"THAT's What I Was Looking For": Comparing User-Rated Relevance with Search Engine Rankings

Sameer Patil[1], Sherman R. Alpert[2], John Karat[2], and Catherine Wolf[2]

[1] Department of Informatics, Donald Bren School of Information and Computer Sciences, University of California, Irvine CA 92697 USA
patil@uci.edu
[2] I.B.M. T. J. Watson Research Center, Hawthorne, NY 10532
{salpert, jkarat, cwolf}@us.ibm.com

Abstract. We present a lightweight tool to compare the relevance ranking provided by a search engine to the relevance as actually judged by the user performing the query. Using the tool, we conducted a user study with two different versions of the search engine for a large corporate web site with more than 1.8 million pages, and with the popular search engine Google™. Our tool provides an inexpensive and efficient way to do this comparison, and can be easily extended to any search engine that provides an API. Relevance feedback from actual users can be used to assess precision and recall of a search engine's retrieval algorithms and, perhaps more importantly, to tune its relevance ranking algorithms to better match user needs. We found the tool to be quite effective at comparing different versions of the same search engine, and for benchmarking by comparing against a standard.

1 Introduction

Finding information is a basic task on the Internet. Looking for information on the Internet or on any particular site generally involves a mixture of navigation and search. In order to find information, users typically start at a search engine [13]. Making search better can significantly improve the user experience.

Search engines have typically been assessed in terms of precision and recall. Recall refers to the ratio of relevant records retrieved to the total number of relevant records in the entire database. Precision measures the ratio of relevant records retrieved to the total number of retrieved records. However, as many search developers and researchers now understand:

> "While precision and recall are very helpful in talking about how good search systems are, they are nightmarishly difficult to actually use, quantitatively. First of all, the notion of 'relevance' is definitely in the eye of the beholder, and not, in the real world, a mechanical yes/no decision. Secondly, any information base big enough to make search engines interesting is going to be too big to actually compute recall numbers (to compute recall, you have to know how many matches there are, and if you did, you wouldn't need a search engine)" [2].

M.F. Costabile and F. Paternò (Eds.): INTERACT 2005, LNCS 3585 , pp. 117 – 129, 2005.

Internet users are impatient, and rarely look beyond the first page of search results. In fact, many may scan only those results "above the fold" (visible without scrolling) [10]. Continued improvements in search engine technology have led to a steady rise in user expectations, further decreasing user willingness to look beyond the first few results [10]. Thus, the overall precision and recall of a search engine may be rather meaningless in the face of the user's judgment regarding "Does this result seem to provide the information I am looking for?" In other words, the effectiveness of the search engine is determined by whether the topmost results that are shown on the first page of search results are relevant to the user's goal(s) behind performing the query.

As a result, it is critical that the search engine's determination of which pages are the most relevant to the user's query coincide to the maximum extent with what the user himself or herself judges to be relevant to the task. We believe that the best judge of relevance is the user himself or herself. Given the limitations of keywords in completely and accurately capturing intention [11], the user has a clear advantage of knowing the goal behind the query. Therefore, we argue that asking the user is the most effective way of determining the extent of match between search-engine-rated relevance with actual user-perceived relevance. The tool we describe in this paper provides a lightweight mechanism to capture user-perceived relevance.

2 Related Work

"Search" has always been an important topic in Information Retrieval (IR) research. Initial research focused on algorithmic improvements for better precision and recall in order to retrieve relevant results

Various measures and techniques have been proposed and utilized for evaluating the performance of a search engine, and for comparing different search engines. A typical approach is to use a sample of queries and/or a sample of tasks to compare the relevance of search engine results with the ratings of "expert judges" [20]. As noted before, this approach suffers from the inherent limitation of lacking knowledge regarding user intent and context. Even though independent judges become more proficient with training, as Janes [6] found, "Clearly, there are differences between ratings of users and others, and from this we may infer that there are different processes at work in their judging." Further, the wide range of user queries – a large percentage of which are unique [18] – makes it difficult to select an appropriate sample of queries for evaluation. Indeed, it has been found that such measures may not reflect real world performance [7].

