least one stage.
- 7 (31%) users shifted to a *previous* information seeking stage.
- 5 (22%) users stayed in the *same* information seeking stage.

7.1.3. Change in uncertainty level

Different users experienced different levels of change in the *uncertainty level* of their topic due to their Inquirus interaction.

- 7 (31%) users shifted one uncertainty level.
- 4 (19%) users shifted to a *previous* uncertainty level.
- 11 (50%) users stayed at the *same* uncertainty level.

Table 6
Questionnaire data

User no.	Search precision (%)	User pre-search information problem stage (1–4)	User post-search information problem stage (1–4)	User information problem stage shift	User change in information problem understanding from pre- to post-search (1–58)	User pre-search information seeking stage (1–6)	User post-search information seeking stage (1–6)	User change in information seeking stage from pre- to post-search	User change in personal knowledge from pre- to post-search (1–58)	User judgment of Inquirus contribution to their information problem resolution (1–58)
1	15	1	1	Same stage	43	1	4	3 stage +	37	40
2	55	2	3	1 stage +	9	1	5	4 stage +	12	38
3	21	1	1	Same stage	0	1	3	2 stage +	11	21
4	22	1	1	Same stage	2	6	3	3 stage −	30	29
5	10	3	3	Same stage	1	1	5	4 stage +	32	16
6	15	2	2	Same stage	37	4	3	1 stage −	2	3
7	15	1	1	Same stage	33	2	3	1 stage +	31	33
8	15	2	3	1 stage +	45	2	4	2 stage +	44	34
9	22	3	2	1 stage −	2	2	2	Same stage	2	3
10	10	3	3	Same stage	5	6	5	1 stage −	6	29
11	63	2	2	Same stage	9	3	4	1 stage +	21	28
12	5	3	2	1 stage −	26	4	4	Same stage	33	12
13	60	2	2	Same stage	36	6	3	3 stage −	33	39
14	0	1	1	Same stage	28	2	1	1 stage −	37	36
15	60	1	1	Same stage	15	4	5	1 stage +	43	44
16	35	2	1	1 stage −	9	4	1	3 stage −	5	7
17	52	3	4	1 stage +	43	6	6	Same stage	51	53
18	25	3	3	Same stage	2	5	5	Same stage	14	28
19	11	1	2	1 stage +	56	2	5	3 stage +	55	54
20	10	2	3	1 stage +	7	3	3	Same stage	22	34
21		3	3	Same stage	4	1	5	4 stage +	8	7
22	10	2	1	1 stage −	26	2	1	1 stage −	21	19

7.1.4. Change in understanding of information problem due to the interaction

As Table 7 shows different users experienced different levels of change in their *understanding of their information problem* due to their Inquirus interaction.

7.2. Contribution of Inquirus interaction to information problem resolution

Table 8 shows the contribution of Inquirus interaction on a *user's information problem resolution*.

- All users reported that their Inquirus interaction contributed at some level to the resolution of their information problem.

7.2.1. Change in personal knowledge on their topic

Different users experienced different levels of *change in their personal knowledge* on their topic due to their Inquirus interaction (Table 9).

- All users reported a change in their personal knowledge on their topic.

7.3. Significant correlations between pre- and post-search assessments

Table 10 shows significant correlations (<0.05) between pre- and post-search user assessments.

Table 7
User change in information problem understanding due to the Inquirus interaction (scale: 0–58)

Level of change	Number of users	% of users
0	1	4.5
1–9	10	46
10–19	1	4.5
20–29	3	13.5
30–39	3	13.5
40–49	3	13.5
50–58	1	4.5
	22	100

Table 8
Contribution of Inquirus interaction to user information problem resolution (scale: 0–58)

Level of contribution	Number of users	% of users
0	0	0
1–9	4	18
10–19	3	13.5
29–30	5	23
31–40	6	27.5
41–49	2	9
50–58	2	9
	22	100

