Revisiting IR Techniques for Collaborative Search Strategies

Hideo Joho, David Hannah, and Joemon M. Jose

Department of Computing Science, University of Glasgow
Sir Alwyn Williams Building, Lilybank Gardens, Glasgow G12 8QQ, UK.
{hideo,davidh,jj}@dcs.gla.ac.uk

Abstract. This paper revisits some of the established Information Retrieval (IR) techniques to investigate effective collaborative search strategies. We devised eight search strategies that divided labour and shared knowledge in teams using relevance feedback and clustering. We evaluated the performance of strategies with a user simulation enhanced by a query-pooling method. Our results show that relevance feedback is successful at formulating effective collaborative strategies while further effort is needed for clustering. We also measured the extent to which additional members improved the performance and an effect of search progress on the improvement.

1 Introduction

There has been a growing interest in the development of collaborative search technologies in Information Retrieval (IR). A fundamental issue of collaborative search is that existing IR models are not designed to be aware of collaboration. There seems to be at least three interweaving conceptual approaches to addressing this issue: Models, Techniques, and Interfaces. The first level concerns the development of new IR models that can take collaboration into account in retrieval. The second level aims to leverage IR techniques such as relevance feedback, clustering, profiling, and data fusion to support collaborative search while using conventional IR models. The third level is to develop search interfaces that allow people to perform search tasks in collaboration.

This paper addresses the second level (i.e., Technique) by revisiting some of the established IR techniques to formulate collaborative search strategies. We focus on synchronous collaborative search where a team performs search tasks together. We devised eight collaborative search strategies that aimed to divide the labour and share knowledge in a team. While there are many types of collaborative search tasks, we were particularly interested in recall-oriented tasks. An example of collaborative search in a recall-oriented task is found in an information intensive domain [1] such as the intellectual property (IP) community. When companies consider an investment of a new product or technology, they assign a team of searchers to survey the IP coverage of existing patents [2]. This is a highly exhaustive task since the cost of patent infringement can be devastating. A high level of efficiency is also crucial in competitive markets. Their work task motivated us to investigate the effectiveness of collaborative search strategies in a recall-oriented task. It should also be noted that work tasks of other professionals such as doctors, academics, and lawyers are often recall-oriented and collaborative.

The contributions of this paper is as follows. First, we evaluated eight collaborative search strategies based on user simulation. Second, we discussed the utility of IR techniques from an application perspective. Finally, we provided a use case of a query-pooling method for user simulation. The rest of the paper is structured as follows. Section 2 reviews existing approaches to collaborative search. Section 3 presents the research questions being addressed and experimental design of our study. Section 4 presents the results of our experiments. Section 5 discusses the implications of the results on the design of effective collaborative search strategies. Finally, Section 6 concludes the paper with future directions.

2 Approaches to collaborative search

A categorisation of collaborative work was proposed by [3] based on two dimensions: time and space. For example, face to face interactions share both time and space while coordinations via emails share neither of them. Continuous tasks share space but not time, while remote interactions share time but not space. This categorisation applies to the existing approaches to collaborative search. For example, the users of the I-SPY system [4] were not necessarily sharing time nor location, but their click through information was exploited to re-rank the documents retrieved by a similar query. The users of a table-top based interface on the other hand shared time and location to complete a task [5]. Others [6, 7] assume one of either time or location to be shared in their use.

Much research and development in this area has been on the interface level. For example, SearchTogther [7] was designed to facilitate sharing knowledge and communication during collaborative search tasks. Smeaton, et al. [5] measured the performance of collaborative tasks using a tangible search interface. A collaborative interface was developed by Villa, et al. [8] where users can monitor and interact with a team members' activity in video retrieval. Little work has been carried out on the Model level. An algorithmic mediation proposed by Pickens, et al. [6] was designed to optimise the weighting of queries based on relevance and freshness, determined by the analysis of collaborative activity of a team and influenced the ranking of candidate queries and retrieved documents.