However, relatively little research exists comparing real-life user judgments with search engine relevance rankings [15, 16, 17]. Most notably, Spink [15] conducted a study in which users used a search engine for their own information topics, and rated the relevance of the results. However, the study focused on precision of the first 20 results and ignored how the results were ranked by the engine. Moreover, the study used only a 4-point relevance classification.

The tool we present provides an improvement with a low-burden, cost-effective, and automatic mechanism to capture and compare user relevance ratings with the rankings provided by a search engine. It has been shown that a simple binary classification of "relevant/not relevant" is not adequate to judge the relevance of a document

to the user need [1, 5]. Our tool allows the user to make a relevance judgment at a finer grain with a 10-point scale. Also, the user can provide additional input regarding duplicate or non-existing pages as well as include brief comments. While Spink's [15] study involved a single search engine, the user study we conducted with the tool encompassed 2 search engines.

3 SQUARE: Search QUality Analysis by Relevance Evaluation

SQUARE is a tool we have developed to gather user-perceived relevance information. SQUARE wraps itself "around" a normal user search query. It can thus be used to obtain user assessments of relevance for the results of any Web-based search engine. SQUARE executes the user's query using the underlying search engine, presents the results returned by the search engine to the user along with small questionnaire forms, and gathers feedback from the user regarding the perceived relevance ratings of each result entry and the corresponding target document.

SQUARE runs in a standard Web browser, so it can be invoked simply by navigating to its URL. It starts by asking the user to enter a free-form description of his or her information-seeking task, and a search query to find information related to the task. No additional constraints are imposed on user queries, i.e., the form of the query is identical to that the user would enter when using the underlying search engine directly. SQUARE then programmatically queries the underlying search engine using the keywords entered by the user. The results are identical to those obtained if the query is performed directly via the search engine. The top 10 results (or fewer if the query results in fewer than 10 hits) returned by the search engine are collected.

However, instead of presenting the user with the results in the order in which they were ranked by the search engine – as is the case in an actual search – SQUARE presents the results to the user one at a time in random order. First result entries are presented, individually, in random order. By *result entries*, we mean the entries that would normally be displayed by a search engine in the search result hitlist. These entries typically include a title, search-engine-generated snippet of the document, URL, metadata, and classification. The elements that comprise the entry (and the order in which they occur) may vary between search engines, and is also dependent on the target page in question. SQUARE presents all information provided by the search engine.

As shown in Figure 1, each result entry is presented on a page that asks the user to rate its relevance to his or her task on a 10 point scale. Optionally, users can provide an open-ended reason for their evaluation. Users are also asked to indicate whether the entry appears to be a duplicate of one previously presented.

After all result entries have been presented, SQUARE presents the actual target documents associated with the result entries (see Figure 2). As with the result entries, the target documents are shown one after another in random order[1]. The documents are rated in a manner similar to the result entries. Additionally, users can specify whether the document did not display (i.e., an HTTP 404 error).

SQUARE offers several advantages that make it attractive for evaluation of an individual search engine, and comparison of effectiveness of across search engines.

[1] The randomization of target documents is independent of the randomization of the result entries.

