Searching for Relevance in the Relevance of Search

Elaine G. Toms[1], Heather L. O'Brien[1], Rick Kopak[2] & Luanne Freund[3]

[1]Faculty of Management, Dalhousie, Halifax, Nova Scotia B3H 4H8 Canada
etoms@dal.ca, hlobrien@dal.ca
[2]SLAIS, University of British Columbia, Vancouver, British Columbia V6T 1Z3 'Canada
rkopak@interchange.ubc.ca
[3]Faculty of Information Studies, University of Toronto, Toronto, Ontario M5S 1A1 Canada
freund@fis.utoronto.ca

Abstract. Discussion of relevance has permeated the information science literature for the past 50+ years, and yet we are no closer to resolution of the matter. In this research we developed a set of measures to operationalize the dimensions underpinning Saracevic's manifestations of relevance. We used an existing data set collected from 48 participants who used a web search engine to complete four search tasks that represent four subject domains. From this study which had assessed multiple aspects of the search process – from cognitive to behavioural – we derived a set of measures for cognitive, motivational, situational, topical and system relevances. Using regression analysis, we demonstrate how the measures partially predict search success, and additionally use factor analysis to identify the underlying constructs of relevance. The results show that Saracevic's five manifestations may be merged into three types that represent the user, system and the task.

1 Introduction

Much has been written about the assessment of search output. Many models and frameworks for evaluation have been introduced, many measures suggested, and many solutions proposed. Nevertheless results (for better or worse) seem to be tied to traditional precision and recall measures based on some notion of relevance. While precision and recall are concrete, fully operationalized concepts, the underlying concept of relevance is aloof, by its very definition controversial and difficult to measure.

In common usage, *relevance* describes a relationship in which one thing has a direct bearing on another. There are two sources of ambiguity when this concept is used in information retrieval (IR). First, on what basis can we say that two things are directly related, and second, which two things are we relating? The traditional approach to relevance in IR uses features of the text as indicators of a relationship between queries and documents. In contrast, the information behaviour community claims that relevance is in the eye of the user, and is a subjective measure of the applicability of information to a user's need, problem, situation and context.

One of the few frameworks to offer a more precise interpretation of the various meanings of relevance is Saracevic's [18] "manifestations of relevance." However, his conceptual constructs remain largely unmeasured and few metrics have been devised and/or validated. In this research, we propose measures to operationalize each of Saracevic's manifestations, test each with a real life set of data, and examine the underlying dimensions of these manifestations as well as relationships among the measures.

2 Previous Work

Measuring search output begs the question: what is the purpose of the measurement? In the 1950s, initial developers of IR systems established the goal to retrieve relevant output [18]. Relevance has remained the holy grail of success for IR system research and retrieving only relevant information has also been the holy grail of IR system development; yet achieving relevance and knowing when relevance is achieved continues to challenge both researchers and developers to this day.

2.1 Relevance

Relevance has been debated, researched and reviewed, the most significant of these works being Saracevic [17], Schamber [19], and Mizzaro [15], and more recently, a thoughtful examination by Borlund [2]. For the past fifteen years, many proposals have emerged from the debate, from situational relevance [20], to psychological relevance [9] and task-based relevance [5], [16]. Mizzaro [15] proposed a time-driven logical relevance topology, while Saracevic [18] developed five manifestations of relevance.

2.2 Saracevic's Manifestations of Relevance

Saracevic's set of types (manifestations) of relevance was the first holistic approach to delineate the multi-dimensional nature of relevance, and remains the most comprehensive view, although others [2], [5] have suggested revisions. Each type outlined in this conceptual framework expresses a relationship betweens two elements of the triad: query, document, and user.

System/Algorithmic relevance is indicative of the similarity of a query, in essence its features, to a document. This type of relevance asks the question: how close is the fit between the retrieved set of documents and the user's query as determined by the algorithm? This is normally interpreted as an objective relevance – a comparison, but the questions that emerge are what and whose relevance?

Topical relevance indicates the 'aboutness' of a document. How close is the semantic fit between the query and the topics of the documents retrieved? Borlund [2] distinguishes this from algorithmic relevance by referring to it as "intellectual topicality;" a document may be assessed for aboutness independent of the query.

Cognitive relevance or Pertinence relates to how a document suits the "state of knowledge and cognitive information need of a user." What is the user's judgment about the applicability of the retrieved documents to the matter at hand? Saracevic suggests this is related to novelty, informativeness, information quality, and cognitive correspondence. Borlund [2] points out that pertinence is indicative of the dynamic nature of information needs – an item may be pertinent to a person at a point in time, but not necessarily pertinent to another person with the same problem, or indeed pertinent to the same person at a later (or earlier) point in time.

Motivational/Affective relevance relates to how a document corresponds with a user's intentions, goals and motives in seeking the information. It is also related to the user's emotional state and his or her perceived satisfaction and success with the task. Notably, this represents the human drive for information, and is likely inherent in other relevance types [2].

Situational relevance or Utility refers to the fit between the situation, problem space or task of the user and the documents output by the system. Do retrieved items allow the user to complete the task at hand? This form of relevance is driven by the context of users as well as their motivation for the information, and potentially affect other relevance types [5].

Of these types of relevance, System/Algorithmic and Topical relevances are of a lower order, closer to the system, while Cognitive/Pertinence, Situational and Motivational/Affective relevances are of a higher order of relevance, closer to the user. Underpinning all forms of relevance is the notion of interactivity. It is the relationships between the query, the document and/or the user that will determine relevance, rather than an assessment of any one of these in isolation. Rolled up in the concept of relevance then is the expectation that an IR system is capable of effective query processing and is able to deliver documents that are on the topic of the query, are pertinent to the user, leave the user satisfied, and enable task completion.

