Toward a theoretical framework for information retrieval (IR) evaluation in an information seeking context

Amanda Spink

School of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802

T. D. Wilson

Department of Information Studies
University of Sheffield
Sheffield, UK

Abstract

This paper presents work in progress toward the development of the information retrieval (IR) evaluation measure – *Information Problem Shift* – within a theoretical framework and model for conceptualizing and researching information retrieval (IR) within an information seeking context. The theoretical model consists of a set of situated actions by information-seekers within interactive search sessions with information retrieval (IR) systems over a period of time. First, the paper outlines the IR evaluation measure – *Information Problem Shift* – within the theoretical framework of the model. Second, the paper discusses work in progress in the form of two studies currently collecting data, including the proposed IR evaluation measure - one study of information-seekers using an IR system and another study of information-seekers using the Web. Overall, this research seeks to develop an integrated view of the interactive IR processes by information-seekers and an IR evaluation measure with utility for IR system researchers, designers and information-seekers themselves.

1. INTRODUCTION

This paper presents work in progress toward the development of a user-based IR evaluation measure - *Information Problem Shift* – within the context of an integrated model of information seeking and retrieving. The paper discusses the framework for *Information Problem Shift* and outlines two studies that are currently exploring the utility of this IR evaluation measure. This work is evolving within developments in the field of information retrieval (IR) and IR evaluation, to connect two fairly unrelated bodies of research. One body of research has focused on the systems aspects of IR and IR evaluation. Another body of research has focused on the human, cognitive and interactive aspects of IR and IR evaluation. Concurrently, the field of human information behavior (HIB) is coming closer to interactive IR. HIB research investigates the broader issues related to human processes for seeking and using information. Recent research has seen further integration of interactive IR research and HIB research, based on the mutual interest of both areas in human information related behaviors. The further integration of research and elements of both fields is emerging as an important area of research (Spink, 1998; Vakkari, 1998; Wilson, 1998), particularly to the development of more effective Web and IR systems design and evaluation.

This paper reflects the move to appropriate integration of interactive IR and HIB research to further the area of IR evaluation. The relationship between interactive IR and HIB will further develop as both fields mature theoretically and intellectually, and the emergence of models and research findings from empirical studies of human information behavior and seeking (Wilson, 1997; Ellis, 1989; Kuhlthau, 1993) and interactive IR (Belkin, Cool, Stein & Theil, 1995; Ingwersen, 1992, 1996; Saracevic, 1996a, 1997). The authors of this paper are also conducting two large-scale studies of IR system and Web interaction to extend the development and integration of elements of both fields (Spink, Wilson, Mira '99

Ellis & Ford, 1998) and further develop the proposed IR evaluation measure.

This paper outlines our proposed IR evaluation measure – *Information Problem Shift* – including the limitations of existing measures, criteria for an IR evaluation measure and a theoretical framework and model as the basis of the IR evaluation measure.

2. IR EVALUATION MEASURE

2.1 Limitations of Existing IR Evaluation Measures

The explosive growth of Web search engines, digital libraries and IR systems, and the development of new Web meta-search tools, requires the development of new and more effective IR evaluation measures. Precision and recall present a limited perspective on the value of an information-seeker's interaction with an IR system. Various IR evaluation measures have been proposed, including Su's (1994) "value of the search results as a whole" and Tague and Schultz's (1989) "informativeness". Limited studies have subsequently developed using these measures to have shown their value. Su's proposes that a measure of the value of search results as a whole has some utility as an IR evaluation measure. However, it is not clear what "value" means in relation to an information-seeker. If the value of my search results as a whole is 8, what does that mean? Value in relation to what? The same can be said of the Tague and Schultz's (1989#) measure of "informativeness". Informativeness in relation to what? These measures lack relational strength and validity.

We propose that effective IR evaluation measures should be based on the reality of human interaction with IR systems. Therefore, we focus on the following questions – *What are meaningful criteria for IR evaluation measures? What is a meaningful IR evaluation measures for information-seekers? What is important to measure and how to measure it?*

The next of the paper discusses criteria for the development of useful IR evaluation measures.

2.2 Criteria for IR Evaluation Measures

We propose that a meaningful IR evaluation measure must be useful to IR researchers and designers, and should be meaningful and useful for people using IR systems by measuring what is important to those information-seekers in the form of a self-assessment tool. We propose that:

1. Effective IR evaluation measures must be meaningful and important for information-seekers
2. What is important to information-seekers is the resolution of their information problem
3. To resolve their information problem, information-seekers move through the *changes or shifts* in their information seeking process.
4. If information-seekers interact with IR systems, then an IR evaluation measure must relate the effectiveness of their IR system interaction to shifts or changes in their information problem due to their interaction with the IR system.
5. An IR evaluation measure must also be a self-assessment tool .
6. An important IR evaluation measure for information-seekers is their *Information Problem Shift*.

