Beyond Traditional Collaborative Search: Understanding the Effect of Awareness on Multi-Level Collaborative Information Retrieval

Nyi Nyi Htun[1], Martin Halvey[2] & Lynne Baillie[3]
[1]Glasgow Caledonian University, Glasgow, UK (E-mail: nyinyi.htun@gcu.ac.uk);
[2]University of Strathclyde, Glasgow, UK (E-mail: martin.halvey@strath.ac.uk);
[3]Heriot-Watt University, Edinburgh, UK (E-mail: l.baillie@hw.ac.uk)

Abstract

Although there has been a great deal of research into Collaborative Information Retrieval (CIR) and Collaborative Information Seeking (CIS), the majority has assumed that team members have the same level of unrestricted access to underlying information. However, observations from different domains (e.g. healthcare, business, etc.) have suggested that collaboration sometimes involves people with differing levels of access to underlying information. This type of scenario has been referred to as Multi-Level Collaborative Information Retrieval (MLCIR). To the best of our knowledge, no studies have been conducted to investigate the effect of awareness, an existing CIR/CIS concept, on MLCIR. To address this gap in current knowledge, we conducted two separate user studies using a total of 5 different collaborative search interfaces and 3 information access scenarios. A number of Information Retrieval (IR), CIS and CIR evaluation metrics, as well as questionnaires were used to compare the interfaces. Design interviews were also conducted after evaluations to obtain qualitative feedback from participants. Results suggested that query properties such as *time spent on query*, *query popularity* and *query effectiveness* could allow users to obtain information about team's search performance and implicitly suggest better queries without disclosing sensitive data. Besides, having access to a history of intersecting viewed, relevant and bookmarked documents could provide similar positive effect as query properties. Also, it was found that being able to easily identify different team members and their actions is important for users in MLCIR. Based on our findings, we provide important design recommendations to help develop new CIR and MLCIR interfaces.

Key words: awareness; collaborative search; information access; multi-level collaboration; non-uniform access

1. Introduction

A great deal of research in Collaborative Information Retrieval (CIR) and Collaborative Information Seeking (CIS), e.g. (Capra et al., 2012; Halvey, Vallet, Hannah, Feng, & Jose, 2010; Morris, 2013; Shah, 2016; Soulier, Tamine, & Shah, 2016; Tamine & Soulier, 2016), assumes that team members in a collaborative search team have equal and non-restrictive access to underlying information. However, in practice, for a number of reasons such as security, privacy, etc., team members may not always have equal access to underlying information. For example, as Handel and Wang (2011) outlined, a signal intelligence specialist and a human intelligence specialist could be working together to understand a new threat. Due to their lack of equal access to underlying information such as intelligence databases, the two

specialists may have differing knowledge but most importantly, they may not be able to share any or part of it between each other. Despite this, the two specialists must somehow work together to understand the threat. This type of scenario has been referred to as Multi-Level Collaborative Information Retrieval (MLCIR), a term first proposed by Handel and Wang (2011). Day to day activities such as searching for health information online may also introduce similar problems. De Choudhury et al. (2014) surveyed 210 people to find out how they choose between search engines and social media to search for health information. De Choudhury et al. found that people are less likely to share their health-related information with others on stigmatic conditions. This is closely related to another MLCIR example highlighted by Handel and Wang (2011) where an individual with a health issue does not want to disclose the entire range of symptoms to other people in a group. Thus, MLCIR can occur not just in certain businesses and organisations, but also in our day to day activities.

Recently, some researchers have begun to realise the complexity and difficulty of collaborative search within important domains such as healthcare (Karunakaran & Reddy, 2012), crisis management (Bjurling & Hansen, 2010) and legal search (Attfield, Blandford, & Makri, 2010); these researchers discussed how unequal distribution of knowledge and organisational hierarchies could hinder collaboration in the respective domains. Handel and Wang (2011) also discussed in detail a number of case studies from several domains including healthcare, business and government highlighting problems that emerged due to non-uniform access to underlying information.

MLCIR is complex and difficult because considerations need to be given to information flow, security and shareability between collaborators in addition to the collaboration itself (Handel & Wang, 2011). Therefore, a number of existing CIR and CIS concepts such as awareness, division of labour and persistence may be inapplicable to MLCIR. Although such concepts have previously been investigated by a number of researchers (Halvey et al., 2010; Morris & Horvitz, 2007; Shah & Marchionini, 2010), to the best of our knowledge, there has not yet been any investigation into the effect of existing CIR and CIS concepts on MLCIR. Previous work presented by Bjurling and Hansen (2010), Attfield et al. (2010), Handel and Wang (2011), and Karunakaran and Reddy (2012) has been based on observations and did not provide a systematic solution to solve the problems with MLCIR. In order to systematically evaluate the impact of non-uniform information access in CIR, we conducted a simulated user study (Htun, Halvey, & Baillie, 2015). However, this work did not go as far as a user study in that actual human feedback was not provided, and not all user interaction could be easily replicated in the simulation.

To address these shortcomings, we conducted a preliminary user study which indicated three awareness types that are usable for MLCIR interfaces (Htun, Halvey, & Baillie, 2017); these are query awareness, result awareness and team awareness. In this paper, we present two separate user studies where we investigated the impacts of different awareness kinds on MLCIR using the MLCIR scenarios that were highlighted by Handel and Wang (2011) and were also utilised in our previous simulated study (Htun et al., 2015). In the first user study, we investigated the impacts of query awareness. In the second user study, we investigated the impacts of result awareness and team awareness. Result awareness and team awareness were investigated as one study because at the time the study was being conducted, not

many different interface components were proposed for either result awareness or team awareness that are usable in MLCIR interfaces. As for query awareness, different variety of components have been utilised in previous collaborative search systems (Amershi & Morris, 2008; Joho, Hannah, & Jose, 2008; Morris & Horvitz, 2007; Shah, 2010a). The main difference between the two studies was the interfaces, and the type of awareness that they support. The combined objectives of the two studies presented in this paper are to:

1) understand the impact of supporting query awareness, result awareness and team awareness on collaborative search outcomes in MLCIR scenarios.
2) understand the impact of supporting query awareness, result awareness and team awareness on individual search outcomes in MLCIR scenarios.
3) understand the impact of supporting query awareness, result awareness and team awareness on users' search experience in MLCIR scenarios.
4) provide design recommendations to help develop new MLCIR interfaces.

Since our studies were the first attempt to investigate different awareness types in MLCIR scenarios, we developed a number of interfaces which used previous research studies as a starting point, e.g. (Amershi & Morris, 2008; Freyne, Farzan, Brusilovsky, Smyth, & Coyle, 2007; Htun et al., 2017; Morris & Horvitz, 2007; Shah, 2010a). Other than the interfaces, the studies shared the same experimental design. Pairs of participants were presented with three different information access scenarios and search interfaces. The participants' collaborative and individual search outcomes were measured using a number of existing evaluation metrics (Freyne et al., 2007; Joho et al., 2008; Shah & González-Ibáñez, 2011; Soulier, Shah, & Tamine, 2014), e.g. some measured performance, some measured collection coverage, etc. Participants were also asked a number of post-task evaluation questions to be able to assess their perception of search tasks, performance, etc. At the end of the study, design interviews were undertaken to obtain participants' feedback related to their search experience and to be able to provide important design recommendations for new MLCIR interfaces.

The remainder of the paper is organised as follows. In Section 2, we discuss related research regarding CIR and CIS, the awareness concept and MLCIR. In Section 3, we present the experimental setup and results of study 1. In Section 4, we present the experimental setup and results of study 2. In Section 5, we discuss the results from both studies, providing design recommendations based on findings from the design interviews. In Section 6, we highlight limitations of the studies. Finally, we conclude this paper in Section 7 and outline possible future work.

2. Background

2.1. Collaborative Information Retrieval/Seeking

Searching for information was often considered a solo activity, but there are many situations where a group of people with shared information need to work together to search for information (Tamine & Soulier, 2016; Tamine et al., 2016). For information searching activities that involve gathering a large amount of information, e.g. patent searching, troubleshoot information searching, etc., collaboration is an effective means compared to individual efforts (Shah, 2012; Tamine & Soulier, 2016). This is because collaboration gives rise to a number of opportunities such as sharing workload, submitting diverse queries, etc. which cannot be achieved during individual

search. For this reason, an increasing number of people in different domains have been engaging in various information searching activities (Morris, 2008; Morris, 2013; Shah, 2015; Spence, Reddy, & Hall, 2005).

The term CIS has been used by some researchers, e.g. (González□Ibáñez & Shah, 2011; Mitsui & Shah, 2016; Shah, 2015) while others have used the term CIR, e.g. (Foley & Smeaton, 2010; Handel & Wang, 2011; Hansen & Järvelin, 2005; Joho et al., 2008; Soulier et al., 2016; Tamine & Soulier, 2016). According to Shah (2010b), CIS is "a process of information seeking that is defined explicitly among the participants, interactive, and mutually beneficial" (pp. 14). As for CIR, according to Foster (2006) it is "the study of the systems and practices that enable individuals to collaborate during the seeking, searching, and retrieval of information" (pp. 329). Nevertheless, to date commonly accepted definitions for both terms do not exist and many researchers have used the terms interchangeably.

In order to support users in collaborative search activities, a number of collaborative search systems have been proposed, e.g. (Amershi & Morris, 2008; Golovchinsky, Adcock, Pickens, Qvarfordt, & Back, 2008; Mitsui & Shah, 2016; Morris & Horvitz, 2007; Morris, Lombardo, & Wigdor, 2010; Shah, 2010a). Most collaborative search systems can be distinguished into UI-only mediated and algorithmic mediated systems (Golovchinsky, Pickens, & Back, 2009). In UI-only mediated systems, collaboration is supported only at the user interface level by utilising UI components such as result visualisation, result recommendation, query visualisation, instant messaging, etc. In algorithmic mediated systems, collaboration is enhanced by an algorithmic layer that re-ranks search results based on users' roles, actions or preferences. Although there are communication-only systems that support collaboration through communication channels such as instant messaging, voice chat and video conferencing, such systems are commonly not considered CIR nor CIS systems (González□Ibáñez & Shah, 2011; Morris & Horvitz, 2007).

There are a number of examples of UI-only mediated systems. CoSearch (Amershi & Morris, 2008) allows synchronous and co-located search over multiple devices e.g. shared computers and Bluetooth enabled mobile devices. SearchTogether (Morris & Horvitz, 2007) allows remote collaboration by providing components such as instant messaging, split-screen search, etc., and asynchronous collaboration by enabling persistence storage. Coagmento (Shah, 2010a) utilised a combination of components from SearchTogether and previous research to provide asynchronous, remote and co-located collaboration on both computers and mobile devices for CIS. Coagmento 2.0 (Mitsui & Shah, 2016) was recently introduced with a number of improvements to the previous version (Shah, 2010a), such as tagging, filtering the tags and searching the tags. Coagmento 2.0 also allows other researchers to extend its functions and components as an open source tool. WeSearch (Morris et al., 2010) allows co-located search for up to four people on a tabletop. Whilst CoSearch, SearchTogether, Coagmento and WeSearch support text and web retrieval, there has also been research conducted into multimedia retrieval. For example, Halvey et al. (2010) developed a collaborative video retrieval system called ViGOR, which allows asynchronous and remote collaboration. Smeaton et al. (2007) developed a synchronous and co-located video retrieval system for a multi-user, touch sensitive tabletops.

Algorithmic mediation is widely used in the recommender systems (e.g. Amazon Shopping Recommendations (Linden, Smith, & York, 2003)). Example algorithmic mediated collaborative search systems include I-SPY (Smyth et al., 2004), Cerchiamo (Golovchinsky et al., 2008), etc. I-SPY (Smyth et al., 2004) is a community-based search engine that takes advantage of previous search behaviour of communities of searchers to re-rank future search results. Cerchiamo (Golovchinsky et al., 2008) is a synchronous collaborative search system that takes advantage of a complex algorithmic layer to leverage different roles within a search team and then splits up work based on the roles. Soulier et al. (2013) proposed an algorithm to re-rank and allocate documents towards the most suitable team member in a collaborative search team using a relevance feedback process. Through a simulated study, Soulier et al. also showed the effectiveness of their algorithm. Whilst certain collaborative search systems have a distinct type of mediation, some researchers have tried combining UI-only and algorithm mediation. For example, Freyne et al. (2007) implemented a system that integrates UI and algorithmic mediation by utilising previous search information to re-rank new results, and interactive icons to augment the results. A great deal of research that has been conducted to support collaborative search activities has focused on providing users not only with better results but also with better communication and collaboration capabilities.

A common assumption in both CIR and CIS is that every team member in a collaborative search team has equal access to underlying information. However, this may not always be the case. In practice, certain collaborative search teams may involve people with differing access to underlying information (Handel & Wang, 2011). An added complication is that such people may also have differing shareability of information between each other due to security and privacy reasons (Handel & Wang, 2011). Thus, many of the existing CIR and CIS concepts such as awareness, division of labour and persistence (Morris, 2007), as well as existing collaboration models such as communication, coordination, etc. (Shah, 2010b) may need to be revised for MLCIR.

2.2. Awareness

In the context of the WWW, awareness alone can be separated into a number of different kinds of issues such as: group awareness, workspace awareness, contextual awareness and peripheral awareness (Liechti & Sumi, 2002). The effect of supporting different kinds of awareness in CIS has been investigated in a number of research studies (McNeese & Reddy, 2015; Shah & Marchionini, 2010; Shah, 2013). Shah and Marchionini (2010) investigated the effects of different awareness types highlighted by Liechti and Sumi (2002) using three different search interfaces. They found that awareness of team actions and history provides advantages for collaborative search without adding new work to users. Shah (2013) examined the effects of awareness on users' coordination in collaborative search using three different search interfaces. Shah found that providing an adequate and appropriate amount of team awareness is beneficial for collaborative search compared to not providing any. McNeese and Reddy (2015) examined the development of team cognition during CIS using observations and interviews of participant teams engaged in co-located CIS tasks. They found that different awareness types: "search, information and social" can help team members obtain teamwork and taskwork knowledge which are important for developing team cognition. These research studies show the importance of awareness for effective collaborative search.

In an effort to support different awareness types for users during collaborative search activities, researchers have developed a number of systems. Example systems include SearchTogether (2007), CoSearch (2008) and Coagmento (2010). To support different awareness types highlighted by Liechti and Sumi (2002): group awareness, workspace awareness, contextual awareness and peripheral awareness, Shah and Marchionini (2010) designed a system for CIS, Coagmento, utilising a number of different interface components including query history with names and different colours, result history, common work space, etc. Morris and Horvitz (2007) implemented SearchTogether that supported awareness via query history with profile pictures, and page-specific metadata such as view information, ratings and comments. Amershi and Morris (2008) developed CoSearch system that included components such as query queue, result queue, user identity region, etc. As such, a number of research studies have focused on supporting different kinds of awareness for collaborative search activities.