2.3 Measures of Search Outcome

While consensus may be emerging in defining the multi-dimensional nature of relevance, there is little consensus in how to measure it [2]. Precision and Recall have been used to measure search outcome since the initial experiments of Cleverdon [4]. The limitations of such an approach have been well documented and will not be addressed here (see for example, [10]). Yuan and Meadow [31] analyzed the use of measures in IR research using an approach analogous to co-citation searching. They examined the works of a set of authors from five research groups and found inconsistencies in the selection of measures, and coverage of the problem, demonstrating the confusion of measurement in this field. More recently, an

ongoing Delphi study [30] is attempting to reach a consensus on appropriate measures of online searching behaviour. To date, they have found that the search outcome measures ranked most highly by more than 50 researchers including: users' criteria for evaluating retrieved items, satisfaction with the search results, and utility/value of search results. Precision and recall measures have been ranked the lowest, not unlike the finding of Su [23].

One of the most significant and systematic attempts to measure the success of IR systems is that of Su [23] who compared 20 measures that included relevance, efficiency, utility, user satisfaction and overall success. She equated precision with relevance which was the generally accepted view of the day. In her study, participants were more concerned with recall than precision. This observation is not surprising given that the usual expected outcome during this period was to include *all* the documents on a topic. In addition measures such as satisfaction, user confidence, value of the search results, and user knowledge were more tightly correlated with the user's overall assessment of success. In this case success was user response to a scaled variable at the conclusion of the test. This study was conducted in an era when intermediaries conducted the search and the user paid for that service.

More recently, Greisdorf and Spink [8], [21] have attempted to map measures to Saracevic's relevance types, but these measures are subjective and potentially confounding. Users were asked whether or not retrieved items had met an information need based on five self-rated statements corresponding to the five types of relevance. No objective measures were used in the study.

A little know measure devised by Tague-Sutcliffe [25], Informativeness, determines the amount of information resulting from the interaction of a user and a document. Notable about this measure is that it captures the interactivity emphasized by Saracevic [18], the time element suggested by Mizzaro [15], and includes a system penalty when the system fails to deliver relevant documents (rather than the reward for success suggested by Vakkari and Sormunen [29]). Except for the work of Tague-Sutcliffe, this measure has languished (see one of the few applications in Tague-Sutcliffe and Toms [26]).

Many approaches to the measurement of relevance exist [3], [10] and the lack of a standard protocol for measurement impacts the conduct of research and development in this area. Having the ability to do systematic system comparisons has been missing in the interactive retrieval area as is evidenced by the TREC Interactive Track (see http://trec.nist.gov). Although the use of recall-precision measures has been strongly criticized, they have served the IR community for decades as a technique for making system comparisons. The long term objective of our work is to identify a parsimonious set of measures that may be used for research experimentation and by developers to assess system success. We place an emphasis on the notion of parsimony; although Schamber [19] devised a list of more than 80 criteria for assessing relevance, we believe that the essence of the problem is identifying the smallest set that will measure system success. Notably Barry and Schamber [1] in a comparative study found that two different contexts shared relevance criteria suggesting that the same criteria may be used in multiple contexts.

The objective of this study was to identify a set of measures for relevance using Saracevic's types of relevance as a framework for the work. The intent was to identify measures that could be interpreted as either subjective or objective, and could either explicitly or implicitly represent the essence of the relevance type. While some of the types such as Topical relevance are clearly understood with an operational definition that easily prescribes a probable measure, others such as Cognitive relevance have not been illuminated to the same degree. In addition, where warranted for each type of relevance, we wanted to explore the relationships among its measures to determine if a single measure could be used to represent that type. Finally, we were interested in the relationships among the types of relevance. For example, given our selected measures, do some relevance types predict others? Lastly, does each type represent an underlying construct?

Using previously collected data, we derived and tested relevance measures for each of Saracevic's types of relevance. While recognizing that searching is dynamic and interactive and some relevance judgements may change over the course of the search, we assessed search outcomes. System success in the context of IR systems can be determined by how successful users are in completing their tasks.

3 Methods

In 2001, we conducted an exploratory and experimental study of web searching behaviour in the context of the TREC 10 Interactive Track [28]. It was exploratory in that we collected a wide range of data (both

qualitative and quantitative, and both objective and subjective) to assess user cognitive and affective behaviours, and to examine how the search was conducted on a process and procedural level. It was experimental because we had several variables with multiple levels including: search tasks from four topical domains and two sources of the search task: researcher-specified or user-personalized. The intent of that work was to examine ways of improving the search process; for the work reported here we focus on our assessment of search outcomes. In this section we elaborate on the design of that study and explain how the data used to explore relevance was collected and analyzed.

3.1 System Used

For the research, we designed a custom search interface to access the Google search engine. The standard Google interface was modified to contain a longer search box of 200 characters with the Google directory categories displayed below. The screen contained the instructions: "please enter your search or select from the directory categories below." Beyond the first page, the standard Google interface screens were retained. The purpose of including the directory was to provide an alternative option – a scan capability – for the user. Choosing Google as the search engine was based on its status as the most widely used search engine.

3.2 Task

Sixteen tasks (which had been devised by the TREC 10 Interactive Track participants) were used in the study. The questions came from four domains: Consumer Health, Research, Travel and Shopping. Of the 16 tasks (four per domain), half were fully specified by the Track (e.g., “Tell me three categories of people who should or should not get a flu shot and why.”) and half could be personalized by participants who were instructed to specify an object or a topic based on their interests. Examples of these personalized tasks include “Name three features to consider in buying a(n) [name of product]” and “List two of the generally recommended treatments for [name of disease or condition].” The former are referenced as “researcher-specified” and the latter as “user-personalized” in subsequent discussions.

3.3 Participants

Participants were adult members of the general public who had used the web but who had not taken a professional online search course. Participants represent a sample of convenience; no formal sampling was done and participants self-selected, i.e., were volunteers. They were recruited through advertisement via printed posters posted on bulletin boards on campus, or in libraries and coffee shops in the downtown area, and via e-mail messages posted to listservs or e-notice boards at the research sites. Thirty-two were from Toronto and sixteen were from Vancouver.

The 29 women and 19 men ranged in age from 18 to 20 to over 65, with 71% per cent between 21 and 35 years old. Most had university level education, mainly at the bachelor (18) or masters (14) level, predominantly from the humanities or social sciences. About half were students; the remainder were from a diverse range of occupations. Almost all participants (94%) had been using the web for two or more years, and most were moderate web users. Overall, they were a relatively young, educated group who were experienced in terms of web use.

