Relevance Judgment: What Do Information Users Consider Beyond Topicality?

Yunjie (Calvin) Xu
Department of Information Systems, School of Computing, National University of Singapore, 3 Science Drive 2, 117543, Singapore. E-mail: xuyj@comp.nus.edu.sg

Zhiwei Chen
Media Data System Pte Ltd., 1 Sims Lane, #08-01, 387355, Singapore. E-mail: chenzw@globalsources.com

How does an information user perceive a document as relevant? The literature on relevance has identified numerous factors affecting such a judgment. Taking a cognitive approach, this study focuses on the criteria users employ in making relevance judgment beyond topicality. On the basis of Grice's theory of communication, we propose a five-factor model of relevance: topicality, novelty, reliability, understandability, and scope. Data are collected from a semicontrolled survey and analyzed by following a psychometric procedure. Topicality and novelty are found to be the two essential relevance criteria. Understandability and reliability are also found to be significant, but scope is not. The theoretical and practical implications of this study are discussed.

Introduction

The rapidly expanding Internet and other digital document depositories have generated a huge amount of textual documents. Consequently, information overload becomes a pressing issue for users of these depositories. Searching for relevant information is increasingly a hard and frustrating task. Among the huge amounts of textual documents retrieved by typical information retrieval (IR) systems nowadays, most are found irrelevant. This phenomenon has raised doubts about the effectiveness of the mechanical approach to relevance definition such as the well-known cosine score in the vector space model. It has also triggered a resurgence of interest in the concept of relevance, which is regarded as the "fundamental and central concept" in information sciences (Saracevic, 1975; Schamber, Eisenberg, & Nilan, 1990). As a result of the inadequacy of system- or algorithm-oriented perspectives on relevance, a user-oriented and subjective perspective is gaining momentum in the research community (e.g., Borlund, 2003; Cosijn & Ingwersen, 2000; Saracevic, 1970; Schamber et al., 1990). For example, Borlund (2003, p. 913) argues that "relevance is a multidimensional cognitive concept whose meaning is largely dependent on users' perceptions of information and their own information need situations." Subjective relevance concepts such as psychological relevance and situational relevance are accepted, at least theoretically, as replacements or extensions of objective and system-determined relevance. In general, relevance is now regarded as a subjective, multidimensional, dynamic, and measurable concept (Schamber et al., 1990).

If relevance is subjective, then what makes a user judge a document as relevant? Many different document attributes have been noted to affect relevance judgment, including recentness, reliability, and topicality. Such a list of document attributes can easily contain more than 20 criteria (e.g., Barry & Schamber, 1998). Though comprehensive, the long lists in the extant research suffer a number of limitations. First, having a large number of factors obscures the key factors. Second, although Barry and Schamber (1998) suggest that there is a core set of user criteria across different situations, no consensus regarding the set and the definition of key factors in the set has been reached. One factor that seems ubiquitous is topicality (e.g., Bateman, 1998; Hirsh, 1999; Schamber & Bateman, 1996; Wang & Soergel, 1998). In fact, topicality has been identified as the first or basic condition of relevance (Greisdorf, 2003). In contrast, there is no agreement on factors beyond topicality, neither in terms of what they should be nor of how important they are. Finally, in regard to methodology, past studies are almost exclusively exploratory and data driven. Naturalistic inquiry with qualitative research methods has been advocated (Park, 1994; Schamber et al., 1990) and adopted (Barry, 1994; Hirsh, 1999; Wang & Soergel, 1998) by many researchers. Exploratory studies are very useful in uncovering an unknown phenomenon (Park, 1994); however, they cannot confirm whether a certain factor so identified is statistically significant in the domain of interest. Comparatively,

Received November 23, 2004; revised April 29, 2005; accepted May 20, 2005

 • Published online 16 March 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20361

confirmatory studies adopt a positivist perspective and employ a statistical hypothesis-testing procedure, which helps further the test of the validity of identified factors and weed out the insignificant ones (Tashakkori & Teddlie, 1998). Unfortunately, almost no study of relevance judgment has adopted a confirmatory approach. Consequently, the importance of many relevance criteria is still unclear.

With a focus on users' relevance judgment, the purposes of this study are (1) to identify a set of core relevance criteria using a theory-driven approach and (2) to test the validity of these factors with a rigorous psychometric approach. The rest of the article is organized as follows: We review the literature on relevance and relevance judgment next. After that, we identify a set of core factors based on Grice's (1989) communication theory, which leads us to our research model and hypotheses. The empirical study is then discussed and the data analysis reported. Finally, we discuss the theoretical and empirical implications of our findings.

Review of Related Work

Subjective Relevance

What is relevance? For more than 50 years, information scientists have attempted to define this concept in different ways (Saracevic, 1975; Schamber, 1994). In recent years, the concept has increasingly come to be regarded as subjective in nature rather than being algorithm determined (Borlund, 2003; Cosijn & Ingwersen, 2000; Mizzaro, 1997; Saracevic, 1975; Schamber, 1994). The term *subjective relevance* is used as an umbrella term to cover the concepts of subjective topicality (or topical relevance) and situational relevance. Here, *subjective topicality* refers to the "aboutness" of a document as perceived by the user in relation to her information need; *situational relevance* refers to the usefulness, value, utility, pragmatic application or pertinence of a document to the task, or problem at hand (Cosijn & Ingwersen, 2000; Hjørland & Christensen, 2002; Mizzaro, 1997; Park, 1997; Saracevic, 1975).

Subjective topicality extends system-determined query-document match, which is known as *system relevance* (Saracevic, 1996). However, it is different from system relevance because whereas system relevance is calculated by mechanical criteria such as cosine similarity in the vector space model, topical relevance is the user's subjective judgment. If the user believes a document is about the topic area of interest, then it is topically relevant. However, relevance is not limited to topicality, as indicated by Bookstein (1979). Boyce (1982) argues that merely hitting the topic area is insufficient; users are looking for informativeness beyond topicality. Hersh's (1994) study in the medical field advocates the recognition of situational factors in defining what is relevant. In the 1990s, more researchers turned to the situational aspects of this concept (e.g., Barry, 1994; Harter, 1992; Park, 1997).

Situational relevance takes a pragmatic perspective and defines relevance as the utility of a document to the user's task or problem at hand (Borland, 2003; Saracevic, 1996). In this view, if a document contributes to problem solving, it is relevant; otherwise, it is irrelevant. Wilson (1973, p. 458) first introduced the concept of situational relevance, defined as "the actual uses and actual effects of information: how people do use information, how their views actually change or fail to change consequent on the receipt of information." Saracevic (1975, 1996) regards the utility perspective of relevance as a cost-benefit trade-off. Saracevic (1975, p. 334) highlights that "it is fine for IR systems to provide relevant information, and the true role is to provide information that has utility—information that helps to directly resolve given problems, that directly bears on given actions, and/or that directly fits into given concerns and interests." Borlund (2003, p. 922) conceptualizes situational relevance as a user-centered, empirically based, realistic, and potentially dynamic type of relevance. Therefore, the key difference between subjective topicality and situational relevance bears on the pragmatic impact of a document.

