Synchronous Collaborative Information Retrieval: Techniques and Evaluation

Colum Foley and Alan F. Smeaton

CLARITY: Centre for Sensor Web Technologies
Dublin City University
Glasnevin, Dublin 9, Ireland.
colum.foley@computing.dcu.ie

Abstract. Synchronous Collaborative Information Retrieval refers to systems that support multiple users searching together at the same time in order to satisfy a shared information need. To date most SCIR systems have focussed on providing various awareness tools in order to enable collaborating users to coordinate the search task. However, requiring users to both search and coordinate the group activity may prove too demanding. On the other hand without effective coordination policies the group search may not be effective. In this paper we propose and evaluate novel system-mediated techniques for coordinating a group search. These techniques allow for an effective *division of labour* across the group whereby each group member can explore a subset of the search space. We also propose and evaluate techniques to support automated *sharing of knowledge* across searchers in SCIR, through novel *collaborative* and *complementary* relevance feedback techniques. In order to evaluate these techniques, we propose a framework for SCIR evaluation based on simulations. To populate these simulations we extract data from TREC interactive search logs. This work represent the first simulations of SCIR to date and the first such use of this TREC data.

1 Introduction

Collaborative information retrieval (CIR) is a phrase which refers to the user-user collaboration which can occur in the information retrieval (IR) process. The vast majority of work to date in this area has concentrated on leveraging the past experiences of users to benefit a new user coming to the system. For example, recommender systems filter items for users based on the recommendations of other users, collaborative footprinting systems allow users to see the trails left by others through an information space and social search engines re-rank query results based on the viewing history of like-minded users. These CIR systems are characterised by an asynchronous, implicit collaboration. The purpose of these systems is to improve the IR process for an individual searcher.

Recently we have begun to see the emergence of a more explicit, engaging collaborative IR experience which we refer to as Synchronous Collaborative Information Retrieval (SCIR). These systems attempt to improve the performance

of a *group* of users who are searching together at the same time in order to satisfy the same, shared information need. As such, these systems represent a significant departure from how we view the IR process, from a single-user to a group perspective. SCIR is an emerging research domain which is gaining pace. SCIR can occur either remotely, where two users communicate across the internet, or in a co-located manner and the development and adoption of such systems is facilitated by developments in both environments. Most early work in the area focussed on improving the *awareness* across a distributed group of collaborating searchers [1]. These systems provided web browsers that were embellished with chat windows, shared whiteboards for brainstorming, and shared bookmark areas where group members could save documents of relevance to the search task, thereby bringing them to the attention of their collaborators. More recently we have seen the development of systems that support co-located SCIR [2]. Bringing people together to search increases awareness across the group as users see what their collaborators are doing.

In both remote and co-located domains, the commonality across systems is their efforts to improve awareness across the collaborating searchers, the motivation being that when users are more aware of their partners' actions they can coordinate the group activity themselves. For example, if a user can see the query terms entered by their search partner they may decide to enter different terms, or a user may decide not to spend time reading a document if it is in the shared bookmark folder. As observed by [3], however, requiring users to both search and coordinate a group activity can be troublesome and distracting, requiring too much of a user's cognitive load to switch between the two tasks. [4] proposed an "algorithmic-mediated" SCIR system which allowed users to work together in a co-located setting under predefined roles, where the system would coordinate the activity across the users. Such a division of users into predefined roles, however, may not be an ideal model for adhoc search common in web searching as some form of user-user coordination is required in order to assign roles.

In this paper we propose system-mediated techniques for adhoc SCIR search, for either a co-located or remote setting, which do not require any user-user coordination during the search task. In order to evaluate the effects of these techniques, we also propose a novel evaluation framework based on simulations of an SCIR task. These simulations are populated with data from previous TREC interactive experiments.

The rest of this paper is organised as follows. In section 2 we will outline our proposed system-mediated techniques for SCIR, namely division of labour and sharing of knowledge. In section 3 we will outline our proposed evaluation methodology for SCIR. In section 4 we present the results from our experiments and finally in section 5 we outline our conclusions.

