The Influence of Recall Feedback in Information Retrieval on User Satisfaction and User Behavior

Eduard Hoenkamp (hoenkamp@acm.org)
Henriette van Vugt (vanvugt@cogsci.kun.nl)
Nijmegen Institute for Cognition and Information; Montessorilaan 3
6525SW Nijmegen, the Netherlands

Abstract

The unprecedented scale-up of the World Wide Web, and the number of people relying on it for information, make it inevitable to reassess the validity of the traditional metrics for quality of information retrieval (IR). Of these, the most widely used metrics are recall and precision.
Users can judge the precision of an information retrieval system by inspecting the retrieved documents. They cannot judge recall, however, which would involve inspecting the whole collection, thus obviating the IR system, and impossible in the case of WWW. How then, can we ascertain whether recall is a valid metric for the quality of an IR system as perceived by the end-user? In a carefully controlled experiment we presented users with a simulated web search engine. Besides the search results, the engine could give a (spurious) recall estimate, presented as a pie chart. We manipulated this recall feedback, and whether the information need was fulfilled with respect to quantification type (the number of documents requested). It seems that fulfillment is a better predictor of user satisfaction and behavior than precision and recall as used to evaluate IR systems. The results reported here also suggest that whereas recall may be a valid metric for designers and evaluators of IR systems, it may lack validity as a metric for search quality as perceived by the end-user.

Introduction

Barely a decade ago, techniques for information retrieval were still in the able hands of librarians, in the case of print, and of data base managers in the case of electronically stored information. The explosive growth of the World Wide Web has changed this situation dramatically and irrevocably. Since then, people in all walks of life depend on automated 'librarians' as provided by search engines such as Google, AltaVista, Yahoo, and many others. Oviously, the end-user of such a system wants information that is relevant, and wants it returned within a reasonable time. The latter is a matter of efficiency, and that is where most of the research effort has gone. For example: how to increase bandwidth, how to index documents, how to encode multimedia. Not surprisingly, the aspect of efficiency is dominated by computer science, and solid metrics are known for these technical aspects.

Effectivenes, on the other hand, can only be gauged by the users of an IR system themselves. We claim that IR is a golden opportunity for cognitive science, with its roots in both psychology and computer science. For this, researchers can pursue two avenues: one is to exploit cognitive principles in modeling the user, the other by evaluating traditional metrics of IR concerning effectiveness through experimental design. The viability of the former approach we demonstrated elsewhere (Hoenkamp, Stegeman, & Schomaker, 1999; Hoenkamp & de Groot, 2000). In this paper we give an example of the latter.

From the the early days of the library sciences until today many metrics have been proposed to evaluate the quality of information retrieval systems (Swets, 1963; Cleverdon, Mills, & Keen, 1966). These metrics are to measure how satisfactory the material is that the system retrieves (the output), with respect to the user's information need presented as a query (the input). After several decades, *recall* (proportion of relevant documents actually retrieved) and *precision* (proportion of the retrieved documents that are relevant) have stabilized as *normative* measures for the quality and thus comparison of IR systems. The evaluation of these metrics has itself become a subject of study regarding both their reliability (Buckley & Voorhees, 2000) and their validity (Hersch, Turpin, Price, Chan, Kraemer, Sacherek, & Olson, 2000). Yet, however respectable and theoretically sound the metrics may be for comparing IR systems, it does not make them automatically apropriate to predict the satisfaction of the end-user with such a system. And given that IR systems are eventually built not for the evaluators but for the end-user, we wanted to investigate whether these metrics are also *valid* measures for quality from *the perspective of the user* conducting the search.

Users can only fare on the documents actually returned, and not on the uncounted documents never found. And as users can determine the relevant documents only among those returned, they can determine precision, but not recall. In addition, if users want to refine a search or provide feedback, again they can only do so on the basis of the documents returned. As precision is the only parameter the user can be aware of, it is the more likely parameter to determine the quality of a search as perceived by the user. So precision can be validated in principle, as one predictor of the user's satisfaction with an IR system. As the user cannot observe recall, there cannot be a corresponding validation for recall. This ends the symmetry between the two metrics that their definitions suggest. Any hope for exploring the relationship between recall and search quality, as perceived by the end-user, would require restoring that symmetry. This is exactly what we

set out to do. In a moment we will describe an experiment where we provided users with recall feedback, and measured the influence on their satisfaction with search results and search machine, and with their subsequent search behavior. Also, the usefulness of recall feedback is measured. It is important to understand that the recall feedback was represented by a slice on a pie chart. The size of the slice was manipulated, and had no relation whatsoever to actual recall.