Although the fine line between situation relevance and topicality has been drawn, the term *relevance* has been used to refer to any position in the continuum from topicality to situational impact by different authors (Harter, 1992). One type of relevance that lies between is cognitive relevance (Saracevic, 1996), that is, that a document should be understandable to the user given the current stock of knowledge the user possesses (Saracevic, 1975, 1996). In this continuum, topicality is viewed as a basic requirement while situational relevance is viewed as the highest requirement because it corresponds more directly to the user's judgment in a real situation (Borlund, 2003). In this sense, situational relevance requires topicality and cognitive relevance. In this study, topicality is regarded as a document attribute rather than relevance itself; the term *relevance* refers to the portion of the relevance continuum beyond topicality; it encompasses both cognitive and situational relevance. We define it as the perceived cognitive and pragmatic impact of the *content* of a document in relation to the user's problem at hand.

Relevance Criteria

Because relevance in general is conceptualized as the user's judgment of the strength of the relationship between a document and her information need (Saracevic, 1975), the question that follows naturally is which criteria the user employs in the judgment. Schamber and associates (1990, p. 771) highlight the importance of relevance criteria studies and suggest that "an understanding of relevance criteria, or the reasons underlying relevance judgment, as observed from the user's perspective, may contribute to a more complete and useful understanding of the dimensions of relevance." As early as in the 1960s, researchers attempted to identify the criteria for relevance judgment. For example, Rees and Schulz (1967) suggested 40 variables that would affect relevance judgment. Cuadra and Katter (1967) found that relevance judgment is affected by 38 factors, including style and level of difficulty of the document. Since 1990,

DOI: 10.1002/asi

many empirical studies have been carried out to identify relevance criteria or factors in different problem domains (e.g., Bateman, 1998; Hirsh, 1999; Schamber & Bateman, 1996; Wang & Soergel, 1998).

Although previous empirical studies provide a rather comprehensive view of relevance criteria, they suffer a few serious limitations. First, the number of factors is very large. If a predictive model is to be built eventually in an IR system, asking the user to provide feedback on all the factors or automatically measuring all of them is impractical. Second, the terminology is confusing. The same criterion is named differently by different authors or users (e.g., accuracy and reliability, utility and usefulness); this condition calls for a streamlining of terms (Greisdorf, 2003). Third, factors have overlapping meanings (e.g., novelty, recentness, temporal issues). Fourth, the judgment of an IR system and the judgment of document content need to be distinguished. For example, the accessibility of a document (Bateman, 1998; Hirsh, 1999) is more a property of the IR system (whether it carries a certain document or not) than of the document per se. The relevance of a document should be judged on the basis of its content rather than its physical properties, such as physical availability or monetary cost (Borlund, 2003). Fifth, document attributes and relevance evaluations are treated at the same level (Fitzgerald & Galloway, 2001; Park, 1997; Spink, Greisdorf, & Bateman, 1998). Variables such as utility, usefulness, pertinence, informativeness, and helpfulness should be treated as surrogates of relevance judgment, i.e., as dependent variables; they should not be treated as independent variables or criteria because they are overall evaluations. Document attributes, whether objectively observed (e.g., date of publication) or subjectively perceived (e.g., novelty), should be independent variables. These attributes measure aspects of the document and contribute to the overall evaluation of it (i.e., relevance). A special aspect of the overall evaluation of a document is its hedonic impact such as enjoyment and happiness after reading. Just as the consumption of a product can have a utilitarian or hedonic impact, the consumption of information can. Our study, however, focuses only on the utilitarian aspect of relevance. Finally, as mentioned, the existing studies are exploratory rather than confirmatory in their methodology. With the myriad explorations carried out, a theory-driven, confirmatory study to integrate and verify the results is in order.

We note here that some of the problems described have been identified by previous research as well. For example, Barry and Schamber (1998) compared the results of their two studies in totally different situations, academic and weather media, and found a considerable overlap of relevance criteria. However, a critical question remains: what is the set of core relevance criteria? For that purpose, we turn to human communication theory.

Theory and Research Model

Departing from the extant research, which adopts an inductive and grounded exploratory methodology, we use a positivist perspective, which starts with hypotheses deduced from the current stock of domain knowledge or existing theories (Tashakkori & Teddlie, 1998). The positivist perspective and its theory-based hypothesis construction and testing procedure offer a number of advantages (Neuman & Kreuger, 2002). First, established theories, such as those of psychology and communication, address fundamental human behavioral patterns. They have also been tested in many different contexts and proved to have good generalizability. Second, established theories have identified the most important factors and their relationships in a more general domain. When applied to a particular problem, these factors provide guidance in identifying the corresponding domain-specific factors. Finally, because established theories postulate relationships among key factors, they provide a foundation on which new hypotheses may be proposed. Such hypotheses serve as the major testing area of a confirmatory study.

To identify the core relevance criteria, we propose that Grice's (1975, 1989) maxims on human communication be used as theoretical foundation. Not only does Grice's framework of maxims address human communication in general (and IR may be regarded as an indirect form of human communication), it is also consistent with many empirical studies in the IR area, as we will discuss shortly. Grice's work established the foundation of the inferential model in human communication, which is more general than Shannon's code model of communication (Sperber & Wilson, 1986). Grice (1975, 1989) posits that the essential feature of human communication, both verbal and nonverbal, is the recognition of the speaker's intention. In this model, a hearer infers a speaker's meaning or intention on the basis of the words or information provided. The communication is successful when both parties cooperate in making their meanings clear, i.e., the principle of cooperation. What kind of communication is cooperative? Grice further describes the hearer's expectation of the speaker's message in terms of conversational maxims: quantity, quality, relation, and manner.

The maxim of quantity has two submaxims. In Grice's words, to contribute an appropriate amount of information to communication is to "make your contribution to the conversation as informative as is required" and "not make your contribution to the conversation more informative than is required." Although Grice focuses on the appropriate amount of information in conversational communication, a more appropriate term in written communication via documents would be *scope.* We therefore identify scope as one relevance criterion. The concept of scope can be described in terms of two components: breadth and depth. Levitin and Redman (1995) suggest that scope and level of detail are the two important dimensions of data quality. They argue that a user needs the data to be broad enough to satisfy all the intended uses, and at the same time to include no unnecessary information. We define *scope* as the extent to which the topic or content covered in a retrieved document is appropriate to user's need: i.e., both the breadth and depth of the document are suitable. This definition encompasses similar concepts such as specificity (Cool, Belkin, & Kantor, 1993; Fitzgerald

DOI: 10.1002/asi

& Galloway, 2001; Schamber, 1991), depth/scope (Barry, 1994) and depth/breadth (Tang & Solomon, 1998). We hypothesize the following:

H1: Document Scope Is Positively Associated With Relevance

Notice that although Grice's maxim of quantity focuses on amount of information, it suggests that new information should be supplied; therefore, the conversation is "informative." In a user study, Wang and Soergel (1998) suggest that novelty and the resultant epistemic value are implied in any functional value of a document. Therefore, we add the criterion *novelty.*

Psychology researchers define *novelty* as a property of a stimulus that has not been previously presented to or observed by and is thus unfamiliar to the subject. In psychology literature, novelty-seeking behavior is regarded as the internal drive or motivational force of a human being (Cattell, 1975). Harter (1992) notices that normally a citation corresponding to an article already known to the requester cannot be psychologically relevant because it will not produce cognitive change in the requester. However, it may serve as a reminder. Therefore, novelty should be regarded as a matter of degree. We define *novelty* as the extent to which the content of a retrieved document is new to the user or different from what the user has known before. Novelty, as we define it, unifies similar concepts such as content novelty (Barry, 1994; Park, 1997) and divergent and strange content (Fitzgerald & Galloway, 2001). A few previous studies (e.g., Cool et al., 1993; Tang & Solomon, 1998) have mentioned recentness of a document. However, when a user comments on recentness, she implicitly assumes that it leads to novelty. Recentness can be regarded as one possible way of ensuring novelty, but not the only one. Therefore, recentness is conceptually different from novelty, though related to it. We therefore hypothesize the following:

H2: Document Novelty Is Positively Associated With Relevance

The maxim of quality also has two submaxims: "Do not say what you believe to be false" and "Do not say that for which you lack adequate evidence." We use the term *reliability* because the term *quality* in IR implies more than what Grice means. For example, quality could mean presentation, formatting, and even print quality, which are not properties of document content per se. However, Grice's maxim clearly refers to reliability.