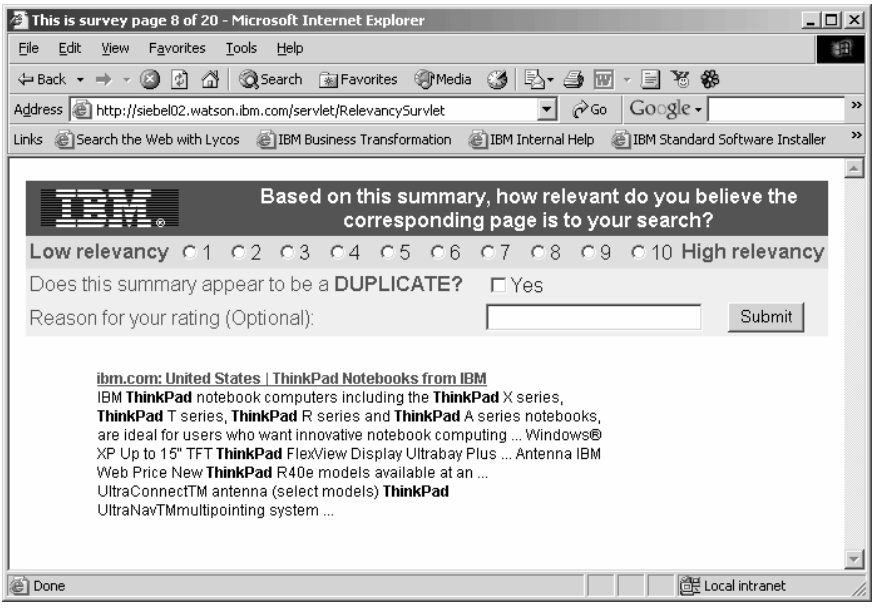

Fig. 1. Sample screen for rating the relevance of a result entry returned by the search engine

Lightweight and Simple. As is evident from the preceding discussion, SQUARE operates in a rather straightforward manner, and fits within the same framework that is used for "normal" searches – using the same Web browser, and the same keywords. Moreover, its simple design allows the user to indicate the rating quickly with minimal effort.

Easily Extensible. To perform searches, SQUARE programmatically forwards queries to the underlying search engine. Thus, it can be extended to any search engine of interest merely by modifying the API (Application Programming Interface) it uses to submit a query, and collect the returned results.

As currently implemented, the search results are formatted in a generic (neutral) fashion, such that any search-engine-specific interface aspects (e.g. text colors, indentation etc.) are ignored, while the search terms that occur within the result entry are bolded. We decided to adopt a neutral approach to eliminate potential bias regarding user ratings of relevance, since a user's relevance perceptions are influenced by form and style as well as the textual content [19]. However, it is relatively straightforward to present a search result "as-is", or to even embed the interface of SQUARE within the results list.

Low-cost, Low-effort. Because SQUARE is deployed via the Web, it eliminates the efforts and costs of arranging for a usability laboratory setting. Users are simply directed to a URL to start the search with SQUARE. SQUARE captures all relevant data to a file for later analysis.

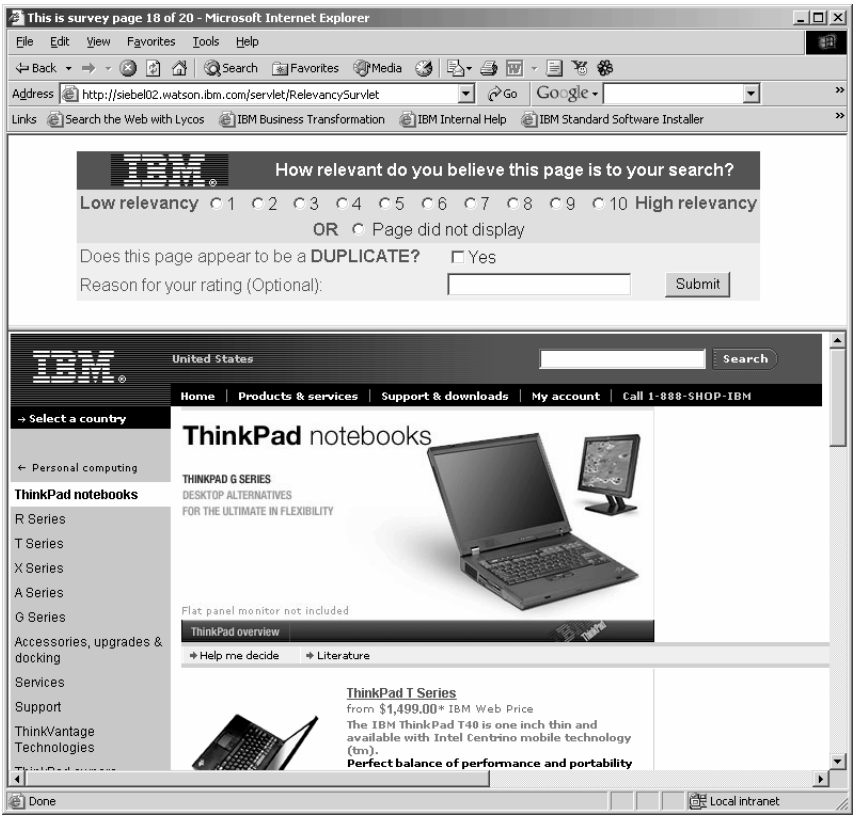

Fig. 2. Sample screen for rating the relevance of the target document of a result entry

Automatic. As SQUARE uses the same framework that is used to conduct actual searches, it is able to simultaneously and automatically capture all relevant metadata along with user input. This results in considerable reduction of effort that would be required for manual capture of this information.

