UALBANY
State University of New York

Institute of Informatics
Logics and Security Studies

Achieving Improved Retrieval Performance Assessment for an Experimental Question-Answering System

Ben-Ami Lipetz

June 2007

ILS Tech Report: 001

Available for download from:
`http:www.ils.albany.edu`

ILS Tech Report: 001 (4/9/07)

ACHIEVING IMPROVED RETRIEVAL PERFORMANCE ASSESSMENT FOR AN EXPERIMENTAL QUESTION-ANSWERING SYSTEM

Ben-Ami Lipetz
University at Albany, USA

Abstract

Assessing the performance of a complex computer-based information retrieval system is desirable but not simple. Measurements that indicate quality of retrieval, although not the only consideration in assessing system performance, are central. However, measurements of precision and recall are difficult to validate because they involve the subjective concept of relevance, and may also be difficult to obtain because they involve labor-intensive procedures for judging relevance that are outside of normal system use. Performance evaluation for an interactive question-answering system presents further difficulties because the user-driven retrieval criteria do not remain fixed throughout a search session. An approach is described that is being developed and implemented for assessing retrieval performance of the HITIQA experimental question-answering system which is configured to facilitate the on-line generation of a search report as an integral part of a search session. Each item in the system's database has a unique identification code, which is carried into a session record when that item is cited in the session report. Extensive data are collected automatically by the HITIQA system; reports produced by searchers are added to the data collected automatically. All of this can be used to generate a variety of measurements reflecting retrieval performance from different points of view regarding relevance. In one retrieval assessment approach, items used in searchers' reports are considered relevant by definition, making the calculation, comparison, or tracking of searchers' retrieval performance with respect to reporting more-or-less objective and programmable. Alternative programmable definitions of relevance may be used to assess aspects of retrieval performance from different viewpoints; measurements may be chosen to focus on performance of searchers, or on evaluation of the retrieval system or its specific components, or on performance with different user access provisions, different search question topics and formats, or comparisons of retrieval performance when different sources are used for stocking the system's database..

1. Introduction

The rational development and improvement of any system intended to select and provide useful information from assemblages of possibly-useful records requires above all a capability to assess reliably, and preferably in a quantitative manner, just how well the system performs its retrieval functions. There are usually other important aspects of an information system to consider in its evaluation and development, such as its capacity, costs, coverage, accessibility, and user acceptance; but none of these may matter unless the retrieval performance is adequate or superior.

While assessment or measurement of retrieval performance is essential for system development, it is also highly problematic. This applies to any information retrieval system, and can be especially troublesome for a large and complex computerized system, such as a question-answering system, as will be explained below.

2. Measures of Information Retrieval

The basic functions that an information retrieval system should perform are two-fold: 1) to retrieve (i.e., identify, select, copy, tag, move, and/or deliver) from an available pool or stream of records (e.g., from a database) records that match a specification from someone who is attempting to use the system, and 2) to avoid retrieving records that do not match the user's specification. Desirable measures that would, in theory, characterize the retrieval performance of a system with respect to these basic functions are:

1. Selection. Number of items retrieved (however that may be defined).

2. Rejection: Number of items withheld from retrieval.

3. Precision: Proportion of retrieved items that are relevant by the user's specification.

4. Recall: Proportion of relevant items within the system's available records pool that are retrieved.

The first measure in this list can usually be obtained easily and objectively by simply counting the units retrieved as they are retrieved. The second is measured by subtracting the first from either the presumably known size of the system's records pool or from a count of records streamed or considered in any other manner in the process of performing retrieval.

In contrast to the first two measurements , the final two listed above, precision and recall, are deceptive and problematic. They are attractive to think of as purely objective concepts, but are not. Difficulties with precision and recall were recognized in the Cranfield College of Aeronautics studies (Cleverdon 1959, 1962, 1965, 1967, Cleverdon et al. 1966, Keen 1966, Rees 1965), widely regarded as the earliest attempts to conduct scientific evaluations and comparisons of information retrieval systems. Precision and recall both make use of the notion of "relevance," which is often poorly defined, or left undefined, by users of a system, and which is subject to different interpretations by different linguistic matching programs or by different non-users who may be called upon to judge relevance after a search or to predict it in anticipation of a search. To get fully exact and authoritative measurements of precision and recall, one would have to get system users to determine the relevance or not of every record in the system, whether retrieved or not, which is, of course, almost always not feasible. Indirect methods must be employed instead, such as extrapolating from small samplings of user judgment, or accepting as meaningful the results of limited tests in which some of the records contained in the system have been pre-judged by presumably authoritative persons to be relevant to artificially composed test questions that are to be put to the system (Cuadra 1967, Cuadra and Katter 1967a, 1967b, Lancaster 1968, Vickery 1965). Recall is especially difficult to measure accurately, since it requires determining the total number of relevant records available in the system. With any sizable system, direct record-by-record review is not possible. Only indirect and approximate measures can be employed, such as sampling and extrapolation, or projection of a ceiling from comparison of overlaps and underlaps when the same question is searched in the same system by multiple users or by using multiple search procedures.