Our work differs from these previous studies in terms of the focused level and experimental methodology. This paper concerns the Technique level to exploit some of the established IR techniques in supporting collaborative search. Also, we measure the performance of different search strategies based on user simulation, allowing us to investigate many different strategies. However, we address the common issues such as division of labour and sharing knowledge since they are important factors for successful collaborative work [3]. In particular, we investigate relevance feedback as a means of implicit sharing of knowledge about

Table 1. Collaborative Search Strategies

Code	Strategy
SS1	Team members performs search independently
SS2	SS1 with unjudged documents only
SS3	SS2 with independent relevance feedback
SS4	SS3 with shared relevance feedback
SS5	Team submits the same queries and divides the results with round-robin
SS6	SS5 with clustering for result division
SS8	SS4 and SS5 (Shared relevance feedback on round-robin division)
SS10	SS4 and SS6 (Shared relevance feedback on clustering division)

topical relevance, and clustering as a means of effective division of labour such as browsing and judging retrieved documents.

3 Experiment

This section first presents the research questions being addressed in this paper. Then, we discuss the experimental design of our study.

3.1 Research questions

The main research hypothesis proposed in this work is *IR techniques such as relevance feedback and clustering can be effective for supporting collaborative search.* We were also interested in an impact of team size on the performance of search. More specifically, we address the following research questions in this paper:

RQ_1 Is relevance feedback effective at accumulating and sharing knowledge of topical relevance among the team?
RQ_2 Is document grouping effective at dividing the labour of document browsing and relevance assessments among the team?
RQ_3 Are the two techniques complementary in collaborative search?
RQ_4 To what extent do we gain by adding extra members to a team?
RQ_5 To what extent is the gain of extra members affected by the progress of a search session?

3.2 Search Strategies

We devised eight search strategies to address the research questions and they are shown in Table 1 along with reference codes. To explain the behaviour of search strategies, we use a hypothetical component called the *Agent* who controls the flow of interaction between the system and searchers. In Search Strategy 1, or `SS1`, the Agent did nothing. Team members submitted a query independently and judged the top 20 retrieved documents to find relevant documents. If the same query was submitted, the same 20 documents were returned. In SS2, the Agent recorded the documents judged by the team and only returned non-judged documents in every query. We considered SS1 and SS2 as the baseline strategies.

In SS3 and SS4, the Agent performed query expansion based on relevance feedback. Individual profiles (i.e., a set of relevant documents found) were formulated for query expansion in SS3, while SS4 created a team profile where relevant documents found by all members was recorded. The former can be seen as an accumulation of topical relevance knowledge for team members, and the latter can be seen as implicit sharing of those accumulated knowledge among the team. In both strategies, when a query was submitted, an expanded query was generated based on judged relevant documents. When the submitted query terms were not found in the expanded queries, we added them. Otherwise, we gave the highest weight to the submitted query terms in the expanded query before submission to the system. This ensured that submitted query terms were emphasised in expanded queries.

In SS5 and SS6, the Agent submitted a common query for the team and grouped the retrieved documents. In our previous user study [9], browsing a different set of documents was employed as a frequent strategy of the division of labour in collaborative search. SS5 and SS6 simulated a case where this strategy was supported by grouping the retrieved documents. A round-robin approach was used in SS5 and the group-average clustering method was used in SS6. We used the group-average method because of its robust performance on retrieved documents [10]. In these strategies, we assumed that a list of candidate queries were formulated in advance and the Agent submitted the queries to the retrieval system. The top 300 retrieved documents were then divided by an underlying technique and distributed to each of the team members. The number of generated clusters was set to the team size. Since we used a hierarchical clustering method, some clusters were smaller than 20 documents (See Section 3.4 for this size). However this rarely occurred; SS6 had 1.3% fewer documents judged when compared to other strategies.

SS8 and SS10 were combinations of strategies. In SS8, the Agent performed query expansion based on shared relevance feedback and divided the retrieved documents using a round-robin approach. SS10 on the other hand performed query expansion but clustering was used to divide the retrieved documents.

3.3 Query pool

We decided to run a user simulation (as opposed to a user study) due to the variety of search strategies shown above. One way to run a user simulation is to use a test collection to simulate a user's query re/formulation, browsing, and relevance judgements [11–13]. A limitation of this approach is the lack of diversity in queries per topic. Often it uses only the title of the topic description as the single query of the topic. This is not ideal given that searches are often iterative in a recall-oriented task. Ruthven [14] applied a range of query expansion techniques to create a query pool which contained a large set of queries per topic.

Table 2. Topics and number of unique queries (N=993).