3.4 Procedure

The participants were recruited in August and September of 2001. Each participant was given four search tasks, one from each of the four subject domains. Of the four assigned tasks, two were research-specified and two could be personalized. We used a modified Latin square method to distribute the questions among the participants. Search tasks were given to participants one at a time. Because we were interested in the full range of searching and browsing behaviour of web searchers, participants were free to use either the search box or the directory categories for all tasks. In anticipation of questions regarding search syntax, we

printed a "cheat sheet" of basic search instructions for Google and placed it at the computer. Very few read it.

In each two-hour session, participants, first completed a demographic and web/search experience questionnaire and were assigned four search tasks in sequence. For each search task, they completed four steps as follows:

1) They completed pre-search questionnaire containing a scaled set of questions about their familiarity and expertise with the search topic.

2) They searched for the topic using the web interface. Participants were left uninterrupted for this part of the session. During this time, screen capture video recorded the search activity and a transaction log stored user actions. Participants were requested to print pages they believed useful to the task; these print commands were recorded in the transaction log along with other actions such as the query, categories selected and pages examined.

3) They responded to a post-search questionnaire containing a scaled set of questions about their perception of the search including their satisfaction with the results, the amount of time they had been assigned and their overall assessment of the results.

4) They participated in a semi-structured talk-after interview while reviewing the on-screen video of the search. In this part of the session, the screen capture video was re-played and paused while participants narrated the search process they had undertaken. Participants identified decisions, problems and issues with task completion. A series of probing questions were used to help the participant articulate the process.

When all search tasks were completed, participants participated in a short structured interview about the problems and challenges of searching the web.

Data were collected in the following ways:

a) on paper for questionnaire type data, namely the demographics and web/search experience, and the pre- and post-search questionnaires;

b) by audio tape recorder for talk-after interviews and the final interview; these were later transcribed;

c) using a transaction log to capture keystroke data; *WinWhatWhere* software captured the titles, URLs of all sites visited, and all keystrokes entered, and time stamped each action. *WinWhatWhere* is a 'spy' software that works at the operating system level (see http://www.winwhatwhere.com for more information).

d) using a screen capture application to capture all events on the screen, and thus to record the user process in a video form. *Lotus ScreenCam* was used for this aspect, and it is no longer being updated.

3.5 Data Analysis

Because of the myriad types of data, we first had to prepare the data for analysis. Some of this preparation was relatively straightforward such as transcribing the paper-based questionnaires into digital form, and some required more substantive actions such as the preparation required for the transaction log files.

First we cleaned the *WinWhatWhere* files, removing duplicate and esoteric data unrelated to our problem, to isolate selected actions including the queries and categories used, and to delimit the four task segments. Some of the data that we deemed important to our further analyses could not easily be identified in the cleaned files. For example, we wanted to determine how many different actions were used in the context of a search and how much time was spent in the results list. This type of process data could not be automatically identified in the logs, and thus we manually coded each participant session. To do this, we reviewed the activity on the ScreenCam video screen capture files to identify and/or verify the nature of recorded action in the *WinWhatWhere* files. We labeled each action from a pre-specified set of codes including: using a query, using a category, examining the results list, reviewing a URL selected from the results list, and viewing a URL selected from a link. In addition, for each website examined we noted its rank on results list, and the verbatim text of the query. This process resulted in a single file per user per task with time and date stamps, queries submitted, and coded actions. These logs were summarized by action within a participant's task to create measures such as Time-in-List (the amount of time spend scanning the results pages), Rank (the average rank of items identified as relevant by the participants), Not-on-List (the proportion of relevant pages examined that did not come from the results pages, and Modified Queries (number of queries used to complete the task).

In addition, we assessed the results – the pages indicated by participants as being useful – for each task using independent judges. Using both the paper printouts from the sessions and the URLs in the transaction log files, we first created a master list of all URLs declared relevant by participants, and saved a copy of that page (to capture the precise text that the user had viewed). Subsequently, each page was examined twice by two external judges. One assessed for aboutness, that is, does the page have anything to do with the search task. A second judge examined the *set* of all pages declared relevant to the task to assess task completeness. No comparison was made between participants to ascertain whether one set was better than another; each was assessed for its own merits. These measures are described in section 4.

Finally data from the pre- and post- search questionnaires and the demographics/experience survey data were combined with summary data from transaction logs, and the results analysis. This resulted in a set of over 80 variables. Data were analyzed using primarily SPSS univariate General Linear Model to assess differences by the experimental factors. In addition, Regression Analysis and Correlation were used to assess the relationships among the variables within relevance type, and among the relevance types. Factor Analysis was used to explore the underlying constructs of the resulting measures. The 'talk after' interviews (which are outside the scope of this paper) were coded using *Qualrus* qualitative analysis software (see http://www.qualrus.com for more information about this software).

4 Results

The first challenge in this research was to match appropriate measures from those we had collected, or create new measures from the collected data. This process was partially inductive and partially deductive: it required examining each variable to determine if it was an appropriate measure of a relevance type, and additionally, examining the each relevance type to determine which measures represent its underlying dimensions. We assumed that no single variable was likely to represent a single relevance type, although that proved incorrect given that we could identify only one measure each for Topical relevance and Situational relevance. In addition, none of these measures are binary; Schamber [19] recommended that binary judgments be avoided, and later Tang and Solomon [27] observed the need for more than two levels while Kekäläinen and Järvelin [13] found that graded relevance assessments more reliably identified the distinction among retrieval methods.

After the measures were selected, we assessed them against our data, examining relationships among the measures for each type of relevance. Next we assessed the relationships among the relevance types and finally analyzed the metrics as a set for underlying constructs. These variables were calculated per individual search session (192 in total).

4.1 Measures of Relevance

Table 1 summarizes the measures used to evaluate each type of relevance. For each measure, we provide a definition and identify the source of the data. Some are *objective*, derived from system observations of user search behaviour. Others are based on the *subjective* responses of participants to pre- and post-task questions, or based on the *external assessment* of expert judges. These measures are intended for post task assessment, rather than interval assessment of within-task behaviours/decisions. Although we agree with Mizarro [15] that relevance may change over time, we assessed the outcome – the success of the user in completing a task.