2 System-Mediated Techniques for SCIR

2.1 Division of Labour

Allowing multiple people to search together at the same time in order to satisfy the same information need can allow the search task to be divided across the users, enabling each user to explore distinct subsets of the collection. As users are searching in order to satisfy the same shared information need, however, unless some form of coordination is provided for them there may be duplication of effort across the users. When searching to satisfy the same information need users often use the same query terms [5], resulting in similar ranked lists being returned to all users which in-turn can cause users to spend time viewing the same documents.

As discussed earlier, a user driven coordination approach may not be the most effective for SCIR search, due to users suffering from cognitive overload. SearchTogether [6] is an example of a state-of-the-art SCIR system which provides support for a simple system-mediated division of labour through its split-search facility, which allows a user's query to be split by the search engine in a round-robin manner across users. However, the coordination of an entire SCIR session may be problematic with such a system. In particular, if one user decides to issue another search, it is not clear how to coordinate this search. Should the results be split again ? Or should the user ask permission first before providing results to their search partner ? By splitting the results again, the user who receives the list is expected to move their attention onto another ranked list, and as the number of independent search results increases this may lead to users becoming overwhelmed with results. On the other hand, coordinating the activity through a chat facility may also be too demanding of users.

We propose a simpler solution which allows users to work more independently whilst the system coordinates an effective division of the search task. At any point in the search, each collaborating searcher will have viewed a number of documents and may be examining a ranked list. An SCIR system can use this information in order to implement a division of labour policy that removes from a user's ranked list:

1. Documents that have already been seen by another user.
2. Documents contained on other user's current ranked list.

By maintaining a list of all documents seen by each searcher during a search, an SCIR system can implement 1 by ensuring that documents seen by one co-searcher are never returned to another. In order to implement 2, the SCIR system needs to decide on the number of documents to assume that a user will examine on their list. This number could correspond to the number of documents being presented on the user's screen. For example, a web search on a standard PC screen would typically return 10 documents, on a large tabletop display, this number could be 30 or 40, while on a mobile phone it could be 5.

Implementing a division of labour policy can improve the performance of SCIR search by replacing redundant documents in a user's list with new material, enabling the group as a whole to view more documents during the search.

2.2 Sharing of Knowledge

A common feature of many state-of-the-art SCIR systems is their use of a shared bookmark facility into which users can save documents they feel are relevant to the search. These bookmarks represent *explicit* relevance judgments from users. Relevance feedback (RF) is a technique used in traditional, single-user, IR to reformulate a user's query in the light of relevance information. Foley et al. [**?**] outlined how the traditional RF process can be transformed into a *collaborative relevance feedback* process, whereby each user's relevance information is combined in the RF process. Such a technique can allow for an implicit *sharing of knowledge* across users collaborating in an SCIR search, as users can benefit from the explicit relevance judgments of their co-searchers in their ranked lists. One approach proposed in [**?**] and shown in equation 1, is to extend the traditional Robertson Spärck-Jones (RSJ) probabilistic relevance weighting formula [7] into a *partial-user* relevance weighting formula. In this approach, the relevance statistics used in the RSJ formula are extended so that the proportions for relevance and non-relevance are composed of a weighted combination of each collaborating searcher's relevance statistics based on their relevance judgments (see [**?**] for a detailed derivation):

$$purw(i) = \log \frac{(\sum_{u=0}^{U-1} \alpha_u \frac{r_{ui}}{R_u})(1 - \sum_{u=0}^{U-1} \alpha_u \frac{n_i - r_{ui}}{N - R_u})}{(\sum_{u=0}^{U-1} \alpha_u \frac{n_i - r_{ui}}{N - R_u})(1 - \sum_{u=0}^{U-1} \alpha_u \frac{r_{ui}}{R_u})} \quad (1)$$

Here n_i refers to the number of documents in the collection in which term i occurs, N refers to the number of documents in the collection, r_{ui} refers to the number of relevant documents identified by user u in which term i occurs, and R_u refers to the number of relevant documents identified by user u. α_u determines the impact of user u's proportions on the final term weight, with $\sum_{u=0}^{U-1} \alpha_u = 1$ Foley et al. [**?**] proposed techniques to extend Robertson's offer weight in a similar manner.

