Comparing Crowd-based, Game-based, and Machine-based Approaches in Initial Query and Query Refinement Tasks

Christopher G. Harris[1] and Padmini Srinivasan[1,2]

[1]Informatics Program, The University of Iowa, Iowa City, IA 52242 USA
christopher-harris@uiowa.edu
[2]Computer Science Department, The University of Iowa, Iowa City, IA 52242 USA
padmini-srinivasan@uiowa.edu

Abstract. Human computation techniques have demonstrated their ability to accomplish portions of tasks that machine-based techniques find difficult. Query refinement is a task that may benefit from human involvement. We conduct an experiment that evaluates the contributions of two user types: student participants and crowdworkers hired from an online labor market. Human participants are assigned to use one of two query interfaces: a traditional web-based interface or a game-based interface. We ask each group to manually construct queries to respond to TREC information needs and calculate their resulting recall and precision. Traditional web interface users are provided feedback on their initial queries and asked to use this information to reformulate their original queries. Game interface users are provided with instant scoring and ask to refine their queries based on their scores. We measure the resulting feedback-based improvement on each group and compare the results from human computation techniques to machine-based algorithms.

1 Introduction

Although searching is a core component of any document retrieval system, few user information needs are satisfied by the initial query. In studies of Web searches, which parallel document searches, more than half of all queries are subsequently reformulated by users after results are returned from an initial query [24]. Query refinement is often necessary due to the presence of over- or under-specified search terms, inappropriate terms retrieving non-relevant documents, and typos. Thus, query refinement is an important step and a core area of study in information retrieval.

The difficulty with the initial query and query refinement may be due to inadequate guidance; most users receive little, if any, instruction on designing effective queries and also have difficulty identifying useful terms for effective query expansion [23]. Since users are typically unaware of the depth or the contents of the document collection in advance, they are neither able to measure (or estimate) their own search success nor are they able to compare their own results with those of others searching the

adfa, p. 1, 2011.

same collection. This results in few opportunities for users to improve their search techniques in an objective manner. This in turn, potentially leads to the perpetuation of these same search-related errors on subsequent queries.

Given how important it is to have an effective query for document retrieval it is not surprising that query design, term expansion strategies, methods for reformulating term weights etc., have been explored over the last several decades. There are many studies involving algorithmic methods (such as the classic Rocchio algorithm [22] and classifiers [15]) and many others exploring human intelligence (using expert searchers and librarians, e.g., [11, 19, 26]). At this point it is almost universally acknowledged that in most cases an initial query refined using a reasonable strategy will yield better results than the initial query. The basis of the refinement may be true or pseudo relevance feedback derived from the documents retrieved by the initial query.

Two recent socio-technological developments charge us to return to query design research. These are the development of crowdsourcing and the development of games with a purpose (GWAP). Crowdsourcing is a framework whereby tasks (such as categorization, image annotation, and relevance assessments) may be accomplished quickly and cheaply by soliciting workers from a largely anonymous pool of participants. GWAP systems are similar except that these devices are also games meant to entertain, reward with scores, be interactive, and in general look and feel like a game. These mechanisms are not error free and so involve strategies for error recognition and correction. Crowdsourcing has gained widespread attention, as illustrated by recent conferences and workshops even in the IR context [3, 9, 18]. GWAP systems, while relatively harder to implement, have also garnered some interest, though not yet as much as with crowdsourcing.

These two developments motivate our goal, which is to assess the use of human intelligence through crowdsourcing and GWAPs both for initial query design and for query refinement in document retrieval. Note that this human intelligence is not that of the original user or of an expert librarian (an angle well-studied in the literature), but of the largely anonymous individuals. As indicated in [14], if the methods examined here are found to be effective then we will have the beginnings of a new approach for assisting searchers with query design. This option may be invoked when a query is particularly difficult and the information need has longevity (e.g., in topic detection and tracking [2]) or where some latency in returning results can be tolerated.

