学校编码：10384　　分类号_____密级_____
学号：X2013231024　　UDC_____

厦门大学

工 程 硕 士 学 位 论 文

军队电子文档信息管理检索系统的设计与实现

Design and Implementation of Electronic Documents Management Retrieval System for Military

黎海燕

指 导 教 师：董 槐 林　教 授

专 业 名 称：软 件 工 程
论文提交日期：2015 年 10 月
论文答辩时间：2015 年 11 月
学位授予日期：2015 年 12 月

指 导 教 师：_____
答辩委员会主席：_____

2015 年 10 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

近年来，随着我军信息化建设的快速发展，军队信息网络基础建设取得了长足的进步，军队各级自动化办公条件明显改观。随着时间的累积，军队电子文档信息将不断增多，对这些电子文档信息的管理和检索已成为急需解决的问题。目前，全文管理和检索技术在对电子文档信息的管理中已得到了广泛的研究和应用。不过由于很多现有的全文管理和检索技术方法都是以明文的形式来存储文档和索引文件的，因此将对文档的管理带来很大的安全问题。

目前，很多人针对全文检索的安全性问题提出了很多解决的方案，但是这些方案要么解决了电子文档的加密问题，但却不能对非结构化的文档进行检索；要么实现了对非结构化文档的检索，但却不能保证文档本身的安全。在全文检索技术和非结构化的密文文档存储技术的实际结合运用上，目前还是空白。针对这一问题和军队各级单位在电子文档管理使用上的实际需要，本文结合军队团级单位的编制和日常业务需求，设计了有针对性的支持密文全文检索访问控制方案和管理的算法，同时基于该访问控制方案，结合电子文档数据的中文分词和加密等技术，实现了针对团级单位的电子文档信息管理检索系统。该系统以军队日常使用的电子文档信息为处理对象、并对文档进行加密处理后以密文形式进行存储管理，检索使用时按照用户的角色和职责权限进行控制。

关键词： 军队机关；访问策略；密文检索

Abstract

Recently, with the rapid development of information construction of the army, the military information network infrastructure has made considerable progress, also the working condition has been improved obviously. With the accumulation of time, electronic document information in army will continue to increase, how to manage and retrieve this electronic document information has become an urgent problem. Currently, the full management and retrieval technology in electronic document information has been widely studied and applied. However, a lot of text management and retrieval techniques store the documents and index files without encryption, thus bring a lot of security risks.

Currently, a lot of researchers has made a lot of work about the security of full-text search, but these are either to solve the encryption of electronic documents, while can't retrieve unstructured documents; or have implement of unstructured documents retrieval, but the safety of the document itself can't be guaranteed. There is rare study on the full-text search combined with the use of the cipher text and unstructured document storage. To solve this problem, This paper, combining the needs of the daily business and the regiment level units of army, designs an algorithm for cipher text retrieval access control policy and management, and based on this access control policy, combined with segmentation technology in electronic document for Chinese, we develop an electronic document retrieval and management system on group level units. The system deals with the army daily as objects, and manages and storage documents with encryption, while its retrieval for encrypted documents based the proposed access control strategy.

Key Words: Military Authorities; Access Strategy; Encrypted Full-text Retrieval

目 录

Contents

第 1 章 绪 论

1.1 研究背景与意义

近年来，随着我军对军队信息化建设的重视程度的提高，军队计算机网络建设的步伐越来越快，信息化建设在提升部队的技术水平，提高我军信息化作战能力的同时，也加剧了我军网络中的信息安全形势。近几年来，国外敌对分子利用网络技术手段加大了对我军重要部门、重要部队的攻击力度。先后有关于一些军队院校及部队在职人员失泄密的情况通报，更有甚者，我军还未正式公布的一些装备信息，在境外网站上就有了报道。当前，我军的信息化建设还处于起步阶段，还存在一些较为突出的矛盾和问题。比如军队最重要的各类电子文档信息，在管理上尽管做到了移动存储介质保存及专人保管，也建立起了很多非常严格的管理制度和措施，但同时也出现了使用电子信息不方便，保密员权利过于集中、安全隐患突出等问题。另外，军队计算机网络上的电子信息也由于广泛共享与方便使用为敌特分子和不法用户非法使用系统资源打开了便利之门。出于对军队信息的保密，军队内部网络与地方网络做到了完全物理隔离，并且也加装了保密设备安装了保密系统，但是由于采用明文存储、传输，因此在部队内部之间以及存储介质丢失的情况下安全隐患还比较突出。另外，由于没有专门针对军队电子文档信息而开发的检索系统，用户在检索所需的信息时只能通过操作系统自身的搜索功能，不仅查找历史文档信息耗费时间，而且查找出来的结果也不够全面或者根本就不是用户真正想查找的文件，给用户使用带来诸多不便，也很大地影响军队各级的办公效率。

