Evaluation of Intelligent Information Access Systems

Jean Scholtz

National Institute of Standards and Technology
100 Bureau Drive, MS 8940
Gaithersburg, MD. 20817
jean.scholtz@nist.gov

Abstract

Evaluation is essential for showing the benefit of intelligent systems. This requires metrics that are user-centered. User-centered metrics also serve to help researchers understand where their work fits within the users' work environment. This paper discusses traditional evaluations for information access systems and proposes metrics more suited for evaluation of intelligent information access systems.

Introduction

The Information Access Division at the National Institute of Standards and Technology has conducted the Text Retrieval Conference (TREC) for a number of years (http://trec.nist.gov/). As stated on the TREC home page the goals of this conference are to:

- encourage research in information retrieval based on large test collections;
- increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

Over the past 13 years TREC has expanded by adding tracks. A track focuses on a particular area of interest within information retrieval and develops the methodology and infrastructure for evaluation. The following tracks are examples of current tracks within TREC:

Cross-Language Track

This track investigates the ability of retrieval systems to find documents that pertain to a topic regardless of the language in which the document is written.

Filtering Track

This track uses a stream of new documents and the system is asked to find information relevant to the user's stable information needs. For each document in the stream, the system must make a binary decision as to whether the document should be retrieved (as opposed to forming a ranked list).

Genome Track

The purpose of the track is to study the retrieval of genomic data, where genomic data is broadly interpreted to mean not just gene sequences but also supporting documentation such as research papers and lab reports.

HARD Track

The goal of HARD is to achieve High Accuracy Retrieval from Documents by leveraging additional information about the searcher and/or the search context, through techniques such as passage retrieval and targeted interaction with the searcher.

Interactive Track

This track studies user interaction with text retrieval systems. Participating groups develop a consensus experimental protocol and carry out studies with real users using a common collection and set of user queries.

Novelty Track

This track to investigate systems' abilities to locate new (i.e., non-redundant) information in successive documents.

Question Answering Track

A track designed to take a step closer to information retrieval rather than document retrieval. Participants interact with the system through more natural questions rather than Boolean search techniques.

The methodology used for the many of these tracks is to compute recall and relevance for each system. Relevance is ground truth. This is supplied by human assessors and the documents returned by each system is compared to these judgments. These metrics have contributed greatly to improvements in the information retrieval community. However, as we begin to explore intelligent information retrieval systems we need to explore other measures. The interactive track, the novelty track, and the question-answering track need methodologies and metrics that go beyond ground truth assessment and precision and recall. Others in human-computer interaction research have suggested that having common scenarios and metrics would help researchers in the field to show progress (Whittaker, Terveen, and Nardi, 2000).

Intelligent Information Access Systems

A number of research programs currently are employing artificial intelligence (AI) techniques to develop systems that work collaboratively with information analysts. Goals for these systems include:

- Use the analysts' evolving mental creations, understandings, and knowledge throughout a tightly coupled interactive and spiraling process.
- Support high-level exploration and interaction with information to free analysts to focus on issues that matter instead of on details and system functions that don't. This includes the concept of engaging in a dialogue with a system to obtain information rather than having to formulate low-level queries to a search engine.
- Reveal new indicators, issues, and/or threats that might be overlooked by the analyst because of the size of the data or biases of the analyst.

Intelligent information retrieval includes a number of research areas:

- Question understanding and contextual interpretation.
- Determining the answer which involves information retrieval and extraction from multiple media/languages and data types, interpretation, synthesis, and the resolution of conflicting information and a justification.
- User models that capture the cognitive preferences, goals, knowledge, and abilities of analysts and use that information to tailor the behavior of the information system to both facilitate what the analyst is attempting to do as well as to overcome any biases or lapses in the analysis process.
- Use of prior and tacit knowledge that an analyst brings to bear. Currently users have to filter information retrieval results to obtain new information. This research examines ways to capture this information from the analyst and to automatically filter retrieved information that is already known.
- Formulation of hypotheses based upon evidence. Analysts use evidence marshalling to assess how much support there is for a particular argument of hypothesis. Can intelligent systems use agent-based technologies to develop hypotheses and assemble evidence or establish contradictory hypotheses to help the analyst assess more possibilities?