However, awareness is a broad topic and investigations into supporting awareness would require several research studies. Since assumptions between CIR/CIS and MLCIR are different in terms of information access and shareability, not all awareness types and interface components used for previous collaborative search systems may be relevant for MLCIR systems. Besides, there is also a trade-off between supporting awareness and enforcing information security. Although providing users with every available piece of information seems like an ideal thing to do in traditional collaborative search, it may be impossible in MLCIR. A user study we recently conducted using 20 participants and 3 different information access scenarios suggested a number of awareness types for MLCIR (Htun et al., 2017); these are:

- query awareness,
- result awareness and
- team awareness

Query awareness includes providing a history of queries submitted by team members. Result awareness includes providing a history of interesting documents that are seen/judged/saved by team members. Team awareness includes providing clearly identified query history and seen/judged/saved documents by each team member. So far, there has not yet been a research study to investigate the impacts of any awareness types on MLCIR. Thus, using two separate user studies, we investigated the effect of query awareness, result awareness and team awareness on MLCIR.

2.3. Multi-Level Collaborative Information Retrieval

Some researchers (e.g. Attfield et al. (2010), Bjurling and Hansen (2010), and Karunakaran and Reddy (2012)) have begun to study the difficulties and complexities that arise in legal, government and healthcare domains. Attfield et al. (2010) presented a case study of a large London law firm, and discussed difficulties and complexities that may arise in current awareness networks, and provided interface design suggestions. Bjurling and Hansen (2010) observed a Swedish crisis management system and discussed inefficiencies in the collaborative network due to different interpretations and sharing of information. Karunakaran and Reddy (2012) described a number of case studies in the healthcare domain, and discussed frequent occurrences of non-uniform knowledge distribution and miscommunication.

However, these research studies did not provide explicit solutions for MLCIR. Handel and Wang (2011) discussed problems with MLCIR in domains such as healthcare, business and government, etc., and suggested a number of design considerations for MLCIR systems. However, these suggestions were based on experience and observations of Handel and Wang within Boeing rather than an empirical study.

In a separate thread of research, researchers such as Pickens et al. (2008), Shah et al. (2010), Soulier et al. (2014), Tamine and Soulier (2015), etc. have begun to study different user roles in collaborative search to help improve search performance. Pickens et al. (2008) developed algorithms to support user roles: "miner and prospector" for collaborative search. Following this, Shah et al. (2010) further developed algorithms to support "gatherer and surveyor" user roles for collaborative search. These research studies showed that supporting two different user roles for collaborative search (i.e. miner and prospector, gatherer and surveyor, etc.) allowed team members to find more relevant information in an efficient and effective way. Soulier et al. (2014) proposed different algorithms that monitor team members' actions and automatically suggest appropriate roles to optimise performance. Tamine and Soulier (2015) recently conducted a user study to understand the impact of role assignment into CIR and found that user roles limited the precision of the search results, demonstrating that user roles may sometimes negate search performance. Nevertheless, the primary focus of these research studies (Pickens et al., 2008; Shah et al., 2010; Soulier et al., 2014; Tamine & Soulier, 2015) have been on division of labour aspect of CIR (Kelly & Payne, 2013). Although some might argue that division of labour and MLCIR are similar, MLCIR is concerned with information security, flow, accessibility and shareability between collaborators (Handel & Wang, 2011) rather than distributing workload between team members.

In order to quantify the impact of non-uniform information access in CIR, we conducted a simulated user study (Htun et al., 2015) using a number of MLCIR scenarios that were highlighted by Handel and Wang (2011). Based on Handel and Wang's (2011) work, we devised four non-uniform information access scenarios, namely:

1) document removal,
2) random term blacklisting,
3) blacklisting most frequent terms in a query pool and
4) blacklisting most frequent terms in a document collection (Htun et al., 2015).

Figure 1 shows the experimental design of the simulated user study. The simulation was carried out based on the approach of Joho et al. (2009). We also used a number of collaborative search strategies and search topics proposed by Joho et al. (2009).

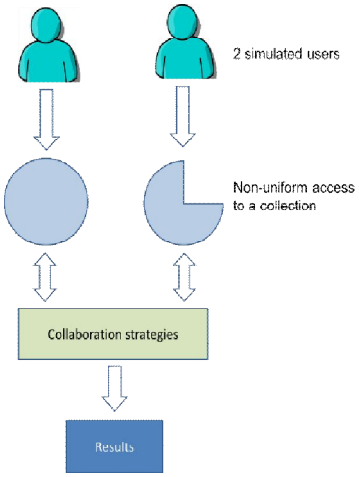


Figure 1. Experimental design of the simulated study (Htun et al., 2015)

The simulation was carried out as follow. For each search topic, individual team members submitted a random query selected from a query pool generated through a user evaluation (Joho et al., 2008). To simulate an actual user's judgement, the top 20 search results of individual team members were selected for each query submission. Individual team members searched 20 iterations per topic (i.e. 20 queries per individual). Thus, individual team members judged a maximum of 400 documents per topic (i.e. 20 documents x 20 iterations), with the team judging a maximum of 800 documents per topic. Search sessions were repeated 10 times in order to reduce randomness and inconsistencies.

For each of the non-uniform information access scenarios, we also formulated a number of possible access combinations for two simulated users (Htun et al., 2015). The access combinations determined the percentage of access level each simulated user had to the document collection in each information access scenario such as 10%-10%, 20%-10%, 20%-20%,...100%-90% and 100%-100%. This resulted in 55 possible access combinations for two users in each information access scenario (i.e. the combinations of 10%-10%, 10%-20%, 10%-30%, 10%-40%, etc. up to 100%-100%). Taking this into account, there were a total of 1,716,000 searches performed by each simulated user (i.e. 13 topics x 20 iterations x 55 access combinations x 4 information access scenarios x 3 search strategies x 10 runs).

Results from our simulated study highlighted the lowest possible access level a team can tolerate in each scenario without having a negative impact on search performance in comparison with the full access combination (i.e. 100%-100%). Although our simulated study was the first attempt to systematically evaluate the impact of MLCIR, it did not investigate the impact of different types of awareness on MLCIR search outcomes. To the best of our knowledge, no user studies have been conducted to investigate the impacts of awareness on MLCIR search outcomes. To address this gap, we conducted two separate user studies. These studies are outlined in the following two sections.

3. Study 1: Impact of Query Awareness on MLCIR

The aim of the first study was to investigate the impacts of query awareness on MLCIR search outcomes. We utilised three different information access scenarios, and three different search interfaces with varying support for query awareness.

Detailed explanations of the scenarios and interfaces utilised in this study are presented in Sections 3.2 and 3.5 respectively. The research questions we attempt to address in this study are:
S1-RQ1: How does support for query awareness impact collaborative search outcomes in MLCIR?
S1-RQ2: How does support for query awareness impact individual search outcomes in MLCIR?
S1-RQ3: How does support for query awareness impact users' search experience?
S1-RQ4: How can query awareness be better provided for MLCIR?

3.1 Document Collection and Search Tasks

With respect to document collection, most research studies in CIR and CIS have utilised either the Web (Amershi & Morris, 2008; Morris & Horvitz, 2007; Shah & González-Ibáñez, 2011) or test collections (Joho et al., 2008; Joho et al., 2009; Shah, Marchionini, & Kelly, 2009). In order to remove access to underlying information, the use of a test collection was more practical for our study. In addition, using test collections allowed us to accurately calculate traditional IR evaluation metrics: precision, recall and f-measure. We used the TREC HARD 2005 track's (Allan, 2005) test collection and topics for our study since they have successfully been utilised by a number of researchers in CIR, e.g. (Capra et al., 2012; Joho et al., 2008; Joho et al., 2009). The test collection used by the track was the AQUAINT corpus[1] which contains a total of 1,033,461 documents (about 3 GB) of newswire text data written in English (Allan, 2005). For 13 out of 50 test topics[2] of the track, Joho et al. (2008; 2009) generated a pool of queries which contains a list of query terms that were submitted by users for each topic. These terms represent the most likely search terms for each of the topics. We were provided with this query pool and it allowed us to blacklist search terms for users in our study (see Section 3.2 for details).

Based on Joho et al.'s (2009) work, we selected 10 out of their 13 topics with medium difficulty, which means these 10 topics had reasonably similar performance outcomes and number of relevant documents within the AQUAINT corpus. By using 10 topics, we had a broad selection of topics for users and were not tied to only certain part of the document collection. During the study, participants were presented with topics that were semi-randomly selected from the 10 candidate topics; while the topics were selected randomly, we manually ensured that the same topic did not repeat within a pair of participants. Table 1 presents the topic numbers (i.e. topic ID) and titles of the 10 topics.

Topic number	Title
303	Hubble telescope achievements
363	transportation tunnel disasters
383	mental illness drugs
393	mercy killing
397	automobile recalls
448	ship losses

[1] https://catalog.ldc.upenn.edu/LDC2002T31

[2] http://trec.nist.gov/data/hard/05/05.50.topics.txt

625	arrests bombing WTC
651	U.S. ethnic population
658	teenage pregnancy
689	family-planning aid

Table 1. Topic numbers and titles of 10 candidate topics

Each topic has a unique number, title, description and narrative; all were presented to study participants. For example, topic number 397 contains:

> **Title:** automobile recalls
>
> **Description:** Identify documents that discuss the reasons for automobile recalls.
>
> **Narrative:** A relevant document will specify major or minor reasons for automobile recalls by car manufacturers. Documents that discuss truck recalls are not relevant.

3.2 Information Access Scenarios and Access Combinations

The MLCIR scenarios used in this study are based on those utilised in our previous study (Htun et al., 2015); these scenarios are:

1) document removal and
2) term blacklisting based on their frequency in a query pool (see Table 2).

The document removal scenario (i.e. DR in Table 2) represents the scenario where access to documents in the collection is removed for some members. The term *blacklisting scenario* (i.e. TR in Table 2) represents the scenario where members do not find results if they search using certain blacklisted terms; the blacklisted terms in this case are the most frequent terms in the query pool. These two scenarios were selected for this study because they are the most likely scenarios in real life according to the MLCIR examples highlighted by Handel and Wang (2011). In addition to the two MLCIR scenarios, we included a full access scenario (i.e. FA in Table 2) which represents the case where both team members have full access to the collection, which is the typical assumed scenario in CIR and CIS research. During the study, pairs of participants performed searches using all 3 scenarios, but in order to avoid any order effects, the scenarios were counterbalanced using a Latin Square counterbalancing measure.

The access combinations in Table 2 represent the percentage of documents or terms left for a pair of searchers in the collection after a certain amount has been removed/blacklisted. In our simulated user study (Htun et al., 2015), for each of the two MLCIR scenarios (i.e. DR & TR), we devised a number of possible access combinations for two simulated users ranging from 10% access to the collection to 100% access to the collection e.g. 10%-10%, 20%-10%, 20%-20%,....100%-90% and 100%-100%. The combinations presented in Table 2 were selected based on the findings from our previous study (Htun et al., 2015) where at access combinations: 100%-60% (of DR) and 100-70% (of TR), search performance dropped significantly from the full access combination (i.e. 100%-100%). For the DR scenario in Table 2, one searcher has full access to the collection (i.e. 100% of documents) while the other has access to 60% of the documents in the collection which is precisely 620,077 documents (i.e. 40% of documents removed). For the TR scenario in Table 2, one

searcher can get results for all terms in the collection (i.e. 100% of terms) while the other can get results for only non-blacklisted terms in the collection (i.e. after 30% of the most frequent terms in the query pool has been blacklisted from the collection, thus represented as 70%). Based on Joho et al.'s (2009) query pool, 30% of the most frequent terms per topic was equivalent to 4 terms per topic on average. The access levels were rotated between a pair of participants after each scenario so that each participant had a chance to experience having full access and non-full access in the MLCIR scenarios.

Code	Access Scenario	Access Combination
DR	Document removal: remove access to documents from collection	100%-60%
TR	Term blacklisting: blacklist the most frequent terms in the query pool from collection	100%-70%
FA	Full access	100%-100%

Table 2. Information access scenarios with their respective access combinations

3.3 Participants

A total of 20 participants were recruited for the study through the university contacts. This sample size is in keeping with similar studies (Pickens et al., 2008; Shah et al., 2010; Smyth et al., 2004; Soulier et al., 2014). The participants were randomly assigned into pairs to form 10 groups. While some previous studies, e.g. (Morris & Horvitz, 2007; Tamine & Soulier, 2015), recruited participant pairs who had prior relationships, others recruited a mixture, e.g. (Joho et al., 2008; Soulier et al., 2014). We followed the latter approach because MLCIR scenarios, as highlight by Handel and Wang (2011), may occur between both known and unknown parties. Each participant received a £10 Amazon voucher; they were informed of this while being recruited. There were 5 females and 15 males. The average age of the participants was 28.2 (σ = 6.6) ranging from 18 to 44 years old. All of the participants were students; they were studying in a number of different subject areas including life science, engineering, computer science, etc. Amongst 20 participants, 5 reported that they usually spend 6 to 10 hours per week using search engines, 6 reported 11 to 15 hours per week, 3 reported 16 to 20 hours per week, and 6 reported more than 20 hours per week. 10 of the participants reported that they had taken part in a collaborative information search at least once, using either Google or Google Scholar.

3.4 Study Procedure

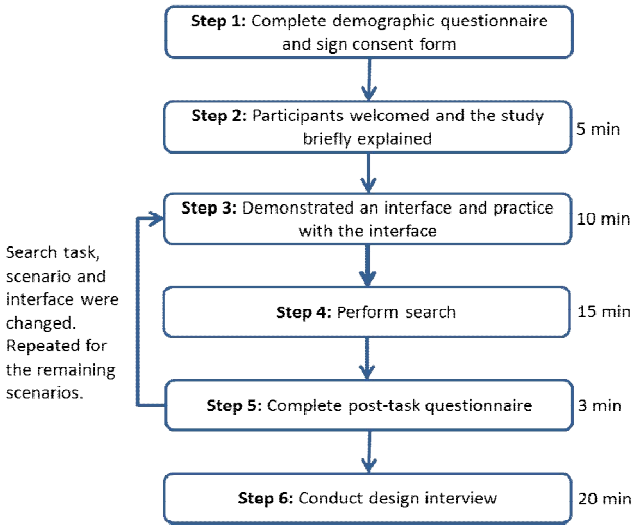


Figure 2. Study procedure for a pair of participants

A summary of the study procedure is highlighted in Figure 2. After a pair of participants was assigned randomly into a group, each participant was sent an email containing an information sheet, a consent form and a link to demographic questionnaire. They were instructed to read the information sheet, sign the consent form and then send back an electronic copy of the consent form. They were also instructed to complete the demographic questionnaire prior to arriving for the study.