System/Algorithmic Relevance

Our intention was to identify an objective measure of the fit between the query and the system output, but this proved difficult. Technically, System/Algorithmic relevance reflects the degree to which the system representation of a document and the user initiated query terms match. But how do we assess the system's ability to do its job without confusing this form of relevance with Topical and Situational relevances? Saracevic [18] notes that system relevance is inferred mainly through comparison, which is also supported by Cosijn and Ingwersen [3]. Thus, we can compare one algorithm to another in terms of the efficiency or effectiveness of this matching process, but this is problematic for the evaluation of a single system. Furthermore, observing that one system achieves a different algorithmic relevance than another, that is,

produces a different match, says nothing about the merits of the difference. It cannot be objectively decided as there is no absolute benchmark by which to compare the outcome.

In TREC-style assessments based on an identified document collection, a set of tasks, and a set of relevance judgments, the "gold standard" is a set of documents based, a priori, on an external judge's assessment of what is relevant for a given query. In essence, due to the nature of the problem space, a very human problem space, the assessment is always determined by a user, or a surrogate of the user, i.e., an external assessor. Algorithmic relevance as defined by Saracevic [18] may be useful in conceptual discussions, but does not lend itself easily to operationalization. Borlund [2] too had difficulty providing an operational definition of this type, concluding that one could have vector space relevance or probabilistic relevance.

In this study we assume that Algorithmic relevance can be externally assessed. That is, at some point, a human being will decide how well the system does its job; an assumption cannot be made that the system has an appropriate Algorithmic relevance because the algorithm can make a match, irregardless of the quality of the match. Thus, we will do this indirectly using human selection or human workload. In essence, a system that makes a good match will highly rank the documents that are relevant to the query, and will reduce the user's effort. Implicitly, if the system cannot do both, then it cannot be argued that the system has made a good match, and thus has attained a high System/Algorithmic relevance. Much of past work focuses on System/Algorithmic relevance as being related specifically and only to the algorithm; this is an outdated perspective. Despite Saracevic's original definition, and subsequent discussions by many others, the *system* is much more than its algorithm; how results are displayed, how the system is enquired and so on are equally important.

Table 1: Measures for Saracevic's Manifestations of Relevance

Relevance Types [18]	Measure	Operational Definitions	Source of Data
System or Algorithmic Relevance	Rank	Average rank on the Google hitlist of all pages declared relevant by the user	System
	Not-on-List	Portion of relevant pages not on hitlist, but found through some other means	System
	Time-in-List	Time spent examining hitlists (in seconds)	System
Topical or Subject Relevance	Aboutness	Average of all pages examined per task on a scale of 1 to 5, as determined by independent coders	External Judgement
Cognitive Relevance or Pertinence	Certainty	Measured on a scale of 1 to 5; asked users, per task, how certain they were they had found an adequate answer	User
	Modified Queries	Number of queries used in the task	System
Motivational or Affective Relevance	Satisfaction	Measured on a scale of 1 to 5; asked users, per task, how satisfied they were with the search	User
	Ease of Use	Measured on a scale of 1 to 5; asked users, per task, how easy it was to do the search task	User
	Perceived Time	Measured on a scale of 1 to 5; asked users, per task, whether they had sufficient time to do the task	User
	Familiarity	Measured on a scale of 1 to 5; asked users, per task, how familiar they were with the topic of the search	User
Situational Relevance or Utility	Completeness	Measure of how complete the task was based on entire set of relevant webpages selected by participants	External Judgement

For System/Algorithmic relevance, we identified three measures:

1) *Rank*: Highly ranked documents are the ones determined by the system as the best match between the query and the document collection. If a user declares that items highly ranked are the relevant items, then one may conclude that the system is doing its job, and conversely, if the user selects items much lower on the list, then the system is not doing its job. Thus, Rank, the average rank of all items declared relevant, is indicative of this type of relevance.

2) *Not-in-list*: If the relevant hits are not on the results list but are found on secondary pages, then similarly the system has not done its job. Relevant hits should be listed on the results page. Not-on-list is thus the proportion of relevant pages acquired through some other means.

3) *Time-in-list*: This measure is indicative of the amount of effort that it takes a user to scan the results list to select a relevant item. If the user must take a considerable amount of time to select an item, then the system is also not doing its job. This may be due to poor ranking of the relevant documents and/or poor representation of the documents (relevant or not) on the results page. In our view, System/Algorithmic relevance is not just about the ranking algorithm, a traditional view of IR evaluation; it is equally about how that system is represented to the user.

These three measures are implicit measures of relevance. That is, none directly measure System/Algorithmic Relevance, but serve as proxies of this type of relevance. On average, participants found relevant items half-way down the results page (n=192, x=4.4, SD=2.8296) and found a small proportion of the relevant items elsewhere (n=188, x=0.3, SD=0.385) elsewhere on the Web, through hyperlinks from the site that appeared on the results list. In addition, they spent a couple of minutes reviewing the results pages (n=192, x=128.31, SD=193.3) per task.

Topical Relevance

Topical relevance also called subject relevance reflects *aboutness*, a generally agreed upon interpretation [18], [15], [5]. While one could assess aboutness independently of the query [2], in the development of this measure, we considered how well the topics in the document matched the topic represented by the query. In this case, external assessors examined the 395 pages declared relevant by our participants to assign a value as illustrated in table 2. No additional measures emerged from our data collection to either implicitly or explicitly represent this relevance type. Overall, the documents printed by participants were rated highly for *aboutness* (n=181, x=4.54, SD=0.866).

Table 2. Aboutness Measure

Code	Definition
5	pages directly related to the topic and containing clear info on the topic,
4	pages that provide some information that is related, or leads directly to the answer
3	pages that about the topic but may be broader or narrower that the topic
2	tangentially related but not really in the topic area
1	pages that are clearly not about the topic at all

Cognitive Relevance or Pertinence

Cognitive relevance is the most poorly defined of the five relevance types. Cosijn and Ingwersen [5] suggest a wide range of measures for this, noting that it is highly subjective and personal. In our work, we consider this form of relevance as the *opposite* of cognitive dissonance, the psychological conflict within the individual caused by inconsistencies between belief and behaviour. Therefore, cognitive correspondence is achieved when there are congruities between the searchers' initial queries and the search results.

