Social and Collaborative Web Search: An Evaluation Study

Kevin McNally, Michael P. O'Mahony, Barry Smyth, Maurice Coyle, Peter Briggs
CLARITY: Centre for Sensor Web Technologies
School of Computer Science and Informatics
University College Dublin, Ireland
{firstname.lastname}@ucd.ie

ABSTRACT
In this paper we describe the results of a live-user study to demonstrate the benefits of using the social search utility HeyStaks, a novel approach to Web search that combines ideas from personalization and social networking to provide a more collaborative search experience.

Author Keywords
Collaborative Web Search, User Evaluation, HeyStaks

ACM Classification Keywords
H.4.0 Information Systems Applications: General

General Terms
Algorithms, Experimentation

INTRODUCTION

Given the recent emphasis on the *social web* and *collaborative computing* it is somewhat surprising that one of our most familiar online tools, the search engine, has remained mainly "anti-social" in terms of the way that it interacts with users. However, there are signs that this is set to change and recently there has been considerable interest in the potential for web search to evolve to become a more *social* activity [2, 4], whereby the search efforts of a user might be influenced by their social graph or the searches of others, potentially leading to a more *collaborative* model of search.

In this paper we focus on HeyStaks (*www.heystaks.com*), a particular type of social search service. Its aim is to help people during mainstream search tasks — that is, when they are using mainstream search engines — by harnessing the recent search experiences of their friends and colleagues via their social networks. The emphasis then is on making the solitary world of web search more collaborative. This relates to recent work in the area of *collaborative information retrieval*, which attempts to capitalize on the potential for collaboration during a variety of information seeking tasks [1, 5, 6, 7, 8].

IUI'11, February 13–16, 2011, Palo Alto, California, USA.

The key contribution of this paper is exploring a number of important questions relating to the benefits of the HeyStaks approach to collaborative/social web search, including: 1) Does search collaboration help individual searchers find more relevant results, when members of shared search staks, than they might have on their own? 2) How does search collaboration influence the efficiency of search sessions? 3) How good are the recommendations made by HeyStaks? To answer these questions, we present the results of a closed, live-user trial. This study complements earlier evaluations of HeyStaks, such as that carried out in [3]. These earlier evaluations had the benefit of being open-ended trials, following users during routine search tasks, but were limited in their ability to evaluate the baseline relevance of HeyStaks recommendations. The benefit of the present (closed) trial is that it facilitates a more detailed comparative evaluation of recommendation relevance by comparing HeyStaks recommendations directly to the default Google results.

HEYSTAKS

A detailed description of the HeyStaks architecture is given in [8]. To provide context for this work, here we briefly review the key concepts and functionality of the system.

The HeyStaks service is designed to add a layer of collaborative/social search on top of mainstream search engines. Importantly, HeyStaks allows for context-sensitive search: users can store their search experiences (queries and associated results) according to topic in repositories called *search staks*. A user can select the current search context (stak) using the HeyStaks Web browser *toolbar*. The toolbar is a key component of HeyStaks, and helps to tightly integrate HeyStaks functionality into mainstream search engines (e.g. Google, Bing and Yahoo), allowing users to continue to use their favourite search engine while benefitting from the social search services provided by HeyStaks.

The key functionality provided by HeyStaks lies in the *recommendations* that can be made to users at search time, which are integrated directly into the search engine's interface. In essence, results that other users have found to be relevant for similar queries in the past are recommended to the target searcher. The staks provide the primary source of these recommendations: when a user submits a query to a search engine, in a given stak context, this query is fed to the HeyStaks back-end server which generates a set of recommendations based on the target stak and, possibly, other staks that the user has joined. Moreover, since HeyStaks facilitates the

Question
1. Who was the last Briton to win the men's singles at Wimbledon?
2. Which Old Testament book is about the sufferings of one man?
3. Which reporter fronted the film footage that sparked off Band Aid?
4. Which space probes failed to find life on Mars?

Table 1. A sample of the user-trial questions.

sharing of staks, users can benefit directly from the previous search experiences of other members.

In additional to selecting or creating new search staks, the toolbar also facilitates users to *tag* or comment on result pages; to *vote* (up or down) on pages; and to directly *share* pages (by email or by posting to their Facebook Wall etc.) with other HeyStaks members. Indeed, all such activities on result pages are logged by the system, and these activities are also employed to further refine the recommendation process.

EVALUATION

Our experiment involved 64 first-year undergraduate university students with varying degrees of search expertise. The students participated in a general knowledge quiz during a supervised laboratory session, and answered as many questions as possible from a set of 20 questions in a 60 minute period. Each student received the same set of questions which were randomly presented to avoid ordering bias. The questions were chosen from specifically for their obscurity and difficulty; see Table 1 for a sample of these questions.

