Cross-Evaluation: A New Model for Information System Evaluation

Ying Sun and Paul B. Kantor
4 Huntington Street, Rutgers University, New Brunswick, NJ 08903. E-mail: {ysun, kantor}@scils.rutgers.edu

In this article, we introduce a new information system evaluation method and report on its application to a collaborative information seeking system, AntWorld. The key innovation of the new method is to use precisely the same group of users who work with the system as judges, a system we call *Cross-Evaluation*. In the new method, we also propose to assess the system at the level of task completion. The obvious potential limitation of this method is that individuals may be inclined to think more highly of the materials that they themselves have found and are almost certain to think more highly of their own work product than they do of the products built by others. The keys to neutralizing this problem are careful design and a corresponding analytical model based on analysis of variance. We model the several measures of task completion with a linear model of five effects, describing the users who interact with the system, the system used to finish the task, the task itself, the behavior of individuals as judges, and the self-judgment bias. Our analytical method successfully isolates the effect of each variable. This approach provides a successful model to make concrete the "three-realities" paradigm, which calls for "real tasks," "real users," and "real systems."

Introduction

Scientific progress in the development of information systems requires measurement and evaluation of those systems. Evaluation plays a role not only in assessing the final result of development, but also in guiding development. Historically, and for reasons of efficiency, in the field of information retrieval we have sought to separate evaluation from the details of any specific user's task. The most prominent example of this separation is the development of the notion of relevance as something that can be assessed by judges distinct from those who need the information. In this sense, the complex notion of relevance is reduced to the relatively objective notion of being "related to the stated topic." This approach was pioneered by Cyril Cleverdon in the Cranfield Experiments (Cleverdon, 1960). It has evolved into the successful and influential Text REtrieval Conference (TREC) evaluation framework. There is no question that the existence of TREC, and the applicability of this method, has contributed substantially to the growth of information systems.

At the same time, there have been efforts to extend this fundamental paradigm. However, in most cases (an anonymous reviewer has reminded us that this was not a constraint of the early Cranfield work) a document is considered to be objectively either relevant or not relevant for a task. This cannot be maintained for more complex situations, such as question-answering and aspectual recall. It is possible to ring changes on the underlying concept without stepping away from the guiding paradigm. For example, in the short-lived TREC confusion track, Kantor and Voorhees (2000) looked at the cumulated utility, and utility measures have been adopted for the adaptive filtering task (Robertson Soboroff, 2002). Fundamentally, however, these are mathematical elaborations of the underlying paradigm: Given the problem, a particular document has an intrinsic score of either 1 or 0.

As the tasks become more complex, we would like to move closer to real judgment by real users. An example of a more complex problem is the one addressed in the TREC Interactive Track. This track has struggled for years with the issue of defining a measure that somehow goes beyond elaborations of binary relevance (Over, 2001). The problem is, when a user uses a system to do a task, how can we separate the user from the system? Conceptually, of course, the answer is clear: We need experimental designs in which the variables representing the user and the system are "crossed." In other words, we must have several users each use several systems. Historically, in the TREC setting, this was not done because systems "lived" at different institutions and it was not convenient for the same individuals to use multiple systems. With the growth of Web-based systems, the problem is expected to resolve itself. It is therefore time to consider seriously the design and analysis of experiments involving multiple users and multiple systems.

Received January 13, 2005; revised February 21, 2005; accepted March 11, 2005

 ● Published online 1 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20324

At the same time, we finally have an opportunity to move away from the artificially imposed dependence on judges who are *not* themselves the users of the systems. The key innovation in the proposed method of Cross-Evaluation is to use precisely the same individuals who work with the system to find answers as the judges of those answers once found. There are obvious justifications for this method. These users will have become familiar to some degree with the problem and its aspects, and they will have gained some knowledge of the corpus or collection through their own work with a particular system. They are, if they act promptly, immersed in the problem and able to provide a more thoughtful and effective evaluation than could an uninvolved judge. This is particularly true if the standard of judgment that has been chosen is not the quality or relevance, often limited to a binary scale, of the entities found, but is rather the quality of an overall work product.

The obvious potential limitation of this method is that individuals may be inclined to think more highly of the materials that they themselves have found and are almost certain to think more highly of their own work product than they do of the products built by others.

We are thus faced with a concept, Cross-Evaluation, that seems to have significant advantages and at the same time a fatal flaw. The key to the method proposed here is to neutralize the flaw through carefully designed statistical analysis, leaving only the benefits.

We note there are three levels of granularity at which the results of using an information system may be judged. One is at the levels of the individual item retrieved, which is the basis for the conventional TREC metrics. The second level of granularity is task completion. The third level assesses the impact of the overall, completed task on the motivating goal of the individual or organization. We focus here on the middle level, task completion. To study task completion, we need three components: (1) a stable of systems, (2) a basket of sufficiently similar tasks, and (3) a collection of suitable users. Let us think of these as the materials of our study. We also need a clear analytical model, which states what is to be measured and states how meaning is to be extracted from the measurements. All of these are detailed in this article.

This article is organized as follows. Related Work is an overview of some related works and the significance of the problem. Cross-Evaluation describes the general model of the Cross-Evaluation method. The Analytical Method sets forth the analytical model that we use to control for all of the factors, including self-preference. A Pilot Study is the description of a pilot experiment, which we have done to validate this approach. It includes the description of the stable of systems and the instruments for gathering data, which we have developed specifically for this project, in a Web-based form. Results of the Pilot Experiment describes the analysis processes and results. Conceptually A Pilot Study and Results of the Pilot Experiment may be thought of as a separate article describing the application of the method proposed here. Conclusions and Discussion contains the summary of the findings and our conclusions.

Related Work

Evaluation of Interactive Information Retrieval Systems

The Cranfield paradigm of evaluation has long been criticized for not taking issues such as the interactive and cognitive aspects of information seeking into consideration. TREC, as the most successful representative of the paradigm, has made an effort to stretch the traditional TREC test collection, and the track concepts, into the field of interactive information retrieval (IR) from the very beginning of TREC experience (Over, 2001). However, the interactive track maintains very little comparability across systems, unlike other TREC tracks.

The TREC interactive track, especially in TREC-5 and TREC-6, has made some good attempts to make comparison across systems, while taking the impact of system, topic, searcher, and their interactions into consideration. The idea of the Matrix Experiment was developed before TREC-5 and largely tested during TREC-6. In this approach each participating site compares its own experimental system to a single control or baseline system. A basic Latin square design allowed reasonable estimation of the difference between the experimental and the control system at each site, while the main effects of topic and searcher are to some degree under control (Lagergren & Over, 1998). Unfortunately, the attempt at cross-site (cross-systems) evaluation was inconclusive, and it was dropped after TREC-7. One guiding principle of the Cranfield/TREC paradigm is that any comparison between systems has to be made on the basis of a single set of queries and judgments. For systems without human interaction, the expert judgments are the only set of judgments and every system is tested against them. The comparison based on the judgments directly reflects the differences among systems. In the interactive track, the expert judgments are used as a standard against which each participating site's output of human-system interaction (in other words, the judgments of the site's user group) is tested. Because each site employs its own group of users, there are multiple sets of judgments involved in the evaluation process: user groups' judgments and expert judgments. The rule of single set of judgments is violated. Thus it is hard to draw conclusions across systems. The differences among system users from different sites naturally lead to disagreement on the relevance judgments. These disagreements not only block direct cross-site comparison but also raise questions about the meaning of any evaluation based on the comparison among multiple sets of judgments.

There has been another direction in the information system evaluation work. It has been primarily seen in research-oriented papers in which the method of evaluation is itself the subject of the research. In the pioneer study following this path, conducted by Saracevic and Kantor (Saracevic et al., 1988a, 1988b, 1988c), the information seeking process was viewed as involving five types of entities: users, questions, intermediaries, search, and retrieved items. Their systematic study clearly demonstrated the five entities' complex relations with effectiveness measures including

DOI: 10.1002/asi

relevance-based precision and recall, and some utility measures. On one hand, that study shows the attractiveness of the underlying principle of this approach, an insistence on doing studies that involve "real users, real problems, and real systems." On the other hand, the results reveal that the approach in itself contains some potential for its own destruction.

