A Model for Quantitative Evaluation of an End-to-End Question-Answering System

Nina Wacholder
School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ 08901. E-mail: nina@scils.rutgers.edu

Diane Kelly
School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

Paul Kantor, Robert Rittman, Ying Sun, and Bing Bai
School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ 08901

Sharon Small, Boris Yamrom, and Tomek Strzalkowski
Department of Computer Science, University at Albany, Albany, NY 12222

We describe a procedure for quantitative evaluation of interactive question-answering systems and illustrate it with application to the High-Quality Interactive Question-Answering (HITIQA) system. Our objectives were (a) to design a method to realistically and reliably assess interactive question-answering systems by comparing the quality of reports produced using different systems, (b) to conduct a pilot test of this method, and (c) to perform a formative evaluation of the HITIQA system. Far more important than the specific information gathered from this pilot evaluation is the development of (a) a protocol for evaluating an emerging technology, (b) reusable assessment instruments, and (c) the knowledge gained in conducting the evaluation. We conclude that this method, which uses a surprisingly small number of subjects and does not rely on predetermined relevance judgments, measures the impact of system change on work produced by users. Therefore this method can be used to compare the product of interactive systems that use different underlying technologies.

Introduction

We describe a method for performing a realistic evaluation of an end-to-end question-answering (QA) system designed for use by professional information seekers. QA systems take as input free-text questions and produce as output information that provides specific and accurate answers to these questions. End-to-end systems support all stages of the QA process, beginning with the user's entry of free-text questions and ending when the user has gathered as much information as desired and has assembled the answers into a document. Since each user's interaction with the system is different, and since there is a variety of arguably correct answers to complex questions, systematic evaluation is notoriously difficult (Voorhees & Tice, 2000b; Voorhees, 2003a; Sparck Jones, 2003).

Our approach incorporates real users, real tasks, and real systems into the evaluation of information access systems (e.g., Saracevic, Kantor, Chamis, & Trivison, 1988; Saracevic & Kantor, 1988a, 1988b; Cool, Belkin, Frieder, & Kantor, 1993; Tague-Sutcliffe, 1996; Jansen, Spink, & Saracevic, 2000; Sparck Jones, 2001). "Real" users have information needs like those the system is designed to satisfy and have experience trying to satisfy such needs; "real" tasks are similar in domain and complexity to the kind of tasks the real users typically take on; "real" (computer) systems are far enough along in their development that they can be used accomplish the real tasks. We refer to this approach as the *R*eal *U*sers, real *T*asks and real *S*ystems (RUTS) paradigm.

The RUTS paradigm measures (a) the quality of the report produced by information seekers at the end of the information-gathering process, and (b) the experience of information seekers during the information-gathering process. This model shifts the emphasis of assessment away from the quality of answers to individual questions and toward measurement of the product of the information-seeking process; this approach circumvents the need to rate systems based on the correctness of paradigmatic answers.

Received March 15, 2005; revised February 1, 2006; accepted July 17, 2006

 ● Published online 23 April 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20560

The key instrument that we use to obtain consistent judgment of the reports, which may be quite different from one another, is cross-evaluation (Sun, Kantor, Strzalkowski, & Wacholder, 2004), in which real users assess each other's reports, including their own. The reasoning behind this approach is that (a) individuals who use a better system will produce better reports than those who use a less good system, and (b) the collective judgments of the users will accurately represent the quality of the report. The cross-evaluation tool was originally developed by two of the authors of this paper for the AntWorld project (Sun & Kantor, 2006). We have supplemented the cross-evaluation with questionnaires that assess how realistic the tasks were and whether the system was ready for use in the participants' normal work environment.

We applied the evaluation method to the High-Quality Interactive Question-Answering (HITIQA) system developed at SUNY Albany (Small, Ting, Shimuzu, & Strzalkowski, 2003; Small et al., 2004) and funded by the Advanced Research and Development Activity (ARDA) as part of the Advanced Question Answering for Intelligence (AQUAINT) program (Maybury, 2002). The work reported here was the first attempt to systematically evaluate an end-to-end QA system sufficiently mature to be tested using real users and real tasks. The evaluation was conducted at two 3-day workshops held a month apart. The results of the evaluation illustrate what can be learned, and provide a baseline for future study of similar systems.

More important than specific information gathered about one system are the development of a protocol for evaluating a complex QA system, the assessment instruments used, and the experience and knowledge obtained in conducting the evaluation. We find that the RUTS approach is effective and efficient. RUTS incurs the expense of involving real users at each iteration of the evaluation. But the experimental design extracts meaningful results from the smallest possible number of participants, and bypasses the arguably impossible requirement to develop a single set of correct answers. We demonstrate that RUTS can be used to compare the effect of system on reports produced with two different versions of the HITIQA system; we anticipate that this approach can also be used to compare systems that employ different underlying technologies.[1]

The HITIQA System

HITIQA (Small et al., 2003, 2004) is a complex end-to-end question-answering system specifically designed to let users ask exploratory, analytical questions with complicated answers that cannot be readily judged as right or wrong. The answers to analytical questions are considerably more complex than the answers to factoid and definitional questions such as those used in the Text Retrieval Conference (TREC) QA track. The answers to factoid questions are relatively short and straightforward and usually are not matters of opinion. For example, the correct answer to the factual question "How many states were in the U.S. in 1974?" should contain the information that there were 50 states. In contrast, an adequate answer to an analytical question such as "How has Russia reacted to the bombing in Kosovo?" may encompass reactions by a variety of individuals and organizations; a brief one-dimensional answer (e.g., "The government was unhappy.") typically does not satisfy the user's information need. An answer of some depth is required; how much depends on the "customer" and on the context in which the question is asked.

HITIQA is designed to be used with a large text corpus whose documents contain partial answers related to one or more dimensions of the question. Users ask open-ended questions to probe the dimensions of the available information and express their information needs using any desired syntax; they then engage in an interactive natural-language dialogue with the system to clarify the information need, and they employ a visual interface to explore the available answer space. HITIQA returns answers in the form of paragraph-size units of text that can be saved into a report.

In support of complex QA, HITIQA incorporates a variety of information-access processes, including information retrieval (to retrieve an initial set of documents that will be subjected to more intense processing), information extraction (to identify references to certain events, to entities such as people, organizations, and locations, and to dates), interactive dialogue (to allow the user and the system to clarify what information the user is interested in), summarization (to briefly represent the content of one or more articles), natural-language generation (to "translate" the answer into a form easily understood by people) and basic report preparation functionality (to keep track of which documents the user intends to use in the draft report and to make it easy to transfer this information into a word-processing program) (Strzalkowski et al., 2005).

From the user's perspective, HITIQA uses two main mechanisms to support QA: (a) an interactive dialogue whose goal is to clarify the scope of the question, and (b) a visual display that helps users explore the answer space. In the example in Figure 1, the user initiates the dialogue with a question about South Africa's weapon programs (1). HITIQA responds by asking about the user's interest in potentially related information (2, 4, 6, 8, 10, 12). Finally, HITIQA directs the user to the available answers (14).

The visual interface, shown in Figure 2, represents graphically the possible answers that the system has found. The balls stand for clusters of related answers and the spokes stand for entities that are discussed in some of the answers. The color of the balls (not shown in this figure) represent HITIQA's estimate of the likelihood that the user will find the answer helpful. After inspecting the available answers, the user may decide to explore a particular dimension in more depth or to shift the focus of the questions. The user controls which answers are displayed by using the mouse to manipulate the

[1]In fact, this method was used in July 2004 to compare several QA systems, including HITIQA, at the ARDA Challenge Workshop.

DOI: 10.1002/asi

(1) User: What is the status of South Africa's chemical, biological, and nuclear programs?
(2) System: We have found another type of information retrieved for your query. Can we also add material on the transfer of weapons and intelligence between South Africa and other countries to your answer?
(3) User: Yes
(4) System: Should we add material on imports of weapon materials to South Africa to your answer?
(5) User: Yes
(6) System: Do you want to see material on transferring weapons between South Africa and other countries?
(7) User: Yes
(8) System: Would you be interested in supplying of weapons between South Africa and other countries?
(9) User: Yes
(10) System: Are you also interested in background information on the weapons development program in South Africa?
(11) User: Yes
(12) System: May we also show you information on possible treaty violations involving South Africa?
(13) User: No
(14) System: Please go to Answer tab.

FIG. 1. Sample interactive dialogue.

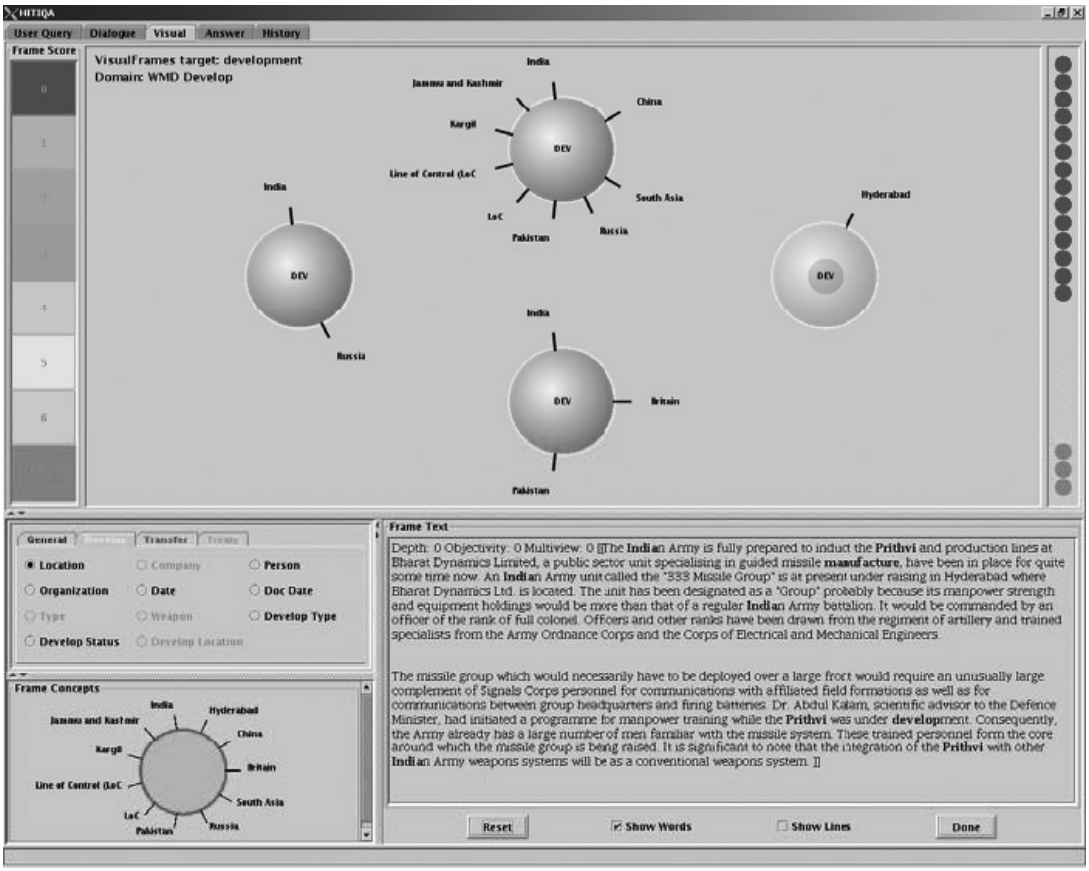


FIG. 2. Sample visual display of the answer space.