The IR evaluation measure – *Information Problem Shift* – is based on a model developed from interactive IR research that reflects the process of human interaction with IR systems as they progress through their information-seeking process. The next section of the paper outlines the model upon which we base the framework for our IR evaluation measure.

3. MODEL

The integrated model, displayed in Figure 1, includes a set of situated actions by information-seekers within interactive search sessions with IR systems over a period of time.

Figure 1. Model of situated actions, interactive sessions and time..

This model extends and integrates a model of relevance level, region and time developed by Spink (1998), Spink, Greisdorf and Bateman (1998a,b) and a model of human information seeking developed by Wilson (1997; 1998).

- *Time* is represented by movements or shifts during interactive search episodes, include tactics, information problem, strategies, terms, feedback, goal states, or uncertainty, and between searches.
- *Interactive search episodes* are represented by interactive IR models, including Ingwersen (1992,

1996), Belkin, Cool, Stein and Theil (1995), and Saracevic (1996b, 1997).

- The *set of situated actions* includes actions, decisions and judgments during an interactive search episode, e.g., relevance, magnitude or strategy feedback, tactics, search strategies, or search terms within a search episode. Sets of situated actions that occur during interactions.

Therefore, sets of situated actions may occur during each interactive search episodes that take place over a period of time.

In more general terms, the development of an integrated model provides an important framework for the development of theoretical and empirical research to integrate interactive IR research and develop IR evaluation measures within information-seeking contexts, and explore their interactive search episodes within their changing information-seeking contexts. We seek to develop this model further. *However, we claim that effective IR evaluation measures must take account incorporate research findings that show IR interactions taking place within the context of information-seeking behaviors.*

Therefore, we now discuss each facet of the model to develop a framework for an integrated view of the interactive search processes within changing information-seeking contexts as the basis for the IR evaluation measure – *Information Problem Shift*.

4. TIME

An IR evaluation measure must account for the element of time in information-seeking behavior. The measure – *Information Problem Shift* – includes consideration of time and accounts for the effect of the changes and shifts that occur at the IR interaction level (Robins, 1998; Xie, 1997) that affect the shifts at the information problem level. The set of situated actions during IR interactions occur over a period of time. For example, judgments during an evolving information-seeking process or during successive search episodes. Each set of situated actions may be plotted within four attributes: (1) interaction time, (2) successive searching time, (3) information-seeking time, and (4) problem solving time.

As shown in the model, a *period of time* may be represented by an information-seekers:

1. *Problem solving processes*, represented in Wilson's (1998) problem-solving model of information-seeking behavior in which interactive search episodes provide the information inputs to the problem solving process through which the information-seeker's uncertainty level is reduced,
2. *Information seeking stages*, represented in the model by Kuhlthau's Information Search Process Model (1993), or
3. *Successive searches over time* related to the same or evolving information problem (Spink, 1996).

Time may be plotted from the initiation of an information-seeker's information problem, including the measures associated with the attributes of searches and judgments, in a visual model.

4.1 Problem-Solving Time

Information seeking is integrated with information-seeking time in Wilson's Problem-solving Model (1997, 1998) and related to Kuhlthau's Search Process Model (1993) as a highly developed model of the information seeking process. Other information-seeking or behavior models (e.g., Ellis, 1989) could be adapted and integrated in the time dimension. Wilson's Problem-Solving Model presents information-seeking behavior as goal-directed behavior, with the resolution of the problem and/or the presentation of the solution as the goal. In moving from each of the stages of *problem identification*, *problem definition*, *problem resolution*, and *solution presentation*, 'uncertainty' must be resolved and individuals are seen as engaging in interaction episodes with information sources (including people and other sources as well as IR systems) to resolve uncertainty. Of, course, the attempt to resolve uncertainty may actually increase it and, therefore, the model provides for feedback at each stage.

We argue that shifts and changes occur at different levels for information-seekers over time. First, a single or series of

successive search occur during an information-seeking process within a problem solving process. One level feeds into the other toward the information problem level. Therefore, the measure – *Information Problem Shift* – is measuring a change in the human problem solving process that is affected by actions at the search and information-seeking level.

To extend the model, the next level within the facet of time is the interactive search session.