In this section we will extend this work in two ways. Firstly we will examine the application of a user-biased *authority weighting* scheme to the collaborative RF formulae. Following that we will propose a technique for using relevance judgments in SCIR through a novel *complementary relevance feedback* process.

Authority Weighting When multiple users search together, each user may have different levels of expertise with the search task. Poor relevance judgments, unless recognised and dealt with, may pollute an RF process which attempts to combine relevance information from multiple users. The collaborative RF techniques outlined in [**?**] allow for a biasing of each user's relevance statistics. Referring to equation 1 above, this can be achieved by adjusting the α_u value associated with each user. Using an authority weighting mechanism we can exploit this weighted combination in order to favour the RF documents of more authoritative users. There are several ways in which this authority weight can be assigned. For example, if a topic expert is searching with a novice, the users themselves may decide on the biasing prior to searching. The weight could also

be calculated and assigned dynamically each time RF is performed during the search, based on the estimated quality of each user's relevance judgments. In section 4.2 we investigate how an authority weighting scheme performs against a unbiased method when we simulate users making poor relevance judgments.

Complementary Relevance Feedback Foley et al. in [**?**] observed that a collaborative RF process, such as outlined above, can cause collaborating users' reformulated queries to become so similar that diversity is lost across their ranked lists. Another method of utilising relevance information in an SCIR environment is through a *complementary* RF process. Figure 1 provides a conceptual overview of the two techniques. Unlike a collaborative RF process which attempts to aggregate users' relevance information, a complementary RF process will try to increase the diversity across collaborating users' RF processes. There are several

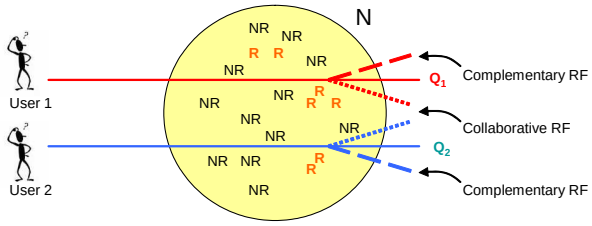

Fig. 1. Comparison of collaborative and complementary relevance feedback process

methods by which an RF process can be extended in order to allow collaborating users' relevance feedback information to complement their partners'. A simple approach we have developed, referred to as *complementary query expansion*, removes, from a user's query, any query expansion terms that appear in their search partners' queries. Another approach we have developed is clustering the set of RF documents found and terms from them, using k-means clustering, into k partitions where k is the number of collaborating searchers. In section 4.2 we investigate the effects of such techniques in SCIR.

3 Evaluation Methodology for SCIR

In our work, we have developed a novel framework for evaluating SCIR based on simulations. Simulations have been used previously in interactive IR evaluation in an attempt to model a user's interactions with an IR system [8]. However to-date no simulations have attempted to model an SCIR environment where two or more users collaborate to search for information.

Our simulations are populated with data extracted from the user interaction logs of the TREC 6 to TREC 8 interactive track experiments. Groups submitting

runs for evaluation to these tracks were also required to submit so-called rich-format data along with their submissions. For each interactive search session this data recorded significant user interactions during the search task such as queries, documents viewed, and relevance judgments made along with timing data associated with each of these actions. Originally these users would have completed these topics separately, in our work we simulate these users searching together synchronously. In our work we are interested in evaluating the effects of our proposed system-mediated techniques to coordinate the search activity, therefore in our simulations we assume that users do not communicate during the search and that coordination is performed in the back-end. Our SCIR simulations comprises *two* collaborating searchers as a recent study on the collaborative nature of search have shown a group size of two to be the most popular size [6]. In order to simulate users searching together we synchronise the timing data by aligning the time at the start of the original session and use the timing offset information to interleave the significant events. In total we extracted rich-format data from 10 different experimental systems across TREC 6 to TREC 8. This resulted in a total of 591 paired user simulations across 20 search topics. This data set provides a rich and diverse range of systems and users on which to evaluate the performance of our techniques.