There are yet more problems relating to relevance: When dealing with measurement of precision and recall, the concept of relevance is usually treated as an absolute, yes-or-no quality, when in fact system users very often regard relevance as quantifiable, with some relevant records being identifiable as significantly more relevant or less relevant than others. Also, the user's notion of what constitutes relevance may change throughout the course of a search session on some topic, as new understanding is gained and as new ideas take shape. This is a very important consideration in interactive question-answering systems, which are deliberately designed to facilitate changing of specifications as a search progresses. It is a factor in traditional retrieval systems as well, however, since, in reviewing candidates for retrieval, searchers can and do reject clearly pertinent items that are seen as redundant and no longer desirable—that is, they alter the specifications for relevance.

It should be acknowledged that, despite the difficulties stated here, there has been progress in deriving measurements of precision and recall that are helpful in evaluating large information systems. Of particular note is TREC (Text Retrieval Conference), which has since 1992 provided a vehicle whereby multiple researchers and organizations that construct well-reviewed computerized test collections of many kinds can make them available to each other for experimentation and comparisons (Voorhees 2005). In TREC, relevance is determined by non-ser judges; thus the difficulties stated above are accommodated and not overcome.

3. HITIQA

The HITIQA (High Quality Interactive Question Answering) system, currently under development at the University at Albany, is intended to assist intelligence analysts in searching for pertinent textual material on topics of concern, and is also intended to assist the searcher in assembling a report on the findings in a search. A descriptive introduction to HITIQA has been prepared by Strzalkowski, Small, Hardy, Yamron, Liu, Kantor, Ng, and Wacholder. HITIQA executes a variety of procedures to retrieve items from a database in which all of the included texts have previously been analyzed and characterized ("framed") in a manner that makes possible the rapid identification of texts that tend to match the characteristics called for in a question entered later by a searcher. The development and validation of this framing method are described by Small et al, and Ng et al. In matching a text item to a question, the number and types of matching characteristics are noted by the system, and the degree of match may be expressed quantitatively.

The texts in its database that HITIQA operates on are typically documents that have been selected from other available databases by conventional search methods in order to build a broad pool of material considered to be of potential relevance to a range of likely future topics of interest. When long documents are entered into this pool they are automatically segmented (usually by paragraphs) into smaller pieces that are analyzed and characterized as independent text items. Each document and each segment in the pool is given a unique identification number that accompanies it in every subsequent HITIQA procedure.

In a HITIQA search session, the searcher first enters a question, and the system responds by identifying matching text items and ranking them by closeness of match. It presents to the searcher at once the texts of the most highly ranked items, and makes the items with weaker matches (down to some pre-set threshold) known and available to the searcher through notification and linking devices that include annotated listings and also displays of icons representing text items in which the icons are color-coded to indicate degree of match and displayed as clusters according similarities in attributes of their frames. The searcher may decide to open and read, or not open, any item that has been offered but not yet displayed as text. The searcher may copy any text item that has been displayed into a work file, for review at a later time and especially when composing the search report.

The searcher may enter a new or modified question at any time, thus triggering a new response by HITIQA. In the application for which HITIQA is designed, there will typically be many questions in a search session, because search topics are generally either fuzzy or multifaceted. HITIQA is programmed to suggest new questions to the searcher as a session progresses; these arise from statistical determination that texts being read and retained by the searcher have common attributes that were not specifically requested and that could be invoked to retrieve still other somewhat matching items. The searcher may accept or reject a suggested question, or modify it. The searcher may switch back to an earlier question, to review what HITIQA offered or to resume his work on that set of offerings. Yet another HITIQA feature is that it allows the searcher to comment on the degree of matching (relevance) that it has determined for any text item offered in response to a question; the searcher, if he chooses, may register agreement or may assign a lower or higher rating to the item.

