DOI:10.13878/j.cnki.jnuist.2017.05.011

沈蒙[1]　赵梦蕉[1]　祝烈煌[1]　马宝利[1]

支持近似最短距离查询的高效图加密机制

摘要

近似最短距离查询是图检索的基本模式.为了保护外包数据安全,通常对图数据进行加密.已有加密方案使用两跳覆盖模型构建加密图索引,导致索引结构复杂,降低了查询效率.本文提出了一种基于图压缩的加密机制,可以提高图的检索效率,并且支持加密图最短路径查询.该机制使用 K-mediods 聚类使得图中的节点按照距离分成 K 个簇,每个簇内的节点使用其中心节点代理,当查询2个点间最短距离时,对于相同簇内的点直接查询,对于簇间的点使用代理节点查询距离.实验结果表明该机制有效地减少了查询时间,提高了查询效率,且查询结果误差度在可接受范围内.

关键词

近似最短距离;K-mediods 聚类;图压缩

中图分类号 TP301

文献标志码 A

0　引言

随着信息时代的迅猛发展,云计算更加方便和普及,个人和企业将大量的数据外包给云服务器,使用第三方的云平台服务器进行计算处理和数据存储.但是,由于云服务器是半可信的,外包数据存在安全隐患.因此,需要既能够保持云计算的优势,又要保护数据的隐私不被泄露.图结构数据在现实生活中具有广泛应用,如道路信息或社交网络[1-3]等.因此,关于图的加密搜索成为研究的重点内容.

由于云服务器的半可信特征,常常将数据本地加密上传到云服务器,在密文下进行查询,在服务器端计算复杂度较高,如何有效地提高云服务器的查询效率是研究的难点问题.现有的加密图查询研究,将图使用哈希函数或二叉树等形式构建查询索引,在云服务器查询时仍需要耗费大量时间,需要优化查询搜索的构建方案.

为了解决上述问题,本文提出了支持近似最短距离查询的高效图加密机制(Efficient graph encryption mechanism for Approximate shortest distance Search,EAS).该机制对于原始图中的节点进行预处理,根据节点间距离分成 K 个簇,同一个簇内的节点使用其中心节点代理.当查询2个簇内的点的最短距离时,使用代理点间的距离代替.这样使得在云服务器上的查询计算时间大大缩短,且降低的精确度在可接受的范围内.实验结果表明,与准确查询方法相比,EAS 可以有效地减少查询时间,提高查询效率.

1　相关工作

1.1　图加密搜索技术

图加密是一种结构化加密,是将图的数据结构加密,且能在隐私保护的情况下进行查询.结构化加密首先是由 Chase 等[4]提出的.由于最短路径的查询是最常使用的关于图的查询,近年来,对于最短路径查询的方案取得了很多研究成果.Cash 等[5]提出了支持大量数据进行可搜索加密的方案,但是该方案只支持布尔查询.Zhu 等[6]提出一个使用干扰算法以及支持同义词查询的最短路径搜索加密机制(SPSQ).Meng 等[7]提出了在对图加密之前构造距离预言机(distance oracle)的数据结构,有效地提高了检索的效率,并且支持近似地查询最短距离.该方法通过牺牲一定准确度来达到有效计算,且没有支持图的动态更新.Wang 等[8]和 Haynberg 等[9]针对以上问题提出了一种支持有效更新的确切

收稿日期 2017-07-01

资助项目 北京市自然科学基金(4164098);国家自然科学基金(61602039);国家重点研发计划(2016YFB0800301)

作者简介

沈蒙,男,博士,讲师,硕士生导师,研究方向为云计算隐私保护.shenmeng@ bit.edu.cn

祝烈煌(通信作者),男,博士,教授,博士生导师,主要从事网络与信息安全方向的研究工作.liehuangz@ bit.edu.cn

1　北京理工大学　计算机学院,北京,100081

最短路径查询的图加密搜索方案,使用额外的存储空间存储图的相关信息(节点邻居信息和连接表),便于在密文上进行可修改的同态加密,并利用历史查询信息作为缓存,可以快速返回查询过的信息.Wu 等[10]提出了对于原始街道地图信息压缩,使用基于隐私信息查询 PIR(Private Information Retrieval)和混乱电路(garbled circuits)支持导航隐私保护的加密方案.但是该方案中涉及到计算查询的每一个中间跳时都需要客户与服务器的交互.

