"I can't tell you what I found:" Problems in Multi-level Collaborative Information Retrieval

Mark Handel, Emily Wang
Boeing Research & Technology
M/C 7L-70, POBox 3707
Seattle, WA, 98124 USA
+1 734 516 6907
{mark.j.handel, emily.y.wang} @boeing.com

ABSTRACT

In this position paper, we describe the unique challenges of collaborative information retrieval (MLCIR) in an environment where information access and visibility is not equal across all participants. We call this multi-level collaborative information retrieval, after the term used to describe an environment with differing classification levels. In contrast to traditional collaborative information retrieval, which focuses on technologies to facilitate dynamics in a group, MLCIR must be aware of the information flow in a group, making sure there is not information contamination and inadvertent disclosure of information, while still allowing for collaboration. We cover some of the challenges, and discuss impact on not just formal MLCIR domains, but also more traditional collaborative information seeking domains.

Categories and Subject Descriptors

H5.3. Information interfaces and presentation (e.g., HCI): Group and Organization Interfaces: Computer-supported cooperative work.

General Terms

Design

Keywords

Classified Information, multi-level security, collaborative information retrieval

1. INTRODUCTION

Previous work in collaborative information retrieval (CIR) has mainly focused on the use case where all of the collaborators have similar access to the underlying information repositories. For instance, two people searching for a suitable restaurant [2] both have similar access to the internet, search engines, and mapping tools. However, in certain circumstances, this assumption does not hold. We have been doing initial work in this domain, calling it *multi-level* CIR (MLCIR). We begin by defining the multi-level environment as well as terminology from the world of classified information since this environment has a specialized language that isn't always obvious. Next, we define the multi-level environment, describe some common, non-classified situations that have similarities with MLCIR, and look at the minimal research done to date. We then discuss some of the implications of this environment for design, and conclude with how this may apply to the larger CIR research.

1.1 Terminology

In the military, diplomatic, and intelligence worlds, there is often a need to protect information, sources, and methods from disclosure to a wider public. The description and definitions here are for the United States, although most other countries have similar laws and procedures to protect strategic information [2, 5]. Also, this is a greatly simplified account of how classification works; as with almost any government process, the caveats and subtleties fill volumes. This national security-sensitive information is protected by being *classified,* according to the degree of impact to national security if the information is revealed. Classification is generally at one of several *levels*, with different organizations having different terms. For instance, the US Department of Defense uses "TOP SECRET," "SECRET," "CONFIDENTIAL," and "UNCLASSIFIED" [4]. In addition, there can be additional layers of access control, such as *compartmentalized* that further restrict access on an individual or program basis.

Access to classified information is governed by a) the individual's highest level of access, called *clearance* (e.g. John may have a "TOP SECRET" clearance), and also by b) a need to know (e.g. John may have a need to know about specific human intelligence sources in Eastern Europe). Specifically, it is not enough that John has just the TS clearance, but that he also has a need to know the specific piece of TS-classified data. This means that disclosing top secret information to an individual cleared TS, but who does not have a need to know is still a

CIR'11, October 28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0951-6/11/10...$10.00.

classified “leak” just as much as TS information being disclosed to an individual without any clearance.

In addition to the regulations governing who has access to the data, there is a large set of federal laws and regulations dictating how classified data is stored, manipulated, and transmitted. Generally, classified data must be protected, whether in digital or hard-copy formats. Physical devices like safes and restricted access buildings are used to ensure security. Also, computers with classified information are never connected to the public internet, and may have limited ability to read or write removable media. There also exists a “black” Internet that is used solely for classified information transmittal and processing.

Classification of data happens at the lowest possible granularity, and each piece of information is classified at the lowest possible classification. In practice, this means that documents are generally classified at the paragraph level. So, one paragraph of a document may be unclassified, while another one may be at a TS level. This allows easier redaction of a document if it must be shared, and an understanding of what is sensitive.

Of course, with five basic levels of classification, and classification happening at the lowest possible level of granularity, information environments often have data that spans multiple classification levels. These situations are called a *multi-level* environment, that is, an environment where more than one security classification is currently in play. For instance, an environment where some data is unclassified, while other chunks are classified as Secret would be considered a multi-level environment. Multi-level environments are complex, because three things have to be managed: the classification of the data, the clearance of the individuals, and the individual’s need to know. Of course, this glosses over issues around *operational security:* the procedures to ensure that day-to-day tasks and movements do not disclose crucial information. Operational security is often as great of a concern as the underlying information being protected.