In this case, we interpreted cognitive relevance as the *certainty* with which participants felt that they had done a good job – a perception of a good match, and a perception of personal success. From a user perspective, a strong measure of certainty may be equated with a perception of overall success. Secondly, we noted the number of times users felt it necessary to modify a query (*modified queries*) as a signal of a probable mismatch.

Participants reported being fairly certain that they had found adequate information to satisfy their queries (n=191, x=3.9, SD=1.069) and modified a small number of queries (n=168, x=0.9, SD=1.428). Pearson's Correlation Coefficient was used to examine the relationships between these two variables. Certainty was

negatively correlated with the number of modified queries (R^2=-0.167, p<0.05). The more queries created resulted in a lower degree of certainty.

Motivational or Affective Relevance

As defined by Saracevic [18], motivational relevance deals with intentions and goals and as such is an *a priori* construct that potentially changes over the course of doing a search. This too is subjective and personal [5]. *Familiarity*, the degree to which the topic matter is known to the user (or prior knowledge) can be a powerfully influential force in affecting both motivations and intentions. Affective behaviours, on the other hand, may change over the course of a search, but their state at the end of a search may be related to the cognitive state of the user at the end of the search activity. At the end of each search task, participants indicated the *ease* with which the task was accomplished, the suitability of the amount of time – the *timeframe* – assigned to do the task, and their *satisfaction* with the task. This relevance type contains measures of both pre-search and post-search behaviours.

Participants similarly rated their levels of satisfaction (n=191, x=3.63, SD =1.121), ease of use (n=191, x=3.7, SD =1.139), and perceived time frame (n=191, x=3.47, SD =1.06) for the search tasks. It is, therefore, not surprising that all three variables are significantly correlated: satisfaction and ease of use (R^2=0.76, p<0.001), ease of use and timeframe (R^2=0.527, p<0.001), and satisfaction and timeframe (R^2=0.523, p<0.001).

Situational Relevance or Utility

Situational relevance is a context specific dimension that examines the fit between the documents retrieved and the task. In this study, this perceived fit was artificial in that the tasks were not personal to the participant, and participants did not have to process the information post the search. To assess this aspect, we created a measure called *completeness* which was an expert assessment of the proportion of the task that could be completed with the set of documents declared relevant by the participant. Spink, Greisdorf and Bateman [22] found that partially relevant documents added new knowledge to users' understanding of their problem. Thus all pages identified by a participant as useful to the search task were included in the set, irregardless of the document's aboutness rating, as partially relevant document may contribute some aspect to task completion. The same set of webpages examined for aboutness was re-evaluated. In this case, for each participant task, the *set* of documents identified as relevant was assessed. When taken together as a set, how much of the task could be completed? The scale is illustrated in Table 3.

Table 3. Completeness Measure

Code	Definition
5	100% of the problem has been answered.
4	about 75% of the problem has been answered/responded to
3	about 50% of the problem has been answered/responded to
2	about 25% of the problem has been answered/responded to
1	0% of the problem has been answered/responded to

The selected web pages retrieved by users received high ratings (n=176, x=4.39, SD =1.2) according to completeness. In other words, on average, participants selected pages that could be used to satisfy at least 75% of the assigned task.

4.3 Predicting Success in IR Systems

In this section, we examine relationships among the identified measures using multiple regression. While the measures can be defined in terms of the types of relevance and represent underlying dimensions of each type, we wondered how much these measures contributed to search success. Success is an elusive construct in IR evaluation. Among our set of measures, we hypothesized that users in our study would declare success according to certainty – the degree to which they feel they had achieved an appropriate response, a measure of Cognitive relevance. Because of the significant correlations among the various measures used for each relevance type, one measure was selected to represent the types with multiple measures. Thus,

rank and satisfaction were selected for System/Algorithmic relevance and Motivational/Affective relevance respectively. Two of the relevances, Topical and Situational had only a single measure: aboutness and completeness, respectively. These four measures were found to significantly predict certainty ($F(4, 146)=25.077$, $p<0.001$), and to explain 39% of the variability in certainty, or as we consider it – success.

Among these five variables, two are responses from the search engine to users' queries: rank and aboutness, and three are based on human judgments: certainty, completeness, and satisfaction. We were interested in the interplay between these two general types of variables. In essence, do user oriented measures predict the system measures, or vice versa? We regressed average page rank and average aboutness score with certainty ($F(2,151)=4.626$, $p<0.05$), completeness ($F(2,149)=67.137$, $p<0.001$), and satisfaction ($F(2,151)=4.081$, $p<0.05$) to triangulate the system and user constructs used to measure relevance. Although significant, they account for only 4.5 to 5% of the variability in certainty. With such a low percentage, we wondered if the relationships among these variables represented a different set of constructs other than the initial five types of relevance.

4.4 Identifying Components of Relevance

Because of the mixed results in looking at the five types of relevance as potential predictors of success, we used factor analysis to ascertain which of the measures might form coherent groupings that are relatively independent of one another (see [24] for an excellent explanation of factor analysis). Were there any potential underlying factors of relevance that were not evident in our previous analyses?

All measures listed in Table 1 were loaded initially, but *familiarity* and *time-in-list* were removed because they were poorly correlated with the other measures. The Kaiser Meyer Olkin measure (.737) and the Bartlett's Test of Sphericity ($\chi 2=429.627$, df=28, $p< .0001$) indicated that the sample was adequate and that the measures were likely to be related. These two tests are conducted to determine if factor analysis is an appropriate technique for this data set.

