

Improving Performance Support Systems through Information Retrieval Evaluation

STEVEN SCHATZ
University of Hartford, USA
schatz@hartford.edu

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 whom? When?* and *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 (tag-based and text-based), comparing these new measures with traditional measures and each other. The open-source Swish text-based search engine and a self-constructed tag-based search tool were used. 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. This study specifically analyzes small information systems. The design concepts would be untenable for large systems. Results indicated that there was a significant correlation between the three measures, indicating that the new measures provide an equivalent method of evaluating systems and have some significant advantages, which include not requiring relevance judgments and the ability to use the measures *in situ*.

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, for example, by providing better information during the course of performance rather than designing a training intervention. 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). EPSSs are part online help, part online tutorial, part database, part application program, and part expert systems. EPSSs quickly and easily provide answers to the questions workers have when performing a job, and address workers' concerns (Carliner, 2002).

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).

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" (1992). Much of the work concerning meta tags (Dublin core, IMS, SCORM, GEM, etc.) 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 investigate how the constructs and measures of information retrieval evaluation may inform the research agenda of performance support and if new con-

structs 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 1) A collection of documents, 2) A collection of questions (queries) to be asked of the document collection, and 3) A set of judgments of which documents are relevant to each question. To evaluate the system, one queries the document collection 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 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. If the system retrieved 100 documents, of which 20 were relevant, the precision would be 20/100 or 20%. 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. There is no consideration of users. The focus is on the system. In addition, underlying this method of evaluation is the assumption that more is better. Better systems retrieve more relevant documents. 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). This study is a case study that examines two new 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 familiar engines are web-based systems, such as Google or Yahoo. This study uses the open-source Swish search engine for text search. Search engines have three parts. The first is a spider or crawler. This is an automated program that goes out on the web, following links and reading web pages.

When the crawler or a controlled system reads a document, the system then breaks the documents into words and phrases and constructs an index (the second part of a search engine). This index contains a list of vocabulary words and phrases 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 on the web, 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? The reason you will get different results using different search engines is that they use different decision sets to weight the documents, deciding which documents to display toward the top of the list. 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 way it decides how to order the results.

Tag-based Searching

A card catalogue in a library contains information about the books - the catalog number, the title, the author, when it was printed, a brief description and where to find the book. Meta tags do the same thing for a web-based object. A tag-based search tool does not match the words in your search with words in the document. It matches words in your search only with the tags. In this research, we used a set of controlled vocabulary, so that documents were tagged by checking boxes with pertinent words (for example, *lesson plan, math, 3-5 grade*). So, when searching, one checks the pertinent tags and the search tool matches tags *without reading the document*. If, for example, you were searching for a lesson plan on stars and, using a text search tool, entered *star lesson plan*, you would retrieve a news article about the foibles of a drunken actor - "Star learns Lesson. Plans to enter rehab." However, a document would have to be tagged as a lesson plan with star in the title or description in order to be retrieved by a tag-based search engine.

Scale of Systems

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 huge document sets 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. This study shall be explicitly studying small systems. The design concepts evaluated would not be tenable for large systems.

Review of Pertinent Literature

Traditional Information Retrieval Evaluation

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 of those returned, most pertained to the question asked (Cleverdon, Mills, & Keen, 1966; Sparck Jones, 1981). The best system would return all possible documents that had something to do with the ques-

tion and no documents that were extraneous.

The above theoretical assumptions led to the constructs of recall (quantity) and precision (cleanliness). 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), it is a discussion of the *measure* relevance.

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 concerning this model. The question that has been explored most extensively examines issues of relevance.

What is Relevance?

While very simple on first view, relevance (whether or not a document answers a query) has proven to be a very difficult construct, mostly because of the difficulty in developing a widely accepted relevance measure. The only definition that can be agreed upon in literature is that relevance is a relation between two entities, hardly a robust measure. (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). 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.

At the heart of these fundamental disagreements is the question, “Relevant when and to whom?” 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, Greisdorf, & Bateman, 1998). The conclusion is that different users find items relevant, depending upon their own information need and their existent knowledge, preferences, and understandings.

However, user-centric measures of relevance make measuring recall difficult. If relevance depends on users, it is not possible for researchers to make advance relevance judgments for the document set, so it is not possible to calculate recall (number returned divided by relevant returned). This is one explanation for the persistence of the traditional measures. Those involved with building systems seek nice, clean measures – a number by which they can compare systems. The traditional measures provide that method. Research using user-defined relevance measures, although they give a richer view of searches, does not.