We study the value of using largely anonymous people via crowdsourcing for query design; this includes both initial query formulation and query refinement given some relevance feedback. We study this anonymous people approach in game (GWAP) and non-game settings. This allows us to tease out, for example, the effects of offering entertainment on quality and cost. As a contrast we also study query design with a more homogenous and not so anonymous group of individuals; namely students in a campus. Finally we compare performance with an algorithmic baseline. We compare retrieval results obtained using all of these query design methods applied to a com-

mon set of topics and by running the resulting queries with the same retrieval algorithms and against the same collection. We ask the following research questions:

1. Does retrieval performance differ when the initial query is designed by humans versus the machine?
2. Does retrieval performance differ when feedback-based query refinement is done by humans versus the machine?
3. Does retrieval performance differ for humans using the non-game (basic web interface) versus the game interface? (Note this question is asked both for initial query design and for query refinement with feedback).
4. For each type of interface (game and non-game) does retrieval performance differ between student participants and crowdworkers? (Note this question is asked both for initial query design and for query refinement with feedback).

This is the first controlled study we know of that assesses the value of crowdsourcing and online games for query design and to compare these with query design by humans recruited from more traditional settings and by algorithms. Our long-term goal is to explore mechanisms for involving crowdsourcing and games (relatively new socio-technological developments) in information retrieval. Here we focus on query design - a core step in information retrieval.

The remainder of this paper is organized as follows. In the next section, we discuss the background of our approaches. In Section 3, we provide a description of our experimental methods. In Section 4, we provide our results. Section 5 provides some topic-specific analysis and is followed by a discussion of our general findings in Section 6. We conclude and briefly discuss future directions of our work in Section 7.

2 Background and Motivation

2.1 Crowdsourcing-based approaches

To date, most crowdsourcing studies in information retrieval have examined relevance assessment. Several studies, such as [4, 19] have compared the crowd to experts in document assessment, concluding there is little difference in quality, particularly when multiple assessors are used. Few evaluations have been conducted to compare crowd-based and lab-based participants on search performance. One study compared crowd and lab participants on multimedia search results in [13], concluding that the two groups were indistinguishable in quality.

Integrating the crowd is becoming more commonplace for the difficult searches, perhaps indicating the crowd represents a nice tradeoff between speed, cost, and quality. Bozzon *et. al.* describe a tool called CrowdSearcher, which utilizes the crowd for difficult searches in [7]. A study by Yan *et. al.* described a mobile search application in [27]; claiming a search precision of 95%. Ageev *et. al.* conducted an experiment to

evaluate crowd search techniques in [1], but do not compare the crowd's performance with other groups. These studies provide the premise that the crowd can be used to search effectively and deliver results with reasonable precision.

2.2 Game-based approaches

Only a few games with a purpose (GWAP) have been constructed to address initial query and query reformulation effectiveness. Thumbs-up [10] is a GWAP that uses output-agreement mechanism to gather relevance data. This game asks players to evaluate search terms and attempt to independently determine the most relevant document to a given query. Search War [17] is another game used to obtain data on search relevance and intent for a user-provided query. Players are paired and each given a unique search query and the objective of guessing their opponent's search query first. The design relies on the premise that players will select the least relevant webpage w.r.t. the search query, to provide to their opponent as hints, which implicitly provides a relevance judgment.

Koru [20], the most similar game to the one we use in our study, allows users to assess their search skills relative to other searchers and evaluate how their own searches might be improved. Like other GWAPs, it is intended to be both fun and to create valuable output on query refinement behavior in a controlled information task. However, it does not make a comparison between different approaches and it is limited to a small document collection from a single source (the New York Times).

2.3 Machine-based Approaches

There have been a number of studies that examine interactive query expansion versus automatic query expansion and reformulation. Interactive query expansion and reformulation can be used as an effective means of improving a search. Efthimiadis [12] found system-provided terms, on average, when selected, improved retrieval performance. Conversely, Belkin, et al. [6] found that humans rarely used relevance feedback features and were often puzzled by some machine-suggested terms. Ruthven [23] demonstrated that human searchers are less likely than machine-based systems to make good expansion and reformulation decisions. Anick [5] found that users made little use of machine-suggested terms to expand and refine their queries, but when they did it improved retrieval performance. Thus, there are mixed performance results from machine-provided query reformulation and these machine-based approaches have not been adequately evaluated against human computation-based methods.