balls, spokes, and colors.[2] As a result of this interaction, the user's notion of a suitable answer evolves over the course of an extended question-answering process. Users can end their session either when they are satisfied with the answers they have obtained, when they feel that they have found all the useful answers available from the text collection, or when time runs out. For more details about the HITIQA system architecture the reader is referred to Strzalkowski et al., 2005.

Related Work

In this section, we review previous work in evaluation of information-access systems. We argue that the Cranfield paradigm, so successfully used to evaluate and stimulate progress

[2]Because the focus of this article is on the method of evaluation, we do not provide further details about the display. See Strzalkowski et al. (2005) for more details.

DOI: 10.1002/asi

in information retrieval (IR) systems, as well as some other applications, cannot be used to assess complex analytical QA systems because they depend on accurate prejudgment of answers.

The Cranfield Paradigm

The Cranfield method for evaluation of IR systems equates effectiveness with good document ranking: systems that rank relevant documents most highly are considered most effective (Cleverdon, 1967/1997). Efficient and economical, the Cranfield method has been used in the National Institute of Standards (NIST) sponsored TREC evaluations (Harman, 1992; Voorhees, 2003a). NIST has assembled and made available at no cost to researchers a sizeable corpus of reusable materials that include clean documents, thoughtfully designed topics, and carefully prepared relevance judgments (Baeza-Yates & Ribiero-Neto, 1999). Numerous organizations have participated in TREC and the materials have been used extensively in other IR and natural-language-processing research. By making it possible to economically compare systems using vastly different technology, the TREC-Cranfield activity has been credited with stimulating significant advances in IR systems.

However, several properties of the Cranfield paradigm limit its value for evaluation of complex analytical QA systems such as HITIQA: (a) Real users are a confounding variable; (b) So far, attempts to develop a reusable answer key for QA evaluation have not succeeded, even for factoid and definitional questions; and (c) Fixed, mutually independent assignment of relevance values are generally not suitable for interactive environments where users' information needs change in response to the system.

Real users as a confounding variable. The Cranfield model is convenient; it is conducted in "a laboratory environment … freed as far as possible from the contamination of operational variables" (Cleverdon, 1991, p. 7). Exactly three components, a corpus, a set of topics, and a set of relevance judgments for each topic for each document, are needed to conduct an evaluation. Real users, whose variability may confound results, are represented by a single set of relevance judgments. Many researchers have noted the potential significance of the absence of users in the Cranfield model; among those who have demonstrated that characteristics of the information seeker and the information-seeking context significantly affect judgments about document relevance are Saracevic (1975/1997), Beaulieu, Robertson, and Rasmussen (1996), Harter (1996), Spink (1996), Harter and Hert (1997), Vakkari and Hakala (2000), Borlund (2003), and Vakkari (2003).

The addition of the interactive track to the TREC tasks represented an effort to devise a method to evaluate interactive systems based on paradigmatic relevance judgments (Hersh & Over, 2000, 2001, 2002). But Hersh et al. (2000) and Turpin and Hersh (2001) show that although better IR systems produce better results, as measured by precision and recall, they do not translate into better performance on an information-seeking task by individuals using the "better" IR systems.

By TREC 7 (1998), the attempt to conduct cross-site comparison of systems based on a single answer set was dropped in favor of an evaluation method that simply counted the instances of the desired information (Over, 2001); in TREC 10 (2001), the last time the interactive track was held, the participating teams conducted observational studies and attempted separately to assess "some aspect of user searching performance" (Hersh & Over, 2002, p. 3). For a system (such as HITIQA) that aims to interact with users, it makes little sense to base an evaluation on a paradigm that treats real users as a confounding variable.

The difficulty of creating a reusable evaluation corpus. Voorhees and Tice (2000a, 2000b, 2000c) and Voorhees (2002, 2003a, 2003b) extensively document the complications of producing a reusable answer key for factoid and definitional questions. The fundamental problem is that accurate measurement of precision depends upon a reliable count of correct answers. Complications in identifying correct answers arise from the unexpected ambiguity of questions, from the multitude of partial answers, and from differences in the use of implicit information or world knowledge to determine the correct answer. Voorhees (2003a) and Sparck Jones (2003) provide numerous examples of the problem, even for questions with apparently straightforward factual answers. These reports are especially persuasive because they have been produced by individuals who contributed to the successful use of the Cranfield paradigm for IR evaluation and for TREC. Voorhees (2003a) concludes that efforts to develop a high-quality, reusable evaluation corpus for factoid and definitional questions has not yet succeeded and suggests that figuring out how to do this is a key priority. Given the difficulty of preparing a reusable answer key for factoid questions, the effort to develop one for analytical questions seems doomed from the outset.

Idealized assumptions about document usefulness. The Cranfield method depends on fixed judgments about document relevance; these judgments are idealized in that they are independent of any actual information-seeking task. This idealization relies on two assumptions: (a) the relevance of each document is assumed to be independent of that of all other documents, and (b) the user's needs are assumed to be static (Voorhees, 2001). These idealized assumptions are justified when a single document or chunk of text can be assumed to provide the whole answer, so that the users' information needs do not evolve during the information-seeking process. But the answers to complex analytical questions are multidimensional, and typically include information from multiple sources. Users work with the system over an extended period of time, during which their information needs naturally change. HITIQA's interactive dialogue and visual display are specifically designed to support and encourage this process.

Considering all of these limitations, we conclude that the Cranfield model is unsuited to evaluation of a system like

DOI: 10.1002/asi

HITIQA that has been specifically designed to respond to the evolution of the user's information needs during the process of answering a complex question.

Automatic Assessment of QA Systems

Another potential method for efficiently assessing the product of QA systems is automatic answer evaluation. For example, Breck et al. (2000) developed Qaviar, a system for automatically evaluating QA systems. Qaviar assesses answers by comparing them to answers provided by a (TREC) assessor. This system therefore automatically judges as "wrong" answers that are significantly different from those provided by the assessors. This evaluation method is useful for development purposes; developers can measure the impact of changes in the system on its performance on a fixed set of questions. But this evaluation, which is not concept-based, cannot recognize answers not originally included in the list developed by the TREC assessors.

Hirschman, Breck, and Burger (1999) use a reading comprehension system to test QA. They obtain graded reading tests with answers and evaluate the QA system's ability to find the correct answers. The problem with this method of evaluation is that it is "closed": the QA system needs only to find a single correct answer in a relatively short document.

Automatic essay-grading systems are another method of evaluating open-answer questions. For example, a method for automatic grading of essays is described by Burstein and Chodorow (1999), who consider their system successful if it assigns a score that is either identical to or just one point different from the rating (on a scale of 1 to 6) assigned by a human judge. By this measure, their system succeeds 92% of the time. This method depends on the preparation, for each question, of a set of sample essays of varying degrees of quality; learning algorithms then use a set of features to automatically grade the essays.

These automatic evaluation methods are all relatively useful when answers are known or fixed, but they are not capable of recognizing high-quality answers not foreseen by designers of the evaluation.

The Wizard of Oz Approach

Because of the anticipated difficulty of assessing the QA systems sponsored in the AQUAINT program, ARDA held an initial evaluation called the Wizard of Oz assessment (Dahlback, Jonsson, & Ahrenberg, 1993; J. Scholtz, personal communication October 4, 2004). Analysts sat at computer terminals located at NIST and used a Web interface to ask questions of the system. A human intermediary translated the question into a query and submitted it to the system. If the system returned a correct answer, the intermediary relayed the answer to the analyst; if the answer was partially correct, the intermediary modified the answer and relayed it to the analyst; if the system's answer did not seem appropriate, the intermediary passed on to the analyst an answer that the system would have returned if it had successfully answered the question. The Wizard of Oz method can be very useful for assessing interfaces, but it does not provide data for evaluating reports written by users.

In conclusion, neither the Cranfield paradigm nor the other evaluation methods discussed above can be adapted to the evaluation of complex QA systems because of the apparently insurmountable difficulty of producing prior sets of answers to complex questions.

Toward Realistic Evaluation

The RUTS paradigm circumvents the need to rate systems based on idealized assumptions about answer correctness by comparing the reports produced by the interaction of the user and the QA system. In principle, this method allows comparison of output of systems that use widely different technologies. The validity of the assessment is strengthened by the requirement that the interaction involve real users, real tasks, and at least two real systems.

Real Users, Real Tasks, and Real Systems

Conducting a scientific evaluation under the RUTS paradigm entails operationalization of the three major components: real users, real tasks, and real systems. It also depends on the availability of a tool that reliably measures the reports. Cross-evaluation is a black-box method in which participants rate the quality of their own and their colleagues' reports. The rationale for this technique rests on two assumptions: (a) individuals who use a better information-access system will produce a better product than those who use a less good system, and (b) experts are able to recognize good work when they see it.

