



COVER SHEET

Spink, Amanda and Griesdorf, Howard and Bateman, Judy (1998) From highly relevant to not relevant: examining different regions of relevance. *Information Processing and Management* 34(5):pp. 599-622.

Accessed from <http://eprints.qut.edu.au>

Copyright 1998 Elsevier.

Information Processing and Management (1998), 34(5), 599-622.

**FROM HIGHLY RELEVANT TO NOT RELEVANT: EXAMINING DIFFERENT
REGIONS OF RELEVANCE**

Amanda Spink, Howard Greisdorf and Judy Bateman

School of Library and Information Sciences

University of North Texas

P.O. Box 311068

Denton TX 76203

USA

*To whom all correspondence should be addressed: spink@lis.admin.unt.edu

Abstract

User relevance judgments are central to both the systems and user-oriented approaches to information retrieval (IR) systems research and development. User-oriented relevance research has also operated on two largely unconnected tracks. First, a relevance level track that examines users' criteria for relevance judgments. Second, a regions of relevance track that examines the measurement of users' relevance judgments. Users judgments and criteria for highly relevant items have been central issues for much of the relevance research. Findings are presented from four separate studies of relevance judgments by 55 users, conducting their initial online search on a particular information problem. In three studies, the number of items judged "partially" relevant (on a scale of relevant, partially relevant or not relevant) was positively correlated with different aspects of changes in users', including: (1) information problem definition, (2) search intermediaries' perceptions that a user's question and information problem has changed during the mediated search interaction, (3) personal knowledge due to the search interaction, and (4) criteria for making relevance judgments. Users with high knowledge and topic levels were more likely to judge items as highly relevant. Differences between users' criteria for highly, partially and non-relevant items are also identified. Findings suggest the need to expand the framework for relevance research and further identify the characteristics of the middle region of relevance or partial relevance as: (1) partially relevant items may play an important role in the early stages of a user's information seeking process over time for a particular information problem and (2) a relationship may exist between partially relevant items retrieved and changes in users' information problems during an information seeking process. Results also suggest that partially relevant items may be useful at the early stages of users' information seeking processes. We propose a useful concept of relevance as a relationship and an effect on the movement of a user through the iterative stages of their information seeking process. Users' relevance judgments can also be plotted on a three-dimensional spatial model of relevance level, region and time. Implications for the development of IR systems, searching practice and relevance research are also discussed.

1. Introduction

The basic objective of information retrieval (IR) is often stated as the retrieval of relevant items (texts, images, sounds) matched to a user query, which in turn represents a user's information need evolving from a problem-at-hand. Thus, the notion of relevance and users' relevance judgments are critical to the theory and research of IR. In practice, users' relevance judgments exist on a continuum of relevance regions from highly relevant, through partially relevant to non-relevant. However, retrieval by exact-match systems assumes a binary, yes–no, relevance decision: focusing on two regions of relevance — highly relevant and not relevant. Retrieval by best-match systems provides a ranking according to a probability of relevance, but still in the end assumes a binary decision as to a cutoff. When either type of system is evaluated, using precision and recall as measures, a binary relevance judgment by users or their surrogates is incorporated. User-centered research and evaluation, generally asks users to assess retrieved items as relevant (i.e., highly relevant), partially relevant, or not relevant. However, in the analysis of results, the three points (or more depending on the study) of relevance are collapsed into two sets: relevant and not relevant. In other words, retrieval is most often presented in two sets. One set of highly relevant items and the other not relevant items, with highly and partially relevant items combined.

Although relevance is not a concrete binary concept for IR system users, IR system researchers have been primarily concerned with matching a user's query to the items stored in textual databases and retrieving highly relevant items. As mentioned, the measures traditionally employed to show the relative success or failure of this “matching” — recall and precision — have relied on binary user relevance judgments of the retrieved information. Automatic relevance feedback systems that incorporate users' relevance judgments have also been shown to improve retrieval (Harman, 1992; Spink and Losee, 1996). Alternatively, the actual study of user relevance judgments has generally been conducted by user-oriented IR researchers who seek to model the nature of user-IR system interaction (Schamber, 1994). This user-oriented research has generally been distinct from the IR systems research, with little impact on actual IR system design. The IR systems and user-oriented research have largely been operating on separate and unconnected tracks (Saracevic, 1996a).

We seek to examine the characteristics the four regions of relevance. In particular we examine the fuzzy middle or partial regions of relevance. This article reports results from four studies that examine the characteristics of user relevance judgments within relevance regions. In the process, it questions an assumption of both user and systems oriented IR research that users always require the most highly relevant items when submitting a query to

an IR system. This has been an underlying assumption of both exact- and best match IR systems, and automatic relevance feedback techniques. However, findings from four studies reported in this paper suggest the need to further investigate and compare the characteristics of the three regions (high, partial and not relevance) of users' relevance judgements further. The findings also suggest that partially relevant items may be useful in the early stages of users' information seeking processes. These studies show that for users conducting their initial online search on a particular information problem, the items judged partially relevant (on a scale of relevant, partially relevant or not relevant) relate to important changes in users' information problems and information seeking processes.

The next section of this paper outlines a theoretical framework for the study of relevance regions within the development of a three-dimensional spatial model of relevance level, region and time.

2. Theoretical framework

From the 1960's, the definition and measurement of relevance has been widely debated in information science literature. Saracevic (1975) identified relevance as a key concept in the emergence of information science as a discipline and as a critical factor in development of information science theory and experimentation. He described relevance as an intuitive concept within information science where the meaning and use of relevance is widely understood — at least within the context of IR system evaluation. Recently Schamber et al. (1990) identified relevance as the “most fundamental concept” of information science. Although relevance has been debated for more than three decades, a clear definition or viable operationalization within the context of IR system evaluation has not emerged. The limitations of relevance and assumptions regarding relevance as a basis for IR evaluation have also been challenged by many researchers (Belkin et al., 1982a and Belkin et al., 1982b; Cooper, 1973a and Cooper, 1973b; Doyle, 1963; Ellis, 1984; Meadow, 1985; Newby, 1992).

User oriented relevance research within information science has also progressed within two largely unconnected areas. The first area of research has focused on the level and criteria of user relevance judgments. A second area of study has focused on the regions of users' relevance judgments from highly to not relevant. The next section of this paper examines research focused on investigating the levels of relevance.

2.1. Levels of relevance

Recently, Saracevic (1996b) proposed a stratified IR interaction model depicting IR interaction as the interplay between user levels: cognitive, affective and situational, and computer levels: engineering, processing and content, through an interface level at a surface level (Fig. 1).

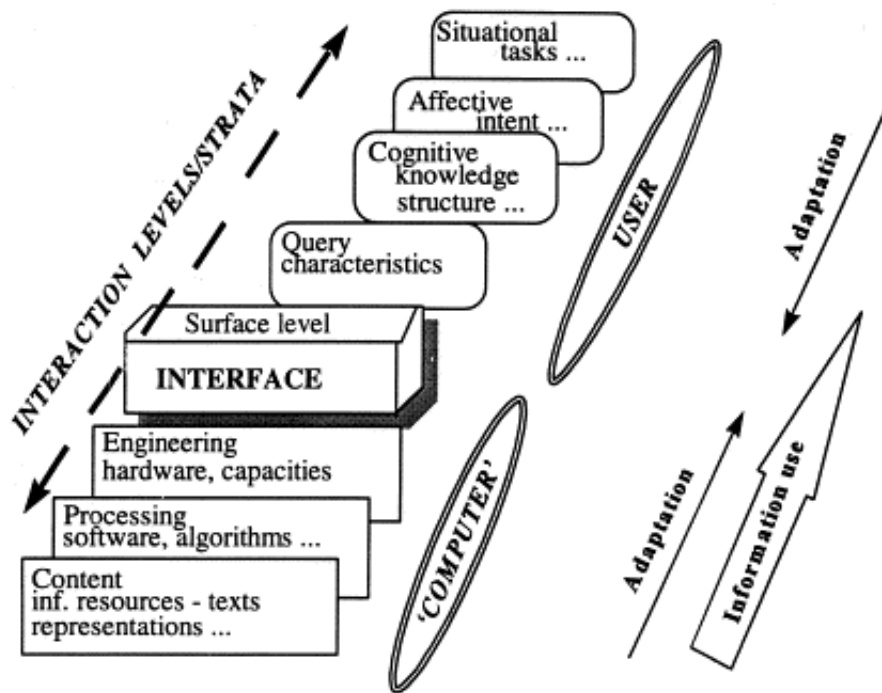


Fig. 1. Elements in the stratified model of IR interaction

Within the cognitive level of IR interaction, Saracevic (1996a) proposed an interdependent system of relevances based on five manifestations or levels of relevance:

(1) *Systems or algorithmic relevance*: relation between a query and information objects (texts) in the file of a system as retrieved, or as failed to be retrieved, by a given procedure or algorithm. Each system represents, organizes and matches to a query using specific methods and algorithms. These methods and algorithms encompass an assumption of relevance, in that the intent is to retrieve a set of texts that the system inferred as relevant to the query. Comparative effectiveness in inferring relevance is the criterion for system relevance.