Table 9
Change in users' personal knowledge on their topic

Level of change	Number of users	% of users
0	0	0
1–9	5	23
10–19	3	13.5
20–29	3	13.5
30–39	7	32
40–49	2	9
50–58	2	9
	22	100

Table 10
Significant correlations between pre- and post-search user assessments

Pre-search variable	Post-search variable	Significance level (<0.05)
Clearer or more focused thinking on the problem	Better definition of the information problem	0.76
	Higher level of definition of the degree of a real information problem	0.7
	More familiar with the language of the problem	0.83
	Higher level of definition of the degree of a real information problem	0.65
Higher level of recognition of a real information problem	Higher level of definition of the degree of the intended use for the information	0.73
Higher at the problem defining or resolving state	More familiar with the language of the problem	0.68
Higher level of specific knowledge or expertise of the problem		0.73
More certain about the progress of the work on the information problem		0.69
Clearer about the progress of the work on the information problem		0.73
Less importance for monitoring of particular sources for maintaining awareness of developments in the related topic		−0.71
Lower familiarity with the language of the problem	More significant changes in relevance judgment criteria	−0.65
	More difficult to evaluate the results of the search	−0.67
Clearer about the progress of the work on the information problem	More certain that an effective way of presenting the results can be found	0.71
Less certainty in the definition of the problem	More serendipitous results are retrieved	−0.76

The findings show that those users with a clearer understanding of their information problem in the pre-search stage had a higher level of understanding of their information problem, its definition and language at the post-search stage. A lower familiarity with the language and definition of their information problem led to more change and serendipitous results. The level of language knowledge and information problem definition is linked with a level of certainty and change due to the interaction with Inquirus.

8. Discussion

Overall, users found Inquirus to be a usable Web searching tool. Inquirus was fairly highly rated for a complex Web searching tool, although users did comment on limitations of the system. Users rated Inquirus highly on the amount of information retrieved and the arrangement of the information on the screen. User suggestions also provided ideas for improving the Inquirus features and capabilities.

In addition, from an evaluation perspective, each user experienced some level of change in their information problem, personal knowledge, and information seeking stage due to their Inquirus interaction. Different users experienced different levels of change in their information problems and information seeking process. Results show that search precision did not correlate with the user-based evaluation measures or users' perceptions of change in their information problem and information seeking stage, e.g., some users experience major changes in their information problem and information seeking stages with a search of low precision and vice versa. Each user shifted or changed their information problem. Different users experienced different shifts. As each user was searching on a different information problem, it is not possible to compare users' shifts in detail. Each user has their own interpretation of their shifts. However, the dimension of these shifts is important, particularly shifting at the information problem/seeking level and interaction level.

Those users with a less clear understanding of their information problem in the pre-search stage had a lower level of understanding of their information problem, its definition and language at the post-search stage. Less familiarity with the language and definition of their information problem led to less change and serendipitous results. The level of language knowledge and information problem definition was related to the level of certainty and change due to the interaction with Inquirus.

The results of this study led to the development of an approach to IR evaluation that is discussed in the next section of the paper.

8.1. IR evaluation measures

IR evaluation measures can be based on the reality of human interaction with IR systems. In line with this proposition, the following questions are proposed:

What are meaningful criteria for IR evaluation measures?

What is a meaningful IR evaluation measure for information seekers?

What is important to measure and how to measure it?

8.1.1. Criteria for evaluation measures

Meaningful evaluation measures could be useful to Web/IR researchers, designers, and people using such systems by measuring what is important to information seekers in the form of a self-assessment tool. This paper proposes that:

- Effective IR evaluation measures can be meaningful and important for information seekers.
- What is important to information seekers is the resolution of their information problems.
- To resolve their information problem, information seekers move through the *changes/shifts* in their information seeking process.
- If information seekers interact with IR systems, then an IR evaluation measure can relate the effectiveness of their IR system interaction to shifts or changes in their information problem due to their interaction with the IR system.
- An IR evaluation measure can be a self-assessment tool.
- An important IR evaluation measure for information seekers is their *information problem shift*.

