Software Engineering and Applications 软件工程与应用, 2015, 4(6), 129-135
Published Online December 2015 in Hans. http://www.hanspub.org/journal/sea
http://dx.doi.org/10.12677/sea.2015.46017

Hans汉斯

Research on the Mechanism of Encrypted Domain Information Retrieval in the Cloud

Tiankai Sun, Rong Bao, Yao Chen

College of Electrical Engineering, Xuzhou Institute of Technology, Xuzhou Jiangsu
Email: strongtiankai@sina.com

Received: Dec. 8th, 2015; accepted: Dec. 22nd, 2015; published: Dec. 28th, 2015

Abstract

Cloud storage platform, as the third party, its users' core information is stored in the encrypted form. Compared to the plaintext, the ciphertext is lack of the structural features which increase the difficulty of the ciphertext retrieval. On the basis of the existing full-text index and database index, an efficient index mechanism of ciphertext is designed by using the Luene. A series of experimental analysis shows that the ciphertext index mechanism has good security and secrecy. Method will be of great practical value.

Keywords

Cipher Domain, Full Text Search, Inverted Index, Index of Cipher Text

云端信息密文域内检索机制的研究

孙天凯，鲍　蓉，陈　尧

徐州工程学院，信电工程学院，江苏 徐州
Email: strongtiankai@sina.com

收稿日期：2015年12月8日；录用日期：2015年12月22日；发布日期：2015年12月28日

摘　要

云端作为第三方的存储平台，用户的核心信息以密文形式存储。信息的密文状态与明文相比缺少了信息

文章引用: 孙天凯, 鲍蓉, 陈尧. 云端信息密文域内检索机制的研究[J]. 软件工程与应用, 2015, 4(6): 129-135.
http://dx.doi.org/10.12677/sea.2015.46017

内部的结构化特征，增加了密文检索的难度。在现有全文索引和数据库索引的基础上，利用Luene设计了一套高效的密文全文索引机制。一系列的实验分析表明，该密文索引机制，具有较好的安全性、保密性。具有较高的实际实用价值。

关键词

密文域，全文检索，倒排，密文索引

1. 引言

随着云计算技术的快速发展，云存储已渗透到了生活的各个领域。第三方公共存储平台的快速应用，在一定程度上减少了用户的存储成本。云端提供便捷存储的同时，其存储内容的安全性、稳定性以及核心信息的隐私性成为当下关注的热点。为有效地保护个人隐私数据，一些用户选择将信息加密存储。信息以密文形式存储，增加了其存储的安全性与此同时增加了其检索的难度。当前，信息资源日益膨胀，如何从大量的数据中快速准确的查找用户所需要的信息，成为了当下日益严峻的课题[1] [2]。

Dan Boneh 等人提出了基于关键词的公钥加密方法，首先使用私钥，对存储的明文关键词通过公钥方式进行加密，然后用得到的密文信息检索，该方法只适用于小范围内数据的检索。Park 等人也提出了一种安全的索引机制，其原理为用事先生成的逆 Hash 序列作为密钥，在 Bloom Filter 中保存加密后的索引，在检索时，先利用逆 Hash 序列得到多个陷门，然后再运行布隆检测，通过得到的密文文件进行解密，以此得到用户所需要的明文文件。Swaminathan 等人提出了保护隐私的排序搜索算法。但是这种方法只使用词频，不适合多个关键字查询，也不能利用它的逆文档频率，效率不高[3] [4]。

密文全文检索技术能够在信息以密文形式存储后，检索出用户所需要的信息，提高了密文检索的效率。在现有全文索引和数据库索引的基础上，利用 Luene 设计了一套高效的密文全文索引机制。一系列的实验分析表明，该密文索引机制，具有较好的安全性、保密性。具有较高的实际实用价值。

2. Lucene

Lucene 是一个著名的、开源全文搜索引擎，拥有灵活的接口设计，其为面向对象设计的典范[5]-[7]。Lucene 的工作过程描述如下：

文章 1：David lives in Nanjing, I live in Nanjing too.

文章 2：She once lived in Nanjing.

