

Measuring Web Search Effectiveness: Rutgers at Interactive TREC

Nicholas J. Belkin, Gheorghe Muresan
School of Communication, Information & Library Studies
Rutgers University
New Brunswick, NJ 08901-1071
nick@belkin.rutgers.edu, muresan@scils.rutgers.edu

Abstract

The purpose of this paper is multi-fold. We discuss methodologies and measures of effectiveness that, in our experience, mainly in the TREC Interactive track, have proven successful in painting an accurate picture of the user interaction when seeking information on the Web. We classify the measures and discuss the contexts when they can be used. We attempt to provide guidelines as to which measures are appropriate in certain conditions.

Measures of effectiveness

Evaluation methods were not born with or created solely for Information Retrieval. The field of IR borrowed methodologies of scientific investigation from other areas of research, including experimental design, statistical analysis and a range of qualitative and quantitative measures, and also created its own, targeted at the specific tasks that it investigates.

Effectiveness of a retrieval system is an ecosystem measure, related to the user's satisfaction with the system output and the interaction outcome (Korfhage, 1997). Modern IR systems are complex systems expected to support the user in at least three ways (Figure 1):

- it should support the user's exploration of the problem domain, the clarification and refinement of an information need, the source selection, and the articulation of the information need in the appropriate input language or syntax;
- it should use algorithms that retrieve as many relevant documents and as few non-relevant documents as possible;
- it should support the user's exploration of the retrieved documents, the extraction of relevant information, and the integration with the task at hand.

It is apparent that evaluating such a complex process is an incredibly challenging enterprise. One possible approach is to evaluate the overall outcome of the information-seeking process, such as the completeness of the user's task, the amount of information gathered or the incidental learning, and to ignore the contribution of each step to the overall success. Another approach is to try and isolate each process that takes place and evaluate its contribution to the overall outcome. Each approach has its own challenges: for example, even if we follow a systemic view and concentrate solely on the search engine, measuring its effectiveness is non-trivial, mainly due to the elusive nature of the concept of relevance. A document may be perceived as relevant when it offers a precise answer to a question, or when it gives a partial answer, or when it suggest a source for more information, or when it gives background information, or when it reminds the user of other knowledge; its relevance may depend on the user's situation, background or domain knowledge, and even on whether the user has seen a document with similar content recently.

While some researchers attempt to investigate the various aspects of relevance (Saracevic, 1975; Mizzaro, 1998), a common simplification is to equate relevance with the degree (which can be a binary or n-ary measure) to which the document answers a query. Related, but more subjective and more rarely used, are the measures of pertinence (how well the document responds to a user information need) and usefulness (how useful the document is to a user).

In the context of this workshop, we will focus on measures appropriate for investigating effectiveness of interactive Information Retrieval on the Web.