Content reliability is different from source reliability. Reliability is first and foremost determined by document content. However, in addition to that, "source status, by influencing perceptions of source credibility, competence, or trustworthiness, can provide message recipients with a simple rule as to whether or not to agree with the message" (Petty, Priester, & Wegender, 1994, p. 103). Therefore, the credibility of the source can be regarded as an external cue of document reliability (e.g., Bateman, 1998; Hirsh, 1999; Spink et al., 1998). In this study, we define *reliability* as the degree to which the content of a retrieved document is perceived to be true, accurate, or believable. Similar concepts in the literature are accuracy (Schamber, 1991), validity (Barry, 1994), and agreement/disagreement (Fitzgerald & Galloway, 2001). We hypothesize the following:

H3: Document Reliability Is Positively Associated With Relevance

The maxim of relation is defined as "to be relevant." However, the term *relevant* in the maxim is used in its everyday sense—whether a response is about the topic being discussed or an abrupt switch to a different topic. In that sense, the term refers to *topicality* as defined in information science. The importance of topicality is widely recognized in relevance literature. Maron (1977) suggests that aboutness is the heart of indexing. Boyce (1982) indicates that users first judge the topicality of a document, and then consider other factors in their relevance judgment. Greisdorf (2003) also acknowledges topicality as the first or basic condition of relevance. Harter (1992) treats topicality as a weak kind of relevance. On the basis of early studies, Froehlich (1994) posits the nuclear role of topicality for relevance.

We adopt a subjective view and define *topicality* as the extent to which a retrieved document is perceived by the user to be related to her current topic of interest. Topicality, as we define it, unifies concepts that have been proposed in previous studies, such as subject area, discipline, focus, and aboutness. Because of the fundamental role of topicality in situational relevance, we have aligned our understanding of it with most prior exploratory studies and hypothesize the following:

H4: Document Topicality Is Positively Associated With Relevance

Finally, the maxim of manner is "avoid obscurity of expression," "avoid ambiguity," "be brief," and "be orderly." The meaning of this maxim is that conversation should be perspicuous so as to reduce the cognitive load of the hearer. We term it *understandability* in the context of written documents. Studies in communication and education show that the use of jargon or technical language reduces the clarity of a message and leads to significantly lower evaluation of the message than use of jargon-free language (Dwyer, 1999). For example, in client-professional exchange of information, the use of sophisticated language may affect the acceptance of the professional's advice (Elsbach & Elofson, 2000). Both experts and nonexperts are sensitive to the use of jargon in documents (Brown, Braskamp, & Newman, 1978). However, background knowledge is not the only factor that affects understandability. Presentation of the content, use of examples in explanation, and inclusion of graphics can all affect understandability (Bateman, 1998).

We define *understandability* as the extent to which the content of a retrieved document is perceived by the user as

DOI: 10.1002/asi

easy to read and understand. Understandability, as we define it, summarizes the effect of related concepts such as clarity, provision of examples (e.g., Bateman, 1998; Schamber, 1991), and special prerequisites (Wang & Soergel, 1998). We hypothesize the following:

H5: Document Understandability Is Positively Associated With Relevance

In summary, on the basis of Grice's maxims, we identify *scope, novelty, topicality, reliability,* and *understandability* as five core relevance criteria. These criteria correspond very well to the empirical findings in relevance research; many user studies have found them to be important (Barry, 1994; Bateman, 1998; Choi & Rasmussen, 2002; Fitzgerald & Galloway, 2001; Hirsh, 1999; Maglaughlin & Sonnewald, 2002; Park, 1993, 1997; Schamber, 1991; Spink et al., 1998; Tang & Solomon, 1998; Wang & Soergel; 1998; Wang & White, 1999).

Methodology

In order to test the proposed hypotheses, we used a survey method in our empirical study, followed by rigorous psychometric analysis, as proposed by the methodologists Anderson and Gerbing (1988). In psychometric analysis, structural equation modeling is a well-established and dominant quantitative data analysis method. It is widely used in education, marketing, information systems research, organization behavior, and many other disciplines (Kline, 1998). It is particularly suitable for testing causal relationships among psychological perceptions that are not directly observable to researchers.

Instrument Development

In psychometric analysis, in order to test hypotheses, the first step is to develop an instrument (e.g., questionnaire) to measure human perceptions of interest. Human perceptions of an object are called *constructs* or *latent factors* (e.g., topicality, relevance). A construct is assumed to be unobservable directly, but manifested in different ways. For instance, we cannot directly observe how novel a document is to a user, but we can ask questions regarding its novelty. Responses to those questions are the manifestations of the novelty construct. The questions, known as *items* or *scales*, reflect the different aspects of the construct. For example, in order to measure novelty, one might ask survey participants about the amount of new information in it, the amount of unique information in it, and its similarity to previous knowledge. Multiple questions are better than a single one. With multiple questions, the latent meaning underlying all the questions can be extracted by using statistical procedures such as factor analysis to produce a more accurate measure of a subject's true perceptions (the reader is referred to Nunnally & Bernstein, 1994, for detailed treatment of methodology). With a single question—for example, the amount of new information—the concept of novelty is reduced to a single aspect, and other possible manifestations of novelty are not taken into account. In fact, the concept of the construct or latent factor is not new to information scientists. Latent semantic analysis leverages on exactly the same reasoning. When all constructs are measured with multiple questions in a survey, the relationships among them can be tested.

It is recommended that researchers should reuse existing quality questions designed by others for the same construct (Nunnally & Bernstein, 1994). Because there is no prior study using a psychometric instrument, we developed all questions in this study on the basis of our definitions of the five core concepts. Items (i.e., questions) for topicality, novelty, understandability, reliability, and relevance were designed with an 8-point scale. There were three anchoring points on the scale: one on value 0, one on value 1, and one on value 7. For example, a question measuring relevance was, My opinion/view of the current topic has been significantly changed or strengthened by this document (0—totally disagree, 1—strongly disagree, 7—strongly agree). Although Tang and colleagues (1999) suggest the 7-point scale as the best choice, we included 0 in the scale to allow for possible binary judgment by users. However, scope and background knowledge were designed as 7-point scales because we believe it is meaningless for them to assume 0 value. The complete questionnaire is given in Appendix A.

To ensure that items reflect the intended construct, content validity should be checked first. Assuming that all the possible manifestations of a construct collectively form the population of questions, content validity is the degree to which the questions used in the survey for a construct provide representative coverage of the population. The questions we used were to a large degree the rephrasing of different aspects of a construct as defined in the literature, which provided the basis for content validity. In addition, two Ph.D. students familiar with the research project were invited to discuss the phrasing of the questions to ensure that they had at least face validity (Nunnally & Bernstein, 1994, chap. 3). Minor changes were made on the basis of the feedback.