Each user was allocated a desktop computer with Mozilla's Firefox web browser and the HeyStaks toolbar pre-installed; they were permitted to use Google, enhanced by HeyStaks functionality, as an aid in the quiz. The 64 students were randomly divided into search groups. Each group was associated with a newly created search stak, which would act as a repository for the groups' search knowledge. We created 6 *solitary* staks, each containing just a single user, and 4 *shared* staks containing 5, 9, 19, and 25 users. The solitary staks served as a straightforward benchmark to evaluate the search effectiveness of individual users on a non-collaborative search setting, whereas the different sizes of shared staks provided an opportunity to examine the effectiveness of collaborative search across a range of different group sizes. All activity on both Google search results and HeyStaks recommendations was logged, as well as all queries submitted during the experiment.

Methodology

All activity on both Google search results and HeyStaks recommendations was logged, as well as all queries submitted during the experiment. The following event/activity information was logged during the trial for later analysis: 1) The time at which the activity occurred; 2) The ID of the user who acted on a result and the stak ID in which the action was taken; 3) The URL of the page acted on; 4) The type of action (result selection, tag, vote or share); and 5) The type of result acted on, i.e. either an organic Google result or a HeyStaks recommended result.

For the purpose of establishing a ground-truth for result relevance, each result page was examined post-trial by a number of experts and classified as *relevant*, *partially relevant* or *not relevant* depending on whether the result helped the user to directly or indirectly answer the question at hand, or whether is contained no useful information for the question. Approximately 66% of result pages acted on were classified as not relevant, while only 14% were deemed relevant, thereby demonstrating the difficulty of the quiz questions.

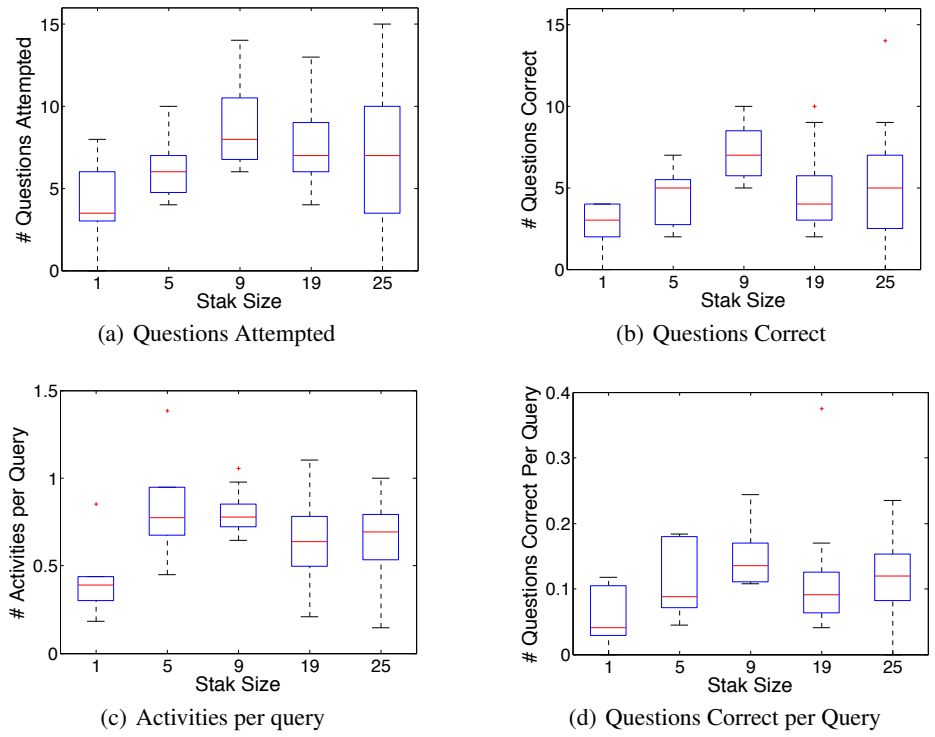
During the 60 minute trial a total of 3,124 queries and 1,998 result activities (selections, tagging, voting, popouts) were logged, and 724 unique results were selected. As expected, during the course of the trial, result selections — the typical form of search activity — dominated over HeyStaks-specific activities such as tagging and voting. Result selections, on average, accounted for over 81% of all activities, with tagging accounting for just 12% and voting for only 6%.