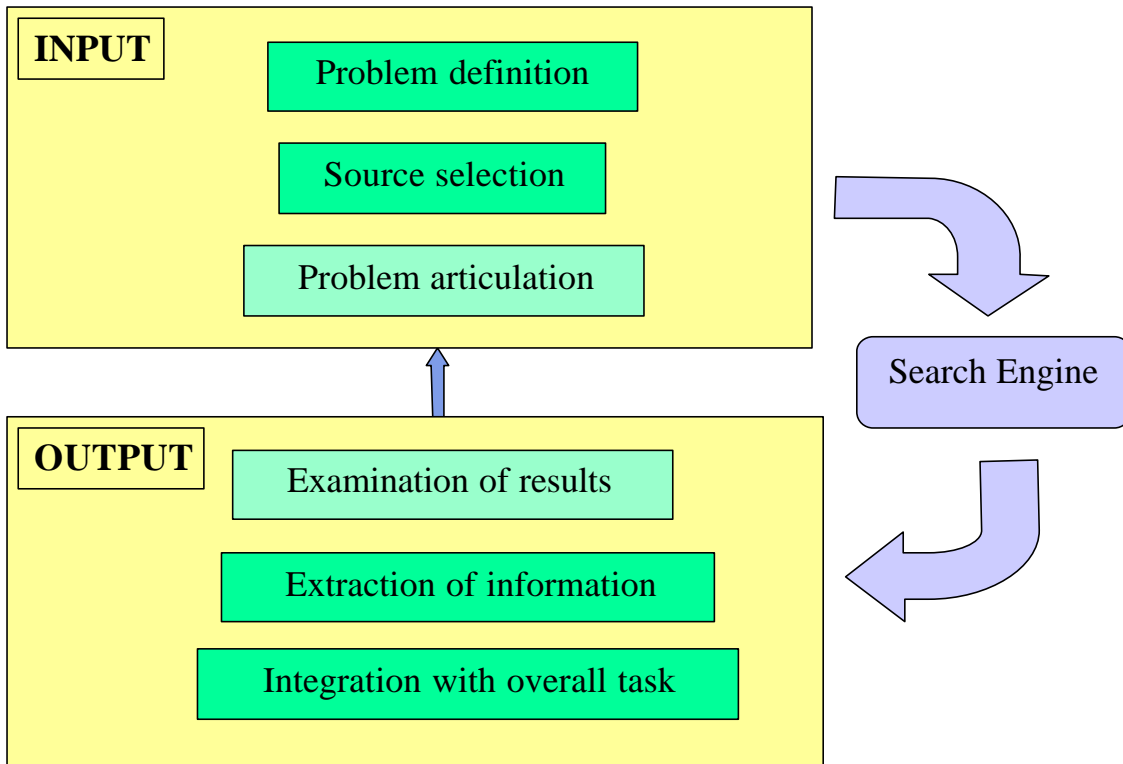


Figure 1. A view of interactive information retrieval

Quantitative vs. qualitative measures

The strong IR tradition of lab experiments in controlled conditions, started by the Cranfield experiments and best exemplified by TREC¹ ensures comparability between systems, components, algorithms or parameters: by varying just one element and keeping the others constant, the effect of that element on the final outcome (retrieval effectiveness) can be measured. In a rich experimental environment, with a large number of documents, test topics and relevance judgments, where sufficient data is available for statistical significance, several elements may be varied at one time, and a statistical analysis will be able to detect the effect of the independent variables on effectiveness, as well as possible interaction effects.

Historically, Information Retrieval Systems (IRS) have evolved from algorithms that applied some mathematical model and returned, in batch mode, a set or list of documents for each query submitted, into highly interactive systems designed to support the human user throughout the information-seeking process. Consequently, evaluation methodologies and measures have been adopted from the Human Factors and Human Computer Interaction communities and adapted to indicate how these systems support the exploration of a certain domain, the accurate formulation of an information need, the exploration of the retrieved documents, the completion of some information-related task etc.

The move from a pure systemic approach to IR and IR evaluation towards a more cognitive, user-centered approach has brought about the adoption of more qualitative methods of exploration and measures. Indeed, the usability guru Jakob Nielsen (2004) strongly rejects the use of quantitative measures in usability studies: “Number fetishism leads usability studies astray by focusing on statistical analyses that are often false, biased,

¹ <http://trec.nist.gov>

misleading, or overly narrow. Better to emphasize insights and qualitative research". On the one hand we take his view that heuristic evaluation, expert reviews, cognitive walkthroughs and other such usability methods are preferred to quantitative methods in terms of cost and accuracy if the purpose of the study is to establish the usability of a system; also, ethnographic studies are probably more accurate than questionnaires if the purpose of the study is to capture the behavior or preferences of a group of people in a certain setting. The combinatorial explosion of taking into account all possible influences would make a quantitative experiment too complex and expensive; ignoring some influences would open the door for bias and inaccuracy, as Nielsen complains. On the other hand, once the usability problems have been ironed out and the user task is well understood, a systematic quantitative study can produce invaluable insight into the effect of various parameters, mathematical models, interaction models, or even of interface elements such as the query formulation mechanism or the layout of the search results.