8.1.2. Information problem shift

The concept of an information problem shift is based on interactive IR research that reflects the process of interaction of humans with IR systems as they progress through their information seeking process. The key issue for IR system users is not the number of items retrieved or the precision of the search. What information seekers care about is how they are progressing toward resolving their information problem. Information seekers primarily care about their own personal information problems. An IR system interaction may lead to a complete resolution of their information problem, or a partial resolution, or a slight or major change in their information problem. The IR system itself is relatively secondary in the reality of their information behaviors. This paper proposes that:

> The effectiveness of an IR system can be measured in terms of the change or shift in the human information problems due to IR system interaction.

IR system effectiveness can be measured as a shift by an individual information seeker or an aggregate of information seekers.

Information problem shift may be assessed and operationalized by measuring the change in an information seeker's information problem stage by measuring their information problem stage before and after their interaction with an IR system. A major weakness of existing IR evaluation measures is their inability to reflect changes or shifts, e.g., changes in an information seeker's understanding of his/her information problem due to interaction with an IR system.

Meaningful measures must involve data collected from information seekers BEFORE and AFTER their IR interaction. In this case, we are actually measuring a change, not just collecting data after an interaction. Collecting data before their IR interaction provides a benchmark for comparison with the data collected after the IR interaction. IR evaluation measures that ONLY measure AFTER an IR interaction are relatively limited. By measuring the information problem stage before and after their interaction with an IR system, we can measure the impact of the IR system interaction on the information problem solving process. Of

course, the effectiveness of the IR system is realized in the interaction in the context of specific situated actions and cognitive, problem and knowledge states during the interaction. However, if an information seeker does not experience some type of shift in information problem process – represented by shifts in cognitive, problem and knowledge states – then the IR system interaction has not been effective.

Researchers seek to form and develop conceptualizations of the phenomena observed. This paper presents a model of a theoretical framework for the IR evaluation measure – information problem shift – within an information seeking context.

8.1.3. Operationalization

This paper proposes that the IR evaluation measure – information problem shift – be conceptualized and operationalized as:

Information problem shift (IPS) = Information seekers' information problem stage after their IR Interaction (AIPST) subtracted from their information problem stage before their IR interaction (BIPST). IPS = AIPST – BIPST.

For example, on an 100 mm line, if the information seeker's information problem stage before their IR interaction (BIPST) was 45/100 and their information problem stage after their IR interaction (AIPST) was 85/100 – then their IPS would be 40.

These data will allow researchers to examine the relationship between these variables, including the before- and after-IR-interaction assessments, to examine the utility of the measure *information problem shift*, particularly in relation to other IR evaluation measures such as precision.

9. Conclusion

This exploratory study has contributed to general IR evaluation theory, models, measures and techniques, and offers an approach to the development of a series of IR evaluation measures of value as user self-assessment tools. Future research is needed to further test and evaluate the value of the measures proposed. The strength of an IR evaluation tool is based on the strength of the models that underpin its development. The approach underpinning this research is based on the model, theories and empirical research at the nexus of IR and human information behavior research. Further research is needed that looks beyond traditional approaches to IR evaluation to consider the information seeking context of the user.

Acknowledgements

The author acknowledges the assistance of Steve Lawrence from NEC Research and C. Lee Giles from Penn State for the provision of the Inquirus meta-search tool for this study.

Appendix A. Questionnaire contents

Users' information problems

Before their Inquirus search users were asked to describe their information problem.

- Please write a description of your *information problem* below:
- Please list the *search terms* you plan to use:

Various measures collected data on aspects of users' information problems.

Users' problem solving stage

Before their Inquirus search users indicated if this was new problem area for them.

- Is this a new problem area for you?

Likert scale questions

Likert-type scales were used to obtain users' data on many aspects of their information seeking, problem solving and other cognitive variables (see questions listed below) before and after the Inquirus interaction. The following type of interval scale was used that moves from left to right as the intensity of users definition increases.

How familiar are you with the specific domain or problem-oriented terminology in current use?

