

Improving Learning Object Schemas and Performance Support Systems through Information Retrieval Evaluation

Steven Craig Schatz
University of Hartford
Schatz@hartford.edu

Abstract

This study examines existent and new methods for evaluating the success of information retrieval systems. The theory underlying current methods is not robust enough to allow testing retrieval using different meta tagging schemas. Traditional measures rely on judgments of whether a document is relevant to a particular question. A good system returns all the relevant documents and no extraneous documents. There is a rich literature questioning the efficacy of relevance judgments. Such questions as: Relevant to who? When? To what purpose? are not well answered in traditional theory.

In this study, two new measures (Spink's Information Need and Cooper's Utility) are used in evaluating two search tools (a tag-based and a text based), comparing these new measures with traditional measures and each other. Thirty-four educators searched for information using both search engines and evaluated the information retrieved by each. Construct measures, derived by multiplying each of the three measures (traditional, information need, and utility) by a rating of satisfaction were compared using two way analysis of variance.

Results indicated that there was a significant correlation between the three measures – so the new measures provided an equivalent method of evaluating systems and have some significant advantages – including no need for relevance judgments and easy application in situ. While the main focus of the study was on the methods of evaluation, the evaluation in this case showed that the text system was better than the tag based system.

Improving Learning Object Schemas and Performance Support Systems through Information Retrieval Evaluation

Situating the Research: Human Performance Technology and Performance Support Systems

A robust sub domain of study within instructional systems is human performance technology (HPT). HPT focuses on terminal performance as its unit of measure and explores three areas to seek leverage in improving performance – information, instrumentation, and motivation. (Gilbert, 1996) To engineer performance, one strives to use strategies that have the greatest impact for the least cost. Accordingly, one may have a greater impact on improved performance by (for example) providing better information during the course of performance rather than designing a training intervention. Rossett (Rossett, 1996) distinguishes between training interventions and performance support by identifying the goal of training as building capacity, occurring before a need arises. Performance interventions, in contrast, provide support at the point of need (both in time and in place).

Information provided just in time offers a powerful tool for improving outcomes. Computer-based systems that can provide just in time information have been investigated since the 1980s. This class of interventions has variously been called electronic performance support systems (EPSS), performance support systems (PSS), and performance support tools (of which a performance portal is a subset). One may view these systems on a continuum of how embedded they are within the work they support. For example, a highly embedded or intrinsic EPSS is the Turbo Tax software, where the system asks a series of questions and does a series of calculations to help the user complete a tax form. It is possible in this case to fill in the forms without leaving the performance support system. In contrast, more extrinsic PSS are “help” functions. Performance Portals, which provide a single point of access to training, information, and support tools, are completely extrinsic, stand-alone systems., EPSSs are, “part online help, part online tutorial, part database, part application program, and part expert system...EPSSs quickly and easily provide answers to the questions workers have when performing a job, and address workers’ concerns.” (Carliner, 2002 p. 400)

While there is a great diversity of functionalities across EPSS applications, one of the most common functions is a database of information that may be searched. In more embedded systems, this search is less obvious, often completed by the system based on context. However, in the more extrinsic systems, the search/retrieval function is an obvious and essential part of the system. “The primary design goal of an EPSS is that the knowledge it contains be easily retrievable by the users at the time they need it.” (Cole, Fischer, & Saltzman, 1997 p. 50) A survey of readers of CBT Solutions magazine conducted in 1996 listed “searchable

reference” as the most common feature of EPSSs either built or being built for their companies. (Benson, 1997).

However, while the search/retrieval function or system has been shown to be an important function of performance systems, there is little known about within the field of instructional technology about evaluation of information retrieval systems. This is a problem, as there have been ongoing attempts to improve search systems. In 1992, Carr wrote, “Despite years of work, our methods for retrieving information from data bases remain relatively rigid and primitive” (Carr, 1992 p. 35). Much of the work concerning meta tags (Dublin core, IMS, SCORM, GEM and more) is predicated on the idea that adding meta tags can improve retrieval systems. Unfortunately, without being able to measure the effect of different search/retrieval systems, there is no way to know if efforts to improve practice are yielding results. The goal of this research is to see how the constructs and measures of information retrieval evaluation may inform the research agenda of performance support and if new constructs and measures may enrich and expand the theory.

Evaluation of information retrieval systems (whose practitioners are usually located in schools of library and information science) is a relatively mature field, with a literature dating to the mid 1960's (further if one considers pre-computer methods). While this literature offers a rich foundation, the constructs and measures underlying the theory are thin. For one wishing to evaluate information retrieval within performance support systems, additional constructs and measures that reflect the unique needs of performance support are necessary.

The traditional method for evaluating information retrieval systems relies on the relevance based measures, recall and precision. To accomplish this type of evaluation requires:

- A collection of documents

- A collection of questions (queries) to be asked of the document collection

- A set of judgments of which documents are relevant to each question.

To evaluate the system, one queries the document collection (ask the questions and see what documents are retrieved) and calculates recall and precision. Recall is the measure of how many relevant documents were actually retrieved. For example, if one question was 'How many angels can dance on the head of a pin?' and upon studying the document collection, it was judged that there were 50 documents that were relevant to that question, and a system retrieved 20 of those documents when it searched for 'How many angels...', then the recall rating for that system would be the number of relevant documents retrieved divided by the total number of relevant documents expressed as a percentage. (In this case 20/50 or 40%.) Precision is the measure of

how many relevant documents were retrieved divided by the total number of retrieved documents. It is a measure of how “noisy” the results are. It would be easy to retrieve the 50 relevant documents – just retrieve all the documents in the collection. A system that retrieved 30 documents of which 20 were relevant to the question is of more value than a system that retrieves 100 documents of which 30 are relevant. While the recall of the second system is higher (60% vs. 40%), the precision is lower (System 1 precision = $20/30 = 66\%$ and System 2 precision = $30/100 = 30\%$). Both measures are reported in the traditional evaluation paradigm. They are generally assumed to be inversely proportional – the better the recall (the more relevant items retrieved) the lower the precision (the “noisier the result”).

The traditional method of evaluation is system-centric. It is possible to test several information retrieval systems using a batch mode – using a program that runs all the queries against a system and calculates the recall and precision of the system. There is no consideration of users. The focus is on the system.

Statement of Traditional theory

While the theory underlying information retrieval is rather thin, it has been the basis of hundreds of research papers. The theoretical constructs, measures, relationships between constructs and assumptions of relevance-based evaluation are:

Assumptions

- Relevance judgments are valid and meaningful indicators of the effectiveness of an information retrieval system. (Harter & Hert, 1997)
- Relevance is a meaningful construct on which to base the evaluation process.
- Systems can be evaluated using the combined measures of recall and precision.
- A successful retrieval system retrieves the greatest number of relevant documents with the least extraneous documents.

Constructs

- Relevance – a judgment of whether a particular object answers a particular question.
- Precision – how many extraneous or non-relevant documents are retrieved for a particular question by a particular system. A measure of accuracy or quality of retrieval.
- Recall – how many documents drawn from a particular set of documents relevant to a particular question are retrieved with a particular system. A measure of scope or quantity of retrieval.

Relationships between Constructs

- Recall and Precision have an inverse relationship. Generally, the greater the recall (number of relevant documents retrieved), the lower the precision (more noise).

Measures

- Relevance judgments can be made through a variety of methods, including automatic text matching and expert judgments.
- Recall is measured by dividing the number of documents that have been judged relevant to a particular query in a document set by the number of relevant documents retrieved for that query.
- Precision is measured by dividing the number of documents retrieved for a query by the number of relevant documents retrieved.

It is an indication of the state of the theory of evaluation that recall, precision, and relevance are both constructs and measures. This encourages confusion. To provide an example of the confusion inherent in these discussions, we may look at the relevance construct. Relevance is a construct that underlies the way the measures recall and precision are calculated. However, there are fundamental concerns about the relevance *measure*. All the discussions that follow detailing concerns with relevance judgments are actually concerns about the relevance *measure* (specifically, how and who can decide if a document is relevant to a question for a person). These concerns, then affect the interpretation of recall and precision *constructs*. There is not a problem with the *measures* of recall and precision, the actual formulas make sense. However, the constructs are brought into question because of questions surrounding the relevance measures.

There are additional problems with the theory of evaluation when it is used to evaluate performance systems. “The main assumption behind the use of measures (constructs) such as recall and precision is that the average user is interested in retrieving large amounts of relevant materials..., while at the same time rejecting a large proportion of the extraneous items ...” (Salton, 1992 p. 442). However, for a performance system that has the goal of providing a fast, specific answer to a problem, the best result is a small number of items with high precision. This is a fundamental difference between an information retrieval system (like a web search engine) and a performance support system.

Research questions

The current state of theory for the evaluation of information retrieval systems is not rich enough to easily map to performance systems. Some have said it does not map all that well to information retrieval systems (Cooper, 1981; Harter, 1996; Harter & Hert, 1997; Schamber, 1994). In this study, I have used a case study to examine two new constructs and their measures as well as the traditional constructs and their measures to attempt to obtain a richer view of comparative evaluation of two systems. Specifically, the following questions are addressed.

- 1) Can a richer comparative evaluation be provided by multiple constructs of success?

