Towards Quantifying the Impact of Non-Uniform Information Access in Collaborative Information Retrieval

Nyi Nyi Htun
SEBE
Glasgow Caledonian University
Glasgow, G4 0BA, Scotland, UK
nyinyi.htun@gcu.ac.uk

Martin Halvey
Department of CIS
University of Strathclyde
Glasgow, G1 1XQ, Scotland, UK
martin.halvey@strath.ac.uk

Lynne Baillie
Department of CS
Heriot-Watt University
Edinburgh, EH14 4AS, Scotland, UK
lynne.baillie@hw.ac.uk

ABSTRACT
The majority of research into Collaborative Information Retrieval (CIR) has assumed a uniformity of information access and visibility between collaborators. However in a number of real world scenarios, information access is not uniform between all collaborators in a team e.g. security, health etc. This can be referred to as Multi-Level Collaborative Information Retrieval (MLCIR). To the best of our knowledge, there has not yet been any systematic investigation of the effect of MLCIR on search outcomes. To address this shortcoming, in this paper, we present the results of a simulated evaluation conducted over 4 different non-uniform information access scenarios and 3 different collaborative search strategies. Results indicate that there is some tolerance to removing access to the collection and that there may not always be a negative impact on performance. We also highlight how different access scenarios and search strategies impact on search outcomes.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval

General Terms
Measurement, Performance, Experimentation.

Keywords
Collaborative search, non-uniform access, effectiveness measures

1. INTRODUCTION
Collaborative Information Retrieval (CIR) involves people with common information needs working together, exploring and collecting useful information, and collectively making decisions that help them move toward their common goal. A simple example might be of a group of colleagues collaborating for a project where they may, individually or together, go through a number of information resources and then discuss their results, exchanging information and knowledge in order to contribute to the project.

A common assumption in much of the research in CIR is that all members of a team have equal access to the information sources, tools etc., and that they may share any relevant information they find with each other without any restriction [4, 5, 11]. However, in reality it may not always be the case that all searchers have equal information access. There are numerous situations where societal, legal or security reasons may prevent a searcher from sharing information within or out with a group. Handel and Wang [6] presented an example of such a scenario involving two intelligence analysts engaged in collaborative search, where one analyst is a signal intelligence specialist and the other a human intelligence specialist. Despite their unequal access to intelligence databases and underlying intelligence, as well as differing information needs and shareability, the two analysts must collaborate to achieve an outcome. This type of scenario was referred to as Multi-Level Collaborative Information Retrieval (MLCIR) [6]. Similar scenarios have been examined by other researchers who have looked at the effect of organisational structure in legal search [2], crisis management [3] and healthcare [10] to gain a better understanding of how these can impede collaboration. Others have considered how different roles within a search team might be leveraged to assist with CIR. For example, Pickens et al. [12] studied the impact of having two different roles in a collaborative exploratory search team, and looked into developing algorithms to support this. However, the main focus of these studies has been on the *division of labour* in CIR and although, to date, having different roles has been viewed as positive in collaborative search tasks, it might not always be. In fact, MLCIR is different from *division of labour* in that any system that supports MLCIR has to be aware of information flow, accessibility and shareability between collaborators [6]. Thus many of the concepts previously used to support CIR such as awareness, sense-making and persistence [4, 5, 11] may need to be revised.

Previous research [2, 3, 9, 10] has focused primarily on qualitative observations which may not be completely applicable in all non-uniform information access scenarios. To the best of our knowledge, there has yet to be a systematic evaluation on the impact of non-uniform information access within a team of searchers. We attempt to overcome this shortcoming by conducting a simulated user evaluation where we investigate the impact of two different kinds of non-uniformity in access, namely removing document access and search-term blacklisting for team members (Details are presented in Section 2.2). There are three main research questions that we attempt to answer in this paper:

1. What is the impact of non-uniform information access on the outcomes of CIR?
2. Do different types of non-uniformity have different impacts on CIR outcomes?
3. Are there scenarios where non-uniform access may be beneficial to CIR outcomes?

2. EXPERIMENTAL DESIGN
As there are a number of potential parameters for collaboration and non-uniformity in information access, we decided to use a simulated study. This approach means that we can more easily compare different variables and combinations than in a user evaluation. In future work, we anticipate exploring the findings from this study in more depth with a user evaluation.