Fast. Finally, SQUARE is able to gather feedback from the user relatively quickly, i.e., in a mere 10-15 minutes per query. As a result, it is easier to attract larger numbers of subjects for a nominal incentive. We were able to gather data from hundreds of users in a very short time frame.

4 Description of Study

We tested SQUARE by conducting a user study that involved two different search engines – the search engine used for intra-site search on the Web-site of a large multinational corporation, and the widely popular Internet search engine Google. In case of the corporate search engine, SQUARE was used to compare two different versions (referred to as Version 1.0 and Version 1.1 in the paper). The two versions of the cor-

porate search engine varied in minor algorithmic details, and in the indexes that they searched. Version 1.1 was, among other things, intended to reduce the number of duplicate pages returned in the result list. Google was chosen because it is widely perceived as the standard in Internet search, and also because it provides a public API making it easy to incorporate it within SQUARE.

4.1 Participants

The study was conducted in three parts – part 1 with Corporate Search 1.0 (CS 1.0), part 2 with Corporate Search 1.1 (CS 1.1), and part 3 with Google. For each of the three parts, an email message was sent to various email lists comprised of employees within the research division of a large technology company. The participants were offered a nominal incentive (a $5 Gift Certificate for the cafeteria) for participation. For the first two parts of the study (i.e. CS 1.0, and CS 1.1), the same email was sent to a different set of employee mailing lists without offering the incentive. Participation rate was about 15% for the employees offered the incentive, and about 2% when no incentive was offered. We received 67 responses for the first part of the study (CS 1.0), 64 for the second part (CS 1.1), and 53 for the third part (Google). As employees of a cutting-edge technology company, all participants are reasonably expected to be highly experienced computer users (and online searchers).

4.2 Methodology

At the beginning, participants were presented with an instruction screen that described the purpose of the study. Participants then proceeded to describe an actual task they might do and to enter keywords to be used for a search engine query. The keywords were then used by SQUARE to gather relevance feedback regarding the search results as described in the previous section. After all result entries and target documents were individually displayed to, and evaluated by a participant, a final screen was shown on which the participants could enter optional overall comments.

Thus, for each participant SQUARE collected participant-specified task description, participant's keywords, evaluations for the result entries (up to 10), and evaluations for target documents of the result entries. Result entry evaluations were comprised of the user's numeric relevance rating (on a 1-10 scale), an optional reason for the rating, and the participant's determination of whether the entry appeared to be a duplicate. Document evaluations were comprised of the same information regarding the target documents with the addition of an indication regarding whether the document did not display.

Along with user evaluations, SQUARE recorded pertinent information such as contents of the result entry, its position in the results list, the URL and the classification of the target document, "keyword" metadata associated with the document, the total number of results returned for the query, and any overall comments from each participant.

4.3 Analysis

After analyzing the queries and tasks that participants had entered for CS 1.0, we found that some were not appropriate for the public (external) corporate Web site, but

rather were intended for either the intranet (internal) site (e.g., searches for internal company project names), or for general search engines (e.g. searches for cartoon websites). Such queries were flagged via independent review by three judges, and excluded from analysis. As a result, in CS 1.1, we slightly modified the wording on the instruction screen to ask specifically for tasks relevant to the external corporate Web site. As before, all queries were reviewed for appropriateness independently by three judges. A few non-relevant queries were still encountered. However, possibly due to the modification in instructions, the percentage of inappropriate queries was much lower compared those for CS 1.0