2) Can user-centric constructs provide time and cost effective comparative evaluation of two systems?

3) Do non-relevance based constructs, specifically change in information need and Cooper's utility, provide useful insight into comparative evaluation?

Background Information

While information retrieval is a necessary part of design and development of a performance support system, it is not an area of research common to instructional systems. So, a bit of background in the field of information retrieval is necessary.

What are search engines

A search engine is an information retrieval system. The most common engines most of us use are web-based systems, such as Google or Yahoo or Excite. These search engines search the web, which is an information or document set that is large and constantly growing. The growth of the web is not controlled by the search engines or system designers. This is in contrast to information systems such as ERIC or Lexus/Nexus that have documents or document summaries that have been added into a database by someone associated with the service. While the web is often referred to as an entity, it is an uncontrolled source. Services such as ERIC or Lexus/Nexus are controlled – the contents and access are controlled by systems designers.

An important part of a web search systems have is a spider or crawler. This is an automated program that goes out on the web, following links and reading web pages. Controlled systems do not find their document set in this manner; they are added to the systems by the system maintainers based on the system's set of criteria.

When the crawler or a controlled system reads a document, the system then breaks the documents into words or phrases and constructs an index. This index contains a list of vocabulary words and a list of all documents that contain that word (or phrase). So, when you search (for example for "horse"), the system is not actually searching documents, it is searching its own index. This is much faster.

The third part of a search engine is a set of decisions about weighting or ranking of documents. If you search for horse, the engine may return 1,000,000 records. How does it decide which ones to display first? Perhaps you have entered "horse race" and are interested in how to become a jockey. The words horse and race may return documents on horse breeding, horse blankets, dude ranches, race relations, and many other items you do not want. The reason you will get different results using different search engines is that they use different decision sets to index the documents and to weight the documents, deciding which documents display toward the top of the list. It is good to keep in mind that these are all automatic processes, designed to handle large numbers of documents.

So, when evaluating information retrieval systems, what we are evaluating, particularly in the traditional methodology, is how the system divides the

documents (parses the document), how it constructs the index, and the decision set for ordering the results.

History of Information Retrieval Evaluation

Of course, there has been information retrieval of a sort since there have been stockpiles of information (libraries). Information systems emerged as tools of scientific communication. With the ability to search and retrieve information came the desire to know if one method of search and retrieval is better than another. In this paper, I delineate three conceptual eras in the history of search/retrieval systems which serve as a lens to help understand the history. For each era, there are two foci – technical and theoretical. While the technological foci do follow a linear, temporal course, the boundaries of the theoretical considerations are not as clear. However, with the reader's indulgence, this division should prove to be worthwhile in making sense of a short summary of 50 years of research and development. The purpose of this history is to attempt to illuminate the backdrop against which the theoretical manifestations take place.

The modern age of evaluation dates back to the late 1950s. Sparck Jones references the 1958 International Conference on Scientific Information as the "Beginning of a new era in information processing" (Sparck Jones, 1981). Others point to the experiments by Cleverdon at Cranfield, known as Cranfield 2, that were published in the early to mid 1960s (Cleverdon, 1962; Cleverdon, Mills, & Keen, 1966).

Cranfield established several important foci, many of which are still used today. These include a system view for evaluation, a reliance on quantitative, experimental methods, and the foundational belief that the measure of success is the retrieval of the most relevant documents. The system is viewed as a black box. To evaluate systems, you need a set of questions, a set of documents, and a set of judgments as to the relevancy of the documents to the questions. If you have those things, you can write a program that uses different systems to query the documents. Then, it is possible to calculate the recall and precision measures for each system. These constructs and measures test the efficacy of the system – did it retrieve as many as possible relevant documents with as few extraneous documents as possible? Relevance was often decided by judges (as opposed to automatic word matching) and was always done prior to testing.

This traditional theoretical view continues to hold sway in much of the information retrieval community. The development of the theory was driven by technological advances in data storage and data processing. With the advent of relatively reliable computers, it was possible to store large amounts of data. However, the challenges inherent in creating these systems led naturally to the system centric view of evaluation. The question, on the "bleeding edge" of technology was, "Can this be done?", not "How do we make this work for users?" The explorers into these realms unknown, like trappers of old, were rarely particularly interested in or good at the niceties of human

interaction. The questions they focused on were the needs of the system. Those who used the system were really viewed as no more than a type of input/output device.

Just as explorers inevitably had to give way to townspeople, who could never survive without stores, power, and utilities, a new focus was necessary as systems become used, to meet the needs of a new type of users who were less hardy and resilient. While these systems were computer-based, the data in them was rarely full text and almost never anything but text. The documents were represented by abstracts, indices, or phrases representing documents. Storage was still expensive and online versions of journals were rare. So, most systems returned references to paper journals.

The second conceptual era was, like the first, theory based, but was driven by technology. Beginning in mid 1970s, with continued ebbs and flows of interest (the latest having been in the mid-1990s), the hallmark of this era was a focus on the user and the result of the search/retrieval – expanding the box to encompass the user as more than an input/output device, and investigating the inner workings of the box. Continuing the exploration metaphor, this era was the time of the homesteaders. The question was no longer whether or not the system *could* work, but *how well* it worked *for the user*. The research of this era was often qualitative in method and tended toward a constructionist epistemology. Discussions of the dynamic nature of relevance belong to this era. (Note that while the author is using the past tense, this era is still very much viable, with user-centric research still being conducted and published.) Technically, this era was the time of larger databases, more linkages, and more storage. As time went on, information could be retrieved from a great number of sources and those sources were beginning to interconnect. In addition, more of the information was full text. This meant that the information retrieval systems could search more exactly – indexing the entire document, not just an abstract or other document representation. More systems were also open to use by neophytes, if not the general public. More computers were available to more people.

The third era is driven by technology – specifically, the web. Since the 1980's personal computers have become increasingly available. Mosaic, the first widely distributed browser, was released in 1994 and both surfing the web and putting up web sites became increasingly common. By the late 1990s, most K-12 schools, nearly all universities and many community libraries had computers and high-speed access to the internet. The result of this dissemination of technology has been greater access by greater numbers of people to more and more information available. Instead of wondering *if* information can be found on a subject, one must develop strategies for picking through thousands of hits. In addition, fundamentally different kinds of information systems are being developed. In addition to the web, there are web-based databases of information not publicly available (sometimes called the hidden web). Databases (both public and “hidden”) of millions of full

text records that grow by thousands of records a week exist.

Theoretically, there is a growing diversity of interests as the field of evaluation expands and matures. Researchers involved in studying information retrieval have come to specialize in fundamentally different areas, often with little common ground between specializations. For example, this study, involving a very small documents set (in the range of 500), operates under a completely different set of interests and constraints than a researcher developing retrieval tools for PubMed or similar huge databases where it is essential to develop mechanisms that automatically index documents to cope with the ceaseless torrent of new information.

There are crucial differences in the framing of research questions and the approach to systems development and evaluation between those working on small and huge systems. However, within the literature, these differences are rarely mentioned or recognized. This lack of awareness of fundamentally different systems is certain to be problematic. Petroski, who writes of mistaken paradigms, says of the problem of scale, “Perhaps no principle in design is so well known and yet so frequently forgotten as the effect of size or scale on performance.” (Petroski, 1994 p.29) Designs for boats and bridges that worked fine suddenly fall apart when “supersized”. This unrecognized problem of scale is pervasive in instructional design – studying the effect of an intervention on a single person or small group and assuming that the intervention may then be scaled up with no further study (Schwen, 2001). In this study, I shall be explicitly studying small systems. The design concepts I hope to evaluate would not be tenable for large systems.

In addition to the difference between large and small systems, there are an increasing number of custom systems that are designed for specific groups of users and specific types of information. Some systems, such as popular web search engines have “one size fits all” interfaces that access a huge and growing number of documents. Others are directed toward specific users who share a vocabulary and information need and require the use of thesauri and/or training for effective use. Researchers are exploring new methods for accessing information through not only the internal workings of the systems, but through different approaches to user interface. As this era proceeds, there is little doubt that more systems that are fundamentally different will be developed. It will be interesting to see how evaluation theory grows to encompass this plethora of systems and diversity of research agendas.

Review of Pertinent Literature

Traditional Information Retrieval Evaluation

Salton makes the point that to properly evaluate a system, one must look to the assumptions underlying judgments of success for that system (Salton, 1992). The assumptions of success established for the large scale

evaluation projects in Cranfield, England in the 1960s were that a “good” system retrieved as many documents as possible and that the returned set of documents had mostly documents that pertained to the question asked (Cleverdon et al., 1966; Sparck Jones, 1981). The best system would return all possible documents that had something to do with the question and no documents that were extraneous.

The above theoretical assumptions led to the constructs of recall (how many of the documents that had something to do with the question were retrieved) and precision (how many of the retrieved documents actually do have something to do with the question). Underlying these constructs is the construct of relevance – a judgment as to which documents do actually pertain to each question. When the literature talks of relevance judgments (Harter, 1996; Hersh, 1994; Mizzaro, 1997; Park, 1994), this is a discussion of the *measure relevance*.

Precision as a construct is trouble free. The construct Recall is slightly more controversial, as it ties to a theoretical assumption that it is good to retrieve as many of the relevant documents in a document set as possible.