There have been a few trials to date in which actual intelligence analysts are asked to use HITIQA to search and prepare reports on various topics that are intended to be similar to some of their actual assignments; findings are still being compiled (Strzalkowski et al.). Many more such trials are in prospect. Future trials are expected to involve numerous individual searchers and numerous topics, a variety of arrangements for collaboration among multiple searchers working on a common topic, and may also introduce experimental variation of the ways in which HITIQA is programmed to characterize texts and to respond to questions.

4. Assessing HITIQA Retrieval Performance

HITIQA search sessions produce voluminous research data, almost all of it captured automatically. Texts, identification codes, and frames (characterizations) of all items in the searched database are retained. For each search session of each participating searcher, HITIQA retains a time/transaction record of each question or instruction and each system determination and retrieval offering. Electronic communications between collaborating searchers are captured, as well as inputs to the searchers' work files and the texts of reports on searches. Data are derived also from computer-administered questionnaires that are presented at the end of each search session and each trial series. When several searchers have reported on the same topic, they are asked to read each other's reports and rate them for over-all quality and for particular aspects of report content. Acoustic communications during searches (as in same-room collaborations) are recorded, but are not currently automatically analyzable. Recording of searcher behavior other than computer use through assignment of observers to record such activity has been tried and may be included again in future trial arrangements when searchers' off-line actions are considered important to study.

The approach being taken at HITIQA to develop effective methods for retrieval performance assessment is to both acknowledge and to mitigate the difficulties (discussed in Section 1) that are inherent in achieving accurate and affordable determinations of relevance—the necessary precursor to calculating both precision and recall in information retrieval. More specifically, the approach taken avoids relying on non-searchers to judge relevance, instead deriving the desired performance measures through objective, probably programmable, analysis of the rich data that HITIQA collects automatically from search sessions and trial series. Since relevance is inherently a subjective quality, the approach does not arrogate a single "correct" standard of relevance, but rather allows relevance to be defined for automatic determination in multiple ways that may each be valid to different people for different evaluative purposes.

To illustrate how retrieval performance measurements may be derived that use different criteria for relevance, some sample data has been extracted from HITIQA trials to prepare the tables and figures presented here. Table 1 shows data from four searches on a single topic (two by pairs of collaborators, and two by single searchers). They are given in descending order of the peer-rating of the reports resulting from these searches. The aim here is to cast light on comparative performance of searchers and search arrangements (collaborative vs. single searches). Here, any text item cited in a search report is considered relevant by definition. Thus, all reports have 100% precision. Recall is calculated using the total number of different text items cited by all reports as the total number of relevant items available; while this is obviously (probably) not true, the resulting figures may still be useful for comparative evaluations. In Figure 1, the recall measurements are plotted against peer ratings. The result is not what one would intuitively expect (i.e., a positive relationship of recall and quality rating), but the numbers here are extremely small and conclusions are not justified, but might be with larger samples.

In Table 2, data on the same trial (four reports from six searchers) are presented more fully and organized differently. The aim here is to cast light on the productivity of the HITIQA search process as it relates to length of search sessions. The table shows, cumulatively, how many items were offered to all searchers by HITIQA over five different periods of elapsed search time, and shows how many of these items were copied to the work files of the searchers during these periods. If we now define as relevant all items that were copied to a work file and define the total of such items as the total of relevant items for measurement purposes, we can determine both precision and recall for each time period. A plot of precision and recall, Figure 2, shows, after 60 minutes of elapsed search time, the classic pattern of inverse precision/recall relationship with increasing search effort; this may be helpful in determining optimal search-session length.

Admittedly, in both examples above, the values used for total number of relevant items are lower than the "true" values that would be expected if the searchers had looked at every item in the database. But these "true" numbers can be derived programmably from the available data (especially with larger samples) in at least two ways by using extrapolative reasoning: One way is to determine from actual retrieval figures for searchers how much the recall total would probably increase as more searchers were added—up to some given number of searchers or some given rate of diminishing return. A second way would be to determine from actual retrieval figures for elapsed search times how much the total recall would increase as more search time was added—up to some given amount of time or some given rate of diminishing return.

