选题报告

一、选题的来源、研究的目的意义（包括在我国应用的前景）、学术和应用价值、创新点以及国内外研究现状及水平：

**1.研究背景意义**

在当今信息时代，互联网已经渗入到人们生活的各个领域。随着互联网技术不断的升级，人们工作与生活产生的数据也呈现指数形式的增长，对大量数据的存储与处理已经成为了当今社会最迫切的需求之一、为了解决这一现象，云计算技术也正不断革新。云计算，按照美国国家标准与技术研究院定义：云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问，进入可配置的计算资源共享池（资源包括[网络](https://baike.baidu.com/item/%E7%BD%91%E7%BB%9C/143243)，服务器，存储，[应用软件](https://baike.baidu.com/item/%E5%BA%94%E7%94%A8%E8%BD%AF%E4%BB%B6/216367)，服务），这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。在云计算基础上延伸和发展出来的新型网络存储技术云存储可以很好的解决大量数据的存储问题。作为云计算服务的一种，云存储是指将接入互联网的大量终端通过集群应用、网络技术或分布式文件系统集合起来，借助应用软件虚拟化后对外提供数据存储和访问功能。

随着移动终端市场的兴起，现在人们通常会拥有多种类型的智能设备。不同类型的智能设备让我们生活和工作也变得十分便捷。越来越多的人借助智能终端发送与接收各种各样的文件，而一般用户希望能够在不同终端进行浏览与修改文件，因此基于云存储技术的网盘软件正好符合需求。网盘，又名网络U盘、网络硬盘，是由互联网公司推出的在线存储服务，服务器商为用户划分一定的磁盘空间，为用户提供免费或收费的文件的存储、访问、下载等文件管理等功能。随着网盘软件的发展与便利性，越来越多人选择将各种文件存储在云端。这样可以随时随地的使用PC、便携式笔记本或各种智能终端在线查看和管理文件。这种文件管理方式减少了繁杂的文件存储和备份的操作，人们从而更专心的专注于文件本身的内容。

一般的网盘软件给用户提供了文件的存储和管理等一系列功能，这些功能能够让用户像在本地使用文件一样方便快捷。但是如果从数据的安全以及隐私方面来评价网盘软件时，我们可能需要考虑更多的因素。首先，数据存储的安全是我们尤为关心的，而数据的安全在网盘软件中具体表现为数据存储安全与数据传输安全。通常，网盘软件并没有给用户提供有效的存储保护措施，则他人可能通过系统漏洞非法获取到用户内容。而由于不重视数据存储安全而带来巨大经济与声誉损失的案例比比皆是。例如苹果公司的iCloud数据被黑客攻击，导致大量的用户数据泄露，给用户带来经济和精神上的巨大损失。在传输安全方面，很多网盘软件在数据传输时并没有引入加密措施。在当今互联网环境下，网络信道非常复杂且不可信，因此在传输过程中引入加密手段也是必要的。

其次，从隐私方面考虑时，大部分情况下用户数据存储在服务器是以明文状态存储。这种存储方式是基于服务端可信的情况下进行。然而现实中的服务器并不是完全可信，甚至可以认为完全不可信。服务器不可信的原因主要有两点:第一，在当今大数据技术日益成熟的状况下，服务端可能会检索并搜集用户存储的数据来用作数据分析。第二，如果在网盘中存储了大量文件后，用户有时需要通过特定关键词进行文件搜索，而这些搜索记录也可能会被服务端收集用作分析。所以在这种不可信的服务端情况下，我们也需要引进新的方法来确保用户的隐私不被泄露。之前2018年Facebook就被爆出泄露用户隐私数据给其他公司分析使用，Facebook公司也因为这件事被推上舆论高峰，遭受了巨大的经济损失。

综上所述，我们讨论了网盘软件的定义、发展历史和一些潜在的问题。这些潜在的问题主要分为两个方面，即安全和隐私。在安全方面，我们主要列举了两个方面的安全问题，一是文件存储在服务端时的安全，二是文件在网络信道传输时候的安全。后续在我们设计的网盘软件中可以通过引入加密技术来保证用户存储数据的安全性。对于隐私方面，我们引入加密搜索技术来确保用户在使用过程中不会暴露隐私行为给服务端，

# 2国内外现状

**（1）加密搜索**

加密搜索，又称可搜索加密（searchable encryption, SE），可搜索加密主要解决在服务端不完全可信的情况下对加密数据的关键字安全搜索。2000年，Song等人首次提出可搜索加密的概念。在借助这一技术下，用户借用服务器强大的计算资源进行关键字查询同时不会向服务器泄露任何隐私信息。这种模式下，不仅仅保护了用户数据的隐私，也防止他人窃取用户的信息，最后还能极大的提高对加密数据的查询效率。目前，可搜索加密在云存储技术上也有很大进展。