Information problem

- How well is your information problem currently defined?
- How would you describe your thinking about the problem at this stage?
- How would you describe your level of interest in your information problem at this stage?
- Indicate how would you describe the degree to which your intended use was well defined.
- Please indicate on the scale below how you would describe the degree to which your intended use for the information is well defined.
- How would you rank the amount of knowledge you possess and the gaps of knowledge?
- How would you rank the amount of knowledge you possess in relation to the broader domain to which your problem is related?
- How familiar are you with the *language* of your information problem?

Uncertainty

How certain are you that:

1. You have recognized a real information problem to investigate?
2. You have defined the information problem appropriately?
3. Your information problem *can* be resolved?
4. An effective way of presenting the results can be found?
5. Relevant information is available and can be found on the Web?

Previous information seeking

- Maintaining awareness of developments in relation to this topic through the monitoring of particular sources.
- Systematically working through a particular source to locate material of interest.
- Verifying and checking the accuracy of information.

At this point in your search which have you engaged in?

- Browsing or semi-directed Web searching in an area of potential interest.
- Differentiating Web sources of information on the basis of the nature and quality of the material examined.
- Following Web links.

- Maintaining awareness of developments in relation to this topic through the monitoring of particular sources on the Web.
- Systematically working through a particular source on the Web to locate material of interest.
- Verifying and checking the accuracy of information on the Web.

Did the searching involve (checklist for interviewer – client can indicate as many as necessary):

Web-based database(s) or information sources
On-line database(s), searching done by a librarian (intermediary)
On-line database(s), searching done on your own
Printed index(es)
Library catalogue(s)
Library collection without use of a catalogue
Own collection
Colleague's collection
Other: please specify
None

Various question formats were used to obtain information from the user on their previous use of Web searching tools.

- Web search engines you have ever used?

Excite_____ Alta Vista_____ Snap_____ HotBot_____ Direct Hit_____ Google_____ MSN_____ GoTo_____ Infoseek_____ Lycos_____ Northern Light_____ Yahoo_____ Yahoo Inktomi_____ Euroseek_____ Other search engines_____

∗ Web meta-search tools you have ever used?

No:_____ Yes:_____ If yes, indicate which ones listed below: Dogpile_____ Metacrawler_____ Others (List:_____)

- How successful was the previous Web search or searches in finding desired information?
- Did any of the above previous Web searches contribute to the formulation of your present question? If yes, in what way:
- Did anything arise out of previous Web searches that were useful in carrying out a search for you? If yes, in what way:

On the scale below, please indicate how comprehensive you would like your Web search to be:

Language: are you able to use any Web material you might find in languages other than English?
Dates: How far back in time would it be useful to take the Web search? Five years, ten years, longer? How recent would you like the material to be?
Kinds of information: Are there any particular kinds of information that will be of interest to you?
Web sites to avoid: Are there any Web sites that you already know that are unlikely to be of value to you?

Information seeking stage

Users were asked to indicate their current information seeking stage before and after their Inquirus interaction. Information seeking stages were drawn from Kuhlthau's information searching process (ISP) model (1991), but not presented in order of the model.

Please select one of the following categories as best matches your current information seeking stage.

_____*Collection* – Having focused my problem I am now collecting specific relevant information.

_____*Exploration* – I am now identifying specific information sources that I think will be of use to me.

_____*Formulation* – The information I have found has enabled me to form a clearer focus on the problem.

_____*Initiation* – I have recognized that I need information at this stage of my work.

_____*Presentation* – I am in the process of finishing the collection of information for this stage of my work.

_____*Selection* – I have identified the general area in which I need information.

Users were asked to indicate their work progress on their information problem on various levels.

How do you feel about the progress of your work on your information problem at this point?