3.1 Simulated SCIR System Type

An important consideration for SCIR simulations is deciding on how to initiate the SCIR search. Should we assume that each user enters their own query ? Or should we assume that one shared query is issued between the group ? Our simulations assume the former, and begin with one shared query across the group. When users search for the same information need they often use the same, or very similar, query terms and therefore any benefit gained, in terms of diversity introduced through multiple queries, may be minimal [5]. Further to this point, in our simulations users do not manually reformulate the shared query during the search; rather we assume that users receive new ranked lists automatically through an RF process. In particular, we have implemented an *incremental relevance feedback* environment [9] in which users receive a new ranked list after *each* relevance judgment is made. Figure 2 provides an overview of our SCIR simulations as described. As we can see, the search session begins with one shared group query. In order to construct this query, we concatenate the unique terms from the users' initial queries as extracted from the TREC rich-format data. The search then proceeds with users making relevance judgments where the order of these judgments is based on the timing offsets from the TREC logs. As each user makes a relevance judgment, the ranked list returned is influenced by both the division of labour policy and sharing of knowledge policy implemented.

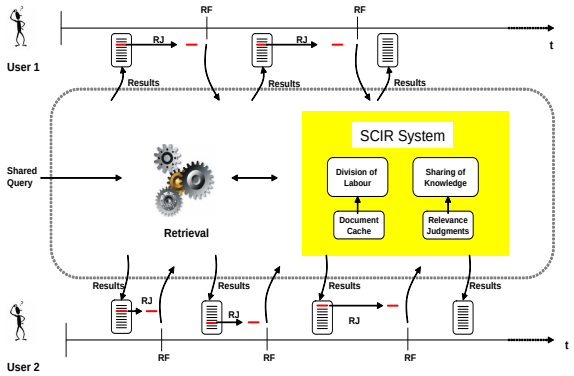

Fig. 2. Simulated session overview

3.2 Dynamic Relevance Judgments

Before we can finalise our simulations we need to decide on how to simulate a user making relevance judgments. The interaction logs used to populate our simulations recorded the original documents saved by users during the search. It would not have been appropriate, however, to use these same documents in our simulations given that we are implementing our own search system and therefore cannot assume that users would have saved the same documents. Instead, we replaced these static relevance judgments with dynamic relevance judgments from the ranked lists being presented to simulated searchers. In our experiments we model two types of relevance feedback environments. A best-case RF environment in which we assume users always make relevance judgments on relevant documents and an RF environment in which we model users making mistakes in their judgments.

To model the best-case, we simulate a user looking down through their ranked list and making a relevance judgment on the first relevant document encountered. Considering that searchers tend to examine a ranked list from top to bottom [10] we feel that this approximation is reasonable. In order to simulate an environment in which users can make mistakes in their relevance judgments, our approach is to build a pool of *perceived relevant* documents, where this pool consists of non-relevant documents (according to the TREC qrels) that were saved by at least two real users during the original TREC experiments from which we extracted our simulation data. These perceived relevant documents represent documents that users could realistically mistake for relevant documents. The simulation then proceeds as before with a user looking through the ranked list and marking as relevant the first relevant or perceived relevant document in the list, whichever comes first. In these experiments we limit the number of documents that a simulated user will examine to the top 30 documents in their ranked list.

3.3 Evaluation Metric

IR is generally evaluated in terms of the quality of a ranked list, where this quality can be measured using standard metrics such as average precision (AP). The novel domain of SCIR presents challenges in terms of developing appropriate metrics. Obviously, rather than having one list to evaluate (as in traditional IR), at any point in an SCIR search there are several ranked lists to evaluate, one for each user. One potential method for SCIR evaluation would be to take a standard IR measure such as AP and average across each user's ranked list. Unfortunately this approach makes no attempt to determine the overlap of documents across users' lists. For example, if two separate collaborating groups of users had the same averaged AP score, but the members of the first group had ranked lists which contained many of the same documents, while the second group had ranked lists with a greater diversity of relevant documents, then the performance of the second collaborating group should be considered better than the first as, across the group, the total amount of relevant material found across collaborating users' lists is greater in the second group. By simply averaging each ranked list's AP score, however, this information would be lost.