Topic	Query	Topic	Query	Topic	Query	Topic	Query	Topic	Query
303	87	367	124	397	76	625	107	689	19
344	72	383	9	439	151	651	131		
363	82	393	13	448	57	658	65		

Our approach was similar to Ruthven's but our query pool was derived from an actual user study. The study, referred as to the *original study* in this paper, had twelve pairs of searchers performing three TREC HARD Track topics both in an independent and collaborative conditions [9]. This generated a total of 1298 queries across 13 topics. The basic statistics of the original study are shown in Table 2. The different number of queries available was due to the take-up rate in the original study where participants were allowed to select three topics from 15 candidate topics based on their interest. The candidate topics were selected from the 50 TREC topics based on the number of relevant documents in the qrels. The number of relevant documents in the 50 TREC topics ranged from 9 to 376, from which we removed those topics with too few and too many relevant documents. As a result, we selected 15 topics with the range from 86 to 152 relevant documents. There was a difference from a conventional use of pooling in our design. We left duplications in the query pool. This allowed us to submit popular queries more frequently in simulation. The details of simulation is given next.

3.4 Simulation

For each strategy, we simulated 100 teams with varied size of one to five (i.e., 20 teams per size). Each team performed searches for 13 topics that lasted up to 20 iterations. In the original study, participants had on average 14.4 iterations to complete a recall-oriented task. Therefore, we considered that 20 iterations were sufficient to assess the performance of collaborative search strategies in our simulation. Queries were randomly selected from the query pool with replacement at every iteration. Each team member was assumed to judge 20 documents at every iteration. Therefore, a one person team would judge 400 documents while a five persons team would judge 2000 documents by the end of a session. The selection of 20 documents depended on the behaviour of search strategies. Like other simulation work, we assumed that searchers judged the relevance of documents as the TREC official judges did. In other words, all relevant documents appearing in the 20 documents were counted as the relevant documents found by a team member. We used the test collection of the TREC HARD Track 2005 [15] as in the original study. The track used the Aquaint collection which contains over 1 million documents (3GB) of news articles. The collection was indexed and retrieved by the Terrier system with the out-of-box setting [16]. As discussed above, we used 13 (out of 50) topics which were selected by the participants of the original study.

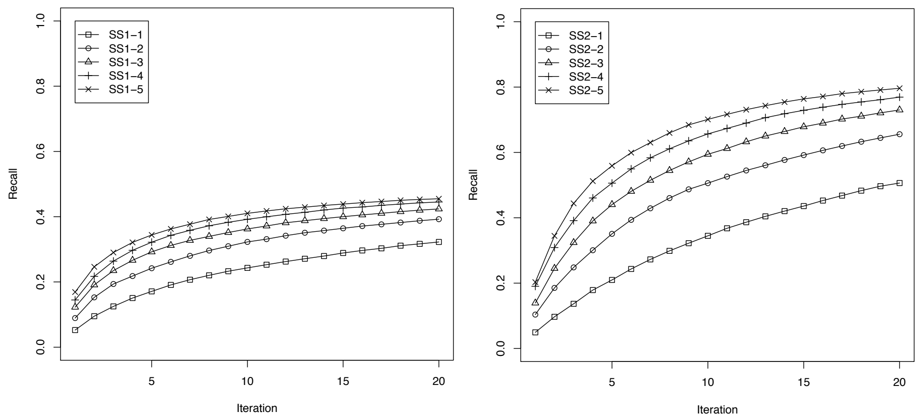

Fig. 1. Recall of baseline strategies: SS1 (Left) and SS2 (Right).

We used recall as the measure of the search performance throughout the experiment, since a recall-oriented task often performed by professionals was our research interest in this paper. Since we used the TREC official judgements, the results reported in the next section should be seen as an upper-bound of the performance of search strategies.

4 Results

This section presents the results of the experiments, structured to answer the research questions defined in Section 3.1. A code was used to represent a search strategy and team size. For example, `SS2-3` means Search Strategy 2 performed by a 3 person team. Each data point in the figures is a mean of 260 samples (i.e., 13 topics by 20 teams) throughout the section unless otherwise stated. The standard deviation varied but was consistently low (less than half of a mean value), and thus, not reported unless appropriate.

We first looked at the performance of two baseline strategies (`SS1` and `SS2`). The results are shown in Figure 1. As can be seen, both strategies improved in performance as the team size increased. With `SS1` where the Agent did nothing, the strategy reached a recall of just above 0.4 even when the team size was five. An expected result was the performance of `SS2` where the Agent filtered out the judged documents from the retrieval results across the team members. With this simple strategy, one person team performed equivalently to the five person team of `SS1` at the 20th iteration. Moreover, this strategy was able to reach a recall of 0.8 at the 20th iteration when the team size was five.