In case of Google, we excluded six queries. Three queries had returned less than 10 results (2, 5 and 6 respectively), and were eliminated for the sake of consistency with the other two parts in which no queries had fewer than 10 results. Of the remaining three, one had a spelling error due to which all results ended up being marked as highly irrelevant. In an actual search using Google, the alternate spelling suggestion feature would have helped to prevent situation. In the second case, the participant had violated the protocol of the study by simultaneously conducting the same search in a separate browser window. This was indicated in the comments: "*I was first presented with 10 URLS one at a time each of which I followed in another browser window so that I could rate them. I was then presented with the actual pages for each of the previously presented URLs. This second set of pages duplicated the URLs.*" In the third case, the participant had performed a search on the names of one of the authors. Based on the participant's unfamiliarity with the author in question, we judged that it is unlikely this task is representative of an actual task that the user might perform via Google, as was asked in the study. (This is also evident in some of the comments the participant included with their evaluations for result entries and documents).

Finally, we ended up with 48 queries (72% of total queries performed) for CS 1.0, 54 queries (86% of total performed) for CS 1.1, and 47 queries (89% of total performed) for Google. The findings reported in the next section are based on analysis of these queries.

5 Findings

Using the data gathered by SQUARE, we compared each part of the study. Table 1 summarizes the findings from the numeric portions of the data (Note that the scale presented by SQUARE interface is reversed when the data is recorded so in the data 1 represents "Highly relevant" and 10 represents "Not relevant"). Result entries and target documents that were marked as duplicates by the participants have been excluded from the numeric analyses. The numeric ratings coupled with user justifications and comments lead to a number of interesting insights as discussed below:

5.1 Correlation Between Rated Relevance of Result Entry and Target Document

As Table 1 shows, there was a very high level of agreement between ratings for result entries and corresponding target documents ($r = 0.78$ for CS 1.0 and 1.1, and $r = 0.76$ for Google, all $p \sim 0$). This suggests that the mechanisms being used for generating

document snippets for result entries worked adequately in all three cases. Because the correlations were very high, we have restricted the analyses to document ratings only. In general, the two separate ratings are mostly useful in identifying places where there is a discrepancy between the summary and document rating. Such cases could be utilized to help identify where improvements might be made in the generation of result entries.

Table 1. Summary of findings showing comparison of user ratings for CS 1, CS 1.1 and Google (1 = Highly relevant, 10 = Not relevant)

Factor	CS 1.0	CS 1.1	Google
Mean entry relevance	6.14	6.61	5.85
Median entry relevance	7.00	8.00	6.00
Mode entry relevance	10.00	10.00	10.00
Mean document relevance	6.24	6.91	6.18
Median document relevance	7.00	8.00	6.00
Mode document relevance	10.00	10.00	10.00
Duplicates	26.30%	18.30%	13.80%
Correlation between entry & document rating	0.78 ($p \sim 0.0$)	0.78 ($p \sim 0.0$)	0.76 ($p \sim 0.0$)
Correlation between document relevance & ranking	Not. Sig.	0.15 ($p \sim 0.002$)	0.18 ($p \sim 0.0$)
Correlation between document relevance & number of keywords	0.22 ($p \sim 0.0$)	0.17 ($p \sim 0.0$)	0.15 ($p \sim 0.003$)
Average words/Query	2.69	2.98	3.53
Median words/Query	3.00	2.00	3.00

5.2 Correlation Between Search-Engine Ranking and Rated Relevance

Given that the result list is ranked by relevance, a good search engine will place the documents with the highest relevance at the top. As one moves down the list, relevance can be expected to decrease. Thus, for a well-performing search engine, one would expect a correlation between ranking and relevance. As seen in Table 1, relevance was essentially uncorrelated with ranking in CS 1.0, but was positively correlated in CS 1.1. This suggests that although average result quality did not improve in CS 1.1, higher quality results were placed higher in the list. Since most users rarely examine more than a screen (i.e., above the fold), or page of results, this indicates an improvement. The average relevance of the results may not be as important a measure to consider if a highly relevant result appears high in the list. Google results also show a positive correlation of the rankings with relevance. More over, Google performed better than CS 1.1 ($p < 0.01$), although the magnitude of improvement is marginal.