尽管对于全文检索，目前已经有很多技术研究得比较深入了, 而且也出现了很多比较成熟的检索系统，但是把电子文档信息进行加密处理，并在密文状态下进行索引构建，同时针对这些索引进行全文检索还存在一些空白。如何对军队电子文档信息进行加密，如何在密文条件下实现全文检索，以便为用户更好地使用和管理好电子文档信息，目前还没有相关研究和针对性的产品，在军队内部更没有看到相关的报道。因此，部队目前存在的这些问题如果不能及时得到解决，不仅会使军队失

泄密事件和网络被攻击事件增多，更会严重影响到计算机网络技术在军队信息化建设中的推广和应用，使我军信息化建设停滞不前。为此，如果能结合部队实际情况，从最小一级指挥机关的团级单位现有编制体制出发，通过对明文文档进行索引构建并进行加密处理，同时对各级用户进行权限控制和安全审计，研究并设计一个可以对电子文档信息进行集中存储管理，能对这些文档进行检索使用的系统，将能很好地解决军队各级在管理使用电子文档信息时存在的问题。

军队电子文档信息管理检索系统的设计与实现意义是非常重大的：

1、带来办公方式的重大改变

系统的开发使用将使机关人员在使用电子文档信息时不再需要进行复杂而且繁琐的使用申请、首长审批、归还注销等过程，也不需要再为每一个业务部门配置存储介质，并对存储介质进行统一保管，只需要将所有电子文档信息存储到系统服务器上，并为每一名机关人员设置相应的用户名和密码，人员在操作使用过程中只需要使用自身的用户名和密码进行登录即可，系统的使用将使电子文档信息的存储管理彻底摆脱存储介质的束缚。

2、极大地保障信息存储管理的安全

由于不需要为每一个部门配置存储介质，这将大大减少存储介质丢失和损坏带来的安全隐患；另外，由于所有信息均存储在服务器上，将大大减小保密员私自使用电子文档信息以及变节带来的安全风险；而且所有电子文档信息均通过加密处理并以密文状态存储在系统服务器上，即使军队内部的不法用户或者国外敌特分子通过各种途径获取了经过处理的文档信息，但也无法掌握文档的明文信息。

3、有效提高机关的办公效率

系统的使用将使机关人员存储电子文档信息变得轻而易举，也很方便的对历史文档信息进行检索查询和操作使用，操作过程在数秒中内即可完成，大大提高机关的办公效率。另外，系统的使用也将使各级首长和部门领导能方便快捷地了解机关人员的日常办公情况，对人员办公情况进行有效地监督，促使机关人员能高效负责地开展自身工作，促进机关办公质量地提高。

1.2 国内外研究现状

1.2.1 全文检索技术

全文检索系统是向用户提供全文检索服务的软件系统，用户在检索时只需要输入检索的关键字，系统即可查找对应的文档并反馈给用户，全文检索系统事先对文档进行分词，接着对每个分词词语建立对应的索引，同时记录每个分词在该文档里出现的位置以及次数等相关信息，当用户输入关键字等进行查询时，系统则先将用户输入的关键字进行预处理，再与系统中事先已经建立的索引库进行相应地查找匹配，最终将查找的结果以用户友好的方式反馈给用户[1]。一个功能较为完善的全文系统应该具备以下三个条件：

1、用户的检索要简洁，即用户可以使用自然语言查找。

2、用户可以输入任意的字段进行检索，比如文章的题目、文种的词组等。

3、用户可以通过检索系统获取到文档全文。

全文检索不仅可以对结构化数据进行检索，而且还被大量用于对非结构化的数据内容直接进行检索。

世界上第一个可实际使用的全文检索系统是法律情报检索系统，它由美国匹兹堡大学卫生法律中心在上世纪五十年代建成。在随后的半个世纪时间里，全文检索技术发展迅速，其应用范围也已经遍布各个领域。特别是 80 年代以后，全文检索技术已经逐渐成为国外产业界检索文字类型信息的主流[2]。

目前，全文检索技术在国内外的商业市场也已经发展得较为成熟，市场也已经出现了很多具有较大影响力的大型全文检索系统，国内的比如百度（Baidu），国外的有 Google、Index、Bing、Yahoo 等等。这些检索系统事先通过网络爬虫爬取互联网数以亿计的各类网页信息，并且定期更新，同时对收集的网页信息建立全文检索索引。用户可通过浏览器输入关键字来使用这些检索系统进行快速查找[3]。