What we need instead is a measure which captures the quality and diversity across collaborating users' ranked lists. Our solution is to count the *total number of unique relevant documents across user's ranked lists* at a certain cutoff. In our simulations, we set this cutoff at 30 documents. We use this evaluation metric in order to produce two different views on the experimental results. We produce a plot of this figure over the entire search, which allows us to show how the figure changes after each relevance judgment is made in the search. We also produce a single figure performance measure for the entire group search which we calculate by averaging this group score metric across all RF iterations over the search. This single figure is then used in order to run significance testing. In our experiments we use randomisation testing to test for statistical significance and use a significance threshold of $p < 0.05$.

In this section we have described the process of evaluating SCIR using simulations populated with rich-format data extracted from TREC submissions, for a more in-depth description of the process the reader is referred to [5].

4 Experimental Results

4.1 Division of Labour

We experimented with three different variants of a division of labour policy as shown in Figure 3. The first is one in which no attempt is made to divide the search task (*No Division*), another which removes those documents seen by others (*Docs Seen Removed*), and a final one that removes both those documents seen by others and those contained on a collaborator's current ranked list (*Full Div*). Alongside our comparisons of the performance of these SCIR systems, we also compare the performance of these collaborative systems with two baseline systems showing users searching independently without any collaboration in

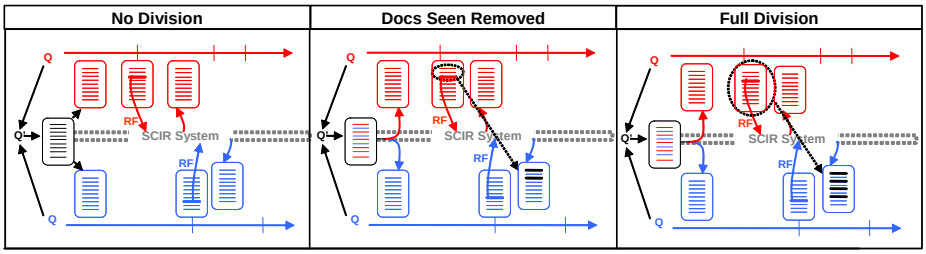

Fig. 3. Division of labour experimental systems

terms of division of labour. The *Independent Group* baseline evaluates how the group of users perform without any collaboration in terms of the initial query or dividing of search results. The *Best Individual* baseline shows how, for each pair of users searching, the best user performs when searching on their own, using their own initial query and the incremental feedback system. The results

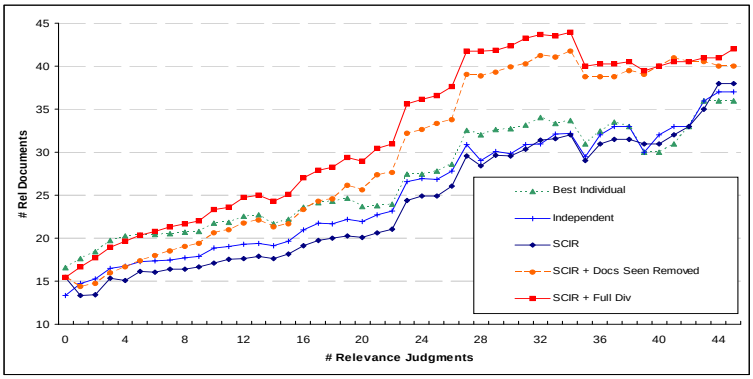

Fig. 4. Division of labour experimental results

from these experiments are presented in Figure 4, where we can see that by implementing a full division of labour policy we can improve the performance of SCIR substantially. Clear improvements are gained as the level of division is increased. The full division system is significantly better than both baseline systems. Significance tests also reveal that the SCIR system without any division is significantly worse than the baseline of users searching independently.