RQ_1 addressed an effect of relevance feedback as a means of accumulating and sharing knowledge of topical relevance among the team. In `SS3` the Agent kept track of the relevant documents of individual members and expanded a new query with the accumulated profile. In `SS4`, on the other hand, the Agent

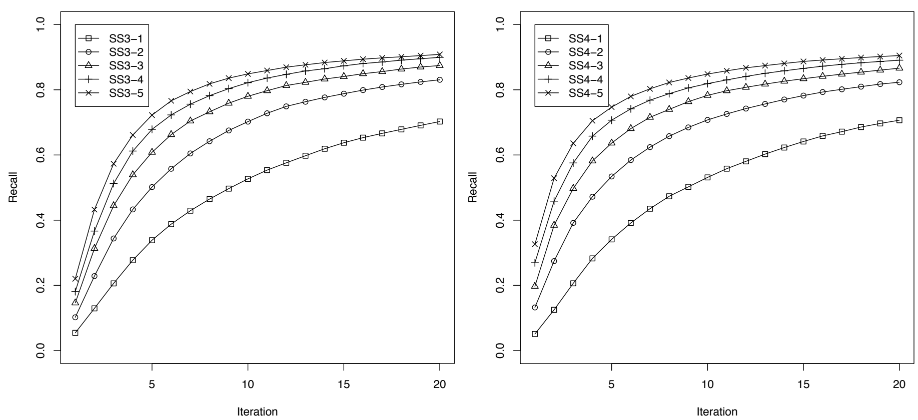

Fig. 2. Recall of RF-based strategies: SS3 (Left) and SS4 (Right).

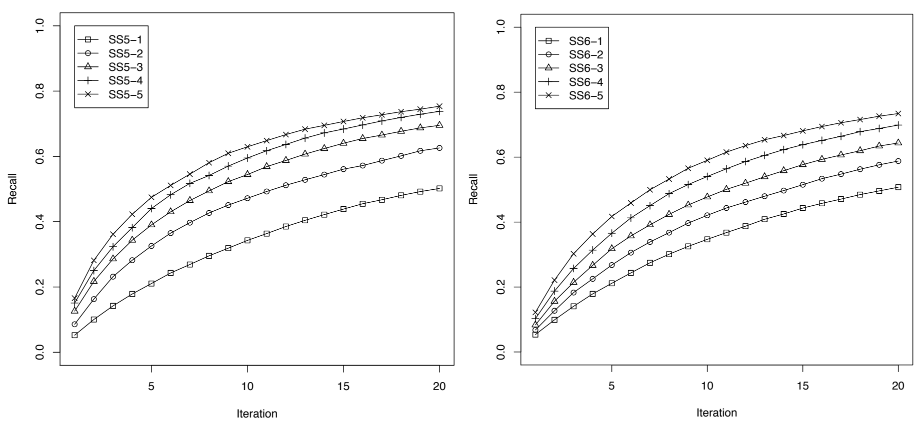

Fig. 3. Recall of grouping-based strategies: SS5 (Left) and SS6 (Right).

had a team profile based on relevant documents found by all team members. The team profile served as implicit sharing of knowledge in collaborative search. The results of the two strategies are shown in Figure 2. As can be seen, both strategies were successful at improving the performance over `SS2`, and reached a recall of 0.9 at the 20th iteration when the team size was five. An effect of sharing relevance information (`SS4`) was found at the early stage of search sessions where the effectiveness of query expansion appeared to depend on the number of relevant documents available up to a point. This helped the team to find more relevant documents at early stages. The performance of both strategies became comparable at the 10th iteration and onwards.

RQ_2 addressed the effect of document grouping as a means of dividing the labour of document browsing and relevance assessments among the team. In `SS5` retrieved documents were grouped by a round-robin approach to each member of