1. Very uncertain	|————————|	Very certain
2. Pessimistic	|————————|	Optimistic
3. Confused	|————————|	Clear
4. Frustrated	|————————|	Relieved
5. Doubtful	|————————|	Confident
6. Dissatisfied	|————————|	Satisfied
7. Disappointed	|————————|	Pleased

Changes due to Inquirus interaction

- After the Inquirus search users were asked to indicate on three Likert-types scales any levels of change they experienced.
- Did any changes occur in your understanding or definition of your information problem as the result of your Inquirus search?
- Did any changes occur in your personal or internal knowledge of your topic due to the Inquirus search?
- Did you change your criteria for relevance judgments of items retrieved due to the Inquirus search?
- How difficult was it to evaluate the results of the search?
- How focused or precise were the search results in relation to your information problem?
- How much extraneous or non-relevant information was retrieved?
- Estimate the completeness of the search in retrieving information on your information problem.
- Estimate the contribution the information retrieved has made to resolving your information problem.

How certain are you that:

You have recognized a real information problem to investigate?

You have defined the information problem appropriately?

You have an information problem that can be resolved?

An effective way of presenting the results can be found?

Relevant information is available and can be found?

To what degree was:

Your expectancy fulfilled or exceeded by the retrieved information?
The information retrieved novel or new to your topic?
The information retrieved serendipitous for you, i.e., unexpected?

User satisfaction

How satisfied are you with the results of the Inquirus search?
Various questions elicited information from the users on Inquirus usability.
How would you rate the worth of the Inquirus search in relation to your time:
5 Worth much more than the time taken
4 Worth somewhat more than the time taken
3 Worth about as much as the time taken
2 Worth less than the time taken
1 Practically worthless

Overall reactions to Inquirus

Terrible							Wonderful	
1	2	3	4	5	6	7	8	9
Frustrating							Satisfying	
Dull							Stimulating	
Difficult							Easy	
Inadequate							Adequate	
Rigid							Flexible	

Users were also asked to evaluate the Inquirus interface based on the following aspects:
Amount of information displayed on the screen

Arrangement of information displayed on the screen
Illogical Logical
Screen layout
Inadequate Adequate
Screen terminology
Not helpful Helpful
Messages that appear on the screen
Not helpful Helpful
Time to learn to use Inquirus
Lengthy Easy
System speed
Too slow Fast Enough
Response time
Too slow Fast Enough
Ease of searching
Difficult Easy

Finally, users were asked to write general comments on Inquirus searching.

References

Belkin, N., Cool, C., Stein, A., & Theil, C. (1995). Cases, scripts, and information seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, *9*(3), 379–395.

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, *53*, 225–250.

Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: The retrieval effectiveness of search engines. *Information Processing and Management*, *35*, 141–180.

Greisdorf, H., Spink, A. (2001). Median measure: An approach to IR systems evaluation. *Information Processing and Management*, *37*(6), 843–857.

Harter, S., & Hert, C. (1997). Evaluation of information retrieval systems: Approaches, issues and methods. *Annual Review of Information Science and Technology*, *32*, 1–94.

Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., & Olson, D. (2000). Do batch and user evaluations give the same results? In *Proceedings of the 23rd ACM–SIGIR international conference on research and development in information retrieval, Association of Computing Machinery, Athens, Greece* (pp. 17–24).

Ingwersen, P. (1992). *Information retrieval interaction*. London: Taylor Graham.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, *52*(1), 3–50.

Kuhlthau, C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, *42*, 361–371.

Lawrence, S., & Giles, C. L. (1998a). Context and page analysis for improved Web search. *IEEE Internet Computing*, *2*(4), 38–46.

Lawrence, S., & Giles, C. L. (1998b). Searching the world wide web. *Science*, *280*(5360), 98–100.

Leighton, H. V., & Srivastava, J. (1999). First 20 precision among World Wide Web search services (search engines). *Journal of the American Society for Information Science*, *50*(10), 870–881.

Losee, R. M., & Paris, L. H. (1999). Measuring search-engine quality and query difficulty: Ranking with target and freestyle. *Journal of the American Society for Information Science*, *50*(10), 882–889.

Rees, A. (1966). The measurability of relevance. *Proceedings of the American Documentation Institute, Washington, DC* (Vol. 3) (pp. 225–234).