It is useful to look first at our intuitions in order to appreciate what we learned through the experiment.

Intuitions

For the evaluator of IR systems, the intuitive trade-off between recall and precision is well-known: High recall can be achieved trivially by returning all documents, as this will include all relevant documents. Obviously, this goes at the expense of precision as many irrelevant documents are returned as well. Similarly, high precision can be achieved by stringent conditions on relevance, at the cost of missing potentially relevant documents. The end-user has also intuitions about recall (which we will capture below under hypotheses 2 and 7): When a search engine returns many relevant documents, the recall is perceived as high (especially when the precision is high). That is, the user thinks that the search engine retrieved a large proportion of the relevant documents. Consequently, the user is satisfied with such a search engine. Conversely, if very few documents, or none at all are returned the recall is seen as low, and the user is less satisfied with the search engine. Note, however, that the actual recall can be opposite to these intuitions. Especially when recall feedback violates these intuitions, this should influence the user's satisfaction with the search engine.

Focusing on the user's satisfaction with the search results, we intuit that it will depend on the degree to which the user's information need is met, and not on the mere number of returned relevant documents. For example: if the user wants to know whether the latest "Harry Potter" is out, just one document could meet this information need. If he wants to know which retailer on the web has the lowest price or the fastest delivery for the book, he needs as many sites as possible to choose from. Finally, if he needs the name of a bookseller nearby, a few documents may suffice to find one. Following Cooper (1968) we refer to these numbers as the *quantification type* of an information need, and call them quantification type 1, 2, and 3 (for one, all, or several documents). We expect the user to be most satisfied with the search result if the number of relevant documents returned matches the quantification type, at a high precision rate (this intuition leads to hypotheses 3 and 8). The satisfaction with the *search engine* we gather will depend on the type, the documents returned and, as the system is evaluated a whole, the *recall* (this leads to hypotheses 4 and 9).

These intuitions about the hypothetical relationship between the satisfaction and the compromise between recall and precision are visualized in figure 1. The figure shows the three quantification types. For example, for quantification type 3, the user would be unhappy with only one relevant document, satisfied with, say five documents, and again less satisfied when many more are returned as they will subsume more and more be irrelevant ones. The figure represents cases with, say, 200 relevant documents. The numbers on the x-axis are fictitious but are meant to indicate recall and precision. From left to right recall increases and precision decreases (recall and precision can easily be calculated, using the numbers in the figure). At the top of each curve the information need is fulfilled at the highest precision rate. The figure represents our prediction that no universally best compromise between recall and precision exists, as satisfaction will depend on the number of documents needed.

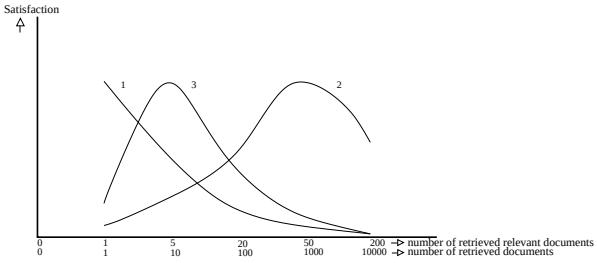

Figure 1: The compromise between recall and precision when *one* (1), *all* (2), or *several* (3) documents are needed. The curves represent the qualitative relationships hypothesized in this paper, between the satisfaction with retrieved documents and recall and precision. Note that the numbers along the x-axis are fictitious.

In this paper we assume that that user is looking for information, as opposed to entertainment. Hence we assume that the user's behavior, i.e. to continue or to stop searching, depends on whether his information need is fulfilled (which leads to hypotheses 5 and 10). Finally, we expect that also the usefulness of recall feedback depends on whether the the information need is fulfilled. More precisely, recall feedback will only be important if the users' need is not (yet) fulfilled. The usefulness may increase with increasing amount of documents searched for (these intuitions lead to hypotheses 6 and 11). we do not expect them to continue searching. Nor to

In line with the above, we expect strong interactions among the dependent variables. For example, when the information need of users is fulfilled, we expect them to be highly satisfied, to stop searching, and not to care much about recall feedback.