3 Experimental Methods

We evaluated performance on three treatments: two different *query types* (initial queries and queries refined based on feedback), three different *approaches* (crowdsourcing, game and machine) and, for crowdsourcing and game approaches, two different

user types (undergraduate students recruited on campus and crowdworkers recruited through an online labor market).

3.1 Datasets

We randomly selected 20 topics used in the TREC-7 ad hoc task. Since the collection involved some topics that were outdated, we discarded those topics from our list of selected topics. The 20 topic numbers chosen were: 351, 354, 355, 358, 359, 363, 364, 369, 374, 375, 379, 380, 388, 389, 390, 393, 395, 396, 399, and 400. These topics were presented to each user in the same order. We used the relevance judgments provided by TREC assessors as our gold standard. The number of relevant documents per topic ranged from 7 (for topic 380) to 361 (for topic 354), with an average of 87.9 relevant documents per topic.

3.2 Query Design Approaches

Seek-o-rama (Data Collection Web Interface)
To examine queries issued through standard browser interface, we invited participants to use Seek-o-rama, a PHP-based data collection interface.[1]

Initial Query Formulation
Users were provided with the title, the description, and the narrative for each of the 20 topics. Participants were given a large text box to input their query, with a pop-up help screen available to them throughout the task. We provided detailed instructions and examples of how to construct queries using terms and simple operators (AND, OR and NOT), and provided the following objective to participants: "The objective of Seek-o-rama is to construct queries that will bring back as many relevant documents as possible while excluding non-relevant documents".

Query Refinement
Once a user had provided initial input for each of the 20 topics, they were instructed to return after two hours to allow us time to run the provided queries against our document collection, provide the recall and precision for each query for the second round. The user's original search terms were pre-loaded in the input text boxes for each topic, allowing easy modification to their original query. Also, in the second round, we provided users with the highest-ranked relevant and non-relevant document from the collection to aid them in their query refinement.

Seekgame (Game Interface)
Some users invited to participate in this exercise were randomly selected to use Seekgame, a PHP-based game, instead of the Seek-o-rama interface.

[1] Screenshots are available at the following URL: http://irgames.org/seekorama/

Initial Query Formulation.
Users selected to use Seekgame were given a different URL, and were presented with the same initial screen outlining the game's objectives, instructions on term and operator rules as the Seek-o-rama interface participants. Participants were asked to enter the initial query. The game instructions also had the following additions. First, there was a time-based constraint that required search terms to be entered within 30 seconds. Second, scoring was provided instantly (explained soon). Third, participants had musical sound effects to enhance the interface's game-like feel. Last, a leaderboard and badges, or icons, were awarded for superior game performance.

Query Refinement.
Unlike Seek-o-rama, the Seekgame did not provide users with precision and recall information from their initial round as they began their second round. This was because the calculation of this information was not integrated into the game interface and would take away from the feeling of engagement. Instead once a user entered a set of terms for a topic, these terms were parsed to remove stopwords, stemmed, and compared against a weighted list of stemmed terms obtained from documents judged relevant for that topic. A pop-up screen provided scoring and bonus information to each player after they submitted their query. A higher score was awarded for the use of relevant terms not commonly used by other participants. This score was immediately calculated and issued to the user, along with a time-based bonus for completing the search quickly. Once a user completed the first round, they could begin the query refinement round without delay. Users were instructed to refine their initial query based on their score and a relevant and non-relevant document provided to them to aid their refinement, subject to the same 30-second time restriction.

Stars were awarded to users who scored above a certain threshold. Badges were given to users having the highest overall score, and a leaderboard was shown to the users, providing the option for top scorers to add their names for "bragging rights".

3.3 Algorithmic Baseline

Initial Query Formulation
The machine-based queries used the title and the description, as provided from the TREC topics data. Similar to Seek-o-rama and Seekgame, this input had stopwords removed using the same stopword list and were stemmed using the Porter stemmer.

Query Refinement
Using the ranked list returned by Indri [25], we selected the highest-ranked document from the results of the initial query. We added the terms contained within the headline and byline of the retrieved document as additional inputs to the query, applied the stemming and stopword list to the added terms. This became our refined query.