4.2 Sharing of Knowledge

Authority Weighting We now explore the effectiveness of an authority weighting scheme operating in an environment where users can make mistakes in their relevance judgments. In these experiments we do not attempt to estimate which

user's relevance judgments are better. We are interested in exploring the potential usefulness of such information and therefore we use an oracle to determine which user's relevance judgments are better at any point in the search. In order to develop this oracle, for each search topic we calculate the relevance weight associated with each term from all relevant documents for that topic as extracted from the qrels, i.e. we run one batch RF process using all relevant documents. In order to calculate which user of the two is the more *authoritative* at any point in the simulated search, we calculate the relevance weights of all terms for each user using their own relevance judgments. We then calculate the correlation between this weighting and the oracle relevance weighting of terms. The user with the higher correlation value is considered the more authoritative.

Having decided on which user is the more authoritative we then need to bias the RF process in their favour, by changing the α value associated with this user in the collaborative RF formulae. In these experiments we have investigated two techniques. Using the *static* weighting scheme, the authority value assigned to the more authoritative user is decided *a priori*, and is then applied to this user's relevance information when performing feedback. Although this scheme does allow the recipient of the authority bias to change mid-search, the amount of weighting which occurs remains static throughout the search. We have therefore also experimented with a *dynamic* authority scheme which proportions the authority value based on the differences in the correlation figures returned from the oracle. For the static runs, we experimented with values of 0.6 - 1 for the authoritative user's α value, and for completeness we also experimented with an inverted authority weight (i.e. assigning the higher α value to the poorer user).

Table 1. Authority weighting experiment results

	Unbiased	**Static Authority Weight**										**Dynamic**
	Combo RF	**0.6**	**0.7**	**0.8**	**0.9**	**1**	**0.4**	**0.3**	**0.2**	**0.1**	**0**	**Authority Weight**
Average Per Topic	20.14	20.21	20.17	20.09	19.98	19.88	10.19	9.39	8.78	8.66	8.62	20.24

Table 1 presents the results from our authority experiments. As we can see the static runs peak at an authority value of 0.6. Not surprisingly the inverted runs perform poorly. The dynamic weighting scheme is the best performer overall providing a significant improvement over an unbiased collaborative RF approach.

Complementary Relevance Feedback One way of maintaining diversity across users through the RF process is by ensuring that the expansion terms assigned to each user are unique through a *complementary query expansion* technique. In Figure 5 we compare the performance of the SCIR system with just a division of labour policy (SCIR + Full Div), with an SCIR system implementing a division of labour policy and a complementary query expansion process. As we can see, the complementary expansion approach performs worse than the SCIR with full division. Running significance tests over the associated single figure performance measure confirms this result to be significant across topics. As Fig-

ure 5 shows, the complementary query expansion technique indeed introduces more unique documents into user's ranked lists, but due to the poor performance of the technique, this diversity is obviously being achieved at a cost of a significant degradation in the quality of user's lists. Another, more sophisticated form

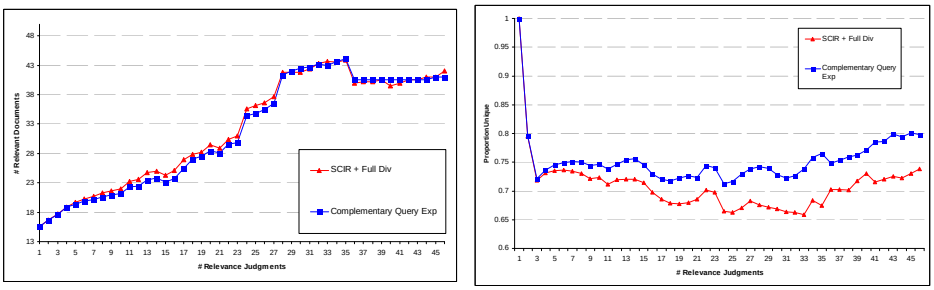

Fig. 5. Comparison of SCIR + Full Div and complementary query expansion in terms of performance (left) and uniqueness across users' ranked lists (right)

of complementary RF we have developed is through the use of clustering. In our work we used the k-means clustering algorithm in order to cluster: (1) the set of relevant documents found by the group so far, and (2) the terms contained within these documents, into two distinct clusters, one for each user, prior to performing feedback. The motivation for both techniques is that by partitioning either the document or term space into two, we should generate more distinct RF queries than is produced by the collaborative RF technique, while producing better quality queries than those produced by a simple complementary query expansion technique. Comparing the single figure performance measure, however, we found that neither the document clustering (*20.74*) or term clustering (*20.40*) technique performs as well as the partial-user collaborative RF technique (*20.95*).