Once a pair of participants arrived for the study, they were welcomed and the study was briefly explained. They were also instructed not to communicate during the search sessions because discussion of results, search strategies and documents could violate the evaluation of MLCIR scenarios. The participants were then told that their goal was to find as many relevant documents as possible for a given task within 15 minutes. They were informed of non-uniform access but not informed of which team member had more (or less) access to the collection than the other. The former was to inform the participants what was involved in the study and to be able to assess their perception of access during the study (see post-task questionnaire in Table 5). Not informing the participants which team member had more (or less) access reduced the possibility of a bias when answering question 1 of the post-task questionnaire which assessed the participants' perception of their access level relative to their partner's. Part of the script used for step was as follow:

> "In some scenarios, one of you will have less access to the results than the other. What that means is if you are the one with less access, it is likely that some of your search keywords will give you very little or no results. In that case, maybe try using different keywords."

Next, the participants were provided with a short demonstration of an interface using a search task which was randomly selected from the 10 candidate topics. The lead researcher explained in detail each component of the interface to both participants. The components were explained from left to right of the interface. To ensure consistency, a script was used. The participants were then given a few minutes to practice with the interface. The participants then searched for a maximum of 15 minutes in the same room using separate computers facing opposite directions (see Figure 3). Observations highlighted by Handel and Wang (2011) indicated that remote collaborations are common in MLCIR. To simulate a remote collaboration, we

followed the approach of Morris and Horvitz (2007) where participants were instructed not to communicate directly and to pretend that they were in different places. After 15 minutes, each participant was provided with a post-task questionnaire which was designed to obtain subjective assessments of individual participant's perception of their access level, search task, search performance and certain interface components (see Table 5 for the questions). The participants were given 3 minutes to complete the questionnaire. Once completed, they were given up to 5 minutes to rest in order to reduce any fatigue effects (note that a counterbalancing measure was already being used to control for amongst other things any fatigue effects). The scenario, task and interface were then changed and the participants were provided with a demonstration of a new interface using a new task. After practicing with the new interface, the participants performed another search session for a maximum of 15 minutes. The rest of the steps for this search session were as presented in Figure 2. Steps 3 to 5 were repeated for the remaining scenario and interface.

After all 3 sessions were completed, a design interview (see Section 3.6) was conducted with the pair of participants. The intention of the design interview was to obtain participants' qualitative feedback about individual interface components and also garner suggestions for new interface components.

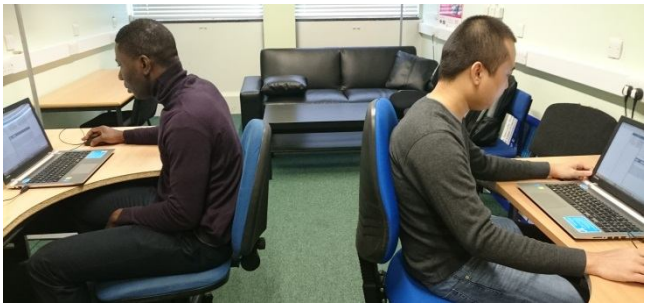


Figure 3. Experimental setup for a pair of participants

Table 3 highlights one complete rotation of the scenarios, access combinations and interfaces amongst the participants. The scenarios and interfaces were both rotated using a Latin-square counterbalancing measure whereas the access combinations between a pair of participants were rotated manually. Three types of interface evaluated in this study are explained in Section 3.5.

Pair ID	Scenario	Access combination	Interface Type
1	DR	100%-60%	1
	TR	70%-100%	2
	FA	100%-100%	3
2	TR	100%-70%	3
	FA	100%-100%	1
	DR	60%-100%	2
3	FA	100%-100%	2
	DR	100%-60%	3
	TR	70%-100%	1

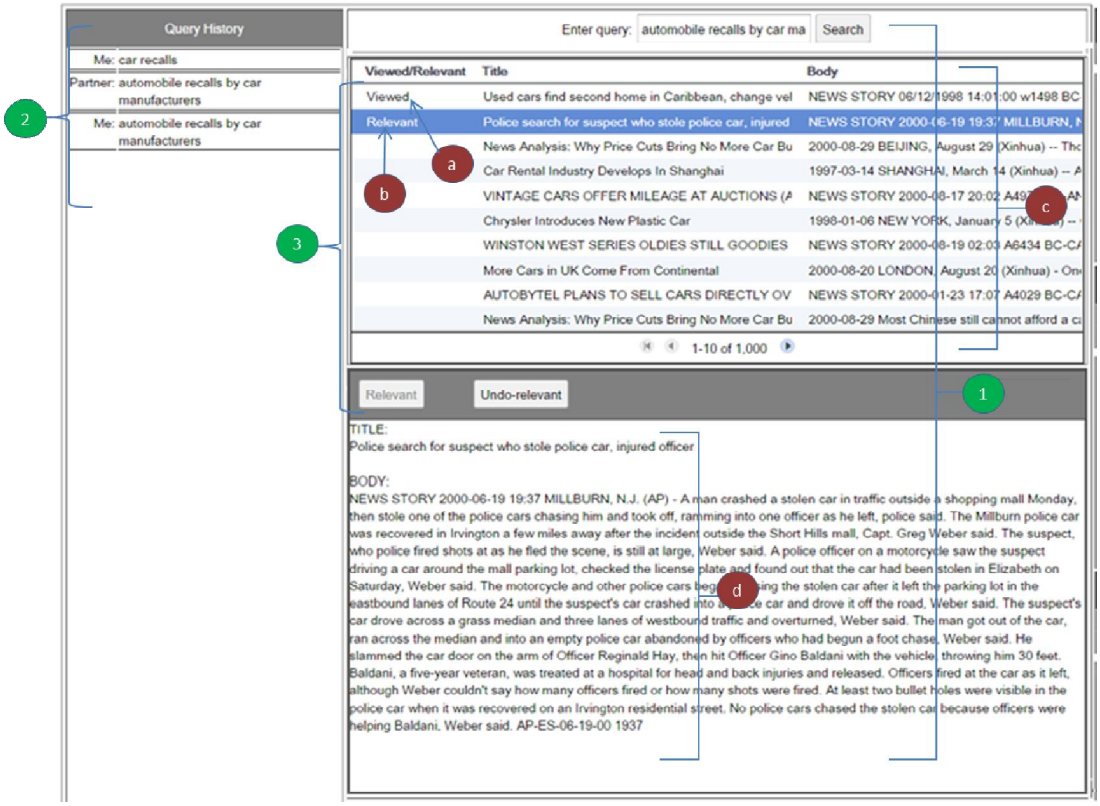
Table 3. One complete rotation of the scenarios, access combinations and interfaces

3.5 Interfaces

For this study, we implemented three different collaborative search interfaces with varying support for query awareness, which allowed a pair of users to judge documents synchronously. Each interface was designed to display a team's shared query history in a different way so that their effects can be compared (see Figure 5 (1, 2, 3)). The interfaces were implemented using Google Web Toolkit[3] and Terrier Toolkit[4]. During the study, participants were presented with a different interface in each access scenario as presented in Table 3. Details of the interfaces are explained in the following sub-Sections (3.5.1, 3.5.2 and 3.5.3).

3.5.1 Baseline Interface (Baseline)

The baseline interface contains three main components: 1) search component, 2) query history component and 3) viewed/judged component (see Figure 4). The search component (i.e. Figure 4 (1)) allows users to enter queries; the results are then displayed in the result list (i.e. Figure 4 (c)). Clicking on any result in the result list will display its contents in result detail (i.e. Figure 4 (d)). Query history component displays a list of shared query history (i.e. Figure 4 (2)). Users can resubmit any queries in the query history by simply clicking on them. They do not get any result if the submitted query is blacklisted. The viewed/judged component (i.e. Figure 4 (3)) provides functionalities to judge documents and see already judged documents (i.e. Figure 4 (b)) or already viewed documents (i.e. Figure 4 (a)) in their search results. Documents that are removed or documents that contain blacklisted keywords do not appear in search results (i.e. Figure 4 (c, d)).



[3] http://gwtproject.org

[4] http://terrier.org

Figure 4. Baseline interface: 1) search component, 2) query history component, 3) viewed/relevant component, a) "viewed" marking, b) "relevant" marking, c) result list, d) result detail

3.5.2 Interface with Icons (IWI)

The query history component of the interface with icons (IWI) interface adds three different query property icons in addition to the query history component of the Baseline interface. This component is shown in Figure 5 (2) where a number of example queries are displayed together with their property icons. Figure 5 (c) represents *query effectiveness* property, Figure 5 (d) represents *query popularity* property, and Figure 5 (e) represents *time spent on query* property. *Query effectiveness* property is measured the number of relevant documents found for a particular query. *Query popularity* property is measured by the number of times a particular query is used by the team. *Time spent on query* property is measured by the duration spent on a particular query before a new query is issued. The icons appear with varying levels of filling to represent the levels of effectiveness, popularity and time spent relative to the rest of the queries throughout an entire search session. A simple mouse over on each icon reveals its detail as shown in Figure 5 (f, g, h). A similar approach has also been utilised in a system implemented by Freyne et al. (2007) where icons augmented results-related properties, e.g. result popularity. Unlike the approach of Freyne et al. (2007), the icons in our interface augmented queries.

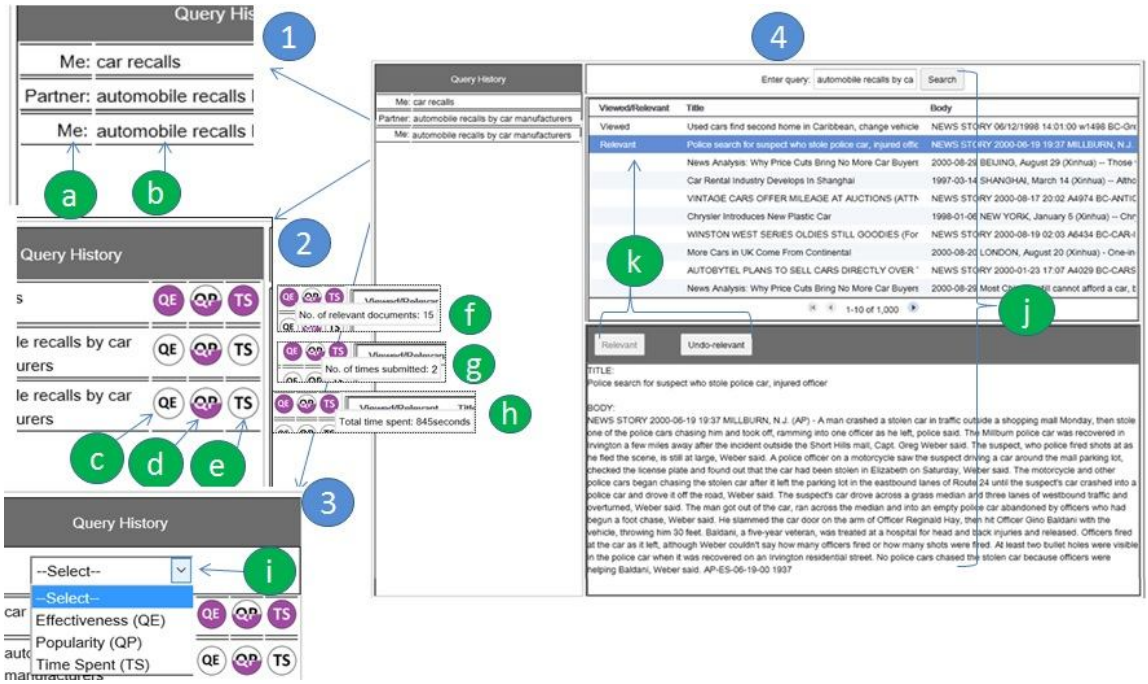


Figure 5. Search interface with various query history components: 1) query history component of the Baseline Interface, 2) query history component of the Interface with Icons, 3) query history component of the Interface with Icons and Sorting, 4) search interface, a) team members, b) queries, c) query effectiveness icons, d) query popularity icons, e) time spent on query icons, f) tooltip of the query effectiveness icons, g) tooltip of the query popularity icons, h) tooltip of the time spent on query icons, i) dropdown-list with three different sort criteria, j) search component, k) viewed/relevant component

3.5.3 Interface with Icons and Sorting (IWIS)

The query history component of the interface with icons and sorting (IWIS) interface adds a sort function in addition to the query history component of the IWI interface.

This component is shown in Figure 5 (3) where a number of example queries are displayed together with their property icons, and a dropdown-list with three different sort criteria (i.e. Figure 5 (i)). The sort criteria allow sorting of query history according to their properties: query effectiveness, query popularity and time spent on query.

3.6 Design Interview

The purpose of the design interview was to understand in detail how each of the query awareness components affected participants during the search sessions. The design interview also captured suggestions, related to query awareness, from participants so that the interface components could be improved for MLCIR. The interview was conducted with pairs of participants after evaluation. The interview consisted of 2 parts. In part 1, pairs of participants were asked a series of questions related to each of the query awareness components (i.e. in what way each component affected their search, and in what way each component could be redesigned and improved). In part 2, the pairs were asked to suggest new components and/or functionalities in order to improve query awareness. During the interview, participants were also provided with printouts of interface components, and empty sheets where they could sketch or annotate their ideas. Throughout the interview, participants' responses were recorded on an audio recorder, which were later transcribed for analysis. The transcripts were analysed using the Constant Comparative Method (CCM) (Glaser, Strauss, & Strutzel, 1968), a data analysis method of the Grounded Theory approach.

3.7 Data Gathering

The interfaces captured a log of participants' interaction with each component in the interface. This log was then used to calculate the evaluation metrics presented in Table 4.

Evaluation metrics	**Interfaces**
Recall	Baseline, IWI, IWIS
Precision	Baseline, IWI, IWIS
F-measure	Baseline, IWI, IWIS
Coverage	Baseline, IWI, IWIS
Relevant coverage	Baseline, IWI, IWIS
Unique coverage	Baseline, IWI, IWIS
Unique relevant coverage	Baseline, IWI, IWIS
Number of queries	Baseline, IWI, IWIS
Average query length	Baseline, IWI, IWIS
Query success	Baseline, IWI, IWIS
Number of viewed documents	Baseline, IWI, IWIS
Number of viewed documents by query	Baseline, IWI, IWIS
Number of clicks on query history	Baseline, IWI, IWIS
Duration spent on icons	IWI, IWIS
Number of times query history was sorted	IWIS

Table 4. Quantitative evaluation metrics measured in different interfaces

The TREC HARD 2005 (Allan, 2005) topics have a non-exhaustive list of relevant and non-relevant documents against the AQUAINT corpus. This list is known as qrel

(query relevance)[5] and it was used to calculate a number of evaluation metrics as explained in the followings. To measure search performance, we used traditional IR evaluation metrics:

- *recall*,
- *precision* and
- *f-measure*

Recall is the number of true positive documents amongst all the documents judged by a team/an individual in each search session divided by the number of relevant documents in the qrel. *Precision* is the number of true positive documents amongst all the documents judged by a team/an individual in each search session divided by the number of all of the documents that are judged by the team/the individual as relevant. *F-measure* is a harmonic mean of recall and precision which is represented by the formula: $\frac{2\,.\,\text{precision}\,.\,\text{recall}}{(\text{precision} + \text{recall})}$.

