Privacy Preserving Document Indexing Infrastructure for a Distributed Environment

Sergej Zerr
L3S Research Center
Hanover, Germany
zerr@L3S.de

Supervised by:
Prof. Dr. Wolfgang Nejdl
L3S Research Center
Hanover, Germany
nejdl@L3S.de

ABSTRACT

To carry out work assignments, small groups distributed within a larger enterprise or collaborative community often need to share documents among themselves while shielding those documents from others' eyes. In this situation, users need an indexing facility that can quickly locate relevant documents that they are allowed to access, without (1) leaking information about the remaining documents, (2) imposing a large management burden as users, groups, and documents evolve, or (3) requiring users to agree on a central completely trusted authority. In order to achieve this aim user access levels and access control have to be reflected in the index structures and/or retrieval algorithms as well as in ranking the search results.

My Ph.D. work focuses on building up an indexing infrastructure which supports confidential indexing, sharing and retrieval of unstructured information which is spread over a number of distributed access-controlled collections. In order to allow for effective and efficient indexing and retrieval in these settings, it considers aspects of confidentiality preservation within an outsourced inverted index, a DHT index structure in P2P networks as well as confidential top-k information retrieval.

1. INTRODUCTION

In the last years, popularity of systems for collaborative work and file sharing increased significantly. The need for effective information sharing and search within the growing amount of documents in this context pushed forward further development of search infrastructures for enterprise data management systems, P2P networks as well as collaborative platforms such as Wikis and Blogs. Software allows the creation of the real, as well as, of the virtual communities connecting people working on similar topics or sharing interests. Supporting privacy preserving interconnections and confidential information exchange within enterprises, working groups and communities in general remains a challenging research topic. User access levels and access controls have to be reflected in the index structures and/or retrieval algorithms, as well as the ranking of search results.

My Ph.D. thesis will be both, theoretically and pragmatically driven. On the one hand, I want to discover and treat new research issues in privacy preserving information sharing and retrieval in order to construct a suitable framework for secure collaborative information exchange within communities. On the other hand, the realization of open source software for privacy preserving information exchange will support acceptance of the research results in a large user community. My current work focuses on building up an open source indexing infrastructure which provides confidential indexing and retrieval of unstructured text spread over a number of distributed access-controlled collections. This infrastructure is suitable, for: work groups sharing files and documents, such as in company intranet, within distributed desktop search, or university environment. It combines existing database and information retrieval techniques in order to allow for secure and efficient sharing and retrieval of unstructured data. Retrieval properties of the resulting system should be comparable with existing solutions, which do not consider confidentiality of stored information and query privacy.

The main contributions of my Ph.D. thesis will be as follows: (i) a confidential infrastructure will be created, which allows the indexing, sharing and search of information stored on untrusted servers; (ii) this infrastructure will offer tunable tradeoffs between security and retrieval properties of the index; (iii) it will allow an untrusted server to perform top-k retrieval; (iv) the resulting confidential index structure will be distributable over a P2P network.

PVLDB '08, August 23-28, 2008, Auckland, New Zealand
Copyright 2008 VLDB Endowment, ACM 978-1-60558-306-8/08/08

2. SCENARIO

PCC (Production Control Company) creates adoptive solutions for production process controlling in manufacturing. These solutions include special software and hardware, which is adapted according to the needs of every specific customer. A large number of electronic documents, such as project and scientific documentation, stuff management, e-mail correspondence, presentations, collaborative documents have to be shared among the partners in the specific projects of PCC.

John is a leader of several projects within PCC. Each project corresponds to a customer manufacture. In order to always obtain up-to-date documents for his projects and share appropriate information with team members of a specific project, John requires a privacy-preserving centralized sharing and search facility. PCC is a large company and different projects within this company are often competitive against each other. Thus the information shared within one project should be made accessible only to the project members. Employees are involved in more than one project and thus need access to the documents of all groups in which they work. Due to the very sensitive nature of shared data, the advisory board of PCC decides to design a special indexing system. Such system needs to support selective sharing of access-controlled documents and deliver precise search results, while simultaneously preserving the confidentiality of the shared data. It should be not possible to reconstruct the content of a particular document from the knowledge of the index structure that can be obtained directly by compromising the server, or by analyzing server backups. The content of the documents shared within a project evolves over time, and dynamic group membership has to be directly reflected on the search results provided by the system. Working groups come and go relatively quickly, as projects start and finish, the system has to efficiently update the group documents.