Real users. For the HITIQA pilot, identifying appropriate users was straightforward. HITIQA was developed specifically for use by intelligence analysts, i.e., professional information seekers who collect information about and perform analyses of complex events, problems, and technology in the national interest. Krizan (1999) defines intelligence as "carefully analyzed information tailored to specific users" (p. 1) and draws an analogy between national intelligence and business intelligence. Initially we knew relatively little about the analysts for whom the system was being developed; it was therefore particularly important to have professional analysts perform the assessment. From the point of view of the study of human information seeking, the behavior of analysts is distinguished by the urgency of the task and the analysts' correspondingly high motivation to quickly make an accurate analysis of very complex issues, based on information of varying degrees of reliability and completeness.

We were fortunate to have intelligence analysts employed by the U.S. government participate in the evaluation: four attended the first workshop and three of these four returned for the second. The analysts participated in the evaluation as part of their jobs; specifically they were assigned the task of helping to determine whether the HITIQA system would be

DOI: 10.1002/asi

useful in their normal work environment. We do not claim that these users comprise a representative sample. The number of participants is small due both to our desire to interview and observe each participant and to the expense of using professional analysts in the process. A major advantage of the cross-evaluation method (described below) is that each report is evaluated by multiple professionals in the field, thereby increasing the power of the analysis.

Real tasks. The analysts' primary task at work is the preparation of reports, usually written, that are submitted to their managers for revision and approval and then passed up the management chain; these reports sometimes reach the highest level of government. The preparation of scenario-based reports was therefore a natural task for the analysts. They spent most of their time at the workshops preparing reports in response to "scenarios": complex questions that entail multiple subquestions.

To ensure that the scenarios were like those that analysts investigate at work, several government agencies collaborated in the development of five primary scenarios used in this workshop, and several back-ups.[3] The collaboration with agencies that have access to the classified reports that analysts actually prepare gave us confidence that the scenarios were suitable. The combination of work task and scenario are similar to Borlund and Ingwersen's (1997) simulated work tasks.

Although the scenarios covered a range of issues, all were on the general topic of weapons of mass destruction, a choice based on the document collection (described below) used in the evaluation. One scenario is shown in Figure 3. The other four scenarios used in the evaluation are presented in Appendix A.

To increase the realism of the experiment, we also needed an appropriate work environment and corpus. This was a particular challenge because the analysts' usual resources, including text corpora, software tools, and reference works are classified and therefore not accessible to us. To obtain an adequate supply of appropriate text to support an assortment of scenarios and to compensate for the absence of classified data, we created a new collection. Taking as a starting point data from the Center for Non-Proliferation Studies (CNS) collected for the AQUAINT Program,[4] we used Google to mine the Web for similar subject matter, by submitting hundreds of queries having to do with international politics and global security (Small et al., 2004). We initially retrieved approximately 3 GBytes of text; after removal of duplicates and filtering, the final corpus was about 1.2 GBytes. This proved sufficient to support two to three hours of use of HITIQA to "solve" the scenarios. After each scenario, participants completed a questionnaire that asked whether the tasks performed at the workshops were like those they performed at work. We see below (Table 2) that participants agreed that they were.

Real systems. In order to simulate comparison of "different" systems, we used two versions of HITIQA in two workshops that took place a month apart. Changes were made to the system during the intervening time in response to participants' reactions at the first workshop. Thus we effectively had a somewhat different system for the second workshop.

The Department of Homeland Security has requested a complete report on the chemical weapon, sarin. This report is due in 5 hours. In your report, include its potency and potential impact on a community, what countries and organizations have been involved in producing it, where these locations are, the production method and how it has developed, who possesses it now, who distributed it (if through trade, what was traded for it?), potential means of use, how can this be integrated into warheads, any known defenses against it, and who is at the greatest threat. Provide any other information that you see relevant.

FIG. 3. Sample scenario.

Evaluation of Report Quality

For our user group, preparation of reports is a natural task. The central question, therefore, is how to evaluate the quality of these reports. We note that the problem facing those who judge the quality of such reports is similar to the problem that faces teachers who grade essays or research papers: on the one hand, the quality of some of these pieces of work is evidently poor; on the other hand, many others are of high quality even though they are quite different. We therefore adopted cross-evaluation, which rests on the assumption that reliable judgments can be made by experts, who recognize good work when they see it. To control for variation in individual judgments, we use multiple judges, multiple scenarios, and multiple reports. We analyze the results with the General Linear Model (GLM) to determine which differences in scores are significant, and to remove the effect of differences among judges, authors, and tasks.

The cross-evaluation model (and its supporting tools) was developed to study the effectiveness and efficiency of a collaborative tool for searching the World Wide Web (Sun & Kantor, in press). That tool, AntWorld, supports the work of subsequent participants by providing access to work by prior participants when that prior work is "sufficiently similar" to the current work (Kantor et al., 1999; Kantor, Boros, Melamed, & Menkov, 1999). The cross-evaluation method was applied to compare two versions of the AntWorld system and one control system.

Three practical issues arise in conducting a cross-evaluation: (a) What criteria are used to assess the quality of the reports?

[3]National Institute of Standards (NIST), Air Force Research Laboratory (AFRL), Center for Non-Proliferation Studies (CNS) and Advanced Research Development Activity (ARDA).

[4]http://cns.miis.edu/dbinfo/about.htm#wmdt

DOI: 10.1002/asi

(b) Who performs the assessment? and (c) How is the effect of the system measured?

Criteria for assessing reports. We (the researchers) had no way of knowing what constitutes a high-quality report. We therefore asked the judges to assess the overall quality of each report. In addition, we applied four criteria that are frequently used to grade student papers:

1. The report covers important ground, i.e., it includes information about aspects of the question that can be obtained from the available data.
2. The report does not include irrelevant materials, i.e., it does not include information that is not helpful in answering the questions.
3. The report is well organized.
4. The report reads clearly and easily.

We used a 6-point scale, where 0 means *worthless* and 5 means *excellent*.

Identification of judges. Because the resources that the analysts use to produce a report, as well as the reports themselves, are classified, we began with the presumption that only individuals who have access to the classified reports are qualified to judge the quality of the reports. In an ideal situation, the analysts' managers, for whom the reports are normally produced, might provide the evaluations. However, analysts are a precious resource and we were fortunate to have access to a small convenience sample consisting of four individuals assigned by their supervisor to participate in the workshops. The decision that the analysts themselves were appropriate judges was based on two factors: (a) The analysts are professional information seekers, hired because of their ability to produce analytical reports; and (b) Individually and collectively, the analysts have extensive experience in producing reports that satisfy their supervisors and so are capable of recognizing good reports when they see them. We asked all analysts to assess their own reports, as well as those of their peers, and estimated the self-judgment bias. To reliably measure the effect of change in the system over time, the same analysts participated in both workshops.

Extracting the effect of the system. In order to find out how each factor affects the analysts' responses, we use an analytical model based on the analysis of variance. We use the General Linear Model available in SPSS software to extract from judgment scores the effects of task, judge, system, and self-bias (details in Sun & Kantor, in press). The operationalization of the RUTS paradigm with real users, real tasks, real systems, and evaluation of system effect on reports produced by multiple experts provides the foundation for a realistic assessment of HITIQA.

The next challenge was to design a valid protocol. We were especially concerned with collecting information about issues that our research team might not have anticipated in advance. In the next section, we report on this protocol.

Experimental Protocol

Having intelligence analysts participate in the evaluation gave us an opportunity to learn how useful they felt HITIQA was, to observe their information seeking behavior, and to observe their reactions to the evaluation method. The evaluation was conducted in two 3-day workshops held a month apart.

Common Elements at Both Workshops

Participants. Four United States Naval Reservists participated in Workshop I. Three of these reservists completed Workshop I in its entirety; one had to leave unexpectedly and was not present on the final day. The three who were present during the whole of Workshop I also returned for Workshop II, which took place the following month.

All four participants were intelligence analysts for the United States government. They were chosen by their supervisors and their participation was treated as a work assignment. There were three males and one female, with varying degrees of experience, ranging from quite junior to quite senior. We were able to observe all of the analysts' interactions with HITIQA and to interview them at various points during the workshop.

Location. The workshops were held at the Institute for Informatics, Logics, and Security Studies at the State University of New York at Albany, where HITIQA was developed. Two rooms were set up with two computers each for the analysts, and individual interviews were conducted in private offices. We conducted the evaluations "live" both to equalize the training to which participants were exposed and to observe participants' reaction to the evaluation process.

Training. Since HITIQA's advanced functionality was unfamiliar to the analysts, we devoted the first day of the first workshop to a group training session that included oral instruction, individual practice, and informal conversation. The oral instruction consisted primarily of a formal presentation of the User's Guide, which had been developed to serve such a purpose. Analysts received a print copy of this guide for reference throughout the workshop. At the end of the training session, analysts completed a "Check Ride," a short, multiple-choice quiz designed to assess their proficiency at using the system. After analysts completed the Check Ride, the group went over the answers, discussing problems and resolving discrepancies. At the beginning of Workshop II, analysts participated in a refresher session, in which changes in the system were described and the analysts had the opportunity to warm up and ask questions. In evaluations of systems with which the users are already familiar, this training would obviously not be necessary.

Scenarios. We used four scenarios at Workshop I; each analyst was assigned two of these scenarios. At Workshop II, each analyst worked on the two scenarios they had not used

DOI: 10.1002/asi

before, and all worked on the fifth scenario. The text of the five scenarios appears in Appendix A.

Report preparation. In defining the report-writing task, we had to acknowledge the differences between their normal work environment and the evaluation situation, and provide analysts with enough instruction to proceed with the task. We asked them to prepare a report as similar as possible to the reports they prepare in their normal work environment. They were told to indicate in their report if they would normally have used a resource not available to them at the workshop, and to compensate for the absence of reliable material by focusing on obtaining answers, regardless of the source of information. The instructions to analysts are shown in Figure 4. Based on discussion with the analysts, the absence of reliable material of was, indeed, perceived as a limitation on report preparation.