5 Conclusions

In this paper we explored the effects of system-mediated techniques for SCIR. We also proposed and developed an evaluation framework for rapid experimentation in SCIR based on simulations of a collaborative search session, the first examples of SCIR simulations to date. Our experiments have shown that a system-mediated division of labour in an SCIR search can significantly improve the performance of the group search. Furthermore our results show that the quality of SCIR search without such a policy can be worse than a group of users searching independently. We extended the work in [**?**] by investigating the effects of an authority weighting scheme on the performance of a collaborative RF process operating in an environment in which users can make mistakes in their relevance assessments and proved its effectiveness. We proposed two techniques

for complementary relevance feedback which attempt to introduce diversity into a relevance feedback process operating in an SCIR environment. Although both techniques introduced more diversity across users' ranked lists, these techniques failed to improve over a collaborative RF process.

These simulated experiments have enabled us to explore many aspects of SCIR search inexpensively, however, in order to fully evaluate the effects of these techniques it will be necessary to evaluate them in the context of an interactive collaborative search involving real users, and this we leave for future work.

We believe SCIR will become more important as people continue to use computers more collaboratively, as such we believe the work presented here represents an important initial contribution to the development of effective SCIR systems.

Acknowledgements

This work was partly supported by the Irish Research Council for Science, Engineering and Technology and by Science Foundation Ireland under grant numbers 03/IN.3/I361 and 07/CE/I1147.

References

1. Gianoutsos, S., Grundy, J.: Collaborative work with the World Wide Web: adding CSCW support to a Web browser. In: Procedingss of Oz-CSCW'96, DSTC Technical Workshop Series, University of Queensland, Brisbane, Australia (1996)
2. Smeaton, A.F., Lee, H., Foley, C., Mc Givney., S.: Collaborative Video Searching on a Tabletop. Multimedia Systems Journal **12**(4) (2006) 375–391
3. Adcock, J., Pickens, J., Cooper, M., Anthony, L., Chen, F., Qvarfordt, P.: FXPAL Interactive Search Experiments for TRECVID 2007. In: TRECVid2007 - Text REtrieval Conference TRECVID Workshop, Gaithersburg, MD, USA. (2007)
4. Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., Back, M.: Algorithmic mediation for collaborative exploratory search. In: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, Singapore, ACM Press (2008) 315–322
5. Foley, C.: Division of Labour and Sharing of Knowledge for Synchronous Collaborative Information Retrieval. PhD thesis, School of Computing, Dublin City University, Dublin, Ireland (2008)
6. Morris, M.R., Horvitz, E.: SearchTogether: an interface for collaborative web search. In: UIST '07: Proceedings of the 20th annual ACM symposium on User interface software and technology, Newport, Rhode Island, USA, ACM Press (2007) 3–12
7. Foley, C., Smeaton, A.F., Jones, G.J.F.: Combining Relevance Information in a Synchronous Collaborative Information Retrieval Environment. In: Collaborative and Social Information Retrieval and Access Techniques for Improved User Modelling. IGI Global (2008)
8. Robertson, S.E., Spärck Jones, K.: Relevance weighting of search terms. Journal of the American Society for Information Science **27**(3) (1976) 129–146

9. White, R.W., Ruthven, I., Jose, J.M., Rijsbergen, C.J.V.: Evaluating implicit feedback models using searcher simulations. ACM Trans. Inf. Syst. **23**(3) (2005) 325–361
10. Aalbersberg, I.J.: Incremental relevance feedback. In: SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, Copenhagen, Denmark, ACM Press (1992) 11–22
11. Craswell, N., Zoeter, O., Taylor, M., Ramsey, B.: An experimental comparison of click position-bias models. In: WSDM '08: Proceedings of the international conference on Web search and web data mining, Palo Alto, California, USA, ACM Press (2008) 87–94