In addition, we also adapted a number of evaluation metrics proposed by Shah and González-Ibáñez (2011) for CIS:

- *coverage*,
- *relevant coverage*,
- *unique coverage* and
- *unique relevant coverage*

Coverage is the number of distinct documents discovered by a team/an individual in each search session. Shah and González-Ibáñez (2011) calculated coverage by using the number of distinct documents viewed by participants. We used the documents discovered by participants which are different from viewed documents. For example, if a participant browsed up to page 2 and viewed only one document, we assumed that the participant discovered 20 documents (i.e. 10 documents per page). *Relevant coverage* is the number of documents in the *coverage* that intersect with relevant documents in the qrel. *Unique coverage* is the number of distinct documents discovered by a team/an individual only in a given search session (e.g. for the DR scenario with the IWI interface), and not in any others. Unlike us, Shah and González-Ibáñez (2011) defined unique coverage as a unique region within coverage that were viewed only by a team/an individual and not by any others. *Unique relevant coverage* is the total number of documents in the *unique coverage* that intersect with relevant documents in the qrel.

Other evaluation metrics we adapted were based on those proposed by Soulier et al. (2014):

- *number of queries*,
- *average query length*,
- *query success*,
- *number of viewed documents* and
- *number of viewed documents by query*

Number of queries is the total number of queries submitted by a team/an individual in each search session. *Average query length* is the average number of words within the *number of queries*. *Query success* is the number of true positive documents (successful documents) divided by the *number of queries* submitted by a team/an

[5] http://trec.nist.gov/data/hard/05/TREC2005.qrels.txt

individual in each search session. The true positive documents were calculated based on the qrel. Unlike us, Soulier et al. (2014) assumed that the documents where participants spent over 30 seconds as the true positive documents (successful documents). This was because Soulier et al. (2014) used the web and did not have access to a qrel to precisely calculate true positive documents. *Number of viewed documents* is the number of documents that were clicked to read by a team/an individual in each search session. This is different from *coverage* in which all the documents up to the lowest possible rank that a team/an individual scrolled were considered. *Number of viewed documents by query* is the *number of viewed documents* divided by the *number of queries* submitted by a team/an individual in each search session.

We also analysed the *number of clicks on each query history* for all the interfaces. For the IWI and IWIS interfaces, the *duration spent (hovering mouse) on icons* was also analysed. In addition, for the IWIS interface, we measured the *number of times query history was sorted*.

The post-task questionnaire provided at the end of each search session was used to obtain subjective assessments of individual participant's perception of their access, search tasks, search performance and interface components (see Table 5 for the questions). The questionnaire was in the form of 5-point Likert scales and the answers ranged from 1 (strongly disagree) to 3 (neither) to 5 (strongly agree). Most of the questions were based on a number of similar research (Freyne et al., 2007; Joho et al., 2008), and each question provided a different understanding of participants' perception of their access, search tasks, search performance and interface components. The first question of the post-task questionnaire (Q1) captured participants' perception of their access level relative to their partner. Q2 to Q4 captured participants' perceptions of search tasks. Q5 to Q8 captured participants' perceptions of search performance. Q9 to Q11 captured participants' perceptions of query property icons. Q12 captured participants' perceptions of query sort function. Q2 to Q6 were based on those investigated by Joho et al. (2008) whereas Q9 to Q11 were based on those investigated by Freyne et al. (2007). (Please note that for the Baseline interface, only Q1 to Q8 were presented to the participants whereas for the IWI interface, Q1 to Q11 were presented. For IWIS, all 12 questions were presented).

		Questions	**Interfaces**
Assessment of participants' access	Q1	I think I had higher access than my partner.	Baseline, IWI, IWIS
Assessment of search task	Q2	The instruction of this task is easy to understand.	Baseline, IWI, IWIS
	Q3	The topic of this task is interesting.	Baseline, IWI, IWIS
	Q4	I was familiar with the topic of this task.	Baseline, IWI, IWIS
Assessment of search performance	Q5	I am satisfied with the documents obtained for my queries for this task.	Baseline, IWI, IWIS
	Q6	I am confident with the documents I judged for this task.	Baseline, IWI, IWIS

		Questions	Interfaces
	Q7	I think my team found a lot of relevant documents for this task.	Baseline, IWI, IWIS
	Q8	I think I found more relevant documents than my partner for this task.	Baseline, IWI, IWIS
Assessment of query property icons	Q9	The *'query effectiveness'* (QE) icons were helpful for this task.	IWI, IWIS
	Q10	The 'query popularity' (QP) icons were helpful for this task.	IWI, IWIS
	Q11	The *'time spent on query'* (TS) icons were helpful for this task.	IWI, IWIS
Assessment of query sort function	Q12	The ability to sort query history was helpful for this task	IWIS

Table 5. Questions of the post-task questionnaire. Q1 = assessment of participants' access. Q2 to Q4 = assessment of search task. Q5 to Q8 = assessment of search performance. Q9 to Q11 = assessment of query property icons. Q12 = assessment of query sort function

3.8 Study 1 Results

For the evaluation metrics described in Section 3.7, comparisons were made between the three interfaces within each scenario (e.g. Baseline vs. IWI vs. IWIS within the DR scenario) and within the full access and non-full access of the non-uniform access scenarios (e.g. Baseline vs. IWI vs. IWIS of individuals with non-full access within the DR scenario). The independent variable was the interface with three levels: Baseline, IWI and IWIS. The dependent variables included all evaluation metrics presented in Table 4 (except for *number of times query history was sorted*) and all questions from the post-task evaluation questionnaire presented in Table 5 (except for Q12).

A one-way ANOVA was used for normally distributed data. Prior to one-way ANOVA, Levene's test was carried out to check for homogeneity of variance. The standard one-way ANOVA assumes that different tested sets of data have similar (or homogeneous) internal levels of variance. Where this assumption did not hold (i.e. if homogeneity of variances assumption was violated), we employed a Welch's ANOVA instead. Welch's ANOVA is an alternative analysis of variance method to the standard one-way ANOVA, and is used when homogeneity of variances assumption is violated. When doing the post-hoc analyses of significant ANOVA results, the standard Tukey test was used following the standard one-way ANOVA, while a Games-Howell test was used following Welch's ANOVA. In addition, Bonferroni corrections were applied to control the type-1 error rate.

For non-normally distributed data, a Kruskal-Wallis H test was used. SPSS automatically performs the post-hoc analyses for Kruskal-Wallis H test using Dunn-Bonferroni post-hoc test which is based on Dunn's (1964) work and controls the type-1 error rate (IBM, 2014). In the following sub-sections, we present detailed results of the statistical analysis and design interview.

3.8.1 Search Performance

Results of *recall*, *precision* and *f-measure* indicated that there was no significant difference in search performance between the three interfaces for the DR and TR

scenarios (S1-RQ1). However, for the FA scenario, results indicated that *precision* was significantly different between the interfaces (Welch's F (2, 3.51) = 15.1, p = 0.019). As shown in Table 6, the Baseline interface had the highest *precision* whereas the IWI interface had the lowest. However, post-hoc analysis, using the Games-Howell test with Bonferroni correction, revealed that *precision* was not statistically different between the interfaces (S1-RQ1). *Recall* and *f-measure* for the FA scenario had no significant difference between the three interfaces (S1-RQ1). For full access and non-full access within DR and TR scenarios, there was no significant difference in *recall*, *precision* and *f-measure* between the three interfaces (S1-RQ2). Mean values of all search performance metrics are presented in Appendix A, Table A.1.
In addition, to visualise the highest possible recall that the participants could have achieved during the study, we analysed *highest possible recall* calculated as *relevant coverage* divided by the number of relevant documents in the qrel for a given task (see Appendix A, Table A.1). The distribution between *recall* and *highest possible recall* metrics showed that perfect recall was not achieved by the participants and that more relevant documents could be found. This means that the participants missed quite a number of relevant documents although they browsed through these documents. A similar finding was reported by Joho et al. (2008) who explained that the time constraint (i.e. 15 minutes) could be a factor.

Further, we analysed the average *percentage of relevant documents in the collection* for each information access scenario (see Appendix A, Table A.1). Note that for the DR scenario, the collections had 620077 documents (60% out of 1033461 documents) whereas for the TR and FA scenarios, the collections had 1033461 documents. The TR scenario had the same number of documents as the FA scenario because unlike in DR, only terms were removed in TR, thus the document count remained the same as FA. The results showed that the percentage of relevant documents available in the collections was generally the same between the conditions. Some differences (e.g. 0.0117% and 0.0094%) could be accounted for the random distribution of topics. However, this did not have an impact on search performance between the information access scenarios (see recall, precision and f-measure).

3.8.2 Query Submission and Documents Viewed

Results showed that for the DR and TR scenarios (both collaboratively (S1-RQ1) and individually (S1-RQ2)), there was no significant difference between the three interfaces in terms of *number of queries*, *average query length*, *query success*, *number of viewed documents* and *number of viewed documents by query*. For the FA scenario, however, results indicated that *query success* and *number of viewed documents by query* were significantly different between the interfaces ($\chi^2(2) = 6.587$, p = 0.037 & ANOVA F (2,7) = 9.85, p = 0.009 respectively). As shown in Table 6, the Baseline interface had the highest *query success* and *number of viewed documents by quer*y.

		FA scenario		
		Baseline	IWI	IWIS
Precision	Mean	0.70	0.05	0.39
	S.D	0.20	0.05	0.36
	Median	0.80	0.05	0.40
Query success	Mean	**1.04**	**0.03**	0.23
	S.D	**0.14**	**0.03**	0.23
	Median	**1.09**	**0.05**	0.24

		FA scenario		
		Baseline	IWI	IWIS
No. of viewed documents by query	Mean	**6.82**	3.05	**2.35**
	S.D	**1.88**	0.43	**1.40**
	Median	**6.71**	3.09	**2.10**

Table 6. Comparison between the interfaces for the Full Access scenario (S1-RQ1). Bold = statistically different pairs (p < .05). S.D = standard deviation.

Post-hoc analysis using Dunn-Bonferroni test revealed that the Baseline interface had significantly better *query success* than the IWI interface (p = 0.036), as well as significantly higher *number of viewed documents by query* than the IWIS interface (p = 0.027) (S1-RQ1). These results demonstrate that the participants submitted a similar number of queries between the three interfaces for all scenarios. They, however, had the least query success within the IWI interface and read the lowest number of documents within the IWIS interface for the full access scenario. Mean values of all query submission and documents viewed metrics are presented in Appendix A, Table A.1.

3.8.3 Collection Coverage

Results showed that for all cases (both collaboratively (S1-RQ1) and individually (S1-RQ2)), there was no significant difference between the interfaces in terms of *coverage*, *relevant coverage*, *unique coverage* and *unique relevant coverage*. This suggests that the participants had similar collection coverage outcomes between the three interfaces for all scenarios. Mean values of all collection coverage metrics are presented in Appendix A, Table A.1.

3.8.4 Usage

The log also recorded usage of each interface such as: *number of clicks on query history*, *duration spent on icons* and *number of times query history was sorted*. Results indicated that in all cases (both collaboratively (S1-RQ1) and individually (S1-RQ2)), *number of clicks on query history* was not significantly different between the three interfaces. Similarly, *duration spent on icons* was not statistically different between IWI and IWIS. Thus, the participants used the common components between the three interfaces in a similar manner. No pairwise comparisons were made for *number of times query history was sorted* since sorting is present only in IWIS. Mean values of all interface usage metrics are presented in Appendix A, Table A.1.

3.8.5 Participants' Perceptions

Participants' perceptions of their access, search task, search performance and interface components were captured by the post-task questionnaire (see Table 5 for questions). Results indicated that for the DR and TR scenarios (both collaboratively (S1-RQ1) and individually (S1-RQ2)), all questionnaires had no significantly different answers between the interfaces. However, for the FA scenario, the results indicated that answers for *perception of higher access* (Q1: $\chi^2(2) = 6.61$, p = 0.037), *result satisfaction* (Q5: $\chi^2(2) = 7.68$, p = 0.021), *confidence in judgement* (Q6: $\chi^2(2) = 6.07$, p = 0.048) and *perception of team performance* (Q7: $\chi^2(2) = 9.6$, p = 0.008) were significantly different (S1-RQ1). As shown in Table 7, the IWI interface had the highest *perception of higher access* (Q1) whereas the Baseline interface had the highest *result satisfaction* (Q5), *confidence in judgement* (Q6) and *perception of team performance* (Q7). Post-hoc analysis using Dunn-Bonferroni test revealed that the

IWI interface had significantly higher *perception of higher access* (Q1) than the IWIS interface (p = 0.041). The Baseline interface had significantly higher *result satisfaction* (Q5), *confidence in judgement* (Q6) and *perception of team performance* (Q7) than the IWIS interface (p = 0.017, 0.043 & 0.021 respectively). Besides, the Baseline interface had significantly higher *perception of team performance* (Q7) than the IWI interface (p = 0.019). A result summary of all the questions is provided in Appendix A, Table A.2.

		FA scenario		
		Baseline	IWI	IWIS
Q1	Mean	3.7	**4.0**	**2.8**
	S.D	0.5	**0.6**	**1.0**
	Median	4.0	**4.0**	**3.0**
Q5	Mean	**4.2**	3.2	**2.5**
	S.D	**0.8**	1.2	**1.1**
	Median	**4.0**	3.5	**3.0**
Q6	Mean	**4.3**	3.8	**3.1**
	S.D	**0.5**	0.8	**1.1**
	Median	**4.0**	4.0	**3.0**
Q7	Mean	**4.2**	**2.5**	2.6
	S.D	**0.8**	**0.8**	0.9
	Median	**4.0**	**3.0**	3.0

Table 7. Comparison between the interfaces for the Full Access scenario (S1-RQ1). (see Table 5 for questions). 1 = strongly disagree, 3 = neither, 5 = strongly agree. Bold or underlined = statistically different pairs (p < .05). S.D = standard deviation

3.8.6 Design Interview

A qualitative analysis of design interview responses using CCM (Glaser et al., 1968) resulted in 3 main themes: knowledge of partner's performance, knowledge of better queries, and improvements (addressing our research questions: S1-RQ3 & S1-RQ4). Details of these themes are presented in the following sub-sections.

3.8.6.1 Knowledge of partner's performance

It appears that certain properties of a query can help users obtain knowledge of their team members' search performance without sharing any documents. 6 of the participants reported that just by using the *time spent on query* property, they were able to tell that their partners were finding relevant documents. On the other hand, 2 other participants reported that the *query effectiveness* property helped them understand their partners' performance. In addition, 1 participant reported that the *time spent on query* and *query effectiveness* properties, when combined, were most helpful for this particular case. For example, the participant explained: "I can see how much time my partner has spent on this query and check QE (query effectiveness) to see if my partner has found relevant documents." (P16).

3.8.6.2 Knowledge of better queries

We found that the *query effectiveness* and *query popularity* properties helped the participants improve their queries. This indicates that the *query effectiveness* and *query popularity* properties can help users obtain knowledge of better queries without sharing any documents. 14 of the participants reported that the *query effectiveness* property was most helpful for improving their queries. As one of the participants

explained, "The first thing I looked at when I searched. This (query effectiveness) gave me an idea of what other terms to use." (P6). On the other hand, 4 of the participants reported that *query popularity* property was most helpful. For example, one of the participants made the following comment: "It (query popularity) helps me find more results because I know this query is popular." (P20).

