

一种密文全文检索系统的安全索引结构

彭 霖, 李瑞轩, 宋 赛, 辜希武, 文坤梅, 卢正鼎

(华中科技大学 计算机科学与技术学院, 湖北 武汉 430074)

摘 要: 文中提出了一种改进的安全全文索引结构. 通过在倒排索引中对词条进行加密以抵抗语义分析攻击, 屏蔽词条位置和频率信息以抵抗统计攻击, 使用分块加密策略以抵抗已知明文攻击, 使用两级索引结构保证了索引安全和检索效率的平衡. 基于该安全全文索引结构设计了密文全文索引加密方案, 并给出了一种多级密钥管理策略.

关键词: 密文全文检索; 倒排索引; 安全索引结构; 分块加密

中图分类号: TP391.3

文献标识码: A

文章编号: 1000-7180(2012)09-0027-04

A Secure Index Structure in Cipher-text Retrieval System

PENG Lin, LI Rui-xuan, SONG Sai, GU Xi-wu, WEN Kun-mei, LU Zheng-ding

(School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: To cope with the potential security hazard of current inverted index, we propose an improved secure index structure. We encrypt the terms of the inverted index to defend the semantic analysis attack. We omit the position offset and frequency of terms to resist the statistics attack. We use block-based encryption to resist the known plaintext attack. We also introduce two-level index structure to balance security and efficiency of the index. Based on the proposed secure index structure, we design the encryption strategy for the secure index and give a multi-level key management policy.

Key words: cipher text retrieval; inverted index; secure index structure; block-based encryption

1 引言

全文检索技术作为信息检索技术发展史上的一次飞跃, 为信息资源的检索利用提供了一种高效的手段^[1], 然而信息安全问题也日益凸显. 基于全文检索技术的通用搜索引擎满足了广大普通用户的一般信息检索需求, 而电子党务、电子政务和电子商务领域也同样存在着特定领域的信息检索需求. 在这些有着高安全需求的领域, 文献信息资源如政府内部公文、商业机密文件等必须进行加密存储, 在此基础上进行的检索必须具有高安全的索引结构.

直接使用传统的倒排索引方法来管理涉密或敏感的文档信息, 其安全隐患主要包括: (1) 索引文件的索引词表须保序以实现高效检索, 攻击者可以利用索引词序来进行字典攻击. (2) 为了保证全文检

索功能的全面性和可扩展性, 倒排文件保留了很多原文信息, 必须采取强有力的措施保护倒排文件中存储的内容. (3) 高安全需求环境中, 针对不同级别的用户应用得到不同的检索结果, 而传统的倒排索引结构不能支持对检索行为的控制, 无法满足密文全文检索的需求.

国外在这方面的研究^[2-5]主要集中在分布式文件系统中的加密文本检索问题, 他们主要通过文本中的关键词而不是全文来建立安全索引, 从而实现快速查询. 国内有研究将全文索引中的明文索引词替换成密文索引词, 但并没有提出考虑安全控制的索引结构设计^[6-9]. 本文的目标是提出一种支持密文状态下全文检索的安全索引结构, 使用户在权限许可范围内对加密文档进行安全高效的检索利用, 从而实现高效的全文检索功能.

收稿日期: 2012-02-20; **修回日期:** 2012-03-25

基金项目: 国家自然科学基金(61173170, 60873225); 国家“八六三”计划项目(2007AA01Z403)

2 改进的安全倒排索引结构

针对传统倒排索引结构存在的安全隐患,本文将分 5 个步骤对传统倒排索引结构进行改进:

(1) 从全文索引倒排词表角度分析,其有序性是保证高效全文检索的前提,不能屏蔽.但是这种顺序容易遭受语义分析攻击.可以事先对文档进行中文分词并逐条加密分词结果,然后针对密文词条构建倒排索引,其保持的就是密文序,该密文序在语义上没有任何意义,这样就屏蔽了倒排词表明文语义上的顺序,这也是密文全文检索系统研究领域现阶段的做法.本次改进后的基于密文词条的索引结构如图 1 所示.