2 问题定义与描述

2.1 系统模型

支持近似最短距离查询的高效图加密机制是图拥有者将图预处理后加密发送到云服务器,用户通过查询条件从云服务器返回加密的图信息[11].通常这类系统模型中主要涉及的实体至少有 3 个,其中包括图的拥有者(Graph Owner)、云服务器(Cloud Server)以及查询用户(User),其模型如图 1 所示.

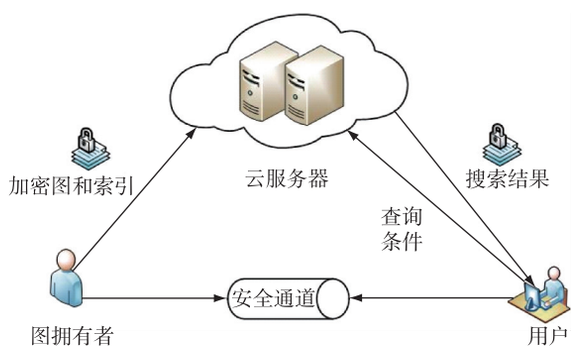

图 1 加密图查询最短路径的系统架构

Fig. 1 Architecture of the shortest path search over encrypted graph

在该模型图中,各部分实体的功能介绍如下:

1)图拥有者是图结构的提供者.通常,图结构只有图的拥有者可以知道,因此在将图外包上传给云服务器之前,为了保护图结构的隐私,将对图结构进行加密,同时为了提高查询效率,会考虑本地完成索引结构,并加密后一并交给云服务器进行管理,提供给自己或他人进行图结构的相关搜索.

2)云服务器主要对图结构拥有者提供的加密图结构和加密索引表进行管理,以及对查询用户提交的图相关的检索请求进行检索并返回结果.通常来说,云服务器无法了解到图的明文结构,但是云服务器是半可信的.

3)查询用户将待查询问题加密后提交给云服务器进行检索,并对云返回的检索结果图像进行解密得到结果.

2.2 相关定义

2.2.1 问题描述

给予有向图 $G=(V,E)$,节点总数为 $n=|V|$,边的总数为 $m=|E|$.最短路径的查询条件为 $q=(u,v)$,是要求的点 u 和点 v 之间的最短路径长度,记为 dist_q.问题描述为:对于给定的图 G,比如代表道路或社交网络,以加密的形式外包到云服务器,用户需要在隐私保护的情况下查询从源点 s 到目的点 t 的最短距离 dist_q[12].

2.2.2 保序加密算法方案

本文对于图中的边加密使用保序加密算法方案,其中 ORE(Order-Revealing Encryption)是一种保序加密方案且安全性很高[13].该方案包含 $\prod^{\mathrm{ore}}=(\mathrm{Gen},\mathrm{Enc},\mathrm{Cmp})$ 3 个多项式时间算法:

1) $\mathrm{Gen}(1^{\lambda})\rightarrow sk$:是以安全参数 λ 为输入的概率函数,输出密钥 sk;

2) $\mathrm{Enc}(sk,m)\rightarrow ct$:是输入密钥 sk 和消息 m,输出密文的概率函数;

3) $\mathrm{Cmp}(ct_1,ct_2)\rightarrow z$:是输入 2 个密文,比较它们明文的大小,并输出一个比特 $r\in\{0,1\}$.

3 支持近似最短距离查询的高效图加密机制

本文使用的加密图的最短路径检索机制为 Enc-GraphSearch,加密方案 $\prod=(\mathrm{KeyGen},\mathrm{Preprocess},\mathrm{Setup},\mathrm{Query})$ 由以下 4 个算法构成:

1) 密钥生成算法 $\mathrm{KeyGen}(\lambda)\rightarrow sk$:是以安全参数 λ 为输入的概率函数,输出密钥 sk;

2) 图预处理算法 $\mathrm{Preprocess}(K,G)\rightarrow G'$:是将原图 G 中所有的节点聚类并分成 K 个簇,得到新的图 G';

3) 图加密算法 $\mathrm{Setup}(sk,G')\rightarrow\triangle$:是使用密钥 sk 对图 G' 进行加密,输出加密索引;

4) 检索算法 $\mathrm{Query}((sk,q),\triangle)\rightarrow\mathrm{dist}_q$:是以加密的查询条件和加密索引为输入,输出所求 2 点的最短距离 dist_q.

3.1 图预处理算法

在将图外包到云之前,对于明文下的图中的节点和边作出相应的处理,使用压缩图进行检索,提高外包到云后图的检索效率.