对于中文检索系统而言，由于汉子的特点以及中文表达的特殊性，国外现有的全

文检索技术无法较好的用于处理中文，因此，在研究中文全文检索技术时需要通常需要结合汉字的特点来考虑[4,5]。当前，如何有效地解决中文检索系统中对于中文的分词问题以及中文语义理解、句法理解问题，如何提高系统检索时的性能；结合全文检索与人工智能设计基于知识库推理机制的信息检索系统；结合现有的全文检索技术与加密技术实现安全的全文检索系统等等都是全文检索系统发展的主要趋势。

1.2.2 密文检索技术

密文检索是指为被加密的数据进行检索，一般处于安全考虑，不直接存储原始数据，而是将数据经过一定的加密算法加密之后再进行存储，这样提高了数据的安全性，但同时系统的性能也会因为加密以及解密过程受到影响。

1、加密数据库检索

数据库系统在存储需要被加密的数据，其性能会受到很大的影响。对于存储加密的字符型数据而言，这种性能影响更为突出。我们使用数据库的字符串精确查找运算符“=”时，当数据库存储的是加密型数据，其可以先将我们输入的查询关键字进行同样的加密，接着直接对加密数据直接查询而不用先将数据库中被加密的数据进行解密。然而，当在数据被加密的数据库上进行大量模糊匹配查询操作（如：like，>，< 等）时，数据库系统需要事先解密被加密的数据然后才能进行模糊查询。此时数据库无法在被加密数据上直接进行模糊查询操作，因为当数据被加密后，原始数据本身所具有的比如可比性（Comparability）、相似性（Similarity）、有序性（Ordering）等一些固有特性[6]会发生变化。如果直接对密文数据进行比较判断，那么查询的结果精确度将会大大降低。如果在查询匹配操作之前对被加密的数据进行解密，数据库的性能会因为数据解密过程带来的系统开销而受到很大的影响，查询效率也会变得很低下，因此这种方法无法较好的用于实际情况。

为了提高数据库在存储被加密数据方面的适应能力，国内外学者进行了大量的研究，并取得了一定的成果。目前对于被加密数据库的操作处理方案主要可以分为下面这三类：

（1）不事先解密数据而直接操作被加密数据。这类方法直接对数据库中的被加密数据进行常规的数据库操作而不用事先解密被加密数据。理论上，这种一种最为理想的方案，由于不需要事先对数据库的加密数据进行解密操作，数据库性能不会受到影响，此时就如同操作数据未被加密的数据库。对于此类方案，影响较大的有保持有序的加密技术以及秘密同态加密技术等。

（2）对被加密数据进行快速解密。对于加密数据库而言，主要性能消耗集中在对已被加密的数据的解密操作上，这类方法主要通过解决如何提高解密加密数据过程中的效率来提高系统在查询操作上的性能。智能卡加密技术、子密钥加密技术等则是这类方法的代表。

（3）通过缩小待解密的数据范围。此类方案通过缩小数据库中需要被解密的数据的范围，只对数据库中部分被加密的数据进行解密操作，减少因解密被加密数据而带来的系统开销，最终提高数据库在查询操作上的性能。索引技术以及过滤技术则是通过这类方法实现。

2、加密文本检索

国外对于加密文本的检索问题研究主要集中在分布式文件系统中。对于加密文本中的数据检索问题而言，其主要技术路线是事先通过语法分析等分析文档并提取其中的关键词，并对这些关键词建立对应的索引，再对索引也进行加密，从而达到快速检索的目的。

Dan Boneh，Giovanni Di Crescenzo 等人[7]针对加密文本的检索问题提出一种基于公钥系统的方案。考虑这样的问题：用户 B 用用户 A 发布的公钥将其发送给 A 的邮件进行加密。Email 网关服务器想验证此邮件中是否含有关键词“urgent”，从而决定其转发。而用户 A 并不希望网关服务器解密他的所有信息。他们试图设计一种机制，使得邮件服务器存储他人用 A 的公钥加密的发给 A 的邮件，同时通过此机制，A 提交给服务器某些关键词进行查询，而服务器就能确认哪些邮件包含 A 提交的关键词，并且服务器不会知道任何其它的信息。为了解决这种问题，其基于公钥加密技术提出了一种关键词搜索 PEKS 方案，同时针对该方案给出了两种构建方法。在构建 PEKS 的过程中，文章利用到了基于身份的加密 IBE[8]。IBE（Identity Based

Encryption）是一个公钥加密系统，在这个系统中，任何字符串都可作为合法的公钥。其主要应用在于 IBE 邮件系统，比如使用邮件地址或是日期作为加密公钥。然而这种方案对于大数量的加密文本检索问题则显得力不从心。