**（2）网盘现状**

当今社会也不乏一些主流的网盘服务，例如百度公司的百度云、115网盘和国外Dropbox公司的Dropbox。这些网盘服务确实给用户提供了很好的服务体验，但是如果从数据的安全性和用户的隐私性来看，这些网盘服务可能存在一些安全漏洞和隐私问题。

就当今社会主流软件网盘对比，按照对用户安全和隐私的保护，不同公司推出的网盘产品可以主要划分为三大类。

第一类网盘是对用户的数据没有任何加密和保护，即用户数据明文上传和明文保存。这类网盘产品代表有115网盘、百度网盘等。在这种情况下，首先，用户的任意数据内容可以被服务商获知，用户的隐私得不到保证。其次，明文上传和明文保存在网络信道不可信的情况下会被他人截取数据内容，用户数据的安全也受到一定的的威胁。因此这种情况下既不能保护用户的隐私也不能保证用户数据的安全。

第二类网盘是服务商提供商采用对用户保存的文件进行加密和解密的安全措施。这类网盘的代表是Dropbox公司推出的Dropbox网盘。对文件进行加密首先解决了用户的数据存储安全问题，只要服务商加密解密的密钥不丢失，那么服务商可以防止他人盗取数据内容。由于这种加密是由服务商提供，密钥也是由运营商生成和保管，导致服务商也具有检索与收集用户数据的能力，所以用户的隐私问题还是没有很好地解决。此外，这种加密手段是服务商在云端进行，在网络信道不可信情况下，用户上传文件时仍有可能被截取数据内容。

第三类网盘是由客户端在本地加密，加密完成后上传到服务端，密钥由用户进行管理。这类网盘代表是一个叫“隐形云”的产品，是由是杭州奕锐电子有限公司和阿里云合作推出的一个云盘。首先，对数据的加密可以防止他人非法获取数据内容，从而保证了用户数据的安全。其次这种产品可以保护用户的隐私，由于密钥由用户进行保管而服务商无法查看用户的数据内容。但是“隐形云”缺点在于两点。第一点是使用时必须要先下载客户端才能使用这个网盘。第二点是用户在网盘使用中对自身数据进行检索时，网盘并没有对用户搜索信息进行加密，而这样有可能造成用户的部分信息泄露，对数据存储的安全也造成了很大的挑战。因此一个既能安全存储用户数据又能支持对搜索记录加密的网盘是符合用户需求的。

**（3）新型网盘系统模型**

基于对主流网盘产品的分析与总结，我们提出了一个新的网盘系统模型，相比较之前主流的网盘系统，新的系统具有以下的优点。

一、数据上云的全生命周期均处于加密状态，且加密、解密过程不受运营商干预。文件上传之前会先进行加密，文件下载后在本地进行解密，文件加解密操作在用户侧完成，文件加密解密的密钥由用户自身保管。确保了用户文件存储的安全性和隐私性。确保了用户数据在上传与存储的安全。

二、新的网盘系统设计了基于UDP的可靠传输协议，使用UDP文件传输方式，将大文件分片传输，后续对文件片进行校验更是确保文件传输过程的安全与完整性，更能节省处理器性能。

三、保护用户隐私，我们引入可搜索加密技术。实现对用户加密数据的检索，从而保护用户隐私。可搜索加密（searchable encryption, SE），可搜索加密允许用户在密文上进行关键字查找，在服务商不可信的情况下，使用这种方式来处理用户的检索信息可以保证用户搜索时的隐私。

四、针对用户的特定类型文件我们提供版本控制的功能。对一些重要的文件可以追踪其变化，如果重要的文件不小心被删除或者修改就可以回溯到之前的版本。

**（3）创新点分析**

1、将加密搜索概念引入到网盘系统中，并实现该系统。

2、对加密搜索过程中搜索算法进行改进，优化加密搜索的性能。

3、基于UDP可靠传输协议设计。

二、论文研究的主要内容，方案和拟采用的研究方法、手段。已进行的科研工作基础和已具备的科学研究条件（包括文献资料及主要实验仪器设备准备情况等），对其它单位的协作要求。论文总工作量（估计），论文初稿的进度以及预期结果：

