

基于云存储的密文数据检索方法研究与实现

刘伯红,王 刚,王娟娟,史慧丽

(重庆邮电大学 计算机学院,重庆 400065)

摘 要: 为解决云存储服务海量加密数据检索问题,提出一个新的云存储密文数据检索方法,新方法采用倒排索引结构,首先用户输入检索词,在密文索引中匹配,然后将词转化成其他形式和纠正错误的词之后进行再次查询,在返回结果低于先前设定的临界值的情况下继续执行以上步骤,这种基于云存储的密文数据检索方法提高了存储数据的隐私性,能够使用户得到他们所期望的结果,实验结果证明新的检索方法是有效的。

关键词: 云存储;倒排索引;元数据;密文数据检索

中图分类号: TN402

文献标识码: A

文章编号: 1000-7180(2014)10-0056-04

Research and Implementation of Cipher-text Data Retrieval Method Based on Cloud Storage

LIU Bo-hong, WANG Gang, WANG Juan-juan, SHI Hui-li

(School of Computer Science, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: To solve the problem of mass data encryption retrieval in the development of cloud storage service, a new cipher-text data retrieval method is proposed. In this method, a new retrieval method is designed which use inverted index as the index structure. Firstly, user inputs the query words and matches on the cipher-text index precisely and transfer the word into other forms and correct the wrong word to query again. And then repeat the above steps under the circumstance if the returned result is lower than the pre-set threshold. This cipher-text data retrieval method improves the privacy of stored data to make users obtain the results what they expect. Finally, experimental results show that the scheme is effective.

Key words: cloud storage; inverted index; metadata; cipher-text data retrieval

1 引言

加密存储数据是云存储服务中的一个核心功能. 对于加密信息检索的研究,主要的算法包括线性搜索算法^[1],带关键字公钥的加密搜索算法^[2],安全指数算法^[3]等等,上述方法将产生高成本的大规模数据,另外,整理相关的文件又是另外一个问题,这两个问题有望在云存储环境下得到解决^[4-5].

加密存储就意味着数据在云计算系统中是加密的和非个人可鉴别的. 明文和密文的顺序是不同的,所以对于加密数据的检索存在一定难度. 因此,对加密数据的检索成为一个迫切需要解决的问题.

为了解决上述问题,本文提出了一种密文数据检索方法,新方法采用倒排索引结构,这种基于云存储的密文数据检索方法提高了存储数据的隐私性,可以高效率高性能检索密文数据.

2 系统框架

2.1 用户上传文件

以前,用户担心文件是否能够安全地存储在云端服务器,除了服务器和上传文件的用户,文件对于第三方是保密的. 在我们的系统模型中,假设云服务器端是不可信的,服务器和明文信息没有联系. 因此,在向服务器上传文件的过程中,必须在本地数据

收稿日期: 2013-12-08; 修回日期: 2014-01-29

基金项目: 重庆市自然科学基金(CSTC2009BB-2287)

库建立加密索引文件,而不是在服务器端建立。

用户上传文件到服务器数据库的过程见图 1。

(1)被授权的上传者申请上传文件,云服务器分配给上传者一个 id 文件(DocID)。上传者从本地数据库上传明文文件和密文文件。

(2)本地数据库提取关键字并创建一个索引文件。上传加密的索引文件和密文文件到加密的云数据库。

(3) 密文数据库发送加密文件到云存储系统。

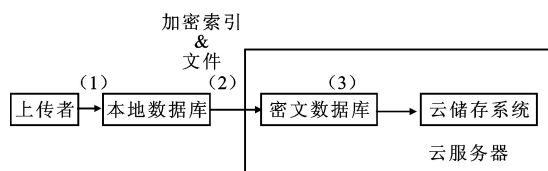


图 1 用户上传文件流程图

2.2 建立索引文件

在云计算环境下,服务器端将为授权用户存储大量的数据,所以索引文件的存储将消耗大量的资源。对于海量数据的存储,应该简化索引文件的结构,这不仅确保了高速的检索,而且减少了索引服务器的压力。

这里,我们建立了一个称作“倒排索引”的索引,它包括两个部分。

(1)索引表:索引表是一个不同索引关键字的集合,关键字被设置为主键。

(2)记录表:记录表记录的文件包含索引词和一些关于这些文件的信息。DocID 是主键。

倒排索引的文件结构如表 1、表 2 所示。

表 1 索引表

Word ID	Word	Ndocs	Offset
N	W	d	a

表 2 记录表

Doc ID	Nhits
Doc	h

为了简化存储结构,减少索引服务器的开销,索引表包含关键字 id(WordID),关键字(Word),这些文档包含一个特定的关键字(NDocs)和字段偏移量(Offset)。字段偏移量其实是一个指向记录表的指针,用来记录包含关键字的文档在记录表中的位置。记录表只包含文件 id,以及在文档(NHits)中特定单词的数量。NHits 用于文件排序,使得相关度高的文档能够首先返回给用户,为了改进文档信息记录,索引服务器可以存储其他特定地字段,以满足用户的需求。

2.3 加密方法

在用户上传文件的整个过程中一些部分需要

加密。

首先,用户上传的密文文件需要使用用户的密钥通过对称加密算法进行加密。经过加密之后,上传者需要向可信的第三方提交密钥 K 和文档编号 DocID。考虑到密钥 K 的安全性,可信第三方可以采用公钥加密算法加密 K 从而生成新的密钥 K'。为了获得 K 的解密文件,上传者需要使用私人密钥来解密密钥 K'。密钥包含在用户授权证书中。

其次,已经建立的密文索引也需要加密,通常的数据库加密方法是以字段为单位的加密方法,以记录为单位的加密方法^[4]。为了使查询更加高效,我们采用细粒度加密,字段分级加密。