Reid, J. (2000). A task oriented non-interactive evaluation methodology for information retrieval systems. *Information Retrieval*, *2*, 115–129.

Robins, D. (2000). Shifts of focus on various aspects of user information problems during interactive information retrieval. *Journal of the American Society for Information Science*, *51*(10), 913–928.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In *Proceedings of the 18th ACM–SIGIR international conference on research and development in information retrieval, Association of Computing Machinery, Seattle, WA* (pp. 138–146).

Saracevic, T. (1996a). Modeling interaction in information retrieval (IR): A review and proposal. In *Proceedings of the annual meeting of the American Society for Information Science: Vol. 33* (pp. 3–9).

Saracevic, T. (1996b). Relevance reconsidered. Information science: Integration in perspectives. In *Proceedings of the second conference on conceptions of library and information science, Copenhagen, Denmark* (pp. 201–218).

Saracevic, T. (1997). Extension and application of the stratified model of information retrieval interaction. In *Proceedings of the annual meeting of the American Society for Information Science: Vol. 34* (pp. 3–9).

Saracevic, T., Kantor, P., Chamis, A. Y., & Trivison, D. (1988). A study of information seeking and retrieving: Background and methodology. *Journal of the American Society for Information Science*, *39*(3), 161–176.

Sparck Jones, K. (1995). Reflections on TREC. *Information Processing and Management*, *31*(3), 291–314.

Sparck Jones, K. (1999). Further reflections on TREC. *Information Processing and Management*, *36*(1), 37–85.

Sparck Jones, K., & Willett, P. (1997). *Readings in information retrieval*. London: Morgan Kaufmann.

Spink, A. (1996). A multiple search session model of end-user behavior: An exploratory study. *Journal of the American Society for Information Science*, *47*(8), 603–609.

Spink, A. (1998). Toward a theoretical framework for information retrieval in an information seeking context. In *Proceedings of the second international conference on research in information needs, seeking and use in differing contexts, Sheffield, August 13–15, 2000* (pp. 21–34). London: Taylor Graham.

Spink, A., & Greisdorf, H. (2001). Regions and levels: Mapping and measuring users' relevance judgments. *Journal of the American Society for Information Science*, *52*(13), 161–173.

Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to non-relevant: Examining different regions of relevance. *Information Processing and Management*, *34*(5), 599–622.

Spink, A., Wilson, T. D. (1999). Toward a theoretical framework for information retrieval (IR) evaluation in an information seeking context. In *Proceedings of MIRA 99: Evaluation frameworks for multimedia information retrieval applications*, *Department of Computing Science*, *University of Glasgow*, *Scotland*, *April 14–16*, *1999* (pp. 75–92). Available: http://www.ewic.org.uk/ewic/workshop/view.cfm/MIRA-99.

Spink, A., Wilson, T. D., Ellis, D., & Ford, N. (1998). Modeling users' successive searching: A National Science Foundation/British Library study. *D-Lib Magazine*, *4*(4) (http://www.dlib.org).

Spink, A., Wilson, T.D., Ford, N., Foster, A., & Ellis, D. (forthcoming). Information seeking and mediated searching: Part I. Background and theoretical framework. *Journal of the American Society for Information Science and Technology*.

Spink, A., Wolfram, D., Jansen, B.J., & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, *53*(2), 226–284.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.

Su, L. T. (1998). Value of search results as a whole as the best single measure of information retrieval performance. *Information Processing and Management*, *34*(5), 557–579.

Tague, J., & Schultz, R. (1989). Evaluation of the user interface in an information retrieval system: A model. *Information Processing and Management*, *25*(4), 377–389.

Wilson, T. D. (1997). Information behavior: An interdisciplinary perspective. *Information Processing and Management*, *33*(4), 551–572.

Xie, H. (2000). Shifts of interactive intentions and information seeking strategies in interactive information retrieval. *Journal of the American Society for Information Science*, *51*(9), 841–857.