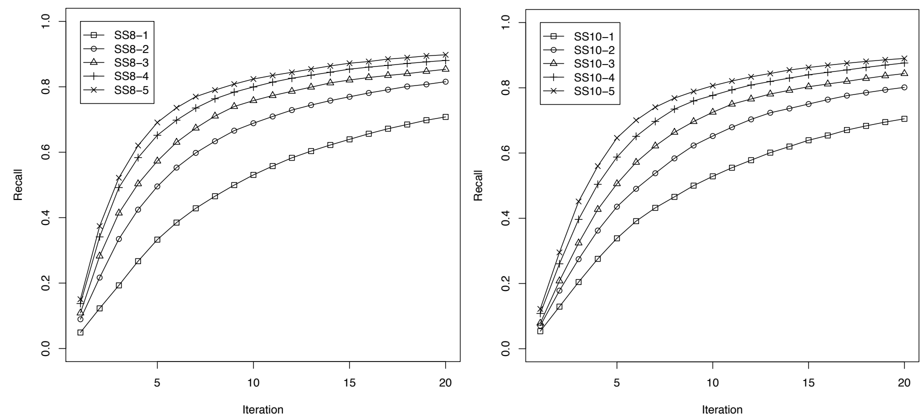

Fig. 4. Recall of combined strategies: SS8 (Left) and SS10 (Right).

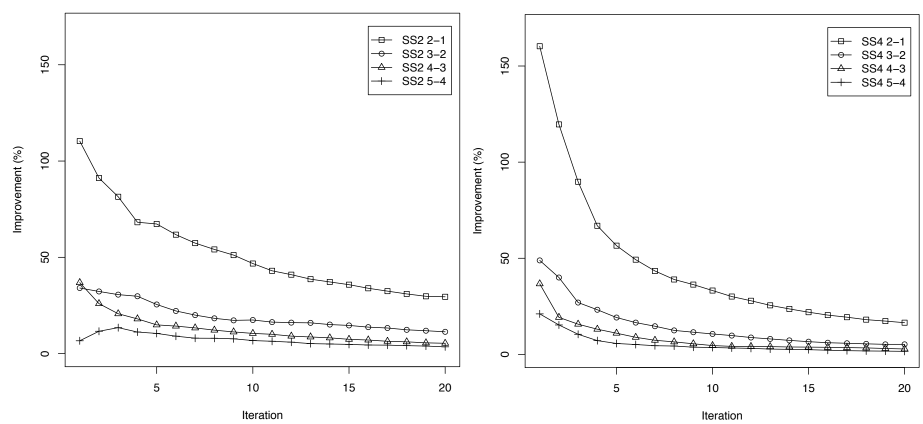

Fig. 5. Effect of team size: SS2 (Left) and SS4 (Right).

the team. In `SS6`, on the other hand, a clustering was performed on the retrieved documents and the top ranked documents in each cluster were distributed to each member of the team. No relevance feedback was performed. The results of the two strategies were shown in Figure 3. As can be seen, both strategies were unsuccessful at improving the performance over `SS2`. The round-robin approach to document grouping was found to perform better than the clustering, although the performance of the two became similar at the 20th iteration with a recall of 0.73 (`SS6`) and 0.75 (`SS5`) when the team size was five.

RQ_3 addressed a complementary effect of relevance feedback and document grouping which was essentially a combination of the strategies presented so far. `SS8` combined `SS4` and `SS5` to perform a shared relevance feedback with a round-robin division, while `SS10` combined `SS4` and `SS6` to perform a shared relevance feedback with a clustering division. The results are shown in Figure 4. At early stages of search, both strategies performed better than `SS2` but worse than the RF-based strategies. With a team size of five, the difference between the combined strategies and RF-based strategies became similar towards the 20th iteration. Given that we expected an improvement in the combined strategies, the results were disappointing. We will discuss the implications of these results in Section 5.

As we have seen so far, the performance was improved as the team size increased. However, the extent of the improvement appeared to vary. RQ_4 looked at the effect of team size on the performance of collaborative search strategies, while RQ_5 concerned the benefit of new members across the progress of search sessions. The results are shown in Figure 5 for `SS2` and `SS4`. A slightly different code was used in the figures. `SS2 5-4` denotes the improvement of team size 5 over team size 4 in Search Strategy 2, that is, a benefit of the fifth member in a team. As can be seen, the benefit of an extra member was the largest on the second member. There was an improvement of 52.1% (SD: 22.3) in `SS2` and 45.8% (SD: 37.8) in `SS4` on average over the 20 iterations. The benefit of the third member was 19.4% in `SS2` (SD: 7.2) and 14.6% (SD: 12.0) in `SS4`, a smaller but still encouraging result. The fourth and fifth member added 12.5% and 7% of improvement in `SS2` and 8.3% and 5.2% of improvement in `SS4`, respectively. The lower improvement in `SS4` is an artifact of its better performance in a smaller team compared to `SS2`. As for RQ_5, there is a general inverse relationship between the impact of extra members and search stage. This was particularly evident in the improvement of the second member (i.e., `SS2 2-1` and `SS4 2-1`). The sustainability of benefit also appeared to be shorter as the team size increased, since the performance tended to better at early search sessions in a larger team.