Data Collection

We carried out a semicontrolled survey to collect data. The subjects were undergraduate and graduate students in schools in a major university in southeastern Asia. They were invited to a computer laboratory, where they were given four search topics and asked to choose one that interested them most; the topics were (1) inappropriate intake of vitamins, (2) outlook of the job market related to the subject's main area of study, (3) reasons for the dotcom bubble burst in 2000, (4) a three-day trip plan to Tibet. They were also allowed to define their own topic if none of the given ones interested them. The subjects were then instructed that the search would be done on the Internet, and the results were meant for a group discussion that they should be involved in later. They were also told to search for as much information as possible so that they might feel comfortable

DOI: 10.1002/asi

with their knowledge on the chosen topic. After a topic had been chosen, the subjects were measured on their background knowledge on the topic (refer to Appendix A). Then the search process began. A monitoring program was also started on each computer to collect the browsing history of the subjects and the time they spent on each Web page. When the subjects had completed their search, the monitoring program provided a list of documents they had browsed and the time that they had spent on browsing each. They were then asked to choose freely and evaluate two Web documents that they had browsed. The criteria for these documents were that they must not be navigation pages (pages mainly containing links), and the subjects must have spent more than 1 minute on the documents as recorded by the monitoring program. The subjects were told to evaluate the two documents according to the questionnaire in Appendix A. They generally took 30 to 60 minutes to complete the whole process and were given about $5 as reward for their participation.

A semicontrolled survey suited our study for two main reasons: First, it helped control for many peripheral factors such as document formatting quality (if we assume Web pages have similar visual readability), task motivation (the subjects used their search results for group discussion), and accessibility (all documents were accessible online). Users' background knowledge was measured as a covariate. Controlling for peripheral factors allowed us to focus on the effect of the theoretically interesting factors. Second, although the requirement for the subjects to evaluate only documents that they had read might introduce sampling bias, it helped to ensure that the subjects really knew the content. If a subject did not even read the document, it was impossible for her to make valid evaluations. However, our procedure was likely to have included more relevant documents than irrelevant ones. As will be shown later, the actual sample was upward biased. However, it still had very satisfactory coverage; hence, it should not affect final hypothesis testing.

The survey was carried out in two phases. Phase 1 was a pilot test involving 36 students. The purpose of the pilot test was to ensure the quality of the questionnaire and experiment procedure. After that, the main study was carried out. Both studies followed exactly the procedure described earlier.

Pilot Study

In the pilot test, 72 document evaluations were collected. In order to verify the questionnaire, exploratory factor analysis (Nannally & Bernstein, 1994) was conducted to test the convergent and discriminant validity of the instrument. *Convergent validity* indicates that all questions intended to measure a construct do reflect that construct. *Discriminant validity* indicates that a question should not reflect an unintended construct and that constructs are statistically different. Exploratory factor analysis is an adequate tool because it allows the underlying factors to emerge naturally from the data without imposing any constraint. If the questions for a construct are well designed, they should converge and form a major factor. If a question is problematic, it can be detected and excluded from later study.

Exploratory factor analysis with principal component analysis was used to extract factors in our study. By following the recommended procedures (Hair, Anderson, Tatham, & Black, 1995), major principal components were extracted as constructs; minor principal components with eigenvalue less than 1 were ignored as noise; an item and the intended construct correlation (also known as *factor loading*) should be greater than 0.5 to satisfy convergent validity; an item and the unintended construct correlation should be less than 0.4 for discriminant validity. We extracted seven factors corresponding to the seven constructs. The items NOVEL4, SCOPE4, and RELEV2 were dropped because they did not satisfy the discriminant and convergent requirements. The remaining items showed appropriate validity and were kept for the main study. Table 1 reports the principal component analysis results with varimax rotation using SPSS11.0 after the items that failed to satisfy validity requirements were dropped.

Main Study and Data Analysis

The main study involved 132 student participants, yielding a total of 264 document evaluations. However, two evaluations were incomplete and had to be discarded, resulting in 262 document evaluations for use. The majority of the subjects were male ($M = 72.7\%$, $F = 27.3\%$) undergraduate students (undergraduate = 96.7%, graduate = 3.3%). The mean age was 20.6 (standard deviation = 1.5).

Among the 262 questionnaires returned, most items did not have a 0 score, even though we allowed topicality, novelty, reliability, and relevance to assume 0 value. The number of instances in which an item had 0 score was 53, or 0.6% of the total data. The 0-score items were from 20 documents, representing 7.6% of 262 documents. There were only three cases in which all items of one construct received 0 score. We discarded the 20 documents that had 0-score items because a 0 score indicated a binary decision that is qualitatively different from that of scores between 1 and 7 and are incremental and continuous. Binary and continuous evaluations should not be mixed in data analysis. With the 20 documents discarded, we had 242 usable document evaluations left. Among them, 73 were based on the topic of inappropriate vitamin intake (30.2%), 17 on the topic of the job market (7%), 10 on the topic of the dotcom bubble (4.1%), 78 on the topic of the 3-day trip to Tibet (32.2%) and 64 self-selected topics (26.4%).

Most of the constructs had a mean between 4.0 and 5.0, indicating an upward bias in the document sample, comparable to the midpoint of 4. One possible reason for this bias is that we required the subjects to evaluate only the documents they had read for more than 1 minute. Because structural equation modeling requires that the data be normally distributed (Nunnally & Bernstein, 1994; Hair et al., 1995), to ensure that the upward bias would not jeopardize normality, we made a univariate normality test for all items on skewness and kurtosis. A 95% confidence level was

DOI: 10.1002/asi

TABLE 1. Factor loading table for exploratory factor analysis of pilot data.

		Component						
Construct	Items	1	2	3	4	5	6	7
Topicality	TOPIC1	.851	.172	−.084	−.008	.115	.129	−.114
	TOPIC2	.886	.043	.131	.062	.009	.056	−.141
	TOPIC3	.896	−.004	.189	.115	.074	.124	−.007
	TOPIC4	.787	−.032	.009	.159	.285	.156	−.017
Reliability	RELIA1	.212	.681	−.236	.180	.193	.054	−.071
	RELIA2	−.065	.850	−.082	.048	.112	.143	.024
	RELIA3	.119	.883	.055	.020	.091	.099	.004
	RELIA4	−.048	.740	.364	−.025	−.135	−.026	.076
Understandability	UNDER1	.134	.030	.919	−.062	.067	.009	−.048
	UNDER2	.030	.012	.944	.017	.026	.089	−.043
	UNDER3	.047	−.013	.865	.044	.001	.099	−.107
Novelty	NOVEL1	.341	−.069	−.346	.582	.106	.303	−.177
	NOVEL2	.069	.150	.161	.670	.058	−.059	−.048
	NOVEL3	.053	.011	−.084	.784	.014	.034	−.173
Scope	SCOPE1	.146	.056	.124	.016	.899	.092	−.004
	SCOPE2	.064	.101	.048	.046	.887	.112	−.130
	SCOPE3	.301	.108	−.162	.119	.523	.249	.048
Relevance	RELEV1	.191	−.073	.068	−.033	.090	.668	−.161
	RELEV3	.242	.121	.249	.333	.189	.607	.207
	RELEV4	.127	.132	.117	−.004	.108	.861	.120
	RELEV5	−.005	.181	−.083	.007	.089	.876	.173
Prior knowledge	KNOW1	−.096	−.037	−.065	−.116	.023	.087	.941
	KNOW2	−.095	.041	−.150	−.062	−.082	.031	.944
	KNOW3	−.059	.028	.023	−.141	−.052	.070	.935
	Eigenvalue	5.330	3.345	3.004	2.364	1.722	1.480	1.195
	Variance %	22.21	13.94	12.52	9.85	7.17	6.17	4.98
	Cumulative Variance %	22.21	36.15	48.67	58.52	65.70	71.86	76.84

TABLE 2. Descriptive statistics and factor correlation.