Questionnaires. We supplemented the cross-evaluation with two questionnaires designed to elicit information about the analysts' experiences, a post-scenario questionnaire and an exit questionnaire. The post-scenario questionnaire was filled out after each scenario session, and asked analysts to evaluate the scenario and their experiences using HITIQA to complete a report on the scenario. Analysts were asked to assess (a) the realism and difficulty of the scenarios of completing the task; (b) the time required to complete the work task; (c) the usability of the system; and (d) their comfort, confidence, and satisfaction in using the system. There were 16 questions, all using 5-point Likert-type scales. A response option of *not applicable* was also provided. The anchors for these questions depended on the specific content of the question.

The exit questionnaire, which analysts completed at the end of each workshop, included 17 statements designed to assess analysts' overall experiences using the HITIQA system. In particular, these questions asked analysts to evaluate (a) the training process, (b) the evolution of their skill and comfort at using HITIQA over the course of the workshop, (c) system usability and functionality, and (d) the readiness of the system for use as an actual work tool for intelligence analysts. Analysts were asked to respond on a 5-point Likert-type scale, with 5 for *strongly agree* and 1 for *strongly disagree*. Analysts also could choose a *not applicable* response option.

Report Preparation

- Use variant spelling when you look for information. Example: al Qaeda and al Qaida.
- Try to prepare a report like what you would do in your normal work environment.
- If you would normally consult a resource that isn't available, simply say so in the report. But the objective of this workshop is to improve Hitiqa, so spend most of your time using the Hitiqa interface.
- Since we're testing performance of Hitiqa, choose information based on content, not reliability. This is obviously very different than what you would normally do.

FIG. 4. Instructions on report preparation.

Qualitative assessment. We conducted focus groups and individual interviews to obtain a more in-depth understanding of what the participants liked about the system and what parts could be improved. We mention these tools here to provide a full picture of the structure of the evaluation; generally the results are consistent with and add to the information that we obtained from the quantitative tools. An in-depth analysis of the qualitative results will be reported in Kelly, Wacholder, Rittman, Sun, Kantor, et al. (2007).[5]

Improvements to the System Between Workshops I and II

In the month between workshops, the changes to HITIQA primarily involved correction of bugs and improvements in the interface. The brief interval between workshops was a by-product of the two-year funding provided by the ARDA program. The first 21 months of this period were spent developing the system; only at the end of this time was a usable system available for initial testing by the intended users.

From the analysts' perspective, the most significant changes were intended to help them better manage system answers. For example, in Workshop I, analysts complained about having to deal repeatedly with identical portions of text that were returned by the system in response to queries for a given scenario. HITIQA was changed for Workshop II so that the system would not return the same portion of text more than once for a given scenario session. In response to other feedback, links to source documents were added for Workshop II so that analysts could evaluate system answers by inspecting the complete source document, and improvements were made to HITIQA's Save feature. Changes to the visual display gave analysts the option of viewing the visual answer space in only three colors, or in the original eight colors.

Design of Evaluation Instruments for System Comparison

During Workshop II, we explored three approaches to comparing the answers obtained by different analysts and to assessing the quality of the support that HITIQA provides. The primary and most successful is the cross-evaluation method described above. In addition, we experimented with two other approaches, identification of information units and merging, which we describe briefly.

Identification of information units. In an effort to put the analysis of complete reports on a more quantitative basis, we suggested that reports are composed of atoms (or perhaps molecules) that we call information units, and we sought to

[5]The qualitative data provide a more in-depth view of specific successes and failures; we omit discussion of these results from this article because they are not directly relevant to the usefulness of the RUTS paradigm for realistic assessment of quality of an interactive QA system.

DOI: 10.1002/asi

learn from the analysts what they consider such a unit to be. We, as researchers, could not provide an intensive definition of an information unit (by stating in an unambiguous fashion what it means to be such a unit) and so we sought an extensive definition. We asked the analysts to provide this definition at Workshop II by taking all of their reports on each topic, and simply marking texts they felt were information units. Although we anticipated that these units might be nested or overlapping, that problem did not arise in practice. We then asked each analyst to review the other analysts' markings, writing a simple "Y" or "N" to indicate whether they did or did not agree that the marked text constituted an information unit. The results, to be reported elsewhere, confirm that identifying information units is not a simple task.

Merging. The quality of information units (if they exist) can in principle be measured by whether or not they are included in a final report. We therefore asked the analysts to compile the best results from each of the individual reports into a single optimal report. It turned out to take the group of analysts an exceedingly long time to reach agreement on which units to include in the optimal report. Under pressure, we aborted the merging task.

Workshop Schedules

The primary objective of the first workshop was to get some idea of the analysts' comfort using the HITIQA system, to run a pilot test of our method, and to obtain baseline data. The schedule is shown in Appendix B. Most of the first day was devoted to orientation and to training analysts in the use of HITIQA. The second day began with a warm-up exercise; then each analyst completed two scenarios, one long (four hours) and one short (one hour). The rest of the workshop was devoted to interviews and focus groups. Initial analysis of the results of Workshop I satisfied us that the tasks as assigned were realistic and appropriate, and that we could accumulate useful quantitative and qualitative data about the analysts' experience in using HITIQA.

At Workshop II, analysts prepared three reports, spending two and a-half to three hours on each. By the end of the workshop, the three analysts had each completed all of the five scenarios, two at Workshop I and three at Workshop II. The schedule of Workshop II is attached as Appendix C. Only two analysts were present for the third day, but the analyst who left early completed some of the final tasks after returning to home base. The major changes in Workshop II were the reduction of the time devoted to training, and the addition of the three methods for evaluating reports: cross-evaluation, information units, and merging.

The first morning of Workshop II was spent reviewing changes that had been made to HITIQA in response to the analysts' feedback from Workshop I. On the afternoon of the first day, analysts prepared a report and performed the first cross-evaluation. The second day involved six hours of report preparation, two scenario assessments, group work on merging, and individual marking of information units. For the analysts, this day was the most tiring. On Day 3, analysts finished the cross-evaluations for each of the scenarios, and completed the exit questionnaire described above. The day concluded with a group discussion about HITIQA and about how to conduct similar evaluations in the future.

Results

In this section we report on the results of the cross-evaluation and the questionnaires. The analysis illustrates how these results can be useful in evaluation of an interactive QA system and can provide a baseline for comparison of future evaluations of similar systems. We first report on the analysts' assessment of the realism of the task, including scenarios and topic preparation, then we report on the results of the cross-evaluation, and finally we examine the analysts' perception of the system.

Realism of the Task

The post-scenario questionnaire asked analysts, after completing each scenario, to assess the experience on a scale of 1 to 5 (5 being the best). In terms of how realistic they were, the scenarios were rated across workshops at an average of 4.24. The average rating of the difficulty of completing the reports, as compared to the analysts' normal work tasks, was 3.12, above the midpoint (on a scale where 1 was *easier than normal work tasks* and 5 was *harder than normal work tasks*). These results suggest that our efforts to develop realistic scenarios of average difficulty succeeded. Given this crucial information, we proceed to discuss the rest of the results.

Cross-Evaluation: The Effect of the System on the End Product

The most notable result is that the evaluation shows the quality of the reports prepared at the second workshop to be significantly better than the quality of those prepared for the first workshop.

For the purpose of comparison, we first conducted a one-way analysis of variance (ANOVA). The result shows that the improvement for three of the criteria for rating reports (overall, covers important ground, is well organized) was significant at a .99 level; the improvement for the other two criteria (avoids irrelevant materials, reads clearly and easily) was significant at a .95 level (These are treated here as separate set of criteria, and we do not correct for multiple comparisons. Effects in a group of 5, significant at .99, will be significant at the .95 level after Bonferroni correction).

Table 1 therefore shows the reports from Workshop II to be significantly better than those from Workshop I on all five criteria. However, we must be careful about drawing conclusions because four other factors may affect the scores assigned by the participants. Each report has been (a) produced by a particular judge, (b) for the work of a particular participant,

DOI: 10.1002/asi

TABLE 1. One-way analysis of variance (ANOVA) in judgment across workshops.

Criterion	Mean – Workshop I	Mean – Workshop II	Difference	F(1,40)	Significance
Covers important ground	3.127	4.085	.9688	12.67	.0100[b]
Avoids irrelevant materials	2.5367	3.6248	.95	5.85	.0402[a]
Is well organized	2.6558	3.886	1.2327	12.23	.00[b]
Reads clearly and easily	2.765	3.726	0.961	7.68	.0201[a]
Overall rating	**2.823**	**3.920**	**1.1007**	**12.07**	**.0100[b]**

[a]Statistically significant at .95 level.
[b]Statistically significant at .99 level.

(c) on a particular topic, and (d) using a particular instance of the system. These are the independent variables.

We used the GLM to isolate the effect of these factors. We introduced four parameters representing the main effects of these variables, which are treated as fixed effect factors, shifting the mean of the scores for the cases to which they apply. We also introduced a self-evaluation factor that distinguishes cases in which an analyst is judging his or her own report from other cases. Specifically, we used the analytical model shown in Equation 1.

Equation 1: Analytical Model

$$V(j, a, t, s, b) = \lambda^0 + \lambda_j^J + \lambda_a^A + \lambda_t^T + \lambda_s^W + \lambda_b^B + e$$

where
V = measurement scores
J = judge variable
A = author variable
T = task variable
W = workshop variable
B = self-judgment bias variable
λ^X = coordinate of the particular value x of variable X, determines the contribution of the independent variable X
e = random error
Lower case letters represent the specific values of the corresponding upper case variables. E.g., $j = 1, 2, 3 \ldots$ labels particular judges.
$b = 1$ if $j = a$ and 0 otherwise.

The results show that three of the five factors have a significant effect on report quality: the author of the report, with $F = 3.42$, $p < .01$; the workshop, with $F = 12.15$, $p = .00$; and the scenario, with $F = 2.54$, $p < .01$. There is no significant difference in judgment of reports by the authors and by other analysts. The difference among different judges is not significant in this case. Details of the method can be found in Sun et al. (2004).