5. INTERACTIVE SEARCH SESSIONS

IR interactions related to the single search episode can be represented in the model by different theoretical interactive IR models – such as Ingwersen’s Cognitive Model of IR Interaction (1992,1996), Belkin, et al., (1995) Episodic Interaction Model, or Saracevic’s Stratified Model of IR Interaction (1996a, 1997), or a combination of elements of all interactive IR models. Therefore, as interactive search sessions occur they exist within the context of time facets such as successive searches, information-seeking process and information problem solving.

To extend the model, the next level within the facet of time and the interactive search session is the set of situated actions.

6. SET OF SITUATED ACTIONS

The set of situated actions includes actions, decisions and judgments during an interactive search episode, e.g., relevance, magnitude or strategy feedback, tactics, search strategies, or search terms within a search episode. Situated actions occur and form part of interactive IR episodes that occur within information-seeking and then problem solving time. A complete model would include all situated actions during an interactive search episode. In the model shown, we explore a specific set of situated actions related to relevance judgments (Spink & Greisdorf, 1997a,b; Spink, Greisdorf & Bateman, 1998a,b). Some specific situated actions, displayed in Figure 1, include relevance judgments.

6.1 Degrees of Relevance

The degrees of relevance are situated within one of four relevance regions in Figure 1 - highly relevant, partially relevant, partially not relevant, and not relevant. Therefore, the region of an information-seeker’s relevance judgment can be situated as to relevance level and relevance degree. For example, an information-seeker may judge a retrieved item highly relevant based on the relevance level of topicality. The ability to plot these cognitive relations by inference is attributes of the second dimension in the set of situated actions, the information-seeker’s region of relevance attributed to these relations or non-relations. This second attribute also contains positive and negative aspects that can be labeled and depicted graphically.

In the model the relevance judgments is depicted within the four regions: (1) highly relevant, (2) partially relevant, (3) partially non-relevant, and (4) not relevant. The distinction between the *partially relevant* quadrant and the *partially not relevant* quadrant in Figure 1 can be operationally defined as follows:

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

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

Researchers within the region of relevance track investigate appropriate ways to measure the region of information-seekers’ relevance judgments - from highly relevant to non-relevant. These judgments are often related to other factors such as the apriori definition of relevance or order of the citations. IR researchers often use triadic categorical scales for relevance judgments (e.g., relevant/partially relevant/not relevant), but collapsed relevance judgments into binary scales

- relevant/not relevant - to simplify the calculation of precision and recall measures. This approach assumes that no information is lost in the process, and that partial relevance is the same as high relevance. Many studies have focused on binary (relevant/not relevant) relevance judgments and measures and collapsed relevant and partially relevant judgments together during their analysis to form the binary scale - relevant and not relevant (Spink, Greisdorf & Bateman, 1998a,b).

For a finer grain analysis, many more regions of relevance can be delineated as the granularity of relevance regions is sharpened. An overlay of the two dimensions (level and region) of a relevance judgment are represented on the set of situated actions. An information-seeker may also makes a relevance decision at a specific point in time during or after the IR interaction, and a graphical representation of such decisions related to retrieved texts can also be plotted.

The next section of the paper outlines facets of the IR evaluation measure – *Information Problem Shift*.

6.2 Information Problem Shift

The bottom line for people who interact with IR systems is not the number of items retrieved or the precision of the search. What information-seekers care about is how they are progressing toward resolving their information problem. Information-seekers' primarily care about their own personal information problem. Maybe the IR system interaction lead to a complete resolution of the their information problem, or a partial resolution, or a slight or major change in their information problem. The IR system itself is relatively secondary in the reality of their information behaviors. A key to understanding this point by IR systems designers may be to refer to *information-seekers* instead of *users*.

We propose that: *the effectiveness of an IR system can be measured in terms of the change or shift in the human information problems due to their IR system interaction*. IR system effectiveness can be measured as a shift by an individual information-seeker or an aggregate of information-seekers. *Information Problem Shift* may be assessed and operationalized by measuring the change in an information-seeker's information problem stage by measuring their information problem stage before and after their interaction with an IR system. A major weakness of existing IR evaluation measures is their inability to reflect changes or shifts, e.g., changes in an information-seeker's understanding of their information problem due to their interaction with an IR system. Meaningful measures must involve data collected from the information-seeker BEFORE and AFTER their IR interaction. In this case, we are actually measuring a change, not just collecting data after an interaction. Collecting data before their IR interaction provides a benchmark for comparison with the data collected after the IR interaction. IR evaluation measures that ONLY measure AFTER an IR interaction are relatively limited.

By measuring the information problem stage before and after their interaction with an IR system, we can measure the impact of the IR system interaction on the information problem solving process. Of course, the effectiveness of the IR system is realized in the interaction in the context of specific situated actions and cognitive, problem and knowledge states during the interaction. However, if an information-seeker does not experience some type of shift in information problem process - represented by shifts in cognitive, problem and knowledge states, then the IR system interaction has not been effective.