Traditional evaluation uses a systems approach – the system is viewed as a black box with no need of user input for evaluation. A batch program runs queries against the test document collection and ratings of recall and precision are calculated. This approach to evaluation is rooted in an objectivist epistemology (Crotty, 1998). There is an external reality. Documents may be judged relevant or not. A clear measurement of effectiveness of a system may be calculated. Systems are evaluated using experimental research methods.

Problem with the traditional model

The traditional model of evaluation, introduced in the 1960s is still very much in use today. However, there are several fundamental questions about this model. The question that has been explored most extensively examines issues of relevance. Questions include: What is a definition of relevance? Who should judge relevance? and Is relevance a static or a dynamic judgment? In addition to relevance issues, some researchers are taking a more user-oriented view, looking at task orientation and interactivity.

What is relevance

While very simple on first view, relevance (does a document answer a query) has proven to be a very difficult construct, mostly because of the difficulty in developing an accepted relevance measure. As definitions often lay the groundwork for a measure, the only part of a definition that can be agreed upon is that relevance is a relation between two entities. (Mizzaro, 1997). Schamber used as a foundation definition “the relationship between a user’s information problem or need and the information that could solve the problem. (Schamber, 1994). It is interesting to note that in defining the construct, she placed the user within the definition, indicating that

measures that do not address user needs are ineffective. As an advocate of relevance measures that depends on the user, the situation, and the information need, Schamber immediately distanced herself from the traditional, systemic measures of the construct.

Saracevic, a leading writer in the field, proposes five different measures of relevance (Saracevic, 1996). They are: System (automatically assigned by an algorithm), Topical (assigned by a judge), Cognitive (user defined based on mental processes Barry & Schamber, 1998; Harter, 1996; Hersh, 1994; Park, 1994; Schamber, 1994; Spink, Greisdorf, & Bateman, 1998), Situational (user defined by the task at hand Barry & Schamber, 1998; Hersh, 1994) and Motivational (user defined based on goals and texts).

In a recent article, Borlund suggests two main classes of relevance measures – objective or system-based and subjective or human (user)-based (Borlund, 2003). System-based relevance considers relevance judgments that do not involve a human judgment, depending on the matching of words or word combinations. While intuitively a weaker method of ascertaining relevance, these methods are essential when dealing with huge systems. In order to make a calculation of recall, one must divide the total number of relevant documents by the number of relevant documents retrieved. In order to evaluate such a system, one must use system-based relevance or turn to some other type of evaluation that either uses normalized precision (Quiroga & Mostafa, 2002), relative recall and precision (Gordon & Pathak, 1999; Su, 2003a) or does not use relevance measures.

Borlund details three types of relevance judgments within the class of user-based relevance – intellectual topicality (judges decide), pertinence (dynamic, user based cognitive), and situational (user identified, task based).

System-based relevance and topical relevance (or intellectual topicality) reflect an objectivist epistemology, assuming that an object can be judged by someone (or a set of algorithms) to be relevant or not. The user-based measures reflect a constructionist epistemology, where the user, and only the user, in situ, is competent to make a relevancy judgment. At the heart of this tension lies the question, “Relevant when and to who?” There is an extensive literature that examines the personal, dynamic nature of relevance (Harter & Hert, 1997; Hersh, 1994; Mizzaro, 1997; Park, 1994; Schamber, 1994; Spink et al., 1998). Different users find items relevant, depending upon their own information need and their existent knowledge, preferences, and understandings. Researchers have found differences in relevance judgments depending on the document’s representation (title only, title and summary, index items, full document). While Barry and Schamber have begun to construct a set of shared relevance criteria for relevance judgments (Barry & Schamber, 1998), they agree that each person uses the criteria to make different judgments. In addition, depending on the specific task and the situation (lack of time, level of interest) a person will judge relevancy

differently (Barry & Schamber, 1998; Harter, 1996; Hersh, 1994; Mizzaro, 1997). This contention is supported by research. It also makes sense if one reflects upon one's own search behavior. When doing a web search, we use different and changing methods for deciding whether or not to look at documents returned by a search.

However, the problem with user-constructed measures of relevance arises when using them to measure recall. If it is not possible to assign relevance to all documents in the document set, it is not possible to calculate recall (the measure of how many of the total relevant documents are returned). Remember, the traditional methodology uses the paired measures of precision and recall. Relying only on precision would give higher marks to a system that returned one relevant document, but ignored 200 other relevant documents. Relying on recall alone would give higher marks to a system that always returned the entire document set, because that would assure 100% recall. The traditional method requires both precision and recall in order to give a richer view. Some researchers have constructed measures of recall and precision based only on the documents returned. Called either normalized (Quiroga & Mostafa, 2002) or relative measures (Gordon & Pathak, 1999; Su, 2003a), these measures ask users to make relevance judgments for the set of returned documents, then calculates precision and recall from that number. This approach has had limited application.

This explains the persistence of the traditional measures. Those involved with building systems seek nice, clean measures...a number by which they can compare systems. It is important to keep in mind that this is the purpose of evaluation – a comparison...a comparison of two systems, or a comparison of alternative manifestations of a single system. Information retrieval systems are typically large and expensive to build. Developers and funding agencies want some method to easily compare systems and manifestations. The traditional measures provides that method, while research using user defined relevance measures, although they give a richer view of searches does not. User centric measures are rarely used to compare systems. A recent pair of articles by Su (Su, 2003a, 2003b) present the first research this author is aware of that compares systems with user-centric measures. For that study, Su used relative recall and relative precision measures (among several other new constructs and measures), having users judge the relevance of a document set of 20 retrieved documents. Most studies that explicate user-based relevance measures are concerned with studying the measures while evaluating a single system, not using the measures for comparison of systems. Even if traditional relevance measures rests are questionable, the constructs of recall and precision are a universally accepted means of reporting system performance and continues to be used for lack of another, similarly clean, alternative. This is analogous to the debate in education about using standardized tests as a measure of success or a predictor of future success. While there are definite questions about

the validity of test scores, they continue to be put forth by elected officials as useful indicators. This is because it is easy to talk about and compare test scores. If scores go up, that is good. If they go down, that is bad. Such news easily reduces to a headline or sound bite. The intricacies of the assumptions upon which the test development rests are not so easy to understand. In the same way, while there are fundamental problems with the measures in the traditional information retrieval evaluation methodology, recall and precision are ubiquitous because they are simple to do and easy to understand and compare.

To this researcher, the tension between the traditional paradigm and those interested in understanding and questioning relevance measures represents the tension between two research epistemologies – objectivism and constructionism. Not a dichotomous shift, there is evidence of a move along a continuum from objectivism toward constructionism. In addition, there is a clear move toward adopting more qualitative methodologies within the field.

The objectivist thinkers see the Cranfield model as providing a clear measure of success. Of course, external judges can assess relevancy. The experimental methods allow batch testing – so each system is tested numerous times with numerous queries, with variables well controlled. To an objectivist view, this gives the best possible evaluation of a system. For those moving toward a more constructionist epistemology, these methods are as the house built on sand – “The rain will fall, and the floods will come, and the winds will blow and beat against that house, and it will fall” (Matthew 7:26-27). However, it is clear that while some researchers have constructivist leanings – looking at a user-centric construction of relevance rooted in an individual, a time, a task, a situation – they have not given up the hope to be able to provide a means to evaluate systems, clearly an objectivist goal.

These “closet constructivists” have certainly expanded the methods of research, investigating task-based, naturalistic studies (Vakkari, 2003). A rich research literature going back at least 30 years points to inherent weaknesses with using relevance based measures (the constructs Recall and Precision), yet all IR systems are *still* evaluated using Cranfield-type experiments. Two groups, two different research traditions, one undergoing an epistemological crisis of faith exist within the field with little understanding or communications between the two sides, and not much realization that there even is a problem. So, like a bad marriage, there is tension. One group studies and questions the traditional constructs and measures. The other group continues to read their morning paper, mumbling, “Yes dear.” while continuing to use Cranfield-like experiments for evaluation. The frustration is particularly evident in the writings of those involved in relevance research. In 1996, Harter wrote,

...the reaction to this research and criticism from experimental researchers who use relevance assessments to conduct Cranfield-like experiments on information retrieval systems, has been mostly

silence...It is as though the enormous experimental and naturalistic research literature discussing factors that affect relevance assessments did not exist, or that the critics of the use of relevance as a foundation for retrieval testing are publishing in remote fields or live in distant lands with which we have no communication. The critics have not been refuted, or even acknowledged. Mostly they have simply been ignored. Researchers conducting experimental work in information retrieval using test collections and relevance assessments *assume* that Cranfield-like evaluation models produce meaningful results. But there is massive evidence that suggest the likelihood of the contrary conclusion (Harter, 1996) p.43).

In Instructional Systems, a similar tension over methodologies played out over the past two decades, with qualitative methods growing in influence and use. The resultant blending of methodologies (Denzin & Lincoln, 2000a, 2000b) provides guidance in this current research. The current problem in evaluation is that on one hand, relevance-based measures using experimental methods provide too thin a picture of a system's efficacy, but on the other hand, alternative constructs and measures have either not provided potential for evaluation, not provided potential for comparative evaluation of systems or have required too many resources to complete. In this research, I seek to apply Yin's (Yin, 2002) case study methods (a qualitative method with an objectivist foundation) to provide a new view of evaluation. By using a case study approach to evaluation, it is possible to examine a smaller number of users in greater detail, attempting to obtain a richer understanding of users' views of two systems. This move away from experimental, batch testing methods allows investigation of new constructs and measures for comparative evaluation.