5.3 Average Relevance

Obviously, the better the search engine, the higher the average relevance of all results in the list taken together. However, average relevance may be of limited use when taken by itself. It is more valuable in comparing across different versions of a search engine or between two different search engines, or with a benchmark. We can note from Table 1, that the average relevance in all three parts of the study was below neutral. Although Google performed better than both CS 1.0 and 1.1 ($p < 0.01$), once again the magnitude of difference seems to be marginal. Similar patterns are observed even if we consider only the top 5 search results returned by the search engines.

A factor that affects the average relevance is the presence of large number of documents marked as not relevant. In fact, the mode of rating in all three cases is 10. It appears that users had little hesitancy in using the lowest rating for documents that seemed irrelevant (CS 1.0: 34.2%, CS 1.1, 31.3%, and Google: 27.4%). This is also reflected in user comments such as: "*Most of the documents found were irrelevant to my specific subject. However, one or two were very helpful.*", "*Most search results are irrelevant, but I did get one result that is exactly what I was looking for", "Search engines continue to optimistically supply huge amounts of irrelevant garbage in their answers."*

The indication of high relevance, on the other hand, was less clear. As a result, we decided to delve a bit deeper into the higher ranked results. Table 2 shows statistics for all documents that were rated within the top 3 (i.e., given a relevance rating of 1, 2, or 3), as well as those that were rated the topmost (i.e., rated 1). We can observe that the percentage of highly relevant documents is much higher in case of Google. In addition, Google seems to do a much better overall job of putting the most relevant results at the very top of the results list, or at least amongst the top 5 results. (Although it may appear as if CS 1.0 is better than Google at listing the highest rated result at the top position, the much lower percentage of such results in CS 1.0 must be taken into account when making an overall comparison.)

Table 2. Comparing highly relevant results

	Top three relevant (Relevance = 1, 2, 3)			Highest relevant (Relevance = 1)		
	CS 1.0	CS 1.1	Google	CS 1.0	CS 1.1	Google
N	100	108	109	34	38	58
%N	20.83	20.00	23.19	7.08	7.03	12.34
Mean rank	5.30	4.77	4.77	4.56	4.68	4.48
Median rank	5	5	4	4	5	4
Mode rank	1	1	1	1	2	1
Range	9	9	9	9	8	9
% at #1	14.00	14.80	16.50	23.50	13.20	17.20
% in top 5	53.00	58.30	58.70	64.70	57.90	65.50

5.4 Duplicates

One frequent user complaint regarding CS 1.0 had been a large amount of "duplicates" in the list, i.e., the same resource being presented as two or more separate results. This

was indeed reflected in SQUARE data. More than a fourth of the results were flagged as duplicates by the participants. User comments such as, "*Lots of duplicates. Everything came up at least twice.*", were quite typical. It should be noted that the reported number is actually *lower* than the actual number of perceived duplicates. A duplicate was flagged by the user if he or she believed that it had already been seen before. However, because the user could not specify which already-seen result the current result was a duplicate of, the very first instance seen remained unflagged as a duplicate.

One of the improvements in CS 1.1 was advertised to be duplicate elimination. Again, SQUARE data shows that CS 1.1 resulted in a reduction in the number of duplicates by about a third – an improvement compared with CS 1.0 but significantly worse compared with Google.

Interestingly, reduction of duplicates in CS 1.1 did not improve the average relevance of results in the list. This could be explained by taking into account that duplicate elimination just results in results further down the list being pushed upwards. Given the positive correlation between ranking and relevance, the results being pushed upwards are likely to be of lower relevance.

5.5 Effect of Number of Keywords

Over the past few years, the average query length logged by the corporate search engine has been steadily increasing. We were interested in examining the potential impact of query length on relevance of retrieved results. We found statistically significant positive correlation between rated relevance and the number of keywords in the query. Thus, longer queries result in documents with lower average rated relevance.

6 Discussion

In the previous section, we illustrated how SQUARE can be used to compare search engine performance across versions, and with a benchmark such as Google. We were able to discover that CS 1.1 lived up to its promise of reducing duplicates and pushing highly relevant results up the ranking. Yet, it failed to improve the average relevance of the top 10 results, and fell short of the benchmark that most users are likely to apply to its performance.