	Mean	*SD*	No. of 0 scores[a]	Min.	Max.	KNOW	TOP	RELI	UND	NOV	SCO	RELE
KNOW	3.25	1.36	—	1	7	0.91						
TOPIC	5.08	1.29	16	1	7	0.00	0.85					
RELIA	5.40	0.97	0	2	7	−0.12	0.62	0.86				
UNDER	5.27	1.23	2	1	7	−0.08	0.35	0.33	0.87			
NOVEL	4.61	1.20	5	1	7	−0.10	0.25	0.25	0.09	0.74		
SCOPE	3.93	1.23	—	1	7	−0.06	0.25	0.24	0.21	0.14	0.78	
RELEV	4.67	1.14	30	1	7	0.02	0.57	0.54	0.35	0.47	0.31	0.81

[a]Number of zero scores is calculated on the basis of the 262 surveys returned. The rest of the statistics are based on the data after dropping the records with zero score.

imposed. All items fell within the corresponding confidence intervals, rendering our data suitable for further analysis. Table 2 shows the descriptive statistics.

Measurement model. According to Anderson and Gerbing (1988), measurement modeling should be carried out as the first step of structural equation modeling. The purpose of measurement modeling is to ensure instrument quality. Unlike exploratory factor analysis, measurement modeling prespecifies construct-item correspondences but leaves correlation coefficients (i.e., factor loadings) free to change. The prespecified construct-item correspondences are then tested for confirmation. Confirmatory factor analysis (CFA) is the conventional statistical method used to specify and test such relationships for the measurement model. With this method, items are expected to be highly correlated with the intended construct only. If an item is not substantially related to the intended construct, or is significantly related to an unintended construct, the prespecified relationship is invalidated and adjustment of the instrument is required. Therefore, the first requirement of confirmatory factor analysis is that the construct-item correlation be significant

DOI: 10.1002/asi

(Anderson & Gerbing, 1988). In addition, for an item, the average variance extracted (AVE) by the latent factor should be greater than 0.5: that is, a construct should explain more than 50% of the item variance (Fornell & Larcker, 1981). Moreover, items of the same construct should be highly correlated. To measure such correlations, two measures, composite factor reliability (CFR) and Cronbach's alpha (α), are required to be greater than 0.7 (Hair et al., 1995). If all these criteria (significant correlation, high AVE, high CFR, and α are satisfied, the convergent validity of the items is said to be confirmed. Table 3 reports the result of convergent validity for our sample using statistical package LISREL v8.5. All criteria were satisfied.

TABLE 3. The convergent validity of the measurement model.

Construct	Item	Std. Loading	*T* value	AVE	CFR	α
Topicality	TOPIC1	0.84	15.72	0.715	0.909	0.907
	TOPIC2	0.86	16.81			
	TOPIC3	0.88	16.86			
	TOPIC4	0.80	14.49			
Reliability	RELIA1	0.83	15.50	0.732	0.916	0.916
	RELIA2	0.89	17.41			
	RELIA3	0.87	16.83			
	RELIA4	0.83	15.54			
Scope	SCOPE1	0.80	13.35	0.606	0.821	0.818
	SCOPE2	0.83	13.85			
	SCOPE3	0.70	11.31			
Understand-ability	UNDER1	0.89	16.94	0.757	0.903	0.903
	UNDER2	0.84	15.52			
	UNDER3	0.88	16.83			
Novelty	NOVEL1	0.76	12.22	0.550	0.784	0.778
	NOVEL2	0.62	9.67			
	NOVEL3	0.83	13.51			
Relevance	RELEV1	0.70	11.95	0.652	0.882	0.877
	RELEV3	0.83	15.53			
	RELEV4	0.88	16.80			
	RELEV5	0.81	14.72			
Prior knowledge	KNOWE1	0.94	19.14	0.823	0.933	0.933
	KNOWE2	0.89	17.52			
	KNOWE3	0.89	17.43			

One way to confirm discriminant validity is to check that interconstruct correlation is less than the square root of AVE (Fornell & Larcker, 1981). The underlying rationale is that an item should be better explained by its intended construct than by other constructs. The correlation among constructs is reported in Table 2. Discriminant validity was confirmed in our sample.

In summary, our measurement model confirmed the difference between all the relevance factors used. It also confirmed the internal consistency of the different aspects of the relevance factors as manifested in different questions (items). With that, we could proceed to test the causal relationship among all the factors.

Structural model. The structural model could be intuitively understood as a regression model, albeit with the variables being latent factors (i.e., the constructs) rather than explicit measures. The measurement model discussed previously can be used to calculate latent factor scores, which are then used as input for the later regression analysis to estimate the relationship between the independent variables (e.g., topicality, novelty) and the dependent variable (i.e., relevance). If the hypotheses are supported, we expect significant regression coefficients for the factors.

Because the measurement model was satisfactory, we proceeded to hypothesis testing. Hypothesis testing is done by creating a structural equation model in LISREL, which specifies both item-construct correspondence and construct-construct causal relationship excluding the control variables. The coefficients are then solved with maximum likelihood estimation. We followed this procedure and arrived at the results summarized in Figure 1.

Before any conclusions can be drawn for hypothesis testing, the model must fit the data well. A few model fitting indices can be employed here. For example, the chi-square and the degree of freedom ratio (a normalized measure of the "badness of model fit") must be less than 3; the root mean square error of approximation (*RMSEA*, a measure of the residual) must be less than 0.10; and the goodness of fit index (GFI) must be greater than 0.9 (Anderson & Gerbing, 1988; Hair et al., 1995; Nunnally & Bernstein, 1994). Our

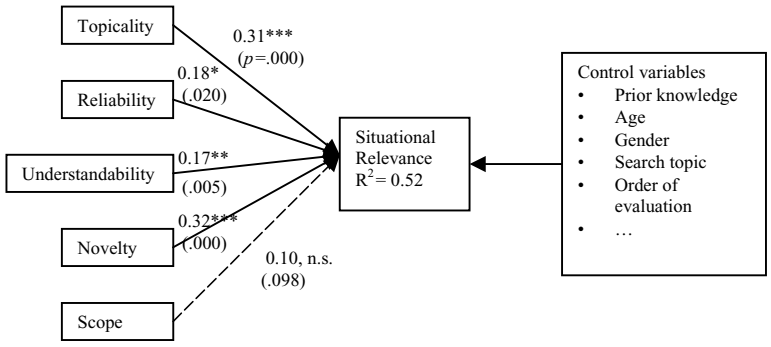

FIG. 1. Standardized LISREL solution for hypothesis testing ($\chi^2 = 272.19$, $df = 174$, $p = 0.0000$, *RMSEA* = 0.048, *NFI* = 0.92, *NNFI* = 0.96, *CFI* = 0.97, *IFI* = 0.97, *RFI* = 0.90, *GFI* = 0.90, *AGFI* = 0.87, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$).