We conclude that there is a real difference in the quality of the reports across workshops. In fact, the effect *workshop variable* has the most significant effect. The cross-evaluation method provides a rigorous way to isolate system effects on the end product of the users' interaction with the system. It can elicit statistically significant differences with as few as four analysts and five tasks, with a total of no more than 20 items evaluated.

We would like to be able to assert that the significant change between workshops can be attributed to the change in the system. Unfortunately, two other factors are confounded with the changes. Specifically, at the first workshop, we did not tell the analysts about the cross-evaluation task, and so they may have been somewhat less diligent in preparing their reports. In addition, at the second workshop, analysts had considerably more experience using the system than at the first workshop. Either of these confounding changes could have contributed some or all of the measured change. Intuitively, it appears that all these factors tend to work in the same direction.

Analysts Perception of the System

The questionnaires provide information about the participants' perception of each version of the system and of system changes.

Workshop I. The normalized results of the post-scenario questionnaires for Workshop I (4 analysts × 2 reports each) are shown in Table 2. Each question was scored on a scale from 1 to 5. The "combined frequency of scores" indicates the number of times each of the five possible answers was selected in one of the post-scenario evaluations by one of the analysts. For example, on Question 1, a rating of 5 was selected five times and a rating of 4 was selected two times. A rating of 1 was selected zero times. The ratings have been normalized, so that higher scores (on the right) represent favorable evaluations, and lower scores (on the left) represent unfavorable ones.

The scores in Table 2 are almost uniformly at the midpoint of the 5-point scale, or above. This indicates that overall, analysts were quite comfortable using the system and satisfied with the quality of the answers the system provided. We interpret these midpoint and above ratings as indications that HITIQA is useful and usable.

The scores for the exit questionnaire completed at the first workshop are also at the higher end of the rating scale, as shown in Table 3; most scores were 3, 4, or 5.

Based on these data, we drew the following tentative conclusions from Workshop I:

- **Realism of the tasks.** The effort to simulate analysts' standard working conditions with scenarios like those they are faced with at work and with a database whose subject matter was fundamentally similar to those they use regularly succeeded.
- **Usability of HITIQA.** Analysts liked the overall design of the system and the general direction in which the system is

DOI: 10.1002/asi

TABLE 2. Normalized results of post-scenario questionnaire for Workshop I.

	Frequency of Combined Scenario Scores					Mean
Question	Score 1	Score 2	Score 3	Score 4	Score 5	Score
1			1	2	5	4.50
2		2	5	1		2.88
3			3	3	2	3.88
4			3	3	2	3.88
5		1	2	4	1	3.63
6		2	2	3	1	3.38
7		1	2	3	2	3.75
8			3	3	1	3.71
9			1	5	2	4.13
10			6	1		3.14
11			1	5	1	4.00
12	1	1	2	3	1	3.25
13a			1	5	2	4.13
13b	1		4	1	2	3.38
13c			2	3	3	4.13
13d			4	3	1	3.63
Total	2	7	42	48	26	3.71

Note. The items on the post-scenario questionnaire were as follows:

1. How realistic was the scenario? In other words, did it resemble tasks you could imagine performing at work?
2. How did the scenario compare in difficulty to tasks that you normally perform at work?
3. How confident were you of your ability to use HITIQA to accomplish the assigned task?
4. Given that you were performing this task outside of your standard work environment, without many of your standard resources, were you comfortable with the process of preparing your report?
5. Given that you were performing this task outside of your standard work environment, with access to a restricted set of documents, were you satisfied with the quality of the report/answers that you were able to find for this scenario?
6. In general, did the display of answers through the Answer Panel help you to navigate the answers in order to see what information was available?
7. In general, did the answers that the system provided make sense in relation to the questions that you asked?
8. In general, was it hard to formulate questions about this scenario that resulted in useful responses from the system?
9. In general, were the answers that the system provided helpful in meeting the goals set forth in the scenario?
10. In general, did the visual interface usefully represent the content of the answers that the system had found for you?
11. For this scenario, did the visual interface help you to find more precise answers than you would have found without it?
12. How would you assess the length of time that it took to perform this task?
13. If you had to perform a task like the one described in the scenario at work, do you think that having access to the HITIQA system would help …
    (a.) Improve your final report?
    (b.) Answer specific questions that you currently have trouble answering?
    (c.) Increase the speed with which you find information?
    (d.) Find information that you have trouble locating?

TABLE 3. Results of exit questionnaire for Workshop I.

	Frequency of Analyst's Scores of Overall Workshop I					Mean
Question	Score 1	Score 2	Score 3	Score 4	Score 5	Score
1		1		3		3.50
2		1	1	1	1	3.50
3				4		4.00
4			1	1	2	4.25
5				2	2	4.50
6	1			3		3.25
7			2	2		3.50
8			3	1		3.25
9			2	2		3.50
10				4		4.00
11				3	1	4.25
12				2	2	4.50
13		1		2		3.33
14				2	1	4.33
15		1	3			2.75
16				3	1	4.25
17	1		1	1		2.67
Total	2	4	13	36	10	3.73

Note. The items on the exit questionnaire were as follows:

1. I feel that I have become pretty proficient at using the HITIQA system.
2. The training on the first day gave me the skills needed to use the system successfully.
3. The training materials are hard to understand.
4. The training materials contain most of the information I needed to learn to use HITIQA.
5. My skill at using HITIQA improved over the course of the workshop.
6. HITIQA is hard to use.
7. I couldn't find enough documents with relevant information.
8. In general, I like the HITIQA interactive dialogue.
9. In general, I like the HITIQA visual interface.
10. In general, I like using HITIQA.
11. HITIQA slows down my process of finding information.
12. HITIQA helps me find important information.
13. Having HITIQA at work would help me find information that I can't currently find.
14. Having HITIQA at work would help me find information faster than I can currently find it.
15. HITIQA is not ready yet to be used in my regular work environment.
16. HITIQA would be a useful addition to the tools that I already have at work.
17. HITIQA would let me stop using some of the tools that I currently use at work.

headed. Analysts liked the interactive dialogue and visual display, but these systems need improvement.

- **Usefulness of HITIQA.** The analysts reported that HITIQA was already quite useful, to the point that the analysts can readily conceive of using it at work. But it is not yet ready to replace their current tools.

Workshop II. The results of the post-scenario questionnaires for Workshop II are displayed in Table 4; the questions are listed in Table 2. The left-hand section of Table 4 displays the distribution of scores given by the three analysts for each question. While the distribution of scores spanned a range of 1 to 5 (5 being best), analysts most frequently used the score 4 to respond to these questions. There were only three questions to which an analyst responded using a score of 1; all these responses came from a single individual. The three

DOI: 10.1002/asi

TABLE 4. Results of post-scenario questionnaires for Workshop II.

Frequency of Combined Scenario Scores in Workshop II						Mean
Question	Score 1	Score 2	Score 3	Score 4	Score 5	Score
1			3	3	3	4.00
2			6	3		3.33
3		1	2	4	2	3.78
4		1	1	6	1	3.78
5		2		5	2	3.78
6		2	3	3	1	3.33
7		3	1	3	2	3.44
8	1	2	1	2	3	3.44
9		3	1	4	1	3.33
10		3	3	2	1	3.11
11		2	3	2	2	3.44
12	1	2		4	2	3.44
13a			6	1	2	3.56
13b		1	3	4	1	3.56
13c	1	1	1	4	2	3.56
13d		2	4	1	2	3.33
Total	3	25	38	51	27	3.51

TABLE 5. Exit questionnaire results for Workshop II.

Frequency of Analyst's Scores of Overall Workshop II						Mean
Question	Score 1	Score 2	Score 3	Score 4	Score 5	Score
1				3		4.00
2				3		4.00
3				2	1	4.33
4				2	1	4.33
5				2	1	4.33
6				2	1	4.33
7		1	1	1		3.00
8			2	1		3.33
9				3		4.00
10				3		4.00
11		1		1	1	3.67
12			1	2		3.67
13			2	1		3.33
14		1		2		3.33
15		1	1	1		3.00
16				3		4.00
17		1		2		3.33
Total	0	5	7	34	5	3.76

questions asked about the time taken to complete work on the scenario, the speed with which information was found (Questions 12 and 13c) and the level of difficulty with formulating questions that resulted in useful responses from the system. The mean results for these questions reflect the impact of a single user, given the small sample size.

As with Workshop I, the primary objective of the Exit Evaluation was to assess the overall usability of HITIQA and to obtain feedback for future refinement. The normalized results of the exit questionnaire are shown in Table 5; the questions are listed in Table 3. As above, higher scores are better.

The left-hand section of Table 5 displays the distribution of scores given by the three analysts for each question. Most analysts used a score of 4 to respond to the questions of the exit questionnaire. Questions with lower mean scores indicate items that deserve future attention, for example, the mean score in response to Question 8, "In general, I like the HITIQA interactive dialogue," was 3.33, which suggests that improvements could be made to HITIQA's dialogue feature.

Based on the data in Table 5, we conclude that analysts continued to be satisfied with their use of the HITIQA system.

Comparison of Workshops I and II. One of our objectives is to develop techniques to track the effect of changes in the system and to compare complex information-access systems. With samples of size 4 and then 3, there are too few data points to justify analyzing the results statistically. (Recall that in cross-evaluation each question for each report was answered by 4 analysts serving as judges, increasing the power of the data.) To explore the change in analysts' evaluation of HITIQA from Workshop I to Workshop II, we use scatter plots to visually compare the results of the exit and post-scenario questionnaires.

Figure 5 represents the changes in responses to the 17 questions on the exit questionnaire. There is one scatter plot for each analyst who attended both workshops. Each point indicates the analyst's answers to one question in the questionnaire with the response for Workshop I as the x (horizontal) value and the response for Workshop II as the y (vertical) value. The higher the score in the plot, the "better" the system with regard to that question. A point on the diagonal indicates that the analyst gave the same response for both workshops on the question. Points above the diagonal represent the questions to which the analyst gave higher scores in the second workshop. Points below the diagonal represent the questions to which the analyst gave lower scores in the second workshop. If most points assigned by a single analyst are above the diagonal, it indicates a generally favorable response to the HITIQA system during the second workshop. If most points are below the diagonal, the analyst had a less favorable response at the second workshop.