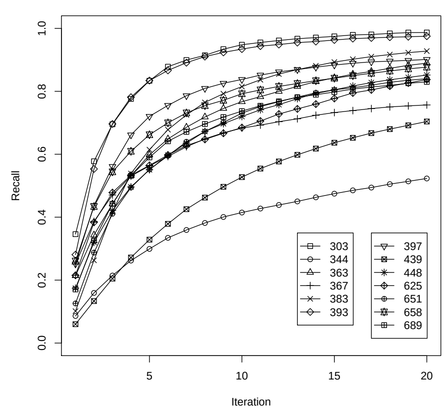

Fig. 6. Topic breakdown of the performance in SS4-5.

We were also interested in the robustness of the strategies and decided to look at the topic breakdown of the performance in SS4-5. The result is shown in Figure 6. There seems to be three groups of the topics: Topic 303 and 393 where almost all relevant documents were found by the 20th iteration; Topic 344, 367, and 439 with a recall between .5 and .75; and the rest with a recall of 0.83 and 0.93. We looked closely at those poor topics. As discussed in Section 3.3, the number of relevant documents in our topic set was reasonably similar. No significant correlation was found between the number of unique queries and the recall, either. We looked at the performance of the three topics in the original user study, and found that participants performed poorly in Topic 344 (Abuses of E-Mail) and 439 (Inventions, scientific discoveries in 90's). Participants expressed a difficulty in finding an *patentable* invention (which was a relevance criterion) in Topic 439. The poor result was also consistent with the top-performing runs in the TREC HARD Track 2005 [15]. Overall, however, it was encouraging that the strategy achieved a recall of over 0.8 in most topics.

5 Discussion

This section discusses the implications of our findings on collaborative search.

5.1 Horizontal and vertical approaches

For a given quantity of resource in collaborative search, one can formulate a horizontal strategy or vertical strategy. In a horizontal strategy, the team submits more queries and judges a shallow depth of retrieved documents. In a vertical strategy, on the other hand, the team submits fewer queries but examines a deeper depth of retrieved documents. The horizontal approach performed better in our simulation. This is in relation to a finding of test collection formation methods [17] where a larger topic size was preferred to a deeper relevance assessment. One of the reasons was that more relevant documents were likely to appear at a higher rank than lower. Similarly, our results show that looking at the top 20 documents in more queries was more effective than looking at the top, say, 100 documents in one fifth the number of queries.

The difference between RF-based and document grouping based strategies also implies the utility of these techniques in collaborative search. In our experimental condition, relevance feedback was found to be robust and easy to formulate an effective strategy. On the other hand, we found it more difficult to formulate an effective strategy using clustering, although there are studies which show the merit of document clustering to improve retrieval effectiveness [10]. However, clustering was not found to be effective as a division of labour among the team members in this study. We also tested a query-biased clustering technique [18] but no significant difference was found from the strategy SS6. Perhaps there were not many cases where retrieved documents had multiple topical aspects to cluster in our condition. However, it should be emphasised that our approach is not the only way to use clustering in interactive IR. A searcher can browse a clustered retrieved documents and select a promising one to find relevant documents [19].

5.2 Limitations

There are some limitations in our study. While we simulated over 10,000 runs in our experiment, the strategies devised still simplified some aspects of the actual collaborative search activity. A fixed number of documents judged per iteration was one such limitation in the simulation. We used a subset of a HARD Track's topics due to the design of the original study although a total of over 1000 queries were formulated and used in our experiment. Another limitation was that we only tested a single relevance feedback and clustering method. While we selected a well known method, there is extensive research on both techniques, and thus, other methods should be investigated to understand their effect on the strategies better. Finally, it should be emphasised that a recall-oriented search is only one task that can be performed in a collaborative fashion. The findings of this study may not apply to other types of task such as a decision making task.