Multiple factors may be at play in determining the overall effectiveness of a search engine. For example, merely improving the correlation of the rankings with user ratings may not be sufficient if the overall percentage of highly relevant documents is low. Given that Google is highly reputed for the quality of its search results, it may be surprising that the magnitude of difference (e.g., average relevance, relevance/ranking correlation) for each factor between Google and CS 1.1 was not very large. This suggests that the cumulative effect of all factors taken together could result in substantial differences. In addition, the average query length was longer in case of Google. As discussed in the previous section, longer query lengths are more challenging for a search engine. (There was also some evidence that users deliberately tried to use harder queries to "challenge" Google. For example, one user commented, "*I may be searching for something that doesn't exist we are evaluating whether to patent the idea...*"). Finally, it also begs the question of what part Google's interface plays in user perceptions regarding its effectiveness. It was discussed earlier how one participant failed to notice his spelling error when the spelling-suggestion functionality was stripped off.

Data captured by SQUARE allows us to explore relevance from such multiple angles by allowing us to frame a variety of questions regarding the distribution of relevance judgments in relation to search engine rankings. For instance, one could easily compare the percentage of participants who rated at least one result as relevant. As the above discussion implies, such a multi-dimension perspective is necessary when evaluating overall effectiveness of the search engine. At the same time, each isolated factor represents an area that may be appropriately fine-tuned for better performance.

6.1 Limitations

Although the results presented are quite interesting, we must acknowledge the limitations of the study. For starters, the sample population of technology professionals is not representative of the average user. As a result, we must caution against treating the numeric data in each part of the study in isolation. However, the fact that the samples were drawn from the same population in all three cases allows us to effectively compare the results from the three parts.

Secondly, SQUARE needs to be extended to support query refinement. As several users commented, iterative query refinement is a typical approach used in real-life search behavior:

"*I would have probably realized after glancing the top 5 results that I needed to refine the search.*"

"*When I search Google I am able to refine my search by learning better search terms as I read relevant hits. Your test allowed just one iteration.*"

"*I guess I should have added the terms "sale" or "buy".*"

7 Implications for Design

We have already highlighted how statistics based on the data captured by SQUARE can be utilized by developers of search algorithms to discover avenues for fine tuning. The findings coupled with user comments provide several additional implications for designers of search engine algorithms and interfaces.

7.1 Algorithmic Implications

Snippet Generation. The high correlations between relevance judgments of result entry and the corresponding target document underscore the importance of generating good document snippets to support effective relevance judgments.

Duplicate Elimination. Frustration expressed by users regarding duplicates coupled with the observed improvement in ranking with reduced duplicate count indicates that eliminating duplicates needs to be paid more attention. In particular, "mirrors" may need to be handled specially.

Personalize by Using Context. A search engine can provide better results if it has some knowledge regarding the user and his or her current context. Such functionality could help avoid frustrating user experiences such as: "*This is in Finland and I am searching from the US. Not useful for me.*", or "*Got lots of warranty stuff when I asked for upgrades. I wanted hardware.*"

7.2 Interface Implications

Iterative Query Refinement. As discussed in the previous section, users typically engage in query refinement. Interface improvements that can facilitate this process could make the process more effective and efficient.

Result List Scanning. Presenting the list of result entries in a manner that facilitates quick scanning ensures that users will be able to quickly recognize the most pertinent results. Quick scanning also aids iterative query refinement.

I**ncorporate User Feedback.** Designers may even wish to incorporate parts of the SQUARE interface within the search result list to support gathering impromptu user feedback.

Indeed, approaches to deal with some of the above factors are already being explored [8, 9, 12, 14]. We believe that significant improvements in user experiences could be achieved if designers of search systems treat search as an activity situated in the larger context, rather than an isolated query or session.

8 Conclusion

We have presented SQUARE, a lightweight tool that provides an inexpensive and efficient mechanism for capturing user perceptions regarding how relevant a search result is to the task. SQUARE is interoperable with any search engine that offers access through APIs, provides a standardized presentation of search results across different search engines, and elicits user input regarding perceived relevance of search results. We have described the utility of SQUARE to compare the effectiveness of different versions of the same search engine, and also for benchmarking by comparing against a known standard. We found that relevance needs to be examined from multiple angles in order to gain a thorough understanding of the various factors that affect search engine effectiveness. Based on our findings, we have suggested that search algorithms should provide effective snippet generation, and duplicate elimination, while search interfaces should support quick scanning and iterative query refinement. We urge designers of search systems to treat search as an activity situated in the larger context, rather than an isolated query or session.