As John has access to a huge number of documents, he is not interested in obtaining all the documents containing query terms, but rather a few documents which are most relevant to the query. Due to the nature of his job John has to travel frequently. Thus, he will use the system's search interface with help of a PDA and GPRS internet connection. In case GPRS is not available, he is even forced to use a modem. Since this kind of connection is very slow, the data volume transmitted over it needs to be minimized and additional data transmission costs should be avoided. In order to reduce the amount of the data transferred over the network, the server needs some means to identify the best documents that match John's query and return only the best top-k search results.

3. STATE OF THE ART

In the literature several areas within the research direction of my work are discussed. In this section I will give an overview over the most important related research areas, namely information retrieval and security as well as collaboration infrastructures and open source software.

3.1 Information Retrieval

Information retrieval provides standards and technologies for the efficient and effective search for: the information within documents, the documents themselves, as well as the metadata which describes the documents, or document search within databases. My thesis aims to enhance the area of information retrieval by developing and extending information retrieval techniques for personal and community data by addressing privacy and security issues.

Specifically, I will address the confidentiality of the inverted index, which is a standard IR data structure, allowing efficient keyword search within a large amount of the documents. An inverted index consists of a list of term–document relations, called *posting lists*. Each document in a posting list is represented by a *posting element* which includes the id of the document. Each posting element in a posting list corresponds to a document which contains a particular term (Figure 1).

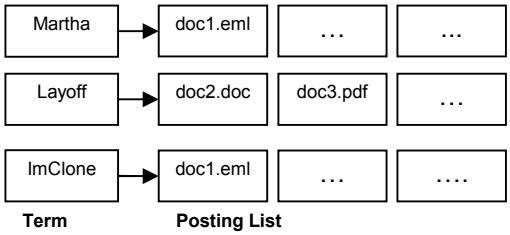

Figure 1: Inverted Index with 9 Elements

For query answering, a server selects the posting list related to the query terms and intersects them in order to find the documents, containing most of the query terms. In order to make search within such an index more efficient, powerful techniques such as sorting the posting elements by their relevance scores for the top-k retrieval, or load distribution among a set of servers were introduced [18]. However, a plain text inverted index is not secure against attacks, because an adversary needs only to scan the posting lists for a particular document id and extract a set of terms belonging to this document. Additional information e.g. term positions in the text, or term frequency can also be stored in the index in order to make the search more precise, this however makes reconstruction of the content of a particular document possible for an adversary.

P2P networks do not depend on a central indexing infrastructure for information exchange. Such network consists of the peer nodes that act as a client and a server at the same time. P2P networks are robust; since information is distributed over several machines and the potential failure of some machines do not necessary lead to a network crash or denial of service. The inverted index in P2P networks is not located on a central server, but is distributed over all peers in a special index structure called DHT - distributed hash table. Chord [20] is one of the original DHT protocols and supports efficient decentralized and scalable infrastructure. The public availability and distribution of the DHT index raises additional security problems which will be considered in my thesis.

3.2 Index Confidentiality

Index security has been addressed by many techniques designed for the outsourcing threat model, where the goal is to secure the index from tampering by the untrusted storage server.

Encryption is a standard technique for storing data confidentially [4, 10, 13]. Work has also been carried out to provide a framework for policy-based protection of XML data by encryption [3, 16] . Other techniques include suppressing and/or

generalizing released data into less specific forms, so that they no longer uniquely represent individuals [9, 12]; k-anonymity is one popular form of generalization (e.g., [2, 14, 15]). However merely encrypting the terms and/or the posting lists is not secure enough. When the document id's are encrypted with the same key in all posting list, it is possible to relate several encrypted terms to a document, by a sequential scan of the posting lists as described in the previous section. Additionally, the length of a posting list corresponds to the number of documents containing specific term, the so called *document frequency*. Document frequency is term specific and therefore can be used for statistical-based attacks and adversary can easily break the encryption scheme or draw conclusions about the corresponding term [8].