There is probably a consensus in the Interactive IR research community that the two approaches to investigating IR systems, ethnographic and laboratory-bound, and the two types of measures, qualitative and quantitative, are complementary and should both be used in order to give an accurate picture of the interaction. In fact, the experience in the TREC Interactive track led the participants to move to a two-year cycle²: in the first year the participants would adopt a rather generic and vaguely-defined task and observe human subjects' attitudes, preferences, and interpretations of the task. The outcome would be a better understanding of human behavior in a certain environment, with a certain type of task, and a set of research hypothesis. In the second year of the cycle, a careful experimental design and a clearly defined task in a controlled environment would allow the systematic investigation of these research hypotheses³.

It is important to distinguish, when discussing about laboratory testing between

- Simulations – the behavior of the ideal or typical user is predicted/simulated, with some amount of variability, and the experiment can be repeated, with variations, any number of times;
- User studies – due to the effect of learning, but also of fatigue and loss of concentration, repetition is difficult and expensive, and unlikely to produce identical results.

Dimensions of evaluation

In IR, the original evaluation measures were intended to be appropriate for evaluating the performance of IR systems in what was considered to be the generic IR task: the effective retrieval of documents relevant to a topic. Effective was operationalized as being able to retrieve all of the relevant documents in a collection (measured by *recall*), and only the relevant documents in a collection (measured by *precision*). Furthermore, this evaluation was intended to be performed in a non-interactive environment, without human intervention, since the goal was to test and understand the effectiveness of the representation and comparison processes of the IR system.

More recently, IR research has recognized that there are IR tasks that are more specific (or different from) the generic task described above. Many of these have been reflected in the different tracks of the TREC series. For instance, the filtering task is different from the generic task in that it is a Boolean decision that needs to be evaluated by a utility measure; the question-answering task requires not a list or set of relevant documents, but rather an answer to a question; the aspect retrieval task requires not all the relevant documents in a collection, but only a set of documents which, taken together, discuss the various aspects of a topic (see, e.g. Voorhees & Harman, 2000). Other IR tasks are easy to imagine.

We suggest that it is useful to consider measures of IR performance according to the two facets implied by this history: Degree of Interaction; and, Task Specificity. Figure 2 displays how the two values of each facet (Non-interactive/Interactive; Generic/Task-Specific) are associated with particular conditions, and with evaluation measures which have been used in each of these conditions. With respect to Figure 2, we believe that evaluation of the performance of Web searching should concentrate in the bottom two quadrants; that is, major features of Web searching are that it is interactive, and that people come to it with a variety of tasks.

² SIGIR2000 workshop on Interactive TREC (http://www.acm.org/sigir/forum/S2000/Interactive_report.pdf)

³ Unfortunately, the experimental model of the Interactive track did not fit the general TREC framework; the track was axed by NIST after the completion of the first two -year cycle.

Our experiences in the TREC Interactive Track have given us some insight into methods and measures for evaluating IR in contexts that we believe are relevant to the Web IR evaluation situation. Below, we briefly describe the experience of the TREC Interactive Track, and discuss the different tasks that were investigated in that Track, the measures that were used to evaluate performance and process, and the advantages and disadvantages of these measures.

Task Specificity	General	Task-specific
Interactivity		
Non-interactive (laboratory evaluation of the retrieval algorithm)	Recall, Precision, E, F Expected search length	Question answering : mean reciprocal rank of the correct answer Filtering : utility Topic distillation : Coverage and Accuracy
Interactive (evaluation of the interaction process and outcome)	User satisfaction User effort (clicks, iterations, scrolling, documents seen, viewed or read) Efficiency: Time to complete task, Precision at N seen Expected search length	Aspect retrieval : Aspectual recall, number of saved documents Question answering : completeness and correctness of answer Topic distillation : Coverage and Accuracy