Factor analysis was conducted using principal components analysis as the method of extraction and varimax as the method of rotation. Principal components analysis looks at linear combinations of variables. The first combination tends to account for the largest amount of variation, the second and subsequent contain successively smaller portions of the total variance, and additionally are independent of one another. Varimax (an orthogonal rotation that results in factors that are uncorrelated) is used to ensure that the resulting factors are interpretable. In essence, do the variables that load together strongly have an identifiable construct? Potentially there can be as many components or factors as there are variables. But that was not the case in this analysis which resulted in three components or factors (see Table 4). In addition, these three factors have internal coherence – they have clearly identifiable meanings. *Ease*, *satisfaction*, *certainty* and *timeframe* are user perception dimensions, indicative of cognitive or user type of relevance. The second factor, *aboutness* and *completeness*, are dimensions of the task relevance, while the third factor contains *not-on-list* and *rank*, dimensions of the system relevance

Table 4. Factor Loadings

	Factors		
Measure	**User Perception**	**Task**	**System**
Ease	.842		
Satisfaction	.829		
Certainty	.809		
TimeFrame	.731		
Aboutness		.753	
Completeness		.746	
Not-on-List			-.787
Rank			.706

In this analysis, the communality values were high – all greater than .70; all variables loading on the three factors together account for 72% of the variance. All three factors had eigenvalues greater than 1; the factor loadings illustrated in Table 4 indicate the correlations of the variables with the factors. All measures

for all factors correlated at greater than .70. These relationships are represented in Figure 1 which has been modified for a two-dimensional presentation.

As illustrated in Figure 1, the first factor (on the extreme right in Figure 1) includes measures of user perceptions and as a set represents variables from both Cognitive and Motivational relevance as defined by Saracevic [18]. The second factor (at the top) includes measures from Topicality and Situational relevance that results in an intersection between the *aboutness* of the document and the task for which it will be used. The third factor (on the lower left) concerns the System/Algorithmic relevance. To summarize, there are three underlying factors represented by the set of measures that were previously identified to fit the five types of relevance. These factors may be interpreted as representing the user (cognition and motivation), the task, and the system.

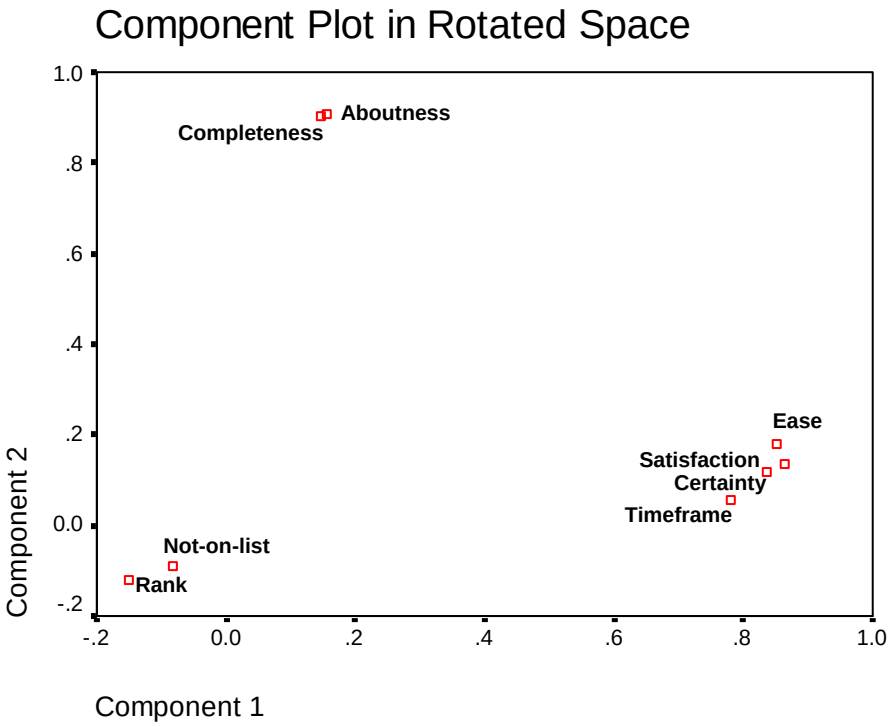

Fig. 1. Relationship between Measure and Factors in a 2-D space

5 Discussion

5.1 Measures for Relevance Types

This research examined the relevance problem, operationalizing the relevance types previously defined by Saracevic [18], and identifying one or more measures for each of the types. Notably, the measures were derived from a holistic study that included 16 web search tasks performed by 48 users; but, the choice of measures was informed primarily by Sarecevic's conceptual framework. Where more than one measure existed for a single type, all of these measures were strongly correlated within that type. While the findings for this particular data set have the usual limitations (e.g., no replication), the relationships among the variables and the underlying factors that emerged from the analysis are noteworthy.

Challenging in the selection of measures was System/Algorithmic relevance. While this type is clearly defined and that definition is widely accepted – the similarity of a query to a document, the definition has to date only been operationalized in TREC style comparative studies that are unable to determine

definitively that a system delivers relevant documents; it can only state that system A delivers more relevant documents than system B. Furthermore in web-based studies the notion of precision and recall are incalculable except in an arbitrary way. With the 16 search tasks used in this study, participants often received a 'set' containing thousands if not millions of documents. Balance this point with the fact that people examine on average 1.8 pages of references (as found in the recent PEW studies (http://www.pewinternet.org)); the system designer's notion of a document set is at odds with user perception. Is precision, therefore, to be measured for the set defined by the algorithm or that perceived by the user? TREC studies may calculate precision-recall values for hundreds to thousands of documents, which is useful for comparing systems on a theoretical level, but has no basis in real world activity. An IR system services human activity and, its human users becomes the assessor of its quality and its success. Thus, for this relevance type, we used measures that implicitly evaluate System/Algorithm relevance. As a final footnote to this type, we believe that the definition of System/Algorithm relevance needs expansion. While the system may provide a good match between a query and a document, a user still may not be able to identify the relevant document because of many other characteristics of the system such as how the document set is presented to the user, how the system is queried and so on.

Because Topical relevance has been in use for so long, the choice of a measure, *Aboutness*, was almost self-evident. We wondered what additional sorts of measures might service this type, but like others (e.g., [5], we did not find additional measures.