**1、论文研究的主要内容**

在本论文中，作者阐述了一种新型的基于加密搜索技术的安全网盘系统的设计与实现。论文工作主要从两方面展开

**1.1可搜索加密文件管理系统设计及实现**

基于对现有网盘系统的分析，我们提出新的网盘系统架构，网盘系统的设计如下

**1.1.1威胁模型**

通过对传统的网盘系统架构进行分析，我们可以看到传统无加密网盘的安全威胁，为了更能清楚的阐述情形，这里采用C/S架构进行描述（B/S情况和C/S架构下是完全相同的）。首先，我们假设服务器与网络环境是不可信的，接下来我们针对网盘客户端的一系列操作进行安全分析：（1）文件在上传与下载时可能他人会窃取数据信息，因此数据在网络环境中的传输是需要进行加密处理。（2）文件存储在服务端应该以密文进行存储，基于服务端不可信的前提下，文件加密与解密的密钥应该由客户端进行保存。（3）用户可能对自身文件进行查询，这种查询一般是基于关键字查询的搜索，在服务端与通信链路都不可靠的环境下为了保护用户的隐私与安全，我们需要引入加密搜索技术。

**1.1.2网盘系统设计**

**1.1.2.1服务端设计**

服务端主要功能有存储功能，业务处理功能、传输、加密解密和加密搜索功能。因此按照不同的功能，我们以三个主要模块化对应不同的功能分开实现。三个模块分别为存储模块，传输模块和加密模块。其中存储模块主要负责文件的存取、文件版本控制和访问控制。传输模块主要负责以UDP形式可靠传输文件，加密模块主要负责对信息的加密与解密、密钥的生成和加密搜索的支持。

**1.1.2.2客户端设计**

客户端按照功能设计，也可以将不同的功能按照模块化设计分为三大模块。分别是前台模块，加密模块和传输模块。前台模块主要是给用户提供界面展示以及各种文件管理的操作。加密模块主要负责文件加密解密，信息的加密与解密和密钥的生成。传输模块主要负责文件以UDP方式可靠传输。

**1.1.3 加密算法**

加密算法通常分为两大类，即对称加密和非对称加密。

**1.1.3.1对称加密**

对称加密算法是应用较早的加密算法，技术成熟。在对称加密算法中，数据发信方将明文（原始数据）和加密密钥一起经过特殊加密算法处理后，使其变成复杂的加密密文发送出去。收信方收到密文后，若想解读原文，则需要使用加密用过的密钥及相同算法的逆算法对密文进行解密，才能使其恢复成可读明文。在对称加密算法中，使用的密钥只有一个，发收信双方都使用这个密钥对数据进行加密和解密，这就要求解密方事先必须知道加密密钥。对称加密算法的特点是算法公开、计算量小、加密速度快、加密效率高。不足之处是，交易双方都使用同样钥匙，安全性得不到保证。此外，每对用户每次使用对称加密算法时，都需要使用其他人不知道的惟一钥匙，这会使得发收信双方所拥有的钥匙数量成几何级数增长，密钥管理成为用户的负担。对称加密算法在分布式网络系统上使用较为困难，主要是因为密钥管理困难，使用成本较高。在计算机专网系统中广泛使用的对称加密算法有DES和AES。DES全称为Data Encryption Standard，即数据加密标准，是一种使用[密钥加密](https://baike.baidu.com/item/%E5%AF%86%E9%92%A5%E5%8A%A0%E5%AF%86/5928903)的块算法，1977年被[美国联邦政府](https://baike.baidu.com/item/%E7%BE%8E%E5%9B%BD%E8%81%94%E9%82%A6%E6%94%BF%E5%BA%9C/8370227)的国家标准局确定为[联邦资料处理标准](https://baike.baidu.com/item/%E8%81%94%E9%82%A6%E8%B5%84%E6%96%99%E5%A4%84%E7%90%86%E6%A0%87%E5%87%86/3940777)（FIPS），并授权在非密级政府通信中使用，随后该算法在国际上广泛流传开来。传统的DES加密只有56位密钥，在1997年RSA公司发起对DES加密挑战时，DES很快被破解，因此DES在计算机运算速度提升后的今天被认为是不安全的。高级加密标准（英语：Advanced Encryption Standard，缩写：AES），在[密码学](https://baike.baidu.com/item/%E5%AF%86%E7%A0%81%E5%AD%A6/480001" \t "_blank)中又称Rijndael加密法，是[美国联邦政府](https://baike.baidu.com/item/%E7%BE%8E%E5%9B%BD%E8%81%94%E9%82%A6%E6%94%BF%E5%BA%9C/8370227" \t "_blank)采用的一种区块加密标准。这个标准用来替代原先的[DES](https://baike.baidu.com/item/DES)，已经被多方分析且广为全世界所使用。经过五年的甄选流程，高级加密标准由[美国国家标准与技术研究院](https://baike.baidu.com/item/%E7%BE%8E%E5%9B%BD%E5%9B%BD%E5%AE%B6%E6%A0%87%E5%87%86%E4%B8%8E%E6%8A%80%E6%9C%AF%E7%A0%94%E7%A9%B6%E9%99%A2/3931459)（NIST）于2001年11月26日发布于FIPS PUB 197，并在2002年5月26日成为有效的标准。2006年，高级加密标准已然成为[对称密钥加密](https://baike.baidu.com/item/%E5%AF%B9%E7%A7%B0%E5%AF%86%E9%92%A5%E5%8A%A0%E5%AF%86)中最流行的算法之一。