(2) *Topical or subject relevance*: relation between the subject or topic expressed in a query, and the topic or subject covered by retrieved texts, or more broadly, by texts in the systems

file, or even in existence. It is assumed that both queries and texts can be identified as about a topic or subject. Aboutness is the criterion by which topicality is inferred.

(3) *Cognitive relevance or pertinence*: relation between the state of knowledge and cognitive information need of the user, and texts retrieved, or in the file of the system, or even in existence. Cognitive correspondence, informativeness, novelty, information quality, and the like are criteria by which cognitive relevance or pertinence is inferred.

(4) *Situational relevance or utility*: relation between the situation, task, or problem at hand, and texts retrieved by a system or in the files of a system, or even in existence. Usefulness in decision making, appropriateness of the information in the resolution of the problem, reduction of uncertainty, and the like are criteria by which situational relevance is inferred.

(5) *Motivational or affective relevance*: relation between the intents, goals, and motivations of a user and the texts retrieved by a system or in the files of a system, or even in existence. Satisfaction, success, accomplishment, and the like are criteria for inferring motivational relevance.

Each user criterion for a relevance judgment could be identified within one of Saracevic's levels of relevance. Some relevance levels (e.g., cognitive relevance) may be hard to measure and some levels may interact with each other (e.g., cognitive, situational and affective) and may be difficult to measure separately.

Saracevic's approach follows previous research exploring users' relevance judgments that have produced many studies examining users' criteria for relevant items retrieved from IR systems. Users have been found to employ many criteria besides topicality when making relevance judgments (Barry, 1994; Park, 1993; Schamber, 1991 and Schamber, 1994). For example, quality criteria (accuracy, journal or author reputation) are frequently mentioned as important in user relevance criteria studies (Bateman, 1997; Barry, 1994; Schamber, 1991). Schamber (1994) lists 80 factors or criteria that affect user relevance judgments:

The statement made earlier — that relevance is a multidimensional phenomenon — is, of course, a gross understatement. In fact so many factors have been suggested as affecting relevance judgments that it is not possible to list them all here. The 80 factors in Table 1, however, represent a reasonable sample (p. 19)

Table 1. Basic data for four studies

	Users	Location
Study A	13 end-user searches	University of North Texas
Study B	18 mediated searches	Rutgers University
Study C	13 end-user searches	University of North Texas
Study D	11 end-user searches	University of North Texas

She suggests the list as only a partial listing of factors. These findings further imply that relevance is multifaceted and may not be suitable to be measured as a binary (relevant/not relevant) variable. However, relevance criteria research has generally focused on investigating users' criteria for highly relevant or not relevant items with a limited focus on the criteria and role of users' partial relevance judgments. This approach is based on the assumption that partially relevant judgments are similar or identical to highly relevant judgments and criteria. The highly relevant paradigm has also underpinned the design of ranked retrieval systems and automatic relevance feedback techniques.

Alternatively, a body of research has been not been concerned users relevance criteria, but with the *regions* of users' relevance judgments.

2.2. Regions of relevance

Researchers within this relevance track investigate appropriate ways to measure the degree of users' relevance judgments — from highly relevant to non relevant. These judgments are often related to other factors such as the *a priori* definition of relevance or order of the citations. IR researchers often use triadic interval or categorical scales for relevance judgments (e.g., relevant/partially relevant/not relevant), but collapsed users' relevance judgments into binary scales — relevant/not relevant — to simplify the calculation of precision and recall measures. This approach assumes that no information is lost in the process (Schamber, 1994), and that partial relevance is the same as high relevance.

Many studies have focused on binary (relevant/not relevant) relevance judgments and measures (Barnydt, 1964; Gull, 1956; Janes and McKinney, 1992; O'Conner, 1969; Rees, 1967; Rees and Schultz, 1967, Schamber et al., 1990) and collapsed users' relevant and partially relevant judgments together during their analysis to form the binary scale — relevant and not relevant (Pao, 1993; Saracevic et al., 1988; Schamber, 1994). Magnitude estimation

continuous scales developed by Eisenberg and Hu (1987), Eisenberg (1988) and Rorvig (1988) were also collapsed into a relevant/non relevant binary scale (Schamber, 1994). Saracevic et al. (1988) dealt with the issue in part by doing two analyses on a limited section of their data: one concentrating on what they called “strong” relevance, where only the highly relevant items were included, and by contrast, the other called “weak” relevance, where the partially relevant items were included with highly relevant ones.

However, several studies show that the user's concept and use of relevance are very dynamic, and can depend on the user's experience with the IR system, the user's knowledge of the subject area, and even the order of presentation of citations and citation elements (Eisenberg and Barry, 1988). How relevance judgments are affected by these factors and how they interact with each other is not well understood. Several researchers have found some information initially judged relevant was never obtained or used, and information judged not relevant sometimes appeared on final bibliographies (Kuhlthau, 1993; Sandore, 1990; Smithson, 1990).

Various IR evaluation measures have also been developed based on: (1) binary relevance, e.g., precision, recall, (2) utility, e.g., usefulness, (3) value added, e.g., cost, time, or (4) user satisfaction. Each measure is based on a criterion the user employs to judge high relevance (usefulness, cost, satisfaction). However, these measures are usually determined *a priori* by the researchers (e.g., Su, 1994) and may or may not represent the individual user's definition or operationalization of relevance. Within this approach, high relevance is in the strictest sense is defined as about the topic of the search query or the information problem. The most limited conceptualization looks at high relevance as an innate part of the document, independent of the user, implying that judgments of high relevance can be made by nonusers. This approach is used in the text retrieval conferences (TREC) experiments comparing the performance of different IR systems (Harman, 1993; Sparck Jones, 1995). A broader conceptualization suggests that only the user can judge the relevance of information to the user's information problem. Both approaches are founded on the idea that relevance judgments are made based on the topicality of the information. The measures of usefulness, value, and satisfaction measure other important factors that users may employ in making relevance judgments and are sometimes used in research as an alternative way to define and measure relevance.

However, IR evaluation measures are used primarily by researchers and only partially inform researchers about users' information problems, and how or if the user has been assisted by their interaction with an IR system. Any measure of quality has been largely ignored within IR evaluation measures of relevance. Measures based on usefulness, satisfaction, and value

also do not always correlate well with binary relevance judgments. Users may be very satisfied with a search but find little that is relevant to their need. They may find information relevant but not useful. Relevant information may be ignored because of its lack of availability, both in the physical and economic sense.

Harter (1996) also suggests that the use of these IR evaluation measures has produced relevance assessments that show considerable variation, calling into question their validity as a basis for IR retrieval evaluation. He states that:

Despite known wide variation in relevance assessments in several experimental test collections, the effect of these variations on the evaluation model on which retrieval performance is assessed — that is on the *measurement instrument* — is almost completely unstudied (p. 37).

Saracevic (1995), Harter (1996) and Ellis (1996) also criticized the continued use of the experimental “Cranfield-like” model of IR evaluation. Harter concludes with a call for new evaluation methods that can accommodate the many individual variations and factors that influence relevance judgments. He further suggests that:

Alternatively it is possible that Cranfield-like models, as valuable as they have been over the years, cannot accommodate the many variables that affect relevance judgments. It may well be that the best solutions to the problems here identified would involve the invention of radically different paradigms for evaluation C the design of brand-new evaluation instruments (Harter, 1996, p. 48).

The current IR evaluation measures are also not designed to assist end-users in evaluation of their information seeking behavior (and an information problem) in relation to their use of an IR system. Thus, these measures have limitations for IR system users and researchers.