3.8.6.3 Improvements

Regarding query property icons, 10 participants suggested displaying the actual numbers (i.e. number of relevant documents, number of times submitted and time spent) on the respective icons instead of our tooltip function. It appears that displaying the numbers on icons could allow users to quickly identify query properties. In addition, the use of different colours (e.g. red and green) was also suggested in the place of fill-up by 7 participants. Colours such as red and green are distinct, and are widely accepted to represent variations e.g. low and high, bad and good, etc. Therefore, a number of different colours can be used in the place of low, median and high fill-up levels (e.g. red, yellow and green).

On the other hand, 8 participants reported that having to check each of the query properties was a visually demanding and time consuming task. In order to address this issue, we suggest a balanced query score that accounts for all our three query properties. While it is based on f-measure, the balanced query score can be interpreted as an average of the three query properties, where the higher a score reaches the better a query is. The balanced score of a query $Score_q$ can be calculated as:

$$Score_{q=} \frac{(R_q \times T_q) + P_q}{R_q + T_q + P_q}$$

where R_q is the number of relevant documents found for query q, T_q the duration spent on query q, and P_q the number of times query q has been submitted. This single score can then be displayed in the place of our existing query property icons. It can also be augmented with different colours (as discussed earlier), and a tooltip function to display details of the three query properties.

With regard to the query sort function, 6 participants reported that the current design was confusing and/or involves quite a lot of steps. To improve it, participants suggested a sort function that is similar to sorting tables by clicking on column headers.

4. Study 2: Impact of Result Awareness and Team Awareness on MLCIR

The aim of the second study was to investigate the impacts of result awareness and team awareness on MLCIR search outcomes. Result awareness and team awareness were investigated as one study because at the time the study was being conducted, not many different CIR and CIS interface components had been proposed for either result awareness or team awareness in comparison with query awareness. Morris and Horvitz (2007), Amershi and Morris (2008), and Shah and Marchionini (2010) implemented some interface components that support result awareness and team awareness. Using some of these components, we implemented two interfaces building on top of our baseline interface (see Section 3.5.1 for the baseline interface):

- one for result awareness and

- one for team awareness

The same document collection, search tasks, information access scenarios, access combinations and study procedure from study 1 were used in study 2. Please refer to Sections 3.1, 3.2 and 3.4 for detailed explanations. The research questions we attempt to address in this study are:
S2-RQ1: How does support for result awareness and team awareness impact collaborative search outcomes in MLCIR?
S2-RQ2: How does support for result awareness and team awareness impact individual search outcomes in MLCIR?
S2-RQ3: How does support for result awareness and team awareness impact users' search experience?
S2-RQ4: How can result awareness and team awareness be better provided for MLCIR?

4.1 Participants

A new set of 20 participants were recruited for this study. The participants were randomly assigned into pairs. Each participant received a £10 Amazon voucher; they were informed of this while being recruited. There were 7 females and 13 male participants. The average age of the participants was 29.8 (σ = 5.97) ranging from 22 to 47 years old. All of the participants were students. They were studying in a number of different subject areas including marine engineering, life science, computer science and business. Of 20 participants, 1 reported that he/she usually spends less than 6 hours per week using search engines, 5 reported 6 to 10 hours per week, 2 reported 11 to 15 hours per week, 5 reported 16 to 20 hours per week and 7 reported more than 20 hours per week. 16 participants reported that they had taken part in a collaborative search at least once using tool such as Google, Facebook, phone, email and university library systems.

4.2 Interfaces

Three different collaborative search interfaces were used for this study; these are: baseline interface, result awareness interface and team awareness interface. Details of the interfaces are explained in the following sub-sections (4.2.1, 4.2.2 and 4.2.3). During the study, the three interfaces were counterbalanced using a Latin-square counterbalancing measure.

4.2.1 Baseline Interface (Baseline)

The baseline interface comprised three main components:
1) search component,
2) query history component and
3) viewed/judged component.
It is exactly the same as the baseline interface from study 1. A full explanation of the baseline interface is provided in Section 3.5.1.

4.2.2 Result Awareness Interface (RA)

Figure 6 highlights components of the result awareness interface. The documents viewed and/or marked as relevant by users are kept in two separate lists (i.e. “Viewed documents” list and “Relevant documents” list) as shown in Figure 6 (1). Documents can also be bookmarked by clicking on the “Bookmark” button (i.e. Figure 6 (a)). The

documents that are bookmarked by both users are kept in the "Bookmarked documents" (i.e. Figure 6 (2)). These three lists can be seen as a history of viewed, relevant and bookmarked documents. Each user can only see documents that they have access to, or documents that do not contain blacklisted keywords for them. Clicking on a document in any lists displays the full contents of the respective document as shown in Figure 7. We also considered implementing a result recommendation component that was successfully used by other researchers (Morris & Horvitz, 2007; Shah, 2010a). However, given that users in MLCIR scenarios are unaware of their access limitation, as well as other team members', a result recommendation could be misleading.

Search results can be sorted by using the dropdown list as shown in Figure 6 (3). The sorting criteria are default, viewed, and relevant. In Figure 6, the results are sorted by a "viewed" criteria. Previous research has utilised approaches such as re-ranking results based on previous search information (Freyne et al., 2007) and re-ranking results based on user roles (Shah et al., 2010). However, unlike these approaches, our sort function is explicit (i.e. triggered by users) and utilises viewed/relevant properties of the documents.

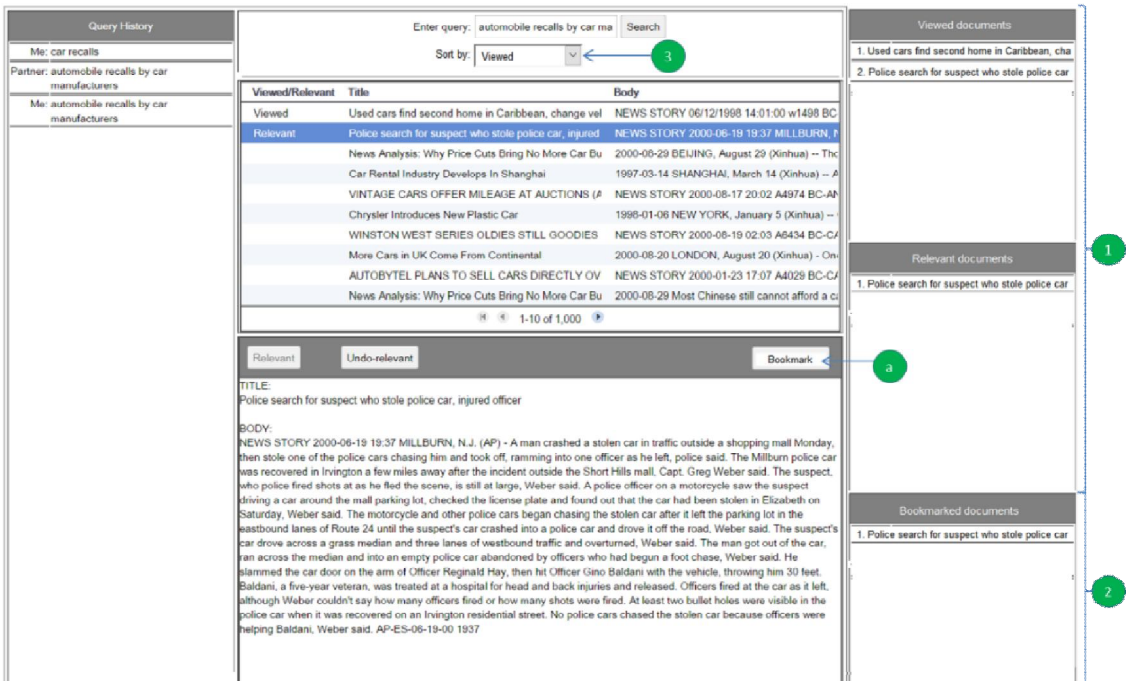


Figure 6. Result awareness interface: 1) viewed and relevant documents lists, 2) bookmarked documents list, 3) result sorting function, and a) bookmark button

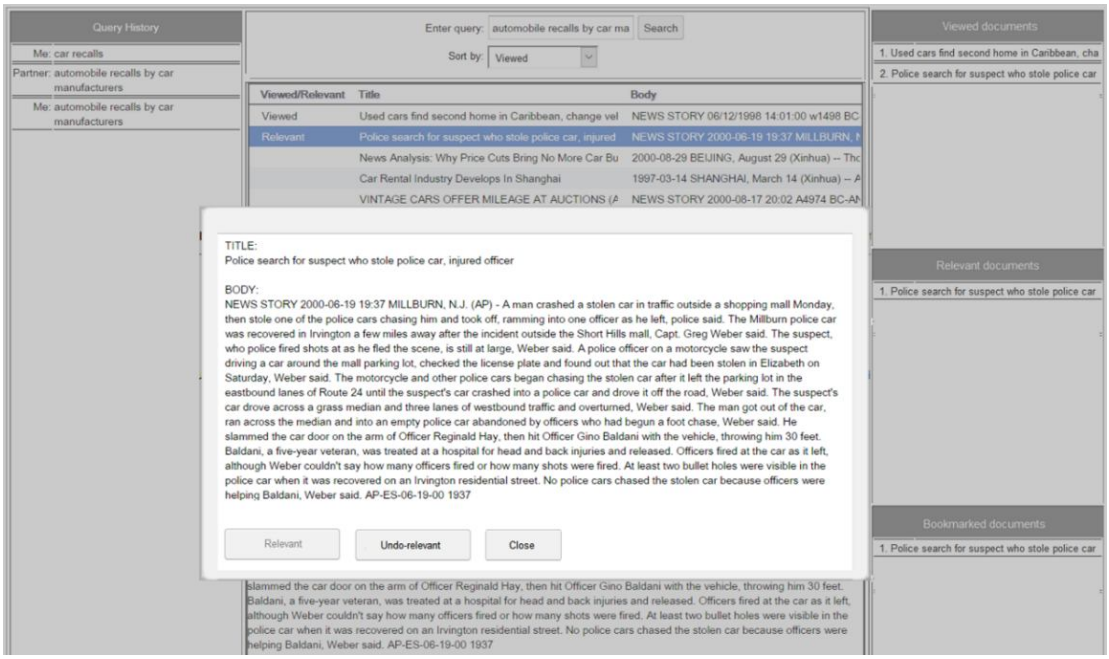


Figure 7. Result awareness interface 2: full contents of a document are displayed once it is clicked in any of the lists (i.e. "Viewed documents", "Relevant documents" and "Bookmarked documents" lists)

4.2.3 Team Awareness Interface (TA)

Figure 8 highlights components of the team awareness interface. As shown in Figure 8 (a), query histories are displayed in two separate lists and are differentiated by different colours according to team members. For the viewed/judged component, team members who viewed and judged documents are differentiated using their respective colours and initials (see Figure 8 (b)). A collaborative search system: CoSearch (2008) successfully utilised different colours and names to highlight different users in both query history and search result.

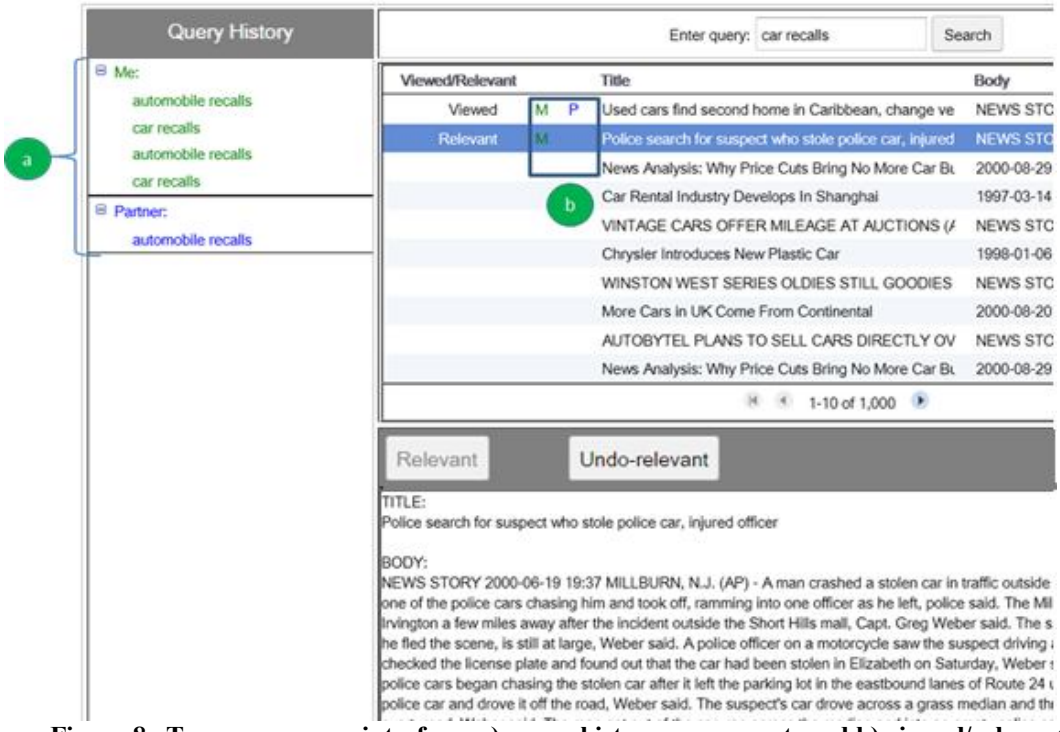


Figure 8. Team awareness interface: a) query history component, and b) viewed/relevant component with initials of "Me" and "Partner"

4.3 Design Interview

The design interview in study 2 was aimed to capture qualitative feedback related to result awareness and team awareness interfaces. The interview was conducted with pairs of participants after evaluation. The interview comprised 3 parts: in part 1, the participant pairs were asked questions related to each of the result awareness components. In part 2, the pairs were asked questions related to each of the team awareness components. The questions asked in parts 1 and 2 include: in what way each component affected their search, and in what way each component could be redesigned and improved. In part 3, the pairs were asked to suggest new components and/or functionalities to improve result awareness and team awareness. Participants were provided with printouts of interface components, and empty sheets to sketch or annotate their ideas. Participants' responses were recorded on an audio recorder. The responses were later transcribed and analysed using the CCM (Glaser et al., 1968).