3.1.1 图聚类

由于图的节点数庞大,在对图进行加密和检索

的时候都需要消耗较长的时间,即使是在云平台上,这个时间也是不可忽视的.在没有严格精度要求的情况下,对于本文中定义的问题,求解图中 2 点间的最短距离取近似的结果是可以接受的.

本文采取图聚类方法基于 K-mediods 聚类,即对于图中欧氏距离较近的点聚成一簇,最终形成 K 个簇.图聚类也就是把图 $G=<V,E>$划分成为 K 个不相交的子图 $G_i=<V,E>$.

K-mediods 聚类算法思想如下:

1) 初始化中心点:随机选取图中的 K 个点作为中心点;

2) 更新中心点:寻找新的中心点,使得新的中心点到该类簇的其他点距离最小;

3) 分配样本点到中心点:分配其他点到距离最近的中心点,形成新簇,计算误差平方和,若与上一次迭代的误差平方和相同,则停止,否则继续执行第 2)步.

3.1.2 构建代理点

本文提出了"代理点"的概念,即对于在同一个簇中的点,都可以使用其类簇的中心点来代替.使用代理点能够有效地减少查询时间.比如说,当查询条件为 $q=(u,v)$时,如果顶点 u,v 在不同的 2 个类簇 c_1,c_2 中,那么使用类簇的中心点 m_1,m_2 来代替,而此时 m_1,m_2 的距离即可近似地看作查询的距离.另外,当查询条件为 $q=(u,v)$时,如果顶点 u,v 在相同的 2 个类簇 c_1,c_2 中,就可以减少查询的点到某一个特定的类簇中,这样极大地减少了查询的时间.

3.1.3 图预处理算法

基于以上图聚类和构建代理点的思想,提出了如下算法 1 的图预处理算法.通过输入原始图和聚类簇的数目 K 值,得到由中心点构成的新图.

算法 1 Preprocess algorithm for EncGraphSearch

输入:聚类成簇的数目 K 和原始图 G
输出:由中心点构成的新图 G'
1:对于原始图 G 中的顶点进行 K-mediods 算法,得到 K 个类簇分类结果集 $c_i \in Clusters$ 和中心点集 $m_i \in mediods$
2:**for each** $c_i \in Clusters$ **do**
3: **for each** $c_{i+1} \in Clusters$ **do**
4: **if** c_i 与 c_{i+1}之间有边 **then**
5: **Compute** $\mathrm{dist}_{c_i,c_2}=D(m_i,m_{i+1})$
6: **Insert** $(m_i,m_{i+1},\mathrm{dist}_{c_i,c_{i+1}})$ **into graph** G'
7: **break**
8: **end if**
9: **end for**
10:**end for**
11:**return** G'

在算法 1 中,首先对原始图中的节点聚类得到结果集和中心点集,然后判断分成的簇内的点之间是否有边,若有边则计算中心点距离并加入到新生成的图中,如第 2—10 行所示.

3.2 图加密算法

在加密图的时候,包括对于聚类后代理点表、图和预处理后图的加密.其中原图和预处理后图加密的原理相同,都使用如算法 2 进行加密.算法 2 中输入原图和密钥,得到加密的图索引.其中使用到 2 个伪随机函数 h 和 g 以及一个带有安全参数的同态加密的函数 f,例如 AES 算法[14].

算法 2 Preprocess algorithm for EncGraphSearch

Input:密钥 sk,原图 G.
Output:加密图索引($\triangle$)
1:生成原图 G 的两跳索引结构$\triangle$,设置 $sk_d=h(sk,1)$
2:**for each** $u \in G$ **do**
3: **Set** $T_u=h(sk,u||1)$
4: **for each** $(v,d_{u,v}) \in \triangle(u)$ **do**
5: **Compute** $V=h(sk,v||0)$.
6: **Compute** $D^1_{u,v}=f(sk,d_{u,v})$.
7: **Compute** $D^2_{u,v}=\mathrm{ORE.Enc}(sk_d,d_{u,v})$.
8: **Insert** $(V,D^1_{u,v},D^2_{u,v})$ **into the dictionary** $I[T_u]$.
9: **end for**
10:**end for**
11:**return** $(\triangle)=I$

在算法 2 中,生成图的索引结构[15]为 u $(v,d_{u,v})$,分别对密钥和节点 u 取哈希值加密.对于每个与顶点 u 有边的顶点 v,对他们之间的距离使用 AES 加密和 ORE 加密,并将加密结果保存到索引中,如第 4—9 行所示.距离使用 2 种加密方式,原因是需要对加密距离值进行比较并满足返回到用户处可以解密的要求.