In summary, with the strong research focus on users' high relevance and not relevant judgments, users' judgments of partial relevance during interactive IR have also not been the subject of much investigation. IR systems and user oriented research have largely operated on two separate and unconnected tracks. User oriented relevance researchers have also operated on two largely unconnected tracks — the relevance level/criteria orientation and a region of relevance orientation. We seek to explore the middle fuzzy region of relevance or partial relevance.

3. Two-dimensional model of relevance level and region

To provide a framework for the connection between these two areas of relevance research, we propose a two-dimensional model of relevance level and region (Fig. 2). This model includes a plane of judgment with both negative and positive aspects of Saracevic's five levels of relevance on the vertical axis and regions of relevance on the horizontal axis.

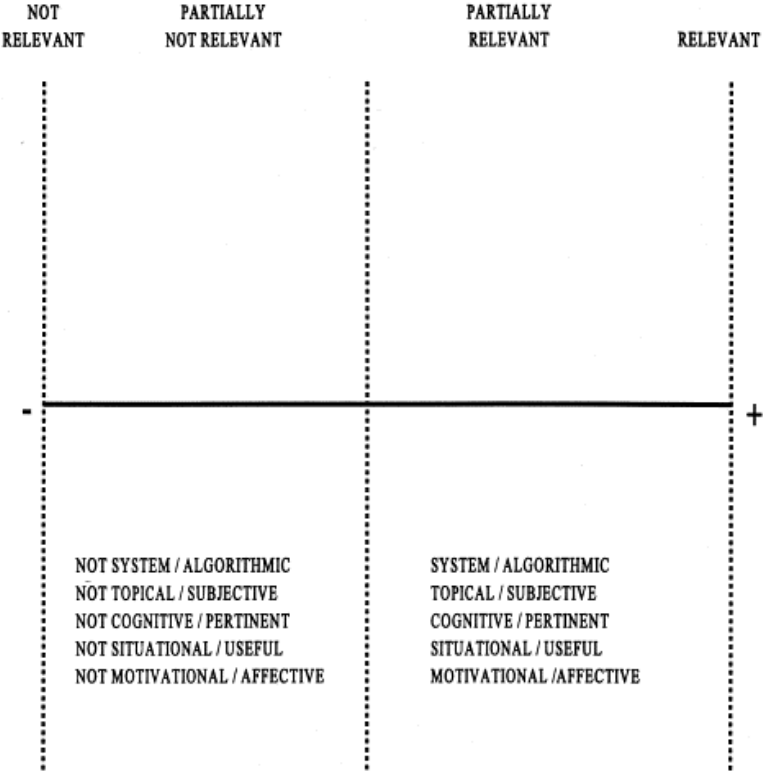


Fig. 2. Plane of judgement

Saracevic's five levels of relevance exist as regions along the vertical axis. Their placement along the vertical axis does not imply a relationship between the levels. Each level exists as a discrete category — not on an interval scale. Each user relevance criterion could also be situated within one or more of Saracevic's five levels of relevance along the vertical axis. Relevance judgments can be situated within one of four relevance regions — highly relevant, partially relevant, partially not relevant, and not relevant. These regions may overlap along the Plane of Judgment. Therefore, a user's relevance judgment can be situated on the dimension of relevance level and relevance region. For example, a user may judge a retrieved item as highly relevant based on the relevance level of topicality. Sometimes multiple criteria are used. For example, the user may have several other criteria that represent several levels to judge high relevance — affective (“I like the author”) or situational (“I can locate or access this easily”).

To further extend the theoretical framework for our examination of the characteristics of different relevance regions, and in particular partial relevance judgments, we include the time dimension of users' relevance judgments.

3.1. Time dimension of relevance

IR researchers have generally not examined users' relevance judgments in the context of time or changes in users' information seeking process. A key assumption is that users conduct only one IR search when seeking information on a particular information problem. If they conduct second and subsequent searches on the same problem, these are considered independent of each other and any previous relevance judgments. Recent research is exploring the contextual and time dimensions of users' IR interactions and relevance judgments.

3.1.1. Single search assumption

Robertson and Hancock-Beaulieu (1992) and Spink (1996) and Spink (1997) show the limitations to this one search paradigm by demonstrating that users often conduct successive searches over time during their information seeking on the same information problem, and that these searches are connected. Spink (1996) identified two groups of IR system users — the initial search user or the user conducting their first search on a particular topic, and the successive search user with more than one search on their particular topic. This finding is also supported by the extensive information behavior and seeking research that shows humans seek information on a particular information problem in stages over time (Ellis, 1989; Kuhlthau, 1993; Wilson, 1981 and Wilson, 1997). Users' information problems are also modified during an information seeking process and shifts that occur in the interactive IR search (Robins, 1997; Saracevic et al., 1997; Xie, 1997). A shift is underway towards a more contextual view of interactive IR and a dynamic information seeking approach to relevance that includes the time dimension of relevance judgments. In particular, two recent papers provide an exploration of more dynamic approaches to relevance.

Schamber et al. (1990) posit a dynamic, situational approach to a conceptualization of relevance. Their primary focus is the criteria that users employ when making relevance judgments, since these criteria “as observed from the user's perspective, may contribute to a more comprehensive and useful understanding of the dimensions of relevance (p. 771)”.

They suggest that “the literature points to a view of relevance as a multidimensional cognitive concept that plays a role in the dynamic process of information evaluation (p. 774)”. Three issues characterize the limitations of the static approach to defining and measuring relevance. The first issue is the lack of research into users' cognitive perceptions of their situational environment and their use of an IR system. A second issue is the assumption that IR system users are passive and engage in static information evaluation behaviors. A final issue is the extent of individual variation when making relevance judgments. Although Schamber et al. (1990) suggest an approach to reexamining relevance in a dynamic context, they do not offer a definition, operationalization, or evaluation measure based on a dynamic notion of relevance. They do not develop a measure that more adequately and accurately capture changes and variations in user relevance judgments over time.

A recent paper by Harter (1992) presents a cognitive or psychological view of relevance for IR evaluation. Pointing to the limitations of topical relevance to users, Harter suggests that exposure to information should have a measurable effect on the user's cognitive state and that users prefer information “that will cause a change of some kind — information that will have an effect on their cognitive state (p. 603)”. Harter defines relevance as “a theoretical concept of psychology, as a relation between an assumption (premise) and a context (p. 605)”. Although Harter views relevance as a stimulus to a subject judgment by a user that causes a measurable change in the user's dynamic, contextually determined cognitive state, he also does not offer an approach to measuring these changes. He theorizes that the user's information need acts as the context at the time the user interacts with an IR system and thus defines his or her cognitive state. Harter acknowledges that the user's information need is dynamic but does not further explore psychological relevance within the user's information seeking process.

We suggest that the existence of users' successive search episodes highlight the need to question the highly relevant assumption and further explore the role of users' partial relevance judgments to their successive search process and information seeking process. Borlund and Ingwersen (1997) also suggest that the concepts of relevance needs to include users “relative” and “partial” non-binary judgments.

The framework for an information-seeking approach to relevance includes consideration of level, region and time of a user's relevance judgment. An information seeking approach understands relevance in a way that applies the findings of information seeking and relevance research within the context of IR system evaluation. To extend and explore this framework for relevance, this paper reports results from four studies of users' relevance judgments that examine the role of highly, partially and not relevant judgments, and the

user's perceptions of changes in their information seeking process and information problem following an interactive IR search.

4. Research objective

The objective of this research is to begin to explore the fuzzy middle region of relevance or partial relevance. An information seeking approach to relevance suggests that a relationship exists between users' relevance judgments and their information seeking process. Therefore, a specific objective of this research is to examine the relationship between users' relevance judgments (in particular judgments of partial relevance), and changes in users' information problem and information seeking processes. A specific goal of the research presented in this paper was to examine if:

Partially relevant items selected by initial users are related to: (i) users' level of knowledge about the problem underlying the search; (ii) changes in users' information problem during or after the search; and (iii) changes in relevance criteria employed.

Another objective of this research was to investigate end-users criteria for relevant, partially relevant and not relevant items retrieved. Concentration was on *initial users*, i.e., the users engaged in the *first* search related to their given information problem.

5. Research design

5.1. Data collection

The data analyzed in this research was taken from four studies of user behavior during interactive online IR (Table 1).