相比较于DES，AES加密算法会更加安全，因此我们在本论文的工作中也是采用了AES加密技术，确保加密的可靠性。

**1.1.3.2 非对称加密**

不对称加密算法使用两把完全不同但又是完全匹配的一对钥匙—公钥和私钥。在使用不对称加密算法加密文件时，只有使用匹配的一对公钥和私钥，才能完成对明文的加密和解密过程。加密明文时采用公钥加密，解密密文时使用私钥才能完成，而且发信方（加密者）知道收信方的公钥，只有收信方（解密者）才是唯一知道自己私钥的人。不对称加密算法的基本原理是，如果发信方想发送只有收信方才能解读的加密信息，发信方必须首先知道收信方的公钥，然后利用收信方的公钥来加密原文；收信方收到加密密文后，使用自己的私钥才能解密密文。显然，采用不对称加密算法，收发信双方在通信之前，收信方必须将自己早已随机生成的公钥送给发信方，而自己保留私钥。由于不对称算法拥有两个密钥，因而特别适用于分布式系统中的数据加密。广泛应用的不对称加密算法有RSA算法和美国国家标准局提出的DSA。以不对称加密算法为基础的加密技术应用非常广泛。

DSA（Digital Signature Algorithm）是Schnorr和ElGamal签名算法的变种，被[美国](https://www.baidu.com/s?wd=%E7%BE%8E%E5%9B%BD&tn=24004469_oem_dg&rsv_dl=gh_pl_sl_csd" \t "_blank)NIST作为DSS(DigitalSignature Standard)。 DSA是基于整数有限域离散对数难题的。DSA被常用作数字签名，并且DSA不具有加密解密功能。相比较之下，RSA具有加密和解密功能，同时也能用作数字签名。RSA是一种基于大数做因数分解的加密算法，RSA是1977年由[罗纳德·李维斯特](https://baike.baidu.com/item/%E7%BD%97%E7%BA%B3%E5%BE%B7%C2%B7%E6%9D%8E%E7%BB%B4%E6%96%AF%E7%89%B9/700199" \t "_blank)（Ron Rivest）、[阿迪·萨莫尔](https://baike.baidu.com/item/%E9%98%BF%E8%BF%AA%C2%B7%E8%90%A8%E8%8E%AB%E5%B0%94" \t "_blank)（Adi Shamir）和[伦纳德·阿德曼](https://baike.baidu.com/item/%E4%BC%A6%E7%BA%B3%E5%BE%B7%C2%B7%E9%98%BF%E5%BE%B7%E6%9B%BC/12575612" \t "_blank)（Leonard Adleman）一起提出的。RSA就是他们三人姓氏开头字母拼在一起组成的。

在论文中，我们采用RSA加密算法，这样既可以对信息进行加密也能用作签名认证。

**1.2加密搜索搜索算法优化**

**1.2.1可搜索加密实现**

可搜索加密的介绍上文已经阐述，接下内容我们聚焦于加密搜索的具体实现方式。

SE（加密搜索）的实现方法较多，但是基础的SE的机制主要包括4种算法，分别是密钥产生（Gen）、搜索凭证的生成（GenToken）、索引建立（BuildIndex）和搜索执行（Search）。

（1）Gen：该算法主要由数据所有者或者权威的公信机构运行，主要用来产生密钥，算法会根据输入的安全系数生成相应的密钥。

（2）GenToken：该算法以用户的输入的搜索关键字为输入，产应相应的搜索凭证，主要用于后续搜索执行。

（3）BuildIndex：该算法是由数据拥有者执行，数据拥有者基于数据选出关键字，并在可搜索加密机制上建立索引表。在基于公钥加密的SE机制中，数据拥有者会使用公钥对关键字进行加密，在对称加密的SE机制中，数据拥有者使用对称密钥或者使用哈希算法对关键字集体进行加密。