Challenges for Evaluation

Previous evaluations for information retrieval relied upon the concept of a large corpus of data, the ability to establish “ground truth”, and the metrics of precision and recall. Evaluations for interactive, intelligent systems cannot be conducted in this manner. Ultimately the evaluations must be conducted empirically to assess the impact on the user. Evaluations in a real-world environment preclude the assessment of “ground truth” and therefore, demand the development of metrics other than precision and recall. Moreover, precision and recall are not sufficient to measure user impact.

In addition to the development of new metrics, we need to develop methodologies for evaluation. Typically, real-world evaluations are conducted as field studies. These are used mostly for usability evaluations (Wixon and Ramey, 1996) and rely heavily on observations to collect data about the usability of software in the context of use.

While laboratory studies try to replicate context of use, many issues such as interruptions, interoperability with other software, and an understanding of actual workflow can only be assessed in this type of environment. Laboratory studies of usability also do not take into account learning curves. Field studies conducted at different intervals in time can help assess how users have learned and adopted the software into their work process.

We want to go beyond assessing the usability of the software and measure the utility of the intelligent information systems. Furthermore, the research work we are interested in evaluating is being conducted not just to support a current work practice but to support an envisioned work practice.

We propose a three level approach to evaluation: evaluation of the science, evaluation of component technology, and impact evaluation. Different methodologies are needed for evaluation of each of these levels. In some cases, these methodologies already exist. The challenge is to develop metrics for each type of evaluation as well as a mapping between the metrics for the three types of evaluations.

Metrics for Intelligent Information Access Systems

Our goal is to develop metrics that are user-centered. Precision and recall are technology-centric or scientific metrics. They measure the performance of algorithms but the values for these metrics don't necessarily ensure a benefit to the end user. We want to develop metrics that can be useful both for system improvement but will also contribute to a substantial benefit from the perspective of the end user. In addition to developing appropriate metrics we will also need to consider what data can be obtained and used to compute the necessary measures.

Impact Metrics

Potential system impact metrics include:

- Trust in the system
- Shift in user time
- Increased quality of product
- Increased confidence in product

Measurement of these attributes will encompass both quantitative and qualitative data. Trust in the system can be measured by the amount of system suggestions that information analysts further explore and by user ratings and comments. Trust in the system may also be a measure of the understanding the user has of how the system actually works. In that respect assessing the user's mental model of the system is appropriate (Carroll and Olsen, 1987; Norman, 1988).

Intelligent systems should alleviate some of the work that users currently do for themselves. This would hopefully allow a user to concentrate on aspects of the process that are a better match for the user's skill. The shift of time distributions can be acquired by doing a baseline of a process. Currently we suspect that a high percentage of an information analyst's time is spent in data collection. If data collection can be offloaded to intelligent software agents, then an information analyst could spend more time synthesizing data and developing arguments to support various hypotheses.

Better deliverables may also result from being able to spend the available time more wisely and from being able to cover more material. Product quality can be obtained through judgments of subject matter experts. Better coverage of data may result in an increased confidence in the analyst's recommendations. Increased confidence in the product can be measured by users' ratings.

Other overall metrics might include process improvements. To measure this, a cognitive task analysis can be conducted to provide insight about obstacles in the process. Noting where and what those obstacles are, we can use our baseline data to determine if these obstacles are removed once research tools are inserted.

While overall process metrics are the long-term goal, we also want to be able to provide component based metrics. That is, if we find that the time shift now allows users to concentrate on higher value tasks, we would like to be able to attribute this shift to the component(s) responsible.

Component Level Metrics

Component metrics are, of course, specific to the particular functionality of the component. To provide some examples of possible metrics, we go back to some current research efforts. In particular, we consider potential metrics for question and answer dialogues, user modeling, hypotheses generation, prior and tacit knowledge.

Question and answer systems currently being developed in the research community go beyond simple Q&A systems. These systems might help a user construct a travel itinerary, with flight reservations, hotel bookings and car rentals. Evaluation efforts for dialogue systems have used user satisfaction as the impact metric. Component metrics for question and answer dialogues might include:

- Completeness of answer
- Accuracy of answer
- Effort required on part of user engaging in dialogue

Recent evaluations have been done of question and answer dialogues in speech research and in the information retrieval world (Walker et al. 2001). If the question is

reasonably simplistic, a template can be constructed of the different variables that need to be filled in to constitute an answer. Upon completion of the dialogue, the percentage of variables filled in might be used as a metric. While this is a reasonable component metric, the dialogue may be of no value to the user if the answer is not complete. However, if we compared components based on the percentage of the template filled in, we might predict that components achieving a higher percentage of success will contribute more to impact metrics than components with a low completion rate. Metrics used in these evaluations were:

- Dialogue efficiency metrics such as total elapsed time, time on task, user turns, system turns, and turns on task.
- Dialogue quality metrics including sentence error rate, word error rate
- Task success metric – user perception of success.