Further, there is generally a requirement to provide traceability of derived information back to the source material. Classification generally follows the underlying information, so a derivative document will be classified at the same level as the source document. This helps to avoid contamination and inadvertent disclosure of information. In addition, in the US Government case, there are a limited number of people within a given project or program who have classification authority to give a new document the appropriate classification level. Therefore, creating new classified documents is often a time-consuming and difficult task.

1.2 MULTI-LEVEL COLLABORATIVE IR

Our area of interest is how to best support the task of *multi-level collaborative information retrieval.* That is, multiple participants, each possibly with a different clearance level and need to know, working together to search on-line resources to answer an information need or question. Critically, there is more than one person participating, and access to information is not uniformly distributed across all of the participants.

1.2.1 Example: Theatre Intelligence Analysts

A classic example of this problem happens with intelligence analysts. A common situation is where two intelligence analysts are collaborating to understand a new threat. One analyst is a SIGINT (signal intelligence) specialist, and the other analyst is a HUMINT (human intelligence) specialist. Each of them has access to different intelligence databases, and different access to the underlying intelligence. In both cases, they are able to share some of the intelligence information, but are limited in other aspects. For instance, the SIGINT may be able to share the text of the messages that have been recorded, but for purposes of operational security, can’t reveal how the messages were captured, nor between whom the messages were exchanged. On the other side, the HUMINT has reports from an informant in the theatre about a new threat, but is restricted from discussing details about who the informant is. As they search for collaborating information, the two analysts are unsure if they have two independent pieces of corroborating information, or if the HUMINT source is one of the participants in the message exchange.

1.2.2 Example: Cyber Attack

Another example of MLCIR is the investigation of a cyber attack. Many cyber attacks target government websites and facilities, and originate from one or more foreign countries. In order to investigate these cyber attacks , intelligence and government officials often have to work with small companies as well as the foreign governments to identify the origins of attacks, stop the on-going attack, and prevent future, similar attacks. In the process of the investigation and resolution, collaboration across many different entities is required, some of whom may not necessarily even be friendly parties. Again, the protection of classified information is critical. This also illustrates one possible extreme of the problem: in communicating with a neutral (or hostile) government, there may be extremely limited ability to explain the context of the problem: “There’s a lot of internet traffic coming from your country” may be as much as can be revealed.

1.3 A UNIQUE PROBLEM?

Although the military / intelligence domain is one of the most extreme examples of the MLCIR problem, we want to suggest that some aspects of multi-level search occurs frequently in common collaborative search scenarios in both professional / business and personal settings. These scenarios relax different requirements, such as the degree that provenance needs to be maintained, the certification of the systems, or the levels of security required, but some of the core issues and concerns remain.

For instance, in the business world, one participant is often on a company intranet, possibly with access to specialized databases, while the other participant(s) are outside the intranet (e.g. customers or suppliers). In this scenario, the internal participant may find relevant documents, but the other participants don't have direct access to the results. Depending on the nature of the documents, like a trouble-shooting document that has proprietary information about another customer's installation, the internal participant may not even be able to share the documents through side channels. Other common situations where access to information and the ability to generally share the found information include proprietary information, trade secrets, export controlled data such as International Traffic in Arms Regulations (ITAR) and Export Administration Regulations (EAR), and medical records, such as those regulated under HIPPA.

One other frequent business scenario where aspects of the MLCIR problem arise is discovery in legal proceedings. In this case, one party is making a request for relevant documents, without necessarily knowing anything about the document corpus they are querying against. The owner of the document corpus is the opposing party in a legal dispute, and as such has unique incentives to be as unhelpful as possible in the collaborative search (either by "burying" the requester with all possible documents, or construing the search as narrowly as possible.) Freedom of Information Act (FOIA) requests often have similar problems: the requester must provide detailed requests, without necessarily knowing search terms or keywords.

Even in scenarios involving individuals, aspects of the MLCIR scenario can play out. An example is a search around a health issue. The individual with the health issue may not want to reveal the entire range of symptoms or problems to the other participants in the collaboration. As they find results that pertain to this "secret" part of the problem space, they may not want to reveal these results. Another example is in shopping or finance-related searches. A user may not want to reveal their personal account information, such as price range, current income, or savings.

Although these are simple examples, it raises the issue that MLCIR isn't an esoteric concern of a highly specialized set of users. Rather, there are frequently situations where people will have different levels of both knowledge and access to results, and not necessarily be able to (or want to) share this knowledge.