As mentioned, in all four studies users were involved in their initial search. Three studies involved graduate students at the University of North Texas, each doing their own online searching ("end-users"). The second study incorporated data from 18 users (faculty and graduate students) at Rutgers University that involved a mediated online search, i.e., the users were present while a professional searcher did the interviewing and searching. A fourth study examined 11 end-users' criteria for relevant, partially relevant and not relevant judgments of items retrieved. In the four studies, the users' information problems were more complex than searching for a specific document or fact, and involved faculty, dissertation, thesis or student research. Encouraging findings from the first study led researchers to

analyze data from the second study, and further conduct a third and fourth one. Thus, data from 55 initial search users were analyzed, involving both mediated and non-mediated searching. Details of the four separate studies are listed below.

5.2. Overall design

The first three studies followed the same general design, as described in Saracevic and Su (1990), thus only a brief description is provided here. This study involved real users who were searching for information to resolve their real information problems. Online searching was done on the DIALOG Information Services. Data collection included:

- (a) Videotaping the interaction between users and searchers in Study B, and end-user searching in Study A — users in studies C and D were not videotaped;
- (b) Capture of the search logs in all four studies;
- (c) Users' judgment of retrieved items on a three-point scale — relevant/partially relevant/not relevant in all four studies.
- (d) End-users' criteria for retrieved items judged relevant, partially relevant and not relevant.

In the first three studies users completed questionnaires, at the end of the search interaction, first developed during Study B by Saracevic and Su (1990) and subsequently modified for end-user studies A and C. Items in the questionnaire were rated by respondents on a 5-point scale, where a rating of “1”=low or “5”=high was related to the users' perceptions of the degree of their *knowledge of the problem* for which the search was conducted.

- (a) Users' assessment of degree of *focus (on target) of the retrieved information* in relation to their problem at hand.
- (b) Users' perception of the degree of change in their *personal or internal knowledge* about the domain or problem-at-hand due to search interaction and/or the feedback process during the ongoing search.
- (c) Users' assessment of any change that occurred in their *criteria for relevance judgments* of items retrieved due to interaction and/or the feedback process during the ongoing search.

(d) Users' assessment of the degree of change in their own *problem definition* due to interaction and/or the feedback process during the ongoing search.

(e) Users' assessment of the degree of change in their *knowledge of the problem* due to interaction and/or the feedback process during the ongoing search.

In addition, Study B involved professional searchers as mediators, questions were also asked of searchers:

(a) Searchers' perception of the degree of change in the original *users' question* due to interaction with the searcher and the feedback process during the ongoing search.

(b) Searchers' perception of the degree of change in the users' *criteria for relevance judgments* of items retrieved due to interaction and/or the feedback process during the ongoing search.

Table 2 provides the basic data on the items retrieved for each of the first three studies.

Table 2. Basic data from users' relevance judgments in studies A, B and C

Variable	Mean No.items per search	Std.deviation	Minimum itemsper search	Maximum itemsper search	Total(%)
Study A: (13 end-users)					
No. partially relevant items	4.9	4.1	0	11	64 (13.5%)
No. relevant items	10.7	10.3	1	38	140 (29.5%)
No. not relevant items	20.7	22.7	0	58	270 (57%)
Total	36.4	27.1			474 (100%)
Precision	56.8%				
Study B: (18 users; 4 mediating search intermediaries)					
No. partially relevant items	50.3	47.8	8	230	906 (29%)
No. relevant items	50.9	66.7	9	290	917 (29.4%)
No. not relevant items	72.0	47.3	12	164	1297 (41.6%)
Total					3120 (100%)
Precision	55%				
Study C: (13 end-users)					
No. partially relevant items	9.9	11	0	37	129 (17.8%)
No. relevant items	17.9	26.2	0	99	233 (32.2%)
No. not relevant items	27.7	40.8	4	133	361 (49.9%)
Total	55.6				723 (100%)
Precision	56%				

Table 3 provides the basic data for Study D.

Table 3. Basic data for study D

Variable	Mean no. itemsper search	Minimum itemsper search	Maximum itemsper search	Total(%)
Study D: (11 end-users)				
No. of partially relevant items	15.8	1	91	174 (28.6%)
No. of relevant items	16.9	2	42	186 (30.5%)
No. of not relevant items	22.6	0	98	249 (40.9%)
Total	55.4	6	231	609 (100%)
Precision	59%			

5.3. Data analysis

The data from each of the first three studies was analyzed separately. Correlation analysis (Williams, 1992) was conducted for selected variables related to: changes in users' relevance criteria; changes in a user's personal knowledge; changes in users' problem definition, and changes in a user's specific knowledge of the problem-at-hand. For Study B, additional correlations included searcher perception of changes in users' questions and relevance criteria. These variables were correlated with the number of items judged partially relevant by users and the number of items judged relevant. Due to space considerations only statistically significant correlations are reported in the results section below. Further analyses and correlations with other variables has been or will be reported elsewhere.

6. Results

6.1. Study A: 13 end-users

The first study was conducted during Spring Semester 1994 at the University of North Texas to explore the patterns of end-user searching behavior, including the use of search terms and strategies. All 13 end-users were graduate students conducting an initial online search on their particular information problem. The most interesting and surprising finding from the analysis was the positive correlation between the number of partially relevant items retrieved and many other variables. Due to the intriguing nature of this finding, the researchers decided to examine data from a previous study of mediated online searching by Saracevic and Su (1990) to further test the finding related to partially relevant items in Study B.

6.2. Study B: mediated search study — 18 users

The mediated searching study was conducted by Saracevic and Su (1990) at Rutgers University during 1989 and 1990. Forty users and four search intermediaries were included in this study. Twenty-two users had previous searches on their topic and eighteen users (45%) were conducting initial mediated searches. To correspond with Study A, only the data from the 18 initial search users was included in the analysis. Significant correlation was found between the number of partially relevant items retrieved. Subsequently, the researchers further tested this finding with an additional data set from a group of end-users in Study C.

6.3. Study C: 13 end-users

The second set of 13 end-users was conducted during fall semester 1996 at the University of North Texas. All 13 end-users were graduate students conducting their initial online search on their particular information problem also using the DIALOG Information Services as described. Again a result of the data analysis were significant correlations related to both the number of partially relevant items retrieved and the number relevant items retrieved.

In the first three studies significant correlations were found between questionnaire variables and the number of partially relevant and relevant items retrieved. The significant results for each of the three studies are displayed in two separate tables below. Table 4 shows the significant correlations related to the number of items judged partially relevant.

Table 4. Significant correlations from studies A, B, and C related to the number of items judged *partially relevant*

Variable	P-Value	Statistically significant ($P < 0.05$)
Study A: (13 end-users)		
Change in end-user relevance criteria	0.000	yes
Change in end-user personal knowledge	0.003	yes
Study B: (18 users; 4 mediating search intermediaries)		
Change in user relevance criteria	0.030	yes
Searcher's perception that user changed question	0.030	yes
Searcher's perception that user changed relevance criteria	0.038	yes
Study C: (13 end-users)		
Change in end-user problem definition	0.024	yes
End-user specific knowledge of the problem-at-hand	-0.34	yes

Table 5 shows the significant correlations related to the number of items judged relevant

Table 5. Significant correlations related to the number of items judged *relevant*

Variables	P-value	Statistically significant ($P < 0.05$)
User Study B: (18 users; 4 mediating search intermediaries)		
User knowledge of the problem-at-hand	0.046	yes
User assessment of focus in retrieved items	0.042	yes

In Study A, involving 13 end-users, the number of items end-users' judged partially relevant correlated positively with the end-user's perception that changes had occurred due to the search interaction and/or feedback process during the ongoing search. This included changes in end-user personal knowledge of their criteria for relevance judgments. In other words, as the number of partially relevant items retrieved increased, the degree of change in

an end-user's personal knowledge about the problem-at-hand and the initial criteria they used for assessing the relevance of retrieved items changed. These findings were further confirmed and extended in the findings from the mediated searches in Study B.