2.1 Data, Topic and Search Strategies

Our evaluation followed the same procedure as Joho et al.'s simulation of collaborative search [7], with some small changes as outlined below. We utilised the TREC HARD 2005 [1] collection (AQUAINT corpus) and topics. For their study, Joho et al. [8] generated a query pool through a user evaluation for 13 of the topics. We were provided with this query pool and thus use the same 13 topics (303, 344, 363, 367, 383, 393, 397, 439, 448, 625, 651, 658 689). The query pool has a total of 1157 queries across the 13 topics and each query contains up to 9 terms.

Joho et al. [7] simulated teams of searchers (of variable size from 2 to 5) to carry out collaborative search tasks. Each team had 20 search iterations per topic. During each iteration, a team member selected a random query from the query pool and was assumed to judge 20 documents per iteration. For simplicity in our evaluation we simulate a pair of users rather than vary team size, as this would introduce extra complexity, whereby combining a multitude of possible access combinations could become intractable. In other words, we assume that there are always 2 people in a search team for any given search session and the team performs 20 search iterations. Thus each individual in a team would judge a maximum of 400 documents per topic, with a team judging a maximum of 800 documents. One of the goals of Joho et al. was to compare a number of collaborative search strategies [7]; we utilise 3 of these search strategies for our study. These 3 strategies are:

1) *Independent Search (IS)*: team members judge documents independently without any interaction between each other, and have their results merged at the end of each search iteration.

2) *Independent Relevance Feedback (IRF)*: same as (1) but query expansion is performed based on their independent relevance feedback and then the expanded queries are resubmitted independently to the system. Team members do not share any knowledge on relevancy of documents.

3) *Shared Relevance Feedback (SRF)*: same as (2) but the query expansion is performed based on the relevance feedback of both members. Thus, team members share knowledge on relevancy of the documents.

For Joho et al. [7], *IS* was the most basic and simplest search strategy whereas the other two were the most effective. Due to its simplicity, *IS* is also the easiest to compare directly with any other search strategies in terms of performance, collection coverage, etc. The other two strategies chosen were the best performing in their experiments.

2.2 Access Scenarios and Combinations

We devised 4 scenarios to simulate non-uniform information access amongst team members completing a collaborative search task; these are summarised in Table 1 and outlined in detail below. For each scenario, we assumed that each of the two searchers have access to more or less of the collection relative to their search partner. For example, in one case, one searcher might be able to access only 10% of the collection while their partner can access 20% of the collection. Also, there is a possibility that one searcher cannot retrieve any documents that contain certain phrases or terms.

Therefore, starting with S1 (document removal), we began by indexing a random selection of 10% of the documents from the document collection. Then an iterative process was adopted whereby we increased the percentage of documents indexed by 10% until 100% of the collection had been indexed. This resulted

in 10 different indexes for each person and 55 possible access combinations of indexes for two people (i.e. combinations of 10%-10%, 10%-20%, 10%-30%, 10%-40%; up to 100%-100%). This simulates a scenario laid out by Handel and Wang [6] where a person with higher security clearance may have access to more documents than a subordinate.

Table 1. Information access scenarios

Code	Scenario
S1	Remove access to documents from collection
S2	Term blacklisting – remove access to random terms from the collection
S3	Term blacklisting – remove access to terms based on their frequency in documents
S4	Term blacklisting - remove access to terms based on their frequency in query pool

Scenarios S2, S3 and S4 simulate term blacklisting, this is a major problem highlighted by Handel and Wang [6]. For S2, we began by analysing the collection for a list of terms. After that, we indexed the entire corpus meaning there is complete access. We then created other indexes by iteratively removing 10% of the terms randomly, until only 10% remained. This also resulted in 55 possible combinations of indexes for 2 individuals. Scenarios S3 and S4 took a more systematic approach. We analysed term frequencies in both collection and query pool, which contain 841498 and 591 unique terms respectively. We then followed the same procedure as S2 but instead of removing random terms we removed terms based on their frequencies in the collection and in the query pool respectively for S3 and S4. Therefore, for S3 the first 10% removed were the most frequent terms in the collection whereas for S4 those were the most frequent terms in the query pool. In each scenario we had 10 indexes for each team member and 55 different access combinations, although the indexes in S4 are of different size to S1, S2 and S3 because in S1, S2 and S3 we can theoretically exclude everything from the collection whereas for S4 this is dependent on the query pool.

Thus for each scenario, there are 55 possible combinations; for each of these combinations, we conducted each search simulation 10 times in order to reduce randomness and inconsistencies. In total, there were 1,716,000 search sessions performed by teams in our simulation (i.e. 3 search strategies x 4 access scenarios x 10 runs x 55 combinations x 13 topics x 20 iterations). For all of the indexing and retrieval, we used the Inverted File indexing method and BM25 retrieval algorithm, these were developed using the Terrier[1] library with out of the box settings.