3.4 Participants

Crowdsourcing workers (N=58) were recruited using Amazon Mechanical Turk. We structured the task such that, to receive any compensation, these crowdworkers would have to complete both rounds of initial query design and query refinement. We discarded the inputs for those workers who did not complete all 20 topics in both rounds. We paid $0.20 for crowdworkers to complete both rounds, regardless of interface. Undergraduate student volunteers (N=47) were recruited from several sections of an undergraduate business course from a small Midwestern university in September 2012. Participants from this group, which we call our *student participants*, were randomly assigned to use one of two interfaces and they were not compensated.

3.5 Assigning Participants to Interfaces

Student and crowd participants were assigned randomly to either Seek-o-rama or Seekgame, but not both. Of the student participants, 7 failed to complete both rounds of the task; of the crowdworkers, 18 failed to complete both rounds. In each case, those participants who did not complete both rounds and the two surveys had their inputs removed from our dataset. Participants in each of the human participant groups were split equally between the game and non-game treatments.

3.6 Retrieval Algorithms

We used two standard retrieval algorithms implemented by the widely-used Indri [25] system. The first uses *tf-idf* scoring to rank documents against queries [16]. The second uses the Okapi algorithm [21]. For tf-idf, we used parameter values k1 = 1.2 and b = 1.2; for Okapi we used parameter values k1= 0.75, b = 0.75, and k3 = 7.

4 Results

The results from our study, comparing the different human-based approaches and interfaces to the machine algorithm baseline, are summarized below in Table 1.

Table 1. Overall results from our study, comparing human approaches to the machine baseline

Approach	**Initial Query**				**Query Reformulation**			
	MAP		P@10		MAP		P@10	
	Okapi	*tf-idf*	Okapi	*tf-idf*	Okapi	*tf-idf*	Okapi	*tf-idf*
Students - Non-game	0.106	0.104	0.203	0.198	0.089	0.093	0.231	0.225
Students - Game	0.114	0.102	0.179	0.175	0.135	0.131	0.206	0.201
Crowd - Non-game	0.098	0.094	0.183	0.178	0.110	0.111	0.215	0.209
Crowd - Game	0.131	0.121	0.179	0.174	0.136	0.128	0.203	0.197
Algorithm	0.076	0.073	0.145	0.141	0.079	0.076	0.160	0.155

We conducted tests to examine each of our four research questions, which are provided in Table 2. In each test described below, we provide two-tailed t-tests at the $p<0.05$ level of significance for the Okapi results. Our *tf-idf* results provided the same

conclusions with only minor differences in statistical significance, so they are not reported here.

Table 2. Summary of findings on the 20 topics for our four research questions, based on two-tailed t-tests ($p<0.05$). Standard deviation is given in parentheses next to each mean value. An asterisk indicates it is statistically significant at $p<0.05$.

Research Question	**Mean Average Precision (MAP)**		**Top 10 Precision (P@10)**	
	Initial Query	**Query Refinement**	**Initial Query**	**Query Refinement**
RQ1 and **RQ2**: Humans (A) vs. Machine (B)	A: 0.110 (0.183) B: 0.076 (0.146) p=0.041*	A: 0.120 (0.182) B: 0.075 (0.149) p=0.041*	A: 0.186 (0.151) B: 0.145 (0.176) p=0.034*	A: 0.214 (0.147) B: 0.160 (0.179) p=0.033*
RQ3: Game (A) vs. Non-Game (B)	A: 0.117 (0.197) B: 0.102 (0.201) p=0.063	A: 0.135 (0.163) B: 0.105 (0.114) p=0.044*	A: 0.179 (0.148) B: 0.193 (0.102) p=0.040*	A: 0.205 (0.136) B: 0.224 (0.099) p=0.036*
RQ4: Crowd (A) vs. Students (B)	A: 0.110 (0.167) B: 0.110 (0.161) p=0.052	A: 0.123 (0.181) B: 0.118 (0.169) p=0.055	A: 0.181 (0.182) B: 0.191 (0.176) p=0.056	A: 0.219 (0.137) B: 0.214 (0.153) p=0.057

Our first two research questions compared human-based and machine approaches on mean average precision (MAP) and p@10 for both initial query formulation and query refinement across all 20 topics (See [8] for further discussion of these parameters). We found a significant difference for both initial query and query refinements between the two, indicating that for the 20 topics we examined, humans provided better mean average precision as well as precision over the top 10 documents retrieved using the same retrieval algorithms as compared with the machine approach.