Other techniques protect the outsourced data probabilistically, adding faked information to the inverted index. E.g. μ-Serv is a system developed at IBM which is used for indexing distributed access-controlled documents [1]. μ-Serv has a centralized index that responds to a keyword search by returning a list of *sites* that have at least an *x*% probability of having documents which contain query keywords. Users then repeat their query at each suggested site in order to obtain the documents. The lack of precision in the search results from the central index represents a tradeoff between the search efficiency and confidentiality preservation. Further, μ-Serv does not support centralized ranking; the user must get ranked search results from individual sites and combine them.

While many researchers have addressed aspects of data confidentiality, none of their schemes are intended for an environment with many dynamic collaboration groups as in the scenario described in Section 2. For example, researchers have suggested ways to search encrypted text or tables stored on a remote untrusted server (e.g., [5, 6, 11, 19]). In a situation with many collaboration groups with dynamic membership, these approaches are not easy to use or manage. Document owners and/or project group managers must generate and distribute keying material for all group members, so that they can encrypt keywords and decrypt search results. If a key is lost, stolen, or even published, the index entries encrypted with it are compromised. When a key is compromised, or a member leaves a group, the key must be revoked and all the content associated with that key must be re-encrypted and re-indexed. Modern group key management schemes, such as logical key trees [7] and broadcast encryption, reduce the costs associated with giving keys to members, but still require content re-encryption. Some approaches also require that the entire index for a particular collection of documents be regenerated by the collection owner every time an entry is added to, or deleted from the index.

3.3 Collaboration Infrastructures

Collaboration, knowledge acquisition and dissemination infrastructures like Wikis, Forums and Blogs provide the foundation for joint collaborative knowledge creation and dissemination. My thesis will extend this area by developing privacy-preserving collaboration infrastructure, which does not require participants to fully trust centralized instances. The goal is to provide community members with a homogeneous view on protected as well as on public resources available to them. In short, the user can find all resources within the distributed work space, which are accessible to her and satisfy the search conditions.

3.4 Open Source Software

Open-source software makes it possible to reuse and build on top of existing sophisticated systems and create open collaborative infrastructures. This thesis will establish a crystallization point for the privacy-preserving information exchange within the open-source software development community. The tools for privacy preserving social networking and knowledge exchange will be made available as an open source software package implemented in a platform-independent manner in order to ensure its extensibility and integration.

4. CHALLENGES

Protecting the confidentiality of indexed information when the index is outsourced to a largely untrusted server raises several challenges. Unfortunately, it is not possible to directly apply the security techniques, proposed in the literature so far, to make an inverted index confidential enough, while preserving its retrieval effectiveness and efficiency. In this section, I outline the confidentiality challenges I would like to address in my work and discuss an ideal confidential indexing scheme.

4.1 General Problem

An index over a set of documents is a subset of the information contained in these documents. Since an index is typically outsourced, to a largely untrusted server, an interesting challenge arises when the stored information is confidential. On the one hand, the index should serve an authorized user with information about the dataset that is necessary for efficient document retrieval with respect to a particular user query. On the other hand, an adversary should not be able to extract sensitive information out of the index even if the server is compromised and the whole index is visible to an adversary. A tradeoff between the amount of information that can be extracted from an outsourced inverted index and retrieval efficiency of the index structure has to be investigated.