Cognitive relevance or Pertinence is a multi-dimensional type of relevance, and has been defined as a series of qualities from informativeness to novelty. The one consistent quality that seems to be in general agreement is cognitive correspondence, which we interpret as the opposite of cognitive dissonance. Motivational/Affective relevance, on the other hand, is much more clearly understood although with some dispute as to whether it is a mutually exclusive relevance type (see the argument put forth by Borlund [2]) with regard to the other types. Both of these types of relevance stem from the user, and as seen from our results share the same underlying construct and thus are related.

In the context of our study, a laboratory experiment, Situational relevance was more difficult to assess. The tasks were assigned tasks, although half could be personalized by participants. This form of relevance is generally interpreted as a post system-interaction assessment. Are the results useful in decision-making? Does it reduce uncertainty? Are the results useful? In this case we looked at task completion from an external assessor perspective which was a good surrogate measure for this relevance type. Additional measures are conceivable in non-laboratory studies that see the work task to completion.

5.2 Examining Relationship among the Measures and Relevance Types

Using the 'standard bearer' measure – rank, aboutness, satisfaction, completeness and certainty – defined for each relevance type, we explored the relationships among the five types. Some of the measures are systems-oriented and some user-oriented; we compared the systems-oriented ones with the user-oriented measures. Interestingly, the System/Algorithmic and Topical relevance measures – aboutness and rank – predict measures for the user specified relevances: Motivational, Situational and Cognitive. This was unexpected, as it is at odds with the current belief concerning system relevance – that relevance is human-driven (see [2], [5], [18], [19]). However, the contribution to variability in those user-specified measures (certainty, satisfaction and completeness) was small, leading us to conclude that success at the systems level is not sufficient to predict success at the user level.

In addition, we proposed that certainty from the user's perspective is the ultimate goal not unlike the success variable used by Su [23]. We may hypothetically have perfect relevance matches in all types, but if the user is uncertain about the results, then appropriate matches either have not been made or have not clearly communicated. Our results demonstrated that the System, Motivational, Topical and Situational relevances predict the Cognitive. Thus a person's level of success with a search – the certainty with which participants believed that had a good response is determined by the satisfaction with which they performed the search, the aboutness of the documents retrieved, the average rank of useful items, and the completeness of the task. Troubling about this finding is that only 39% of the variability in certainty could be explained by these four variables. Notably this is significantly higher than the relationship between systems-oriented and user-oriented measures discussed above. This finding though not ideal brings us a step closer to defining a parsimonious set of measures for relevance, and in particular for measuring system success.

5.3 Re-examining Relevance Types

Initially we examined associations among measures that reflect relationships between any two of query, document and user. These were founded on the five pre-defined relationships described by Saracevic's [18] relevance types. However, once we explored the associations amongst the measures unencumbered by the relevance types, a different pattern became apparent. Rather than five underlying constructs, three emerged from our data: system, task and user. On further inspection, we concluded that Saracevic's original relevance types are heavily oriented toward the user, while Mizzaro's [15] typology is more heavily weighted toward the system; the outcome from our work provides a more balanced blend of the two.

System: it is not surprising that the measures we used – *Rank* and *Not-on-list* – emerged as a single construct, considering our earlier discussion of System/Algorithmic relevance. The definition for this type does not change substantially. It remains a match between query and document according to the system's ability to highly rank useful items.

Task: this was not the case for Topical and Situational relevances; that *Aboutness* and *Completeness* would form a single construct was unexpected. While both were assessed by external judges, the judges for each measure differed. The first measure is used at the level of an individual document while the second is based on a document set. One is a match between query and document while the second is a match between document and the work task. The merging of these two measures suggests a (work) task relevance type. Although task is often separated from situation in discussions of relevance, a situation dictates a task, and a situation may require multiple tasks, each of which in turn may require multiple search tasks. In the case of our study for example, few of the tasks could be handled with a single query. Conceivably, task is unlikely to be mutually exclusive from situation, and thus will inherit many of its characteristics from the situation. Our finding is not unlike the fourth dimension of Mizzaro [15]'s model which contains topic, task and context components, and supports earlier work [16]. Situation, it could be argued, will impact not only the task, but other types of relevance as well [5]. Thus, this type of relevance becomes a match between the documents and the task including both a topical match, and task completion.

User: Cognitive and Motivational relevances form a single construct representing multiple characteristics of the user. Borlund [2] suggests that Motivational relevance is not independent of other types, and in particular of other subjective types; Cosijn and Ingwersen [5] isolate affect as a time-based dimension. Both groups had concerns with these two separate but clearly interrelated types of relevance. Mizzaro [15] has no user-specific component in his model. However, creating a type that combines both Cognitive and Motivational/Affective relevances is not atypical; it corresponds to the 'ABCs' of cognitive psychology: Affect, Behaviour and Cognition. Thus, this type of relevance is a holistic one that includes multiple user dimensions.

Of particular note in this work is the contribution to our understanding of the underlying constructs within the sea of potential measures of relevance. That our probing of relevance would reveal that user, task and system relevances emerge from all of the measures we employed is on one hand unexpected, and on the other, predictable. Our findings are clearly in line with Ingwersen's [11] cognitive model of interactive IR (although interestingly this model diverges from his relevance model [5]); He proposed that interactive IR contained three elements: systems, users and the environment. Similarly Borlund [2] identifies a process that includes a user level, a system level and a surface level. These two frameworks are closely aligned with the one that emerged from our data. Relevance can be defined in terms of three components each with its own dimensions and measures: user, system and task. Although there is much yet to explore concerning these three constructs, together they form a much simpler model than that of Saracevic and Mizzaro.

7 Conclusion

The findings from our study are not unlike that of Delone and McLean [6], [7] who examined many studies of systems evaluation in business to develop a model that predicts systems success. In their model, system, information and data quality affect system use and user satisfaction which, in turn, affect the net benefits of the system. We have not yet explored our data to examine the multiple effects as presented by Delone and McLean.

Another form of evaluation which is rarely mentioned in the context of IR systems is that of usability, a concept well-known in human-computer interaction and often used in the assessment of interfaces.