The Paradise framework (Bonneau-Maynard, Devillers, and Rosset, 2000; Hirschman, 2000; Hirschman et al., 1993; Price et al., 1992) was used in these evaluations to integrate the above metrics to predict the ultimate metric of user satisfaction.

Researchers in user modeling expect that by understanding more of what the users are doing, intelligent systems will be able to provide appropriate help. Metrics for consideration here include time, accuracy, benefit:

- time user spends explicitly entering data/ user critique of model/ time for model to adjust to changes in interest
- comparison with current system and what information is delivered to user
- time saved/time spent (comparison)
- number of helpful interactions/total number of interactions

Components scoring well on these metrics will likely contribute to the time shift measure proposed for an impact metric.

Researchers looking at hypotheses generation intend to develop systems that can suggest hypotheses to the analyst that may have been overlooked. Assuming that we have knowledge of the data collection being used, we could use human assessors to judge the relevance of generated hypotheses. Another measure might be the percentage of hypotheses generated by the system that cause the analyst to do some further exploration. The generation of a large number of hypotheses for the information analyst to consider might increase confidence in the end report. However, having to sort through a large number of hypotheses might not produce the time shift distribution that we desire. This illustrates the necessity of looking at the mapping between component and impact metrics.

Evaluation components that make use of prior and tacit knowledge are more problematic. A subject matter expert in a particular area would be pleased to have information filtered out that he or she already knows. However, intelligence information is dynamic. Just because a country's economy was stable several months ago doesn't mean it is currently. Therefore temporal aspects of the prior knowledge have to be considered as well. One measure could be the percentage of previously known information/ all the information returned with the appropriate time intervals factored in.

Another aspect of evaluation of intelligent information systems is the length of time over which evaluation must be conducted. As systems learn and adapt to the user, the system should become more useful over time. How do we decide what the appropriate window is for evaluation? Moreover, if the system takes a long time to become useful, users may well abandon usage prematurely. Adaptive systems must also react to changes of focus for users. If an analyst is suddenly assigned to new analytic tasks, how long does it take the system to react?

Scientific Metrics

Scientific metrics are still important and need to be the first evaluation applied. Consider the same topics discussed in component metrics: question and answer dialogues, user modeling, hypothesis generation, prior and tacit knowledge.

Scientific metrics for question and answer dialogues might include:

- Recognition and handling of ambiguous questions or statements
- Context tracking for threads of dialogue
- Ability to engage in mixed initiative dialogue with end user
- Percentage of simple questions recognized and accurately answered

The domains for the discourse would have to be considered as well as any constraints on the vocabulary of the end user.

Scientific measures for user modeling might include:

- Ability to construct user model based on recognition of user activities
- Accuracy of user model
- Predictability of user model

The various types of user activities that can be recognized and used by the system would be a consideration in this scientific evaluation.

Algorithms that generate and track hypotheses would have to demonstrate the ability to:

- generate hypotheses from data
- Recognize user generated hypotheses
- Assemble evidence or contradictory information from data for multiple hypotheses

These capabilities should be evaluated using data collections of various sizes and with different signal to noise ratios.

The TREC novelty track might be a reasonable starting point for scientific evaluation for prior and tacit knowledge. Given a number of documents, can algorithms find only the new information in successive documents.

Mapping Metrics

Intelligent information systems need to demonstrate good science first. Scientific evaluations are controlled, laboratory type evaluations. Different algorithms can be evaluated and compared. Items such as number of domains, vocabulary, size of data collections, signal to noise ratio can be manipulated to test the algorithms under different conditions.

Before integrating these algorithms directly into a system for evaluation we recommend a component level evaluation that uses user-centric metrics. These evaluations may or may not be done empirically. The metrics used in these evaluations will also be used in the final impact evaluations.

For example, consider hypothesis generation and tracking algorithms. Suppose that a scientific evaluation showed that the algorithm was capable of recognizing hypotheses that an information analyst was investigating and could find alternative hypotheses in large, noisy data collections. The component metrics showed that a high percentage of the hypotheses brought to the attention of the analyst were used for further exploration. An impact evaluation showed that the analyst confidence in the final report was much higher than usual. If all other parts of the analysts' environment were constant, we could attribute this increase in confidence to the ability to explore more alternative hypotheses.