Figure 2. A categorization of effectiveness measures

The TREC Interactive Track

The TREC Interactive Track was initiated already in TREC-3 (Harman, 1995). Its goal was to study interactive IR as a process, as well as to investigate methods of improving IR systems in their support for interactive IR. The track eventually settled on a general methodology for accomplishing these goals. This methodology was, in common with TREC in general, experimental, with each participating site comparing various aspects of subjects' performance and behavior in two (or more) different IR systems specific to that site. In order to maintain some measure of comparability amongst the work at the different sites, the standard TREC model was used, of having a common task set for all participants, with a common set of topics to be searched for, in a test collection consisting of a frozen database with external judgments as to the relevance (or other evaluative criterion) or not of a subset of the documents to the topics. There were three major differences between the Interactive Track and the other TREC tracks: one was that, after some initial attempts to do cross-site comparison, this was given up as impractical and so the emphasis was placed on what each site learned about its specific system(s), rather than on comparing performance between different sites. Another was that new tasks, and, correspondingly, new measures, had to be developed in order to do the sorts of evaluation that were desired. And, quite different sorts of data had to be collected and analyzed than in the other tracks, which were concerned with evaluation of IR in non-interactive situations.

Until 2001, the Interactive Track used for its test collections the TREC databases of newswire and newspaper articles and associated topics, generally topics which were being constructed for other tracks (usually modified to fit the specific tasks of the Interactive Track). In TREC-9, this procedure was slightly modified, in that the Interactive Track made up its own topics, because ordinary TREC topics were not well suited to the question-answering task of the Track for that year.

As of TREC-10, the Interactive Track moved from using one of the typical TREC databases to using the Web as the database to be searched. For TREC-10, there were some topics set by the Track organizers, but there was no requirement to conduct an experiment. Rather, each of the sites was encouraged to do some exploratory

observational research, in order to develop hypotheses which could be tested in the next year's Interactive Track. Thus, most sites that participated had subjects who searched on one or more publicly available Web search engines, recorded their searches in various ways, and then analyzed the data in order to identify some issue(s) to address in TREC 2002. Then, TREC 2002 was conducted in the by now classic experimental mode, but using a database that was a crawl of the .gov domain, and a specific task and related topics for the Interactive Track. TREC 2002 was the last year that the Interactive Track existed as a separate entity; in TREC 2003, there was an Interactive sub-track of the Web track, which used the same database as for TREC 2002, but with a different task, which was related to one of the tasks of the Web Track, the so-called "topic distillation" task. This was again carried out in experimental mode.

Uses, advantages, disadvantages of evaluation measures and procedures in the TREC Interactive Track

The first TREC Interactive Track adopted the TREC routing task. The goal of this task was for a system, or a person, to construct the best possible query for a set of topics, given information about the relevance (or not) of documents in a training database to those topics. Performance was measured according to recall and precision results for these queries, when they were applied to a new, test database. At Rutgers, this task was performed by experienced, professional searchers, who we thought were familiar with this general type of task. Relevance, or not, in both training and test collections, was determined not by the searchers, but by the relevance assessors at NIST. Thus, in the training condition, what the subjects had to learn was how the relevance assessors had evaluated relevance, and apply this knowledge to the construction of their queries. This caused some problems, due to the well-known problem of disagreement in relevance assessment among different judges. Also, the searchers tended to treat the task as roughly equivalent to the TREC ad hoc task, which was to generate a query that would find as many relevant documents as possible in the first 1000, with as high a density of relevant documents at the top of the list as possible. Thus, the actions that the searchers engaged in were not what they would have done in an equivalent real-life situation, and the actual task that they performed was not what it was intended to be.