A recent paper by Diane Kelly et al. surveys various ways in which the performance of a question-answering system might be assessed. The authors are skeptical of the practicality of using captured transaction data, in part because of anticipated high labor costs. It remains to be seen whether the approach being taken at HITIQA will allay this skepticism. In the same publication, the authors conclude that a "user-centered" approach is the best approach, and express high enthusiasm for cross evaluation of search products by peer searchers. As was noted above in Section 3, peer evaluations are conducted routinely in HITIQA trials and their results are part of the transaction record. It is planned to exploit that data in performance studies (see, e.g., Table 1 and Figure 1) and to learn whether peer evaluations can be reliably predicted to correlate with more mechanical, less subjective measures of performance.

5. Discussion

The purpose of this paper is solely to show that, in a question-answering system in which there is routine capture of data on system transactions and user decisions, it is possible to devise plausible and, in principle, easily attainable measures of retrieval performance that do not involve use of non-participant judges. There is no intention to imply that the use of such judges is therefore unnecessary or unjustified in all circumstances or for all reasonable purposes when assessing retrieval performance. However, as work on the development of the approach described here progresses, there is expectation that it will gain substantial acceptance because of its flexibility and programmability, and suitability for standardization across multiple systems.

6. Acknowledgments

Prof. Tomasz Strzalkowski has my thanks for inviting me to participate in the HITIQA adventure and to help devise performance assessment methods. I thank Nick Webb for encouraging me to shape this paper as a clarification of basic issues in retrieval assessment. I am grateful to Hilda Hardy, Samira Shaikh, and Sharon Small for acquainting me with the intricacies of HITIQA, and especially to Samira Shaikh for extracting the data and preparing the tables and figures used here for illustration of assessment approaches.

7. References

Cleverdon, C.W. (1959) The evaluation of systems used in information retrieval. Proceedings of the International Conference on Scientific Information. Vol. 1. pp.687-698. National Academy of Science—National Research Council, Washington.

Cleverdon, C.W. (1962) Report on Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. College of Aeronautics, Cranfield, UK.

Cleverdon, C.W. (1965) The Cranfield hypotheses. Library Quarterly **36**(2):121-124.

Cleverdon, C.W., et al. (1966) Factors Determining the Performance of Index Languages. 3 Vol. College of Aeronautics, Cranfield, UK.

Cleverdon, C.W. (1967) The Cranfield tests on index language devices. Aslib Proceedings **19**(6):173-194.

Cuadra, C.A. (1967) Toward a scientific approach to relevance judgments. Proceedings of the 33rd Conference of FID and International Congress on Documentation.

Cuadra, C.A. and Katter, R.V. (1967a) Opening the black box of 'relevance.' Journal of Documentation **23**:291-303.

Cuadra, C.A. and Katter, R.V. (1967b) The relevance of relevance assessments. Proceedings of the American Documentation Institute **4**:95-99.

Keen, E.M. (1966) Measures and Averaging Methods Used in Performance Testing of Indexing Systems. ASLIB Cranfield Research Project, Cranfield, UK.

Kelly, D., Kantor, P., Morse, E., Schultz, J., Sun, Y. User-centered evaluation of interactive question answering systems (2006) In: Interactive Question Answering: Proceedings of the Workshop, 8-9 June 2006, New York. Association for Computational Linguistics.

Lancaster, F.W. (1968) Information Retrieval Systems: Characteristics, Testing, and Evaluation, pp. 120-128. Wiley.

Ng, K.B., Kantor, P., Strzalkowski, T., et al. (2006) Automated judgment of document qualities. Journal of the American Society for Information Science and Technology **57**(9):1155-1164.

Rees, A.M. (1965) The evaluation of retrieval systems. In: A.W. Elias, editor, Technical Information Center Administration (TICA 2). pp. 129-144. Spartan-Macmillan. Reprinted in: A.W. Elias, editor, Key Papers in Information Science, 1971, pp.154-169.

Rees, A.M. (1966) The relevance of relevance to the testing and evaluation of document retrieval systems. Aslib Proceedings **18**(11):316-324.

Rees, A.M. and Saracevic, T. (1966) The measurability of relevance. Proceedings of the American Documentation Institute **3**: 225-234.

Rees, A.M. and Schultz, D.G. (1967) A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching: Final Report. 2 vol. Case Western Reserve University, Cleveland, USA.

Small, S., Strzalkowski, T., Liu, T., Ryan, S., Salkin, R., Shimizu, N., Kantor, P., Kelly, D., Rirrman, R., Wacholder, N. (2004) HITIQA: Towards analytical question answering. In: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, August 2004.