Evaluation Issues for Mobile and Ubiquitous Computing

Other examples of intelligent system evaluations are in ubiquitous computing. Of particular interest are context-aware applications. In these applications, sensory information is used to determine the context of a situation and the application modifies its behavior accordingly. The behavior could be an action performed automatically or information that is modified based on the context of the user. For example, a cell phone equipped with GPS could determine that you are in a movie theater and automatically set the notification method to vibrate rather than ring. A system might recognize that you are in a social situation and display a message in text rather than using audio as the delivery method. Location and social surroundings are the most commonly used context. Other researchers in augmented cognition and affective computing are attempting to understand the cognitive state of the user and modify the delivery of information based on what the user can currently handle.

Evaluations of such systems cannot be done solely through empirical methods as there are too many situations to test. Simulation could be used as a supplement though human judgment of the appropriateness of the modified behavior needs to be the ultimate metric (Bylund and Espinoza, F. 2002).

One possibility is to classify modified behaviors according to a severity scale and test representatives of the various classes. Possible measures include: time savings by having the action automatically performed, reduction in user cognitive load by having the action automatically performed, the time needed to "undo" an inappropriate action or the severity of not "undoing" the action, the percentage of appropriate actions in a given experiment.

Mobile computing and ubiquitous computing deliver information to users who are conducting primary tasks. The information delivery is beneficial but is a secondary task. A possible measure here is the time taken from the primary task to obtain this information. (Smailagic et al. 2001) suggest using levels of distraction as measures of a secondary activity. They propose 4 classes of distractions: snap, pause, tangent and extended. Snap activities take a few seconds and does not interrupt the primary activity. Pause activities take several minutes and will require the user to stop a primary activity but should not cause severe problems in restarting the primary task. Tangent and extended activities cause the user to switch from the primary task and concentrate on the secondary task.

Workshops on evaluation of Ubiquitous computing were held at Ubicomp 01 (www.nist.gov/ubicomp01) and Ubicomp 2002 (www.nist.gov/ubicomp02). At the earlier workshop participants identified four axes for evaluation along with what made this different for ubiquitous computing (Scholtz and Richter, 2002). These axes and suggested impact metrics are shown in Figure 1.

Others have also suggested various metrics for evaluation. Jameson (Jameson 2003) proposes five usability challenges for adaptive interfaces (i.e., systems that learn from the user's behavior and react accordingly): (1) predictability and transparency, (2) controllability, (3) unobtrusiveness, (4) privacy, and (5) breadth of experience. Friedman and Kahn (Friedman and Kahn, 2003) suggest 12 key human values with ethical import: (1) human welfare, (2) ownership and property, (3)

freedom from bias, (4) privacy, (5) universal usability, (6) trust, (7) autonomy, (8) informed consent, (9) accountability, (10) identity, (11) calmness, and (12) environmental sustainability. Bellotti et. al. (2002) suggest five interaction challenges for designers and researchers of sensing systems: (1) address—"directing communication to a system," (2) attention—"establishing that the system is attending," (3) action—"defining what is to be done with the system," (4) alignment—"monitoring system response," and (5) accident—"avoiding or recovering from errors or misunderstandings." These challenges could also be used for evaluation of context-aware systems but these pertain only to interaction.

AXIS	**Challenge for Evaluation**
Universality *Definition*: For who and in what domain *Metrics:* Personal Info, training	New class of users and new use cases where workload, metrics, stress points are not identified. No baseline information available.
Utility *Definition:* Benefit to users *Metrics:* Correctness of system inferences	Benefits have to measure within context so systems need to be evaluated in use. Activities must be evaluated, not just interactions.
Usability *Definition:* effort/ utility per unit *Metrics:* configuration, predictability, distraction, mixed initiative, cost of reversing a decision	How can user intent be captured to determine appropriateness of system interactions?
Ubiquity Definition: Points and times of delivery in physical world Metrics: graceful degradation, trust	Must evaluate within a larger set of degraded operating modes.