Because of the difficulties encountered in TREC 3, the TREC-4 Interactive Track changed to the ad hoc task, again using recall and precision based on the evaluations of the NIST assessors. This task required the searchers to generate a list of documents retrieved with respect to a query, which contained as many relevant documents as possible, within a constrained time. This was thought to be a reasonable interactive IR task, with appropriate measures to evaluate performance. However, evaluation of performance in this task according to standard TREC procedure turned out to be not a measure of performance, but rather a measure of agreement in relevance assessment between the searchers and the external evaluators. Furthermore, the task itself turned out not to be as ecologically valid as was initially thought, since the evaluation was based on the topic 1000 documents retrieved by the searchers' final queries. None of the searchers could actually comprehend how their query might work in this way, and instead concentrated on making the top of the list contain as many of what they thought were relevant documents as possible. It was also thought by the various participants in the Interactive Track that the general TREC ad hoc task was really not one that most people searching for information would actually want to accomplish.

Because of the general nature of TREC, in the TREC 5 Interactive Track there was an attempt to develop a method which would allow comparison in performance between participating sites. For this purpose, there was no change in the task or measure. The result of this effort, repeated in TREC 6, was to give up on the idea of cross-site comparison for interactive IR, primarily because of pragmatically uncontrollable variations in subject-topic-system interactions between sites.

The problems in TREC-4 with the task and its evaluation led to the Interactive Track identifying a completely new task for TREC-6 the so-called *aspectual retrieval* task. This task was thought to be more ecologically valid than the TREC routing or ad hoc tasks, with respect to people searching for information. The goal of this task was for searchers to identify the different aspects or instances of a given topic, as exemplified by documents in the database. For instance, if topic were "Methods of reducing high blood pressure", the task would be to identify as many different methods as possible, by finding documents which discussed one or more of the methods. Now, the task was not to find and save all of the documents which discussed methods, but rather only documents which discussed different methods. For this task, a new measure, *aspectual recall*, was constructed.

This measure was based on judgments by the NIST assessors of what the different aspects of each of the topics were, as determined from the pooled set of documents saved by all of the searchers at each participating site for each topic, and to specify which documents discussed which aspects. Then, aspectual recall was the number of aspects that were discussed by the documents saved by the searcher for a given topic, divided by the total number of aspects for that topic. Searchers were instructed that they should not save more than one document per aspect, although they were not penalized for saving more. This task turned out to be relatively easy for the subjects in the experiments (in our case at Rutgers, graduate students in library and information science) to understand, but there were still problems with the evaluation measure. Once again, they centered on the issue of agreement, this time not in terms of relevance, but in terms of what constituted an aspect of a topic, and in particular, the granularity of aspects. The problem was not as severe as the general relevance agreement problem, but still led to problems in interpretation of the results.

TRECs 7 and 8 Interactive Tracks continued with the aspectual retrieval task, using aspectual recall as the evaluation measure, but concentrating ever more on measures of the process of the interaction. These measures included such factors as number of document titles viewed, number of documents actually opened, number of documents saved, number of queries issued during a search, time to task completion, subjective assessments by the subjects of various aspects of the interaction and of their task performance, and use of various system features. These factors, which could be grouped generally under performance (time to complete task, subjective assessment) or effort (number of queries, number of documents saved as a proportion of documents viewed) variables, could then be used in conjunction with the more “objective” performance measure, aspectual recall, to consider the difference in overall performance of a system in support of interactive IR. For instance, if the two systems being compared had no significant difference in aspectual recall, but in one subjects had to exert significantly less effort than in the other to achieve the same level of aspectual recall, then the former could be considered to be more effective. Another important result from TREC-8 was that there was found to be a significant correlation between aspectual recall (based on external assessments) and the number of documents saved per topic. This result suggested that, at least for this task, it might be possible to evaluate system performance in supporting the searcher in the task solely on what the searcher did, without external validation.