全文分析：Lucene 为基于分词的索引搜索，首先须获得文章 1 和文章 2 的分词，分词的获取采取如下方法：

先要找出文章中的所有单词，即分词。英文用空格做分隔符，中文做特殊字符处理。文章中 in too 等词无实际意义，通常过滤。存在搜索 HE 的情况，希望把 he 搜索出来。为此，需要将所有的词转化成大写或小写，文中的标点符号无意义，可以忽略。

分词处理后：

文章 1 分词：[David] [live] [Nanjing] [I] [live] [Nanjing]

文章 2 分词：[She] [live] [Nanjing]

分词处理后，由此可建立倒排索引。关系：“文章号”——“文章中所有分词”。把关系倒过来，变成：“分词”——“拥有该分词的文章号”如表 1 所示。

仅仅知道分词出现在文章中的信息还远远不够，还需要知道分词出现的次数和所在位置。一般有 2

Table 1. Inverted index table
表 1. 倒排索引表

分词	文章号
Nanjing	1
She	2
I	1
Live	1,2
Nanjing	2
David	1

种位置：一种是字符，即字符出现在了文章的何处；第二种为该分词是文中第几个分词。将这些数据(位置、次数)加上“出现频率”和“出现位置”信息存入 Lucene 中后，索引结构信息表变为如表 2。

3. Lucene 功能模块实现倒排索引

密文搜索算法具体描述为：

(1) 索引加密。设计加密搜索算法的关键在于索引的加密，为达到搜索效果，需预先指明关键字的形式。客户端在获得关键字之后，对这些关键字进行处理。建立关键字链表，存储每个关键字包含的文件和文件标识。将关键字链表组成的索引调整为倒排索引。为了确保服务器无法从索引中获取有用的信息，倒排索引使用随机函数加密，并将每个上链表的头节点存储在字典 Search Table 上，加密后的索引存储在数组 SearchArray 的随机位置上[8] [9]，索引加密流程图，如图 1 所示。

(2) 文件搜索。因为 SearchArray 和查找表内的元素都是经过加密的，所以服务器无法从 SearchArray 和查找表内获取相关的明文信息。用户请求搜索时，客服端对关键字进行加密得到搜索令牌，其指明关键字对应的密文在何处。服务器收到用户的搜索请求时，获取用户的加密索引进行查找，查找得到对应的文件标识，最后将对应的密文传给用户。文件搜索过程如图 2 所示。

(3) 文件动态添加、删除。为了方便云存储用户能够随时随地的添加和删除云端的文件，研究开发了一种能够有效的方便用户进行添加和删除文件的解决方案。构造加密索引是搜索的核心，为了确保用户在添加和删除之后还能够进行高效的检索操作，必须时时更新的索引。添加文件时，文件的关键字可能并未出现在索引中，只需在关键字链表上进行简单的更新就可以了。删除时，必须遍历关键字链表上的结点，查找与其关键字相同的结点将其删除，删除之后还要保证其顺序，其操作相对较复杂。为了能高效的进行删除操作，将每个文件中的关键字存储在文件链表中，文件存储在文件索引中，并将其加密后存储在数组上。并且将链表中的头结点存储在字典中。在删除文件时，需要先找到文件对应的关键字在搜索数组中的位置，之后再搜索数组中进行更新，并且从删除数组中删除对应的文件链表。为了确保服务器无法从数组的使用情况来获取文件信息，所以将数组中空的单元用随机的字符来填充。为了在添加文件时快速的找到空闲的结点，所以需要记录数组中空闲的结点。文件动态添加、删除过程如图 3 所示。

根据系统的运行环境，系统的索引文件的生成是在客户端进行的，也就是由用户来执行索引文件生成，并对文档进行加密处理生成密文文档，再一起提交给服务器。索引文件构建的流程图如图 4 所示。

检索流程图如图 5 所示。

在密文全文检索模块中，第一次检索结果只是密文文档的名称集合，如图 6 所示。通过对密文文档名称进行解密，获取明文文档名称，选择需要的明文文档，发送密文文档名称下载请求，下载密文文档再进行解密获取相应的明文文档。

Table 2. The index structure information table
表 2. 索引结构信息表

分词	文章号	出现频率	出现位置
Nanjing	1	[2]	3, 6
he	2	[1]	1
I	1	[1]	4
live	1	[1]	2, 5
Nanjing	2	[1]	3
David	1	[1]	1