4.2 Ideal Confidential Indexing Scheme

Given a keyword query, the *ideal* indexing scheme's answer will be identical to that of a trusted centralized ordinary inverted index that incorporates an access control list check on the ranked document list identifying top-k documents just before returning it to the user. The ideal indexing scheme will answer queries and handle updates as fast as an ordinary inverted index, and with no greater network bandwidth or storage usage. Changes in group membership will be immediately reflected in the query answers of the ideal indexing scheme. The ideal indexing scheme will impose no management burden on group members, beyond the requirement that: the group coordinator maintains a list of the identities of the people in the group; the group members know how to authenticate to those identities; and the group members have trusted desktop or local web servers where they can upload their sensitive content into appropriate, access-controlled directories and have a daemon automatically ensure that the corresponding index updates are carried out quickly. If the server or servers containing the ideal scheme's index are compromised, no information about the content of the documents should be revealed. No user or super user on a non-trusted machine should be able to obtain any information about the content of sensitive indexed documents that they are not authorized to access.

The ideal indexing scheme will be unattainable in practice; however in the thesis possibilities to maximize the index efficiency under the confidentiality conditions will be considered.

5. WORK PLAN

An indexing technique resulting from this work has to preserve confidentiality of the indexed information, allow for top-k retrieval and it has to be suitable for the dynamic working group scenario. Specifically, it has to be efficient not only in terms of search, but also for updates and deletes. Index should be fresh, thus reflect the changes within the document set which it represents. In following the issues are described in more detail in the order they can be solved.

5.1 Confidentiality of an Inverted Index

The work I have done up to date has investigated how inverted index can be protected in order to deny statistical attacks, since only encryption is not sufficient. The main issue here is to hide the direct term-document correspondence while still being able to serve the user with the documents which are relevant to a particular query. Proposed was a measure of confidentiality of a merged inverted index [22]. Thereby the tradeoff between search efficiency and confidentiality is regulated by merging some of the posting lists together. Merging posting lists means combining the posting elements from the several posting lists into one bigger list. Each posting element contains a document id, an id of the term it represents and the ranking information, e.g. the normalized term frequency in an encrypted form. The term – document relation in a merged posting list becomes more uncertain for an adversary the more posting lists are merged together.

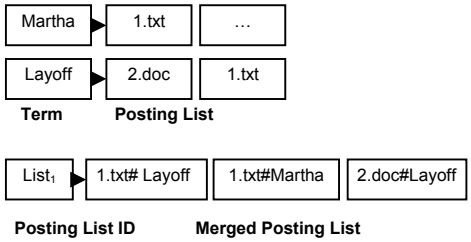

Figure 2: Unmerged and Merged Unencrypted Posting Lists

Figure 2 shows a merged posting list $List_1$, which combines the document id's corresponding to the terms "Martha" and "Layoff".

An authorized user can query the index, obtain the merged posting list, decrypt the posting elements and retract the merging. Then the querying user can sort the posting elements using ranking information contained in the posting elements and ask the corresponding information provider directly for the documents. Several merging strategies were investigated and applied in order to improve the tradeoff between the search efficiency and the confidentiality. Currently only single term queries are supported on the server side as the posting list intersection cannot be done on the server directly. In order to process multiple term queries, a separate search for each query term has to be executed first and the intersection occurs on the client side after the posting elements are decrypted. I consider multiple term queries on a confidential index as an interesting future work direction.

5.2 Top-k Retrieval from an Outsourced Confidential Inverted Index

Server-side ranking of the documents in an outsourced confidential inverted index with respect to a particular query can give an adversary a possibility for additional attacks depending on the function, which is used for ranking. For example, if confidentiality of the index is based on the posting list merging as described in the previous section and ranking algorithm is based on the term frequency, an adversary could claim that better ranked elements correspond to the more frequent terms. Distribution of the normalized term frequency used by the state of the art ranking techniques is term specific and provides additional attack possibilities. In [23] we proposed a qualitative importance measure of particular term to a particular document. This measure does not require any term specific information and uses the transformed term frequency distribution scores. Thereby a term frequency distribution pattern is analyzed on a training set of documents and a function which can destroy this pattern is developed. Multiple terms queries for this approach are also within the future work.

5.3 Confidential Index for P2P Systems

P2P networks are often used for the exchange of information within communities. My work will enable confidential sharing of the sensitive information within P2P networks in terms of securing DHT – the index structure which is used for P2P networks. The DHT is a distributed inverted index where the posting lists are copied and distributed over the network. Typically to secure some entries in the DHT, one uses the same techniques as for an ordinary inverted index. These protection techniques share the disadvantages described earlier.