It will be clear that there are many potential interactions among the variables we introduced, hence the rather complex design. The hypotheses in the design section below are only a more detailed expression of the hypotheses we have just introduced informally.

Description of the experiment

Method

Participants The thirty two participants, fourteen female, were almost all acquaintances of one of the experimenters and volunteered to participate. Thirty of them had at least college education. They varied in age from 18 to 72, with a mean of 26. Their computer experience is presented in table 1.

Table 1: The participants' familiarity with computer activities. The columns indicate frequency the activity. a: never, b: once a year, c: once a month, d: once a week, e: several times a week, and f: daily.

	a	**b**	**c**	**d**	**e**	**f**
computer use	0	0	0	0	8	24
internet use	0	0	1	3	12	16
use of search engine	1	0	4	7	13	7

Design The experiment followed a within subjects design. Three variables were manipulated (1) fulfillment of the information need (2) quantification type (one, all, several), and (3) presence or absence of recall feedback, represented by a pie chart. If feedback was present, three ranges were used: low, medium, and high. These were depicted as slices of respectively 10, 20, and 30% of the pie, 40, 50, and 60%, and 70, 80, and 90%. Let us reiterate that the recall value had nothing to do with actual recall. It was only used to give the subjects the impression that the search engine produced this value. In reality a search engine cannot give such a precise number to a user, that would be a paradoxical situation where it would need the user to evaluate the relevance of documents it has not shown to the user.

In our pilot study we had prepared the material such that for each query we could give a number of relevant documents to match the quantification type. As the search engine would always return ten documents, we in effect controlled the *precision@10* (the proportion of relevant documents in the first ten documents).

The dependent variables were (1) satisfaction with the documents, (2) satisfaction with the search engine, (3) usefulness of the chart, and (4) tendency to continue to search. In a questionaire, the first three varialbels were scored on an 11-point Likert scale and the fourth was the answer to a yes/no question.

The hypotheses we investigated are an elaboration of the intuitions we described previously. Especially because they are so intuitively appealing, they have to be carefully laid out.

H1 Fulfillment of information need will be the dominating factor influencing the dependent variables.

Hence we split the other hypotheses up in two cases.

When the information need is fulfilled:

H2 The intuition of participants has the following effect: high recall causes a higher satisfaction, a higher usefulness of recall, and a higher stop rate than low or average recall.

H3 The satisfaction with the documents is high irrespective of quantification type and recall feedback.

H4 The satisfaction with the search engine is high and increases with magnitude of recall. There is no influence of quantification type.

H5 Users do not want to continue to search. Yet, a low recall may persuade them to do so.

H6 The usefulness of recall feedback is low and does not depend on its magnitude. If it would change at all, it would increase in the order of quantification type 1, 3, and 2.

When the information is not fulfilled:

H7 The intuition of participants has the following effect: low recall causes a lower satisfaction, a higher usefulness and a lower stop rate than average or high recall.

H8 The satisfaction with the documents is low irrespectively of quantification type and recall feedback.

H9 The satisfaction with the search engine is low, but increases with magnitude of recall. There is no influence of quantification type.

H10 Users want to continue to search. Yet, a high recall may persuade them to stop searching.

H11 The usefulness of recall feedback is high and does not depend on its magnitude. If it increases at all, it would be in the order of quantification type 1, 3, and 2.

Apparatus Participants interacted individually with Netscape 4.7 on a Macintosh G3. The HTML pages used in the experiment were stored locally to avoid network delays. Several toolbars ('navigation', 'location', and 'personal') were turned off to maximize window area as well as prevent interfering or unneeded interaction. The simulated search engine had the unadorned look and feel of the 'Google' search engine. The advantage of the simulation is obviously that all variables could be carefully controlled. Besides the query page, there was a page with search results (including documents and possible recall feedback) and a questionnaire existing of four questions and a box in which remarks could be written. For each search task we returned exactly ten documents. The participants were provided with pencil and paper to jot down the search task at hand. It had a circle printed on it, where they could copy the pie chart.

Procedure Each participant completed one practice task, and 24 randomized experimental search tasks that included a broad range of topics. The quantification type of each search task was obvious (e.g. the task to find a particular home page, is of type 1). The participants had to read the instructions from the screen. They were told

that we wanted to evaluate a search engine that used a novel search strategy. After the instructions, they had to explain the meaning of a pie chart, so we could check whether it was correctly understood (namely as recall information). For each task they went through the following cycle: (1) read the task printed on paper, which represented the information need, (2) indicate the quantification type, (3) input the keywords to the search machine, (4) inspect the search result, write down the number of relevant documents and copy the pie chart, if any, on paper and (5) fill in the questionnaire.