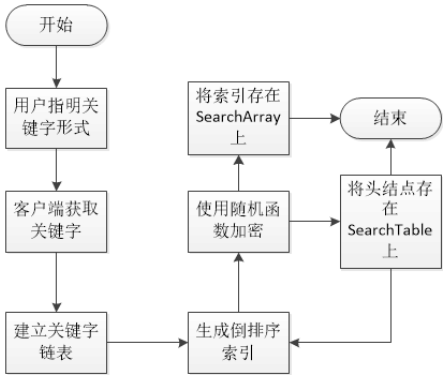

Figure 1. The index encryption process
图 1. 索引加密过程

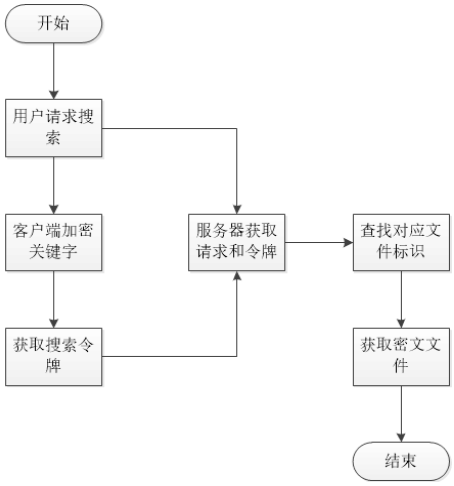

Figure 2. The document retrieval process
图 2. 文件检索过程

4. 仿真测试

以 JAVA 为编程语言，以 Eclipse 作为开发平台，辅助使用 lucene4.0。使用 desAPI 对数据加解密操作，其中文件加密采用 AES 加密算法[10] (Advanced Encryption Standard)。

4.1. 密文索引构建时间测试结果和分析

为了排除其它因素的影响，将测试数据的文件数量和文件大小定义为相同数量。通过密文和明文文

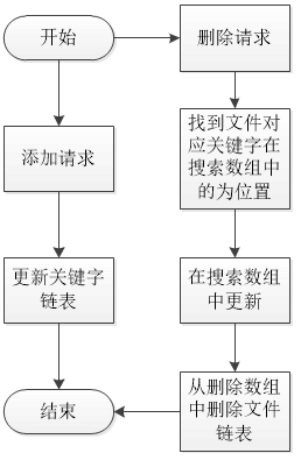

Figure 3. The files dynamically adding, deleting
图 3. 文件动态添加、删除过程图

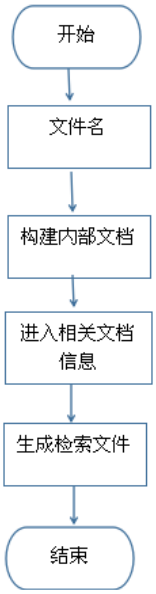

Figure 4. The flow chart of index file construction
图 4. 索引文件构建的流程图

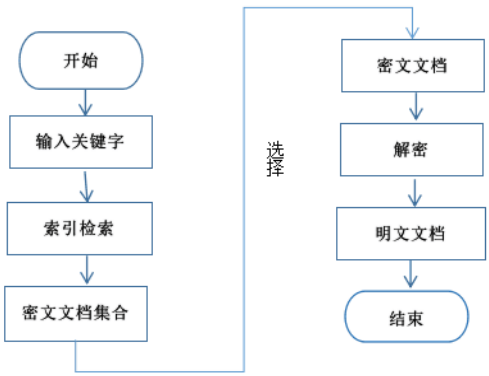

Figure 5. The search flow chart
图 5. 检索流程图

件构建时间的对比来体现测试结果，结果如图 7 所示。通过实验数据可知，该系统可以高效的构建索引文件但随着文件数量的增加效率会相对明文文件略有降低。

4.2. 密文检索性能测试结果和分析

为了排除其它因素的影响，将测试数据的检索关键词长度都设为 2。通过密文和明文文件检索时间的对比来体现测试结果，结果如图 8 所示。由实验数据可知检索时间与包含关键的文件数目成正比，且与明文检索的时间相差不大。