DOI: 10.1002/asi

results indicated satisfactory model fit. From the detailed results, we concluded that all our hypotheses were supported except scope, which was nonsignificant at the $p = 0.05$ level but significant at the $p = 0.10$ level.

We had also added age, gender, order of a document, search topic, and user's knowledge as control variables to the model. They were all insignificant except topic 3: age ($\beta = -.03$, $p = .59$), gender ($\beta = -.03$, $p = .52$), order in which a document is evaluated by a subject ($\beta = -.02$, $p = .75$), search topics (topic 1: $\beta = -.03$, $p = .71$; topic 2: $\beta = -.03$, $p = .62$; topic 3: $\beta = -.11$, $p = .05$; topic 4: $\beta = -.08$, $p = .23$), and prior knowledge ($\beta = 0.09$, $p = .09$). No significance change was observed for the main factors. Therefore, the hypotheses were robust across variation in the control variables.

Discussion and Implications

Summary of Data Analysis

The objective of this study was to identify and confirm a set of key relevance judgment criteria. We have identified five such criteria, i.e., topicality, novelty, understandability, reliability, and scope, from Grice's maxims and previous literature. We are particularly interested in the factors that a user considers beyond topicality.

The results of our exploratory and confirmatory factor analysis show that our five constructs satisfy discriminant validity (i.e., they are distinct concepts) and convergent validity requirements. For each construct, different phrasings that emphasize different aspects of the same construct should be unified. For example, topicality can be termed as relatedness to information need or subject area aboutness; yet all such terms load on a single factor because of their shared meaning. A first attempt at the same goal was made by Bateman (1998) using confirmatory factor analysis. However, because the items used in that study were not designed on the basis of theory, the conceptualization was incomplete. Our study goes one step further by refining and verifying the key relevance antecedents.

Among the five criteria that we have proposed on the basis of Grice's theory, only the criterion scope is not supported by the data. Our results show that topicality and novelty are the two factors that are most significant to relevance judgment ($\beta_{\text{Topicality}} = 0.31$, $\beta_{\text{Novelty}} = 0.32$). Understandability and reliability are also significant but to a smaller degree. Together, all our criteria explain 52% of relevance variance in our results, suggesting that these factors are quite comprehensive in explaining relevance judgment. Inclusion of control variables does not appear to change the significance of the hypotheses. Our results are therefore robust across the demographic and topical differences in our context of study.

Limitations and Future Directions

As ours is the first confirmatory study of relevance judgment that follows a psychometric procedure, logically we should point out the key limitations of our effort before drawing any implications. First, although we adopted a hypothesis confirmation process in our study, our findings may not be valid in some other contexts because relevance is context dependent and dynamic. Second, the use of structural equation modeling assumes an additive model: i.e., the contribution of each criterion to relevance is additive. This assumption might be viable when minimal topicality is assumed, in which case other criteria can be considered extra premiums to the basic topicality requirement. In our study, because the document sample was upward biased and totally irrelevant documents were discarded, we may assume topicality to be present for most documents. In other contexts, if this assumption is not met, other types of models such as a multiplicative or stepwise model might be considered.

Relevance Revisited

Our findings suggest that topicality and novelty are the two major underlying dimensions of relevance. If they are, then the concept of relevance can be depicted with different combinations of topicality and novelty levels. If we classify topicality and novelty into low and high levels, we will have four quadrants: low topicality–low novelty, low topicality–high novelty, high topicality–low novelty, and high topicality–high novelty.

In the low topicality–low novelty quadrant, a document is neither on topic nor new to the reader. It is thus most likely to be dismissed as *irrelevant.* In some cases, a document may be outright off topic or may provide duplicate information (already known to the user); in other cases, it could be a document slightly related to the user's information need even though the information may already be known to the user. For example, Salton and McGill's (Salton & McGill, 1983) classic textbook *Introduction to Modern Information Retrieval* is at best marginally related to this article and not new to us; its relevance to us is very low if not nonexistent. In the high topicality–low novelty quadrant, a document is on topic but already known to the user. Imagine that we are writing another article to address the limitations of this study; Saracevic (1975) would then fall into this category. Saracevic (1975) is a classic article on the topic of subjective relevance; hence, it has topicality. However, as we write the follow-up article, we are already familiar with the content of Saracevic (1975). We may still treat it as relevant because we need to cite it occasionally and to check some concepts defined there or to quote some sentences. Such a document would be useful, pertinent, and relevant to the research at hand, but it would be used as a *tool.*

The low topicality–high novelty quadrant deals with documents that are unclear in topicality yet provide certain new information that attracts the reader's attention. As Harter (1992) points out, there is no absolutely fixed information need in a search process. Information needs are typically vague and multiple in nature. The interaction between new information in a document and the current cognitive state of the user helps clarify the information need and create future

DOI: 10.1002/asi

needs. Consequently, a document might be regarded as relevant because the reader anticipates its future value rather than its current value. Harter (1992) even draws a surprising conclusion that a relevant document not on topic is more important than a relevant document on topic. We attribute the potential value of such off-topic documents to their novelty. Researchers share the common experience that a not-so-relevant reference previously collected turns out to be a major reference in later research. It is such *potential relevance* that characterizes the low topicality–high novelty quadrant. Finally, the high topicality–high novelty quadrant contains the ideal documents that inform and satisfy the current need. Such documents might help the reader clarify an information need or offer new solutions or a new method to evaluate different possible solutions. In either case, they enrich the reader's cognitive state about the topic of interest, or the information they contain can be applied directly to solve the problem at hand. If the utility of the document is more related to cognitive enlightening, we can term this type of relevance *informativeness.* If the utility of the document also bears on actual problem solving, it has high *situational relevance.* In this quadrant, when the current information need is satisfied, it spurs new information needs and new rounds of search. Taking the two dimensions together, a relevant document is not one that is just on topic. The change it introduces to the current cognitive state is also indispensable. Such cognitive change is the heart of relevance as a potentially dynamic subjective notion. Because cognitive change hinges on novelty, relevance would be a static concept without novelty. Topicality and novelty are thus the two pillars of the concept of relevance.

Contribution of Reliability, Understandability, and Scope

Why is scope nonsignificant in our findings? One plausible explanation is that scope is a premium factor. Users may not have a keen expectation of scope and turn to other criteria first. Hence, although they could be happier with documents of appropriate scope, they would not see that consideration as critical to their relevance judgment. Wang and Soergel (1998) posit that information seeking is a phased process. The first step is to select relevant documents for focused reading. The relevance judgment we have investigated falls into this phase. Here, because a user faces pressure to go through a lengthy list of documents, quick decision is required to filter out the irrelevant ones. Less effort is therefore exerted to judge the appropriateness of scope. However, scope might be a significant factor when we move to the later stages of information-seeking behavior, e.g., focused reading or citation.