Next, performed a test to examine our third research question; that is, compare the game and non-game interfaces for our human participants on average precision and p@10 across all 20 topics. For the initial query formulation, we did not find a significant difference related to the type of interface used on mean average precision. For query refinement, however, we found a significant effect on average precision, with game interfaces providing a higher mean average precision than non-games. For p@10, our test also indicated a significant difference for both initial queries and query refinements, but in contrast to our finding on average precision, the non-game interfaces provided better precision in the top 10 retrieved documents.

Last, to examine our fourth research question, we ran tests to compare the crowd and student subject groups on average precision and p@10 across all 20 topics. For the initial query formulation, no significant effect was found on average precision related to the group used for initial query formulation or query refinement. Likewise, we found no significant difference between the crowd and students for p@10. This indicates that there was no significant difference in either the average precision or in precision for the top 10 documents retrieved between the two human subject groups.

5 Topic Specific Analysis

Tables 3 and 4 below provide an overview of the number of topics favoring the different treatments we examined for average precision and p@10, respectively.

Table 3. Preference determination for each topic based on average precision (AvgP)

AvgP	Number of Topics					
	A > B		A = B		A < B	
	Initial Query	Query Refinement	Initial Query	Query Refinement	Initial Query	Query Refinement
Human (A) vs. Machine (B)	17	12	3	3	0	5
Game (A) vs. Non-Game (B)	16	15	3	3	1	2
Crowd (A) vs. Students (B)	11	10	3	3	6	7

Table 4. Preference determination for each topic based on precision of the top 10 retrieved documents (p@10)

p@10	Number of Topics					
	A > B		A = B		A < B	
	Initial Query	Query Refinement	Initial Query	Query Refinement	Initial Query	Query Refinement
Human (A) vs. Machine (B)	11	12	8	7	1	1
Game (A) vs. Non-Game (B)	1	4	9	6	10	10
Crowd (A) vs. Students (B)	5	6	7	7	8	7

In the previous section, we examined the effects of each treatment across all 20 topics as a single test. From Tables 3 and 4, we can observe that when observing the best approach per topic, the majority of topics were best resolved using human approaches over machine approaches, which is consistent with Table 2. Likewise, we get a larger number of topics favoring game approaches for average precision, but more topics favor non-games in an evaluation of p@10. However, the results show some interesting contrasts in Tables 3 and 4 that are not apparent in Table 2. For example, human-based approaches, we see that more topics favor the crowd for average precision, but this is reversed (albeit slightly) in our examination of p@10 on these topics.

6 Analysis and Discussion

Consistent with a number of earlier studies on Web logs, only four percent of queries written by humans using the non-game approach used term operators, such as ‘AND’, ‘OR’ and ‘NOT’. Although the instructions, examples, and help were made available to users and instructed them on the advantages and proper use of these operators, we believe the influence of Internet-based search techniques (that only require a set of terms without operators) has likely influenced the user’s non-operator-based querying technique. Humans supplied fewer terms, on average, than machine approaches (5.1 terms vs. 8.3 terms for initial query; 6.7 terms vs. 15.6 terms for the query refinement). Game participants supplied fewer terms than non-game participants (6.3 terms for non-game vs. 3.4 terms for game in the initial query; 7.2 terms vs. 4.6 terms for

the query refinement). Supplying more terms did not necessarily provide more precise results. We found that the average precision on these 20 topics from machine-based methods showed a better correlation with the number of TREC assessor-determined relevant documents than humans in the initial query ($r=0.683$, $p<0.01$) and with the query refinement ($r=0.614$, $p=0.04$). This may indicate that supplying more terms work better when the pool of relevant documents is large, but this approach does a poor job at finding relevant documents when the number of relevant documents is small.