2.3 Evaluation Measures

For the evaluation we utilised traditional IR evaluation metrics: recall, precision and f-measure in conjunction with specific metrics for CIR proposed by Shah and González-Ibáñez [13]: coverage, relevant coverage, unique coverage and unique relevant coverage. Coverage is the average number of distinct documents discovered by the team throughout the entire search session. Relevant coverage is the average number of documents in coverage that are actually relevant. Unique coverage is the average number of distinct documents that are only discovered in a given access combination, and not in any other. Unique relevant coverage is the average number of documents in unique coverage that are actually relevant.

---

[1] http://terrier.org

3. RESULTS

Table 2 shows the access combinations which yield the highest values for recall, precision and f-measure across all access scenarios and search strategies and Table 3 shows those for coverage, relevant coverage, unique coverage and unique relevant coverage. As our data was not normally distributed, for each measure across 4 access scenarios and 3 search strategies, we conducted a Friedman analysis to compare the 55 access combinations (i.e. 10-10, 20-10, 20-20, 30-10, etc.) and found that there was a statistically significant difference in every case. Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at p<0.00003367. We present more detailed results of the pairwise comparisons in the following sub-sections. For reasons of space as there were many comparisons we do not present all of these comparisons.

3.1 Search Performance

Our first research question examined the impact of non-uniform information access on the outcomes of CIR. First of all, statistical analysis of recall, precision and f-measure values showed a number of access combinations that were not significantly different from the best performing access combinations. However, what was interesting among these is that for S1, S2 and S4, relevance feedback search strategies had a very high number of combinations that are not significantly different from their best performing access combinations (ranging from 50-20 to 90-60 for S1, 70-70 to 90-80 for S2, and 70-70 to 100-80 for S4) whereas the *IS* strategy had only a few (90-80, 90-90, 100-80, 100-100 for S1; 90-90, 100-10, 100-60 100-90 for S2; 90-90, 100-90 for S4). It suggests that in terms of recall, precision and f-measure non-uniform access for S1, S2 and S4 had very little effect when relevance feedback strategies were employed.

Table 2. Highest recall, precision and f-measure values with their respective access combinations. * indicates those values at full access (i.e. 100-100)

	Recall	Precision	F-measure
Independent Search			
S1	0.0859 (100-90) 0.0829*	0.2459 (100-90) 0.23898*	0.1270 (100-90) 0.1227*
S2	0.0813 (100-100)	0.2349 (100-100)	0.1204 (100-100)
S3	0.0818 (100-100)	0.2446 (100-20) 0.2353*	0.1210 (100-100)
S4	0.0830 (100-100)	0.2389 (100-100)	0.1228 (100-100)
Independent Relevance Feedback			
S1	0.1210 (90-90) 0.0383*	0.3576 (90-90) 0.1302*	0.1802 (90-90) 0.0604*
S2	0.1110 (90-90) 0.0376*	0.3273 (90-90) 0.1266*	0.1653 (90-90) 0.0588*
S3	0.1241 (90-90) 0.0370*	0.3931 (90-90) 0.1244*	0.1878 (90-90) 0.0572*
S4	0.0904 (90-90) 0.0376*	0.2711 (90-90) 0.1295*	0.1350 (90-90) 0.0580*
Shared Relevance Feedback			
S1	0.1001 (90-30) 0.0325*	0.3317 (80-70) 0.1756*	0.1502 (90-30) 0.0548*
S2	0.0836 (90-90) 0.0324*	0.4197 (90-90) 0.1748*	0.1391 (90-90) 0.0554*
S3	0.1006 (90-90) 0.0323*	0.5208 (90-90) 0.1745*	0.1683 (90-90) 0.0544*
S4	0.0762 (100-90) 0.0324*	0.3570 (90-90) 0.1748*	0.1173 (100-90) 0.0551*

Looking at Table 2, we found that when the *IS* strategy was employed for S1, the values of the 3 measures (recall, precision and f-measure) were highest at non-full access (i.e. 100-90) whereas for the rest of the scenarios (S2, S3 and S4) the values reached the highest at full access. When relevance feedback strategies were employed, however, it was found that the values reached the highest at non-full access (mostly at 90-90) for all 4 scenarios (S1, S2, S3 and S4). This suggests that there is some tolerance to removing access from the collection, and while it was expected that there would be a decrease in performance when access had been reduced, there were some cases which indicate that there may not always be a negative impact on performance. In addition, as mentioned earlier, our statistical test results revealed a number of combinations that are not significantly different from the best performing access combinations, which suggests that there are certain combinations that allow search performance to be comparable to the best performing access combination regardless of the users' unequal, or equal but not full (e.g. 90-90) access to the collection. This finding addresses our third research question. Moreover, the statistical test results also showed us that depending on the type of access scenario and search strategies being utilised, the resulting combinations were different, and thus resulted in different outcomes, addressing our second research question.