Acknowledgements

We would like to thank Pat Velderman, Alfred Kobsa, and all the users who participated in the study.

References

1. Borlund, P. (2003) The Concept of Relevance in IR, Journal of The American Society for Information Science and Technology, 54 (10), pp. 913-925.
2. Bray, T. (2003) On Search: Precision and Recall, http://www.tbray.org/ongoing/When/200x/2003/06/22/PandR

3. Della Mea, V. and Mizzaro, S. (2004) Measuring Retrieval Effectiveness: A New Proposal and a First Experiemental Validation, Journal of the American Society for Information Science and Technology, 55(6), pp. 530-543.
4. Dziadosz, S. and Chandrasekar, R. (2002) Do Thumbnails Previews Help Users Make Better Relevance Decisions about Web Search Results?, In Proc. SIGIR 2002.
5. Eisenberg, E. (1988) Measuring Relevance Judgments, Information Processing and Management, 24(4), pp. 373-389
6. Janes, J. W. (1994) Other People's Judgments: A Comparison of Users' and Others' Judgments of Document Relevance, Topicality, and Utility, Journal of the American Society for Information Science, 45 (3), pp.160-171.
7. Hersh, W. Turpin, A., Price, S. , Chan, B., Kramer, D., Sacherek, L., and, Olson, D. (2000) Do Batch and User Evaluations Give the Same Results?, In Proc. SIGIR 2000.
8. Leroy, G., Lally, A., and Chen, H. (2003) The Use of Dynamic Contexts to Improve Casual Internet Searching, ACM Transactions on Information Systems, 21 (3), pp. 229-253.
9. Liu, F., Yu, C., and Meng, W. (2002) Personalized Web Search by Mapping User Queries to Categories, In Proc. IKM 2002, pp. 558-565.
10. Karat, J. Wolf, C., Alpert, S. R., Velderman, P., and Patil, S. (2003) "Improving Search on IBM.com", IBM Research Technical Report.
11. Kyung-Sun, K. and Allen, B. (2002) Cognitive and Task Influence on Web Searching Behavior, Journal of the American Society for Information Science and Technology, 53(2), pp. 109-119.
12. Muramatsu, J. and Pratt, W. (2001) Transparent Queries: Investigating Users' Mental Models of Search Engines, In Proc. SIGIR 2001.
13. Nielsen, J. (2004) When Search Engines Become Answer Engines, Alertbox, August, 24, 2004, http://www.useit.com/alertbox/20040816.html
14. Paek, T., Dumais, S., and Logan, R. (2004) Wavelens: A New View onto Internet Search Results, In Proc. CHI 2004.
15. Spink, A. (2002) A User-Centered Approach to Evaluating Human Interaction with Web Search Engines: An Exploratory Study, Information Processing and Management: An International Journal, 38(3), pp.401-426.
16. Spink, A. and Greisdorf, H. (2001) Regions and Levels: Measuring and Mapping Users' Relevance Judgments, Journal of the American Society for Information Science and Technology, 52(2), pp. 161-173.
17. Spink, A., and Saracevic, T. (1997) Interaction in Information Retrieval: Selection and Effectiveness of Search Terms, Journal of the American Society for Information Science, 48(8), pp.741-761.
18. Spink, A., Wolfram, D., Jansen, M. B. J., and Saracevic, T. (2001) Searching the Web: The Public and Their Queries, Journal of the American Society for Information Science and Technology, 52(3), pp.226-234.
19. Wolf, C. G., Alpert, S. R., Vergo, J. G., Kozakov, L., and Doganata, Y (2004) Summarizing Technical Support Documents for Search: Expert and User Studies, IBM Systems Journal, 43(3), http://www.research.ibm.com/journal/sj/433/wolf.pdf.
20. Yaltaghian, B., and Chignell, M. (2002) Re-ranking Search using Network Analysis: A Case Study with Google, In Proc. IBM Centre for Advanced Studies Conference.