The potential for destruction lies in two of the three realities: users and questions. Although "real systems" are something that we know how to deal with, particularly in the world of information retrieval research, in which the systems are usually studied by those who have built them, the other two components, users and problems, are open ended. The study of users can carry us into the study of cognitive science, the study of personal psychology, the study of educational theory, the study of socially constructed behavior, and so forth.

The study of "real problems" also poses enormous potential hazards. Problems, as can questions, can reflect the entire range of human knowledge or ignorance. The classification of problems or questions then becomes no simpler than the long-standing problem of classifying all of knowledge. These ontological questions have occupied librarians for generations, and anyone familiar with the available solutions knows that they are driven more by the pressures of practical reality than by any compelling logic. Thus, people who have taken up the issue of classifying problems find themselves looking at the distressingly large and elusive notions of "task" and "context." To make matters worse (or better, from the perspective of the aspiring academic), the notion of context easily works its way back to connect to all of complexities related to the notion of the "user."

The challenge is to extend the notion of real users and real problems through the idea of a Cross-Evaluation design to the realm of concrete measurement that will isolate the system effect while taking into consideration user and problem effects.

Measures for Information Retrieval Systems

There are many ways to classify the measures available for evaluating an information system. Here we try to organize the measures along the dimension of the measures' objective or subjective nature.

As shown in Figure 1, measures at one end (the left end) of the dimension are the most objective measures, such as the time spent on a task or a search, the number of queries entered, the number of documents saved. These measures can be easily obtained by an automatic analysis of system logs. The limitation of this group of measures is that they measure either the system performance alone or the *efficiency* (speed, cost, etc.) of the interaction. It is hard to assess the interactive system's *effectiveness* with these measures.

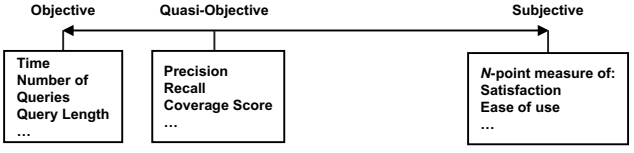

FIG. 1. Objective/subjective dimension of evaluation measures.

At the other end of the axis we place *N*-point subjective measures. Such measures are widely used in interactive system evaluation work. Normally, they are collected by questionnaires that use *N*-point Likert scales. The questions asked often isolate the users' perceptions of various aspects (utility or usability) of their experience with the systems. Such measures are important in helping to understand what the users expect in a "good" system and, and thus give suggestions for improving the current system (formative evaluation). But they are not satisfactory, on their own, in the situation of comparing two or more systems (summative evaluation).

In the middle of the axis are located what we call *quasi*-objective measures. Many popular measures are representatives of this type. They share three common characteristics. First, they measure systems by the direct outputs of systems or interactions. The measures of precision and recall, for example, are based on the sets of documents retrieved. Other areas, such as Question Answering (QA) System evaluation, have adopted similar measures. The TREC QA track is the representative of current QA system evaluation work. The TREC QA track uses a recall like measure based on the completeness and correctness of the answers. The "correctness" assessment is based on certain predefined criteria (TREC QA Track). The Document Understanding Conference (DUC), run also by the National Institute of Standards and Technology (NIST), compares each of the machine-generated summaries with an "ideal" model summary written by NIST assessors and computes a coverage score called *C* (DUC, 2002). The ideal model summary is decomposed into a list of units, and the machine-generated summaries are matched against it. In simple words, C is the percentage of the ideal model units that is included in the machine-generated summary.

The second common feature of all these measures is that they rely fundamentally on some subjective processes. The measures of precision and recall are based on relevance judgment, which is widely accepted to be subjective, dynamic, and multidimensional (Barry, 1994; Schamber, 1994). The growth of QA research has now expanded far beyond just factual (or "factoid," as it is sometimes called) QA (Burger et al., 2001). With research dimensions such as multiple question types and interactive dialog with the user to refine/guide the QA process, objective criteria for correctness are not easily extended to analytical QA.

In DUC, to prevent the evaluation variability introduced by human differences, researchers sought to replace human experts with some absolutely objective machine matching technique. But it is clear that the judgments of relevance and correctness of texts/passages/summaries inevitably involve the assessors' subjective behavior. So the measures based on them are not objective measures but, as we have noted, quasi-objective.

DOI: 10.1002/asi

These quasi-objective measures are "normalized" in two senses. The first is a kind of mathematical normalization. The original number based on subjective judgments (number of relevant texts retrieved) is normalized, to produce, for example, p_{100}, the precision of the first 100 retrieved documents, average precision, and so forth. In the second sense, the judgments themselves are normalized. This normalization is implemented by using expert judges and/or by defining objective and static criteria for judges to follow. The assumption is that the experts are more likely to give ideal judgments and/or that there are some objective and static criteria for the required quality (relevance, correctness, etc.) of the documents. The normalization carries this group of measures closer to the objective end of the dimension. To summarize, we call them quasi-objective because they originate in subjective judgments and are regularized to some degree by the "objectifying" normalization.

Objective and quasi-objective measures have worked well in comparing noninteractive information systems. However, as mentioned, there are problems in adapting them directly for evaluating highly interactive systems. *N*-point subjective measures satisfy the requirements for interactive system evaluation from a purely user-centered perspective. However, most currently asked questions only measure systems indirectly, through users' perception. But users' perceptions of a system are driven by some combination of both the usability aspects and the utility aspects of the system. In particular, a high user satisfaction score does not necessarily mean that the system is more effective. The user may be satisfied because the system is simple enough, or pleasant enough, to use. In addition, these measures often do not really isolate different aspects of the system. That is, factor analysis shows that the answers to several dozen questions may really depend on just two or three underlying factors. Generally, a comparison of systems based on such measures is not as persuasive as one based on measurement of the results produced by users of the system.

In this study, we evaluate interactive systems by the final products–analytical reports, which the users produce. Our measure combines the advantages of both quasi-objective and subjective measures: The evaluation is based on real products, not on abstract perceptions, and the measurement is made by real users in the form of *N*-point scales without any normalization.

We also tried evaluation at the document level granularity. However, two arguments make us believe that to measure a system's effect on complex analytical tasks, the report-based measure is more appropriate and more accurate. First, collecting information is not the final purpose of using the systems. Second, users save information for various purposes. The saved information is not necessarily useful to the task, but the fact that it was saved may represent a good work habit of the analyst.

Our proposed evaluation model focuses on realistic analyst tasks and provides a concrete measurement of overall performance, through the quality of generated analytical reports.

Cross-Evaluation

As noted, the two key features of the Cross-Evaluation method are (1) letting the users be the judges and (2) measuring systems at the level of task completion. The method has been applied to evaluate a collaborative Web information seeking system and several QA systems. One common feature of these systems is support of complex analytical tasks. Collecting information and preparing an analytical report on an assigned topic is the representative task of the target user groups of the systems. Therefore, in the general model we present here, the task for which a system is used is to collect information and complete a draft report. The effectiveness of the system is measured by the quality of the report. We note that the method can be used for systems supporting other types of tasks. As long as some measures of task completion are defined and in the form of numerical scores, the model will work.

From a participant's point of view, for each particular topic, the evaluation procedure involves two stages. In stage one, the participant uses the assigned system to collect information and prepare a draft report on the topic. In stage two, the participant works as an assessor of the report's quality. We developed a set of five report quality judgment criteria:

- Covers the important ground
- Avoids irrelevant materials
- Is well organized
- Reads clearly and easily
- Overall rating this report

Each criterion is scored on an *n*-point Likert scale, and each participant judges all of the reports prepared for a given task.