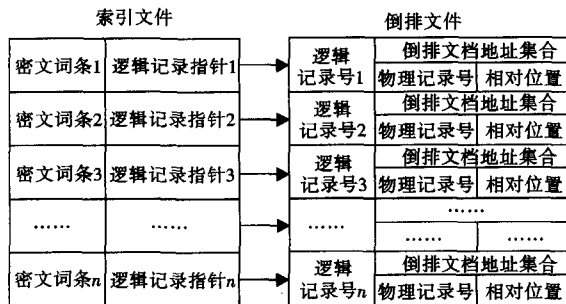


图 1 基于密文词条构建的倒排索引示意图

(2) 从全文索引静态存储角度分析,虽然上述索引结构中索引文件里存储的都是密文词条,但索引结构本身并没有加密,攻击者依然可以通过统计分析攻击获得关键信息.例如,倒排文件里的地址串越长,其对应的索引词就一定是该应用场景中的热点词.基于第一步的改进,可以进一步将索引文件中的密文词条数据与指向倒排文件的地址指针进行捆绑加密,并将索引文件与倒排文件分离存储,同时将倒排文件中的词条位置信息和频率信息在倒排文档地址指针集合中屏蔽掉,这样就可以完全防范统计攻击了.本次改进后的词条和指针捆绑加密的索引结构如图 2 所示.

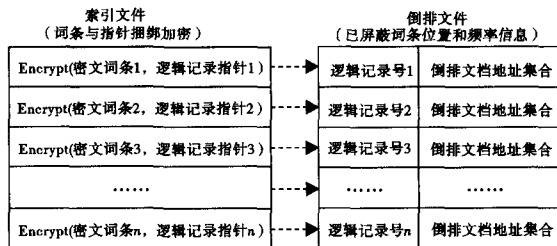


图 2 词条和指针捆绑加密的倒排索引示意图

(3) 从全文索引动态访问角度分析,如果攻击者能动态跟踪全文检索系统的访问过程,则还是有

可能找出一系列词条与该词条在索引中相关信息的对应关系,从而破解部分关键信息.基于第二步的改进,可以进一步将索引文件中的倒排词表划分区间,一个区间里包含 M 个词条,然后再对词条区间进行整体加密,这样就将单个词条的安全性转化为该词条所在区间的整体安全性,在一定程度上防范了动态访问过程中的明密文比对攻击.本次改进后的按词条区间分块加密的索引结构如图 3 所示.

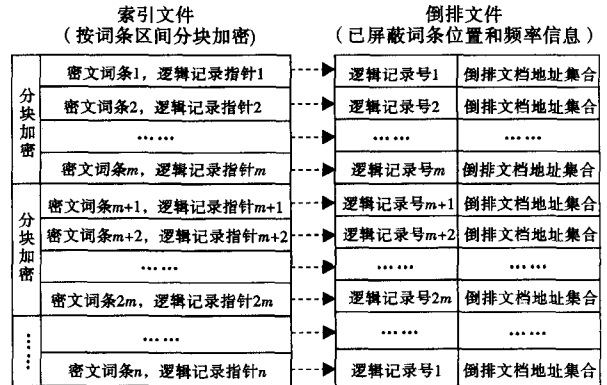


图 3 按词条区间分块加密的倒排索引示意图

(4) 经过上述改进后,为了进一步提高性能,基于第三步的改进,可以进一步引入两级索引结构.一级索引包含全部密文词条元组集合,每一个元组由密文词条和指向倒排地址文件的指针组成,然后再将这些元组按词条区间分块加密.取每一区间的首个词条再次建立索引即得二级索引,其规模是一级索引的 $1/M$,可在系统初始化时一次性导入内存以加快索引的访问速度,从而弥补索引加密带来的性能损失.二级索引的每个元组由密文词条和指向一级索引中该词条所处区间的首地址指针构成.二级索引采取整体加密的方式存储.经过最终改进后的索引结构如图 4 所示.

(5) 文档在构建密文索引时,其权限信息以密态形式被安全地注入到索引结构中;用户在全文检索时,其权限信息被注入提交的查询请求中;系统在密文索引库中检索时,对文档和用户的权限信息进行计算,保证了检索操作在用户权限范围内进行.

3 密文全文索引的加密方案

密文全文检索系统需要重点解决的是静态数据加密问题,其加密方案设计包含加密层次、加密粒度、加密算法三个方面:

(1) 加密层次

如图 5 所示,密文全文检索系统的加密层次也可以有两种:全文检索引擎内部加密和全文检索引

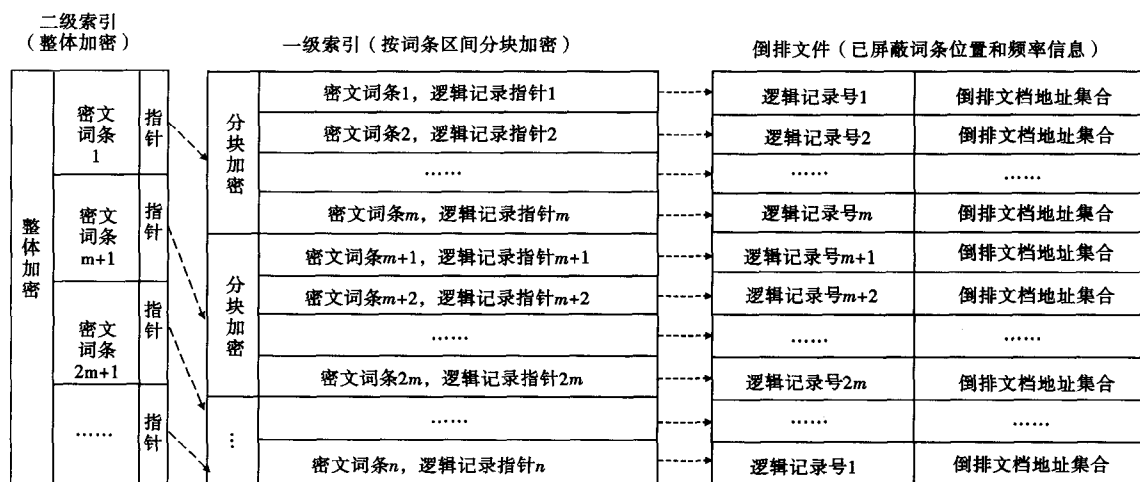


图4 两级索引结构的倒排索引示意图

擎外部加密。内部加密可以基于索引逻辑结构或者存储结构进行,加密方式比较灵活,也便于进行代码级的控制,对上层应用透明。外部加密就只能在操作系统层次,利用其文件管理功能直接对文件系统中存储的文件进行加密,加密方式不灵活,对性能的影响依赖于文件系统。Lucene 作为一个开源的全文检索引擎工具包,便于进行内核级的扩展,所以本文选择在全文检索引擎内部基于索引逻辑结构加密方式,以提供灵活高效的密文全文检索服务。

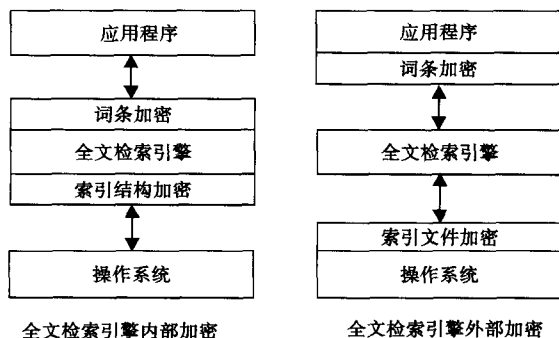


图5 密文全文检索系统加密层次示意图

(2) 加密粒度

与加密数据库不同,全文检索系统的操作对象是整个索引库,其内部逻辑结构高度耦合。从索引结构本身考虑,密文全文索引有以下不同粒度的加密需求:

① 倒排词条:为了屏蔽索引中词条的明文语义顺序,词条必须先加密再倒排。检索时,先将用户提交的查询请求进行中文分词并逐条加密,再将密文词条投入索引库检索,这样就实现了不脱密的全文检索。

② 索引文件:对于本文设计的两级索引文件,构建索引时,一级索引按词条区间分块加密存储,二

级索引整体加密存储;全文检索时,一级索引在需要时按词条区间分块解密后调入内存或缓存起来,二级索引在系统初始化时整体解密后一次性导入内存并使之常驻内存。

③ 文档:在密文文档库中,原始文档和对应的XML文档整体加密存储,在需要时整体解密调入内存。

(3) 加密算法

与加密数据库相同,密文全文检索系统需要重点解决的也是静态数据加密问题,也对加密算法的运算速度有较高的要求,所以一般采用对称加密算法。为保证加密算法的可置换性,应尽量降低密文全文检索系统与加密算法的耦合度。

4 索引的安全性分析

在密码学中,密码分析或者说密码攻击手段有以下五种情况:惟密文攻击、已知明文攻击、选择明文攻击、选择密文攻击、选择文本攻击。

本文基于这样三个假设来分析密文全文索引的安全性:系统采用的加密算法是公开的;存储在硬件加密卡中的密钥是无法被攻击者获得的;系统运行过程中内存数据是安全的。则对于密文全文检索系统而言,其可能遭受的攻击主要有两种:惟密文攻击和选择明文攻击。下面基于前面提出的索引结构进行其安全性分析:

(1) 惟密文攻击分析

假想攻击者已知系统公开的加密算法,然后又通过一定手段获取了存储在硬盘上的密文索引库。结合加密算法,针对静态存储的密文索引库可以有两种攻击:基于明文词汇表顺序的语义分析攻击和基于词汇出现频率的统计分析攻击。