What to study?

In order to obtain a richer comparison, this study looks at four constructs and their attendant measures – the paired traditional constructs of precision and recall and two measures that do not use relevance judgments – information need and utility. The tradition constructs of recall and precision will be used to provide a benchmark. Using the traditional constructs requires a document set, a set of queries or tasks, and relevance judgments about which documents match which tasks.

This study does not explore constructs that utilize user-based relevance measures. These constructs have been extensively researched elsewhere and do not address the underlying theoretical assumption that “success” means retrieving as many relevant documents as possible. For retrieval in a performance support context, where more is *not* better, examination of non-relevance based constructs and measures provides a more effective investigation.

Following Hersh, this study evaluates the outcomes experienced by users, not the system. (Hersh,

1994) Working in the area of medical information retrieval, Hersh makes the distinction that information found using systems be not only relevant, but also correct. Hersh argues that we need to develop new constructs as well as measures of the relevance construct.

...systems should ultimately be judged by how well the help users in their task for which they consult the system, whether it is to make correct decisions or lead to some improved state in what they do with the information. An outcomes-based approach provides a different perspective on the topical versus situational relevance debate, indicating that neither is adequate for a complete assessment of retrieval systems. (p.202)

In addition to the traditional method of evaluation, I turn to evaluation models that use different initial assumptions of success. In this research, I have used measures based on Spink's Information Need and Cooper's Utility. Using these methods in addition to traditional relevance measures within a case study methodology will, it is hoped, provide a richer comparison of two systems.

Different assumptions for success

Spink (Spink, 2002) has developed a methodology which includes the construct of shift or change in information need. Spink has subjects self-rate themselves before and after a search as to which stage in the information search process they feel they are based on Kuhlthau's 6 stages of initiation, selection, exploration, formulation, collection and presentation. (Kuhlthau, 1997)

Kuhlthau's stages are rather gross measures for rating the efficacy of a search, particularly within a performance support system, where the problem is already established and presentation is not applicable. So, while the specific measure is limited in this context, the construct of information need and the idea of measuring the change before and after the search is certainly useful. By asking users to self-rate information need before and after a search with a system without attaching Kuhlthau's labels, we have a useful measure for the construct.

William S. Cooper's View of Evaluation

Finally, I draw from the writings of William Cooper (Cooper, 1973a, 1973b, 1981). Cooper challenged some of the foundational assumptions of information retrieval. In an article in 1973, that won the “Best Article of the Year” award from the most important journal in the field (The Journal of the American Society for Information Science), Cooper questioned the basic assumptions of retrieval evaluation. He proposed a new construct - utility, based on the value a user puts on a document retrieved. How valuable or useful are the documents retrieved by a system for the *user*? That, Cooper says, should be the judgment of success. Relevance is a good construct for information scientists

focused on designing and testing systems, as it provides an easy method for designing experimental tests that do not require human subjects. However, using such constructs lead to systems designed to optimize the goals of information scientists (a systems design version of teaching to the test). Cooper feels this is an unworthy direction. "Systems should be designed so as to optimize the satisfaction of their users, not the information scientists." (Cooper, 1973a p. 92). One must look at utility, the usefulness of *documents for the user*, not relevance, even if relevance is the easier way to judge a system. Cooper suggests that using only relevance to measure system retrieval is similar to a mechanic who judges the performance of a car on the basis of a wheel alignment test because he only knows how to do wheel alignments. This limited construct was patently ridiculous when Cooper wrote in 1973. Unfortunately, thirty years later, with larger and much more complex "cars", the system mechanics are still only checking the alignment to evaluate systems. Relevance is not a bad construct; it merely offers a partial view.

In place of relevance, he suggests the "utile" as a measure of the utility construct. The utile is measured by asking users to rate individual documents based on how much they value them. The entire system's utility may be measured by averaging the utile of each document. Cooper is much cited when those who question relevance are suggesting different measures for evaluation. However, I have found no research that attempts to use utility as a measure. Cooper himself is also not aware of any such research (Cooper, 2003). It is my belief that the reason Cooper has not been used is that the measure is more amenable to qualitative methods, and in the 1970s, such an approach was not be considered good science in the field of information retrieval evaluation. Indeed, Cooper not only attempted (in a second article) to apply a quantitative methodology to this underlying theory (Cooper, 1973b), he also called it a "naïve evaluation methodology" and spent two thirds of the paper suggesting objections and answering them. With the current general acceptance of qualitative methods in IST and the beginnings of acceptance in the field of information retrieval evaluation, the time has come to use Cooper's naïve methodology to inform the methods of retrieval evaluation.

The goal of Cooper's construct is to "measure somehow the retrieval systems' ultimate worth' to its users" (Cooper, 1973a p.88), or to develop a means to measure this construct. He proposes a series of questions to ascertain the measure of value (or utile). In his article, Cooper used money as the utile. At that time, searching was not generally available. It was common to have to pay for expert searchers. So, using money as the valuation of a search was natural.

Cooper suggests the following method for measuring the utility of a system. A search is done and the user reviews the first document. The researcher asks the user to make a judgment as to whether that document has been helpful or not, and then to quantify that judgment. If it has been helpful, how much would they be willing to

pay to have had this experience with the document? If the document was not helpful to the user, the question becomes how much the user would pay to avoid such an experience. If the user feels nothing was gained or lost, the researcher records a zero. Cooper even suggests questions to elicit this type of information. This number, either positive or negative is the "document-utility". The user then goes on to evaluate the next document. Cooper suggests that the utility of the second document is dependent on the first. For example, if the second document is a near duplicate of the first, its utility will be low or even negative.

Cooper defines the end of the evaluation process as the point when the user has either "satisfied his information need or given up the search". So, after each document is evaluated, the user is asked whether they are satisfied, want to give up, or want to continue. This is significantly difference from relevance-based measures of evaluation. The number of documents evaluated is not necessarily the number of documents the system returns. It depends upon the wishes of the user. Once the information need is satisfied or abandoned, the evaluation is complete. This can mirror the activity of the user when using a performance system, where the user seeks an answer and stops once an acceptable answer is found. This is user-centric instead of system-centric evaluation. When the user does not want to continue, the measure of system utility is calculated by totaling the document utility for each document used and averaging.

In this research, I define the utile by time, not money. So, instead of taking the actual monetary value, I ask users to rate how much time they have saved by seeing the useful documents on a scale from 1 to 7. In today's online environment, where a large amount of information is freely available, the cost is rarely actual money, but the time necessary to sift the wheat from the chafe. Cooper agrees that time is a useful measure in this case (Cooper, 2003). With this change, Cooper's naïve methodology reflects the needs of evaluation in a performance support system. As it uses the mean of the utility of each document, the number of documents retrieved is not as important as how useful they are. In this methodology, a system that returns a very few highly useful documents would be rated higher than a system that returns a great many documents of questionable usefulness.

So, in order to get a richer comparative evaluation, I begin with the traditional method, using the quantitative measures that date to Cleverdon's work in Cranfield in the early 1960s. I then turn to Spink's Information Need. Finally, I operationalize Cooper's naïve methodology from 1973.

Methods

Traditional information retrieval evaluation is objectivist, experimental and quantitative. Cooper (Cooper, 1981) has questioned the efficacy of this model, suggesting that "full-scale retrieval tests are difficult,

expensive, unreliable, and often inconclusive". A reason for this is because of the lack of a strong theoretical basis.

In fact, in the search for a general theory it is hard to do much better than to give some elaboration of the vague rule that a system should retrieve for the user those documents most likely to satisfy him. As scientific theories go this truism is not very impressive, but it is the only wisp of general theory we have. What was said of a recent political candidate can be said of document retrieval theory: Deep down inside it's shallow. (p. 201)

The current state of information retrieval evaluation is not only theoretically shallow; there are some theoretical problems, specifically with the construct of relevance and how it is measured. Relevance based constructs of recall and precision do not take into account the cognitive processes of the users because of the way relevance is measured. Recall and precision do not connect the user and the evaluation of results. However, recall and precision have been used in hundreds of studies. They are still the standard evaluative tool for information systems and have been for over 50 years. So, in this study I am going to look at two constructs that do look at the consequences of cognitive processes to see if those measures may give a richer view for evaluation.

New Constructs

This study seeks to look at new constructs and new measures that are not relevance-based, freeing evaluation from the paradigm requiring relevance judgments, which makes situ research on authentic search tasks impossible. One new construct is Utility, based on Cooper's writing (Cooper, 1973a, 1973b). The measure of utility is based on the *utile* – what the searchers value. Originally, the *utile* was money. In this research, the *utile* is time. Each document's utility is measured in *utiles* and the total System Utility is simply a mean of the individual document utility scores. The second construct is Information Need – the notion that the amount of information a user desires to answer the query they bring to a search tool changes between the time they begin to search and the time they stop searching. Information need is measured by self-rating of time needed to respond to a task both before and after a search and, in this research, is then mapped to a 7 point scale (discussed in more detail later). The use of these measures is examined in a case study, as detailed by Yin (Yin, 2002). This research examines whether these new constructs may provide additional insight into evaluation of information retrieval systems. In this specialized sense, I am generalizing to theory. The new constructs studied here are independent of each other, so do not have relationships between the constructs (as recall and precision do). The new constructs, measures of constructs and assumptions examined for this research are:

Assumptions

Search Evaluation

End user consideration of document content can provide a useful measure of evaluation.