4.4 Data Gathering

A log of participants' interaction with each of the interfaces was recorded, which was later used to calculate the evaluation metrics presented in Table 8. Most of the evaluation metrics used in study 2 are exactly the same as those used in study 1. For the RA interface, we additionally calculated:

- *number of times results were sorted*,
- *number of clicks on viewed documents list*,
- *number of clicks on relevant documents list*,
- *number of clicks on bookmarked documents list* and
- *number of clicks on bookmark button*

Evaluation metrics	Interfaces
Recall	Baseline, RA, TA
Precision	Baseline, RA, TA
F-measure	Baseline, RA, TA
Coverage	Baseline, RA, TA
Relevant coverage	Baseline, RA, TA
Unique coverage	Baseline, RA, TA
Unique relevant coverage	Baseline, RA, TA
Number of queries	Baseline, RA, TA
Average query length	Baseline, RA, TA
Query success	Baseline, RA, TA
Number of viewed documents	Baseline, RA, TA
Number of viewed documents by query	Baseline, RA, TA
Number of clicks on query history	Baseline, RA, TA
Number of times results were sorted	RA
Number of clicks on viewed documents list	RA
Number of clicks on relevant documents list	RA
Number of clicks on bookmarked documents list	RA
Number of clicks on bookmark button	RA

Table 8. Quantitative evaluation metrics measured in different interfaces

As in study 1, a post-task questionnaire was provided at the end of each search session. The aim of the questionnaire was to capture subjective assessments of individual participant's perception of their access, search tasks, search performance and interface components (see Table 9 for the questions). The first question (Q1) captured participants' perception of their access-level relative to their partner, Q2 to Q4 captured their perception of the search tasks, Q5 to Q8 captured their perception of their search performance. Q9 to Q12 were presented to the participants only in the RA interface and captured participants' perception of the RA interface components. Similarly, Q13 to Q16 were presented only in the TA interface and captured participants' perception of the TA interface components.

		Questions	Interfaces
Assessment of participants' access	Q1	I think I had higher access than my partner.	Baseline, RA, TA
Assessment of search task	Q2	The instruction of this task is easy to understand.	Baseline, RA, TA
	Q3	The topic of this task is interesting.	Baseline, RA, TA
	Q4	I was familiar with the topic of this task.	Baseline, RA, TA
Assessment of search performance	Q5	I am satisfied with the documents obtained for my queries for this task.	Baseline, RA, TA
	Q6	I am confident with the documents I judged for this task.	Baseline, RA, TA
	Q7	I think my team found a lot of relevant documents for this task.	Baseline, RA, TA
	Q8	I think I found more relevant documents than my partner for this task.	Baseline, RA, TA
Assessment of result awareness interface	Q9	The result-sort function was helpful.	RA
	Q10	Having a list of viewed documents was helpful.	RA
	Q11	Having a list of relevant documents was helpful.	RA
	Q12	The bookmark function was helpful.	RA
Assessment of team awareness interface	Q13	Having different colours for me and my partner in query history was helpful.	TA
	Q14	Having separated lists of query history for me and my partner was helpful.	TA
	Q15	Having different colours for me and my partner for viewed/relevant documents was helpful.	TA
	Q16	Having initials for me and my partner for viewed/relevant documents was helpful.	TA

Table 9. Questions of the post-task questionnaire. Q1 = assessment of participants' access. Q2 to Q4 = assessment of search task. Q5 to Q8 = assessment of search performance. Q9 to Q12 = assessment of result awareness interface. Q13 to Q16 = assessment of team awareness interface.

4.5 Study 2 Results

Pairwise comparisons were made between the three interfaces (i.e. baseline vs. result awareness vs. team awareness) within each scenario, as well as within the full access and non-full access of the non-uniform access scenarios. The same statistical analysis approach as study 1 was used which includes: an ANOVA test and a Kruskal-Wallis H test, together with their respective post-hoc comparisons (see Section 3.8). Detailed results of the statistical analysis and design interview are presented in the following sub-sections.

4.5.1 Search Performance

Comparisons of recall, precision and f-measure between the interfaces (i.e. Baseline, RA & TA) showed no significant difference within the scenarios (S2-RQ1). However, there was a significant difference between the interfaces within the full access of term blacklisting (TR) scenario in terms of *precision* ($\chi^2(2) = 6.635$, $p = 0.036$). As shown in Table 10, post-hoc analysis using Dunn-Bonferroni test revealed that the RA interface has significantly higher precision than the Baseline interface (p

= 0.045) (S2-RQ2). Mean values of all search performance metrics are presented in Appendix B, Table B.1.

To visualise the highest possible recall that the participants could have achieved during the study, we analysed *highest possible recall* calculated as *relevant coverage* divided by the number of relevant documents in the qrel for a given task (see Appendix B, Table B.1). The distribution between *recall* and *highest possible recall* metrics indicated that perfect recall was not achieved by the participants and that the participants missed quite a number of relevant documents although they browsed through these documents (i.e. the same finding as study 1 (see Section 3.8.1)).

In addition, we analysed the *percentage of relevant documents in the collection* (see Appendix B, Table B.1). The result showed that the percentage of relevant documents available in the collection was generally the same between the conditions. Note that the *percentage of relevant documents in the collection* was the same as study 1 (see Section 3.8.1) because we use the same materials for both studies.

		Full access of TR scenario		
		Baseline	RA	TA
Precision	Mean	**0.0**	**0.73**	0.21
	S.D	**0.0**	**0.39**	0.37
	Median	**0.0**	**0.88**	0.0

Table 10. Comparison between the interfaces within the full access of the term *blacklisting scenario* (S2-RQ2). Bold = statistically different pair (p < .05). S.D = standard deviation

4.5.2 Query Submission and Documents Viewed

Results showed that for all cases (both collaboratively (S2-RQ1) and individually (S2-RQ2)), there was no significant difference between the interfaces in terms of query submission and documents viewed, measured by: *number of queries*, *average query length*, *query success*, *number of viewed documents* and *number of viewed documents by query*. This suggests that the number of queries submitted and documents read was similar between the three interfaces for all scenarios, demonstrating that input from the participants was the same between the interfaces. Mean values of all query submission and documents viewed metrics are presented in Appendix B, Table B.1.

4.5.3 Collection Coverage

No significant difference was found between the interfaces for all cases (both collaboratively (S2-RQ1) and individually (S2-RQ2)) in terms of collection coverage, measured by: *coverage*, *relevant coverage*, *unique coverage* and *unique relevant coverage*. Again, this suggests that the participants had similar collection coverage outcomes between the three interfaces for all scenarios. Mean values of all collection coverage metrics are presented in Appendix B, Table B.1.

4.5.4 Usage

Usage of the interfaces was measured by the metrics: *number of clicks on query history, number of times results were sorted*, *number of clicks on viewed documents list*, *number of clicks on relevant documents list*, the *number of clicks on bookmarked documents list*, and the *number of clicks on bookmark button*. No significant difference was found between the interfaces for all cases in terms of *number of clicks on query history* (S2-RQ1 and S2-RQ2). It means the usage of the three interfaces

was the same for all the scenarios. No pairwise comparisons were made for the rest of the usage metrics since they were measured only in the RA interface. However, mean values of all interface usage metrics can be found in Appendix B, Table B.1.

4.5.5 Participants' Perceptions

Post-task questionnaire questions utilised in this study are presented in Table 9 . Statistical analysis results showed that questions Q1 to Q8 had no significant different answers between the interfaces within all three scenarios (S2-RQ1). For individuals with full access of term blacklisting (TR) scenario, however, the results indicated that scores for *task's easiness to understand* (Q2: $\chi^2(2) = 6.644$, $p = 0.036$) and *task's familiarity* (Q4: $\chi^2(2) = 7.094$, $p = 0.029$) were significantly different (S2-RQ2). As shown in Table 11, the RA interface had the highest scores in both Q2 ($p = 0.032$) and Q4 ($p = 0.025$). A result summary of all the questions is provided in Appendix B, Table B.2.

		Full access of TR scenario		
		Baseline	RA	TA
Q2	Mean	**3.33**	**5.00**	4.33
	S.D	**1.15**	**0.00**	0.58
	Median	**4.00**	**5.00**	4.00
Q4	Mean	3.33	**4.50**	**1.67**
	S.D	0.58	**0.58**	**1.15**
	Median	3.00	**4.50**	**1.00**

Table 11. Comparison between the interfaces within the full access of the term *blacklisting scenario* (S2-RQ2) (see Table 9 for questions). 1 = strongly disagree, 3 = neither, 5 = strongly agree. Bold = statistically different pair ($p < .05$). S.D = standard deviation

4.5.6 Design Interview

A qualitative analysis of design interview responses using CCM (Glaser et al., 1968) resulted in 3 main themes: knowledge of better queries and results, reducing visual load, and improvements (addressing our research questions: S2-RQ3 & S2-RQ4). Details of these themes are presented in the following sub-sections.

4.5.6.1 Knowledge of better queries and results

We found that having access to a history of viewed, relevant and bookmarked documents helped the participants obtain better queries and results. While all 20 participants provided positive remarks for having access to the result history, 8 participants explicitly described their experience where they obtained better queries and search results through the result history. For example, one participant mentioned: "When I was searching through, once I couldn't find [results], I was able to go back and re-read [the result history], and then looked for the keywords." (P11). This might have allowed the participants to reformulate their queries which in turn led them to find more related results for their search topics. One participant explained: "I read our previous relevant documents and based on that, I found other similar documents as well." (P13). This finding indicates that having access to a history of intersecting viewed, relevant and bookmarked documents can help users in MLCIR scenarios to share their expertise without the necessity to disclose sensitive information.

4.5.6.2 Reducing visual load

Although participants liked having access to their result history, the result awareness interface was criticised for its complex design. For example, one participant described: "It's very clunky, it almost kind of distracts you from what you are doing, so maybe having an option to click [the result history] and maybe just have less information." (P19). When presented with the team awareness interface, all 20 participants noted that they preferred its simple design and the ability to easily identify other team members. For example, one participant explained: "I found [team awareness] interface way better to use. It is clearer. If you could add [the result history] panel to [the team awareness] interface, that would overall make [the team awareness] interface the best to use." (P12). Another participant described: "It was good to reduce the visual load. [team awareness interface] gave most instantaneous impression of what's going on. The colour coding makes it stand out more." (P2). It is possible that in the result awareness interface, participants felt overwhelmed by a large amount of information being presented without much team information. Thus, obtaining a balance between different awareness kinds may be crucial for MLCIR systems since too much information means more cognitive load for users and could also lead to an unintended disclosure of sensitive information.

4.5.6.3 Improvements

With regard to the result awareness interface, 6 participants suggested displaying the person viewed, judged or bookmarked the documents in the respective components. For example, one participant explained: "Maybe if the 'partner and me' [identification] function, if that was included here, that would make this more efficient." (P11). In addition, 4 participants suggested displaying the query terms that were used to obtain the documents displayed in the viewed, relevant and bookmarked lists. One participant clarified: "For example, if you show the keyword that my partner used to [obtain] relevant documents, I might be able to use it because I know this keyword is leading [my partner] to find this relevant document." (P1). On the other hand, 8 of the participants thought that the viewed, relevant and bookmarked lists were taking an unnecessarily large amount of space. Therefore, they suggested that users should be able to hide the lists if they wish to. One participant said: "I think users should be able to hide [the lists]. If someone wants to see it, they can expand it." (P5). Finally, 10 participants suggested that the result sorting function should also allow criteria such as: sort by person, sort by popularity, sort by bookmark and sort by date. For example, one participant explained: "The way sort would have been important to me is if there was a way I could sort those [results] based on popularity, date, etc." (P2).

With regard to the team awareness interface, 5 participants suggested that the query history could also display total results returned, number of viewed documents and number of relevant documents for each query. Some of these suggestions have already been implemented and investigated in study 1. 8 of the participants raised a concern regarding viewed/relevant component of the team awareness interface, in which one participant said: "When you have many people, I don't know how you could fit initials into the columns" (P20). For this particular scenario, the same 8 participants also provided suggestions such as: to "use the first two initials instead of just one" (P15), and to "distinguish me and the other people" (P3).

5. DISCUSSION

5.1 Awareness vs. Collaborative Search Outcomes

In relation to the first research question for study 1 (S1-RQ1): "How does support for query awareness impact collaborative search outcomes in MLCIR?", it was found that within the full access scenario (FA), the baseline interface had significantly higher *query success* than the interface with icons (IWI) and a significantly higher *number of viewed documents by query* than the interface with icons & sorting (IWIS). In terms of participants' perception, within the FA scenario, participants had significantly higher perception of access in the IWI interface condition than they did in the IWIS interface condition. For the same scenario (i.e. FA), in terms of result satisfaction and confidence in judgement, the baseline interface had a significantly higher score than the IWIS interface. Besides, participants' perception of team performance was significantly higher in the baseline interface condition than both IWI and IWIS conditions for the FA scenario. These findings suggest that all three interfaces (i.e. Baseline, IWI & IWIS) had similar search performance and collection coverage within our three access scenarios. In relation to the first research question for study 2 (S2-RQ1): "How does support for result awareness and team awareness impact collaborative search outcomes in MLCIR?", the results again suggested that there were no significantly different collaborative search outcomes between the interfaces (i.e. Baseline, result awareness (RA) & team awareness (TA)). Thus, current awareness interfaces did not have considerable improvements over our baseline interface in terms of collaborative search outcomes. Comparisons between the IWI, IWIS, RA and TA interfaces were made to find out whether certain awareness type helps most in terms of collaborative search outcomes. None of the three awareness types significantly outperformed each other.

Study 1 had six participant pairs who knew each other prior to the study whereas study 2 had seven participant pairs who knew each other prior to the study. To check whether the two categories (i.e. the participants who knew each other and those who did not prior to the study) had any significant differences in search outcomes, comparisons were made between the two categories. In neither study did we find any significant differences in search outcomes between the two categories. Since the participants were not allowed to communicate, it appears that their prior knowledge about each other did not have a significant impact on search outcomes. In the next section, we discuss the impact on individual search outcomes.

5.2 Awareness vs. Individual Search Outcomes

In relation to the second research question for study 1 (S1-RQ2): "How does support for query awareness impact individual search outcomes in MLCIR?", our results showed that there were no significantly different outcomes between the interfaces (i.e. Baseline, IWI & IWIS) within full access and non-full access of both document removal (DR) and term blacklisting (TR) scenarios. In relation to the second research question for study 2 (S2-RQ2): "How does support for result awareness and team awareness impact individual search outcomes in MLCIR?", results showed that for individuals with full access of the TR scenario, the RA interface had significantly higher *precision* compared to the baseline interface This could be related to the high scores on *task's easiness to understand* and *task's familiarity* (see Table 11). Comparisons between the IWI, IWIS, RA and TA interfaces showed that none of the

three awareness types significantly outperformed each other in terms of individual search outcomes.

So far, we found that in almost every case, the awareness interfaces did not outperform our baseline interface. This is possibly due to the simplicity of the awareness interfaces we utilised. Nevertheless, these interfaces have formed a starting point for further investigations on MLCIR, and our findings show that there is room for improvement for these interfaces. To understand how each of the awareness interfaces effect users' search experience and to provide design recommendations for further improvements, we discuss findings from the design interviews in the following sub-sections.