(2) 选择明文攻击分析

假设攻击者已知系统公开的加密算法,然后还能入侵系统以文档管理员的身份提交明文文档,并能知晓密文索引库中所产生的相应更新,进而可能分析出一系列词条与该词条在索引中相关信息的对应关系,从而破解部分关键信息.对于这种情况,该索引结构中索引文件是按词条区间分块加密的,一个区间包含 N 个词条,从而将单个词条的安全性转化为该词条所在区间的整体安全性,攻击者虽然可以通过提交某个明文词条定位到该词条所在的区间,但并不能准确确定为到某个具体的词条,这种一对多的关系不足以对整个词条区间形成威胁.

5 结束语

本文针对当前的倒排索引结构在应用于高安全领域存在的安全隐患,提出了一种改进的安全全文索引结构.通过在倒排索引中对词条进行加密以抵抗语义分析攻击,屏蔽了词条位置和频率信息以抵抗统计攻击,使用分块加密策略以抵抗已知明文攻击,使用两级索引结构保证了索引安全和检索效率的平衡.实验结果显示,本文所提出的安全索引结构从根本上提高了索引的安全性,解决了文献信息资源的检索与保密之间的矛盾,可以很好地适应大规模文档数据环境下的全文检索.

参考文献:

- [1] Melnik S, Raghavan S, Yang B. Building a distributed full-text index for the web[J]. ACM Transactions on Information Systems, 2001, 19(3): 217-241.
- [2] Goh E J, Shacham H, Modadugu N, et al. Securing remote untrusted storage[C]// Proceedings of the 10th Network and Distributed System Security (NDSS 2003). San Diego, California, USA, 2003: 131-145.
- [3] Song D, Wagner D, Perrig A. Practical techniques for searches on encrypted data[C]// Proceedings of IEEE Symposium on Security and Privacy (S&P 2000). Berkeley, California, USA, 2000: 44-55.
- [4] Boneh D, Crescenzo G D, Ostrovsky R, et al. Public-key encryption with keyword search[C]// Proceedings of 23rd Annual Eurocrypt Conference. Interlaken, Switzerland, 2004: 506-522.
- [5] Ning Cao, Cong Wang, Ming Li, et al. Privacy-preserving multi-keyword ranked search over encrypted cloud data [C]// IEEE INFOCOM 2011. Shanghai, China, 2011: 829-837.
- [6] 赵丹枫,高峰,金顺福,等.基于错检期望值的密文索引技术[J].小型微型计算机系统,2010,31(1):113-118.
- [7] Guoqiang Gao, Ruixuan Li, Xiwu Gu, et al. Mimir: term-distributed indexing and search for secret documents[C]// The 6th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2010). Chicago, IL, USA, 2010: 1-9.
- [8] 蔡克,张敏,冯登国.基于单断言的安全的密文区间检索[J].计算机学报,2011(11):2093-2103.
- [9] 王迪,刘国华,于醒兵.基于最佳桶划分策略的密文索引技术[J].小型微型计算机系统,2008,29(4):649-652.

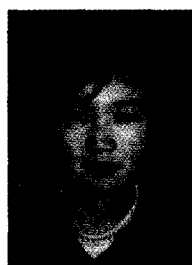
作者简介:

- 彭霖 男,(1964—),博士研究生.研究方向为信息安全.
- 李瑞轩 男,(1974—),博士,教授.研究方向为分布式计算、信息检索、系统安全.
- 宋赛 男,(1982—),硕士研究生.研究方向为信息检索、数据安全.
- 辜希武 男,(1967—),博士,讲师.研究方向为 Web 数据管理、信息检索.
- 文坤梅 女,(1979—),博士,讲师.研究方向为 Web 数据管理、信息检索.
- 卢正鼎 男,(1944—),硕士,教授.研究方向为分布式系统、信息安全.

(上接第 26 页)

- [5] 龚正江,蔡庆超,张学杰.一种基于数据网格的动态副本管理策略研究[J].云南大学学报·自然科学版,2009,31(5):470-476.
- [6] 熊润群,罗军舟,宋爱波,等.云计算环境下 QoS 偏好感知的副本选择策略[J].通信学报,2011,32(7):93-102.
- [7] 施荣荣,周晓明,郭成昊.支持 QoS 和用户需求的 Web 服务选择模糊算法[J].计算机应用研究,2011,28(5):1748-1751.

作者简介:



- 赵秋云 女,(1979—),硕士,讲师.研究方向为网格计算、服务计算