（4）Search：该算法是由服务器端进行，服务器将得到的搜索凭证和索引表作为输入，服务器依靠自身进行计算最后输出结果判断该文件是否满足搜索请求。

本论文中的加密搜索模型也是基于以上四个基本算法展开，在此基础上我们补充一些新的算法用作改进我们加密搜索模型的性能。

**1.2.2 可搜索加密模型**

可搜索加密按照构造算法的不同可以分为两大类，第一种是基于对称加密的可搜索加密模型，第二种是基于公钥加密的的搜索加密模型。两者的区别在于：一、基于对称加密的可搜索加密模型主要使用一些伪随机函数生成器、哈希算法和对称加密算法等构成，而基于公钥加密的可搜索加密模型主要通过一些复杂问题，将安全性建立在复杂性问题的求解难度上。因此相比较于对称可搜索加密模型，公钥加密搜索模型计算开销更大。二、基于对称加密的可搜索加密模型适合单用户创建文件多用户分享的情况，而公钥加密可搜索模型允许文件拥有者之外的用户使用可搜索加密技术生成数据密文并产生新的加密索引表。结合上述分析，新的网盘系统中选用了基于对称加密的可搜索模型。

**1.2.3可搜索加密机制分析**

可搜索加密一个关键用处就是给用户提供安全，隐私的搜索功能。实验中，我们采用了支持单词字的搜索模型。

单词字搜索模型是指对客户端将用户输入的特定关键字转化为搜索凭证并将搜索凭证发送到服务器，服务器用自身的计算能力通过该搜索凭证搜索包含该关键字的文件并将符合条件的结果返回给客户端。之前内容讲到，可搜索加密模型分为四个步骤，即密钥生成、搜索凭证生成、建立索引和搜索执行。加密搜索很好的保护了数据使用者和拥有者的隐私，这是加密搜索的优势所在。因此在搜素执行这一步骤的性能是作为评价加密搜索模型整体性能的重要指标。

搜索执行的过程就是服务器通过将搜索凭证与索引表进行比较来判断当前文件是否符合搜索条件。一些文章在应用加密搜索相关技术在搜索执行这一步用的方法是将搜索凭证与索引表中的关键字进行一一比对。如果假设关键字总数为n，搜索算法的效率为，即搜索算法的时间复杂度与关键字总量呈线性关系。也就是说每次查询都需要浏览一遍所有的关键字，这种方法在关键字数量较少时候对系统性能不会有影响，而当关键字数量达到一个比较大的数量级后这种方法会严重的影响系统的性能。因此对搜索执行过程的优化是非常有必要的。

**1.2.3可搜索加密算法的优化**

考虑到关键字的数量会随着大量文件的添加而急速增加，单纯的线性搜索时间不能够适用于大量关键字的情况。首先，单纯的的关键字搜索类似于对整个关键字集合进行逐一遍历，这种方法是不会有遗漏的，但是也是低效率的。从关键字角度出发，关键字并不是孤立存在的，可能好几个关键字都是相近的或者是指向同一文件的。因此首先我们可以将所有关键字这个大集合按照某种方法划分成小的集合，当我们进行搜索时可以先判断关键字是否属于某一个小集合，然后在小集合里进行精确查找。按照这种方法，我们可以将新的搜索划分为三个步骤：（1）先将关键字按照不同属性划分成不同的小集合（2）每个小集合生成一个特征值（3）查询时，先将由关键字生成的token与小集合的特征值进行比对，找到匹配的特征值后再进入相应的小集合进行精确查找。

在上述搜索步骤中，我们用一个特征值来代替一个小集合所有的关键字进行搜索比对，所以特征值的选取需要能代表整个集合的特征，选取过程中我们引入密码学中累加器概念。一个密码学上的累加器是一个[单向的](https://zh.wikipedia.org/wiki/%E5%96%AE%E5%90%91%E5%87%BD%E6%95%B8)[隶属函数](https://zh.wikipedia.org/wiki/%E9%9A%B8%E5%B1%AC%E5%87%BD%E6%95%B8)。它可用于识别一个候选是否为一个[集合](https://zh.wikipedia.org/wiki/%E9%9B%86%E5%90%88)的成员，且不会在过程中暴露[集合](https://zh.wikipedia.org/wiki/%E9%9B%86%E5%90%88)中的成员。在此基础借助密码上累加器这一特点，我们可以将一组集合中的值累加到一个值，并且对于任意搜索的一个token值判断是否已经在这个集合内。

**1.3版本控制**