In Study B, involving 18 mediated searches, the number of items judged partially relevant was positively correlated with a change in the user's relevance criteria, as assessed by users, the same as Study A. The higher the number of partially relevant items, the greater the change in the criteria used to select relevant items. This was further confirmed by findings in relation to the perceptions of the search intermediaries. The search intermediaries perceived a greater change in both the user's criteria for relevance judgments and a change in the user's question with a higher number of partially relevant items retrieved. This finding is of interest because the search intermediaries did not have access to the user's post search relevance judgments or statistics. The correlation suggests that search intermediaries perceived changes occurring in relation to the users relevance criteria. In this mediated situation, the users did not report the same change in their information problem during the search interaction. However, the search intermediaries did perceive a change in information problem of many users during the discussion and search interaction.

In Study C, involving 13 end-users, the number of items judged partially relevant was positively correlated with end-users' assessment of a change in their own information problem definition. A greater number of partially relevant items selected was a reflection of a greater change in an end-user's understanding of their information problem. Interestingly, users' specific knowledge about the problem-at-hand was negatively correlated with partially relevant items. The less they knew about the problem they searched for, the more items they judged as partially relevant. However, in this study the users did not perceive a change in their relevance criteria. This finding from Study C imply that end-users with more well developed information problems should select more relevant items. Table 5 shows such a relationship.

In Study B, the users with a greater specific knowledge of the problem-at-hand, and higher assessment that retrieved items were focused on their information problem, selected more relevant items. Higher knowledge about the problem and greater focus in retrievals resulted in a higher number of items judged relevant. This finding implies that the more users know about a problem the better they can formulate a better question and subsequent search that results in focused retrievals. More knowledge of the problem may result in better retrieval — which is not surprising, but nice to confirm. However, this relationship was not found in Studies A and C.

Overall, the three studies showed some mixed results. In general:

(a) For initial users, partially relevant items provided new information that often changed their understanding of their information problem and the criteria used to make relevance judgments.

Intermediary searchers independently confirmed such changes.

(b) Partially relevant items also provided information related to changes in users' problem definition.

(c) The less users knew about the problem at hand, the more items they assessed as partially relevant, and the more they knew the more items they assessed as relevant.

(d) Not surprisingly, the more focused the retrieved items were on the problem at hand, the more items were judged as relevant.

(e) Partially relevant items are associated with changes. However, these judgments are fuzzy in nature, and so may be the changes.

6.4. Study D: 11 end-users

To explore end-users' distinctions between relevant and partially relevant items further, Study D was specifically conducted to examine the relevance, partial relevance and negative relevance criteria used by 11 end-users when judging retrieved items. As shown in Table 3, 11 end-users searching on their own information problem retrieved a total of 609 items — 186 (30.5%) relevant, 174 (28.6%) partially relevant and 249 (40.9%) not relevant. Fig. 3 lists the 11 end-users' criteria for items judged relevant, partially relevant and not relevant.

Criteria for Relevant Items	Criteria for Partially Relevant Items	Criteria for Not Relevant Items
It excited me It included my search terms It was specific to my query It answered my question All the concepts I was searching for were included It was an authoritative source My personal image of what I perceived to be a relevant document	Not on the money The chronology (timeliness) Not enough information Dealt only partially with the subject Contained multiple concepts On target, but too technical Lists good resources Lists good references Identifies a different, but related concept (new terms) Information included too brief Future implications (related to current problem) On target, but too narrow Could be helpful (but I don't know yet) Could be other opportunities Duplicate information	Not useful Wrong meaning of search terms Wrong language Don't understand context Duplicate

Fig. 3. End-users criteria for retrieved items judged relevant, partially relevant and not relevant

Relevant items were generally those items that answered the user's question. End-users' criteria for relevant items included; a sense of excitement; the inclusion of the end-user's search terms or concepts; specificity to their query or question or personal image of a relevant item; and the source authority. These relevance criteria conform to those outlined previously and summarized in Schamber (1994).

Not relevant items did not answer the user's question. However, end-users' also identified criteria for partially relevant items.

6.4.1. Partial relevance criteria

Items were judged partially relevant by end-users because they:

- (1) May answer the information problem, but end-users were not able to determine this using the information provided by the IR system.
- (2) Provided insufficient information to determine high relevance to the information problem.
- (3) Not as specific enough or included additional concepts than the items judged relevant.

(4) Provided interesting or new material, but did not directly answer the user's question.

Partially relevant items often included new, but related concepts to the end-user's original concepts and were “helpful” or “related”, or provided “opportunities” to explore new dimensions of the information problem. Also, partially relevant items were those with “good resources” or “good references”, although they were not highly relevant.

Our findings from four separate studies suggest that both partially relevant and highly items may have a potentially important role to play in the evolution of users' information problems. This suggestion is discussed further in Section 7 of the paper.

7. Discussion

Findings from the four studies extend our understanding of the characteristics of users' partial and high relevance judgments. For users conducting their initial search with a low knowledge of their search topic, partially relevant judgments may relate to changes in many aspects of the user's information seeking process and information problem, including level of knowledge, relevance criteria, and other aspects related to users' knowledge and searching. In other words, the retrieval of partially relevant items may have a crucial role in providing users with new information and directions that may lead them through further stages of their information seeking process toward a possible resolution of their information problem. Specifically, partially relevant items may be important for initial search users as entities facilitating the necessary development of a greater understanding of their information problem.

Items considered highly relevant may either be familiar to this type of user or contain information that conforms to the user's current understanding of their information problem or answers a user's current problem. However, highly relevant items may not change the user's cognitive or information space in relation to their information problem. Initial search users who were more advanced in their understanding of their information topic may tend to select more relevant items and experience less change due to the search interaction. Bates (1996) also found that humanities scholars familiar with their topic would often identify retrieved items that were “content relevant” (relevant to the query), indicated these items had little “utility relevance” — as users had seen the items previously. She suggests that users with a greater knowledge of their topic are looking for novel or unfamiliar items.

The findings of our studies suggest that IR and relevance researchers should begin to question the assumption that highly and partially relevant items have the same utility for users. Information seeking research shows that at the beginning stages of an information seeking process (what we considered as initial users), a user's information problem is usually fairly ill-defined and subject to change (Kuhlthau, 1993). If the goal of the user is to resolve their evolving information problem (possibly within a successive search framework) then the assumption that only the most highly relevant items are useful to all users may be questionable. Of course the user's perception of a highly or partially relevant items may change over successive searches. However, highly relevant items may only confirm what the user thinks they need to know or provide the user with what they already know, as these items equate strongly to the current state of the user's information problem. Highly relevant items may not relate to a shift in a user's information problem toward resolution, but may reinforce the current state of the user's information problem and knowledge state. Items retrieved that are partially relevant may be related to shifts in the user's thinking about their information problem by providing new information that may lead the user in new directions toward the resolution of their information problem. We could also suggest that users situation at the initial or exploratory stages of an information problem or research may cause them to judge more documents as partially relevant as they do not have a high topic knowledge. This issue requires further analysis.

Ingwersen (1996) suggests that an overlap exists between elements considered part of a user's cognitive space — including current cognitive state, problem state or uncertainty, and information need. He suggests that a problem state or uncertainty may form part of the user's current cognitive state because the knowledge gained (information judged relevant) has been absorbed into the user's current cognitive state. The findings reported in this paper suggest that initial search low topic knowledge users may experience a shift in their information problem during a search interaction. This may be resolved into the current cognitive state. If more partially relevant items are identified, the problem state or uncertainty is perpetuated, some absorption takes place into the current cognitive state, and successive searching may take place (Spink, 1996). However, the information need is yet to be resolved.

We suggest that an integral relationship exists between a user's relevance judgments and their movement through their information seeking process related to a particular information problem. Traditionally, relevance has been conceptualized as a relationship between an item and a user's information problem (Saracevic, 1996b). However, such a relationship is not a stable and static entity, and therefore relevance can also be conceptualized as both a “relationship” and “effect” on a users information problem and information seeking process.

The findings from this study support this conceptualization. Although recall and precision measures have been used extensively and are based on dichotomous relevant/not relevant judgments, they assume that highly relevant items are the only useful items for all users as they resolve their information problems. That assumption, however, is also challenged by the findings reported in this article. In future studies, partially relevant items should not be collapsed into relevant items, but should be analyzed separately.

Based on the findings from our research, we propose a three-dimensional spatial model of relevance, level, region and time that provides an integrated view of a users' relevance judgments based on their level, region and time.

8. Three-dimensional spatial model of relevance level, region and time

We propose that each user relevance judgement can be plotted within three dimensions: manifestations of inferential relationships (levels of relevance), relevance region, and time (Fig. 4).