Questions Attempted & Correctly Answered

In terms of overall quiz performance, Figure 1(a & b) presents box-plots of the median number of questions attempted and answered correctly *per user* across the different stak sizes; note that for clarity we have grouped the results obtained for the 6 solitary staks and reported the aggregate information as a single solitary stak, indicated as the stak of size 1. These results point to the benefit of sharing and collaboration during this search task. For example, we see that the single-users of the 6 solitary staks attempt a median of 3.5 questions but answer only 3.0 of these questions correctly. By comparison, the median values across shared staks are between 5.5 and 8 questions attempted per user and between 4 to 7 questions correctly answered per user. Overall, there is not a strong correlation between the above measures of performance and stak size. In the 9-person stak, for example, more questions are answered correctly (7) than any of the other shared staks, even compared to much larger 19- and 25-person staks. It is likely that the search expertise of individual users is playing an important role here. As such, a simple measure such as stak size is unlikely to be a powerful predictor of overall performance given the variation in expertise that likely exists between the individual members of a stak. Moreover, the closed-world nature of this trial — staks are limited by people and by topic to a 20-question quiz — limits the value of increasingly large staks, at least beyond some minimum critical mass.

Search Effort

The above results point to better performance for the collaborating searchers compared to solitary searchers. Our key hypothesis is that this is due, at least in part, to the type of search collaboration that HeyStaks facilitates. For example, looking at the level of granular search activity across search staks, we note that solitary searchers generally expend more effort in terms of the number of queries submitted compared to members of shared staks; 52 queries per searcher for the solitary staks versus 39–46 queries per searcher for the shared staks. In other words, solitary searchers are found to submit 13%–33% more queries than their counterparts in shared staks. Moreover, when the number of ac-



(a) Questions Attempted (b) Questions Correct

(c) Activities per query (d) Questions Correct per Query

Figure 1. Boxplots illustrating performance metrics per user, per stak.

tivities registered by solitary and collaborating users is examined — as a preliminary indicator of result relevance — we find that the former have a median of 23 activities (selections, tags etc) across these queries compared to 28–40 activities for the members of the shared staks; this is a relative increase of 22%–74% in favour of the shared staks.

Combining the above findings to look at the median number of activities per query per user across the staks, as per Figure 1(c), can be viewed as a proxy for the relevance (via number of activities) *per unit search effort* (number of queries submitted). We can see a significant difference between the activities per query for the solitary searchers (approximately 0.4 activities per query) and the collaborating searchers from shared staks (approximately 0.6–0.8 activities per query). In other words, 1.5 to 2-times as many queries lead to some form of activity among the collaborating searchers compared to the solitary searchers, suggesting that the former are benefitting significantly from results that are, on the surface at least, more relevant than those experienced by the latter.

A more pragmatic metric of relevance per unit search effort can be calculated by examining the number of correct answers per query per user across the various staks. This is presented in Figure 1(d) and once again we can see a very significant difference between the solitary searchers and the users who are members of shared staks. Solitary searchers correctly answer only 0.04 questions per query compared to up to 0.15 for the collaborating searchers in the 9-person stak. In other words, on a per query basis the collaborating searchers in the 9-person stak are answering more than 3 times as many questions correctly than the solitary searchers, which is a very significant productivity-gain for the members of this shared stak. Similarly, substantial productivity gains are also seen for the other shared staks.

Recommendation Relevance

Given that members of shared staks seem to be enjoying greater search productivity compared to their solitary counterparts, we now consider the likely source of this improvement: the recommendations that are generated by HeyStaks.

Members of shared staks benefitted from more HeyStaks recommendations than solitary searchers. Across shared staks, for example, 40 - 62% of queries lead to recommendations, compared to only 16% of solitary searchers' queries (all of those were resulting from users submitting queries similar to ones they had used previously). However, making lots of recommendations is not the real goal of HeyStaks. Ultimately, success depends on how relevant the recommendations are and, in particular, whether they are more relevant than the top-ranked Google results. To evaluate this we focus on search results that ultimately received user attention (selections, tags etc). There are 724 of these results and, as mentioned previously, we manually categorised each as *relevant*, *partially relevant* or *not relevant*. Figure 2 shows the percentage of activities on these results that are at least partially relevant (*Rel. + Part. Rel.*) and not relevant (*Not Rel.*) for both the default organic results and the HeyStaks recommendations across the shared staks. Here, we exclude single-person staks as we wish to examine the effects of result-sharing, rather than simply result recovery.

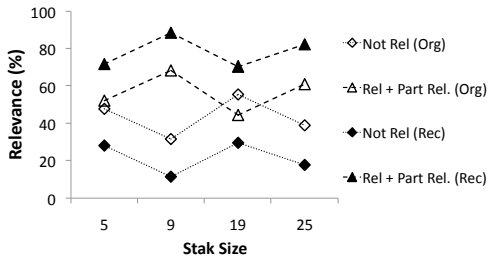


Figure 2. The relevance of organic (Org) and recommended (Rec) results acted on per stak.

