可搜索对称加密技术在气象数据安全领域的研究与进展

郭聪　何恒宏　田征　钟磊
（国家气象信息中心，北京 100081）

摘要：与传统存储模式相比，云存储技术运维成本低廉，服务模式更为弹性可靠，已成为目前应用最广泛的信息技术之一。随着数据采集方式的多样化,数据总量呈指数级增长，数据的隐私泄露问题成为云存储技术发展的重要制约。把数据加密处理后再存储在云服务方，可以有效解决隐私泄露问题，但却给数据的检索和查询带来额外的通信、存储和计算成本。现有的解决方案无法解决用户对多源气象数据的加密搜索需求，因此不能直接应用于气象数据云存储应用中。为了解决该问题，提出了一种新型且安全高效的可搜索对称加密方案。该方案不仅可以满足用户对多数据源数据的加密和检索需求，而且可以保证敌手无法获取用户文档和搜索结果在各数据源中的分布信息，因而可以有效保证用户数据的隐私安全。

关键词：云存储，可搜索对称加密，隐私，安全

DOI：10.3969/j.issn.2095-1973.2018.01.010

Research and Development of Searchable Symmetric Encryption Schemes for the Security of Meteorological Data

Guo Cong, He Henghong, Tian Zheng, Zhong Lei
(National Meteorological Information Centre, Beijing 100081)

Abstract: Compared to the traditional storage pattern, user can effectively reduce the cost of data storage and management by outsourcing data to cloud storage, and enjoy the advantages of cloud storage such as high scalability and management. Researchers have paid much more attentions on this issue, which is one of the most widely applications and techniques in recent years. Data security has become a recognized issue. Using users' own private key to encrypt their own data before outsourcing is an effective way to avoid data private leakage, but it will destroy the query capability of data, and bring much more communication cost, storage and computation overhead. The existing searchable symmetric encryption schemes may solve this problem, but it only can be applied to one single data source. Meteorological data is the scarce and valuable resources in meteorology departments, so we should pay much more attentions to the data protection issues.Here, we propose a novel scheme, which may be applied to multiple data sources and has a high-performing. The test results show the our scheme may provide efficiently search query result without leaking users' information and query information.

Keywords: cloud storage, searchable symmetric encryption, privacy, security

0　引言

据市场调研公司Gartner的预测，全球数据总量将在2020年达到35 ZB，云存储市场潜力巨大。新技术的出现为数据存储和管理提供了极大便利，但数据隐私泄露等安全问题却越发突显[1-3]。数据所有者若将数据存储在云服务方，会对数据失去天然的物理管控能力，同时为黑客和云服务方对数据进行窥探、访问和利用提供便利和可乘之机。数据所有者不但缺乏对存储在云服务方数据的监督和管理能力，而且无法对云服务方的数据采取任何安全保护措施。其中，重要敏感数据面临更为严峻的泄露风险，数据隐私安全问题已逐渐成为云存储技术发展所面临的重大挑战[4]。尽管，云服务商可以与数据所有者签署协议，并在履行协议期间保证其能够对数据所有者的各类数据提供周密的安全保护措施，但这并不意味着从根本上解决了用户数据隐私泄露的问题，数据所有者更无法防止或有效监测云服务商对数据的偷窃、利用和分析。一方面，服务商出于对成本的考虑，可能并未真正履行其数据的安全保护义务。另一方面，假设服务方对托管数据进行了加密等相应的防护措施，但数据所有者并不可能持有云服务方的加密私钥，加密保护对云服务

收稿日期：2017 年 6 月 30 日；修回日期：2017 年 11 月 27 日
第一作者：郭聪（1982—），Email：guoc@cma.gov.cn
资助信息：国家气象信息中心青年科技基金课题（NMICQJ201701）

方自身来说形同虚设。数据所有者对数据进行自加密后将其存储到云服务方，是一种解决数据的隐私泄露问题的有效方法。用户只需确保自身加密秘钥不被泄露即可，这在很大程度上能够抵御恶意敌手或者云服务方对敏感数据的窥探、分析和攻击。