As a next step I plan to investigate, how the DHT index can be better protected. Due to the fact that each peer has a specific document set and therefore specific term distribution, training the system required by the current posting list merging approach could become infeasible and other merging strategies to ensure the confidentiality might be developed. In P2P the DHT index parts are publicly available on untrusted peers. Finally, I would like to see if it is possible to provide a confidential desktop search using P2P technology.

5.4 Index Efficiency under Confidentiality Constraints

As next direction I am looking forward to improve efficiency of a confidential inverted index. Multiple term queries should be processed on the server without violation of confidentiality of the document provider as well as the searcher. Update and deletion of the posting elements should occur efficiently and immediately reflect the changes within the document collection. In the current approach, each term in a document has its own posting element, such that in case posting elements are encrypted, no term-document correspondence can be determined by the server. Thus in case a document is updated all terms in this document need to be re-indexed and resubmitted to the index. This fact makes updates and, in similar way, deletes very expensive within an encrypted inverted index. In my work I will try to reduce the necessary data administration overhead. This can be achieved by combining several ids of the documents in a collection that contains particular term together into one posting element. Thus

updates would be less frequent if several of the combined documents change in the meantime. Depending of the amount of information that is stored in such posting element, accuracy of the search results might be affected. I am looking forward to investigate the tradeoff between retrieval accuracy and update costs on such index. Further steps would be to allow provider selection, such that the index is created not on the document, but on the document collection level and all document ids corresponding to the particular term in a particular collection are combined into one posting element.

6. METHODOLOGY & EVALUATION

In my Ph.D. thesis, I will propose a privacy preserving search infrastructure for confidential information and will systematically evaluate the proposed concepts and algorithms. This section outlines the design principles of my work. It describes datasets selected for the experiments and defines the quality measurement concept for the out coming algorithms. Finally, it explains how the deliverables will be evaluated.

6.1 Dataset Selection

The dataset has to consist of a set of full text documents like they are used in an enterprise. In order to reflect the selected scenario the document collection has to contain documents on various topics. Ideally, this would be the document collections from different working groups, where some documents may belong to many groups as well. The number of the documents should be at least comparable with a number of documents shared within an enterprise.

For the experiments I have chosen two different datasets in order to have a comparison possibility. The first collection is a subset of "Open Directory Project" collection [17] which consists of 250,000 randomly selected web pages on about 100 topics. This collection is a typical web crawl which could represent a protected subset of intranet pages in an enterprise. The documents which belong to one topic can be viewed as a document collection shared within a working group. The second collection is a set of documents from the "Stud IP" [21] from a university. "Stud IP" is a learning management system which is maintained by a number of German universities and allows the sharing of the access-controlled learning materials within groups of students and teachers. In this dataset the working groups are represented by the corresponding courses and thus perfectly reflect the scenario.

6.2 Definition of Confidentiality

First the notion of confidentiality of a protected inverted index has to be defined. For example, it can be defined as a proportion between the knowledge that is gained from the index after it has been attacked and the background knowledge of the adversary. The background knowledge is general knowledge about the dataset e.g. language statistics, which could be known to the adversary in advance. This proportion can be used as a measure of the information leakage from an index. The more information that is leaked to an adversary, the less confidential is the index. The information leakage may be measured in terms of a probability increase of a term being in particular document. This proportion can be adjusted regarding the needs of the index server by means of selective posting list merging [22].

6.3 Evaluation

Evaluation is a measurement of how the proposed index structure meets the earlier defined confidentiality conditions as well as a measurement of the retrieval effectiveness and efficiency of the index. Compared with an ordinary inverted index, a confidential index structure can affect retrieval properties of the index as well as introduce resource usage overhead (e.g. storage overhead as well as increased network usage by update and query operations).

In order to measure the overhead introduced by the index protection, the outcomes of a confidential index have to be compared with the outcomes of an ordinary, unprotected system. For instance, in order to evaluate the retrieval effectiveness and efficiency of the index, an existing query log can be executed and corresponding precision/recall can be measured. In case different index building strategies exist, they can be compared with respect to the produced index confidentiality as well as retrieval properties of the index and resource usage.