What are the roles of understandability, reliability, and scope in relevance decisions? Our findings show that they exert relatively less influence than novelty and topicality. Assume a document is at least marginally on topic. Two scenarios may then be visualized when understandability and reliability take effect. In the first scenario, when a document is outright unreliable (Grice, 1989) or impossible to understand (e.g., in a nonpreferred language, Spink et al., 1998), document relevance vanishes. However, such cases are rare or can be addressed technically (e.g., retrieve only documents in the desired languages). In most cases, users are very unlikely to dismiss the document as totally unreliable or impossible to understand. Regarding reliability, because a user has only limited knowledge of a domain (the reason she needs to search for information), she is unable to judge the reliability of a document with full confidence. She tends to assume a document is reliable until she is confronted by other arguments. In our sample, no subject set reliability score to 0 and the average reliability was 5.4, indicating rather high perceived reliability. As for understandability, the mean was the second highest (5.27). Again, if a user does not have full knowledge of a topic, a certain level of cognitive effort in reading a document is anticipated. In summary, when topicality and novelty are present even at a low level, reliability, understandability, and scope are more likely to be additional values.

Would the contribution of all the five factors be the same to different types of relevance (i.e., cognitive, situational, and affective relevance; Cosijn & Ingwersen, 2000; Saracevic, 1996) and at different stages of document consumption (i.e., collection, reading, and citation, Wang & Soergel, 1998; Wang & White, 1999)? These are the future directions for exploration.

Methodological Implications

As ours is the first study that adopts confirmatory psychometric analysis in relevance judgment, it has multiple methodological implications. First, we have demonstrated how various conceptualizations of relevance criteria can be unified and differentiated with exploratory and confirmatory factor analysis. Our study showcases a pathway to establish a common language. It also provides some building blocks for future relevance research. Second, we recommend rigorous hypothesis testing for relevance study. When a large number of exploratory field studies have been carried out, further verifying or fortifying of the explored territory would provide an established foundation on which further studies could be carried out. Hypothesis testing is a useful tool for verifying exploratory findings and establishing causal relationships among theoretical constructs. Whereas exploratory studies are good for uncovering underlying factors of a domain, the complementary advantage of hypothesis testing lies in its ability to verify the factors in a larger context statistically and hence ensure their generalizability.

System Design Implications

Decades of research effort have been made to improve capture of topicality. What our study suggests is that the next powerhouse of IR system design might be the quantification of novelty. The following are questions that we could consider: How could we capture a reader's cognitive state before document evaluation? How could we measure the

DOI: 10.1002/asi

novelty of a document against the cognitive state? How could we combine novelty and topicality in an overall relevance score? Although this study does not offer any answer to these questions, we suggest that effort in these directions will be rewarding. In fact, at the paragraph or sentence level, the novelty track of Text REtrieval Conference (TREC) has already begun to consider these issues. Document level application of novelty, however, still lacks attention.

Acknowledgment

This research was supported by the School of Computing, National University of Singapore, Research Grant: R253-000-028-112.

References

Anderson, J.C., & Gerbing, D.W. (1998). Structure equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin, 103(3), 411–423.

Barry, C.L. (1994). User-defined relevance criteria: An exploratory study. Journal of the American Society for Information Science, 45(3), 149–159.

Barry, C., & Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. Information Processing & Management, 34, 219–236.

Bateman, J. (1998). Changes in relevance criteria: A longitudinal study. Proceedings of the 61st Annual Meeting of the American Society for Information Science, 35, 23–32.

Bookstein, A. (1979). Relevance. Journal of the American Society for Information Science, 30(5), 269–273.

Borlund, P. (2003). The concept of relevance in IR. Journal of the American Society for Information Science and Technology, 54(10), 913–925.

Boyce, B. (1982). Beyond topicality: A two stage view of relevance and the retrieval process. Information Processing & Management, 18(3), 105–109.

Brown, R.D., Braskamp, L.A., & Newman, D.L. (1978). Evaluator credibility and acceptance as a function of report styles: Do jargon and data make a difference? Evaluation Quarterly, 2(2), 331–341.

Cattell, R.B. (1975). Personality and motivation: Structure and measurement, New York: Harcourt, Brace & World.

Choi, Y., & Rasmussen, E.M. (2002). Users' relevance criteria in image retrieval in American history. Information Processing & Management, 38, 695–726.

Cool, C., Belkin, N.J., & Kantor, P.B. (1993). Characteristics of texts affecting relevance judgments. In M.E. Williams (Ed.), Proceedings of the 14th National Online Meeting (pp. 77–84). Medford, NJ: Learned Information.

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. Information Processing & Management, 36(4), 533–550.

Cuadra, C.A., & Katter, R.V. (1967). Opening the black box of "relevance." Journal of Documentation, 23(4), 291–303.

Dwyer, J. (1999). Communication in business: Strategies and skills. Sydney: Prentice Hall.

Elsbach, K.D., & Elofson, G. (2000). How the packaging of decision explanations affects perceptions of trustworthiness. Academy of Management Journal, 43(1), 83–89.

Fitzgerald, M.A., & Galloway, C. (2001). Relevance judging, evaluation, and decision making in virtual library: A descriptive study. Journal of the American Society for Information Science and Technology, 52(12), 989–1010.

Fornell, C., & Larcker, D.F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. Journal of Marketing Research, 18, 382–388.

Froehlich, T.J. (1994). Relevance reconsidered: Towards an agenda for the 21st century: Introduction to special topic issue on relevance research. Journal of the American Society for Information Science, 45, 124 –134.

Greisdorf, H. (2003). Relevance thresholds: A multi-stage predictive model of how users evaluate information. Information Processing & Management, 39, 403–423.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), Syntax and semantics (Vol. 3, pp. 41–58). New York: Academic Press.

Grice, H.P. (1989). Studies in the way of words. Cambridge, MA: Harvard University Press.

Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (1995). Multivariate data analysis with reading (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Harter, S.P. (1992). Psychological relevance and information science. Journal of the American Society for information Science, 43(9), 602–615.

Hersh, W. (1994). Relevance and retrieval evaluation: Perspective from medicine. Journal of the American Society for Information Science, 45(3), 201–206.

Hirsh, S.G. (1999). Children's relevance criteria and information seeking on electronic resources. Journal of the American Society for Information Science, 50(14), 1265–1283.

Hjørland, B., & Christensen, F.S. (2002). Work tasks and socio-cognitive relevance: A specific example. Journal of the American Society for Information Science and Technology, 53(11), 960–965.

Kline, R.B. (1998). Principles and practice of structural equation modeling. New York: Guilford Press.

Levitin, A., & Redman, T. (1995). Quality dimensions of a conceptual view. Information Processing & Management, 31(1), 81–88.

Maglaughlin, K.L., & Sonnewald, H. (2002). User perspective on relevance criteria: A comparison among relevant, partially relevant, and not-relevant. Journal of the American Society for Information Science and Technology, 53(5), 327–342.

Maron, M.E. (1977). On indexing, retrieval and the meaning of about. Journal of the American Society for Information Science, 28(1), 38–43.

Mizzaro, S. (1997). Relevance: The whole history. Journal of the American Society for Information Science, 48(9), 810–832.

Neuman, W.L., & Kreuger, L.W. (2002). Social work research methods: Qualitative and quantitative approaches. Boston: Allyn & Bacon.

Nunnally, J.C., & Bernstein, I.H. (1994). Psychometric theory. New York: McGraw-Hill.

Park, H. (1997). Relevance of science information: Origins and dimensions of relevance and their implications to information retrieval. Information Processing & Management, 33(3), 339–352.