However, as researchers, we also seek to form and develop conceptualizations of the phenomena we are observing. Therefore, the next section of this paper presents a model as a theoretical framework for the IR evaluation measure – *Information Problem Shift* - within an information-seeking context. In other words, we try to explain the theoretical basis for our measure within the framework of research findings represented in the model we presented.

7. OPERATIONALIZATION STUDIES

We propose that the IR evaluation measure - *Information Problem Shift* - be conceptualized and operationalized as:

Information Problem Shift (IPS) = Information-seekers' information problem stage after their IR Interaction (AIPST) subtracted from their information problem Stage before their IR interaction (BIPST).

$$\text{IPS} = \text{AIPST} - \text{BIPST}$$

For example, on an 100mm line, if the information-seeker's information problem stage before their IR interaction (BIPST) was 45/100 and their information problem stage after their IR interaction (AIPST) was 85/100 - then their *Information Problem Shift* would be 40.

The model and the IR evaluation measure - *Information Problem Shift* - are being further developed by two theoretical and empirical studies. The first study of information-seekers' interaction with IR systems is currently being conducted by Dr. Amanda Spink at the University of North Texas in collaboration with researchers at the University of Sheffield (Spink, Wilson, Ellis & Ford, 1998). This project is based in the School of Library and Information Sciences at the University of North Texas and is funded by a National Science Foundation POWRE Grant and at the Department of Information Studies at the University of Sheffield (UK) and is funded by a grant from the British Library Research and Innovation Center. A second study of information-seekers' interaction with a web meta-search tool users is currently being conducted at the University of North Texas (Spink, Greisdorf, Goz & Chang, 1999), in collaboration with NEC Research Institute (Princeton) using their Web meta-search tool – INQUIRUS (Lawrence & Giles, 1998). These studies are:

1. Exploring the shifts and changes in IR system and Web interactions during information seeking processes, including changes in uncertainty, relevance judgments, information-seeking stage and information problems.
2. Gathering and plotting relevance judgment data, and other user judgment data, on the model for analysis.

In both studies, information-seekers' complete a pre-search assessment form, conduct a search (the search log, relevance judgments and think aloud protocols are captured) and complete a post-search assessment form.

Pre- and post-search assessments of various variables are captured on a 100mm line, including:

1. Information problem
2. Personal knowledge
3. Uncertainty stage

Other measures include:

1. Problem solving stage – four (4) stage measure
2. Information seeking stage – six (6) stage measure

Search characteristics and demographics are also recorded. This data will allow researchers to examine the relationship between these variable, including the before and after IR interaction assessments to examine the utility of the measure – *Information Problem Shift* – particularly in relation to other IR evaluation measures such as precision. These studies are attempting to develop IR evaluation measures that are meaningful to information-seekers, researchers and designers. An outcome of the current studies will be the development and testing of an approach to IR evaluation and measures. The goal of the research is to contribute to general IR evaluation theory, models, measures and techniques, and develop a series of IR evaluation measures of value as self-assessment tools. These tools will be based on the model, theories and empirical research at the nexus of IR and human information behavior research. However, the strength of an IR evaluation tool is based on the strength of the models that underpin its development.

8. CONCLUSIONS

The proposed IR evaluation measure and model have strengths and weaknesses. A major strength of the IR evaluation measure – *Information Problem Shift* – is its firm basis in information-seeking and retrieving behavior research. It is an information-seeker based measure. A current weakness is the need to further develop the operationalization and testing of the measure. An important strength of the model is the focus on the big picture and the drawing together of major information-seeking and retrieving concepts, such as situated actions, relevance, IR interaction and time. Other concepts, such as feedback, representation, information problem, and context, may also be incorporated into the framework. The model can integrate existing and future research and models from IR and information-seeking. A further strength is the framework provided for gathering, plotting and testing data.

Strengths can also be weaknesses. The general framework of the model tends to focus on major dimensions and not specific differences in information-seeking contexts. The model is also specifically related to the IR context, not information-seeking in general. A major issue is the problem of combining aspects from different modeling approaches, including the purposes and intentions of the selected models and their components. Despite these limitations, the authors believe their work in progress toward a framework, model and proposed IR evaluation measure is a reasonable and heuristic approach from which to proceed with theoretical and empirical research in IR evaluation.