As mentioned earlier, there is a tradeoff between the document confidentiality and the retrieval efficiency of the index. This tradeoff can be visualized to allow for adjustments by a system designer.

7. REFERENCES

[1] Bawa, M., Bayardo, Jr. R. J. and Agrawal, R. Privacy preserving indexing of documents on the network. In Proc. of the VLDB, 2003.

[2] Bayardo, R. and Agrawal, R. Data privacy through optimal k-anonymization. In Proc. of ICDE, 2005.

[3] Bertino, E., Castano, S. and Ferrari, E. Securing XML documents with Author-X. In IEEE Internet Computing, May/June 2001.

[4] Blaze, M. A cryptographic file system for UNIX. In Proc. of the CCS, 1993.

[5] Boneh, D., Crescenzo, G. D., Ostrovsky, R., and Persiano, G., Public-key encryption with keyword search. In Proc. of Eurocrypt 2004.

[6] Büttcher, S. and Clarke, C. L.A. A Security Model for Full-Text File System Search in Multi-User Environments. In Proc. of the FAST, 2005.

[7] Chang, Y.-C. and Mitzenmacher, M. Privacy preserving keyword searches on remote encrypted data. Cryptology ePrint Archive, Report 2004/051, Feb 2004. http://eprint.iacr.org/2004/051/

[8] Cho, T., Lee, S. and Kim, W. 2004. A group key recovery mechanism based on logical key hierarchy. J. Comput. Secur. 12, 5 (Sep. 2004), 711-736.

[9] Fung, B. C. M., Wang, K. and Yu, P. S. Top-down specialization for information and privacy preservation. In Proc. of ICDE 2005.

[10] Goodrich, M., Tamassia, R., and Schwerin, A. Implementation of an authenticated dictionary with skip lists and commutative hashing. In DISCEX II, 2001.

[11] Hacigumus, H., Iyer, B. R., Li, C. and Mehrotra, S. Executing SQL over encrypted data in the database-service provider model. In Proc. of the SIGMOD, 2002.

[12] Iyengar, V. Transforming data to satisfy privacy constraints. In Proc. of the SIGKDD, 2002.

[13] Kallahalla, M., Riedel, E., Swaminathan, R., Wang, Q. and Fu, K. Plutus: scalable secure file sharing on untrusted storage. In Proc. of the FAST, 2003.

[14] LeFevre, K., DeWitt, D. J. and Ramakrishnan, R. Mondrian multidimensional k-anonymity. In Proc. of the ICDE 2006.

[15] Machanavajjhala, A., Gehrke, J. and Kifer, D. l-diversity: Privacy beyond k-anonymity. In Proc. of the ICDE 2006.

[16] Miklau, G. and Suciu, D. Controlling Access to Published Data Using Cryptography. In Proc. of the VLDB 2003.

[17] Open Directory Project: http://www.dmoz.org/

[18] Singhal, A. Modern Information Retrieval: A Brief Overview. In IEEE, Data Eng. Bull. 24(4), 2001

[19] Song, D. X., Wagner, D., Perrig, A. Practical Techniques for Searches on Encrypted Data. In Proceedings of IEEE Security and Privacy Symposium, May 2000, 44-55.

[20] Stoica, I., Morris, R., Karger, R., Kaashoek, M. Balakrishnan, H. Chord: A scalable peer-to-peer lookup service for internet applications. In Proc. of the ACM SIGCOMM ’01.

[21] Stud IP LMS. Available at: http://www.studip.de/.

[22] Zerr, S., Demidova, E., Olmedilla, D., Nejdl, W., Winslett M., Mitra, S. Zerber: r-Confidential Indexing for Distributed Documents. In Proc. of the EDBT 2008.

[23] Zerr, S., Olmedilla, D., Nejdl, W. Zerber+R: Top-k Retrieval from an r-Confidential Index. Technical Report, L3S Research Center, March 2008.