Each particular task forms an "evaluation unit." The components and their relationships for a complete evaluation unit are illustrated in Figure 2.

The Analytical Model

Carefully designed statistical analysis is the second key feature of the proposed method.

The measures obtained by Cross-Evaluation can be identified as having been (1) produced by a particular judge for (2) the work of a particular participant (author) on (3) a particular topic and (4) using a particular system. The concept of Cross-Evaluation itself contains a potential fatal flaw: self-judgment bias. We posited that an individual may recognize the products that he or she contributed and may give them higher scores during the evaluation stages. To neutralize this potential flaw, a self-judgment bias effect is included. We build an analytical model based on analysis of variance, which is applied to all of the five measurement scores, with these variables as factors. This leads to the general linear model:

$$V(j, a, t, s, b) = \lambda^0 + \lambda_j^J + \lambda_a^A + \lambda_t^T + \lambda_s^S + \lambda_b^B + e \quad \text{Analytical Model} \quad (1)$$

where V = measurement scores; J = judge variable; A = author variable; T = task variable; S = system variable;

DOI: 10.1002/asi

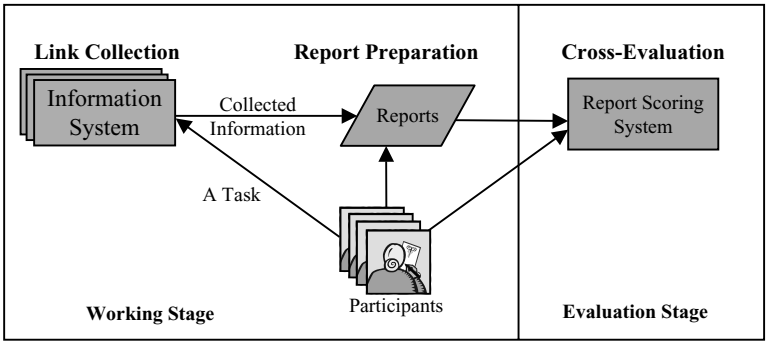

FIG. 2. Cross-Evaluation model.

B = self-judgment bias variable; λ_i^I = coordinate of the particular value i of variable I, determines the contribution of the independent variable I; and e = random error.

The lowercase letters represent the specific values of the corresponding uppercase variable. Thus $j = 1, 2, 3\ldots$ labels particular judges. In the Equation 1 analytical model, we do not include any interaction effects. Typical experimental designs concentrate on getting the best possible estimate of the main effects. Quite large experimental designs would be required to measure the possible interactions. For example, to study all the possible interaction effects in a study including three systems, three tasks and three user groups (as reported in the section A Pilot Study), a $3 \times 3 \times 3$ factorial design, three times the proposed Latin square design, is required.

A Pilot Study—AntWorld Evaluation

The Cross-Evaluation method was developed in the evaluation component of development of a complex collaborative information finding system for the World Wide Web: AntWorld. We report here the study as an illustration of the proposed method.

The Tasks

As noted, we evaluate the systems at the granularity level of task completion. We defined the task as the preparation of a preliminary or draft set of materials for an analytic report on one of three topics. The topics are shown in Table 1.

TABLE 1. Topics.

Task 1. Anthrax (AN). Locate information and integrate it into a briefing report covering these major aspects of anthrax: detection, prophylaxis, treatment, and use as a weapon.
Task 2. Terrorist Sites (TE). Locate information and integrate it into a briefing report covering these major aspects of terrorism on the Web: overtly terrorist sites; sites that covertly support or link to overt sites, under guise of charities; sites that seem to be endorsed by either of the other two kinds of sites.
Task 3. Nuclear Weapons (NW). Locate information and integrate it into a briefing report covering these major aspects of development of nuclear weapons by nongovernmental organizations: reports of loss of nuclear raw materials, reports on capabilities (needed or existing) for making weapons, issues of transporting nuclear weapons to the target locations.

It is a historical accident that our method was developed in the context of the study of collaborative systems, and for this reason the situation is somewhat more complicated than the theoretical framework (Figure 2) requires. Specifically, we were studying the impact of a collaborative tool (AntWorld), and therefore the task had to be done "in collaboration." It follows that the concept of "user" in this particular pilot study is replaced by the concept of the "team of users." (We note that another intriguing layer of analysis, which would result from systematic recombination of individual analysts among teams, is possible. The budget for this project did not permit that method, which would provide a useful topic for future study.)

The Systems

AntWorld systems with and without previous knowledge. There are, essentially, three systems in this pilot study. The first is the AntWorld system itself, whose study motivated the development of this method. AntWorld is a system that supports the work of subsequent analysts by providing access to work by previous analysts when that prior work is "sufficiently similar" to the work of the present analysts. The details have been described in a number of articles (Boros, Kantor, & Neu, 1999; Kantor, Boros, Melamed, & Meñkov, 1999; Kantor et al., 1999). It is clear that the AntWorld system may present itself in two ways. The first is the mature system, in which there are a substantial number of relevant previous traces to be followed. The other is the immature system, in which there are no such traces. Because the primary goal of our study is to assess the value of such traces to subsequent analysts, we arranged the experiment so that access to previous traces could be turned on or off as part of the experimental conditions.

The annotation tool. Finally, because the task is a collaborative one, we needed to build a simple and uniform collaboration tool, which would make the experience of working with AntWorld as similar as possible to a null experience of searching with conventional search engines and no Ant-like sharing. We called this new tool the *Annotation Tool* (AT), and it was developed from scratch by using the Java Server Page (JSP) language. Thus, the stable of systems

DOI: 10.1002/asi

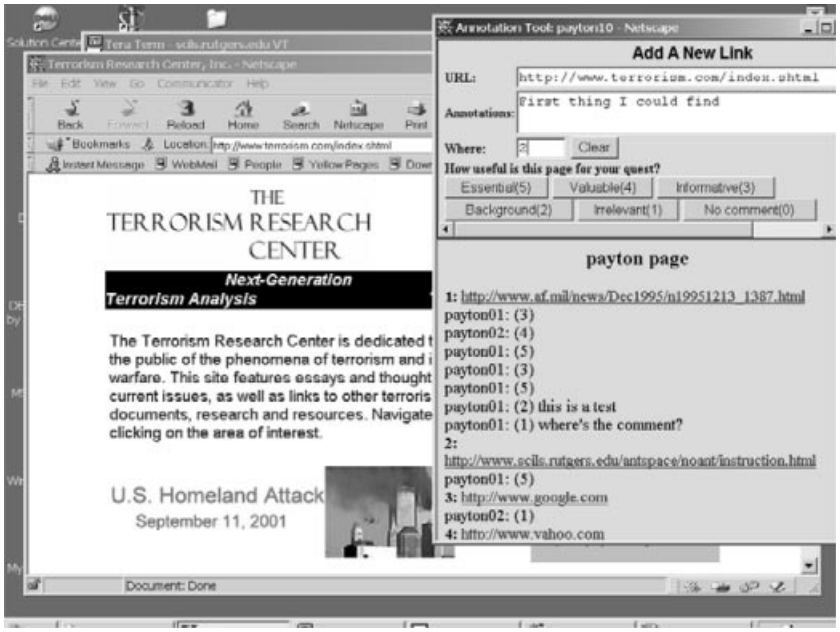

FIG. 3. Interface of the annotation tool.

consists of the bare AntWorld (AW), the populated AntWorld (AW2), and the Annotation Tool (AT).

The Annotation Tool is in essence simple groupware but specially designed for our evaluation work as a null system. It provides a method for users to collect useful information from the Web and to share the information with team members. When a user logs into the system, a Web page kept for the user's team and a regular browser window are shown (Figure 3). The team Web page contains the uniform resource locators (URLs) they have selected and simple annotation information. The tool is multithreaded. Team members can work from different places at the same time.

Team members can search and browse with the regular browser window. If a user finds a piece of useful information and wants to share it with his/her teammates, he or she can add new information (URLs of useful Web pages and brief annotations) on this page. A simple function to organize information is provided by letting the user specify where on the page the new information should appear.