Usefulness for the user of documents retrieved is an important way of ascertaining the effectiveness of a system.

Usefulness is relative, a judgment of each individual user.

A self-assessed measure of information need, taken before and after search/retrieval can provide a useful measure of effectiveness of a system.

Constructs

System utility - a user-based judgment of how useful the documents retrieved are for a particular users for a particular task.

Utile – the measure of utility. In Cooper's original work, the *utile* was money – cost of retrieval. In this research, the *utile* is time – time spent or saved by finding a particular document.

Information Need – the amount of time required to be ready to fulfill a task. By measuring the change in information need from before to after a search, one can ascertain the effectiveness of a system.

Measures

System Utility can be measured by asking users to rate the usefulness of individual documents and averaging the ratings of all documents retrieved that the user wishes to examine.

Information Need can be measured by asking users to self rate their information need measured both in time and on a 7 point scale before and after searching. The difference between the two assessments is the change in information need.

A Case Study

This is case study research based in an objectivist epistemology and a postpositivist methodology (Crotty, 1998; Phillips & Burbules, 2000). This study attempts capture a naturalistic setting. Denzin and Lincoln say, "...qualitative research involves an interpretive, naturalistic approach to the world. This means that qualitative researchers study things in their natural settings, attempting to make sense of, or to interpret, phenomena in terms of the meanings people bring to them" (Denzin & Lincoln, 2000a) p.3). While not a true naturalistic setting, neither is it an experimental setting, as is the format of the traditional method of evaluation. Participants searched for information to meet tasks chosen from a list of authentic tasks. In the continuum from naturalistic to experimental, this study may not have crossed the border into naturalistic, but it lingers amiably at the gates.

Collecting the data

Studies that evaluate using the traditional method use a large number of queries and a large number of searches in automated, batch testing. On the other hand,

Schatz@hartford.edu

user-focused studies investigating relevance decisions often use as little as nine respondents. This study attempts to bridge these methods. The foundation of the study is a comparison of evaluation constructs used to examine a comparison of information retrieval systems. Each respondent evaluated documents retrieved by two different systems, searching twice using each system, for a total of four searches per respondent. Thirty-four educators participated in the research. The goal of the research is a richer view of evaluation through a deeper understanding of theoretical constructs and measures.

This is a three x two study – three types of evaluation (traditional, information need, and utility) and two cases (a tag based and a text based search tool). Participants were recruited via email, posting on Listservs and bulletin boards, and referrals by associates. Sixty three educators agreed to participate, 34 finished the online research tool. Participants were all educators. Most were classroom teachers.

This research used a web-based data collection tool. This allowed a larger number of respondents from a wider geographic area (including educators from Florida, California, Oklahoma, and Massachusetts). If, because of this research, a web-based research protocol tool is shown to provide a useful view of the evaluation process, then this research will have the additional impact of providing an easily replicated method for others to use.

Participant Process

Each participant conducted four searches – two for each of the two systems. Participants were randomly assigned to start with either tag or text based search. After filling out some demographic questions, the participant selected a task from a list of 36 tasks. They rated their information need (pre test for information need).

After viewing instructions on the use of search tools, participants were allowed to search for as long and as many times as they wished. Both search tools (tag and text) displayed the title of the document and a short description of returned documents. Participants could mark a document they wished to keep and then continue searching. Participants could keep a maximum of seven documents. Searching continued until the participant clicked a button saying they were satisfied.

Participants then reviewed documents selected, one at a time. For each document, they were first asked if it had been useful (Yes/No) and then, considering the time they had spent searching, if they were better or worse off. If the document was useful and they were better off, then they were asked to quantify how useful (1-7) and how much time they had saved by finding this document (1-7). These answers were averaged to compute a document utility score. Finally, they were asked if they felt they had enough information – if they would stop looking at documents if this were not a research project. If a participant indicated they would stop, no more document utilities were calculated. However, the participants were asked to review all documents they initially selected, in order to calculate information need and in order to produce a richer data set for further research.

After reviewing all documents, participants again self rated information need to provide a post intervention measurement. In addition, two questions of user satisfaction were asked – satisfaction with results and satisfaction with the system.

Participants repeated the process with the first search tool in order to mitigate differences in results that might have been caused by being unfamiliar with the search tool. Results were averaged between the two searches, resulting in a 1-7 ranking score for traditional method, information need, and utility. Participants were then asked to repeat the process with the second search tool. The entire process took about one hour. Participants were able to leave the questionnaire and return at a later time. Most took advantage of this feature.

The measures of the three constructs

1. Traditional Method

Recall and precision were calculated based on final search results. Because each respondent completed two searches per system, the two measures of recall and precision for the system were averaged. This was repeated for the second system. In order to compare these measures, the percentage score was mapped to a 1-7 scale. Finally, the score for precision and recall for each system was averaged to provide a single P/R score.

2. Information Need

Information Need is not relevance-based measure. This construct is measured by reporting the change in the participant's information need based on interaction with the system. Respondents were asked to self rate information need before and after each search. Scores were ranked on a 1 – 7 scale.

3. Utility

Upon completion of searching, respondents reviewed documents one at a time. For each document, respondents were first asked, "Are you better or worse off having had the experience of looking at this document?" If the answer was worse, then the document was rated as zero. Positive interactions were self - rated 1-7 for a document utility score.

After each document was rated, the user was asked if they now had enough information to complete their task, or if they would like to continue looking at documents. If they wished to continue, they rated the next document. When the participant either indicated that they would stop looking at documents or came to the end of the documents selected during the search, search utility was calculated by taking the mean of the document utility scores.

4. Satisfaction

Some researchers have worked to develop measures of satisfaction, notably Bruce and Chen et.al (Bruce, 1998; Chen et al., 1998). Hersh (Hersh, 1994), has expressed doubts with the construct satisfaction. Just because the user "likes" the information does not mean

that the information is “good for them”. I did not use satisfaction as an evaluation construct, feeling it was not as robust as the other measures. However, self-ranked 1-7 measure of satisfaction was taken to provide some additional insight. This proved to be useful.

Results

Description of Participants

Data was collected from the beginning of January 2004 through the end of February 2004. 34 educators participated – 24 teachers, 3 administrators, 4 librarians, 3 university educators.

Explanation of Descriptive Statistics

The average precision (the percentage of retrieved documents that are relevant – a measure of how “noisy” the results are) for the tag-based system is 51.13% and for the text-based system is it 43.85%. By this measure, tag-based systems are more accurate than text systems.

-Insert Table 1 Here –

However, the differences between systems in this study are not statistically significant. An independent samples t-test was conducted to compare the precision scores for text and tag systems. There was no significant difference in the precision scores for tag systems ($M=51.128$, $SD=32.99$), and text systems ($M=43.85$, $SD=28$, $t(66)=-.981$, $Sig=.33$). The magnitude of the differences in the means was very small ($\eta^2=.014$).

For the recall measure, the systems are even closer, with text-based being slightly better (37.11%) than tag-based systems (35.48%). Again, the differences are not significant between tag systems ($M=35.48$, $SD=23.1$), and text systems ($M=37.11$, $SD=21.8$, $t(66)=-.299$, $Sig=.766$). The magnitude is very small ($\eta^2=.001$).

-Insert Tables 2 & 3 Here –

Development of New Constructs for Evaluation

The two new constructs (information need and utility) and the new representation of precision and recall (the combination measure P/R) were used to compare the two systems (text and tag) using a two way between groups analysis of variance (ANOVA). While the means did show a slight difference in scores between systems, the difference in means again were not statistically significant, just as the difference between the two systems using traditional measures reported in percentages were not significant.

-Insert Table 4 Here -

Therefore, it is not possible, using these new constructs to ascertain a statistically significant difference between these two systems. However, there was a statistically significant difference indicated *between* the measures ($\text{sig} = .035$). On first glance, this researcher was

disappointed, as this statistic seems to disqualify the new, easier constructs as comparable to the traditional constructs. One must note that the effect size is rather small (.03). A Tukey posthoc test indicates that the significant difference is between the P/R measure and the Utility measure.

-Insert Table 5 Here –

Two possible explanations for the lack of distinction between the two systems using these measures are worth consideration. The first is that both systems in this research were searching the same document set which was relatively small (500 documents) when compared to the 4 billion pages that Google has indexed. Searching such a small document set that was completely gathered by a single researcher might well decrease the difference between what documents the two systems would retrieve. This suggestion is given more credence from the finding that the traditional measures of precision and recall also did not indicate a significant difference between the systems. However, developing a significantly larger document set was outside the scope of this research. The second cause suggested for the lack of variance of the measures was that the numbers from all measures tended to be rather closely clustered. The variances observed were too modest to allow a better differentiation in mean comparisons. A solution for this problem did suggest itself.