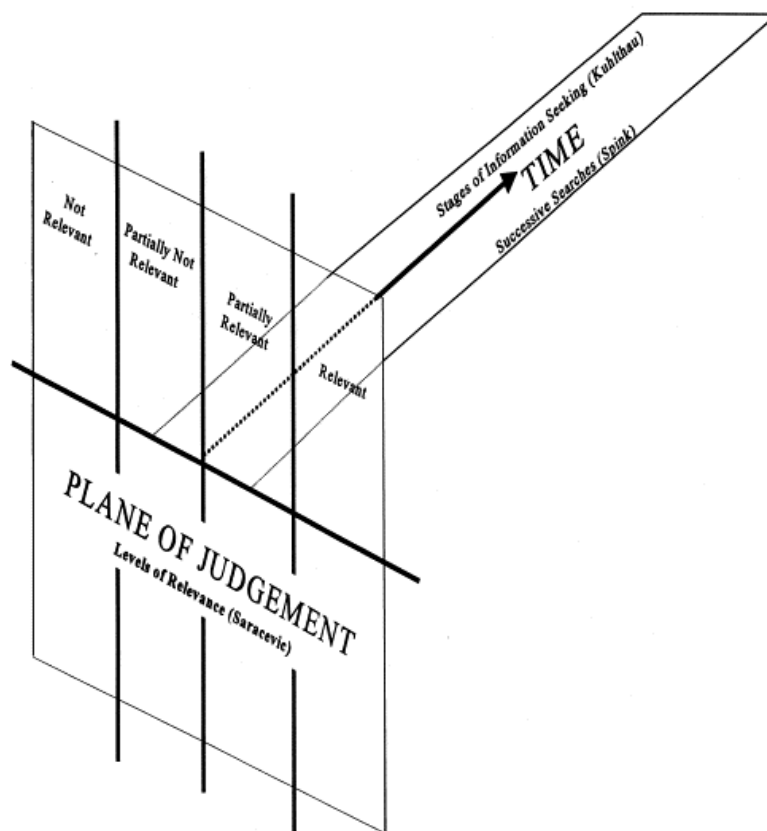


Fig. 4. Three dimensional model of relevance level, region and time

While a ratio measurement associated with time is easily plotted from the initiation of a user's information need, the measures associated with the attributes of relevance level and region are more elusive to operationalize.

8.1. Level of relevance

The first dimension of this spatial model is developed from Saracevic (1996b), who suggested that “as a cognitive notion relevance involves an interactive, dynamic establishment of a relation by inference, with intentions toward a context”. We propose that there also exists a cognitive notion of negativity in that same realm of relevance. Hence, a cognitive notion of relevance also involves an interactive, dynamic establishment of *no relation* or a *partial relation* by inference, based on interactions toward a context. If this negative aspect is added to the manifestations of relevance as defined by Saracevic, we find an expansion of the cognitive plane of judgement to allows further manifestations as follows:

Systematic/Algorithmic Inference: the relation or non-relation between query and information objects;

Topical/Subject Inference: the relation or non-relation between the subject/topic expressed and the retrieved text(s);

Cognitive Inference/Pertinence: the relation or non-relation between a user's state of knowledge and the informativeness of the retrieved text(s);

Utility/Situational Inference: the relation or non-relation between the problem at hand and the retrieved text(s);

Motivational/Affective Inference: the relation or non-relation between the user's goals/intents and the retrieved text(s).

The manifestations of inferential relationships become levels of relevance is shown as the first dimension of a plane of relevance judgement. These levels of relevance imply no hierarchy or measure of strength, they merely exist as possible relational inferences at a specific point in time. The ability to plot these cognitive relations by inference is determined by the second dimension in the plane of judgement, the user's region of relevance attributed to these relations or non-relations. This second dimension also contains positive and negative aspects which can be labeled and depicted graphically.

8.2. Regions of relevance

In the model the second dimension we chose to depict relevance judgments within four regions: (1) highly relevant, (2) partially relevant, (3) partially non relevant, and (4) not relevant. The distinction between the *partially relevant* quadrant and the *partially not relevant* quadrant in Fig. 3 can be operationally defined as follows:

Partially relevant represents a judgement that confirms that some relation by inference exists as a manifestation of relevance, but the relation is weaker than a relevant relation at the time the judgement is made.

Partially not relevant represents a judgement that some non-relation exists by inference as a manifestation of relevance, but the inference is not strong enough to totally reject the relation as not relevant at the time the judgement is made.

For a finer grain analysis, many more regions of relevance of relevance can be delineated as the granularity of relevance regions is sharpened. An overlay of the two dimensions (level and region) of a relevance judgment are represented on a plane of judgement. A user also makes a relevance decision at a specific point in time, and a graphical representation of such decisions related to retrieved texts can also be plotted.

8.3. Time dimension of relevance

We may identify his/her relational or non-relational inferences along with a decision indicating the region of relevance at the time a document is being judged. Although this plane of judgement exists at a specific moment in time as inference melds with context, research has yielded the dynamic aspect of relevance judging related to a user's knowledge state, problem state and cognitive state (Ingwersen, 1996) as time advances. Adding this third dimension to the plane of judgement yields a spatial model of dynamic information seeking and judging. This time dimension can be measured and plotted in formats such as information seeking stage (Kuhlthau, 1993) and successive searches (Spink, 1996). The user's information seeking stage could be plotted within one of Kuhlthau's six stages of a user's information search process: initiation, selection, exploration, formulation, collection, and presentation. Within each of those stages successive searches over time related to the same evolving information problem may take place, as well as the aggregate number of searches that make up the total search process over time to resolve an information need.

In Fig. 4, a user's partial relevance judgment on a document may be plotted on a the first plane of judgment. At a later time and different plane of judgment the user judges the same document as relevant. The ability to plot these relational inferences in a spatial model such as this, could allow researchers as well as system designers to more closely identify how, when and why a user makes particular inferences which lead to specific relevance judgements.

The implication of this spatial model is the potential ability to isolate a user's plane of judgement at a particular point in time based on an assessment of levels of relevance, region of relevance, and time (information seeking stage and successive searches). This could lead to major implications for system design and design criteria. When each user's relevance judgment through successive searches is plotted on these three dimensions, a complex picture of the changes in user relevance judgments could be analyzed. This may be a useful and very practical method for users and researchers alike, particularly if the plotting of judgments can be automated and displayed to the user in a visual way. This model could also be used to map the interests of relevance researchers, by locating each study on the three-dimensional model. We also suggest that most relevance research would probably cluster in the upper right relevant quadrant of the plane of judgment. Few studies have compared the characteristics of different relevance regions, including the middle region of partial relevance, or the time dimension of users relevance judgments, particularly in relation to uncertainty during a user's information seeking process, is an area for further discovery. There lies the future of relevance research.

The next section of the article discusses the implications of the findings for IR system design, searching practice, and relevance research.

9. Implications

9.1. IR system design

Findings from this research have implications for the design and development of IR systems, as IR ranking and relevance feedback systems are important tools for users. Currently, ranked retrieval systems (including Internet Web browsers) and automatic relevance feedback techniques display retrieved items in relation to a user's queries in order of probabilistically calculated relevance. These IR systems do not attempt to understand the user's level of topic knowledge or stage of information seeking. Nor do IR systems classify

retrieved items into categories of relevance (e.g., highly relevant, partially relevant or not relevant) beyond ordering retrieved items with the most highly relevant items first. This ordering is based on the assumption that the most highly relevant items are the only useful items for all users. Users want and need highly relevant items, but the fuzzy region of partial relevance is also important. The findings from the four studies reported in this paper suggest that partially relevant items are also potentially important, particularly for an initial search low topic knowledge user. These items may be more useful to such a user in the long term than highly relevant items, because they produce change in many important dimensions related to the problem-at-hand.

Automatic relevance feedback techniques are another case in point. Studies have shown that users often retrieve large numbers of items in relation to a query (Spink, 1997). Automatic relevance feedback systems are mechanisms to help identify the most highly relevant items from a potential retrieval list (Spink and Losee, 1996). However, for the initial search low topic knowledge user, an automatic relevance feedback system will provide a small number of what it thinks are highly relevant items based on the user's query. The user is then asked to select those they consider highly relevant and the IR system brings back another list of what it considers highly relevant items that match those selected.