From Table 3, we observe that game-based approaches work well for achieving results with a higher mean average precision, but non-game approaches worked better for providing a higher precision in the top 10 retrieved documents. We also examined p@20 and p@50 for these two interface types and found non-game approaches consistently provided better precision than game approaches across the top set of documents retrieved. Game-based approaches may have capitalized on the "fun" aspect in the initial query, but this aspect may have encouraged the wrong type of terms to be provided, increasing the non-relevant documents in the retrieved set. We found that game-based approaches were best at retrieving the "rare" relevant documents missed by other approaches. This pattern was also reflected in TREC-7 ad hoc results.

Three of the topics (369: "anorexia nervosa bulimia", 379: "mainstreaming", and 388: "organic soil enhancement") did not have any relevant documents provided for any of the treatments or our machine approach. These topics had few TREC-assessed relevant documents (13, 16, and 51 respectively); an evaluation of the user-supplied queries indicates that few additional relevant terms for these topics were provided by users. For example, topic 379 "mainstreaming" asked users to "identify documents that discuss mainstreaming children with physical or mental impairments". This required unconventional knowledge of this topic. For topics that were well-covered in the mainstream media, e.g., topic 400: "identify documents which indicate measures being taken by local South American authorities to preserve the Amazon tropical rain forest", a variety of terms were supplied by users resulting in a much higher average precision. The challenge of identifying query terms to find relevant documents for these three topics also occurred with nearly all TREC-7 ad hoc task participants.

7 Conclusion

Although query design, term expansion strategies, methods for reformulating term weights etc., have been studied extensively, two recent socio-technological developments – crowdsourcing and GWAP – have motivated a new investigation of query design research. In this paper, we conduct a study to evaluate different how these developments may impact precision in initial query construction and feedback-based query refinement. Using identical retrieval algorithms, this study examines how human-based query approaches compare with machine-based approaches on 20 TREC topics, concluding that human approaches provide better mean average precision

(MAP) and precision in the top 10 retrieved documents (p@10), as compared with machine approaches. We also compare MAP and p@10 for a web-based interface and a game interface, discovering that game interfaces provide a higher MAP but non-game interfaces provide a higher p@10. This finding likely has to do with the engagement aspect of games affecting a user's term choice. For this reason, we believe the use of games in search techniques begs further examination. Last, we examine how anonymous crowd-based participants compare with undergraduate students, concluding no significant difference in MAP or p@10 for the 20 topics investigated.

Overall, we find approaches that encourage a larger number of terms in a query do not necessarily provide a better MAP or p@10 performance, particularly when the number of relevant documents in the collection is relatively small. Topics that human searchers were less familiar with had lower MAP and p@10 results than those that were more familiar to human searchers.

The research explored in this paper has uncovered some interesting aspects of human computation and search techniques that we have only briefly covered. We anticipate additional work to examine what aspects of games can improve initial query and query refinement performance and look at how this can be integrated to make the user experience in search more engaging and more accurate. Likewise, we hope to examine techniques that integrate the crowd into a user's document search process, and how this might affect query performance.

References

1. Ageev, M., Guo, Q., Lagun, D. and Agichtein, E. (2011) Find it if you can: a game for modeling different types of web search success using interaction data. In Proceedings of SIGIR'11. ACM, New York, NY, USA, 345-354
2. Allan, J., Papka, R. and Lavrenko, V. (1998) On-line new event detection and tracking. In Proceedings of SIGIR'98. ACM, New York, NY, USA, 37-45.
3. Alonso, O. and Lease, M. (2011) Crowdsourcing for information retrieval: principles, methods, and applications. In Proceedings of SIGIR'11. ACM, New York, NY, USA, 1299-1300
4. Alonso, O. and Mizzaro, S. (2012) Using crowdsourcing for TREC relevance assessment. Information Processing & Management. 48:6, 2012, 1053-1066.
5. Anick, P. (2003) Using terminological feedback for web search refinement: a log-based study. In Proceedings of SIGIR '03. ACM, New York, NY, USA, 88-95.
6. Belkin, N. J., Cool, C., Kelly, D., Lin, S.-J., Park, S. Y., Perez-Carballo, J. and Sikora, C. (2001) Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. Inf. Process. Manage., 37:3, 403-434.
7. Bozzon, A., Brambilla, M. and Ceri, S. (2012) Answering search queries with CrowdSearcher. In Proceedings of WWW'12 (Lyon, France). ACM, New York, USA, 1009-1018.
8. Buckley, C. and Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In Proceedings of SIGIR '04. ACM, New York, NY, USA, 25-32.