New Measures

After studying the data, the researcher developed a set of new constructs by considering the measure of satisfaction reported by respondents. While satisfaction may not be a robust enough measure to be considered independently, it can provide important information as well as an indication of how much a system might be used. A respondent who likes a system is more likely to use that system, experimenting and trying to understand it in order to get the most out of the system. Conversely, if the respondent does not like a system, they tend to give up searching more quickly, often to the detriment of the final set of documents retrieved. So, satisfaction is an important indicator of system success. Instead of using it as a unique measure, three new measures were created by multiplying a participant’s satisfaction score for a system with each of the existing three scores for that system. These new measures were called P/R*SAT, IN*SAT, and UT*SAT and were calculated for both the tag-based and the text-based systems. These new measures, which resulted in scores ranging from 2.75 to 49 gave a range of data that allowed a richer view of the measures and of evaluation of the systems. These are the measures whose descriptive statistics are displayed in Table 6 below.

Two-way between groups ANOVA

A two-way between-groups analysis of variance was conducted to explore the scores generated for each of the evaluation constructs (P/R, IN, and UT) for each system. This analysis showed a difference between the

text-based and the tag-based systems. The plot (Figure 1) clearly shows the difference between the two systems.

-Insert Figure 1 Here -

Note that the text system scored better (lower numbers) than the tag-based system. This is the opposite of original expectations, a matter discussed in more later. However, it is important to remember that the main goal of this research is investigating the effectiveness of these new measures for evaluation. From this plot, the difference between the systems using the measures can be clearly seen, indicating the potential usefulness of the new measures.

There was a statistically significant main effect for system $F(1, 198) = 8.43, SIG = .004$, however the effect size was small ($\eta^2 = .041$).

The three measures (UT*SAT, IN*SAT, and P/R*SAT) do not account for a significant difference in scores – $F(2, 198) = 1.14, SIG = .32$. This is an indication that each of the three measures is similar in measuring the difference between the two systems – all three measures are measuring the same thing. If that is the case, we may state that in this case the two new measures (Utility and Information Need) seem to be comparable to the existent, traditional measures as evaluation measures.

- Insert Table 6 Here -

Correlation

To see how closely related the three measures were, a bivariate correlation was run. All three measures were significantly correlated within each system, significant at the .01 level. For tag systems P/R*SAT was correlated to IN*SAT at .704 (79%) and UT*SAT at .806 (80%). These are both very strong correlations. For text systems P/R*SAT was correlated to IN*SAT at .862 (80%) and UT*SAT at .878 (88%), again strong correlations. So, the three measures of constructs (P/R*SAT, UT*SAT, and IN*SAT) are strongly correlated within both systems – they seem to measure the same thing. In other words, there is a strong indication that they are equivalent measures.

-Insert Table 7 Here -

There were also some statistically significant, but much less strong correlations between some tag-based measures and some text-based measures. All these were at the .05 level and ranged from .39 - .42. The IN*SAT measure in tag systems was correlated with the IN*SAT measure in text systems (.42 at .05 level) as well as with the UT*SAT measure in text systems (.40). The UT*SAT measures in tag-based and text-based systems were also correlated (.40). With this correlation of measures across systems that is rather small (less than 10%), but is statistically significant, it would appear that these measures are equivalent, and there is a difference between the two systems. The new measures are useful and effective.

Conclusions

This research has reviewed the process of evaluation of two cases – two systems that use very

different methods to retrieve information. An attempt was made to emulate an authentic situation, with the information system taking the part of a performance support tool. To begin to consider what has been learned, let us return to the original research questions.

Can a richer comparative evaluation be provided by multiple constructs of success?

Can user-centric constructs provide time and cost effective comparative evaluation of two systems?

Do non-relevance based constructs, specifically change in information need and Cooper's utility provide useful insight into comparative evaluation?

In both cases, multiple measures clearly provided a richer view into the evaluation of the systems. While analyzing the data, limitations or confusion around each measure was apparent in respondent's answers. While this researcher would not like to be limited to any of these measures as a single measure, using some in conjunction with others clearly offers a richer view.

The limitations of the traditional constructs of precision and recall were quickly apparent. During relevance judgments, with only two judges, there was a high (over 25%) rate of disagreement. It was also the case that judgments often did not agree with user rating of usefulness. Perhaps more limiting was the limitation of relevance based measures vis a via authentic situations. As mentioned previously, using relevance measures compel evaluations to take place in an experimental setting. This burden became apparent during the research. Analyzing responses of searches undertaken within a quasi-authentic situation highlighted the value of in situ evaluation, using authentic tasks with the needs, drives, limitations, and time constraints of a real user. The difference between evaluation in these circumstances and in an experimental setting, even a setting such as this one, which attempted to evoke authenticity, puts one in mind of car commercials in which a car spins and dances around obstacles with ease while small text at the bottom of the screen states, "Professional driver on closed track. Do not attempt to replicate." Information evaluation in any but authentic situations can only provide images of in situ search behavior of less use and less substance than the shadows dancing on the walls of Plato's cave. One of the greatest advantages of these two new measures is that they can be used in authentic settings.

The Information Need construct provided a good window into the systems. There was some confusion about these measures evidenced with some participant's remarks and scores. In a small number of searches, scores indicated confusion (pre need rating very low or none or post need rating lower than pre need rating). In previous studies examining information need, a researcher was present to aid the participant in assessing information need. As I choose to attempt to implement an online evaluation tool, a researcher was not there to provide such clarification. While a toll-free number and email link for support was provided, less than 5 calls and less than 10 emails asking for support or clarification were received.

Only one of these asked for clarification about this measure. With more refinement of the assessment tool, its instructions, help functions, and interface, this problem may be reduced or negated.

The Utility construct was certainly effective. Particularly promising was the explicit question asked after reviewing each document – “Do you have enough information now? Would you stop or continue?” While the Information Need measure also does not require a review of all documents retrieved, Utility explicitly encourages the participant to behave in an authentic manner. The measure seemed clear to respondents. Several did say they would stop after one or two documents. While not quantifiable, this researcher felt the responses generated from this measure to be qualitatively superior. The scores seemed more closely tied to the actual practice of searching and the individual documents examined. Responses seemed to be more based in respondents’ actions, not on someone else’s view of relevance. The measure did not ask the respondent to try to assess and quantify their information need – a relatively fuzzy concept. It asked the respondent to look at each document and respond whether or not it was useful to them, a task more easily undertaken.

Each of the constructs provided a remarkably similar view of the two systems. Indeed, with the correlation so high between constructs within each system (49% - 82% correlation in the tag system and 74% - 80% for text systems) it seems that these are not new constructs. Instead, *these are new measures for the same constructs*. Instead of identifying the *constructs* as traditional, information need, and utility, let us identify a new construct - *system efficacy*. Then this research indicates that, at least in these cases, there is evidence that traditional, information need, and utility are three *measures* of the same construct – system efficacy. So, the first research question – “Can a richer comparative evaluation be provided by multiple constructs of success?” must be answered in the negative. We have not found multiple constructs, but multiple measures of a single construct.

An important part of this research was the development of three new online tools. The case studies examined not the development of the tools, but the evaluation of systems. However, it is appropriate here to discuss the technical achievements.

For tagging and searching for objects, the findings are conflicting. This is the first time an online tool for tagging web-based objects was built, tested, and used extensively. On the positive side, the tagging tool worked extremely well. With the implementation of this tool, a number of technical decisions were made, including data structure and interface design. A fundamental question of whether to locate the objects in a database or provide pointers to the URL of the object was decided in favor of providing pointers. So, some very fundamental decisions were made and an important first step was taken.

Using the tool to tag over 500 objects was instructive. On a positive note, it was certainly easy and

fast to tag an object. This is a basic question that cuts to the heart of the eventual utility of tagging systems. If it takes too long to tag objects, people will not tag them. A surprising question that arose during the process was one of uniformity. In all tagging systems, there is a problem with having tagging judgments made with some degree of uniformity, particularly when an untrained group of users is allowed to tag objects. I have suggested the use of controlled vocabulary as a means to try to enforce uniformity. However, during this project, I realized that even with controlled vocabulary, tagging will not be uniform. Even with one tagger, over the eight weeks of collecting and tagging objects, my tagging was not uniform. On one hand, this points to the near impossibility of achieving any kind of uniformity. On the other, it shows the importance of explicit, thoughtful, controlled vocabulary as a much stronger method for encouraging uniformity than text fields with no restrictions. In addition, I doubt that perfect uniformity in tagging is either possible or necessary, as different users’ search behavior certainly will not be uniform. Over time, we will see how damaging this lack of uniformity is to the usefulness of tag-based systems. I began this project an advocate of tag-based systems. I end with less enthusiasm, but not a clear decision of their effectiveness. In order to find out more, we must fine tune tagging systems and that will require new measures for evaluation that allow in situ evaluation.

The second tool built was the tag based search tool. Again, this was the first time that such a tool was implemented and used extensively by many people. Many technical and design considerations were answered, at least for this implementation. While there are areas that need improvement, including interface design and testing, this is an excellent first step. Tag based searching holds great promise of much greater accuracy and control than can be imagined with a text based search tool.