3.3 检索算法

本文提出的云服务器上查询算法如算法 3 所示.用户提出要查询的初始点 s 和终点 t,即查询条件为 $q=(s,t)$,使用密钥 sk 进行加密后 $\tau_s=h(sk,s||1)$,$\tau_t=h(sk,t||2)$发送到云服务器.服务器通过查询加密代理表,查看顶点所在类簇信息.如果顶点在不同的 2 个类簇中,则使用聚类后的新图形成的索引$\triangle_2$查询,代理点的距离即可近似地看作查询的距离.另外,若顶点在相同的 2 个类簇中,就可以减少查询的点到某一个特定的类簇中,减小查询图的规模.

算法 3 Query algorithm for EncGraphSearch

Input:加密后的查询条件(τ_s,τ_t),图的加密索引$\triangle_1$,代理点的加密索引$\triangle_2$
Output:查询结果 $dist_\tau$
1:在索引$\triangle_2$里寻找查询条件(τ_s,τ_t)对应索引(τ_s',τ_t'),及其所在簇的编号 Cluster(τ_s'.ClusterID,τ_t'.ClusterID)和簇的中心点 mediods(τ_s'.mediod,τ_t'.mediod)
2: **if** τ_s'.ClusterID=τ_t'.ClusterID **then**
3: 在索引$\triangle_1$中查询(τ_s'.mediod,τ_t'.mediod)的最短路径
4: **break**
5: **end if**
6: **else if then**
7: 在索引$\triangle_1$中查询(τ_s',τ_t')的最短路径
8: **break**
9: **end if**
10:**return**

3.4 安全性分析

对加密图搜索机制 EAS 提出的加密方案$\prod$进行安全性分析.将图结构外包到云服务器时,由于云服务器是半可信的,可以对图结构的密文进行一定操作,那么就有可能泄露一定的信息,经分析本文提出的加密方案$\prod$是隐私安全的.

定义 1 在半可信的云环境下,方案$\prod$是隐私安全的:

1) 方案$\prod$保证在云服务器不能推断出原始的明文或最终结果;

2) 方案$\prod$是抗选择查询攻击的.

由于对图结构的点加密是使用带盐的哈希函数,每个相同的点都具有不同的加密值.对点间距离值的同态加密使用长度为 128 的安全参数,使用穷举法攻击破解时的时间复杂度为 $O(2^n)$,复杂度按照指数增长,因此方案是实际有效安全的.

选择查询攻击的攻击者 A 伪造查询条件和加密索引结构,由于加密函数f,g,h 和 ORE 都是安全的,因此伪造的与真实值是可区分的.那么对于任意多项式时间敌手 A 对于真实查询和模拟查询具有不可区分性.

4 实验结果与性能评价

4.1 实验设置

对上述机制的基本加密函数实现是基于 OpenSSL 数据库完成的,实验环境的配置为处理器 2.5 GHz,内存 8 GB.数据集采用随机生成的 3 个不同规格的无向图,如表 1 所示.实验对比方案涉及明文和密文、不同聚类的 K 值以及不同查询条件下的图最短路径搜索.以构建索引时间和大小,查询时间和准确性 4 个方面作为衡量提出加密图搜索机制 EAS 的效率指标.

表 1 实验所用数据集

Table 1 Datasets for experiments

数据集	节点数	边数	数据集大小/KB
DataSet1	4 625	53 291	1 051
DataSet2	10 876	70 753	1 425
DataSet3	21 721	93 993	1 987

4.2 实验方案

4.2.1 索引时间和大小

在构建索引的时候,包括明文生成图的索引和加密索引 2 部分.定义明文生成索引时间包含对图的聚类,以及对生成聚类后的图和原图的两跳索引的总时间.构建索引的时间和大小如表 2 所示.

表 2 构建索引时间和大小

Table 2 Time and size of index construction

数据集	明文		密文	
	时间/min	大小/MB	时间/min	大小/MB
DataSet1	467.3	6.61	497.7	27.42
DataSet2	8 710	6.11	8 768	23.68
DataSet3	260.45	12.59	278.01	49.10

由表 2 可以看出 3 种数据集构建索引大小和时间有较大的差别,这是因为由于图的结构不同造成的.同一个数据集的密文和明文构建索引时间相近,密文下构建索引的大小是明文的 4 倍左右.