The TREC-9 Interactive Track shifted to a new task, complex question answering. This was motivated by the feeling that we could get no farther with the aspectual retrieval task, and that support for complex question answering in an interactive environment was an interesting and important task. The main TREC Q&A Track was at that time concerned with a very narrow definition of question answering: finding “factoid” answers to very limited factual questions, within a single document. The goal of the Interactive Track was to investigate support for answering questions that would require synthesis of information from several documents. In order to evaluate performance in this task, the same sorts of process and subjective measures were used as in previous TRECs, but new “objective” measures were developed as well. The intent of these measures was to determine whether the system had actually helped the searcher in finding the correct answer to the question, and whether the answer that the person found was complete (the latter was necessary, because many of the questions were of the sort exemplified by: “How many films did Orson Welles act in?”). These two measures were quite different from the measure used in the TREC Q&A Track, which was the mean reciprocal rank of the correct answer in the lists returned by the automatic system of answers. In the Interactive Track, the measures were tailored specifically to the interactive IR situation and the complex question answering task.

In TREC-10, the Interactive Track began its direct involvement with the Web as the database in which searching would take place. Although in general this was to be an exploratory year, at Rutgers we conducted experiments to investigate means for increasing query length, and the effect of query length on performance. There were also general goals of investigating whether there was a difference in searching behavior and performance between topics which were completely specified for searchers, and topics which searchers could tailor to their own interests, and of behaviors and performance in different topic “types”: medical, buying, projects, and travel. These types were based on “typical” reasons for searching identified in studies of Web searchers (Rieh, 2002). The individual topics were couched as complex questions, and the evaluation measures were again correctness and completeness, and the subjective and behavioral measures as before. One basic problem with TREC-10 was that it was carried out in the live Web, which meant that the subjects in the experiment were not all searching the same database, although they were all searching through the same search engine.

The TREC 2002 Interactive Track continued with the same task, the same topic types, and the same measures, but changed in an important way. Rather than searching in the live Web, the Interactive Track made use of the TREC Web Track's database which was constructed from a crawl of the .gov domain. This addressed directly the issue of comparability of results between subjects, and also allowed at least some sites to do their own indexing of the database, and use their own retrieval algorithms. Apart from that, the conditions were the same. TREC 2003 found investigation of interactive IR moved to a sub-track of the Web Track. One of the tasks of the automatic Web Track was so-called "topic distillation", in which systems were to identify a set of sites which were good sources of information on a topic, with little overlap in coverage of the topic between the sites. The Interactive Sub-Track adopted this task, and performance was evaluated by the measures of coverage and overlap. These two measures were based on assessments made by NIST evaluators of the completeness of coverage of the topic by the set of sites (on a 5-point scale from bad to good), and the degree of overlap between the set of sites, again on a 5-point scale. The subjective and process measures were largely equivalent to those in previous Interactive Tracks. The objective performance measures turned out to suffer quite badly from the already known problem of agreement between assessors and searchers, and also from the complexity of the task itself, and the assessors' knowledge of the topic in general.

Conclusions

What have we learned from all of this that's relevant to user-centered evaluation of Web IR? First and foremost, IR, and especially Web IR, should be considered as inherently interactive, and therefore it is insufficient to consider only measures of search effectiveness when evaluating IR system performance. Perceptions of performance by the user are at least as important as "objective" measures, and both need to be interpreted in terms of measures of the search process itself, especially measures of effort and efficiency. Second, the general concept of information seeking consists of many different kinds of activities, and, especially, many different kinds of tasks, associated with many different types of goals. It is therefore incorrect to believe that there will be one (or one set of) measure(s) that will be appropriate for the evaluation of IR systems in general. Rather, when considering the evaluation of a Web IR system, it is first necessary to establish precisely the IR task that the system is supposed to support, with some explicit relationship between the task and the goal that accomplishing that task enables. Doing this will suggest the appropriate performance measure(s) for that task. Third, it is not necessary, or even a good idea, to rely on external judgments for measuring the performance of an IR system. Instead, it can be a good idea, in a variety of ways, to design the specific task in such a way that the *user's* performance in accomplishing that task, and the user's own evaluation of her/his performance, can be used directly (or indirectly) as the criterion on which the measure is based. Finally, we suggest that experimental investigation has had great value in user-centered evaluation of interactive IR, and that it should be a priority to extend this general methodology to the evaluation of Web IR, as a complement to ethnographic studies.