Finally, the design, implementation, use of the online questionnaire tool is a significant achievement. While there were some technical problems, overall, this tool is an important advance. Because of this tool, it was possible in this research to draw on a very wide population. An important benefit is it’s web based delivery. This allows respondents to answer questions in the same milieu, at the same time as they are searching. Participants may search using any search engine or retrieval system from within the system, then have the questionnaire on the screen at the same time, allowing respondents to engage in authentic information seeking behavior with a very minimal intervention or influence by researchers. Coupling this power with the richer understanding of evaluation and the success in this case study of the two measures that do not require relevance judgments opens the possibilities of being able to study information seeking in situ. This opens the possibility of advances in theory and methods of search evaluation that will allow us to fine tune retrieval tools for specific audiences. Moving from a focus on huge systems searching the world, we can look to systems that search very specific document sets for very specific, targeted

needs, retrieving a much smaller set of documents much more accurately meeting the needs of a specific group. These are very exciting possibilities and provide an emphatic affirmative within these cases to the second question – “Can user-centric constructs provide time and cost effective comparative evaluation of two systems?”

The final question addresses the two new measures – information need and utility. Harter, in 1996, wrote, “Our approaches to evaluation must reflect the real world of real users”(Harter, 1996). Cooper suggested his naïve method in 1973. Why have no new ubiquitous measures come to the fore? I believe it is because there have been alternatives that are easy, effective, and inexpensive. Non-relevance based evaluation constructs usually involve measures that require researcher observation and interpretation. Precision and Recall have also required “expensive” measures – requiring relevance judgments. If, as discussed above, we view information need and utility as new measures of the construct System Efficacy, then we have powerful new measures, easily operationalized with the online questionnaire tool, to evaluate this new construct. These new measures certainly do provide useful insight into comparative evaluation within the context of these cases, with the advantage of freeing the researcher from the fundamentally flawed relevance decisions and the need to evaluate in an experimental rather than an authentic setting.

The results, particularly the strong correlations between measures within systems and the weaker correlations across systems, certainly indicate that more research is called for. It is not within the scope of this research to generalize outside of these two cases. The methodology does not support any assertion of application outside of this particular case study. However, the findings of this research do begin to build a theoretical foundation for eventual research that may be able to be more sweeping in its conclusions. What this research shows us is that there is certainly reason for further study. There is reason to continue to examine the application and use of these measures in other settings and other systems.

Implications

Limitations

As with any research, there are limitations to this study. The first is the relatively small document set. The tag-based system used in this study required that all documents be studied and tagged. To increase the validity of these tags, the tags for each document were examined by a second researcher. So, each document in the set had to be examined and considered by two different judges. This made even a document set of 500 documents a significant undertaking. Secondly, relevance judgments were required for each document. A document set of 500 was at the upper limit of available resources and was large enough to reasonably consider the results as tenable. However, the size of the document set may have contributed to the small difference between the two systems.

The second limitation stems from the research method. Because of the convenience method of sampling, the findings of this study cannot be generalized to the population. It is not possible to conclude based on this research that one system is better at returning relevant or useful documents returned than another. It is not possible, based on this research, to claim that the new measures of information need and utility correlate with the traditional measures in any situation except in this case. The intent of this research is to support generalizing to theory, to provide a richer theoretical basis for evaluation. We cannot generalize, but the results certainly allow us to say that the results are promising and further research would be time well spent.

In addition, the tasks and the situation were not authentic. This was not research completed in situ, with the pressures and drives that would normally drive information seeking behavior.

Finally, as has been mentioned, with the development of several new technologies, there were many technical problems, which resulted in frustration for some respondents, loss of some data, and loss of some respondents who were unable or unwilling to complete the research.

Future Studies

The findings of this research do hold great promise for the evaluation of information systems. Enough has been uncovered to make further research worthwhile. With the conclusion that, in this case, the measures of Utility and Information Need were effective in evaluating systems, an important next step would be to add to this foundational work by using these measures in evaluating other systems. One direction would be to study systems with large document sets that have been evaluated using precision and recall, again comparing the new measures with the existing standards. Another extension would be to compare these measures to measures of normalized precision and recall to attempt to draw a comparison. Finally, using these measures in situ with different audiences in order to test efficacy in different populations will extend the value and application of these new measures.

Of particular interest for this researcher is the extension of both the measures and the online tool for application in authentic situations. Evaluating search behavior and results over an extended period of time within the actual settings for information seeking would be invaluable to an understanding of systems and of methods of evaluations. I feel that this holds great promise, not only adding to the theoretical foundation, but also in the creation of tools that will allow other researchers to explore evaluation.

Finally, I hope to continue research into tag-based systems. With this first implementation, new problems presented themselves. In particular, studying ways to make the interface more intuitive and to present explanations and instructions is necessary before an assessment of the efficacy of tag-based systems can be

completed. Rather than putting the question to bed, this research has stirred up more questions, more

considerations, and more possibilities. Can a researcher wish for anything better?

References

1. Barry, C. L., & Schamber, L. (1998). Users' criteria for relevance evaluation: A Cross-situational comparison. *Information Processing and Management*, 34(2/3), 219-236.
2. Benson, G. (1997, January, 1997). A New look at EPSS. *Training & Development*, 43-44.
3. Borland, P. (2003). The Concept of relevance in IR. *Journal of the American Society for Information Science*, 54(10), 913-925.
4. Bruce, H. (1998). User satisfaction with information seeking on the internet. *Journal of the American Society for Information Science*, 49(6), 541-556.
5. Burton, R. R., Brown, J. S., & Fischer, G. (1999). Skiing as a model of instruction. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its Development in social context*. Cambridge: Harvard University Press.
6. Carliner, S. (2002). Read me first: An Introduction to this special issue. *Technical Communication*, 49(4), 399-404.
7. Carr, C. (1992, June, 1992). PSS! Help when you need it. *Training & Development*, 31-38.
8. Chen, H., Houston, A. L., Sewell, R. R., & Schatz, B. R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49(7), 582-603.
9. Cleverdon, C. W. (1962). *Report on the testing and analysis of an investigation in the comparative efficiency of indexing systems*. Cranfield: ASLIB Cranfield Research Projects.
10. Cleverdon, C. W., Mills, j., & Keen, E. M. (1966). *Factors determining the performance of indexing systems* (Vol. 1). Cranfield, UK: Aslib Cranfiled Research Project, College of Aeromautics.
11. Cole, K., Fischer, O., & Saltzman, P. (1997). Just-in-time knowledge delivery. *Communications of the ACM*, 40(7), 49-53.
12. Cooper, W. S. (1973a). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2), 87-100.
13. Cooper, W. S. (1973b). On Selecting a measure of retrieval effectiveness: Part II. Implementation of the philosophy. *Journal of the American Society for Information Science*, 24(6), 413-424.
14. Cooper, W. S. (1981). Gedanken experimentation: An Alternative to traditional system testing? In K. Sparck Jones (Ed.), *Information Retrieval Experiment* (pp. 199-209). London: Butterworths.
15. Cooper, W. S. (2003). email from Cooper about utility. In S. C. Schatz (Ed.).
16. Crotty, M. (1998). *The Foundations of social research: Meaning and perspective in the research process*. London: Sage Publications.
17. Denzin, N. K., & Lincoln, Y. S. (2000a). Introduction: The Discipline and practice of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Handbook of qualitative research* (2nd ed., pp. 1-29). Thousand Oaks: Sage Publications.
18. Denzin, N. K., & Lincoln, Y. S. (2000b). The seventh moment: Out of the past. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of Qualitative Research* (2nd ed., pp. 1047-1065). Thousand Oaks: Sage.
19. Dickelman, G. J. (2000). Performance support in internet time: The State of the practice Discussion between Gloria Gery, Stan Malcom, Janet Cichelli, Hal Christensen, Barry Raybould, and Marc J. Rosenberg. *Performance Improvement*, 39(6), 7-17.
20. Drazin, R., Glynn, M. A., & Kazanjian, R. K. (1999). Multilevel theorizing about creativity in organizations: A Sensemaking perspective. *The Academy of Management Review*, 24(2), 286-307.
21. Gilbert, T. F. (1996). *Human competence: Engineering worthy performance*. Silver Spring, Maryland: ISPI.
22. Gordon, M., & Pathak, P. (1999). Finding information on the world wide web: The Retrieval effectiveness of search engines. *Information Processing and Management*, 35(2), 141-180.
23. Hackos, J. T., & Redish, J. C. (1998). *User and task analysis for interface design*. New York: Wiley Computer Publishing.
24. Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1), 37-49.
25. Harter, S. P., & Hert, C. A. (1997). Evaluation of information retrieval systems: Approaches, issues, and methods. *Annual Review of Information Science and Technology (ARIST)*, 32, 3-94.
26. Hersh, W. (1994). Relevance and retrieval evaluation: Perspectives from medicine. *Journal of the American Society for Information Science*, 46(3), 201-206.
27. Hogan, D. M., & Tudge, J. R. H. (1999). Implications of Vygotsky's theory for peer learning. In A. M. O'Donnell & A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 39-66). Mahwah, New Jersey: Lawrence Erlbaum Associates.
28. IEEE_LOM. (2002). *Position Statement on 1484.12.1-2002 Learning Object Metadata (LOM) Standard Maintenance/Revision*. Retrieved January 12, 2004, from <http://ltsc.ieee.org/wg12/>
29. Jonassen, D. H., Tessmer, M., & Hannum, W. H. (1999). *Task analysis methods for instructional design*. Mahwah, N.J.: Lawrence Erlbaum Associates.
30. Kuhlthau, C. C. (1997). Learning in digital libraries: An Information search process appraoch. *Library Trends*, 45(4).