We note that the changes in responses of the three analysts are quite different. User 3, who had rated the system very high in the first workshop, gave generally lower scores in the second. The scores of User 2 showed a general increase in approval, while the scores of User 1 generally remain close to the diagonal, which indicates unchanged feelings about the system.

The same principle can be applied to the more complex data produced after each scenario session. In Figure 6, the horizontal coordinate of a point indicates the average of the scores an analyst gave to a question for the first workshop, and the length of the horizontal bar represents the standard deviation of that set of scores for the first workshop. The vertical coordinate of a point indicates the average for the second workshop and the length of the vertical bar represents the standard deviation for

DOI: 10.1002/asi

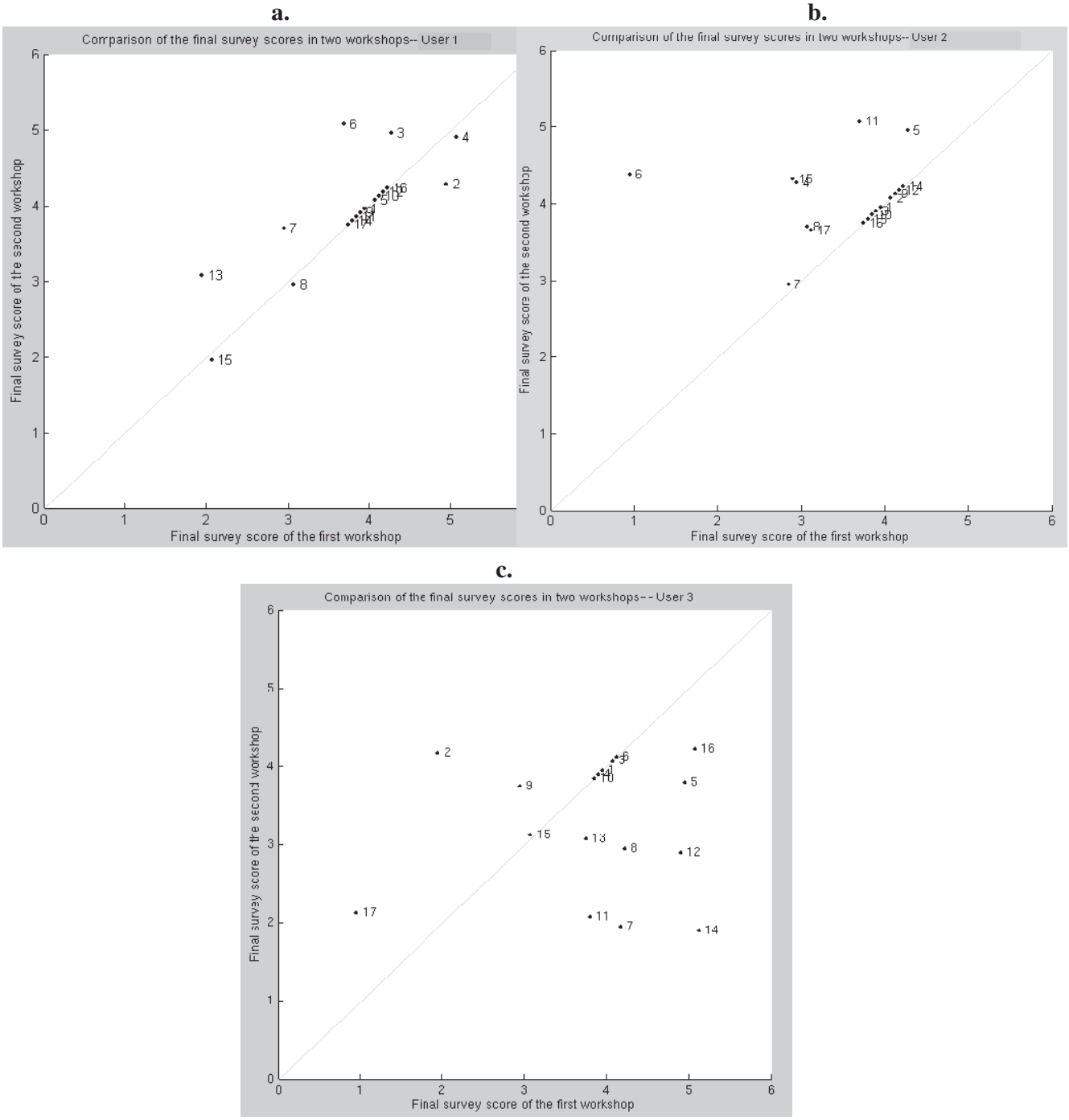


FIG. 5. Scatterplot comparing exit questionnaires for Workshops I and II.

the second workshop. In some cases the vertical or horizontal bar is absent. This simply means that the corresponding responses were identical for all sessions at that workshop, so that the standard deviation is zero.

In the three charts in Figure 6, we see the same general pattern as above. User 1 had relatively stable judgments, which means that the point indicating the answers given to a particular question stays close to the diagonal. One answer moves up to the upper sector. Two answers moved down into the lower sector. This was the answer to Question 2, "How did the scenario compare in difficulty to tasks that you normally perform at work?" which fell from a 5 (*more difficult*) to a 4.

User 2 showed 7 responses that moved up, and none that moved down. Clearly the second experience made a more positive impression in this case. We cannot say whether this is due to the changes in the system, to the confounding factors mentioned above, or, most pessimistically, to a change in analyst's expectations.

User 3 was less pleased with the system at the second session, with 8 responses that moved lower. The observer's notes

DOI: 10.1002/asi

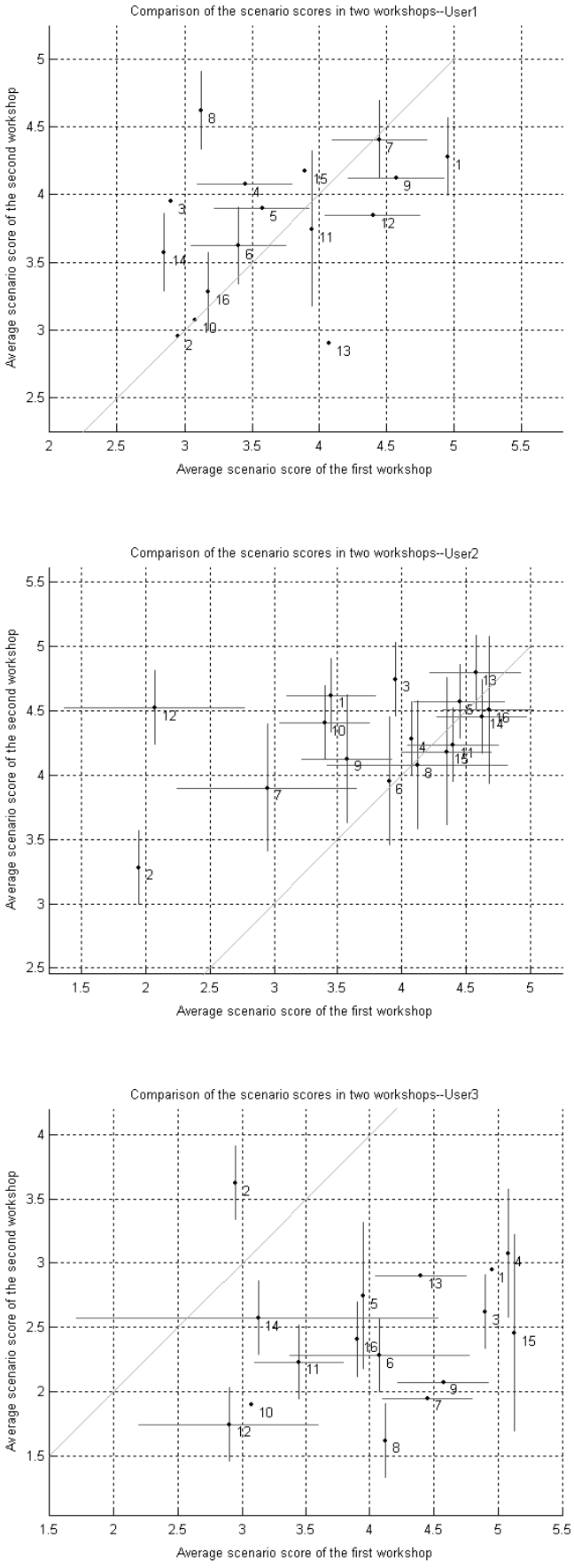


FIG. 6. Comparison of post-scenario questionnaires for Workshops I and II.

indicated that there was a problem in this individual's use of the system. These problems led to some frustrating interaction with the system and were compounded by the analyst's concern to produce high-quality reports for cross-evaluation by colleagues. This underscores the fact that confounding events during a complex experiment may substantially complicate any interpretation of the results, especially with a small number of users.

Use of scatter plots facilitates the comparison of individual user-response patterns across workshops. Even with the small number of participants in this evaluation, it is still possible to identify instructive patterns associated with participants' changing views of the system.

Discussion: Seizing the Evaluation Dilemma by the Horns

A fundamental challenge in the study of complex human behavior is the design of reliable and valid methods for use in environments where researchers lack control over all of the variables. Evaluation of interactive QA systems represents just such a challenge due to the natural variability of human beings, the difficulty of assessing the accuracy of answers to analytical questions, and the complexity of information-access systems that perform multifaceted tasks. As a pragmatic response to this challenge, the RUTS paradigm is designed to maximize what we can learn from involving real tasks, real users, and real systems, by (a) focusing the evaluation on the end product of the user-system interaction, and (b) use of the cross-evaluation method to reliably compare end products produced using different versions of the system.

In terms of the pilot evaluation of the HITIQA system itself, we conclude that the system was generally useful and usable, and was improved by the changes made to it. Participants liked the overall design of the system; they also pointed to specific areas that could use improvement. These results were particularly useful to the system developers because the system's advanced technology had not previously been assessed in a realistic way. With regard to our goal of assigning a real task to the participants, the confirmation from the questionnaire results that the scenarios resembled tasks that the analysts perform at work and were slightly above average in difficulty as compared to their normal tasks encourages confidence in the results of the evaluation.