Results

The four dependent variables were analyzed separately with repeated measures for analysis of variance (GLM). The cohesion between the dependent variables was analysed using linear regression and independent t-tests. We also collected the users' remarks, but we will concentrate here on the summary statistics.

Table 2: Means of the dependent variables in the two conditions *fulfilled* and *unfulfilled* and their levels of significance and F-values.

	Fulfilled (mean)	**Unfulfilled** (mean)	**Sig.**	**F**
Satisfaction documents	9.2	4.2	.000	463.62
Satisfaction search engine	9.0	4.2	.000	387.76
Continue to search	.29	.84	.000	153.40
Chart is useful	6.1	6.9	.072	3.48
Chart might be useful	6.0	7.4	.004	9.86

The influence of fulfillment on the dependent variables is clearly demonstrated in table 2. According to the significance levels, **H1** is confirmed except for the usefulness of the chart.

To avoid a tedious enumeration, we will focus on the main results now. So, instead of giving all the tables for all interactions, we will give table 3 as an example of what the data look like, and then summarize the others (for the reader who wants to study the details, we would be happy to make all the data available).

First we will look at **H2** and **H7**, concerning the intuitions of participants about recall feedback. In the condition unfulfilled, low recall indeed leads to different usefulness (F= 3.81, p= .034), satisfaction with the documents (F= 4.233, p= .013) and search engine (F= 6.803, p= .011). In the condition fulfilled, high recall leads only in type 2 tasks to different usefulness (F= 7.788, p= .007) and satisfaction (documents: F= 11.703, p< .001; search engine: F= 10.067, p= .002). The bahavior, however, is not influenced. This means that intuitions of participants do play a role in evaluation, but not in their subsequent behavior.

Table 3: Satisfaction with the documents, when the information need is fulfilled. 'Q Type' is the quantification type, 'Feedback' the recall feedback. The numbers indicate mean scores on the 11-point scale for user satisfaction.

Feedback / **Q Type**	absent	low	middle	high	overall
1	9.7	10.1	9.7	9.5	9.8
2	9.5	9.0	9.2	9.9	9.4
3	8.9	8.2	8.6	8.3	8.5

Now let's consider **H3-6** (fulfilled condition). The magnitude of the recall did not influence any of the variables. The satisfaction with both the documents and search engine was high but for type 3 lower than for type 1 and 2 (documents: 1-2: p= .276; 1-3: p< .001 and 2-3: p= .002; search engine: 1-2: p= .133; 1-3: p< .001 and 2-3: p= .005). **H3** and **H4** are therefore partly confirmed. As mentioned before, some participants do not agree with us that five relevant documents is enough to fulfill an information need of type 3. As a result, many participants want to continue to search in type 3 tasks of the condition fulfilled (34.4%). Also, in type 2 tasks of this condition many participants want to continue to search (44.5%). This can be explained by the restriction to *ten* documents in our experiment; it is impossible that these always include *all* existing relevant documents. In type 1 tasks, however, 93.0% want to stop searching; most participants obviously fulfilled their information need. Low recall did not cause a larger proportion of participants wanting to stop searching. **H5** is rejected because of these results.

The usefulness of the chart was not as low as expected, but did increase in order of type 1, 3 and 2, confirming **H6**.

Now I will discuss **H8-11** (unfulfilled condition). The satisfaction with both the documents and the search engine was low. Quantification type didn't influence them (p= .397 and p= .512). The satisfaction with the documents was influenced by recall (F= 4.233, p= .013) and was higher in absence of a chart, then in presence. But the satisfaction with the search engine was only in type 2 tasks influenced by the recall feedback low recall causes then a lower satisfaction than average recall (F= 6.803, p= .011), high recall (F= 11.449, p= .001) or no chart (F= 5.666, p= .020). **H8** is just partly confirmed and **H9** is rejected. Participants did want to continue to search (82.3%), confirming **H10**. The usefulness of the chart was not as high as expected, there was an effect

of type (F= 11.07, p< .001); it was highest for type 2, confirming **H11**.