ACKNOWLEDGMENTS

This paper resulted from studies funded by a University of North Texas Research Initiation Grant 1993-94, UNT/SLIS Research Grant 1997-98, a National Science Foundation POWRE Grant 1998-99 and a British Library Research and Innovation Center Grant. The authors also acknowledge the valuable comments by Tefko Saracevic of Rutgers University, and Howard Greisdorf and Judy Bateman of the University of North Texas, and the anonymous MIRA reviewers.

REFERENCES

- Belkin, N. J., Cool, C., Stein, A., & Theil, U. (1995). Cases, scripts and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9(3), 379-395.
- Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation*, 45, 171-212.
- Ingwersen, P. (1992). *Information retrieval interaction*. London: Taylor Graham.
- Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1), 3-50.
- Kuhlthau, C. C. (1993). *Seeking meaning: A process approach to library and information services*. Norwood, N. J: Ablex Publishing.
- Robins, D. (1998). Dynamics and dimensions of user information problems as foci of interaction in information retrieval. *Proceedings of 61st Annual Meeting of the American Society for Information Science*, 35, 327-341.
- Robertson, S. E., & Hancock-Beaulieu, M. M. (1992). On the evaluation of IR systems. *Information Processing and Management*, 28(4), 457-466.
- Saracevic, T. (1996a). Interactive models in information retrieval (IR): A review and proposal. *Proceedings of the 59th Annual Meeting of the American Society for Information Science*, 33, 3-9.
- Saracevic, T. (1996b). Relevance reconsidered:•96. *Proceedings of COLIS 2: Second International Conference on Conceptions of Library and Information Science: Integration in Perspective, October 13-16, 1996*. The Royal School of Librarianship, Copenhagen - Denmark. (pp. 201-218).
- Saracevic, T. (1997). Extension and application of the stratified model of information retrieval interaction. *Proceedings of the Annual Meeting of the American Society for Information Science*, 34,
- Spink, A. (1996). A multiple search session model of end-user behavior: An exploratory study. *Journal of the American Society for Information Science*, 46(8), 603-609.

- Spink, A. (1998). Toward a theoretical framework for information retrieval within an information-seeking context. *Proceedings of the 2nd International Information Seeking in Context Conference, August 12-15, 1998. Sheffield, UK: University of Sheffield, Department of Information Studies.*
- Spink, A., Bateman, J., & Jansen, B. J. (1998). Users searching behavior on the EXCITE Web search engine. *Proceedings of the 1998 National Online Meeting, May 12-14, 1998* (pp.375-386).
- Spink, A., & Greisdorf, H. (1997a). Partial relevance judgments and changes in users information problems during online searching. *Proceedings of the 60th Annual Meeting of the American Society for Information Science, November, 1997. Washington, D.C.* (pp.111-122).
- Spink, A., & Greisdorf, H. (1997b). Users• partial relevance judgments during online searching. *Online & CD-ROM Review, 1*(5), 271-280.
- Spink, A., Greisdorf, H., & Bateman, J. (1998a). From highly relevant to not relevant: Examining different regions of relevance. *Information Processing and Management, 34*(5), 599-622.
- Spink, A., Greisdorf, H., & Bateman, J. (1998b). Successive searching behavior during interactive information retrieval (IR): Development of a new line of research. *Proceedings of the 26th Annual Conference of the Canadian Association for Information Science, Ottawa, Canada, June 3-5, 1998* (pp. 401-415).
- Spink, A., Wilson, T. D., Ellis, D., & Ford, N. (1998). Modeling users• successive searches: A National Science Foundation/British Library Study. *D-Lib Magazine*, April 1998. CNRI.
- Vakkari, P. (1998). Information source use as a function of organizational information requirements, task complexity and information types: Integrating theories of media richness, task complexity and uncertainty of information seeking. *Proceedings of the 2nd International Information-Seeking in Context Conference, August 12-15, 1998. Department of Information Studies, University of Sheffield.*
- Tague, J., & Schultz, R. (1989). Evaluation of the user interface in an information retrieval system: A model. *Information Processing and Management, 25*(4), 377-389.
- Wilson, T. D. (1981). On user studies and information needs. *Journal of Documentation, 37*, 3-15.
- Wilson, T. D. (1997). Information behavior: An interdisciplinary perspective. *Information Processing and Management, 33*(4), 551-572.
- Wilson, T. D. (1998). Models of information-seeking behavior research. *Proceedings of the 2nd International Information-Seeking in Context Conference, August 12-15, 1998. Department of Information Studies, University of Sheffield.*
- Wilson, T. D. (1998). Models in information behavior research. *Journal of Documentation,*
- Xie, H. (1997). Planned and situated aspects in interactive IR: Patterns of user interactive intentions and Information seeking strategies. *Proceedings of the 60th Annual Meeting of the American Society for Information Science, 34*, 101-110.