We note that in the cross-evaluation, reports prepared at the second workshop were scored higher by all analysts. As noted above, this effect can be interpreted as being due to an increase in the effectiveness of the system. However, since the changes in the system are necessarily confounded with the other differences between the workshops, it is possible that they are, in part or entirely, due to the increased focus generated by awareness of cross-evaluation. They may, in addition, be due to increased experience. The truth, quite probably, is some combination of all three factors. Therefore, we must be cautious in interpreting these results. We anticipate that as more evaluations are conducted with the RUTS paradigm, our ability to interpret the findings and draw broader conclusions will improve.

DOI: 10.1002/asi

Of more lasting significance than the specific results obtained from studying four analysts using two versions of a single system are the experience and knowledge gained from conducting this assessment with real users, real tasks, and real systems. The results of the pilot experiment demonstrate that a) the RUTS method offers a practical way to compare interactive systems in a life-like setting, and b) the data collected offers rich possibilities for analysis that can be used both to assess a particular version of a system and to guide future development.

The assessment of the reports provides an overall evaluation of the end product of user-system interaction; the questionnaires supply complementary information about the realism of the task, the analysts' perception of specific components of the system (e.g., the interactive dialogue and the visual display), their comfort using the system, the areas where the system needs improvement, and the readiness of the software tool to be used in the participants' normal work environment. The RUTS method necessarily incurs the expense and inconvenience of involving real users, but the cross-evaluation approach minimizes the number of participants required to achieve meaningful results.

The assessment of the impact of the system (in this case, the workshop) on the product reveals statistically significant effects with surprisingly small samples. For comparison of two instances, the scatter plots developed here are a very effective tool, conveying a great deal of information about overall trends and about specific aspects of a scenario or system.

Once prepared, the task and corpus are fully reusable; this facilitates cross-system comparison. In principle, the paradigm can be used to compare different versions of a single system or quite different systems. In this pilot assessment, two relatively similar versions of the same system were compared; we anticipate that when the systems being compared are more different, the results of the cross-evaluation will be more striking.

The RUTS approach can be readily adopted for use with a variety of end products. The tasks and corpora can be designed as appropriate for the users, task, and system under evaluation. The most important element of the paradigm is the identification of a suitable task that requires a reasonable level of effort from participants and can be fairly assessed in a relatively short period of time. Suitably qualified experts are needed to assess the final product. In grading the reports, experts act like teachers in that they are expected to recognize excellent work when they see it. The use of multiple scenarios, multiple reports and multiple judges offsets the absence of an ideal answer against which system products can be judged. The use of the participants, in our case intelligence analysts, to perform the cross-evaluation is both economical and appropriate in situations where the participants have suitable training and experience. The obvious concern that participants may have a bias in favor of their own work has been managed by introducing a self-bias term into the analysis of variance, under the General Linear Model.

Holding two 3-day workshops to which the participants had to travel to attend and using a participatory process that involved multiple researchers made the evaluation that we describe here relatively costly. However, the assessment can be conducted much more economically online, which also makes it easier to involve more participants and avoid direct contact between them and the developers of the system. In fact, after the 2003 workshops, we successfully conducted additional, online evaluations of HITIQA with a different set of participants, using the scenarios, cross-evaluation, and questionnaires exactly as developed for these workshops. But conducting the workshops in person over a period of several days allowed us to learn a great deal from the analysts, supporting and extending what we learned from the questionnaires and the cross-evaluation. The additional experimenter effort was justified in our case because we wanted to observe the participants' reaction to the evaluation method itself as well as to the system being evaluated.

We also note that our ability to broadly interpret the results of the assessment of HITIQA are restricted by the lack of other studies using the same approach. We hope that use of the RUTS paradigm to compare interactive information-access systems will lead to the refinement of this method and provide a broad range of data that will help to interpret the results of evaluations using different kinds of users, tasks, and systems.

In the future, we plan to test the RUTS paradigm on different kinds of users, systems, and tasks. Our research will identify patterns that represent users' perceptions of different systems and will further our understanding of factors that affect changes in users' judgments; it will also contribute to the development of more useful and useable information-access systems.

Conclusion

The RUTS paradigm is an effective and informative approach to evaluation of information-access systems in which a) users are central to the process, b) creation of a single prototypically correct response is impossible, and c) the ability to adapt to the evolving needs of the user is one of the properties of the system being evaluated. The use of real users, real tasks, and real systems maximizes what can be learned in settings where researchers cannot completely control all variables. The cross-evaluation tool makes meaningful measurements of the impact of system modification on system products using a surprisingly small number of participants. The questionnaires measure the perception of the process of the people who actually perform the task in a real work setting. We recommend the RUTS evaluation model to all who would conduct a user-centered evaluation of information-access systems under realistic circumstances.

Acknowledgments

This article is based on work supported in part by the Advanced Research and Development Activity's (ARDA's)

DOI: 10.1002/asi

Advanced Question-Answering for Intelligence (AQUAINT) Program under a contact to SUNY Albany. Special thanks to Mike Blair, John Rogers, J. Steindl, and USNR analysts for participation in the workshops. Thanks also go to Google for extending their license, to Ralph Weischedel for the use of IdentiFinder, to Chuck Messenger and Peter LaMonica for assistance in development of the analytical scenarios, and to Bruce Croft for the use of the INQUERY system.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. New York: ACM Press.

Beaulieu, M., Robertson, S., & Rasmussen, E. (1996). Evaluating interactive systems in TREC. Journal of the American Society for Information Science, 47(1), 85–94.

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

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. Journal of Documentation, 53(3), 225–250.

Breck, E.J., Burger, J.D., Ferro, L., Hirschman, L., House, D., Light, M., & Mani, I. (2000). How to evaluate your question answering system every day … and still get real work done. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC). Retrieved March 25, 2007, from http://arxiv.org/PS_cache/cs/pdf/0004/0004008.pdf

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for non-native English speakers. In Proceedings of the ACL Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing. Retrieved March 25, 2007, from http://www.ets.org/Media/Research/pdf/erater_acl99rev.pdf.

Cleverdon, C.W. (1997). The Cranfield tests on index language devices. In K. Sparck Jones & P. Willet (Eds.), Readings in information retrieval (pp. 47–59). San Francisco, CA: Morgan Kaufmann. (Reprinted from Aslib Proceedings, 19, 173–192, 1967.

Cleverdon, C.W. (1991). The significance of the Cranfield tests on index languages. In A. Bookstein, Y. Chiaramella, G. Salton, & V.V. Raghavan (Eds.), Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 3–12). New York: ACM.

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

Dahlback, N., Jonsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies: Why and how. In W.D. Gray, W.E. Hefley, & D. Murray (Eds.), Proceedings of the First International Conference on Intelligent User Interfaces (IUI '93) (pp. 193–200). New York: ACM.

Harman, D. (1992). Overview of the First Text REtrieval Conference (TREC-1), NIST. Retrieved March 25, 2007, from http://trec.nist.gov/pubs/trec1/papers/01.txt

Harter, S.P. (1996). Variations in relevance assessments and the measure of retrieval effectiveness. Journal of the American Society of Information Science and Technology, 47(1), 37–51.

Harter, S.P., & Hert, C.A. (1997). Evaluation of information retrieval systems: Approaches, issues and methods. Annual Review of Information Science and Technology, 32, 3–94.

Hersh, W., & Over, P. (2000). TREC-8 Interactive Report. NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8), pp. 57–64. Retrieved March 25, 2007, from http://trec.nist.gov/pubs/trec8/t8_proceedings.html

Hersh, W., & Over, P. (2001). TREC-9 Interactive Report. NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9), pp. 41–50. Retrieved March 25, 2007, from http://trec.nist.gov/pubs/trec9/t9_proceedings.html

Hersh, W., & Over, P. (2002). TREC 2001 Interactive Track Report. NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001), pp. 38–41. Retrieved March 25, 2007, from http://trec.nist.gov/pubs/trec10/t10_proceedings.html

Hersh, W., Turpin, A., Price, S., Kraemer, D., Chan, B., Sacherek, L., & Olson, D. (2000). Do batch and user evaluations give the same results? In N.J. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 17–24). New York: ACM.

Hirschman, L., Breck, E., & Burger, J.D. (1999). Deep Read: A reading comprehension system. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 325–332). Madison, WI: ACL.

Jansen, B., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing & Management, 36(2), 207–227.

Kantor, P., Boros, E., Melamed, B., Neu, D., Menkov, V., Shi, Q., & Kim, M.-H. (1999a). Ant World. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (p. 323). New York: ACM.

Kantor, P., Boros, E., Melamed, B., & Menkov, V. (1999b). The information quest: A dynamic model of user's information needs. In L. Woods (Ed.), Proceedings of the 62nd Annual Meeting of the American Society for Information Science: Vol. 36 (pp. 536–545). Silver Spring, MD: ASIS.

Kelly, D., Wacholder, N., Rittman, R., Sun, Y., Kantor, P., Small, S., et al. (2007). Using interview data to identify evaluation criteria for interactive, analytical question-answering systems. Journal of the American Society for Information Science and Technology, 58(7), 1032–1043.

Krizan, L. (1999). Intelligence essentials for everyone (Occasional Paper No. 6). Joint Military Intelligence College.

Maybury, M.T. (2002). Toward a question answering roadmap. MITRE Technical Papers. Retrieved December 12, 2004, from http://www.mitre.org/work/tech_papers/tech_papers_02/maybury_toward/maybury_toward_qa.pdf

Over, P. (2001). The TREC Interactive Track: An annotated bibliography. Information Processing & Management, 37(3), 161–176.

Saracevic, T. (1997). Relevance: A review of and a framework for the thinking on the notion in information science. In K. Sparck Jones & P. Willett (Eds.), Readings in information retrieval (pp. 143–165). San Francisco, CA: Morgan Kaufmann. (Reprinted from Journal of the American Society for Information Science, 26(6), 321–343, 1975.)