There was a strong cohesion among the variables. Satisfaction with documents and search engine correlate strongly ($\beta = .92, p<.001$), Satisfaction with the documents correlates negatively with the usefulness of the chart ($\beta = -.102, p = .005$), and similarly for the estimated value of the chart, if it was absent ($\beta = .29, p<.001$). Similar values hold for the satisfaction with the search engine. The tendency to continue to search was strongly related to the other three dependent variables.

Discussion and conclusion

Quantification type did influence the results, contrary to our expectations. There are several plausible explanations for this influence. First, in several tasks, some participants did not agree with us about the number of relevant documents among the documents returned. We know that some uncertainty about the relevance of documents existed as some of the participants marked a document as relevant, that we found irrelevant and vice versa. We noticed this uncertainty because the numbers of the documents written on paper did not have complete overlap with the documents we found relevant. Second, there were varying interpretations of the phrase 'several documents' that we used to indicate type 3. Our pilot study indicated that 'several' could stand for 'about five' relevant documents but the participants of our experiment needed more relevant documents to be satisfied. Tasks that were meant to be fulfilled, may, in the eyes of some participants, not have been completely fulfilled and the other way around. Especially for quantification type 3 the satisfaction in the condition fulfilled was lower than expected. The large standard errors, especially in type 1 tasks, also reflect the differences among participants concerning relevancy and fulfillment. Hence, both the conditions fulfilled and unfulfilled are not as unequivocal as expected. Though judgments on relevancy are by definition subjective, more pilot studies could have increased the certainty in interpreting the results for both experimenter and participant.

Nevertheless, the experiment showed us that different types of information needs can be discerned. If search engines can get information about the type of the user's information need, they could adapt the exactness of its search and influence recall and precision (see figure 1)

Second, the results indicate that first, if participants are highly satisfied with the documents, they want to stop searching and they are not interested in the chart, and second, if participants are unsatisfied, they want to continue searching and understand that the chart provides worthwhile information.

To summarize. It seems that fulfillment is a better predictor of user satisfaction and behavior than precision and recall as used to evaluate IR systems. Search results with low precision can indeed result in high satisfaction, e.g. in case of quantification type 1.

Let is briefly comment on the question whether it is worth the effort to see if recall, which is a valid metric to compare the quality of IR systems, is also a valid metric for IR quality as perceived by the end-user. The participants in our experiment found the chart quite useful. This puzzles us, as it was meant to represent recall, and recall had very little overal effect. It might be that participants needed more time to familiarize themselves with the concept of recall feedback. The result is paradoxical enough to warrant further research. We stay with our prediction that recall information indeed will help the searcher in certain cases. But as long as a compromise must be found between recall and precision, precision should get a higher priority; the results suggest that even if the recall is low, the satisfaction can be high.

It is worth the effort to investigate ways to compute recall more precisely than can currently be done (e.g. pseudo recall or relative recall). The present authors are investigating a 'capture-mark-recapture' technique borrowed from biology, to observe in what proportion documents reappear in a search. In addition, we found a few cases where intuition conflicts with experimental findings. These may also be a source for further investigation.

Summarizing the main conclusions: First, among the variables we investigated, the one with the dominant influence on user satisfaction was whether the information need was fulfilled, and second, recall had virtually no influence on satisfaction or search behavior.

References

Buckley, C., & Voorhees, E. M. (2000). Evaluating evalutation measure stability. In *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 33–40). New York: ACM.

Cleverdon, C., Mills, J., & Keen, M. (1966). *Factors Determining the Performance of Indexing Systems* (Tech. Rep.). ASLIB Cranfield Research Project.

Cooper, W. (1968). Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science*, *19*, 30–41.

Hersch, W., Turpin, A., Price, S., Chan, B., Kraemer, D., Sacherek, L., & Olson, D. (2000). Do batch and user evaluations give the same results? In *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 17–24). New York: ACM.

Hoenkamp, E., & de Groot, R. (2000). Finding relevant passages using noun-noun components. In M. Hearst, F. Gey, & R. Tong (Eds.), *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 385–387). New York: ACM.

Hoenkamp, E., Stegeman, O., & Schomaker, L. (1999). Supporting content retrieval from WWW via ‘basic level catogories’. In N. J. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), *Proceedings of the 22rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 301–302). New York: ACM.

Swets, J. A. (1963). Information retrieval systems. *Science, 141*, 245–250.