6 Conclusion and Future Work

This paper presented a simulated evaluation of collaborative search strategies. Eight strategies were devised by incorporating established IR techniques such as relevance feedback and clustering. The results were particularly encouraging when relevance feedback was shared by the team members. This suggests that the knowledge of topical relevance can be implicitly shared in a collaborative search with relevance feedback. On the other hand, we found it difficult to effectively divide the labour of document browsing and judgement by clustering. More work is needed to develop the effective use of clustering for collaborative search.

As discussed in Section 1, we were motivated by the need of a high level of exhaustiveness with efficiency in collaborative search. Although the best strategy reached a recall level of 0.9, the recall curve was tailing off. This suggests that further improvement is needed to reach a recall level of 1.0. An effective combination of relevance feedback and clustering is one such area of future work. We also plan to look into query clarity scores [20] which measure the ambiguity of queries, as a guide for selecting strategies in appropriate context.

7 Acknowledgements

Thanks to Robert Villa for helpful feedback on a draft, and Irakiris Klampanos for valuable conversions on clustering strategies. Funding was provided by the MIAUCE project (Ref: IST-033715).

References

1. Hansen, P., Järvelin, K.: Collaborative information retrieval in an information-intensive domain. Information Processing & Management **41**(5) (2004) 1101–1119
2. IRF, ed.: Proceedings of the First Information Retrieval Facility Symposium (IRFS), Vienna, Austria, Matrixware (2007)
3. Dourish, P., Bellotti, V.: Awareness and coordination in shared workspaces. In: Proceedings of the 1992 ACM CSCW Conference. (1992) 107–114
4. Smyth, B., Balfe, E., Boydell, O., Bradley, K., Briggs, P., Coyle, M., Freyne, J.: A live-user evaluation of collaborative web search. In: Proceedings of the 9th IJCAI Conference. (2005) 1419–1424
5. Smeaton, A.F., Lee, H., Foley, C., McGivney, S.: Collaborative video searching on a tabletop. Multimedia System **12**(4-5) (2007) 375–391
6. Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., Back, M.: Algorithmic mediation for collaborative exploratory search. In: Proceedings of the 31st ACM SIGIR conference. (2008) 315–322
7. Morris, M.R., Horvitz, E.: Searchtogether: an interface for collaborative web search. In: Proceedings of the 20th ACM UIST Conference, ACM (2007) 3–12
8. Villa, R., Gildea, N., Jose, J.M.: A study of awareness in multimedia search. In: Proceedings of the 8th ACM JCDL Conference. (2008) 221–230
9. Joho, H., Hannah, D., Jose, J.M.: Comparing collaborative and independent search in a recall-oriented task. In: Proceedings of the second IIiX Symposium, ACM (2008) 89–96
10. Tombros, A., Villa, R., Rijsbergen, C.J.V.: The effectiveness of query-specific hierarchic clustering in information retrieval. Information Processing & Management **38**(4) (2002) 559–582
11. Harman, D.: Towards interactive query expansion. In: Proceedings of the 11th ACM SIGIR Conference. (1988) 321–331
12. Magennis, M., van Rijsbergen, C.J.: The potential and actual effectiveness of interactive query expansion. In: Proceedings of the 20th ACM SIGIR Conference. (1997) 324–332
13. White, R.W., Jose, J.M., van Rijsbergen, C.J., Ruthven, I.: A simulated study of implicit feedback models. In: Proceedings of the 26th ECIR Conference. (2004) 311–326
14. Ruthven, I.: Re-examining the potential effectiveness of interactive query expansion. In: Proceedings of the 26th ACM SIGIR Conference. (2003) 213–220
15. Allan, J.: Hard track overview in trec 2005 high accuracy retrieval from documents. In: Proceedings of the 14th TREC, NIST Special Publication: SP 500-266. (2005)
16. Ounis, I., Lioma, C., Macdonald, C., Plachouras, V.: Research directions in Terrier: a search engine for advanced retrieval on the web. Novatica/UPGRADE Special Issue on Web Information Access **8**(1) (2007) 49–56
17. Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J.A., Allan, J.: Evaluation over thousands of queries. In: Proceedings of the 31st ACM SIGIR conference. (2008) 651–658
18. Iwayama, M.: Relevance feedback with a small number of relevance judgements: incremental relevance feedback vs. document clustering. In: Proceedings of the 23rd ACM SIGIR conference. (2000) 10–16
19. Zamir, O., Etzioni, O.: Grouper: A dynamic clustering interface to web search results. In: Proceedings of the 8th International WWW Conference. (1999)
20. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of the 25th ACM SIGIR conference. (2002) 299–306