Unfortunately, the partially relevant items are not an element of this process. IR systems without automatic relevance feedback allow the user to scan the retrieved list in total and see potentially "highly", partially and not relevant items. Automatic relevance feedback systems potentially restrict that option for such users. This approach is based on the assumption that items not highly content relevant have little utility for users.

IR system designers should consider interactive IR as a process of assisting users through what can often be a long and complex information seeking process, involving many searches using different search terms and strategies. We need to reconsider the one shot most "highly relevant items" approach to IR system design, as partially relevant items are also important for some classes of users. Specifically, IR systems need to provide both highly and partially relevant items to initial-search-low-topic-knowledge users. User's comparison and distinctions between highly and partially relevant items are an essential process in the determination of levels of relevance (Saracevic, 1996b). Researchers have also found inter-document dependencies (Tiarniyu and Ajiferuke, 1988) and order effects in relation to relevance judgments (Eisenberg and Barry, 1988; Regazzi, 1988). For example, a user may identify a highly relevant item and distinguish a "highly" relevant item because of its relationship to certain partially relevant items.

9.2. Searching practice

Search intermediaries need to consider the user in the context of their information seeking stage and take this information into account when conducting a search. Some recent studies show that search intermediaries do not explicitly elicit information from users regarding their information seeking stage (Kuhlthau et al., 1992; Spink et al., 1998). Search intermediaries need to identify if the user is an initial search low topic knowledge user or a user in a later stage of their information seeking on their particular topic. End-users also need to be aware of the implications of their information seeking stage for the nature of their searching practice. They also need to be aware that partially relevant items are as potentially important for initial-search-low-topic-knowledge users as highly relevant items — in fact they could be more important for eliciting further information to enhance successive searches.

9.3. Relevance research

The research begins the process of characterizing the different regions of relevance, by examining the difference between high and partial relevance judgments. This research also highlights the need to further explore the relationship between interactive IR and information seeking research, and relevance research. An information seeking approach to understanding interactive IR and relevance is emerging within the context of contemporary, user-oriented theory of information seeking. The approach falls within the alternative view articulated by Dervin and Nilan (1986) who posit information as a subjective phenomenon constructed by human beings within a sense-making process. Within this view, meaning is continuously constructed by the individual through internal cognitive processes. As (Schamber et al., 1990) suggest within the alternative paradigm, “because meaning is seen as constantly constructed by the individual, appropriate models for information behavior are complex, contextual and dynamic (p. 769)”. An approach within the context of appropriate models of dynamic information seeking behavior may provide a basis for new insights into the meaning of relevance and IR evaluation.

From this viewpoint, relevance is considered in relation to a users information seeking process. This approach involves an interactional conceptualization of relevance, and an exploration of an operationalization of relevance based on the dynamic context of the human information seeking process. This approach suggests that relevance may be measurable as to its effect on the movement of a user through their information seeking process. Within this framework, relevance has two dynamic processes: judgment and effect. Relevance at its

most basic level can be understood as an “effect”. Relevance must be seen in relation to something else — as relevance is an abstract concept that does not exist independently of its role in judgment and effect. Within this approach, relevance (and relevance judgment) is considered a fundamental property of an IR interaction and the feedback process between user and source, through which a user constructs information (Spink, 1997). Therefore, relevance may be understood as an impetus to movement or an effect on or within the movement in a user's IR interaction and their information seeking process.

Much of the relevance research has focused on relevance judgments and the variables that effect these judgments. More recent research has examined the criteria that users employ when making relevance judgments (Schamber et al., 1990), but relevance criteria do not suggest the function or effect of relevance on the user or his or her information seeking process. Relevance itself may not be a tangible entity, and a relevance judgment may be dependent on the dynamic, situational information need (and the information that the IR system provides). The situational, dynamic approach posits that relevance judgments are made within individual contexts, but what happens to the user or their information seeking process because of these judgments is not known. However, relevance judgments cannot be removed from the context of the information seeking process and the information problem.

Within the information seeking process the user's relevance criteria interact with the document representation or the document and with the IR system. Relevance judgments, and therefore relevance itself, are dependent on how the IR system presents information to the user or the characteristics of the information itself. The way the IR system and the user interact, and the information the IR system presents to the user will influence the user's current relevance definition and judgments. A user finding what they determine to be too little or too much information may change topic formulation and may change their criteria for making relevance judgments. For example, users may initially decide they want only current information within the last five years. However, if the IR system presents them with few citations that meet this criterion, they may broaden the criteria of currency to 10 or 15 years, or drop the criterion entirely.

To better understand relevance and help users make the best relevance judgments for their situation, it is important that the study of end-user relevance be within the entire information seeking context rather than strictly within the IR evaluation context. Users' relevance judgments do not cease after they leave their interaction with the IR system, but continue as they seek, obtain, read, use, and cite information. The user's information problem will be molded by the information he or she does or does not find, by how he or she defines relevance and how that definition is applied in making relevance judgments. Users able to:

- (1) Clearly define their information seeking stage or state,
- (2) Understand the characteristics of the information seeking stages,
- (3) Define how their interaction with an IR system or information affects their relevance judgments and,
- (4) Understand how their relevance judgments affect their information problem, may be better to identify the information they need to resolve their problem. They may also be better able to find the focus for their information problem that helps them most easily resolve their information needs.

When a user does or does not select a particular piece of information, he or she has made a decision that influences the rest of the information process. That decision will take the user down a particular path to resolving or not resolving their information problem. Users may backtrack and traverse the same path again, and it appears that some users do just that (Spink, 1996). However, when the path is repeated the user has often changed his or her relevance definition and information problem. Users able to better judge where they are in an information seeking process and how the IR system information has affected their relevance judgments and information problem, may make better judgments that lead them to the best information in the most effective manner.

The emerging information seeking approach understands relevance as an effect that causes shifts or changes the user's information seeking process. Users themselves may be able to measure a shift or move in their information seeking process and thus better understand this process and the characteristics of the information that can help them through an information seeking process. Such an approach looks beyond user satisfaction to a more multidimensional dynamic approach to evaluation of the interaction of untested relevance elements, including the user and the IR system. The next challenge for evaluation testing and experimentation research in IR is to develop evaluation approaches that incorporate models of human information behavior processes within this framework.

10. Conclusion and further research

This paper proposes a regions of relevance view within an information seeking approach to relevance research. Research is needed to further investigate the characteristics of the different regions of relevance, and the role of partially relevant items in relation to the role of

highly relevant items for users. How are partially relevant items used as opposed to how relevant item? Research is also needed to further identify users' criteria and attributes for partially relevant judgments and how these may differ from highly or non relevant judgments. Studies are also needed to test the ordering of retrieved items for initial search low topic knowledge users and to track the impact of partially relevant retrievals on the user's progression through the stages of their information seeking process. There is also a need for research that identifies what relevance criteria relate to the various levels of relevance.

On the IR systems design side: can we design automatic relevance feedback techniques that retrieve only partially relevant items in relation to a query? What if a user wants only items that were "fringe" to their topic or not precisely topically relevant to lead them in new directions?

A three-dimensional spatial model of relevance level, region and time is also presented to provide a practical and integrated approach to further investigate and model users' relevance judgments. Further research is currently being conducted to gather and plot user relevance judgment data on the three-dimensional model for analysis.

Acknowledgements

This study was funded by a University of North Texas Research Initiation Grant 1993–94. The authors also acknowledge the valuable comments by Tefko Saracevic of Rutgers University and the anonymous reviewers to the development of this paper.