2. PREVIOUS WORK

A brief review of the existing literature[1] showed no prior work on MLCIR, or even mentions of the problem of differing information access among the participants in a collaborative search [6, 8]. However, there has been some work on different user roles [7, 9], which begins to hint at some of the difficulties in our use case. In these works, one user may be a teacher or a guide, and have a different level of knowledge, but still, the assumption is that all users have access to the same databases, and have little or no restriction in sharing what they find. In our preliminary work on MLCIR, we've done interviews with individuals who have been involved in ML collaborations, and discussed their approaches to the problem. Most respondents indicated that the most common approach is a manual approach, where individuals do their own search, generate an appropriate summary, and then share the results as appropriate and needed. Within the multi-level system development community, the bulk of efforts are around information assurance, actively avoiding the issue of collaboration.

Our own work in MLCIR is still very much at the preliminary stage. Boeing research groups have developed and certified cross security domain routers, which provided some important fundamentals of design. We have begun to sketch out ideas for MLCIR, and done early prototypes of possible systems supporting MLCIR, we have not begun the process of certifying these proof-of-concept systems for use with actual classified material. As will be discussed, the task of certifying systems for use with classified material is a significant undertaking.

3. DESIGN CONSIDERATIONS

Our initial work has raised some important design considerations for supporting MLCIR. We categorize these considerations in general areas: considerations that may apply generally across a range of different collaboration systems, and considerations that are specific to the highly regulated information environment. For the regulated environments, we discuss this from the classified context, but other contexts, such as proprietary, personally identifiable information (PII), and medical records are similar in nature, although not all of these considerations will apply, or some of them are relaxed to some degree. The first three considerations apply across all collaboration systems, while the last two are primarily concerns in regulated contexts.

3.1 Results May Not "exist."

Collaboration in a multi-level environment is hard. Between clearance levels, need to know, and operational security concerns, being able to share information between collaborators needs to be seen almost as the exception, not the rule. Obviously, a user at a high clearance level will see more results than a user at a low security level. Of course, the high-clearance level user won't be able to disclose the details of her findings to the lower clearance level user(s). But, more interestingly, the high-clearance user may not even be able to discuss relevant search terms, search strategies, or even that she even found some results!

In many situations, an "access denied" message can be generated to let a user know that relevant results exist, but

[1] : Due to the general distribution of this paper, the literature review was limited to unclassified, public sources.

the user's current permissions do not let them see the results. However, in the classified environment, it may be that these results have to be filtered out and not even shown to the user. This is a very confusing situation to users: in an export-controlled application used within a Boeing business unit, one of the most common calls to the help desk involved records filtered out of search results based on the user's export licenses.

3.2 Social Approaches Don't Work.

A common approach to this general kind of problem is to do a non-technical approach or a workaround / out-of-band solution. Although this definitely addresses some of the problems in MLCIR, it also, to some extent, defeats the purpose. This approach would essentially break the purpose of classifying information (protect sensitive information). Training on handling classified information emphasizes that individuals should not attempt to talk around the classified information. Any possible solution that relies on other channels to help exchange information about what is being found runs the risk of information disclosure. It's not always clear to a novice user what causes a particular classification decision. Sometimes it is the information itself, while other times, it is the methods practices used to collect the information that is classified. In addition, only designated individuals can actually make classification decisions. Together, these two factors suggest that the support for MLCIR needs to be a fundamental part of the design of a tool. Allowing for work-arounds would seriously impact the certifiability (see consideration #3) of a system.

3.3 Provenance is critical

In any CIR system, being able to track the origin, and as metadata (including, but not limited to the classification level) about the results found is useful. This provides users the ability to backtrack and search other paths for relevant information. Even in the personal healthcare scenario discussed earlier, being able to mark results as private or shareable would be a desirable feature for users. In the intelligence scenario, this is not just important functionality, but a critical baseline requirement for any system that would be deployed.

If the system has any functionality for creating new documents, there needs to be the ability to trace back to the initial document which provides the basis classification for that "chunk" of data. Note that here we are also suggesting that the granularity of the data requiring traceability may be at a sub-document level. In addition, although in general, the classification of a new document is the same as the source document, there does need to be a final determination of classification. These classification decisions are all done by humans, and are often time consuming. This means that rapid exchange of results will continue to be bottle-necked by the classification process.

3.4 System must be trustworthy and certifiable.

A challenge in developing prototype systems is that it isn't enough to just say that certain design features support MLCIR. The entire system needs to be tested and certified that it properly stores, marks, and protects classified information. Of course, prototypes can be tested with non-classified test data for usability purposes, but system designers and developers need to be sure that "small" tweaks don't impact the overall functionality and usability of the system.

Unfortunately, the certification process is not easy, and is unlikely to be pursued except by organizations expecting to sell the product or developing it for internal use (e.g. the so-called "three letter agencies," like the CIA, NSA, or NRO). This has the effect of limiting the amount of experimentation and design space exploration possible for a MLCIR system. If each one has to be carefully vetted, there are fewer resources available for experimentation. Further, for classified scenarios, the use of open-source libraries is a concern; these cannot be assumed to be "clean," and must also be verified for compliance with the regulations.