传统的加密技术（如AES、DES等算法）可以对数据提供有效的加密保护，但很难解决云服务方对已加密数据进行合理计算的请求，因此无法满足数据所有者对加密后的数据进行自我查询和检索等操作的需求。如果检索、查询的数据总量较小，数据所有者可以将数据下载至本地，再根据不同需求进行查询和检索。但如果检索、查询的数据总量过于庞大，这种方法会给系统带来巨大的通信、存储和计算开销。因此，为了方便数据所有者能够对数据进行加密，同时又能方便其对密文信息进行查询、检索等操作，一种新型的密码学原语——可搜索对称加密技术（Searchable Symmetric Encryption，SSE）[5]应运而生。数据所有者可以通过本地存储的私钥对数据进行加密，并通过该私钥对用户所搜索的关键词生成搜索令牌（Search Token），云服务商通过搜索令牌可对存储在云端的密文数据计算得到搜索结果，并把结果反馈给用户。在搜索过程中，该技术可以用于代替原有各类应用方案中的传统加密算法，使得云存储服务方无法获知用户数据和用户搜索的关键词等信息，因此有效地保护了用户数据的隐私安全。近年来，研究者们在文献[5]的基础上提出了一系列的技术改进方案[6-10]，并在安全性、高效性和功能性方面做出了权衡，各有利弊，满足了不同应用对安全的需求。但经过笔者分析，该类方案仍存在诸多局限性和不足，比如查询检索过程中可能泄露用户的搜索模式、搜索的关键词及数据源过于单一等问题。

在气象云存储应用场景中，对数据来源单一的假设显示是不合理的。根据技术手段不断演进和物联网技术的不断更迭，数据采集传感设备更新换代频繁，用户数据的来源和分布也越发广泛。比如：1）多设备场景：分布在全国各台站的环境数据采集传感器，7×24 h不间断地自动采集并生成观测数据（风湿压、高空、雷达、酸雨、水汽等多种气象数据资料）；2）多数据中心场景：业务数据分布在国家级主中心、国家级备份中心和各省分中心，数据分布广泛；3）数据共享场景：满足社会发展对数据共享的迫切需求，逐步放开气象数据共享。现有的可搜索对称加密方案在构造的过程中，均隐形地假定了数据源的单一性，因此不适用气象数据种类繁多、分布广泛的多数据源应用场景。

本文主要针对可搜索加密技术不支持多数据源的场景问题进行分析和讨论，对多数据源场景下可搜索对称加密技术在应用中所面临的难点进行重点讨论和分析，给出了一种支持多数据源可搜索加密的静态方案。研究中我们使用严格的密码学安全模型来证明方案的安全性，并根据真实的数据集对方案的性能进行分析和验证，通过试验分析可以证明本方案具备良好的安全性能。

1 系统模型和安全模型

本节只涉及静态可搜索对称加密方案（Multiple Data Source SSE，MDS-SSE）的相关定义、系统模型和安全模型。

定义一 多项式时间算法（Polynomial Time Algorithms）

如果存在一个多项式p，使得对于每个输入x，都存在一个$A(x)$，使得$A(x)$的计算步骤最多在$p(|x|)$步内终止。

定义二 不可区分性（Indistinguishability）

当出现窃听等安全威胁时，一个加密方案$\Pi=(Gen,Enc,Dec)$是不可区分性的加密方案，即任何敌手A在多项式时间内，存在一个可以忽略的函数，使得对于任意n，满足$\Pr[\mathrm{PrivK}^{eav}_{A,\Pi}(n)=1]\leqslant\frac{1}{2}+negl(n)$。也就是说，敌手攻击成功的概率的最大值高于1/2部分的概率是可以忽略掉，即与随机猜测等价。

定义三 静态MDS-SSE（static MDS-SSE，SMDS-SSE）

SMDS-SSE方案Π包含密钥生成、索引建立、索引合并、令牌搜索和搜索五个算法，其定义如下所示：

- $Key\leftarrow\mathrm{KeyGen}(\phi)$：给定参数$\phi$，通过密钥生成算法输出密钥$Key$；
- $Index_{dsid}\leftarrow\mathrm{BuildIndex}(Key,D,dsid)$：给定密钥$Key$，文档集合$D$和数据源身份$dsid$，通过索引表生成算法，输出子索引表$Index_{dsid}$；
- $Index\leftarrow\mathrm{MergIndex}(\{Index_{dsid}\}_{1\leqslant dsid\leqslant l})$：给定$l$个子索引表$\{Index_{dsid}\}_{1\leqslant dsid\leqslant l}$，通过索引表合并算法，输出全局索引表$Index$；
- $\sigma_{kw}\leftarrow\mathrm{SearchToken}(Key,kw)$：给定密钥$Key$和一个关键词$kw$，通过搜索令牌生成算法能够输出搜索令牌$\sigma_{kw}$；
- $ID(kw)\leftarrow\mathrm{Search}(\sigma_{kw},Index)$：给定搜索令牌$\sigma_{kw}$和全局索引表$Index$，通过搜索算法输出包含关键词$kw$的所有文档的标识符的集合$ID(kw)$。