The team page is kept as a Web page to make it convenient for users to look at the information collected by other team members. By clicking the URLs in the team page, they can load Web pages into the regular browser window. The tool also invites the user to make comments on the information currently displayed.

To add a Web page to this "Annotation Page," a user is required to make a relevance judgment according to a 6-point scale: Essential (5), Valuable (4), Informative (3), Background (2), Irrelevant (1), and No Comment (0).

The Evaluation Process

In this pilot study, in addition to the five variables measuring report quality, listed previously, we included a microlevel Cross-Evaluation based on the quality of collected information items. This precedes the macrolevel Cross-Evaluation, which gives a four-stage model as illustrated in Figure 4.

Stage I. Information collection. In the first stage all of the participants, using one of the systems (either one of the two experimental systems or the null system), gather and

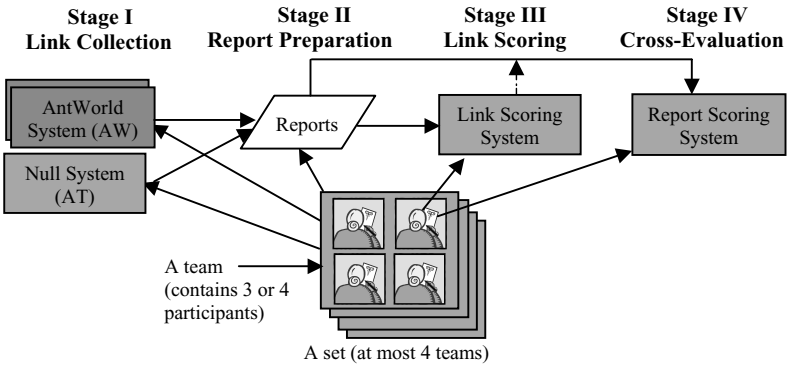

FIG. 4. The four-stage model for AntWorld evaluation.

DOI: 10.1002/asi

roughly annotate information bearing on a particular topic. This stage extends over a 10- to 14-day span, during which individuals are expected to work approximately 10 hours each. The time is intentionally extended, in order to permit the experimental system to accumulate links and comments.

Stage II. Report preparation. In the second stage, users create reports based on the information collected in the first stage. Each team was asked to select a lead editor, who worked with an editing tool (which we developed) to select among the links that had been gathered and organize them into a coherent narrative. Other members could read and make suggestions before the final version of the report was submitted.

Stage III. Link scoring. The third stage proceeds only after all participants in the set of teams working on a task have produced their draft narratives. The links included in any of the narratives on the specific task were merged into a union (i.e., duplicates were removed). Each individual participant was then asked to evaluate the union. The links in the union were presented in random order in a partial effort to control bias toward links that appeared in a judge's own team's report. Participants evaluated each link in terms of its contribution to the overall final report by using a six-category scale. The definitions of the six categories are as follows:

A link is *essential* if the report cannot be good without it. If there are several ways to get the same important information, then each of them is just *valuable*.
Informative is also valuable, but less valuable.
Background is information that fills out the picture but is not key to the recommendation or decision.
Irrelevant indicates that the link should not be in the report, as it will waste the customer's time.
If you have no opinion about a particular link, choose *No Comment*.

Stage IV. Report evaluation. In the fourth and final stage, all participants were asked to judge the draft reports as a whole. The judgment is based on the five characteristics listed. Each characteristic is scored in a range from 0 (worthless!) to 5 (Marvelous!). Each user was asked to judge all of the reports of the task set.

The Evaluation Systems

We built a set of tools to facilitate the four-stage evaluation model. This includes a Cross-Evaluation system to collect judgments and an administration system to facilitate project coordination. Because of the highly distributed nature of the experiment, all the experimental systems are built by using the JAVA Web application tools (JSP and Servlet) to make them accessible through regular Web browsers. In the background, we use the MYSQL database system to store data.

The link scoring tool. The link scoring interface is shown in Figure 5.

On the basis of the login information, the system shows, in random order, all the URLs collected by all of the teams in the set to which the user belongs. The instructions at the top of the window are a part of the instructions given to the participants. Because there are quite a few links to score, the system saves links that have been scored and, when the same

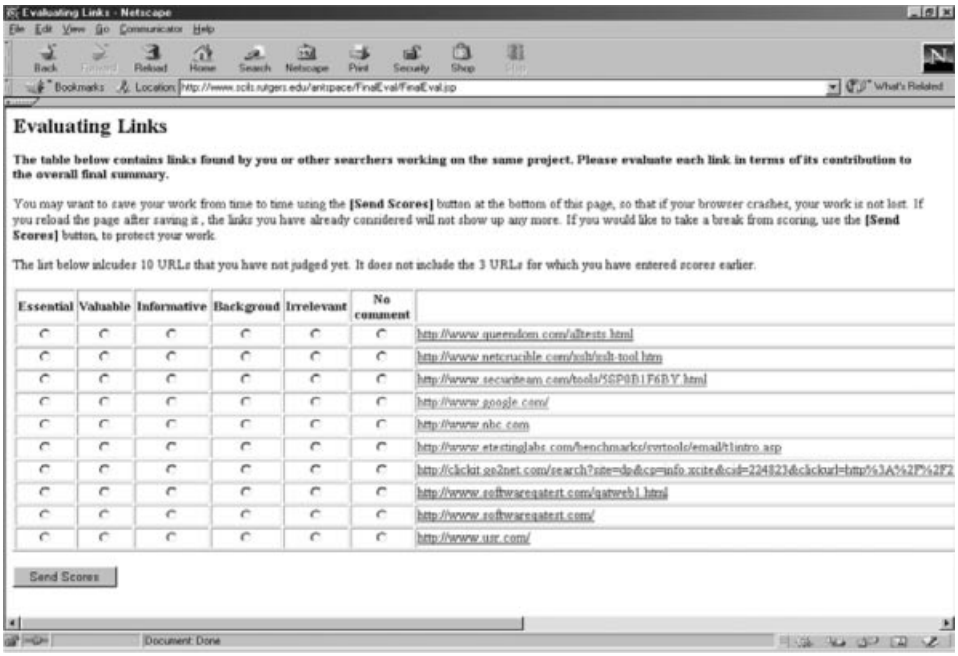

FIG. 5. The interface of link scoring tool.

DOI: 10.1002/asi

FIG. 6. The interface of the report evaluation tool.

participant logs in again (subsequently), it displays only those links not yet judged.

The report evaluation tool. The report evaluation tool is useds for evaluating completed reports, as shown in Figure 6. When the user begins report evaluation, the system automatically checks the status of document-based evaluation. Only after all members in the set finish scoring the links does the report evaluation begin. Before presenting the report the system automatically annotates the reports (details of the annotation are discussed in the next section).

Across the top of the page is a table that displays all of the (team) reports the participant needs to score and how he/she has scored each team's report so far. The team names in the table are links. Following a link will display the corresponding team's report in the table on the lower right side. The participant actually judges a report on the lower left side of the table.

The displayed narrative has its links annotated (by our program) to summarize the results of the scoring exercise just completed. The annotation for each link in the reports includes four statistics:

Avg: the average score assigned by all the judges to the link.
NC: the number of judges who had no comment on the link.
C: a measure of how well the judges agree. The calculation is shown in Equation 2. With V the variance of the judgments, C was used as the confidence measure. N is the total number of judges and a is the average score assigned by all the judges to the link.
MS: my score, the score assigned by the user who is judging the report.

$$V = \frac{\sum_{i=1}^{N} (score\ (i) - a)^2}{N}$$

$$C = \frac{1}{1 + V}$$

Confidence Measure (2)

The participant is able to select reports one after another and review them and may go back to revise scores assigned to an earlier report.