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

Park, T.K. (1994). Toward a theory of user-based relevance: A call for new paradigm of inquiry. Journal of the American Society for Information Science, 45(3), 135–141.

Petty, R., Priester, J., & Wegender, D. (1994). Cognitive processes in attitude change. In R. Wyer & T. Srull (Eds.), Handbook of social cognition (pp. 69–142). Hillsdale, NJ: Erlbaum.

Rees, A.M., & Schultz, D.G. (1967). A field experimental approach to the study of relevance assessments in relation to document searching: I. Final report (NSF Contract No. C-423). Cleveland, OH: Case Western Reserve University.

Salton, G., & McGill, M. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Saracevic, T. (1970). The concept of "relevance" in information science: A historical review. In Saracevic, T. (Ed.), Introduction to information science (pp. 111–151). New York: R.R. Bowker.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26(6), 321–343.

Saracevic, T. (1996). Relevance reconsidered '96. In P. Ingwersen & N.O. Pots (Eds.), Second International Conference on Conceptions of Library and Information Science (CoLIS2) (pp. 201–218). Copenhagen: Royal School of Librarianship.

Schamber, L. (1991). Users' criteria for evaluation in a multimedia environment. In J.M. Griffiths (Ed.), Proceedings of the 54th Annual Meeting

DOI: 10.1002/asi

of the American Society for Information Science (Vol. 28, pp. 126–133). Medford, NJ: Information Today.
Schamber, L. (1994). Relevance and information behavior. Annual Review of Information Science and Technology, 29, 33–48.
Schamber, L., & Bateman, J. (1996). User criteria in relevance evaluation: Toward development of a measurement scale. Proceedings of the 59th Annual Meeting of the American Society for Information Science (33, 218–225). Medford, NJ: Information Today.
Schamber, L., Eisenberg, M.B., & Nilan, M.S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. Information Processing & Management, 26(6), 755–776.
Sperber, D., & Wilson, T.D. (1986). Relevance: Communication and cognition. Cambridge, MA: Harvard University Press.
Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. Information Processing & Management, 34, 599–621.
Tang, R., & Solomon, P. (1998). Towards an understanding of the dynamics of relevance judgments: An analysis of one person's search behavior. Information Processing & Management, 34, 237–256.
Tang, R., Shaw Jr., W.M., & Vevea, J.L. (1999). Towards the identification of the optimal number of relevance categories. Journal of the American Society for Information Science, 50, 254–264.
Tashakkori, A., & Teddlie, C. (1998). Mixed methodology: Combining qualitative and quantitative approaches. Thousand Oaks, CA: Sage.
Wang, P., & Soergel, D. (1998). A cognitive model of document use during a research project: Study I. Document selection. Journal of the American Society for Information Science, 49(2), 115–133.
Wang, P., & White, M.D. (1999). A cognitive model of document use during a research project: Study II. Decisions at the reading and citing stages. Journal of the American Society for Information Science, 50(2), 98–144.
Wilson, P. (1973). Situational relevance. Information Storage and Retrieval, 9(8), 457–471.

Appendix: Major Items of the Questionnaire

	Strongly disagree						Strongly agree
1. I know this topic area very well. (KNOW1)*	1	2	3	4	5	6	7
2. I am able to tell others much about this topic. (KNOW2)	1	2	3	4	5	6	7
3. I am very confident in my knowledge about this topic. (KNOW3)	1	2	3	4	5	6	7

1. The main content of this document ___ my current topic of interest. (TOPIC1)*	Is totally unrelated to	Marginally describes						Substantially describes
	0	1	2	3	4	5	6	7
2. This document _____ the general domain of my current topic of interest. (TOPIC2)	Is totally off	Touches on						Is surely within
	0	1	2	3	4	5	6	7
3. The subject area of the document is ____ to my current topic of interest. (TOPIC3)	Totally unrelated	Marginally related						Substantially related
	0	1	2	3	4	5	6	7
4. ____ details in this document are related to my current topic interest. (TOPIC4)	None of the	Very few of						A large number of
	0	1	2	3	4	5	6	7
5. In this document, the amount of *new information* to me is ____. (NOVEL1)	None	Small						Substantial
	0	1	2	3	4	5	6	7
6. I *already* ___ the phenomena/arguments described in the document. (NOVEL2)	Knew all	Knew a substantial part of						Knew little of
	0	1	2	3	4	5	6	7
7. This document has ___ *unique information* that I am coming across for the first time. (NOVEL3)	No	A small amount of						A substantial amount of
	0	1	2	3	4	5	6	7
8. The content of this document is ____ the content of other document(s) I have read. (NOVEL4)**	Identical to	Very similar to						Very different from
	0	1	2	3	4	5	6	7
9. The content of this document is ____ for me to understand. (UNDER1)	Impossible	Very difficult						Very easy
	0	1	2	3	4	5	6	7

DOI: 10.1002/asi

10. I am able to follow the content of this document _____. (UNDER2)	In no way	With much effort						With little effort
	0	1	2	3	4	5	6	7
11. Readers of my type should find the document _____ to read. (UNDER3)	Impossible	Very difficult						Very easy
	0	1	2	3	4	5	6	7
12. I think the content of this document would be____. (RELIA1)	Totally wrong	Very inaccurate						Very accurate
	0	1	2	3	4	5	6	7
13. I think the content of this document would be ____ facts. (RELIA2)	Totally against	Very inconsistent with						Very consistent with
	0	1	2	3	4	5	6	7
14. I think the content of this document would be____. (RELIA3)	Totally unreliable	Very unreliable						Very reliable
	0	1	2	3	4	5	6	7
15. I think the content of this document would be ____. (RELIA4)	Surely false	Most likely false						Most likely true
	0	1	2	3	4	5	6	7
16. The content of this document is either *too general* or *too specific* for me. (SCOPE1)		Strongly disagree						Strongly agree
		1	2	3	4	5	6	7
17. The coverage of this document is either *too broad* or *too narrow* for me. (SCOPE2)		Strongly disagree						Strongly agree
		1	2	3	4	5	6	7
18. This document gives either *too many* or *too few* details compared to what I expected. (SCOPE3)		Strongly disagree						Strongly agree
		1	2	3	4	5	6	7
19. The scope of this document is *inappropriate* for me. (SCOPE4)**		Strongly disagree						Strongly agree
		1	2	3	4	5	6	7
20. This document can *be used to solve problems in my current topic of interest.* (RELEV1)	Totally disagree	Strongly disagree						Strongly agree
	0	1	2	3	4	5	6	7
21. My opinion/view towards the current topic has been *significantly changed or strengthened* by this document. (RELEV2)**	Totally disagree	Strongly disagree						Strongly agree
	0	1	2	3	4	5	6	7
22. If asked about the current topic, I would *tell people* things based on this document. (RELEV3)	Not at all	Very unlikely						Very likely
	0	1	2	3	4	5	6	7
23. When facing a problem in my current topic of interest, I will really *apply the knowledge* learned from this document to solve it. (RELEV4)	Totally disagree	Strongly disagree						Strongly agree
	0	1	2	3	4	5	6	7
24. When facing a problem in my current topic of interest, I will *take action* according to what is suggested in this document. (RELEV5)	Totally disagree	Strongly disagree						Strongly agree
	0	1	2	3	4	5	6	7

*Item IDs were not in the original questionnaire.
**Items dropped after pilot test.

DOI: 10.1002/asi