1.1 系统模型

一个SMDS-SSE系统主要包含三类主要角色：数据源、数据使用者和云服务方。任一数据源拥有自己的数据集，存储在云服务方。数据使用者和任意被授权的用户可以同数据源共享同一密钥，并能够发出一定的数据搜索请求。云服务方存储所有的用户数据和数据的索引表，并能够应答数据使用者发出的关键词搜索请求。

该系统主要由初始化Setup协议和查询Query协议两部分，如下所示：

1）Setup：协议初始化操作由数据的使用者和服务提供方共同协商完成。数据源通过执行KeyGen算法得到密钥后，可与其他数据源的数据使用者共享密钥。各数据源在执行BuildIndex算法后，可为本地用户建立相应的子索引表，把加密后的密文文档集合和子索引表一同发给云服务方。云服务方通过执行MergIndex算法可将所有子索引表合并为全局索引。

2）Query：协议的查询请求需由数据使用者和云服务方共同协商交互完成。数据使用者通过SearchToken算法生成搜索令牌并发送给云服务方，云服务方收到请求后执行Search操作并将所有结果返回。

1.2 安全模型

我们把云服务方假设为敌手，那么它依照协议的执行过程试图获取用户隐私信息。根据文献[5]提出的CKA2安全性的定义进行修改，简要描述SMDS-SSE方案的CKA2安全性。

定义四 SMDS-SSE的CKA2安全性

假设$\Pi=$(KeyGen, BuildIndex, MergIndex, SearchToken, Search)是一个SMDS-SSE方案。$\varphi=(\varphi_{setup},\varphi_{query})$是方案的泄露函数，给定参数$\phi$，游戏执行者$P$，敌手$A$和模拟器$S$。Real和Ideal的试验游戏如下所示：

1）$\mathrm{Real}_{P,A}(\phi)$：游戏的执行者$P$执行KeyGen($\phi$)算法后可得到密钥$Key$。敌手$A$自行选择$l$个文档集合$D_1,...,D_l$发送给游戏执行者$P$。$P$对所有文档进行加密后，可得到密文文档集合$C$。通过索引算法对所有文档计算得到$l$个子索引表$Index_i \leftarrow \mathrm{BuildIndex}(Key,D_i,i)$，并把$C$和$Index_i$发给敌手$A$。敌手$A$选择若干个关键词并发出搜索请求，针对每个请求，游戏执行者可执行SearchToken(Key,kw)，以便可以得到搜索令牌σ_{kw}。最终，敌手通过抛币决定输出$b\in\{0,1\}$作为游戏的输出。

2）$\mathrm{Ideal}_{P,A,S}(\phi)$：敌手$A$选择$l$个文档集合$D_1,...,D_l$发送给游戏执行者$P$，$A$将$\varphi_{setup}$的输出结果发送给模拟器$S$，$S$模拟$l$个子索引表$\{I_i^*\}_{1\le i\le l}$发送给$A$。$A$发出若干搜索请求，对每个请求，游戏执行者$P$将$\varphi_{setup}$的输出结果发送给模拟器$S$。模拟器模拟搜索令牌$\sigma_{kw}^*$并发送给敌手$A$。最终$A$抛币决定输出$b\in\{0,1\}$作为该游戏的输出。

若敌手A在多项式概率时间内，总存在一个概率多项式的模拟器S和一个可忽略不计的函数$negl$满足：

$$|\Pr[\mathrm{Real}_{P,A}(\phi)=1]-\Pr[\mathrm{Ideal}_{P,A,S}(\phi)=1]|\le negl(\phi), \quad (1)$$

则说明方案Π满足CKA2安全性。

2 静态可搜索加密方案

Cash等[11]中提出了一种动态的可搜索加密方案，但该方案无法有效支持多数据源的应用问题。假使在该方案的基础上进行改进后，能够提供支持多数据源可搜索加密检索方案，但仍会存在诸多问题：1）假设每个数据源均能够访问在线计数表，但为实际应用增加了通信开销和计算开销，并且增加了安全隐患。比如，敌手通过监听在线计数表的更新次数，可以获知数据源包含的关键词的数量；2）该方案的安全性只能到随机预言机（Random Oracle）模型中进行证明，在该模型下是安全的，不一定说明在标准模型下就真正安全。因此，本节主要介绍我们提出的改进的静态可搜索加密方案，该方案允许用户建立本地数据子索引表，云服务商可以将子索引表进行合并，形成全局索引表以应答用户的搜索请求。首先，先介绍下我们方案中使用的符号。