What to Study?

In order to obtain a richer comparison, this study looks at three constructs and their measures – the paired traditional constructs of precision and recall are combined and mapped to a seven point scale for comparison to two new measures that do not use relevance judgments – information need and utility. The tradition constructs of recall and precision are used to provide a benchmark.

Different Assumptions for Success

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 using Kuhlthau's 6 stages of information seeking: 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 without attaching Kuhlthau's labels, we have a useful measure for the construct.

Spink's methodology is free from questions of relevance. It may be used in situ by users. Users may search as often as they wish and view as many or as few documents as they wish – rating the system after they have finished using it. So, it is a user-centric, relevance-free measure. It also allows unobtrusive application. A potential difficulty is having research participants understand and consistently apply the self-assessed measure of information need. It is often difficult to judge one's own need before and after a search. It is certain that there will be questions of consistency between participants. This methodology provides a single point of evaluation – change in need measured at the end of searching. It does not gather feedback when the user looks at each document. This makes the measure less intrusive, but offers a less robust view of the on-going internal evaluation that each user undertakes while viewing the results of a search.

William S. Cooper's View of Evaluation

In 1973, Cooper proposed a new construct, utility, which would be based on the value a user puts on a document retrieved (Cooper, 1973a, 1973b, 1981). How valuable or useful are the documents retrieved by a system for the *user*? That should be the judgment of success. Relevance is a good construct for systems designers, 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 designers, not

users. Using only relevance to measure system retrieval, said Cooper, is akin to a mechanic judging the performance of a car on the basis of a wheel alignment test because that is all he knows how to do. 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 limited view.

In place of relevance, he suggested the *utile* as a measure of the utility construct. The *utile* may be measured by asking users to rate individual retrieved documents based on how much they value them. The entire system's utility may be measured by averaging the *utile* of each document rated. Cooper is much cited when those who question relevance are suggesting different measures for evaluation. However, this author has found no research that attempts to use utility as a measure. Cooper himself is also not aware of any such research (Cooper, 2003). The time has come to use Cooper's naive 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), or to develop a means to measure this construct. 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. As searching is now usually free, this study uses time as the valuation. Searchers rate documents as to how useful each one is to their personal need. Then they decide if they would continue looking at documents or if they have enough information. When they are satisfied, utility is calculated by averaging the utility of the documents rated.

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)

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 *in situ* research on authentic search tasks impossible. One new construct is Utility, mapped to a 7-point scale. The total System Utility is the mean of the individual document utility scores. The second construct is Information Need (based on Spink's model), measured by self-rating of time needed to respond to a task both before and after a search, mapped to a 7-point scale. The use of these measures is examined in a case study, as detailed by Yin (2002). This research examines whether these new constructs may provide additional insight into evaluation of information retrieval systems. In this specialized sense, this report is generalizing to theory.

Because both of these new constructs are mapped to a seven-point scale with no possibility of fractional results (beyond .5), the variance was reduced by the ceiling effect in the scale. Upon consideration, a construct of satisfaction was used to provide a reasonable estimate of the variance that would be seen with an expanded measure. Some researchers have worked to develop measures of satisfaction (Bruce, 1998; Chen, Houston, Sewell, & Schatz, 1998). While satisfaction may not be a robust enough measure to be considered independently, it can provide important information 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 (precision/recall * satisfaction), IN*SAT (information need * satisfaction), and UT*SAT (Utility * satisfaction). 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.

A Case Study

This is case study research based in an objectivist epistemology and a post-positivist methodology (Crotty, 1998; Phillips & Burbules, 2000). This study attempts to replicate a naturalistic setting, but it is not truly naturalistic. Participants searched for information to meet tasks chosen from a list of authentic tasks.

Collecting the Data

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. This is a 3x2 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, allowing for a larger number of respondents from a wider geographic area (including educators from Florida, California, Oklahoma, and Massachusetts).

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. The tag-based search tool was constructed for this research. The text-based search tool was the open-source Swish search engine (<http://www.swish-e.org>). 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 each of the measures traditional method, information need, and utility, as well as satisfaction. 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.

RESULTS

Explanation of Descriptive Statistics

The new measures, P/R*SAT, IN*SAT, and UT*SAT were calculated for both the tag-based and the text-based systems.