The administrative system. Managing this experiment is difficult because of the distributed nature of the experiment. Participants worked remotely from home computers. In addition to all the information posted on the project Web site and instructions provided with the tools, we have a coordinator to instruct the participants. An administrative system was developed to facilitate the coordinator's work. It includes the functions for managing the participant accounts (create new accounts, assign participants into teams, drop participants who cannot continue the experiment, etc.). The coordinator can also use the tool to examine and control the progress of the experiment. For example, after 10 days of collecting information using the assigned system, the coordinator sets a system flag so that the participants cannot add any further links to their team page. Similarly, the final report evaluation cannot begin until all participants finish scoring the links.

The Participants

The original sponsors of the AntWorld system, and of much work of this type, are primarily interested in

DOI: 10.1002/asi

developing tools that will be effective for professional analysts and searchers who spend days or weeks monitoring or researching a particular topic. For users of this type, a 3-hour or even a 10-hour learning process is not a serious barrier, if the results of that process are substantial improvements in efficiency and effectiveness.

We recruited participants by two paths, one yielding experts at search, drawn from the reference library community, and the other yielding experts in intelligence analysis, through contact with a set of retired government analysts. There were altogether 21 participants.

Pilot Experimental Design

In this pilot study, a single instantiation contains data on as many as four "teams" that together form a "task set" or "group," addressing the same task. With the same task (in the same task set), different teams are assigned to use different tools.

Complete study of the collaborative system could in principle involve all of the following variables: the TASK, the SYSTEM, some characteristics of the TEAM as a team, and some characteristics of the INDIVIDUALS who make up the team.

In this study we did not attempt to explore characteristics of the team as a team, although this is a rich area for future investigation. As noted, our budget did not allow the design to regroup participants, a procedure that is necessary if individual factors are to be isolated. Therefore, we intended to keep TEAMS together for the length of the study, so that all phenomena of training and learning could be interpreted as characteristic of the team itself. The proposed design was a Latin square (Table 2).

However, some of the participants had to leave in the middle of the experiment, so that, in fact, the teams were not kept together for the length of the study. The total number of participants was 21. Ultimately, only five individuals participated in all of the three tasks. Seven participated in two tasks and nine users participated in only one task. We tried to keep the same group of users together in one team and constitute new teams as the process evolved. The resulting implemented experimental design is shown in Table 3. The total number of "team members" was 40. Each team was assigned a name from the Chinese animal zodiac.

This pilot design makes it impossible to isolate differences among teams. Therefore, one of the most important issues in the study of collaborative or interactive systems for information finding, the learning effects, is not examined.

TABLE 2. Proposed Latin square design.

	SYSTEM (*n* is the size of the team)		
Task	*AT*	*AW*	*AW*2
AN	Team1	Team 2	Team 3
TE	Team 2	Team 3	Team 1
NW	Team 3	Team 1	Team 2

TABLE 3. Experimental design: Teams by TASK and SYSTEM.

	SYSTEM ((*n*) is the size of the team)			
Task	*AT*	*AW*	*AW*2	Total
AN	Tiger (3)	Horse (3)	Rabbit (4)	10
TE	Monkey (4), ox (4)	Dragon (3)	Rooster (4)	15
NW	Rat (4), snake (4)	Sheep (4)	Dog (3)	15
TOTAL	19	10	11	40

We assert, however, that in a full evaluation they should and must be taken into account.

Pilot Analytical Model

In this pilot study, because of the budget limitation and unexpected loss of participants, there had to be two modifications of the proposed model. First, the author variable was removed from the model. Because the teams are not kept constant through the experiment, we cannot examine this effect. Second, we extended the analytical model to examine interaction effects. We did not ignore the interaction effects because (1) theoretically, we do not know whether in the population the interaction effects are negligible, and (2) practically, our experiment was not a Latin square design, which would assume that they are unimportant. We included in the model all second-order interactions. Then we simplified the model by removing all interactions that do not have a significant impact on the dependent variables. Equation 3 is the model actually used in the pilot AntWorld Evaluation

$$V(j,t,s,b) = \lambda^0 + \lambda_j^J + \lambda_t^T + \lambda_s^S + \lambda_b^B + \sum \lambda_{f_1,f_2}^{F_1,F_2} + e$$
$$F_1, F_2 \in \{J, S, B, T\}$$
$$F_1 \neq F_2$$

AntWorld Evaluation Model (3)

Results of the Pilot Experiment

Report-Based Evaluation

On the report-based evaluation level, the impact of systems on the completion of tasks, that is, on the quality of the final products, was examined. The performance was evaluated by considering the scores assigned to the reports as a whole. This is based on a realistic view of the searching task for the target user group of the system, which is to "gather valuable information and integrate it into a report."

Before reporting the statistical analysis, we look at the figures for report judgments. As can be seen in Table 3, there were five reports created with the null system (AT), and each report was judged by the users who participated in the corresponding task, so there are, in total, 70 judgments. Each judgment includes five 0–5 scores, each on one of the five characteristics. In a similar way, there are 30 judgments on systems AW and AW2, respectively. We broke the scores

DOI: 10.1002/asi

TABLE 4. Distribution of user assignments to each report characteristics grouped by system variable.

Characteristics	Score	System AT ($n = 70$)	AW ($n = 40$)	AW2 ($n = 40$)	Total
Covers the important ground	≥4	54.3%	45.0%	50.0%	50.7%
	<4	45.7%	55.0%	50.0%	49.3%
Avoids irrelevant materials	≥4	41.4%	47.5%	32.5%	40.7%
	<4	58.6%	52.5%	67.5%	59.3%
Is well organized	≥4	54.3%	27.5%	52.5%	46.7%
	<4	45.7%	72.5%	47.5%	53.3%
Reads clearly and easily	≥4	47.1%	30.0%	57.5%	45.3%
	<4	52.9%	70.0%	42.5%	54.7%
Overall rating	≥4	54.2%	37.5%	55.0%	50.0%
	<4	45.8%	62.5%	45.0%	50.0%

into two groups: "high" (if the score is 4 or 5) and "other." The percentage in each group, for each of the five measures, on all reports is grouped by the system variable and summarized in Table 4.

The tabulation shows that 55% of the participants considered the overall rating of reports written with AW2 to be equal to or higher than 4, which is barely better than the control system AT (54.2%). AntWorld without previous knowledge (AW) only has 37.5% of judgments at this level. Because the SYSTEM effect may be confounded with other possible effects, further analysis is necessary to check the accuracy of this ranking.

Analysis of the leading factor of quality. We next performed a factor analysis, to see whether the five measured variables are independent. If our instrument, the five-characteristic report quality measurement, has a balanced set of questions that accurately reflect the decision makers' concerns, then factor analysis is a good way to summarize them.

The results show that there is essentially one concept being judged. The first factor explains 71.5% of the variance with an eigenvalue equal to 3.57. The weights of the first factor on the five characteristics, which are correlations between the five characteristics and the first factor, are shown in Table 5. Apparently, the first factor is generally highly correlated with all five variables. We refer to the first component as the *leading factor* of report quality.

With the leading factor as the dependent variable, we applied the general analytical model of Equation 3, which is a univariate general linear model (GLM) analysis. Our search for interactions revealed only one significant interaction effect (between TASK and SYSTEM). The analytical model is then simplified as shown in Equation 4. (The SPSS syntax is in the Appendix.)

TABLE 5. Weights of the leading factor load on five characteristics (extraction method: principal component analysis).

	Weights on the first factor
Overall rating	.94
Is well organized	.92
Reads clearly and easily	.88
Covers the important ground	.76
Avoids irrelevant materials	.70

TABLE 6. Analysis of variance for the leading factor of report quality.