2.1 符号说明

ϕ：安全参数

K_1, K_2：由密钥生成算法生成的密钥，用于伪随机函数和对称加密方案

Key：生成的密钥元组，其中$Key=(K_1, K_2)$

D：所有数据源文档的集合

KW：关键词集合

$dsid$：数据源的身份ID

L_{dsid}：整数列表

n_{dsid}：各$dsid$本地文档的数量

$Index_{dsid}$：各$dsid$的子索引表

K'：行置换密钥

$Index$：全局索引表

ψ：索引表的键值

γ：索引表的比特串

κ：与关键词和数据源身份相关的密钥

2.2 算法表述

本小节主要介绍协议中使用的具体算法。

2.2.1 密钥生成算法

随机选取长度为ϕ比特的密钥K_1, K_2，分别用于伪随机函数和对称加密方案中。密钥生成后，通过密钥

分发协议分发给其他数据源和数据使用者。

2.2.2 子索引表生成算法

为每个数据源利用子索引表生成算法建立子索引表，子索引表的结构类似元组，每个元组包括两部分：数据源的关键词和长度为n的比特串。n为所有数据源的文档集合，因此单个数据源仅使用n比特的一部分，每个数据源的文档互斥地使用n比特，即某个关键词出现在哪个数据源的哪个文档，则相应的n比特的对应位置为1，否则置为0。我们称长度为n的比特串为“行比特串”。算法具体描述如下：

算法一 BuildIndex

- 输入$(Key, D, KW, dsid, L_{dsid}, n_{dsid})$，其中$Key=(K_1, K_2)$为密钥元组，$D$为文档的集合，$KW$为关键词集合，$dsid$为数据源的身份ID，$L_{dsid}$为整数列表，$n_{dsid}$为本地文档的数量。
- 输出$Index_{dsid}$：子索引表。

1. $Key=(K_1, K_2)$
2. 计算$K' \leftarrow P_r'(K_1, 1)$，其中$P_r' : \{0,1\}^{\phi} \times \{0,1\}^{*} \rightarrow \{0,1\}^{\phi}$
3. 初始化数组$Index_{dsid}$，其长度为$|KW|$
4. 初始化计数器$counter=1$
5. 对给定的$kw \in KW$
   1）初始化计数器$i=1$
   2）初始化长度为n的零比特串γ
   3）对文档的集合D中的文档f，若$kw \in f$，则令$\gamma[L(i)]=1$
   4）$i=i+1$
   5）通过计算$\psi \leftarrow P_r'(K_1, 1 \| kw)$，$\zeta \leftarrow P_r'(K_1, 2 \| kw)$，得到$\psi, \zeta$
   6）通过计算$\kappa_{dsid} \leftarrow P_r(\zeta \| dsid)[1:n]$，其中$P_r : \{0,1\}^{\phi} \times \{0,1\}^{*} \rightarrow \{0,1\}^{n_{\max}}$
   7）令$\gamma = \gamma \oplus \kappa_{dsid}$
   8）令$Index_{dsid}[\mathrm{PR}(K', counter)] = (\psi, \gamma)$，其中$\mathrm{PR}: \{0,1\}^{\phi} \times \{1, ..., KW\} \rightarrow \{1, ..., KW\}$
   9）令$counter = counter + 1$
6. 输出$Index_{dsid}$

2.2.3 子索引表合并算法

当云服务方收到l个子索引表后，使用子索引表合并算法把它们合并为全局索引表。全局索引表设计成类似字典结构，每个条目存储合并后的关键词和比特串元组。对于云服务方来说，子索引表的每一行的l个关键词都需要进行合并。因为l个关键词是相同的，因此云服务方只需要随机选取其中一个作为合并结果即可，其中l行比特串需要通过异或运算进行合并。最终，云服务方会将合并后的元组插入字典中。具体算法描述如下：

算法二 MergIndex

- 输入l个索引表$\{Index_i\}$，其中$1 \leqslant i \leqslant l$。
- 输出全局索引表$Index$。

1. 初始化索引表$Index$
2. 对给定的$1 \leqslant i \leqslant l$，有：
   1）解析出$Index_i = (\psi_{i,1}, \gamma_{i,1}), ..., (\psi_{i,|W|}, \gamma_{i,|W|})$
   2）对于给定的$1 \leqslant j \leqslant |W|$：
      ① 令$\psi = \psi_{1,j}$，其中$\psi_{1,j} = \psi_{2,j} = ... = \psi_{k,j}$
      ② 令$\gamma = \gamma_{1,j} \oplus ... \oplus \gamma_{l,j}$
      ③ 令$Index[\psi] = \gamma$
3. 输出$Index$