3.5 End-To-End Compliance.

In addition to the requirement that the system itself be certifiable, each of the components must also be certified. In many collaborative search systems (e.g. [1]), there is a central server that holds results and allows sharing between the different users. This is just one more part of the system that needs to track clearances, accesses, and enforce the necessary information security.

4. DISCUSSION

4.1 It's a Hard Problem

Our work in this is very much at the preliminary stage, far enough along to have identified some of the considerations discussed earlier. This is a challenging area of work. It's unclear how you can have a successful collaboration when one person in the group can't even discuss what she has found. It's likely that a successful MLCIR system will not just be tools and technologies to support the CIR problem, but also provide automated support for enforcing permissions, possibly advanced support for automatically determining permitted viewers of new documents created in the collaboration, as well as novel ways to allow lower-access participants to know when interesting (but higher-classification) results have been found, without unnecessarily disclosing information.

4.2 It's still Important

Even though this adds a great deal of complexity to the problem of CIR, we think it's an important area to explore. As discussed earlier, the classified domain is an extreme, but clear-cut case of a scenario that seems to have wider relevance. Collaborative search tools have great potential in the workplace, but many of the issues around information protection and security discussed here need to be implemented to some degree for the business market. As we've noted, these problems can even exist on personal searches: one collaborator may not want to reveal all of the relevant information. Solving these issues would provide a much more flexible and versatile tool.

4.3 Keep it Simple

At least in the classified domain, simplicity should trump a complex architecture. Being able to easily certify the overall architecture is critical to getting the system adopted and used. The fewer components and systems involved will reduce the amount of effort needed to certify the system. As well, it should make the system conceptually easier for the users to understand and use, and give them greater trust in the ability of the system to protect the information they add to the system.

5. AUTHORS & ACKNOWLEDGEMENTS

Mark Handel and **Emily Wang** are researchers in the Information, Knowledge and Intelligence Technologies group in Boeing Research and Technology. Mark's previous work has included globally distributed software development teams, supply chain and manufacturing execution, and coordination for coalition operations. Emily has worked in group decision support systems, group coordination, meta-search engine, distributed intelligent systems, and information security.

The authors wish to thank the anonymous reviewers for their comments and suggestions.

6. REFERENCES

1. Saleema Amershi and Meredith Ringel Morris. 2008. CoSearch: a system for co-located collaborative web search. In Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems (CHI '08). ACM, New York, NY, USA, 1647-1656. DOI=10.1145/1357054.1357311 http://doi.acm.org/10.1145/1357054.1357311
2. Anonymous, "Information Security Refresher Training" http://www.robins.af.mil/shared/media/document/AFD-071029-060.ppt, retrieved 13 July 2011
3. Elizabeth Churchill, Elizabeth S. Goodman, and Joseph O'Sullivan. 2008. Mapchat: conversing in place. In CHI '08 extended abstracts on Human factors in computing systems (CHI EA '08). ACM, New York, NY, USA, 3165-3170. DOI=10.1145/1358628.1358825 http://doi.acm.org/10.1145/1358628.1358825
4. Department of Defense, "National Industrial Security Program Operating Manual." DoD 5220.22-M (28 February 2006)
5. Department of Defense, "The Orange Book", http://www.dynamoo.com/orange/, retrieved 13 July 2011
6. Brynn M. Evans and Ed H. Chi. 2008. Towards a model of understanding social search. In Proceedings of the 2008 ACM conference on Computer supported cooperative work (CSCW '08). ACM, New York, NY, USA, 485-494. DOI=10.1145/1460563.1460641 http://doi.acm.org/10.1145/1460563.1460641
7. Jeremy Pickens, Gene Golovchinsky, Chirag Shah, Pernilla Qvarfordt, and Maribeth Back. 2008. Algorithmic mediation for collaborative exploratory search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '08). ACM, New York, NY, USA, 315-322. DOI=10.1145/1390334.1390389 http://doi.acm.org/10.1145/1390334.1390389
8. Meredith Ringel Morris. 2008. A survey of collaborative web search practices. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (CHI '08). ACM, New York, NY, USA, 1657-1660. DOI=10.1145/1357054.1357312 http://doi.acm.org/10.1145/1357054.1357312
9. Heather Wiltse and Jeffrey Nichols. 2009. PlayByPlay: collaborative web browsing for desktop and mobile devices. In *Proceedings of the 27th international conference on Human factors in computing systems* (CHI '09). ACM, New York, NY, USA, 1781-1790. DOI=10.1145/1518701.1518975 http://doi.acm.org/10.1145/1518701.1518975