5.3 Impact on Users' Search Experience

In relation to the third research question for study 1 (S1-RQ3): "How does support for query awareness impact users' search experience?", we found that *time spent on query* and *query effectiveness* properties helped users obtain knowledge of their team members' search performance. Using these properties, information about users' search performance can be exchanged without disclosing any sensitive data. It was also found that *query effectiveness* and *query popularity* properties helped users obtain knowledge of better queries, allowing them to improve their own queries. Thus, these properties could be used to provide an implicit suggestion for better queries without disclosing any sensitive data. Besides, as Harvey et al. (2015) found, looking at high-quality query examples can help users create queries that are highly effective. While the majority of the participants thought that query property icons were helpful during the search sessions, some reported that query sort function was not as useful as the icons themselves. Due to the small team size and the short duration of time allowed for each task (i.e. 15 minutes), it is possible that most participants did not need to use the sort function as much.

In relation to the third research question for study 2 (S2-RQ3): "How does support for result awareness and team awareness impact users' search experience?", we discuss result awareness and team awareness separately. With regard to result awareness, it was found that having access to a history of intersecting viewed, relevant and bookmarked documents helped users to obtain knowledge of better possible queries. The participants also described a number of examples where they actually found better results after looking into the history of documents. Therefore, similar to the query properties, the history of intersecting documents could help users work together without disclosing any sensitive data. In the matter of team awareness, we found that the participants gave positive remarks about the interface due to its simple design and the ability to easily identify other team members. Since being aware of team members also means being aware of their roles and shareable and non-shareable information, it is important for users to have a clear view of team members' information and actions. To sum up, we believe that MLCIR systems should try to obtain a balance between different awareness types as too much information means more cognitive load for users and could also lead to an unintended disclosure of sensitive information. As Handel and Wang (2011) suggested, perhaps MLCIR systems must be "conceptually easier for users to understand and use" (pp. 5). This also suggests that mental load of the users in different MLCIR scenarios need to be assessed systematically in the future.

5.4 Design Recommendations

In relation to the fourth research question for study 1 (S1-RQ4): "How can query awareness be better provided for MLCIR?" and the fourth research question for study 2 (S2-RQ4): "How can result awareness and team awareness be better provided for MLCIR?", we discuss design recommendations based on the findings from our design interviews.

In terms of query awareness, query property icons could display actual numbers instead of a tooltip function. Also instead of fill-up, the icons could use a number of different colours such as red, yellow and green. To make query property icons simpler, it is also possible to combine all the properties into a single balanced score. This score can then be augmented with different colours and/or a tooltip function to display details of the properties. As for the query sort function, instead of using a dropdown list, sorting could be allowed using column headers. For accessibility, we believe that the query history component must be collapsible and expandable.

In terms of result awareness, the result sorting function must provide not only widely used sort criteria such as: date, relevancy, etc., but also other criteria such as: popularity, person, etc. that are specific to collaborative scenarios. To integrate result awareness with query awareness and team awareness, the history of viewed, relevant and bookmarked documents could display the search terms used for each document, and the person who viewed, judged or bookmarked each document. To reduce complexity, we believe that this could be implemented using a tooltip function. For accessibility, the history of viewed, relevant and bookmarked documents must also be collapsible and expandable.

In terms of team awareness, we believe that a summary of team members' information such as their roles and contact must be available in both query history and viewed/judged document components. Such information could help users identify team members easily throughout search sessions. To reduce complexity, this could also be implemented using a tooltip function. This tooltip function could also greatly help larger groups.

Awareness plays an important role in CIR and CIS systems, and there may not be restrictions on how much awareness information can be presented to users. However, when designing MLCIR systems, one must pay great attention not to overwhelm users with too much information since this could lead to information overload and in the worst cases, unintended disclosure of sensitive information. Keeping the interface as simple as possible whilst providing controlled awareness information is important for developing easy-to-use MLCIR systems.

6. LIMITATIONS

Since this was the first study looking to understand the impact of awareness on MLCIR, there were a number of limitations. First, as discussed previously, at the time our studies were being conducted, there were no design recommendations for MLCIR interfaces that had emerged from a user study. Therefore, we designed simple interfaces that provided just enough functionality to support query awareness, result awareness and team awareness to serve as a starting point. Most likely because of this, the awareness interfaces in almost every case did not outperform the baseline

interface. However, results from the design interviews provided us with a number of important design suggestions for improvement. Second, due to limited resources, we were only able to recruit 20 participants for each study. Therefore, our quantitative results may need to be confirmed using a larger sample size. Third, since we utilised the TREC HARD 2005 track's (Allan, 2005) collection, our quantitative results could be slightly different for other document collections. However, it is important to note that the TREC HARD 2005 track's collection and topics are well-established and have been utilised widely to evaluate CIR outcomes (Capra et al., 2012; Joho et al., 2008; Joho et al., 2009). Finally, we did not present in this paper detailed analysis of the impact of each topic. The focus of our studies was mainly on the MLCIR scenarios and interfaces. Others, e.g. (Joho et al., 2008), have followed a similar analysis procedure.

7. CONCLUSION AND FUTURE WORK

This paper presents the first known attempt to investigate the effect of different awareness types on MLCIR. We conducted two separate user studies utilising three different information access scenarios that were highlighted by Handel and Wang (2011) and were also used in our previous study (Htun et al., 2015). The first study investigated query awareness and the second study investigated result awareness and team awareness. The information access scenarios include two non-uniform information access scenarios (document removal & term blacklisting) and one full access scenario. The three interfaces for study 1 were a baseline interface, an interface with icons to illustrate query properties, and an interface that combined those icons with a query sorting function; each interface provided a team's shared query history in a different way. The three interfaces for study 2 were a baseline interface which is the same as study 1, a result awareness interface and a team awareness interface. Retrieval evaluations and design interviews were conducted using pairs of participants.

Generally, evaluation results from study 1 suggested that for the full access scenario, the baseline interface had positive outcomes in comparison to the rest of the interfaces. Evaluation results from study 2 suggested that the result awareness interface had significantly higher *precision* compared to the baseline interface for individuals with full access of the term *blacklisting scenario*. Search outcomes between all of the interfaces we utilised were comparable to each other in the rest of the cases.

Based on the feedback from the design interviews, we also presented a number of findings with regard to users' search experience in MLCIR. We found that query awareness, especially query properties such as *time spent on query*, *query popularity* and *query effectiveness* provided users with information about team members' search performance and an implicit suggestion for better queries without disclosing any sensitive data. Similarly, we found that result awareness such as having access to a history of intersecting viewed, relevant and bookmarked documents had the same positive effect as query awareness. In terms of team awareness, we found that being able to easily identify different actions of different team members in the simplest form was preferred by users.

Finally, we provided a number of design suggestions in terms of query awareness, result awareness and team awareness. In general, an ideal MLCIR system could

integrate all these three awareness types and provide a seamless collaborative search experience for users. However, it is important not to overload users with too much information because this could hinder users' performance and/or unintended disclosure of sensitive information. Therefore, MLCIR systems must be simple, while providing useful awareness information to users.

In the future, we plan to investigate further into MLCIR to provide a well-established framework of concepts that can be used to implement MLCIR systems. Overall, our findings in this paper provide a clear understanding of the effect of different awareness types on MLCIR. We anticipate that the design suggestions we provided will help other researchers develop new MLCIR interfaces and allow further investigation of MLCIR.

8. References

Allan, J. (2005). Hard Track Overview in TREC 2005 High Accuracy Retrieval from Documents. *Proceedings of the 14th Text REtrieval Conference*. NIST Special Publication, pp. 500–266.

Amershi, S., & Morris, M. R. (2008, April). CoSearch: a system for co-located collaborative web search. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1647-1656). ACM.

Attfield, S., Blandford, A., & Makri, S. (2010). Social and interactional practices for disseminating current awareness information in an organisational setting. Information processing & management, 46(6), 632-645.

Bjurling, B., & Hansen, P. (2010, May). Contracts for information sharing in collaborative networks. In Proc. 7th Int'l Conf. Information Systems Crisis Response and Management.

Capra, R., Chen, A. T., Hawthorne, K., Arguello, J., Shaw, L., & Marchionini, G. (2012). Design and evaluation of a system to support collaborative search. Proceedings of the American Society for Information Science and Technology, 49(1), 1-10.

De Choudhury, M., Morris, M. R., & White, R. W. (2014, April). Seeking and sharing health information online: comparing search engines and social media. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems (pp. 1365-1376). ACM.

Dunn, O. J. (1964). Multiple comparisons using rank sums. Technometrics, 6(3), 241-252.

Foley, C., & Smeaton, A. F. (2010). Division of labour and sharing of knowledge for synchronous collaborative information retrieval. Information processing & management, 46(6), 762-772.

Foster, J. (2006). Collaborative information seeking and retrieval. Annual review of information science and technology, 40(1), 329-356.

Freyne, J., Farzan, R., Brusilovsky, P., Smyth, B., & Coyle, M. (2007, January). Collecting community wisdom: integrating social search & social navigation. In Proceedings of the 12th international conference on Intelligent user interfaces (pp. 52-61). ACM.

Glaser, B. G., Strauss, A. L., & Strutzel, E. (1968). The Discovery of Grounded Theory; Strategies for Qualitative Research. Nursing Research, 17(4), 364.

Golovchinsky, G., Adcock, J., Pickens, J., Qvarfordt, P., & Back, M. (2008). Cerchiamo: a collaborative exploratory search tool. Proceedings of Computer Supported Cooperative Work (CSCW), 8-12.

Golovchinsky, G., Pickens, J., & Back, M. (2009). A taxonomy of collaboration in online information seeking. arXiv preprint arXiv:0908.0704.

González□Ibáñez, R., & Shah, C. (2011). Coagmento: A system for supporting collaborative information seeking. proceedings of the American Society for Information Science and Technology, 48(1), 1-4.

Halvey, M., Vallet, D., Hannah, D., Feng, Y., & Jose, J. M. (2010). An asynchronous collaborative search system for online video search. Information processing & management, 46(6), 733-748.

Handel, M. J., & Wang, E. Y. (2011, October). I can't tell you what i found: problems in multi-level collaborative information retrieval. In Proceedings of the 3rd international workshop on Collaborative information retrieval (pp. 1-6). ACM.

Hansen, P., & Järvelin, K. (2005). Collaborative information retrieval in an information-intensive domain. Information Processing & Management, 41(5), 1101-1119.

Harvey, M., Hauff, C., & Elsweiler, D. (2015, August). Learning by Example: training users with high-quality query suggestions. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 133-142). ACM.

Htun, N. N., Halvey, M., & Baillie, L. (2015, August). Towards Quantifying the Impact of Non-Uniform Information Access in Collaborative Information Retrieval. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 843-846). ACM.

Htun, N. N., Halvey, M., & Baillie, L. (2017, March). How can we better support users with non-uniform information access in collaborative information retrieval? *ACM SIGIR Conference on Human Information Interaction & Retrieval (CHIIR)*. ACM.

IBM. (2014). IBM post hoc comparisons for the kruskal-wallis test - United Kingdom. Retrieved from http://www01.ibm.com/support/docview.wss?uid=swg21477370

Joho, H., Hannah, D., & Jose, J. M. (2008, October). Comparing collaborative and independent search in a recall-oriented task. In Proceedings of the second international symposium on Information interaction in context (pp. 89-96). ACM.

Joho, H., Hannah, D., & Jose, J. M. (2009, April). Revisiting IR techniques for collaborative search strategies. In European Conference on Information Retrieval (pp. 66-77). Springer Berlin Heidelberg.

Karunakaran, A., & Reddy, M. (2012, October). The role of narratives in collaborative information seeking. In Proceedings of the 17th ACM international conference on Supporting group work (pp. 273-276). ACM.

Kelly, R., & Payne, S. (2013, February). Division of labour in collaborative information seeking: Current approaches and future directions. In The 3rd International Workshop on Collaborative Information Seeking, held at ACM CSCW 2013. University of Bath.

Liechti, O., & Sumi, Y. (2002). Editorial: Awareness and the WWW. *International Journal of Human-Computer Studies*, *56*(1), 1-5.

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, *7*(1), 76-80.

McNeese, N. J., & Reddy, M. C. (2015). The role of team cognition in collaborative information seeking. *Journal of the Association for Information Science and Technology*.

Mitsui, M., & Shah, C. (2016, June). Coagmento 2.0: A System for Capturing Individual and Group Information Seeking Behavior. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (pp. 233-234). ACM.

Morris, M. R. (2007, January). Collaborating alone and together: Investigating persistent and multi-user web search activities. In *Proceedings of international ACM SIGIR conference on research and development in information retrieval (SIGIR 2007)* (pp. 23-27).

Morris, M. R. (2008, April). A survey of collaborative web search practices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1657-1660). ACM.

Morris, M. R. (2013, February). Collaborative search revisited. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 1181-1192). ACM.

Morris, M. R., & Horvitz, E. (2007, October). SearchTogether: an interface for collaborative web search. In *Proceedings of the 20th annual ACM symposium on User interface software and technology* (pp. 3-12). ACM.

Morris, M. R., Lombardo, J., & Wigdor, D. (2010, February). WeSearch: supporting collaborative search and sensemaking on a tabletop display. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 401-410). ACM.

Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., & Back, M. (2008, July). Algorithmic mediation for collaborative exploratory search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 315-322). ACM.

Shah, C. (2010a). Coagmento-a collaborative information seeking, synthesis and sense-making framework. *Integrated demo at CSCW*, 6-11.

Shah, C. (2010b). Collaborative information seeking: A literature review. *Advances in librarianship*, *32*(2010), 3-33.

Shah, C. (2012). *Collaborative information seeking: The art and science of making the whole greater than the sum of all* (Vol. 34). Springer Science & Business Media.

Shah, C. (2013). Effects of awareness on coordination in collaborative information seeking. *Journal of the American Society for Information Science and Technology*, *64*(6), 1122-1143.

Shah, C. (2015). Collaborative Information Seeking: From 'What?'and 'Why?'to 'How?'and 'So What?'. In *Collaborative Information Seeking* (pp. 3-16). Springer International Publishing.

Shah, C., & González-Ibáñez, R. (2011, July). Evaluating the synergic effect of collaboration in information seeking. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 913-922). ACM.

Shah, C., & Marchionini, G. (2010). Awareness in collaborative information seeking. *Journal of the American Society for Information Science and Technology*, *61*(10), 1970-1986.

Shah, C., Marchionini, G., & Kelly, D. (2009, April). Learning design principles for a collaborative information seeking system. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems* (pp. 3419-3424). ACM.

Shah, C., Pickens, J., & Golovchinsky, G. (2010). Role-based results redistribution for collaborative information retrieval. *Information processing & management*, *46*(6), 773-781.

Shah, C. (2016). The blind leading the blind: Impromptu leaderships influenced by awareness in collaborative search. *Aslib Journal of Information Management*, *68*(2), 212-226.

Smeaton, A. F., Lee, H., Foley, C., & McGivney, S. (2007). Collaborative video searching on a tabletop. *Multimedia Systems*, *12*(4-5), 375-391.

Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M., & Boydell, O. (2004). Exploiting query repetition and regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction*, *14*(5), 383-423.

Soulier, L., Shah, C., & Tamine, L. (2014, July). User-driven system-mediated collaborative information retrieval. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 485-494). ACM.