2.2.4 搜索令牌生成算法

数据拥有者或者使用者通过搜索令牌生成算法为需要搜索的关键词生成搜索令牌。搜索令牌为(a, κ)元组，其中a用于确定全局索引表中关键词的对应条目，κ用于解密相对应的行比特串。

2.2.5 搜索算法

云服务方收到用户的搜索令牌后，首先在全局索引表中查找与a相同的键值，并读取加密的行比特串，再使用κ与其进行异或操作或解密。由索引表合并算法可知，加密的行比特串等于搜索结果与所有数据源的密钥κ异或的结果，因此再与κ进行异或，则可以得到搜索结果。

2.3 协议描述

2.3.1 协议初始化

协议初始化主要包含三个方面的工作：1）本地数据源加密；2）为各个数据源建立子索引表；3）把数据存储至云服务方，合并所有子索引表。

在初始化阶段，要对各个文档进行加密，然后存储至云端。但要考虑敌手监听和被动攻击方式，因此应采取相应的措施，防止敌手获取到文档和数据源的对应官，比如使用匿名技术对文档进行加密传送。通过该技术，每个文档对于敌手来说都来源于随机分配的数据源。

各数据源需要通过相互通信来获取文档总数，大致思路是：各数据源$DS_i(1 \leqslant i \leqslant k)$使用对称加密算法加密本地文档数量$n_i(1 \leqslant i \leqslant k)$并存储至云服务方。云服务方将收到的加密的$n_i$广播给所有$DS_i$。各数据源在收到广播消息后，自行解密$n_i(1 \leqslant i \leqslant k)$并求和得到总和$n$。由于云服务方收到的均为密文，因此无法获取各

数据源n_i的值。

各数据源知道了文档总和n，想进一步得到那些各数据源可以自己使用的比特位，即计算出L_i，则需要使用到伪随机置换，如下所示：

$$P_r:\{0,1\}^{\phi}\times\{1,...,n\}\rightarrow\{1,...,n\},\qquad(2)$$

即将$(1,...,n)$置换为$(a_1,...,a_n)$。比如，DS_1有n_1个文档，则应选取$a_1,...,a_{n_1}$生成L_1。DS_2有n_2个文档，则应选取$a_{n_1+1},...,a_{n_1+n_2}$生成$L_2$，以此类推，则所有$L_i$都互斥在相应的比特位。

使用子索引表生成算法，生成子索引表存储至云端，并进行索引表合并，最终生成全局索引表。

2.3.2 协议请求

用户使用搜索令牌算法对关键词进行查询操作，把算法的执行结果发送给云服务方。服务方收到消息后，通过搜索算法进行搜索，并将查询结果反馈给用户，协议执行完毕。

算法三 Query

1. 用户输入关键词kw，并通过输入kw, n, Key根据搜索令牌生成算法计算τ_{kw}。
2. 用户发送τ_{kw}给存储服务服务提供商。
3. 通过输入τ_{kw}、合并的索引表I和初始化字典FDIC，可根据查询算法计算得出ID_{kw}，并把查询结果转发给用户。

3 试验分析

本节我们比较该方案对不同数据集（表1）产生的索引表的大小和查询的效率，所有试验程序均由Python语言编写完成。搭建的测试环境为MacbookAir，处理器为Intel(R) Core(TM) i5-3427U 1.8 GHz，4.00 GB内存的笔记本电脑，所有测试结果均为试验100次的结果的平均值。

表1 数据集说明

Table 1 Description of the dataset

数据集名称	文档数量	关键词数量
气象数据（地面观测数据文本文件）1	1000	118433
气象数据（地面观测数据文本文件）2	2000	345973
气象数据（地面观测数据文本文件）3	3000	760729
气象数据（雷达图数据png文件）	1000	236853
邮件1	1000	889733

我们选取的三类文本数据分别为：中国地面气象站逐小时观测资料数据集、气象数据雷达图、邮件数据集（这三类是比较具有代表性的试验性质文档数据集）。其中论文下载字中国气象数据网（http://data.cma.cn/），邮件数据集为Enron Email（http://www.cs.cmu.edu/~enron/）的一个子集，网页文档爬取自DBLife数据库[12]，本文的试验中的关键字集合，爬取自Word Frequency Data（https://www.wordfrequency.info/free.asp）。