References

- Barnydt, G. C. (1964). A comparison of relevance assessments by three types of evaluators. *Proceedings of the American Documentation Institute*, October 5–8, 1964, pp. 383–385. American Documentation Institute, Washington, DC
- Barry, C., 1994. User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science* **45** 3, pp. 149–159
- Bateman, J. (1997). Changes in users' relevance criteria: An information seeking study. Unpublished dissertation proposal. School of Library and Information Sciences, University of North Texas
- Bates, M. J., 1996. Document familiarity, relevance and Bradford's Law: The Getty Online Searching Project Report No. 5. *Information Processing and Management* **32** 6, pp. 697–707
- Belkin, N. J., Oddy, R. N. and Brooks, H. M., 1982. ASK for information retrieval: Pt 1. Background and theory. *Journal of Documentation* **38**, pp. 61–71
- Belkin, N. J., Oddy, R. N. and Brooks, H. M., 1982. ASK for information retrieval: Part II. Results of a design study. *Journal of Documentation* **38**, pp. 145–164
- Borlund, P. and Ingwersen, P., 1997. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation* **53** 3, pp. 225–250
- Cooper, W. S., 1973. On selecting a measure of retrieval effectiveness (Part 1). *Journal of the American Society for Information Science* **24** 2, pp. 87–100
- Cooper, W. S., 1973. On selecting a measure of retrieval effectiveness (Part 2). *Journal of the American Society for Information Science* **24** 6, pp. 413–424
- Dervin, B. and Nilan, M., 1986. Information needs and uses. *Annual Review of Information Science and Technology* **21**, pp. 3–33
- Doyle, L. B. (1963). Is relevance an adequate criterion for retrieval system evaluation? In *Automation and scientific communication, short papers*, ed. H. P. Luhn, Pt. 2, pp. 199–200. American Documentation Institute, Washington, DC

- Eisenberg, M. B., 1988. Measuring relevance judgments . *Information Processing and Management* **24** 4, pp. 373–389
- Eisenberg, M. B. and Barry, C. L., 1988. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science* **39** 5, pp. 293–300
- Eisenberg, M. B., and Hu, X. (1987). Dichotomous relevance judgments and the evaluation of information systems. *Proceedings of the Annual Meeting of the American Society for Information Science*, Vol. 24, pp. 66–70
- Ellis, D., 1984. Theory and explanation in information retrieval research. *Journal of Information Science* **8**, pp. 25–38
- Ellis, D., 1989. A behavioral approach to information retrieval system design. *Journal of Documentation* **45** 3, pp. 171–212
- Ellis, D., 1996. The dilemma of measurement in information retrieval research. *Journal of the American Society for Information Science* **47** 1, pp. 23–36
- Gull, C. D., 1956. Seven years work in the organization of materials in a special library. *American Documentation* **7**, pp. 320–329
- Harman, D. (1992). Relevance feedback revisited. *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, June 21–24, Copenhagen, Denmark, pp. 1–10
- Harman, D., 1993. The first text retrieval conference (TREC-1), Rockville, MD, USA, 4–6 November, 1992. *Information Processing and Management* **29** 4, pp. 411–414
- Harter, S., 1992. Psychological relevance and information science. *Journal of the American Society for Information Science* **43** 9, pp. 602–615
- Harter, S. P., 1996. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science* **47** 1, pp. 37–49
- Ingwersen, P., 1996. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation* **52** 1, pp. 3–50

Janes, J. W. and McKinney, R., 1992. Relevance judgments of actual users and secondary judges: A comparative study. *Library Quarterly* **62** 2, pp. 150–168

Kuhlthau, C. C. (1993). *Seeking meaning: A process approach to library and information science*. Ablex Publishing, Norwood, NJ

Kuhlthau, C. C, Spink, A., and Cool, C. (1992). Exploration of stages in the information search process in online information retrieval: Communication between users and intermediaries. *Proceedings of the 55th Annual Meeting of the American Society for Information Science*, Vol. 29. October 1992. Pittsburgh, PA, pp. 67–71

Meadow, C. T., 1985. Relevance? . *Journal of the American Society for Information Science* **36**, pp. 354–355

Newby, G. B. (1992). An investigation of the role of navigation for information retrieval. *Proceedings of the 55th Annual Meeting of the American Society for Information Science*, Vol. 29, pp. 20–25

O'Conner, J., 1969. Some independent agreements and resolved disagreements about answer-providing documents. *American Documentation* **20**, pp. 311–319

Pao, M. L., 1993. Term and citation retrieval: A field study. *Information Processing and Management* **29** 1, pp. 95–112

Park, T. K., 1993. The nature of relevance in information retrieval: An empirical study. *Library Quarterly* **63** 3, pp. 318–351

Rees, A. M., 1967. Evaluation of information systems and services. *Annual Review of Information Science and Technology* **2**, pp. 63–86

Rees, A. M., and Schultz, D. G., (1967). *A field experiment approach to the study of relevance assessments in relation to document searching*, 2 vols. Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University, Cleveland, OH

Regazzi, J. J., 1988. Performance measures for information retrieval systems: An experimental approach. *Journal of the American Society for Information Science* **39** 4, pp. 235–251

- Robertson, S. E. and Hancock-Beaulieu, M., 1992. On the evaluation of IR systems. *Information Processing and Management* **28** 4, pp. 457–466
- Robins, D. (1997). Shifts in focus in information retrieval interaction. *Proceedings of the 60th Annual Meeting of the American Society for Information Science*, November 1997, Washington, DC (pp.)
- Rorvig, M. E., 1988. Psychometric measurement and information retrieval. *Annual Review of Information Science and Technology* **23**, pp. 157–189
- Sandore, B., 1990. Online searching: What measures satisfaction?. *Library and Information Science Research* **12** 1, pp. 33–54
- Saracevic, T., 1975. Relevance: A review of and framework for the thinking on the notion in information science. *Journal of the American Society for Information Science* **26** 6, pp. 321–343
- Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *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). Interactive models in information retrieval (IR): A review and proposal. *Proceedings of the 59th Annual Meeting of the American Society for Information Science*, Vol. 33, pp. 3–9
- Saracevic, T. (1996b). Relevance reconsidered:'96. *Proceedings of COLIS 2: Second International Conference on Conceptions of Library and Information Science: Integration in Perspective*, October 13–16, 1996. The Royal School of Librarianship, Copenhagen, Denmark, pp. 201–218
- Saracevic, T., Kantor, P. B., Chamis, A. Y. and Trivison, D., 1988. A study of information seeking and retrieving: Part I Background and methodology. *Journal of the American Society for Information Science* **39** 3, pp. 161–176
- Saracevic, T., Spink, A., and Wu, M. M. (1997). Users and intermediaries in interactive information retrieval (IR): What are they talking about? *Proceedings of UM97: The 6th International Conference on User Modeling*, ed. A. Jameson, C. Paris and C. Tasso, June 2–5, 1997, Sardinia — Italy. International Center for Mechanical Sciences. Courses and Lectures, Vol. 383, pp. 43–54. Springer Wein, New York

Saracevic, T., and Su, L. (1990). Modeling and measuring the user-intermediary-computer interaction in online searching. *Proceedings of the 52nd Annual Meeting of the American Society for Information Science*, Vol. 26, pp. 75–80

Schamber, L. (1991). User's criteria for evaluation in multimedia information seeking and use situations. Unpublished doctoral dissertation, Syracuse University

Schamber, L., 1994. Relevance and information behavior. *Annual Review of Information Science and Technology* **29**, pp. 3–48

Schamber, L., Eisenberg, M. B. and Nilan, M. S., 1990. A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management* **26**, pp. 755–776

Smithson, S. (1990). The evaluation of information retrieval systems: A case study approach. In *Informetrics 10 — Prospects for Intelligent Retrieval*, ed. K. F. Jones, pp. 75–89. Aslib, London

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

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, pp. 603–609

Spink, A., 1997. A study of interactive feedback during mediated online searching. *Journal of the American Society for Information Science* **48** 5, pp. 382–394

Spink, A., Goodrum, A., and Robins, D. (1998). Elicitation behavior during mediated information retrieval. *Information Processing and Management*, 34(1)

Spink, A. and Losee, R. M., 1996. Feedback in information retrieval. *Annual Review of Information Science and Technology* **31**, pp. 33–78

Su, L., 1994. Evaluation measures for interactive information retrieval. *Information Processing and Management* **28** 4, pp. 503–516

Tiamiyu, M. A. and Ajiferuke, I. Y., 1988. A total relevance and document interaction effects model for the evaluation of information retrieval processes. *Information Processing and Management* **24**, pp. 391–404

Williams, F. W. (1992). *Reasoning with statistics: How to read quantitative research*.
Harcourt Brace Jovanovich, Fort Worth

Wilson, T. D., 1981. On user studies and information needs. *Journal of Documentation* **37**,
pp. 3–15

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

Xie, H. (1997). Planned and situated aspects in interactive IR: Patterns of user interactive
intentions and information seeking. *Proceedings of the 60th Annual Meeting of the American
Society for Information Science*, November 1997, pp. 101–110