Usability, as defined by the International Standards Organization, is the "the effectiveness, efficiency and satisfaction with which a specified set of users can achieve a specified set of tasks in particular environments" [12]. Like relevance, it too suffers from an abundance of potential measures with which to assess its underlying dimensions. Usability contains the concepts of efficiency and effectiveness discussed by Borlund [2]. While efficiency and effectiveness are directly related to use as described by Delone and McLean [6] [7], and task as referenced by the relevance community (see for example, [5], [16]), satisfaction represents affective and cognitive behaviours, that tend to be examined by the information behaviour community [1], [19]. Both the Delone and McLean success model and usability would be fruitful directions to explore in our quest for measuring IR system success.

Our research has demonstrated that a combination of system, user and task measures indicate the outcome of the search. These findings fit with the "interactive framework" [18], [2], [5] within which all relevance types operate, and additionally form a more parsimonious set of relevance types. Our work also points to the importance of including subjective measures in investigations of relevance balanced with quantifiable, tangible metrics. While we achieved some success in identifying useful measures of various relevance types, future work will entail testing and validating these measures as well as the relevance types.

6 Acknowledgements

Work funded by grants to the first author from the Social Sciences and Humanities Council (Canada), the Natural Sciences and Engineering Research Council (Canada), and the Canada Research Chairs Program. Thanks to research assistants Joan Bartlett, Ariel Lebowitz and Louanna Mootoo who worked on the data collection and analysis, and to the three anonymous reviewers whose very helpful suggestions and comments have greatly improved the final manuscript.

References

1. Barry, C.L., Schamber, L.: Users' criteria for relevance evaluation: A cross-situational comparison. INFORM PROCESS MANAG. 34 (1998) 219-236
2. Borlund, P.: The concept of relevance in IR. J AM SOC INFORM SCI. 54(10) (2003) 913-925
3. Borlund, P., Ingwersen, P.: The development of a method for the evaluation of interactive information retrieval systems. J DOC. 53(3) (1997) 225-250
4. Cleverdon, C.W.: Information and its retrieval. ASLIB Proc. 22 (1960) 538-549
5. Cosijn, E., Ingwersen, P: Dimensions of relevance. INFORM PROCESS MANAG. 36 (2000) 533-550
6. DeLone, W.H., McLean, E.R.: Information systems success: the quest for the dependent variable. INFORM SYST RES. 3(1) (1992) 60-95
7. Delone, W.H, McLean, E.R.: Information systems success revisited. 35th HICSS Proc. (2002)
8. Greisdorf, H.: Relevance thresholds: a multi-stage predictive model of how users evaluate information. INFORM PROCESS MANAG. 39(3) (2003) 403-423
9. Harter, S.: Psychological relevance and information science. J AM SOC INFORM SCI. 43 (1992) 602-615
10. Harter, S. P., Hert, C.A.: Evaluation of information retrieval systems: Approaches, issues, and methods. ANNU REV INFORM SCI. 32 (1997) 3-94
11. Ingwersen, P.: Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. J DOC 52(1) (1996) 3-50
12. ISO.: Ergonomic requirements for office work with visual display terminals (VDTs): Part 11.Guidance on usability. ISO 9241-11-1998 (1998)
13. Kekäläinen, J. Järvelin, K:. Using graded relevance assessments in IR evaluation. J AM SOC INFORM SCI. 53(13) (2002) 1120-1129
14. Maglaughlin, K.L., Sonnenwald, D.H.: User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. J AM SOC INFORM SCI. 53(5) (2002) 327-342
15. Mizzaro, S.: How many relevances in information retrieval? INTERACT COMPUT. 10 (3) (1998) 303-320
16. Reid, J. A new task-oriented paradigm for information retrieval: implications for evaluation of information retrieval systems. CoLIS3 Proc. (1999) 97-108
17. Saracevic, T.: Relevance: a review of and a framework for the thinking on the notionsin information science. J AM SOC INFORM SCI. 26 (1975) 321-343
18. Saracevic, T.: Relevance reconsidered. CoLIS Proc. 2 (1996) 201-218

19. Schamber, L.: Relevance and information behavior. ARIST (1994) 3-48
20. Schamber, L. Eisenberg, M.B. Nilan, M.S.: A re-examination of relevance: toward a dynamic, situational definition. INFORM PROCESS MANAG. 26 (1990) 755-775
21. Spink, A., Greisdorf, H.: Regions and levels: Measuring and mapping users' relevance judgments. J AM SOC INFORM SCI. 52(2) (2001) 161-173
22. Spink, A., Greisdorf, H. Bateman, J.: From highly relevant to not relevant: examining different regions of relevance INFORM PROCESS MANAG. 34 (1998) 599-621
23. Su, L.T.: Evaluation measures for interactive information retrieval. INFORM PROCESS MANAG. 28(4) (1992) 503-516
24. Tabachnick, B.G. Fidell, L.S:. Using multivariate statistics, 4th ed. Allyn & Bacon (2001)
25. Tague-Sutcliffe, J.: Measuring Information. Academic Press, New York (1995)
26. Tague-Sutcliffe, J., Toms, E.G.: Information systems design via the quantitative analysis of user transaction logs. Presented at the 5th ICSI, River Forest, Illinois (1995).
27. Tang, R. Solomon, P.: Towards an understanding of the dynamics of relevance judgments: an analysis of one person's search behavior. INFORM PROCESS MANAG 34 (1998) 237-256.
28. Toms, E.G., Freund, L., Kopak, R., Bartlett, J.C.: The effect of task domain on search. CASCON, IBM, Toronto (2003) 303-312
29. Vakkari, P., Sormunen, E.: The influence of relevance levels on the effectiveness of interactive information retrieval. J AM SOC INFORM SCI. 55 (11) (2004) 963-969
30. Wildemuth, B.M., Barry, C., Luo, L., Oh, S.: Establishing a research agenda for studies of online search behaviors: a Delphi sStudy (2004). (see http://ils.unc.edu/sig_use_delphi/ for details of the study, and preliminary reports).
31. Yuan, W., Meadow, C.T.: A study of the use of variables in information retrieval user studies. J AM SOC INFORM SCI. 50 (1999) 140-150