Source	Type III sum of squares	*Df*	Mean square	*F*	Sig.
Corrected model	51.37	29	1.77	2.18	.00
Intercept	.10	1	.10	.13	.72
TASK	.41	2	.21	.25	.78
SYSTEM	2.37	2	1.18	1.46	.24
SELF	3.62	1	3.62	4.45	.04
TASK * SYSTEM	17.80	4	4.45	5.47	.00
JUDGE	24.32	20	1.22	1.49	.10
Error	97.63	120	.81		
Total	149.00	150			
Corrected total	149.00	149			

$$V(j, t, s, b) = \lambda^0 + \lambda_j^J + \lambda_t^T + \lambda_s^S + \lambda_b^B + \lambda_{t,s}^{T,S} + e \quad \text{Simplified Analytical Model} \quad (4)$$

Table 6 shows the impact of each main effect, and of the interaction effect included in the model. The SYSTEM effect is greater than the TASK effect, but still not statistically significant. The interaction effect accounts for a substantial part of the variance and *is* statistically significant. Self-judgment bias is another significant source of the variance. The model shows that with the larger effects of task-by-system interaction, self-judgment bias and the effect of the judge being corrected, the SYSTEM effect can be measured, although in this particular experiment it is not statistically significant.

We can also examine the relative magnitudes of the four main effects and one interaction effect, using the ranges, or differences between the highest and lowest standardized lambda scores for the particular effects.

As seen in Table 7, we can, with only about 75% confidence, draw the conclusion that the SYSTEM variables affect the leading factor of report quality. This is lower than our confidence for the judge (90%), the self-judgment bias (96%), and the interaction effect (100%). However, Table 7 shows that the magnitude of the SYSTEM effect is higher than that of the judge, task, and self-judgment effects, yet lower than that of the system and task interaction effects. The fact that the smaller differences can be more significant than larger ones is caused by the different normal distributions of the measured variables. The shape of the self-judgment bias

TABLE 7. Relative magnitudes of effects analyzed in Table 6.

Effects	Range
SYSTEM and task interaction	4.70
SYSTEM	4.15
Judge	3.82
Self-judgment bias	2.11
Task	1.96

DOI: 10.1002/asi

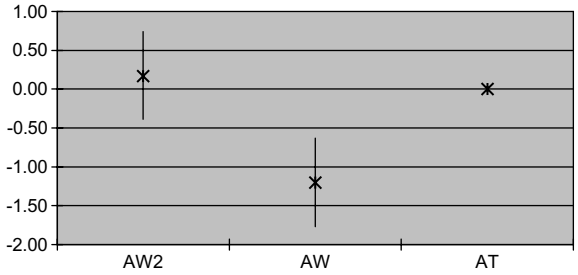

FIG. 7. System effect on leading factor of report quality (bars represent the 95% confidence intervals).

magnitude distribution is sharper and the mean is nearer to the zero point than that for the system effect magnitude.

Our ultimate concern is to learn the system factor's effect on the leading factor of report quality. Because there are three levels of the system factor (AT, AW, and AW2), it is of interest to find the relative relationships among these three levels. We find that the interaction of the system variable and the task variable has an effect on this factor. The system variable contributes to the leading factor in two ways: through the pure system effect and through the interaction effect with the task variables. We can look at the relative relationships among the three systems' contributions to the leading factor in three ways.

First, we examine the parameter (λ) estimated for each value of the SYSTEM variable. This indicates only the pure system effect. The *Statistical Package for the Social Sciences* (*SPSS*) analysis chooses its default settings arbitrarily and has set the control system AT to the 0 point ($\lambda_{AT}^{S} = 0$). The estimated λ of AW is -1.2; the λ of AW2 is 0.17. We found that the systems are separated into two groups. To examine the 95% confidence intervals for the three systems' parameters, λs, we plot the parameters of three systems in Figure 7. It is clear that the estimated mean parameters for AW2 and AT fall outside the 95% confidence interval for AW. This finding indicates that the system effect of AntWorld with relevant information in it (AW2) contributes significantly more to the leading factor of report quality than that of AntWorld without such relevant information (AW) does. We see, also, that the Annotation Tool (AT) is significantly better than AntWorld without relevant information (AW) at the 95% confidence level. Finally, AW2 is not different, at the 95% confidence level, from AT.

Second, we examine the estimated marginal means of three levels of the system variable, which reveals relative relationships among three systems' overall effect on the leading factor. The result shows that AntWorld with previous knowledge (AW2) and naive AntWorld (AW) are not significantly different at the 95% confidence interval (Table 8). Thus, even though the pure system effect shows a significant difference between these two levels, the system and task interaction effect may have opposite effects on the leading factor, which averages out the significant difference between the two versions of the AntWorld system, in this experiment.

TABLE 8. Estimated marginal means comparison.

			95% Confidence interval	
	Mean difference	Sig.	Lower bound	Upper bound
AW2 vs. AW	.35	.095	−.06	.76
AW2 vs. AN	.13	.506	−.25	.50
AW vs. AN	−.22	.252	−.60	.16

Third, to verify this interpretation, we analyzed the system effect for each task separately. Because we care most about finding the difference between the two experimental systems, we consider only the contrast between two specific levels of the system variable: AW and AW2. The results (Table 9) show that AntWorld with previous knowledge has a significant positive effect for only one scenario: Terrorist Sites.

Multifactor analysis. A multivariate GLM with the same set of effects was run to examine the SYSTEM variable's impact on each of the five individual report criteria. Not surprisingly, we found similar patterns—all five criteria can separate AW2 from AW at the 95% confidence level (Figure 8).

The report-based evaluation analyses and results show that we can detect the relative differences among three systems, even though the differences are not statistically significant. Here, we only present the analyses and results for the system effect. A similar procedure can be applied to study the other effects, for example, to answer the question of which tasks are relatively easier.

To sum up, we find that the AntWorld system, when it contains information from previous users, is significantly "better" than the AntWorld system without such information. It is, however, not significantly better than the Annotation Tool that was built to provide a comparison baseline. We find, also, that the significant differences all originate in one of the three tasks. Thus this analysis does not establish a uniform superiority of AW2 over AW.

Link-Based Evaluation

In the link-based evaluation part of the analysis, the impact of the system variable on the quantity and quality of saved links was considered. We measured the performance of a team (and, in fact, the effect of the system used) by computing several numbers: the total number of links that the team collects and keeps in its final report and the total value

TABLE 9. Contrast results between AW2 and AW.

			95% Confidence interval	
Scenario	Contrast estimate (AW2-AW)	Sig.	Lower bound	Upper bound
Anthrax	.08	.84	−.72	.88
Nuclear weapon	−.42	.21	−1.07	.24
Terrorist sties	1.37	.00*	.72	1.03