首先，使用三个同一类型不同文档总数的数据集进行试验。试验结果表明，索引表大小与数据源数量无关，索引大小与文档总数呈线性增长关系。经过分析，发现原因有二：1）方案的索引表大小随关键词与文档标识符对数量的增多而增多；2）该方案索引表中的行比特串长度等于文档总数，因此索引表大小也相对于文档总数呈线性增长趋势。

由此部分试验的结果，结合气象数据的实际进行分析，如果把本文的想法应用于气象数据的云存储应用中，首先要对气象数据进行预处理。对于不能进行直接共享的原始数据和涉密气象数据，可以把数据按气象数据类型进行分类（比如目前气象数据分为14大类200多子类），然后进行加密处理存储在云服务方。试验结果表明，索引的大小与文档的总数成正比，如果文档总数过大，会造成索引的存储和计算量过大。因此，从气象数据文件过多，但单个文件大小又较小的实际情况考虑，我们可以把每个时次的某类数据、每天的某类数据或某旬、月、季的某类数据当做一个大文件来处理。比如，质控后某小时国家级气象自动站的数据是2414小文件，那么可以把这些小文件合并成一个大文件进行处理，或者把2414个小文件看作是一个大文件。当搜索某时次某台站的该类资料时，首先搜到的是这个大类文件，然后再在该类文件的子类中进行搜索，相当于树型结构的层次关系。这样就能解决因气象数据文件数量过大引起的索引表数据量过大所带来的计算和存储负担的问题。

其次，使用三种不同类型相同数量的数据集进行试验。试验结果表明，我们方案的索引表大小保持不变。因为，索引表大小与文档数量有关，与文档关键词无关，因此索引表大小保持不变。

结合该类试验结果和气象数据的实际，我们考虑了如下的解决方案。以质控后的一体化国家站地面气象要素资料为例，进行具体的说明。该类资料的文件名为：Z_SURF_C_BABJ_20170419235955_O_AWS_FTM_PQC.txt。

首先，可以从此资料的文件名中提取关键字Z_SURF、BABJ、20170419235955、AWS_FTM和PQC，作为该文件的关键字。当然，对于20170419235955这类关键字，可以继续细分成年、月、日、时、分、秒等。同时，可以把分布于各省的国家站、区域自动站的站号、所属区县、所属省

等信息作为关键字，进行索引和分析。再比如，以中央台的6小时精细化预报指导产品为例，文件名为：Z_SEVP_C_BABJ_20170420024051_P_RFFC_SNWFD6H_201704201200_02406.txt。

从该类资料的文件名中可以提取关键词：Z_SEVP、BABJ、20170420024051、RFFC、SNWFD6H、201704201200和02406。此外，针对此类文件报文内容，还可以提取出资料编报中心码CCCC、TT、AA、ii、台站号、经度信息、纬度信息等。

由于索引表的大小与文档关键词无关，因此我们可以尽可能多的选取关键词进行索引的建立，以便为用户提供更加全面的数据搜索和检索服务。

我们进一步讨论该方案与文献[11]中的方案在不同文档中的搜索时间消耗问题。由于搜索时间与数据集类型无关，因此我们采用单一的变量方式进行比较，分别测试了搜索时间返回结果占总文档数量的比例、文档数量及数据源数量的变化情况。我们仅针对文本类数据集进行讨论。根据协议的设计，每一个搜索包含搜索令牌生成和搜索两部分，我们分别通过试验记录了它们的执行时间，如图1所示。

每一次搜索查询，包括字典搜索、结果解密和结果提取三个部分的工作。图1a表明，当固定了文档数量和数据源数量后，搜索时间均与返回结果数量线性增长趋势，但我们的方案增长幅度较小。图1b表明，当我们把文档从1000提升至3000，数据源仍为2时，搜索结果返回5%的文档。我们的方案中行比特串的长度取决于文档的数量，比特串越长，计算操作消耗的时间就越多。搜索令牌生成时间和搜索时间虽文档数量线性增长。从图1c可以看出，测试数据量的变化对我们方案的搜索时间的影响非常小，当数据源达到10时，我们的方案的总体搜索耗时仍然很小。