3.2 Collection Coverage

In terms of coverage for the document removing scenario (S1), statistical test results showed that in all 3 search strategies, there were many access combinations which were not significantly different from the best performing access combination and also represent the case where team members had access to a very diverse amount of the collection from each other (these are 50-10, 60-10, 70-10, 80-10, 80-20, 90-10, 90-20, 100-10, 100-20, 100-30). It appears that regardless of the search strategy, reducing access to documents for one member of the team means that a different member can make judgements about different parts of the collection thereby covering similar amount of documents as they would in the best performing access combinations. This finding is in contrast to term blacklisting scenarios (S2, S3 and S4) in which most combinations that are not significantly different from the best performing access combination represent the case where both team members had a higher access to the collection (e.g. 60-60, 100-80, etc.). Next, looking at coverage in Table 3, the fact that the highest values were obtained at non-full access again indicates that there may not always be a negative impact on performance when access has been reduced, addressing our third research question. In addition, statistical test results of coverage also showed that the resulting access combinations are different depending on the type of access scenario and search strategy being utilised which addresses our second research question.

In terms of relevant coverage, Table 3 indicates that when the *IS* strategy was utilised, the highest values were obtained at full access (100-100) for all of the term blacklisting scenarios (S2, S3 and S4). However, statistical test results also indicated that there were non-full-access combinations where relevant coverage was as high as the full access. Besides, it also showed that the resulting access combinations and their outcomes are different depending on the type of access scenario and search strategy being utilised, again addressing our second research question. With respect to unique coverage for S1, it can be seen in Table 3 that across all search strategies the access combination that has highest value is the lowest access (10-10), and this is opposite to S3 where the full access has the highest unique coverage. In addition, it is interesting to note that for all 4 scenarios (S1, S2, S3 and S4) the *SRF* strategy was able to obtain very high unique coverage in all access

combinations compared to the other two strategies. Statistical test results showed that for S2, when the *IS* and *IRF* strategies were utilised, many of the access combinations ranging from 20-10 to 100-100 showed no significant difference from the best performing access combinations (i.e. 50-40 and 10-10 respectively). A similar outcome was also found for S4, but across all 3 search strategies. Unique relevant coverage in Table 3 shows that for all scenarios (other than for S3 of the *IS* strategy), the highest values were not obtained at full access. However, it appears that reducing access to the collection has little or no effect in terms of unique relevant coverage as statistical test results indicated that for almost every access scenario and search strategy, none of the access combinations showed any significant difference from the best performing access combinations.

Table 3. Highest values of different CIR measures with their respective access combinations. * indicates values of those measures at full access (i.e. 100-100)

	Coverage	Relevant Coverage	Unique Coverage	Unique Relevant Coverage
Independent Search				
S1	365.7769 (100-10) 297.7461*	44.6461 (100-80) 42.0769*	8.4923 (10-10) 2.0307*	0.0923 (80-20) 0.0461*
S2	355.7153 (80-80) 296.2615*	42.1615 (100-100)	14.7615 (50-40) 0.3769*	0.1923 (80-70) 0.0*
S3	304.2384 (100-90) 297.9615*	42.4769 (100-100)	9.8846 (100-100)	0.1615 (100-100)
S4	418.6461 (90-60) 296.1538*	42.4307 (100-100)	4.8461 (10-10) 1.8923*	0.1153 (100-30) 0.0538*
Independent Relevance Feedback				
S1	349.2769 (100-80) 290.8846*	48.3769 (90-60) 19.4846*	81.5231 (10-10) 12.4615*	0.3923 (10-10) 0.0*
S2	349.8692 (100-50) 292.4385*	47.5231 (90-80) 19.0538*	12.5923 (10-10) 7.9692*	0.3 (90-60) 0.0*
S3	326.6077 (100-90) 277.8692*	42.7538 (90-80) 18.2769*	17.1385 (100-100)	0.2231 (90-90) 0.0846*
S4	407.6769 (100-60) 281.6154*	40.3308 (90-90) 18.6*	8.7231 (100-100)	0.0846 (100-80) 0.0615*
Shared Relevance Feedback				
S1	353.6153 (100-10) 244.5308*	43.5 (90-40) 17.1*	133.3615 (10-10) 58.6692*	0.7923 (10-10) 0.0385*
S2	361.1615 (100-40) 241.3*	41.4308 (100-90) 17.2385*	74.7538 (40-40) 42.777*	1.2462 (80-20) 0.0769*
S3	304.0615 (100-90) 242.5308*	43.1615 (100-90) 17.0231*	67.2769 (100-100)	1.2692 (100-90) 0.2769*
S4	387.4077 (100-30) 249.3769*	40.1077 (100-90) 17.1308*	47.7692 (100-10) 45.7692*	0.4308 (100-90) 0.3077*