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*SAT, IN*SAT, and UT*SAT) 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.

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 detail 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 $F(1, 198) = 8.43$,

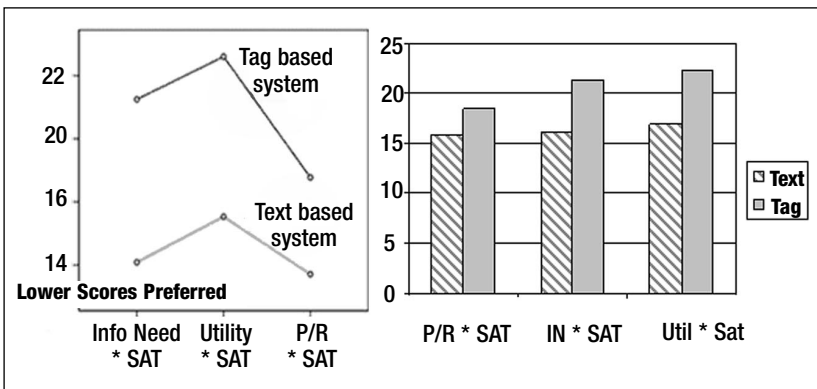


Figure 1. Marginal means comparison between systems using SAT* measures

SIG = .004), however the effect size was small (eta squared = .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.

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.

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

Table 1

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				

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

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. 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. Rather than being 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 structure that 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. 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). 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, the responses generated from this measure were qualitatively superior. The scores seemed more closely tied to the actual practice of searching and the individual documents examined.

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 sys-

tems) it seems that these are not new constructs. Instead, one may propose that these are new measures for the same construct. 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.

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. The author suggested the use of controlled vocabulary as a means to try to enforce uniformity in other works. However, during this project, it became clear that even with controlled vocabulary, tagging will not be uniform. However, it is doubtful 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. 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. Web-based, it allowed

respondents to answer questions in the same milieu, at the same time as they were searching. With such a tool, participants may search using any search engine or retrieval system and 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.

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, freeing the researcher from the fundamentally flawed relevance decisions and the need to evaluate in an experimental rather than an authentic setting.

Implications

As with any research, there are limitations to this study. The first is the relatively small document set. 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 larger, more diverse 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 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. 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.

It is hoped that research will continue 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

- Bruce, H. (1998). User satisfaction with information seeking on the internet. *Journal of the American Society for Information Science*, 49(6), 541-556.
- Carliner, S. (2002). Read me first: An Introduction to this special issue. *Technical Communication*, 49(4), 399-404.
- Carr, C. (June, 1992). PSS! Help when you need it. *Training & Development*, 31-38.
- 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.
- Cleverdon, C. W., Mills, J., & Keen, E. M. (1966). *Factors determining the performance of indexing systems* (Vol. 1). Cranfield, UK: Aslib Cranfield Research Project, College of Aeronautics.
- Cole, K., Fischer, O., & Saltzman, P. (1997). Just-in-time knowledge delivery. *Communications of the ACM*, 40(7), 49-53.
- Cooper, W. S. (1973a). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2), 87-100.

- 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.
- 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.
- Cooper, W. S. (2003). *Personal communication*.
- Crotty, M. (1998). *The Foundations of social research: Meaning and perspective in the research process*. London: Sage Publications.
- Gilbert, T. F. (1996). *Human competence: Engineering worthy performance*. Silver Spring, Maryland: ISPI.
- 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.
- 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.
- Hersh, W. (1994). Relevance and retrieval evaluation: Perspectives from medicine. *Journal of the American Society for Information Science*, 46(3), 201-206.
- Kuhlthau, C. C. (1997). Learning in digital libraries: An Information search process approach. *Library Trends*, 45(4).
- Mizzaro, S. (1997). Relevance: the Whole history. *Journal of the American Society for Information Science*, 48(9), 810-832.
- 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.
- Phillips, C. C., & Burbules, N. C. (2000). *Postpositivism and education research*. Lanham: Rowman & Littlefield.
- 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.
- Salton, G. (1992). The State of retrieval system evaluation. *Information Processing and Management*, 28(4), 441-449.
- 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.
- Sparck Jones, K. (Ed.). (1981). *Information Retrieval Experiment*. London: Butterworths.
- 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.
- Yin, R. K. (2002). *Case study research: Design and methods* (3rd ed. Vol. 5). Newbury Park: Sage Publications.