Soulier, L., Tamine, L., & Bahsoun, W. (2013, December). A collaborative document ranking model for a multi-faceted search. In *Asia Information Retrieval Symposium* (pp. 109-120). Springer Berlin Heidelberg.

Soulier, L., Tamine, L., & Shah, C. (2016). MineRank: Leveraging users' latent roles for unsupervised collaborative information retrieval. *Information Processing & Management*.

Spence, P. R., Reddy, M. C., & Hall, R. (2005, November). A survey of collaborative information seeking practices of academic researchers. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work* (pp. 85-88). ACM.

Tamine, L., & Soulier, L. (2015, October). Understanding the impact of the role factor in collaborative information retrieval. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 43-52). ACM.

Tamine, L., & Soulier, L. (2016, March). Collaborative Information Retrieval: Concepts, Models and Evaluation. In *European Conference on Information Retrieval* (pp. 885-888). Springer International Publishing.

Tamine, L., Soulier, L., Jabeur, L. B., Amblard, F., Hanachi, C., Hubert, G., & Roth, C. (2016, July). Social Media-Based Collaborative Information Access: Analysis of Online Crisis-Related Twitter Conversations. In *ACM 27th Conference on Hypertext & Social Media*.

Appendix A: Study 1 results

		DR	TR	FA	DR		TR	
					Full	Non-full	Full	Non-full
recall	Baseline	0.03	0.02	0.08	0.01	0.02	0.01	0.01
	IWI	0.07	0.04	0.01	0.03	0.04	0.03	0.02
	IWIS	0.07	0.01	0.04	0.04	0.02	0.01	0.00
highest possible recall	Baseline	0.11	0.07	0.23	0.10	0.07	0.04	0.04
	IWI	0.24	0.16	0.10	0.17	0.15	0.12	0.06
	IWIS	0.23	0.06	0.17	0.20	0.08	0.03	0.03
percentage of relevant documents in the collection	Baseline	-	-	-	0.0106%	0.0111%	0.0106%	0.0115%
	IWI	-	-	-	0.0094%	0.0117%	0.0094%	0.0103%
	IWIS	-	-	-	0.0104%	0.0102%	0.0104%	0.0111%
precision	Baseline	0.26	0.47	0.71	0.05	0.24	0.18	0.42
	IWI	0.48	0.40	0.05	0.47	0.48	0.33	0.35
	IWIS	0.46	0.25	0.39	0.29	0.55	0.37	0.00
f-measure	Baseline	0.06	0.04	0.15	0.01	0.04	0.03	0.02
	IWI	0.11	0.08	0.01	0.06	0.07	0.05	0.03
	IWIS	0.11	0.02	0.07	0.07	0.04	0.02	0.00
no. of queries	Baseline	12.00	17.00	8.67	4.25	9.00	9.67	8.67
	IWI	18.33	17.50	18.33	13.67	7.00	12.50	6.25
	IWIS	18.67	19.00	18.00	11.67	10.00	10.33	13.00
average query length	Baseline	3.39	3.74	2.43	3.49	3.17	3.81	3.87
	IWI	2.48	3.49	2.32	2.47	2.59	3.22	3.90
	IWIS	3.19	2.95	3.53	3.45	2.74	2.84	3.04
query success	Baseline	0.37	0.22	1.04	0.25	0.34	0.20	0.16
	IWI	0.68	0.23	0.03	0.58	1.08	0.24	0.19
	IWIS	0.42	0.09	0.23	0.42	0.50	0.23	0.00
number of viewed documents	Baseline	41.00	84.33	57.33	21.25	20.75	25.67	59.67
	IWI	41.67	44.00	55.67	19.67	24.33	18.25	26.75
	IWIS	52.67	48.00	37.50	35.00	18.33	33.33	15.00
number of viewed documents by query	Baseline	6.03	5.82	6.82	6.59	3.43	2.67	10.81
	IWI	3.18	3.24	3.05	2.37	4.66	2.11	7.81
	IWIS	3.36	2.64	2.35	3.37	2.77	4.00	1.20
coverage	Baseline	733.25	909.33	167.67	562.00	483.75	615.67	341.67
	IWI	379.67	860.00	620.00	213.33	229.33	366.00	509.50
	IWIS	905.00	450.67	988.50	329.33	662.33	120.33	343.00
relevant coverage	Baseline	13.00	8.33	25.00	11.25	7.75	4.67	5.00
	IWI	28.33	17.50	9.33	19.33	18.00	13.75	6.00
	IWIS	26.00	7.00	19.75	21.67	9.33	3.67	3.67
unique coverage	Baseline	712.25	906.00	167.67	561.25	483.50	615.67	341.67
	IWI	379.67	858.75	612.33	213.33	229.33	365.25	509.25
	IWIS	900.67	450.67	968.00	329.33	662.33	120.33	343.00
unique relevant coverage	Baseline	13.00	8.33	25.00	11.25	7.75	4.67	5.00
	IWI	28.33	17.50	9.33	19.33	18.00	13.75	6.00
	IWIS	23.00	7.00	19.75	21.67	9.33	3.67	3.67
number of clicks on query history	Baseline	2.25	1.33	1.00	0.25	2.00	0.67	0.67
	IWI	2.67	2.75	2.67	1.67	1.00	2.75	0.00
	IWIS	4.00	5.67	5.50	2.33	1.67	4.00	1.67
time spent on icons (seconds)	Baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IWI	7.00	14.75	2.00	1.33	5.67	14.00	0.75
	IWIS	7.00	7.00	19.50	4.00	3.00	6.67	0.33

		DR	TR	FA	DR		TR	
					Full	Non-full	Full	Non-full
number of clicks on result sorting	Baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IWI	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IWIS	5.33	0.00	8.25	4.67	0.67	0.00	0.00

Table A.1. Mean values of search performance, query submission, documents viewed, collection coverage and usage metrics for 3 access scenarios, and for full access and non-full access within the document removal (DR) and term *blacklisting scenarios* (TR).

			DR	TR	FA
Assessment of participants' access	Q1	Baseline	3.38	2.83	3.67
		IWI	3.00	2.63	4.00
		IWIS	3.33	2.50	2.75
Assessment of search task	Q2	Baseline	4.50	4.17	4.17
		IWI	3.83	4.38	4.00
		IWIS	3.83	4.50	4.75
	Q3	Baseline	4.50	3.83	4.00
		IWI	3.67	3.50	4.50
		IWIS	4.00	3.50	4.25
	Q4	Baseline	2.75	3.00	3.33
		IWI	3.00	3.38	3.67
		IWIS	3.17	2.00	3.63
Assessment of search performance	Q5	Baseline	2.75	2.83	4.17
		IWI	3.33	2.75	3.17
		IWIS	2.50	1.67	2.50
	Q6	Baseline	3.88	3.33	4.33
		IWI	3.83	3.50	3.83
		IWIS	3.67	3.83	3.13
	Q7	Baseline	2.50	2.17	4.17
		IWI	3.17	2.88	2.50
		IWIS	2.67	2.50	2.63
	Q8	Baseline	3.13	2.50	3.00
		IWI	3.33	2.63	3.50
		IWIS	2.83	2.00	2.75
Assessment of query property icons	Q9	Baseline	-	-	-
		IWI	3.67	4.00	4.17
		IWIS	3.50	3.83	3.75
	Q10	Baseline	-	-	-
		IWI	3.33	3.75	3.33
		IWIS	2.83	3.17	3.63
	Q11	Baseline	-	-	-
		IWI	2.83	3.88	3.33
		IWIS	2.83	4.00	3.50
Assessment of query sort function	Q12	Baseline	-	-	-
		IWI	-	-	-
		IWIS	4.17	3.67	3.38

Table A.2. Mean values of 3 access scenarios (see Table 5 for the questions). 1 = strongly disagree, 3 = neither, 5 = strongly agree

Appendix B: Study 2 results

		DR	TR	FA	DR		TR	
					Full	Non-full	Full	Non-full
recall	Baseline	0.04	0.00	0.06	0.01	0.03	0.00	0.00
	RA	0.06	0.03	0.01	0.04	0.05	0.02	0.01
	TA	0.04	0.04	0.06	0.01	0.04	0.02	0.02
highest possible recall	Baseline	0.16	0.13	0.24	0.06	0.12	0.06	0.08
	RA	0.21	0.12	0.24	0.15	0.15	0.04	0.09
	TA	0.21	0.11	0.15	0.06	0.18	0.06	0.06
percentage of relevant documents in the collection	Baseline	-	-	-	0.0106%	0.0111%	0.0106%	0.0115%
	RA	-	-	-	0.0094%	0.0117%	0.0094%	0.0103%
	TA	-	-	-	0.0104%	0.0102%	0.0104%	0.0111%
precision	Baseline	0.30	0.13	0.48	0.08	0.34	0.00	0.22
	RA	0.53	0.73	0.20	0.53	0.54	0.73	0.55
	TA	0.35	0.23	0.64	0.17	0.43	0.21	0.44
f-measure	Baseline	0.07	0.01	0.11	0.01	0.06	0.00	0.01
	RA	0.10	0.05	0.03	0.08	0.08	0.04	0.02
	TA	0.07	0.06	0.10	0.01	0.07	0.04	0.03
no. of queries	Baseline	17.00	18.67	11.67	11.00	8.00	12.33	8.00
	RA	12.00	21.00	15.33	7.00	7.00	14.50	8.75
	TA	18.33	11.33	16.25	10.00	10.00	8.67	5.33
average query length	Baseline	3.45	4.48	3.71	3.29	3.69	4.55	3.85
	RA	3.23	3.12	3.51	3.52	3.04	2.93	3.27
	TA	4.49	4.42	3.75	3.62	4.71	4.02	5.08
query success	Baseline	0.28	0.05	0.69	0.09	0.68	0.00	0.11
	RA	0.81	0.21	0.07	1.89	0.86	0.30	0.21
	TA	0.49	0.51	0.28	0.67	0.52	0.39	0.36
number of viewed documents	Baseline	44.50	52.33	44.00	26.75	19.00	33.33	19.67
	RA	35.33	41.00	63.00	19.33	17.33	11.00	30.25
	TA	64.33	75.67	50.00	12.00	52.67	48.00	28.33
number of viewed documents by query	Baseline	2.52	4.46	3.89	2.29	2.69	3.92	13.58
	RA	3.55	2.20	3.87	4.99	2.62	1.19	3.44
	TA	5.35	8.00	3.43	7.68	6.62	6.82	5.17
coverage	Baseline	693.00	954.67	308.00	225.25	516.50	566.00	429.00
	RA	523.33	563.50	537.67	302.33	339.33	389.00	195.00
	TA	675.33	543.33	550.75	427.00	380.67	276.67	282.00
relevant coverage	Baseline	19.00	16.00	26.67	7.00	13.75	8.00	9.00
	RA	26.33	14.25	23.33	18.00	18.67	5.00	10.50
	TA	23.33	13.67	17.75	6.33	19.67	7.00	7.33
unique coverage	Baseline	678.75	951.67	283.00	225.25	516.50	565.33	429.00
	RA	522.33	563.00	535.00	302.33	339.33	389.00	195.00
	TA	674.33	518.67	536.50	426.33	380.67	276.67	282.00
unique relevant coverage	Baseline	19.00	16.00	26.33	7.00	13.75	8.00	9.00
	RA	26.33	14.25	23.00	18.00	18.67	5.00	10.50
	TA	23.33	13.67	17.75	6.33	19.67	7.00	7.33
number of clicks on query history	Baseline	1.50	1.67	1.33	0.75	0.75	1.33	0.33
	RA	2.33	3.75	1.67	0.33	2.00	2.25	1.50
	TA	2.00	3.33	10.75	0.00	2.00	3.00	0.33
number of times results were sorted	Baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RA	0.33	3.25	1.00	0.33	0.00	1.00	2.25
	TA	0.00	0.00	0.00	0.00	0.00	0.00	0.00

		DR	TR	FA	DR		TR	
					Full	Non-full	Full	Non-full
number of clicks on viewed documents list	Baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RA	0.33	5.00	0.00	0.33	0.00	4.50	0.50
	TA	0.00	0.00	0.00	0.00	0.00	0.00	0.00
number of clicks on relevant documents list	Baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RA	1.00	0.75	0.00	0.00	1.00	0.50	0.25
	TA	0.00	0.00	0.00	0.00	0.00	0.00	0.00
number of clicks on bookmarked documents list	Baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RA	0.33	1.75	0.00	0.00	0.33	1.75	0.00
	TA	0.00	0.00	0.00	0.00	0.00	0.00	0.00
number of clicks on bookmark button	Baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RA	5.33	2.75	0.00	4.00	1.33	1.25	1.50
	TA	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table B.1. Mean values of search performance, query submission, documents viewed, collection coverage and usage metrics for 3 access scenarios, and for full access and non-full access within the document removal (DR) and term *blacklisting scenarios* (TR).

			DR	TR	FA
Assessment of participants' access	Q1	Baseline	2.63	2.67	3.17
		RA	2.50	2.63	2.83
		TA	3.17	2.67	3.13
Assessment of search task	Q2	Baseline	4.38	3.83	4.00
		RA	4.50	4.75	4.67
		TA	4.33	4.50	4.50
	Q3	Baseline	3.38	2.83	3.67
		RA	3.83	4.00	4.00
		TA	2.83	3.67	3.38
	Q4	Baseline	2.50	3.17	3.17
		RA	3.00	4.13	3.00
		TA	2.33	2.67	2.88
Assessment of search performance	Q5	Baseline	2.38	1.67	3.33
		RA	3.50	2.63	3.17
		TA	2.83	2.83	3.00
	Q6	Baseline	3.38	2.50	3.67
		RA	3.83	3.75	3.50
		TA	3.33	3.83	3.25
	Q7	Baseline	2.38	2.50	3.00
		RA	3.00	3.13	3.67
		TA	2.33	3.67	3.25
	Q8	Baseline	2.25	2.00	2.83
		RA	2.50	2.75	2.83
		TA	3.00	2.50	3.13
Assessment of result awareness interface	Q9	Baseline	-	-	-
		RA	4.83	3.50	3.17
		TA	-	-	-
	Q10	Baseline	-	-	-
		RA	4.67	3.88	4.33
		TA	-	-	-
	Q11	Baseline	-	-	-
		RA	4.83	4.63	4.33
		TA	-	-	-
	Q12	Baseline	-	-	-
		RA	4.50	4.25	3.67
		TA	-	-	-

			DR	TR	FA
Assessment of team awareness interface	Q13	Baseline	-	-	-
		RA	-	-	-
		TA	4.17	4.67	4.50
	Q14	Baseline	-	-	-
		RA	-	-	-
		TA	4.17	4.17	4.38
	Q15	Baseline	-	-	-
		RA	-	-	-
		TA	4.50	4.50	4.50
	Q16	Baseline	-	-	-
		RA	-	-	-
		TA	4.00	4.83	4.00

Table B.2. Mean values of 3 access scenarios (see Table 9 for the questions). 1 = strongly disagree, 3 = neither, 5 = strongly agree