Saracevic, T., Kantor, P., Chamis, A.Y., & Trivison, D. (1988). A study of information seeking and retrieving I: Background and methodology. Journal of the American Society for Information Science, 39(3), 161–176.

Saracevic, T., & Kantor, P. (1988a). A study of information seeking and retrieving II: Users, questions, and effectiveness. Journal of the American Society for Information Science, 39(3), 177–196.

Saracevic, T., & Kantor, P. (1988b). A study of information seeking and retrieving III: Searchers, searches, and overlap. Journal of the American Society for Information Science, 39(3), 197–216.

Small, S., Strzalkowski, T., Liu, T., Ryan, S., Salkin, R., Shimizu, N., et al. (2004). HITIQA: Towards analytical question answering. In Proceedings of the 20th International Conference on Computational Linguistics, Article 1291. Morristown, NJ: ACL.

Small, S., Ting, L., Shimuzu, N., & Strzalkowski, T. (2003). HITIQA, An interactive question answering system: A preliminary report. In Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering: Machine Learning and Beyond (pp. 46–53). East Stroudsburg, PA: ACL.

Sparck Jones, K. (2001). Automatic language and information processing: Rethinking evaluation. Natural Language Engineering, 7(1), 29–46.

Sparck Jones, K. (2003). Is question answering a rational task? CoLogNET Area 6: Logic for Natural Language Processing. Retrieved April 5, 2007, from http://www-uilots.let.uu.nl/~ctl/workshops/CES03/On_line_Proceedings/Papers/ sparck.pdf

Spink, A. (1996). Multiple search session model of end-user behavior: An exploratory study. Journal of the American Society for Information Science and Technology, 47(8), 603–609.

DOI: 10.1002/asi

Strzalkowski, T., Small, S., Hardy, H., Yamrom, B., Liu, T., Kantor, P., et al. (2005). HITIQA: A question answering analytical tool. In Proceedings of the 2005 International Conference on Intelligence Analysis. May, 2005. Retrieved April 5, 2007, from https://analysis.mitre.org/proceedings/Final_Papers_Files/21_Camera_Ready_Paper.pdf

Sun, Y., & Kantor, P. (2006). Cross-evaluation: A new model for information system evaluation. Journal of the American Society for Information Science and Technology, 57(5), 614–628.

Sun, Y., Kantor, P., Strzalkowski, T., & Wacholder, N. (2004). Cross Evaluation: a pilot application of a new evaluation mechanism. In Proceedings of the American Society for Information Science and Technology, Vol. 41 (pp. 383–392). New Brunswick, NJ: ASIST.

Tague-Sutcliffe, J.M. (1996) Some perspectives on the evaluation of information retrieval systems. Journal of the American Society for Information Science, 47(1), 1–3.

Turpin, A.H., & Hersh, W. (2001). Why batch and user evaluations do not give the same results. In W.B. Croft, D.J. Harper, D.H. Kraft, & J. Zobel (Eds.), Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 225–231). New York: ACM.

Vakkari, P. (2003). Task-based information searching. Annual Review of Information Science and Technology, 31, 33–78.

Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. Journal of Documentation, 56, 540–562.

Voorhees, E. (2001). The philosophy of information retrieval evaluation: Revised papers from the second workshop of the cross-language evaluation forum on evaluation of cross-language information retrieval systems, pp. 355–370. London: Springer Verlag. Retrieved April 5, 2007, from http://www-nlpir.nist.gov/works/papers/eval_philosophy.ps

Voorhees, E. (2002). The evaluation of question answering systems: Lessons learned from the TREC QA track. In Proceedings of the LREC 2002 Workshop on Question Answering: Strategy and Resources. Retrieved April 5, 2007, from http://www-nlpir.nist.gov/works/papers/lrec02.ps

Voorhees, E. (2003a). Evaluating the evaluation: A case study using the TREC 2002 Question Answering Task. In Proceedings of the HLT-NAACL 2003 (pp. 181–188). Morristown, NJ: ACL.

Voorhees, E. (2003b). Overview of TREC 2003. NIST Special Publication SP 500-255. Retrieved April 5, 2007, from http://trec.nist.gov/pubs/trec12/papers/OVERVIEW.12.pdf

Voorhees, E., & Tice, D. (2000a). Building a question answering test collection. In N.J. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 200–207). New York: ACM.

Voorhees, E., & Tice, D. (2000b). Implementing a question answering evaluation. In Proceedings of "Using Evaluation Within HLT Programs: Results and Trends," a workshop of the Second International LREC Conference. Retrieved April 5, 2007, from http://www.itl.nist.gov/iaui/894.02/works/papers.html

Voorhees, E., & Tice, D. (2000c). The TREC-8 Question Answering Track. In Proceedings of the Eighth Text Retrieval Conference (TREC-8). Retrieved April 5, 2007, from http://nist.gov/pubs/trec8/./papers…qa8.ps

Tague-Sutcliffe, J.M. (1996). Some perspectives on the evaluation of information retrieval systems. Journal of the American Society for Information Science, 47(1), 1–3.

Appendix A: Workshop Scenarios

Chemical Weapon: Sarin

The Department of Homeland Security has requested a complete report on the chemical weapon, sarin. This report is due in 5 hours. In your report, include a discussion of its potency and potential impact on a community, what countries and organizations have been involved in producing it, where these locations are, the production method and how it has developed, who possesses it now, who distributed it (if through trade, what was traded for it?), potential means of use, how it this be integrated into warheads, any known defenses against it, and who is at the greatest risk. Provide any other information that you see relevant.

The al-Qaida Terrorist Group

As an employee of the Central Intelligence Agency, your profession entails knowledge of the al-Qaida terrorist group. Your division chief has ordered a detailed report on the al-Qaida Terrorist Group, due in three weeks. Provide as much information as possible on this militant organization. Eventually, this report should present information regarding the most essential concerns, including who are the key figures involved with al-Qaida along with other organizations, countries, and members that are affiliated, any trades that al-Qaida has made with organizations or countries, what facilities they possess, where they receive their financial support, what capabilities they have (chemical and biological weapons program, other weapons, etc.) and how have they acquired them, what is their possible future activity, how their training program operates, who their new members are. Also, include any other relevant information to your report as you see fit.

South Africa's WMD Program

You have been given 30 days to develop a comprehensive report on the South African chemical, biological, and nuclear warfare program, for your division chief who is to present it to the Secretary of Defense. Your report should include several key elements of the South African WMD Program, including what people, organizations, and countries are involved, what chemicals have been purchased and/or used, where the chemicals have been purchased from and from whom, how their WMD program was financed, where these development locations are, any proposed activity (use, distribution, etc.), any money transactions that have been made between these suspects and other organizations, and any other contacts or travels that have been made by any of the primary figures involved. Supply any further information that can support your documentation.

Nuclear Arms Relationship: Russia and Iraq

The Department of Defense has demanded a report on how Russia has influenced the nuclear arms program in Iraq. The department needs the summary by close of business today. List the extent of the nuclear program in each country including funding, capabilities, quantity, etc. Your report should also include key figures in both the Russia and Iraq nuclear programs, any travels that these key figures have made to other countries in regards to a nuclear program, any weapons that have been used in the past by either country, any purchases or trades that have been made relevant to weapons of mass destruction (possibly oil trade, etc.), any ingredients and chemicals that have been used, any potential weapons that could be under development, other countries

DOI: 10.1002/asi

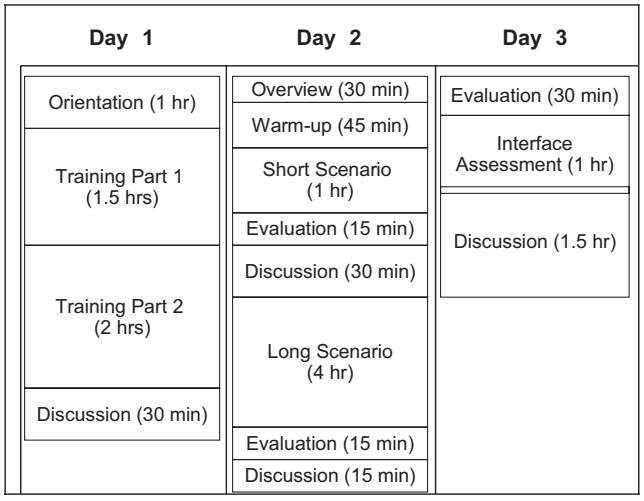
that are involved or have close ties to Russia or Iraq, possible locations of development sites, and possible companies or organizations that these countries work with for their nuclear arms program. Add any other information relating to the Russian and Iraqi Nuclear Arms Programs.

North Korea Nuclear Weapons Program

The Department of Defense has demanded a report concerning the North Korean Nuclear Arms Program. This report is needed within two days and should include any advances that have been made in their program, as well as their current stance on South Korea and the United States. Also include sources of weapons materials and technical assistance received. It should also contain information on any travels that have been made by North Korean government officials (where they travel to, who they visited, etc.), new relationships that North Korea has established, who these countries or organizations are, what capabilities they possess, and any other information that may apply.

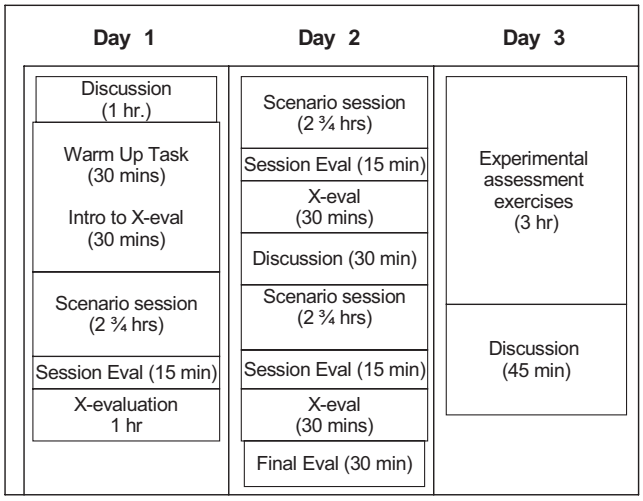
Appendix B

Workshop I Schedule



Appendix C

Workshop II Schedule



DOI: 10.1002/asi