9. Carvalho, V. R., Lease, M. and Yilmaz, E. (2011) Crowdsourcing for search evaluation. SIGIR Forum 44, 2 (January 2011), 17-22.
10. Dasdan, A., Drome, C., Kolay, S., Alpern, M., Han, A., Chi, T., Hoover, J., Davtchev, I. and Verma, S. (2009) Thumbs-Up: a game for playing to rank search results. In Proceedings of the ACM SIGKDD Workshop on Human Computation (Paris, France). ACM, New York, USA, 36-37.
11. Dillon, A. and Song, M. (2006) An empirical comparison of the usability for novice and expert searchers of a textual and a graphic interface to an art-resource database. Journal of Digital Information, 1:1
12. Efthimiadis, E. N. (2000) Interactive query expansion: a user-based evaluation in a relevance feedback environment. J. Am. Soc. Inf. Sci., 51:11, 989-1003.
13. Harris, C. G. (2012) An Evaluation of Search Strategies for User-Generated Video Content. In Proceedings of the WWW Workshop on Crowdsourcing Web search (Lyon, France), 48-53.
14. Harris, C. G. and Srinivasan, P. (2012) Applying Human Computation Mechanisms to Information Retrieval. In Proceedings of 75th Annual Mtg of ASIS&T, Baltimore, MD.
15. Joachims, T. (1996) A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. DTIC Document, 143-151
16. Jones, K. S. (1972) A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28:1, 11-21.
17. Law, E., Ahn, L. v. and Mitchell, T. (2009) Search war: a game for improving web search. In Proceedings of the ACM SIGKDD Workshop on Human Computation (Paris, France). ACM, New York, USA, 31-31.
18. Lease, M. and Yilmaz, E. (2012) Crowdsourcing for information retrieval. ACM, New York, USA, SIGIR Forum 45:2 (January 2012), 66-75.
19. McKibbon, K. A., Haynes, R. B., Walker Dilks, C. J., Ramsden, M. F., Ryan, N. C., Baker, L., Flemming, T. and Fitzgerald, D. (1990) How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searches. Computers and Biomedical Research, 23: 6, 583-593.
20. Milne, D., Nichols, D. M. and Witten, I. H. (2008) A competitive environment for exploratory query expansion. In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL '08). ACM, New York, NY, USA, 197-200.
21. Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M. and Gatford, M. Okapi at TREC-3. (1995) NIST Special Publication SP-1995), 109-121.
22. Rocchio, J. J. (1971) Relevance feedback in information retrieval. In Gerard Salton, editor, The SMART Retrieval System—Experiments in Automatic Document Processing, 313–323, Englewood Cliffs, NJ, 1971. Prentice Hall.
23. Ruthven, I. (2003) Re-examining the potential effectiveness of interactive query expansion. In Proceedings of SIGIR'03. ACM, New York, NY, USA, 213-220.
24. Spink, A., Jansen, B. J., Wolfram, D. and Saracevic, T. (2002) From e-sex to e-commerce: Web search changes. Computer, 35: 3, 107-109.
25. Strohman, T., Metzler, D., Turtle, H. and Croft, W. B. (2005) Indri: A language model-based search engine for complex queries. In Proceedings of the International Conference on Intelligence Analysis, May 2-6, 2005, McLean, VA. Poster.
26. Turtle, H. (1995) Natural language vs. Boolean query evaluation: a comparison of retrieval performance. In Proceedings of SIGIR'95. Springer-Verlag New York, 212-220
27. Yan, T., Kumar, V. and Ganesan, D. (2010) CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones. In Proceedings of MobiSys '10. ACM, New York, NY, USA, 77-90.