Figure 1: Evaluation Axes for Ubiquitous Computing

The three types of metrics for evaluation (scientific, component level, and impact) apply to ubiquitous computing as well. Consider a context-aware tour-guide system. This system recognizes where the user is, gives the user information on a particular attraction, and possibly recommends another attraction that the user would like to visit. The user would also be able to inquire about restaurants and stores in the vicinity. At the scientific level, metrics would address the recognition of a number of attractions and the ability to deliver the information corresponding to the attraction. Evaluation would also look at the ability to recognize and appropriately handle user inquiries. At the component level, we would measure the utility of the information delivered on attractions. We would also evaluate the accuracy and completeness of the information delivered with respect to user inquiries. At the impact level, a possible metric would be the satisfaction rating of the user of their visit to this particular city.

Conclusions

Evaluation of intelligent information systems is an important aspect of the research. These systems must show benefits and improvements to end-users for adoption to take place. Metrics must be developed to show this end-user benefit but must also help researchers focus their efforts. Metrics for intelligent information systems go beyond the traditional information retrieval metrics. We appeal to the research community to start to develop methodologies for evaluation. IR research has shown that having a common set of metrics can benefit the community. We suggest that the IA community could benefit from the development of a common set of metrics as well.

Acknowledgements

This work support in part by Advanced Research and Development Agency (ARDA).

References

Bellotti, V., Back, M., Edwards, W.K., Grinter, R.E., Henderson, A., Lopes, C., 2002. "Making Sense of Sensing Systems: Five Questions for Designers and Researchers," *Proceedings of the Conference on Human Factors and Computing Systems*, pp. 415-422.

Bonneau-Maynard, H., Devillers, L., and Rosset, S. 2000, Predictive performance of dialog systems. In Int. Conf. on Language Resources and Evaluation, LREC 2000.

Bylund, M. and Espinoza, F. 2002. "Testing and demonstrating context-aware services with Quake III Arena," *Communications of the ACM*, ACM Press, New York, pp. 46-8.

Carroll, J. M. and Olsen, J. (Eds). 1987. *Mental models in human-computer interaction: Research issues about what the user knows.* Washington, DC: National Academy Press.

Friedman, B., Kahn, Jr., P.H.. 2003. "Human Values, Ethics, and Design," *The Human-Computer Interaction Handbook*, Lawrence Erlbaum Associates, New Jersey pp. 1177-1201.

Jameson, A. 2003. Adaptive Interfaces and Agents. *The Human-Computer Interaction Handbook.* Lawrence Erlbaum Associates, New Jersey, 316-318.

Hirschman, L. 2000. Evaluating spoken language interaction: Experiences from the darpa spoken language program 1990-1995. In S. Luperfoy (Ed.) *Spoken Language Discourse*. MIT Press, Cambridge, MA..

Hirschman, L, Bates, M., Dahl, D., Fisher, W,., Garofolo, J., Pallett, D., Hunicke-Smith, K., Price, P, Rudnicky, A., and Tzouketmann, E. 1993. Multi-site data collection and evaluation in spoken language understanding. In *Proceedings of the Human Language Technology workshop*, pages 19-24.

Norman, D. A. 1988. The psychology of everyday things. New York: Harper and Row.

Price, P., Hirschman, L., Shriberg, E., and Wade, E. 1992. Subject-based evaluation measures for interactive spoken language systems. In *Proceedings of the DARPA Speech and NL Workshop*, pages 34-39.

Scholtz, J. and Richter, H. 2002. Report from Ubicomp 2001 Workshop: Evaluation Methodologies for Ubiquitous Computing. SIGCHI Bulletin. January/February.

Smailagic, A., Siewiorek, D., Anhalt, J., Gemperle, F., Salber, D., Weber, S., Beck, J., and Jennings, J. 2001. Towards Context Aware Computing: Experiences and Lessons Learned, IEEE Journal on Intelligent Systems, Vol. 16, No. 3, June.

Walker, M., Aberdeen, J., Boland, J., Bratt, E., Garofolo, J., Hirschman, L., Le, A., Lee, S., Narayanan, S., Papineni, K., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., . Rudnicky , A., Sanders, G., Seneff, S., Stallard, D., and Whittaker, S. 2001. DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection, Proceedings of EUROSPEECH 2001.

Whittaker, S., Terveen, L., and Nardi, B. 2000. Let's Stop Pushing the Envelope and Start Addressing it: A Reference Task Agenda for CHI. Human Computer Interaction, vol. 15 (2&3), 75-106.

Wixon, D. and Ramey, J. (Eds). 1996. Field methods casebook for software design. New York: Wiley.