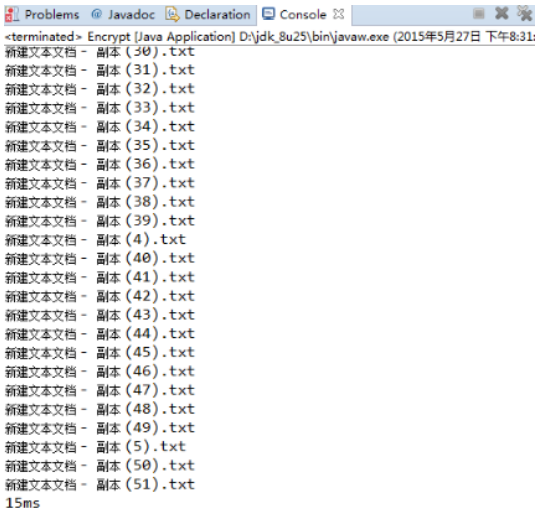

Figure 6. Retrieval results
图 6. 检索结果图

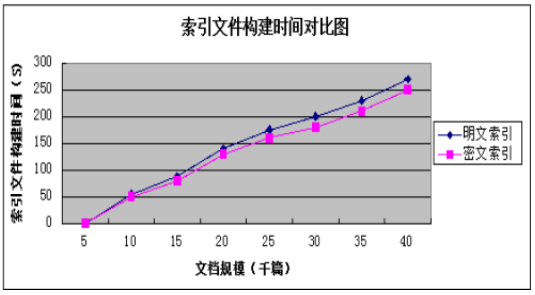

Figure 7. The comparison of index file building time
图 7. 索引文件构建时间对比

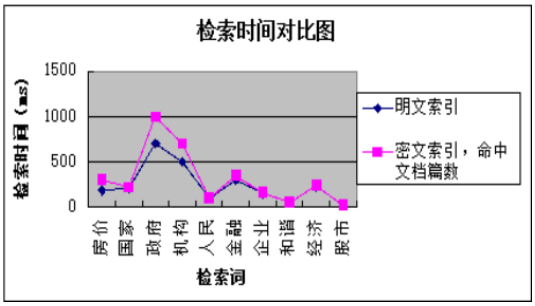

Figure 8. The comparison of retrieval time
图 8. 检索时间对比

5. 结论

数据信息以密文形式存储最大限度地保证了系统和数据信息的安全性，于此同时，快速而准确的找出用户所需要的信息也变的异常困难。通过对密文索引机制的分析，在对现有的国内外的全文检索技术进行分析比较的基础上，以全文索引和数据库密文索引机制为基础，利用 Luene 设计了一套高效安全的密文全文索引机制，包括密文索引的构建，维护以及安全索引。一系列的实验分析表明，该密文索引机制，具有较好的安全性、保密性。具有较高的实际实用价值。

基金项目

国家自然基金(61370145)；江苏省产学研联合创新项目(BY2013020)；徐州市科技计划项目(XM13B126)徐州工程学院青年基金(XKY2015312)。

参考文献 (References)

[1] 王少辉, 韩志杰, 陈丹伟, 等. 云环境下安全密文区间检索方案的新设计[J]. 通信学报, 2015, 36(2): 29-37.

[2] 郭文杰, 张应辉, 郑东. 云存储中支持词频和用户喜好的密文模糊检索[J]. 深圳大学学报理工版, 2015, 32(5): 532-537.

[3] 钱萍, 吴蒙, 刘镇. 面向云计算的同态加密隐私保护方法[J]. 小型微型计算机系统, 2015, 36(4): 840-844.

[4] 刘爱分. 云环境下高效动态的密文搜索方法[D]: [硕士学位论文]. 沈阳: 东北大学, 2013.

[5] 翟永恒. 基于 Lucene 的全文搜索引擎的应用研究[D]: [硕士学位论文]. 贵阳: 贵州大学, 2009.

[6] 蒋毅娜. Lucene 全文检索在网络教学平台中的应用研究[D]: [硕士学位论文]. 北京: 北京邮电大学, 2009.

[7] 励子闰. 基于Lucene搜索引擎的中文全文信息检索技术的研究[D]: [硕士学位论文]. 上海: 华东师范大学, 2009.

[8] 郭利刚. 密文全文检索系统的研究与实现[D]: [硕士学位论文]. 武汉: 武汉理工大学, 2011-04

[9] 冯朝胜, 秦志光, 袁丁. 云数据安全存储技术[J]. 计算机学报, 2015, 38(1): 150-163.

[10] 张国勇. 高级加密标准(AES)算法研究及实现[J]. 湖北师范学院学报(自然科学版), 2006, 26(2): 35-37.