综上所述，经过多方对比，我们的方案更加高效且安全性更高。

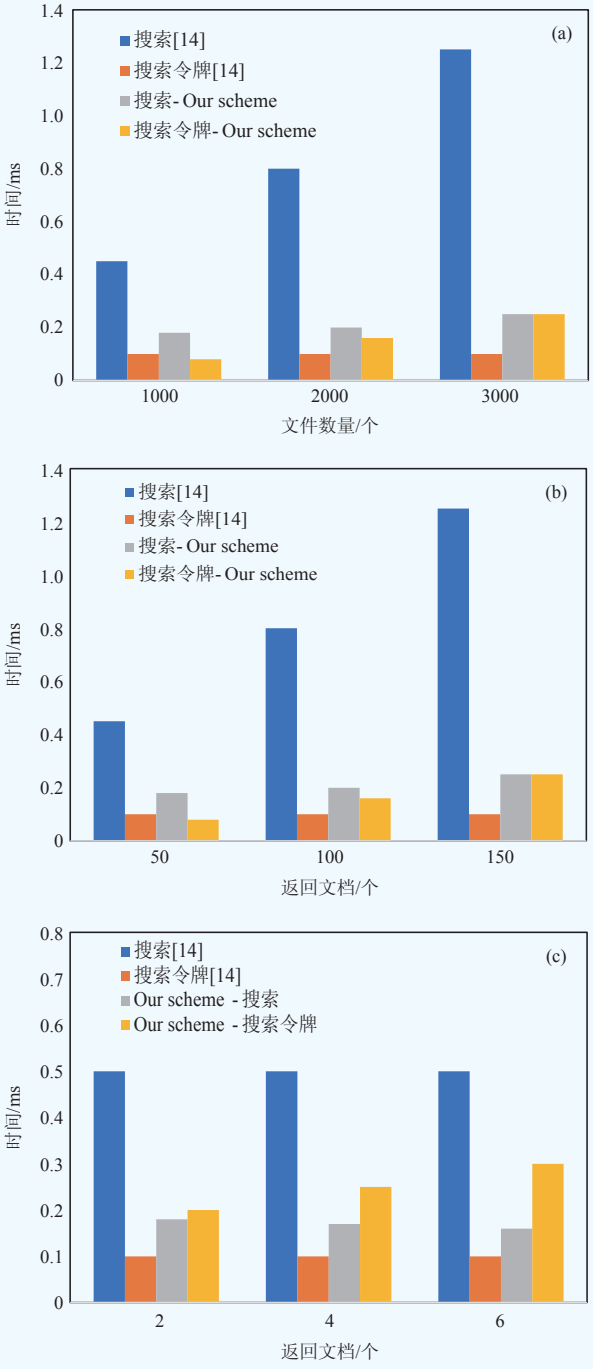


图1 新方案搜索时间的试验结果分析：（a）气象数据1，数据源为2；（b）气象数据1–3，数据源为2；（c）气象数据1

Fig. 1 The searching time by this new scheme: (a) data type 1, data source 2; (b) data type 1–3, data source 2; (c) data type 1

4 未来工作

1）数据更新

在绝大多数实际应用中，对存储在云端的加密数据进行关键词检索都应该是动态更新的。目前，已有部分研究者对动态可搜索对称加密方案进行了研究，但针对多数据源的动态可搜索对称加密技术还十分匮乏，未来可以对该方案在动态多数据源可搜索对称加密方面进行深入研究和拓展，使其更能满足气象数据存储和应用的实际需求。

2）数据类型能力扩展

传统的可搜索对称加密技术是在加密文本上进行关键词的搜索计算，但针对气象部门的数据种类繁多，既有结构化数据，也有非结构化数据的特点，应该深入分析，提出具体的解决办法予以解决。比如，针对雷达图片的数据检索和处理，未来可能会要给图

片类数据添加标签，以方便检索，以及满足各类气象数据云存储安全的需求。

3）支持多用户查询

目前，已有研究者提出了针对多用户的可搜索对称加密技术的概念，该类型的方案主要针对广播群组加密进行设计。该方案允许所有者授权多个用户对其数据进行关键词搜索，并且赋予动态的增删改查操作权限，也就是说该方案支持同一个数据的多个数据使用者，因此更符合目前的实际应用需求，是可搜索对称加密技术的未来发展方向。未来，针对气象数据云存储和数据共享服务的需求，在支持多用户对不同类数据文档的访问权限、下载权限和查询检索方面进行深入研究，以能满足气象大数据上云后，对不同权限用户的数据访问控制，保证气象数据安全，进一步为用户提供更为智能化安全的气象数据共享服务。

参考文献

[1] Seny K, Kristin L. Cryptographic cloud storage. In International Conference on Financial Cryptography and Data Security (FC-10), 2010.