DOI: 10.1002/asi

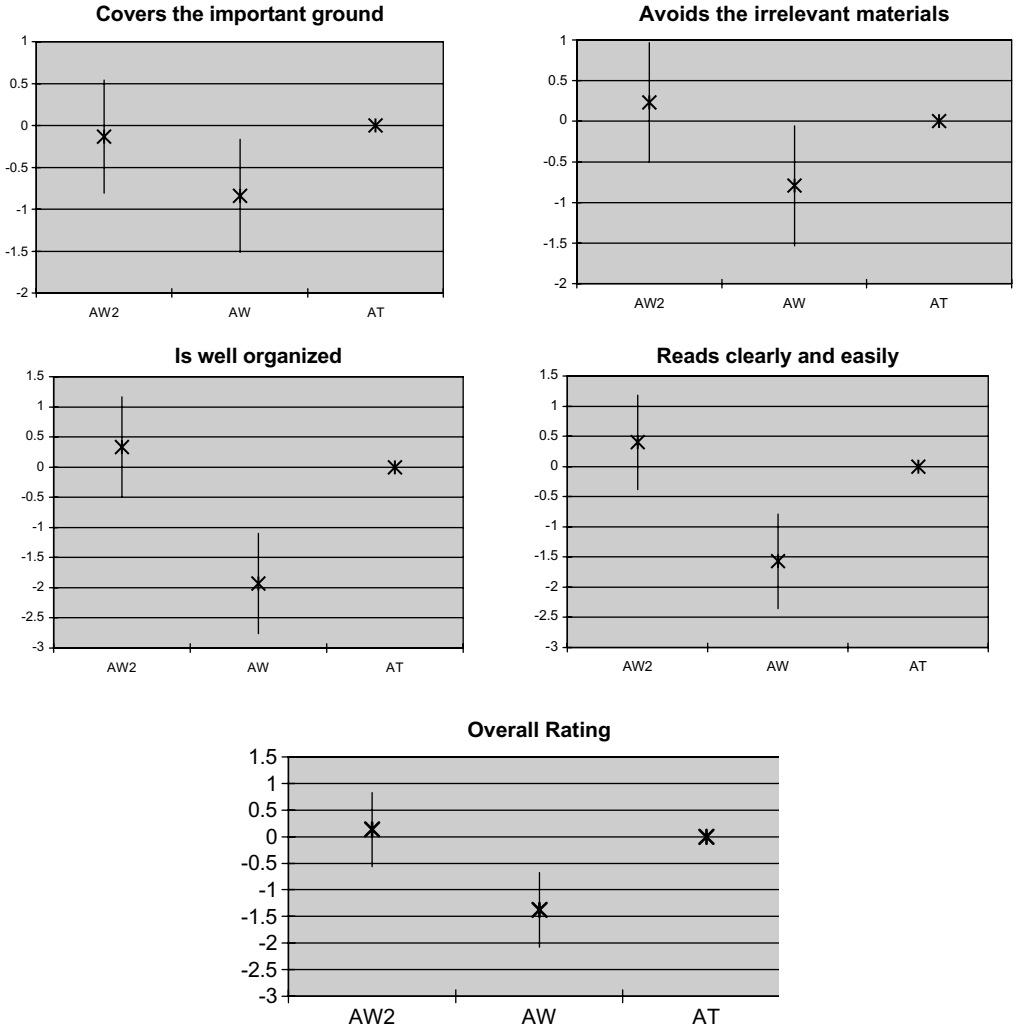

FIG. 8. System effects on five report judgment criteria.

of those links, as computed from the average of all judgments made by the members of *all* teams in the task set. As mentioned, each link was evaluated by using a six-category scale. We recoded the categories into numerical data (Table 10). In recoding, we penalized the irrelevant links.

There are several important questions about the systems that we can ask on this level:

1. What is the relationship between the number of links and the total value of the links saved?
2. What are the impacts of each SYSTEM on the quality of retrieved information?
3. What conclusion can we draw from the comparison?

To answer question 1, we calculate the total value of links in each report as

$$V(t) = \sum_{i=1}^{m} \frac{\sum_{j=1}^{n} S(l_i, j)}{n}$$

where m = the number of links in team t's report, n = the number of participants in the task set in which team t joined, $S(l_i, j)$ = judge j's judgment of the ith link in the report.

TABLE 10. Recode link judgments.

Category	Recode
Essential	4
Valuable	3
Informative	2
Background	1
Irrelevant	−3
No comment	0

We can evaluate a team report by computing both the total number of links that the team keeps in its final report, and the total value of those links. The results are visualized in Figure 9.

We can see that one AW2 team (rabbit) has the highest *total* for both total value and total number of links. The AW2 team (rooster) has high *ratio* of total value to total number of links. But one AW team (sheep) and two AT teams (tiger and ox) also have high ratios.

As with the quality of reports, the values of links may be affected by multiple factors. As earlier, we want to identify the system effect while taking into consideration other possible effects. As shown in Table 11, we found that only one

DOI: 10.1002/asi

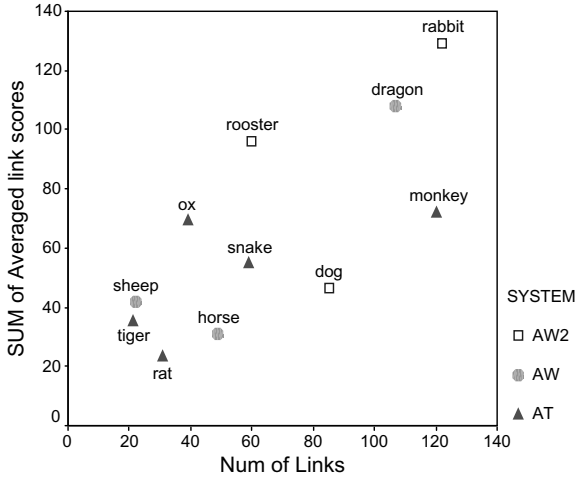

FIG. 9. The total score of links included by each team in their final report, as judged by all members of the same task set, plotted against the total number of links in a report.

interaction (the same one found earlier) has a significant effect on the link scores (the other interaction effects are not shown).

As shown in Figure 10, we see that both versions of the AntWorld system score above the control system AT. This is different from the relation found when studying the quality of reports as a whole. And the 95% confidence interval of the AntWorld with relevant knowledge (AW2) does not have any overlap with the other two systems. Such a result suggests that using the AntWorld system when it contains previous information is more effective, link by link, than using either the Annotation Tool or the AntWorld system without "prior knowledge." We may speculate that the AntWorld alone is a little more effective than the Annotation Tool (a result that is not significant at the 95% confidence level) because it provides more searching tools (a query can be easily submitted to several search engines).

TABLE 11. The impacts of each major effect and interaction effect.

Source	Type III sum of squares	*df*	Mean square	*F*	Sig.
Corrected model	3547.09	29	122.31	30.11	.00
Intercept	5568.12	1	5568.12	1370.56	.00
SELF	347.26	1	347.26	85.48	.00
TASK	3.96	2	1.98	.49	.61
SYSTEM	9.13	2	4.56	1.12	.33
JUDGE	2174.87	20	108.74	26.77	.00
TASK * SYSTEM	772.40	4	193.10	47.53	.00

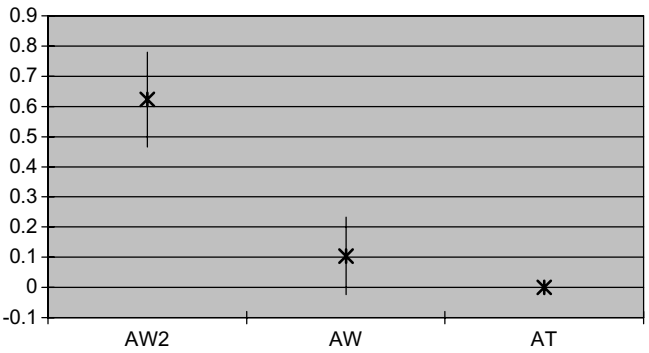

FIG. 10. System effect on the individual link scores.

Conclusions and Discussion

The pilot study described tells us something about the value of the AntWorld system. First, the evaluation was conducted under conditions that are as real as possible. The participants were from the target user group of the AntWorld system. They worked on their own computers in the environments with which they are familiar. The experimental systems, as well as the collection (the Web), are real.

Second, the statistical method successfully isolated the effect of each variable. The data gathered were analyzed at the level of the entire report, which represents the end product of the collaborative task, and at the level of individual links, which are scored for their "relevance to the task." The results indicate that the mechanism of collecting previous searchers' knowledge used in the AntWorld system holds promise to improve professional searchers' work on complicated tasks at both the report-based and the link-based evaluation levels. The present study is able to make statistically significant distinctions on both levels, but the study is too small to be definitive. Because of the lack of continuity in the teams, the results do not reflect the impact of searchers' difference, which remains one of the most challenging issues in evaluating collaborative or interactive systems.

We believe that this study has value well beyond the immediate results. There are four key aspects to this belief. First, this study provides a successful model to make concrete the "three-realities" paradigm without becoming entangled in the troublesome issues of defining users and tasks. In contrast to previous work adopting the "three-realilities" paradigm, we use the searchers themselves as judges, while the analytical model controls the potential hazard of self-judgment bias.