Comparing recommended results versus organic results, we can see a significant relevance benefit for the former. For example, an average of 78% of recommended result activities (averaged across the 4 shared staks) are deemed to be at least partially relevant compared to only 57% for the organic results; in other words, the recommendations that attract user activity are significantly more likely to be relevant than the organic results that attract user activity.

To better quantify this relevance benefit we can compute a *relevance ratio* for organic and recommended results as the ratio of relevant and partially relevant results to not relevant results. Thus, a relevance ratio of less than 1 means that the majority of results are irrelevant, whereas a relevance ratio of more than 1 means that the majority of results are relevant. Figure 3 presents the relevance ratios of organic and recommended results for each stak. For all staks we can see that the recommended results have a much higher relevance ratio than the default organic results. For example, in the case of the 5-person stak, the organic results have a relevance ratio of 1.1, while the relevance ratio for the recommended results in this stak is more than twice as high, at 2.5.

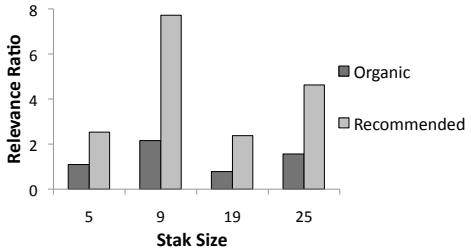


Figure 3. Relevance ratios of organic and recommended results per stak.

The 9-person stak does especially well by this evaluation measure. Although not the largest stak, this stak is the best performer (e.g. more questions answered correctly per user), most likely because its members are better searchers to begin with. The relevance ratio of its recommendations is 7.7, meaning those made are of very high quality. But it is interesting to note that for this stak the relevance ratio of the organic results is also relatively high, at 2.2, at least in comparison to the relevance ratio of organic results in the other staks. This further supports the notion that members of the 9-person stak are better searchers on average compared to members of other staks. And of course if the organic results are more relevant to begin with, then this will ultimately translate into superior recommendations for this stak because these more-relevant organic results ultimately become recommendations as they are acted on by users.

CONCLUSION

This paper discusses the HeyStaks social search service, which is designed to support collaboration during mainstream web search. In particular, we have described the results of a new live-user evaluation of HeyStaks. Our findings show how HeyStaks brings significant benefits to collaborating searchers. Of course this evaluation is limited in scale, which necessarily limits the diversity of staks that can be accommodated. In addition, the search task is similarly limited to a preset list of 20 quiz questions. As such the trial focuses one class of common web search queries, fact-finding queries. Nevertheless, these limitations afforded us an opportunity to take a close look at individual search activities and the actual relevance of search results. Moreover, this study complements other recent studies [3] which focused on more open-ended search tasks but did not support a detailed relevance analysis.

ACKNOWLEDGEMENTS

Based on works supported by Science Foundation Ireland, Grant No. 07/CE/I1147 and HeyStaks Technologies Ltd.

REFERENCES

1. S. Amershi and M. R. Morris. CoSearch: A system for co-located collaborative web search. In *ACM CHI Conference on Human Factors in Computing Systems*, pages 1647–1656. ACM, 2008.
2. B. M. Evans and E. H. Chi. An elaborated model of social search. *Information Processing and Management*, 46(6):656–678. Pergamon Press, Inc., 2010.
3. K. McNally, M. P. O'Mahony, B. Smyth, M. Coyle, and P. Briggs. Towards a reputation-based model of social web search. In *International Conference on Intelligent User Interfaces*, pages 179–188. ACM, 2010.
4. M. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why? A survey study of status message Q and A behavior. In *ACM CHI Conference on Human Factors in Computing Systems*, pages 1739–1748. ACM, 2010.
5. M. R. Morris and E. Horvitz. SearchTogether: An interface for collaborative web search. In *User Interface Software and Technology*, pages 3–12. ACM, 2007.
6. M. C. Reddy and P. R. Spence. Collaborative information seeking: A field study of a multidisciplinary patient care team. *Information Processsing and Management*, 44(1):242–255. Pergamon Press, Inc., 2008.
7. B. Smyth. A community-based approach to personalizing web search. *IEEE Computer*, 40(8):42–50. IEEE Computer Society Press, 2007.
8. B. Smyth, P. Briggs, M. Coyle, and M. P. O'Mahony. Google? Shared! A case-study in social search. In *User Modeling, Adaptation and Personalization*, pages 283–294. Springer-Verlag, June 2009.