Dawn Xiao Song 等[9]人提出采取序列加密（stream cipher）的方法来对需要加密的文本数据进行处理，通过这种方法可以直接对加密文本就行搜索匹配而无需事先解密被加密的数据。他们提出这种方案的基础是在不受信任的服务器上来远程搜索被加密的数据，同时其论证了这种方案可提供安全性支持。Dawn Xiao Song 等人称其方案具有一些关键性的优点：支持隐藏的和可控制的查询；支持查询隔离（query isolation）；具有可证明的安全性；简单、快速（比如对于一份长度为 n 的数据，其加密以及加密后的查询算法时间级为 O(n)）；同时该方案几乎不需要交互开销等；另外其还可通过扩展从而支持更高级的查询操作。但是该方案也存在一些问题：首先，它需要使用特定的加密方法来加密数据，因此也就无法使用市场上一些较为成熟的加密方案。同时，该方案无法处理压缩数据，然而当前很多用户更习惯于在服务器上存放压缩后的文件以节省空间，特别是在一些按存储空间大小收取服务费用的服务器中，这也使得其用户面减少。最后，也如作者他们自己意识到的，他们的方案在针对加密数据的统计分析攻击下并不安全，因为其方法会在攻击者的攻击下泄漏出被加密的关键词在原始文档中的位置等信息。虽然其提出了一些启发式的方法来弥补这个问题，但这说明了他们的安全性证据在理论是不够完备的。

Eu-Jin Goh 等人[10]对于加密文本搜索问题提出了一个新的方案。在文章中，他们定义了一种安全索引，并基于定义的安全索引，其描述了一种抗适应性选择关键词攻击的安全模式（IND-CKA）。另外，Eu-Jin Goh 等人设计了一种能够较好地满足 IND-CKA 的安全索引结构 Z-IDX。Z-IDX 索引使用伪随机函数和 Bloom 过滤器，他们用实际实验验证了其索引能够快速有效地支持在加密数据上的检索。此外，他们声称 Z-IDX 有能力处理一些任意的上传事件，同时对压缩类型文件和对文件采用的加密算法没有限制。其具体的设计方法是在使用 Bloom 过滤器的基础上对每个文件建立关键词索引，同时使用伪随机函数作为哈希函数。当某个用户提交一个文档给服务器时，他同时提交相关的 Bloom 过滤器。其随之而来的问题就是使用 Bloom

过滤器有时会产生错误的结果，因为使用 Bloom 过滤器，系统会使得远程用户潜在地下载一些并不包含其输入的查询关键词的文件。

来自中国科学院计算机网络信息中心的李新[11]提出了一种基于 IDEA 的密文检索方案，并且申请了专利《密文全文检索技术》，专利号为 200410070113.5。此项发明基于全文检索技术，其保留其中大部分内容，只对建立的索引文件的索引词进行加密处理操作，方便于系统的实现以及实际使用。

1.2.3 访问控制技术

访问控制技术起始于 70 年代，其提出一开始主要为了解决大型主机系统中数据共享时的访问控制问题。之后随着网络应用的迅速发展 ，访问控制技术思想以及方法也得到了迅速的发展，并在信息系统的各个领域得到了广泛的应用。访问控制(Access Control)是指针对需要被访问的客体，向被授权的主体以及非授权的主体提供不同访问权限[12] 。访问控制技术通常以用户身份作为前提，然后在此基础上制定各种策略来控制以及规范用户的行为。

目前，访问控制技术主要可分为三类：第一类为强访问控制，其基于安全级，同时采用集中管理；第二类为自主管理的自主访问控制，其基于授权规则；最后一类方案为基于角色与授权规则的、集中管理的访问控制技术。

1、自主访问控制技术

自主访问控制（DAC）是目前被应用最为广泛的一类访问控制技术。该技术的理论基础为访问矩阵模型。访问矩阵理论模型的发展经历了 Lampsno[13]，Grhama 和 Dennni 等人的不断修改与完善，最终发展成为目前较为成熟的 HRU 模型[14]。在访问控制系统的安全性问题上，HRIJ 模型是第一个对其进行形式化分析的模型。随后，Ruzoz，Harrsino 和 ullmna[15]等人在一些特定的访问控制系统上对其安全性问题的复杂度进行了证明。研究人员针对安全性的判定问题也设计出了一些新的可行的访问控制模型，影响比较大的是由 Lipton、JoneS 和 snyder 等人提出的 Take 一 Grnat 模型[16]，之后的许多研究者在 Take-Grnat 模型的基础做了扩展并进行了一定