Second, rather than examining the effects of all variables one at a time, our model looks at all of the effects jointly. Therefore, the isolated effect of each variable (the lambda parameters of the analysis of variance) is the pure effect of the corresponding factor, and the variance of the dependent variables caused by other effects has been removed. We have shown that the Cross-Evaluation method with careful statistical analysis can resolve distinctions among systems.

Through this analysis the distinctions between systems are not confounded with the effects of users, topics, and so on. We believe that the proposed model is of potential interest for all evaluations of information system that are highly interactive or collaborative, or both. The pilot study we report here shows the types of data and analysis that are needed to extract useful conclusions.

Third, we present a method for measuring systems at the level of task completion, as well as at the level of the individual search results. This is significant in three ways: (1) It is theoretically desirable to evaluate at the level of the

DOI: 10.1002/asi

complete work product. As mentioned, it has been widely accepted in the studies of information seeking that searching is not the final goal of users. Searching is in the context of users' tasks at hand. So, searching tools should be evaluated in terms of helping users to finish tasks. Our participants accepted the report-based evaluation as a reasonable method. (2) It is practically economical, especially for studies with the type of users in our study. The time of expert information analysis is expensive. Skipping the time-consuming link-based evaluation would save a great amount of time and money. (3) Report-based evaluation provides a new measurement direction for systems that do not have mature link-based measures. For example, the traditional relevance-based measures do not fit the requirements of interactive QA systems. Of course, reports are not necessarily the final product of all searching tasks. However, as long as final products can be cross-evaluated, they will fit in the models proposed.

Fourth, the evaluation model described can be easily transferred, with minor modifications, to other systems for longtime users with complicated tasks. The design and analytical model reported can be used as a model for testing other complex domain such as QA systems. In fact, while this report was being prepared, the method was extended to the evaluation of QA systems, with a study of the HITIQA system (Sun, Kantor, Strzalkowski, Rittman, & Wacholder, 2004). It has also been applied in a study of metrics for QA systems in general (Kantor & Kelly, 2004; Kantor, Kelly, & Morse, 2004).

In summing up, a key objective of the evaluation model is to bypass the difficult user issue for interactive information system evaluation studies of which users are not the focus. However, as pointed out by one anonymous reviewer, the pilot study reported here could not complete the experimental design as initially planned because participants dropped out. Any experiment that involves users over an extended period will face problems of this kind. The method of analysis proposed here makes it possible to recover useful information even in the face of such problems.

Acknowledgments

This research was supported in part by the Defense Advanced Research Projects Agency (DARPA) under contract number N66001-97-C-8537 to Rutgers University. The application to HITIQA was supported in part by Advanced Research and Development Activity (ARDA) under contract number 2002-H790400-000. The first author acknowledges support of an ARDA Challenge Internship in the summer of 2004.

We thank Robert Rittman and Shelee Saal for their work coordinating our participants, and thank Vladimir Meñkov for his help and insightful comments on system design and implementation and an anonymous reviewer for some very helpful comments. And last, but not least, we offer special thanks to our anonymous participants.

References

Barry, C.L. (1994). User-defined relevance criteria: An exploratory study. Journal of the American Society for Information Science, 45, 149–159.

Boros, E., Kantor, P., & Neu, D.J. (1999). Pheromonic representation of user quests by digital structures. In Proceedings of tthe 62nd Annual Meeting of the American Society for Information Science (Vol. 36, pp. 633–642). Medford, NJ: Information Today.

Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagui, S., Isreal, D., et al. (2001). Issues, tasks and program structures to roadmap research in question & answering (Q&A). Retrieved December 3, 2004, from http://www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc

Cleverdon, C.W. (1960). The ASLIB Cranfield Research Project on the Comparative Efficiency of Indexing Systems. ASLIB Proceedings, 12(12), 421–431.

DUC (2004). The Document Understanding Conference. Retrieved from http://duc.nist.gov

Kantor, P., Boros, E., Melamed B., & Meñkov, V. (1999) The information quest: A dynamic model of user's information needs. Proceedings of the 62nd Annual Meeting of the American Society for Information Science, 36 (pp. 536–545). Medford, NJ: Information Today.

Kantor, P., Boros, E., Melamed, B., Neu, D.J., Meñkov, V., Shi, Q., et al. (1999) AntWorld. In Proceedings of SIGIR'99, 22nd International Conference on Research and Development in Information Retrieval (p. 323). New York: Association for Computer Machinery.

Kantor, P., & Kelly, D. (2004). ARDA Challenge Workshop 2004: Metrics for question answering systems final report. Unpublished report.

Kantor, P., Kelly, D., & Morse, E. (2004). ARDA Challenge Workshop 2004: Metrics for question answering systems. Unpublished manuscript.

Kantor, P., & Voorhees, E. (2000). The TREC-5 confusion track: Comparing retrieval methods for scanned text. Information Retrieval, 2(2/3), 165–176.

Lagergren, E., & Over, P. (1998). Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. In Proceedings of the 21st International Conference on Research and Development in Information Retrieval SIGIR'98 (pp. 164–172).

Over, P. (2001) The TREC interactive track: An annotated bibliography. Information Processing and Management, 37(3), 369–381.

Robertson, S., & Soboroff, I. (2002). The TREC 2002 filtering track report. In the Notebook of the Text REtrieval Conference, 2002. Retrieved December 3, 2004, from http://trec.nist.gov//pubs/trec11/papers/OVER.FILTERING.pdf

Saracevic, T., Kantor, P., Chamis, A.Y., & Trivison, D. (1988a). A study of information seeking and retrieving: 1. Background and methodology. Journal of the American Society for Information Science, 39(3), 161–176.

Saracevic, T., & Kantor. P. (1988b). A study of information seeking and retrieving: 2. Users, questions, and effectiveness. Journal of the American Society for Information Science 39(3), 177–196.

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

Schamber, L. (1994). Relevance and information behavior. Annual Review of Information Science and Technology, 29, 3–48.

Sun, Y., Kantor, P., Strzalkowski, T., Rittman, R., & Wacholder, N. (2004). Cross evaluation—A pilot application of a new evaluation mechanism. In the Proceedings of the 2004 Annual Meeting of American Society for Information Science and Technology (pp. 383–392). Medford, NJ: Information Today.

TREC QA track. Retrieved December 3, 2004, from http://trec.nist.gov/presentations/TREC10/qa/

DOI: 10.1002/asi

Appendix

Statistical Package for the Social Sciences syntax for the report-based analysis.

A: Factor Analysis

The following code extracts the factors underlying the five measured variables.

```
FACTOR
 /VARIABLES cover noirr organ clear overa /MISSING LISTWISE /ANALYSIS cover
 noirr organ clear overa
 /PRINT INITIAL EXTRACTION
 /CRITERIA MINEIGEN(1) ITERATE(25)
 /EXTRACTION PC
 /ROTATION NOROTATE
 /SAVE REG(ALL)
 /METHOD = CORRELATION
```

B: Univariate

The following code examines specific contrasts among the effects of the system and the task, on the leading factor from analysis (A).

```
UNIANOVA
  fac1_1 BY task system self user
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /PRINT  = PARAMETER
  /LMATRIX 'System'
system 1 −1 0 task*system 1 −1 0 0 0 0 0 0 0;
system 1 −1 0 task*system 0 0 0 1 −1 0 0 0 0;
system 1 −1 0 task*system 0 0 0 0 0 0 1 −1 0
  /EMMEANS = TABLES(system) COMPARE ADJ(LSD)
  /CRITERIA = ALPHA(.05)
  /DESIGN = task system self task*system user .
```

C: Multivariate

The following *SPSS* code examines the dependence of all five measured variables on the four factors jointly, with all main effects and only one interaction effect.

```
GLM
cover noirr organ clear overa BY task system self judge
/METHOD = SSTYPE(3)
/INTERCEPT = INCLUDE
/PRINT = PARAMETER
/CRITERIA = ALPHA(.05)
/DESIGN = task system self judge task*system.
```


DOI: 10.1002/asi