4. CONCLUSION AND FUTURE WORK

While a great deal of research has focused on CIR, only a few papers have considered the impact of non-uniform information access on CIR outcomes. This paper is one of the first attempts to quantify the impact of non-uniform information access on CIR outcomes. To that end, we conducted a simulated user evaluation using established scenarios [6] and search strategies [7].

In relation to our first research question it was found that in terms of recall, precision and f-measure that non-uniform access for S1, S2 and S4 had very little impact when relevance feedback strategies were employed. In addition, it was also found that in some cases, one member of the team having a high level of access can compensate for the other team member. Besides, our results have also highlighted that there is some tolerance to removing access from the collection and that there may not always be a negative

impact on performance. This leads us into our second and third research questions. We have found that depending on the type of access scenario and search strategy, access combinations yield different outcomes. Removing access to documents and term blacklisting had different impacts in terms of coverage: for removing document access, coverage remained stable where at least one team member had high access, whereas for blacklisting both members needed high access to retain high coverage. We have also found that in some scenarios, performance is even increased due to non-uniformity. This may in part be because this ensures that parts of the collection which might otherwise be ignored due to overlap in retrieved documents are now examined. Thus, there can be some benefits to non-uniform access depending on the search task.

To address our research questions in this paper we used 3 search strategies, 4 access scenarios, 7 different measures and teams of 2 simulated users. We anticipate extending this study in various ways to be able to produce findings that greatly generalise to a number of real situations. Thus, we intend to look at more complex strategies and access scenarios, and incorporate more users within each team. Furthermore, the findings from this study will be examined further via a user evaluation. To conclude, our findings provide a better understanding on the impact of non-uniform information access amongst searchers in collaborative information retrieval, as well as a roadmap for further user studies.

5. REFERENCES

[1] Allan, J: HARD Track Overview in TREC 2005: High Accuracy Retrieval from Documents. *TREC 2005*, pp. 1-17

[2] Attfield, S., Blandford, A., Makri, S.: Social and interactional practices for disseminating current awareness information in an organisational setting. *IPM*, *46*(6), (2008)

[3] Bjurling, B., Hansen, P.: Contracts for Information Sharing in Collaborative Networks. ISCRAM 2010 (Vol. 1)

[4] González-Ibáñez, R., Shah, C.: Coagmento: A system for supporting collaborative information seeking. *JASIST*, *48*(1), 1-4 (2011)

[5] Halvey, M., Vallet, D., Hannah, D., Feng, Y., Jose, J. M.: An asynchronous collaborative search system for online video search. *IPM*, *46*(6), 733-748 (2010)

[6] Handel, M. J., Wang, E. Y.: I can't tell you what i found: problems in multi-level collaborative information retrieval. *3rd international workshop on Collaborative information retrieval,* pp. 1-6. ACM CIKM 2011

[7] Joho, H., Hannah, D., Jose, J. M.: Revisiting IR techniques for collaborative search strategies. ECIR 2009, pp. 66-77

[8] Joho, H., Hannah, D., Jose, J. M.: Comparing collaborative and independent search in a recall-oriented task. ACM IIiX, 2008 pp. 89-96

[9] Karunakaran, A., Reddy, M.: Barriers to collaborative information seeking in organizations. *JASIST*, *49*(1), (2012)

[10] Karunakaran, A., Reddy, M.: The Role of Narratives in Collaborative Information Seeking. ACM SIGGROUP 2012, pp. 273–276

[11] Morris, M. R., Horvitz, E.: SearchTogether: an interface for collaborative web search. ACM UIST 2007, pp. 3-12

[12] Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., Back, M.: Algorithmic mediation for collaborative exploratory search. ACM SIGIR 2008, pp. 315-322

[13] Shah, C., González-Ibáñez, R.: Evaluating the synergic effect of collaboration in information seeking. ACM SIGIR 2011, pp. 913-922