31. McGee, P. (2004). Fundamental tension between pedagogy and learning objects. In S. C. Schatz (Ed.).
32. Mizzaro, S. (1997). Relevance: the Whole history. *Journal of the American Society for Information Science*, 48(9), 810-832.
33. Morrison, J. B. (2002). *The Right shock to initiate change: A Sensemaking perspective*. Paper presented at the Academy of Management, Denver, CO.
34. Park, T. K. (1994). Toward a theory of user-based relevance: A Call for a new paradigm of inquiry. *Journal of the American Society for Information Science*, 45(3), 135-141.
35. Petroski, H. (1994). *Design paradigms: Case histories of error and judgment in engineering*. Cambridge: Cambridge University Press.
36. Phillips, C. C., & Burbules, N. C. (2000). *Postpositivism and education research*. Lanham: Rowman & Littlefield.
37. Quesenbery, W. (2002). Who is in control? The Logic underlying the intelligent technologies used in performance support. *Technical Communication*, 49(4), 449-457.
38. Quiroga, L. M., & Mostafa, J. (2002). An Experiment in building profiles in information filtering: The Role of context of user relevance feedback. *Information Processing and Management*, 38(5), 671-694.
39. Reid, J. (2000). A Task-oriented non-interactive evaluation methodology for information retrieval systems. *Journal of Information Retrieval*, 2(1), 115-129.
40. Rogoff, B. (1990). *Apprenticeship in thinking*. New York: Oxford University Press.
41. Rogoff, B., & Lave, J. (Eds.). (1984). *Everyday cognition: Development in social context*. Cambridge: Harvard University Press.
42. Rosenberg, M. J. (1995). Performance technology, performance support, and the future of training: A Commentary. *Performance Improvement Quarterly*, 8(1), 94-99.
43. Rossett, A. (1996). Job aids and electronic performance support systems. In R. L. Craig (Ed.), *The ASTD training and development handbook - Fourth edition* (pp. 554-580). New York: McGraw - Hill.
44. Salton, G. (1992). The State of retrieval system evaluation. *Information Processing and Management*, 28(4), 441-449.
45. 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.
46. Saracevic, T. (1996). *Relevance reconsidered*. Paper presented at the Information science: Integration in perspectives. Second conference on conceptions of library and information science (CoLIS 2), Copenhagen, Denmark.
47. Schamber, L. (1994). Relevance and information behavior. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 29, pp. 3 - 48). Medford: Learned Information, Inc.
48. Schneider, S. C. (1997). Interpretation in organizations: Sensemaking and strategy. *European Journal of Work and Organizational Psychology*, 6(1), 93-102.
49. Schwen, T. M. (2001). *The digital age: A Need for additional theory in instructional technology?* Guangzhou, China: South China Normal University.
50. Sparck Jones, K. (Ed.). (1981). *Information Retrieval Experiment*. London: Butterworths.
51. Spink, A. (2002). A User-centered approach to evaluating human interaction with Web search engines: An Exploratory study. *Information Processing and Management*, 38(3), 401-426.
52. Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. *Information Processing and Management*, 34(5), 599-621.
53. Su, L. T. (2003a). A Comprehensive and systematic model of user evaluation of web search engines I: Theory and background. *Journal of the American Society for Information Science*, 54(13), 1175-1192.
54. Su, L. T. (2003b). A Comprehensive and systematic model of user evaluation of web search engines II: An Evaluation by undergraduates. *Journal of the American Society for Information Science*, 54(13), 1193-1223.
55. Vakkari, P. (2003). Task-based information searching. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 37, pp. 413-464): Information Today.
56. Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks: Sage.
57. Weick, K. E. (2001). Making sense of the organization. In. Malden: Blackwell Publishers.
58. Wiley, D. A. (2001). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. In D. A. Wiley (Ed.), *The Instructional Use of Learning Objects* (pp. 1-35). Bloomington: Agency for Instructional Technology.
59. Yin, R. K. (2002). *Case study research: Design and methods* (3rd ed. Vol. 5). Newbury Park: Sage Publications.
60. Ying, O. P. (2002). *Cisco's RLO Strategy*. Retrieved January 12, 2004, from <http://www.ecc.org.sg/cocoon/ecc/website/services/article/ciscostrategy.article>

Table 1. Descriptive statistics for traditional measures

Mean	
Tag	51.13
Tag - Recall	35.48
Text	43.85
Text - Recall	37.11

Recall/Precision reported as a percentage: The traditional measure

Table 2. No significant difference between system means using Recall measure

		Levene's Test for Equality of Variances		t-test for Equality of Means for Recall				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff	Std. Error Diff
Recall as a percentage	Equal var. assumed	.002	.960	-.299	66	.766	-1.62	5.45

Table 3. No significant difference between system means using Precision measure

		Levene's Test		t-test for Equality of Means for Precision				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff	Std. Error Diff
Precision as a percentage	Equal var. assumed	1.163	.285	.98	66	.330	7.28	7.42

Table 4. ANOVA of P/R, UT, and IN showing no significant difference

Dependent Variable: Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Eta ²
Corrected Model	10.96(b)	5	2.19	1.42	.22	.035
Intercept	5335.76	1	5335.76	3457.13	.00	.95
Measure	10.55	2	5.28	3.42	.035	.033
System	.01	1	.01	.01	.95	.000
Measure * System	.41	2	.20	.13	.88	.001
Error	305.59	198	1.54			
Total	5652.32	204				
Corrected Total	316.55	203				

Table 5. Difference between mean of measures – UT, IN, P/R

(I) Measure	(J) Measure	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
P/R	Info Need	-.1728	.21306	.697	-.6759	.3303
	Utility	-.5450(*)	.21306	.030	-1.0481	-.0419
Info Need	P/R	.1728	.21306	.697	-.3303	.6759
	Utility	-.3722	.21306	.190	-.8753	.1309
Utility	P/R	.5450(*)	.21306	.030	.0419	1.0481
	Info Need	.3722	.21306	.190	-.1309	.8753

Based on observed means.

* The mean difference is significant at the .05 level.

Figure 1. Plot of marginal means comparison between systems using SAT*measures

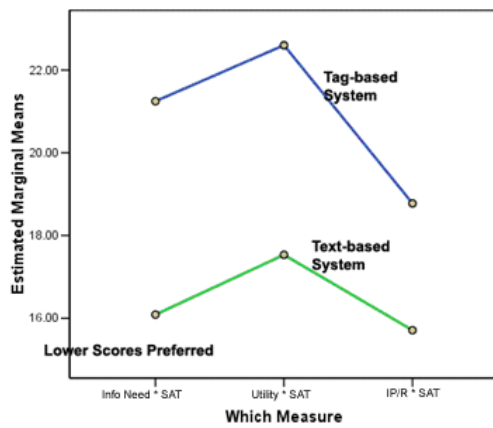


Table 6. ANOVA effects of differences-systems and measures using SAT* measures

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1320.372(a)	5	264.074	2.224	.053	.053
Intercept	71021.407	1	71021.407	598.020	.000	.751
System	1001.573	1	1001.573	8.434	.004	.041
Measure	271.189	2	135.594	1.142	.321	.011
System * Measure	47.610	2	23.805	.200	.819	.002
Error	23514.674	198	118.761			
Total	95856.452	204				
Corrected Total	24835.046	203				

a R Squared = .053 (Adjusted R Squared = .029)

Table 7. Correlations between measures P/R*SAT, IN*SAT, and UT*SAT

		P/R * SAT - Tag	P/R * SAT - Text	IN * Sat - Tag	IN * Sat - text	UT * Sat - Tag	UT * SAT - Text
P/R * SAT - Tag	Pearson Correlation	1	.195	.704(**)	.286	.806(**)	.317
	Sig. (2-tailed)	.	.268	.000	.101	.000	.067
P/R * SAT - Text	Pearson Correlation	.195	1	.311	.862(**)	.282	.878(**)
	Sig. (2-tailed)	.268	.	.074	.000	.106	.000
Info Need * Sat - Tag	Pearson Correlation	.704(**)	.311	1	.420(*)	.936(**)	.395(*)
	Sig. (2-tailed)	.000	.074	.	.013	.000	.021
Info Need * Sat - text	Pearson Correlation	.286	.862(**)	.420(*)	1	.380(*)	.894(**)
	Sig. (2-tailed)	.101	.000	.013	.	.027	.000
Utility * Sat - Tag	Pearson Correlation	.806(**)	.282	.936(**)	.380(*)	1	.397(*)
	Sig. (2-tailed)	.000	.106	.000	.027	.	.020
Utility * SAT - Text	Pearson Correlation	.317	.878(**)	.395(*)	.894(**)	.397(*)	1
	Sig. (2-tailed)	.067	.000	.021	.000	.020	.

**

Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).