References

- Harman, D. ed. (1995) TREC-3. The third Text REtrieval Conference. Washington, D.C.: GPO
- Korfhage, Robert R. (1997) *Information Storage and Retrieval*, Wiley.
- Mizzaro, Stefano (1998) *How Many Relevances in Information Retrieval ?*, *Interacting with Computers*, 10(3): 303-320.
- Nielsen, Jakob (2004) *Risks of Quantitative Studies*, Alertbox – March 1, 2004, useit.com (<http://www.useit.com/alertbox/20040301.html>).
- Rieh, Soo Young (2002) *Judgment of information quality and cognitive authority in the Web*, *JASIST* 53(2): 145-161.
- Saracevic, Tefko (1975) *Relevance: a review of and a framework for the thinking on the topic*, *Journal of the American Society for Information Science*, vol. 26: 321-343.
- Voorhees, E. M. & Harman, D.K. eds. (2000) TREC-9. The ninth Text REtrieval Conference. Washington, D.C.: GPO.

APPENDIX: Measures of retrieval effectiveness – a glossary

Human factors measures:

1. *Time to learn*
How long does it take for typical members of the community to learn relevant task?
2. *Speed of performance*
How long does it take to perform relevant benchmarks?
3. *Rate of errors by users*
How many and what kinds of errors are commonly made during typical applications?
4. *Retention over time*
Frequency of use and ease of learning help make for better user retention
5. *Subjective satisfaction*
Allow for user feedback via interviews, free-form comments and satisfaction scales.

IR-specific effectiveness measures

1. *Recall (R)* = the fraction of relevant documents that was retrieved = nb. of relevant document retrieved / total nb. of relevant documents in the collection.
 - a. Relative recall = nb. of relevant document retrieved / total nb. of known relevant documents in the collection.
 - b. Recall at n (n = 5, 10, 20, 30, ...) = recall computed after a cutoff of n documents is applied.
2. *Precision (P)* = the fraction of retrieved documents that is relevant = nb. of relevant document retrieved / total nb. of relevant documents retrieved.
 - a. *Precision at n* (n = 5, 10, 20, 30, ...) = precision computed after a cutoff of n documents is applied.
 - b. *Average precision at seen relevant documents* = the average of precision figures obtained after each new document is seen in the ranking.
 - c. *R-precision* = precision at the R-th position in the ranking, where R is the total number of relevant documents for the current query.
3. *F-measure (F_β)* = $(\beta^2 + 1) * P * R / (\beta^2 * P + R)$ is the weighted harmonic mean of Precision and Recall.
4. *Fallout (F)* = the proportion of non-relevant documents that are retrieved
5. *Expected search length⁴* = the average number of documents that must be examined to retrieve a given number of documents.
6. *Utility* = (roughly) the value of relevant document retrieved – the cost of relevant documents not being retrieved – the cost of retrieving non-relevant documents.
7. *Mean Reciprocal Rank (MRR)* for each individual query = the reciprocal of the rank at which the first correct response was returned, or 0 if none of the first responses contained a correct answer. The score for a sequence of queries is the mean of the individual query's reciprocal ranks.

User-oriented measures

1. *Coverage ratio* = the proportion of the relevant documents known to the user that are actually retrieved
2. *Novelty ratio* = the proportion of the relevant retrieved documents that were previously unknown to the user
3. *Relative recall* (user-oriented version) = the ratio of the relevant retrieved documents examined by the user to the number of documents the user would have liked to examine.
4. *Recall effort* = the ratio of the number of relevant documents desired to the number of documents examined by the user to find the number of relevant documents desired.

⁴ This measure could also be viewed as a user-oriented measure.