在索引表中,WordID 字段不需要加密,偏移字段响应两个表的映射。这个字段是通过单向陷门处理关键字产生的。nDocs 需要频繁的变化,所以它需要通过秘密同态技术加密^[5],确保能够根据密文直接操作,不影响解密的结果。

在记录表中,(NHits)需要频繁的加密。这个字段和最后文档排序的相关度是相关的,因此它需要使用一个特殊的加密方法。Swaminathan 提出过一个加密排序搜索算法^[6]。这个算法可以分类文档相对应的通过等渗算法加密的频率的词,返回相关度高的文档。但是这个算法只适合单个关键字搜索。为了解决这个问题,Gentry 提出一个全同态加密算法加密词频率信息^[7],该算法可以对这些文档进行多词搜索。

3 查询方法设计

通过使用分词提取关键字进行检索,经过一个单向陷门函数处理允许用户做一个精确的关键字搜索。

检索过程如图 2 所示。

(1)用户向服务器提交检索词。

(2)经过元数据处理检索词生成关键字,云服务器系统使用 MD5 算法来计算关键字的哈希值。

(3)密文数据库查找匹配词,使用偏移指针查找相应的 DocID。

(4)云存储系统根据 DocID 查找相应的密文文件,发送文档和它们的信息返回给用户。

(5)如果返回结果不能够满足用户需求,通过元数据重新处理索引项进行再次检索。

(6)如果结果满足用户需求,授权用户向密钥管理中心提交密钥。

(7)授权用户根据 DocID 从密钥管理中心获取解密文档的密钥,使用密钥解密文档。

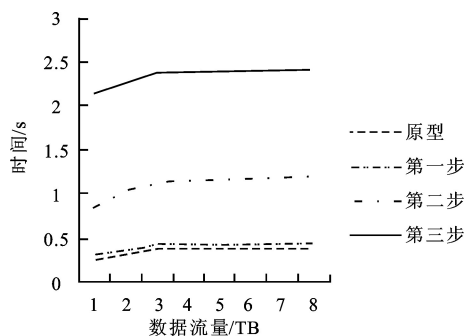


图4 每步处理后的检索时间图

超过3的单词,我们不做考虑,因为这些单词的存在将大大影响检索效率.总之,结果表明密文检索方法的效率和可行性还是令人满意的.

5 结束语

本文提出了新密文数据检索方法,能够确保云存储系统中的数据安全^[9].具体介绍了元数据处理用户输入单词的过程,能够使数据检索变得更加友好.将来的工作将主要分析元素的限制性能和密文数据检索的效率.

参考文献:

- [1] Song D, Wagner D, Perrig A. Practical techniques for searches on encrypted data[C]//The IEEE Symposium on Security and Privacy(S&P2000). Piscataway IEEE, 2000:44-45.
- [2] Boneh D, Di Crescenzo G, Ostrovsky R, et al. Public key encryption with keyword search [C] // EURO-CRYPT. Heidelberg, Springer, 2004(3027):506-522.
- [3] Park D, Kim K, Lee P. Public key encryption with

conjunctive field keyword search [C] // WISA 2004. Heidelberg, Springer, 2005(3325):73-86.

- [4] Dai Y, Shang J, Su Z. Quick index on encrypted database [J]. Journal of Tsinghua University (Sci. & Tech.) 1997(37):24-27.
- [5] Rivest R J, Adleman L, Dertouzos M I. On data banks and privacy homeomorphisms[R]. Foundations of Secure Computation, 1978.
- [6] Swaminathan A, Mao Y, Su G M, et al. Confidentiality preserving rank ordered search [C] // The 2007 ACM Workshop on Storage Security and Survivability (StorageSS2007). New York, ACM, 2007:7-12.
- [7] Gentry C. Fully homomorphic encryption using Ideal lattices[C]//The 41st Annual ACM Symposium on Theory of Computing (STOC 2009). New York, ACM, 2009:169-178.
- [8] Norvig P. How to write a spelling corrector[EB/OL]. [2014-05-22]. <http://norvig.com/spell-correct.html>.
- [9] 彭霖,李瑞轩.一种密文全文检索系统的安全索引结构[J].微电子与计算机,2012,29(1):1-7.

作者简介:

刘伯红 男,(1973-),副教授,硕士生导师.研究方向分布式系统、云计算.

王刚 男,(1989-),硕士研究生.研究方向云计算、密文检索.

王娟娟 女,(1988-),硕士研究生.研究方向云计算、密文检索.

史慧丽 女,(1988-),硕士研究生.研究方向云计算、密文检索.

(上接第55页)

- [4] Wang S S, Yan K Q, Wang S C. Achieving efficient agreement within a dual-failure cloud-computing environment[J]. Expert Systems with Applications, 2011, 38(1): 906-915.
- [5] Chiang M L. Eventually byzantine agreement on CDS-based mobile ad hoc network[J]. Ad Hoc Networks, 2012, 10(3): 388-400.
- [6] 王吉喆,赵蕴龙,吴静. WMN中拜占庭容错网络结构及算法[J]. 计算机工程, 2011(20): 28.
- [7] Klempous R, Nikodem J, Radosz L, et al. Byzantine algorithms in wireless sensors network[C]//Information and Automation, 2006. ICIA. International Con-

ference on. China: Weihai, IEEE, 2006: 319-324.

作者简介:

张钰 男,(1987-),硕士研究生.研究方向为网络信息安全.

王鑫 男,(1976-),副教授,硕士生导师.研究方向为物联网与信息安全.

蒋华 男,(1963-),教授,硕士生导师.研究方向为信息安全.

刘建明 男,(1975-),教授,硕士生导师.研究方向为无线传感器网络.