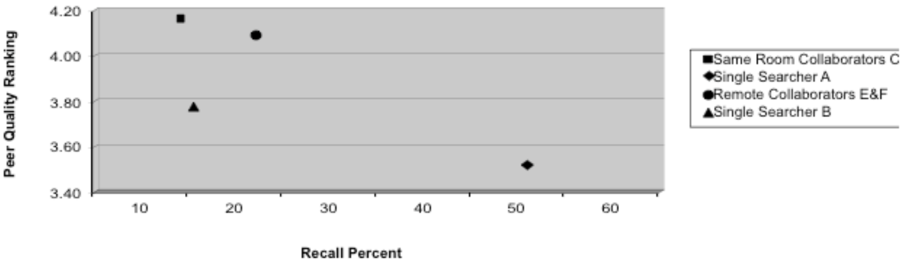
Small, S., Strzalkowski, Liu, T., Shimizu, N., Yamrom, B. (2004) A data drivenapproach to interactive question answering. In: M. Maybury, ed. New Directions in Question Answering. MIT Press.

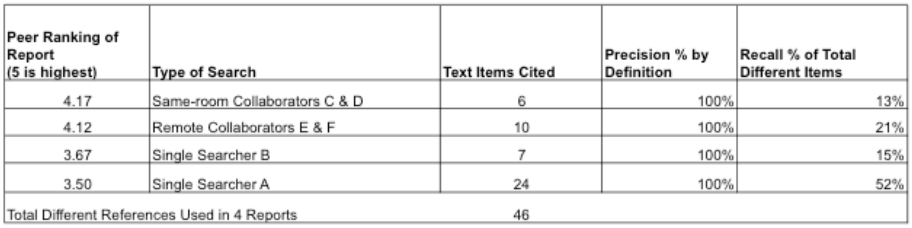
Strzalkowsky, T., Small, S. Taylor, S., Lipetz, B., Hardy, H., Webb, N., et al. (2007) Analytic Workshop with HITIQA Conducted at SUNY Albany on June 13-15, 2006. Preliminary Report. University at Albany, New York.

Vickery, B.C. (1965) On Retrieval System Theory. 2nd ed., pp.165-177.

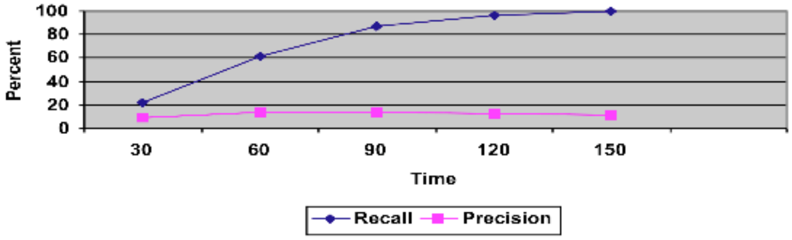
Voorhees, E.M. (2005) TREC: improving information access through evaluation. Bulletin of the American Society for Information Science and Technology **32**(1):16-21.

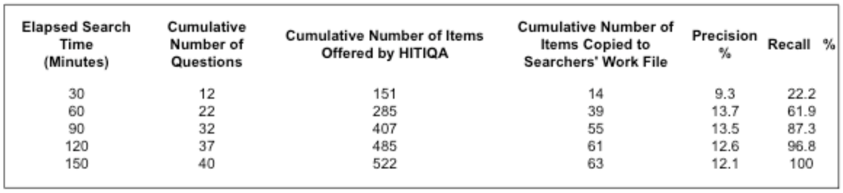
Figure 1: Comparison of Four Reports by Six Searchers on Topic X





Peer Ranking of Report (5 is highest)	Type of Search	Text Items Cited	Precision % by Definition	Recall % of Total Different Items
4.17	Same-room Collaborators C & D	6	100%	13%
4.12	Remote Collaborators E & F	10	100%	21%
3.67	Single Searcher B	7	100%	15%
3.50	Single Searcher A	24	100%	52%
Total Different References Used in 4 Reports		46		


100
80
60
40
20
0
Percent
30
60
90
120
150
Time
Recall
Precision



Elapsed Search Time (Minutes)	Cumulative Number of Questions	Cumulative Number of Items Offered by HITIQA	Cumulative Number of Items Copied to Searchers' Work File	Precision %	Recall %
30	12	151	14	9.3	22.2
60	22	285	39	13.7	61.9
90	32	407	55	13.5	87.3
120	37	485	61	12.6	96.8
150	40	522	63	12.1	100