[2] Wei L, Zhu H, Cao Z, et al. Security and privacy for storage and computation in cloud computing. Information Sciences, 2014, 258: 371-386.

[3] Chang V, Ramachandran M. Towards achieving data security with the cloud computing adoption framework. IEEE Transactions on Services Computing, 2016, 9(1): 138-151.

[4] Rittinghouse J, Ransome J. Cloud computing: Implementation, management, and security. Boca Raton: CRC Press, 2009.

[5] Reza C, Juan G, Seny K. Searchable symmetric encryption: Improved definitions and efficient constructions. In Proceedings of the 2006 ACM SIGSAC Conference on Computer and Communications Security (CCS-06), 2006.

[6] Mehmet K, Mohammad S I, Murat K. Efficient similarity search over encrypted data. In Proceedings of the IEEE 28th International Conference on Data Engineering (ICDE-12), 2012.

[7] Stanislaw J, Charanjit J, Hugo K. Outsourced symmetric private information retrieval. In Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security (CCS-13), 2013.

[8] Emil S, Charalampos P, Elaine S. Practical dynamic searchable encryption with small leakage. In Network and Distributed System Security Symposium (NDSS-14), 2014.

[9] Sophos R B. Forward secure searchable encryption. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS-16), 2016.

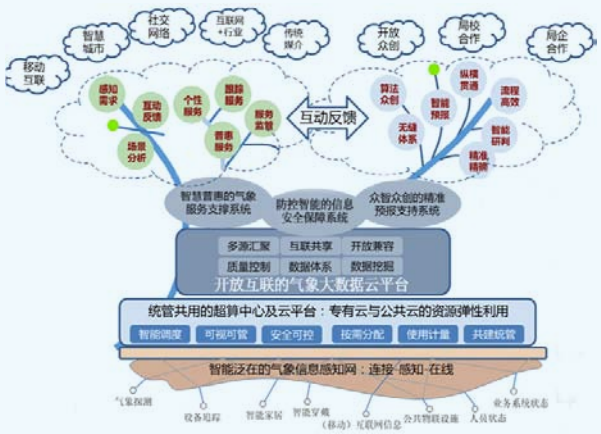
[10] Gilad A, Moni N, Gil S, et al. Searchable symmetric encryption: Optimal locality in linear space via two-dimensional balanced allocations. In European Symposium on Research in Computer Security (ESORICS-16), 2016.

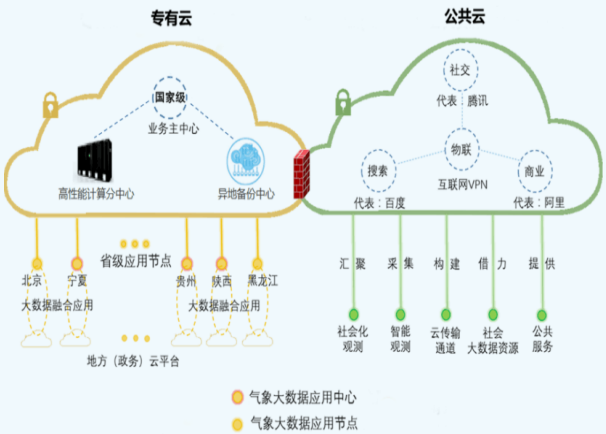
[11] Cash D, Jaeger J, Jarecki S, et al. Dynamic searchable encryption in very large databases: Data structures and implementation. In Network and Distributed System Security Symposium (NDSS-14), 2014.

[12] DeRose P, Shen W, Chen F, et al. Dblife: A community information management platform for the database research community. In CIDR, 2007.

气象信息化系统工程总体结构和布局

气象信息化系统工程在全国以“1主数据中心+1备数据中心+31应用节点”的气象“专有云+公共云”模式布局。气象信息业务系统实现国-省两级部署，市县以终端模式连接云端，打造“云+端”模式，实现气象业务扁平化和气象信息服务均等化。气象信息化系统工程将主中心设于北京，西安分中心承担数据备份以及数据分析的任务，京外分中心与北京业务主中心共同构建高性能计算中心，承载部分高性能计算业务。其他各省级节点与业务主中心通过主干宽带网互联，省内通过省内宽带网实现互联。感知层设备可通过互联网、移动网等多种途径将气象感知数据就近接入省级节点，进而上传到气象云平台。





——摘自《气象信息化发展规划(2018-2022年)》，2017年12月