4.2.2 查询时间和准确性

查询时间体现算法的有效性,定义查询时间为提交查询条件到得到查询结果这一段时间.使用准确查询和近似查询 2 种方法做对比.本实验中随机生成 50 个查询条件,比较数据集 DataSet1 不同 K 值的情况下查询时间的变化.观察 K 值对查询时间的影响,如图 2 所示.

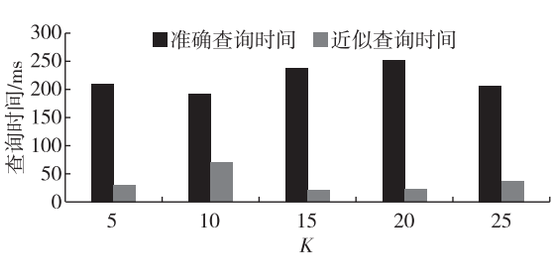

图 2 K 值对查询时间的影响

Fig. 2 Impact of K on query time

由图 2 可以看出,当使用确切查询方法的时候,

耗费查询时间较多,是使用近似距离查询机制 EAS 的 10 倍左右,且 K 值过大或过小都会导致 EAS 机制的查询时间增长,这是因为 K 值较小或较大时都与原图差距较小,聚类的效果不明显.

近似距离查询机制会使结果失去一定精度,通过 K 值的选取控制误差率.定义准确率为近似查询机制的结果与准确查询的比值.如图 3 所示为几个特定的 K 值对查询准确率的影响.

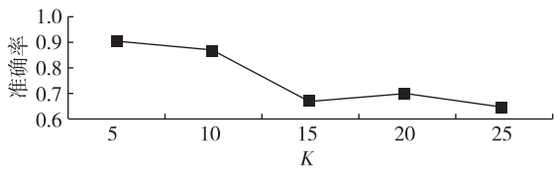

图 3 K 值对查询准确率的影响

Fig. 3 Impact of K on query accuracy

观察图 3 可以看出聚类簇的 K 值越小,查询的准确度越高,越接近确切查询结果值.综合比较 K 值对时间和准确度的影响,使用 K 值为 10 时,能够得到较好的结果,准确度达到 90%左右.

5 结束语

本文针对现有加密图检索时间长的问题,提出一种支持近似最短距离查询的高效图加密机制.该机制使用图聚类的算法和“代理点”的思想,通过可搜索加密的框架对图进行加密,并能够支持近似最短路径的查询.使用真实的数据集实验结果表明,该机制有效地提升了查询效率,并且误差率在可接受范围内.

参考文献

References

[1] Vieira M V, Fonseca B M, Damazio R, et al. Efficient search ranking in social networks[C]// Sixteenth ACM Conference on Information & Knowledge Management, 2007:563-572

[2] Wei F. Tedi: Efficient shortest path query answering on graphs [M] // Sakr S, Pardede E. Graph data management: Techniques and applications. Hershey, PA: IGI Global, 2010:99-110

[3] Yahia S A, Benedikt M, Lakshmanan L V S, et al. Efficient network aware search in collaborative tagging sites[J]. Proceedings of the VLDB Endowment, 2008, 1(1):710-721

[4] Chase M, Kamara S. Structured encryption and controlled disclosure[C]// International Conference on the Theory and Application of Cryptology and Information Security, 2010:577-594

[5] Cash D, Jarecki S, Jutla C S, et al. Highly-scalable searchable symmetric encryption with support for Boolean queries[C]// Advances in Cryptology-CRYPTO, 2013: 353-373

[6] Zhu H, Wu B, Xie M Y, et al. Secure shortest path search over encrypted graph supporting synonym query in cloud computing [C] // IEEE Trustcom/BigDataSE/ISPA, 2016:497-504

[7] Meng X R, Kamara S, Nissim K, et al. GRECS: Graph encryption for approximate shortest distance queries[C]// Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015:504-517

[8] Wang Q, Ren K, Du M X, et al. SecGDB: Graph encryption for exact shortest distance queries with efficient updates[C]// The 6th International Conference on Frontier Computing, 2017, Accepted

[9] Haynberg R, Rill J, Achenbach D, et al. Symmetric searchable encryption for exact pattern matching using directed acyclic word graphs [C] // International Conference on Security and Cryptography, 2013:1-8

[10] Wu D J, Zimmerman J, Planul J, et al. Privacy-preserving shortest path computation [J]. arXiv e-print, 2016, arXiv:1601.02281

[11] 朱旭东,李晖,郭祯.云计算环境下加密图像检索[J].西安电子科技大学学报(自然科学版),2014,41(2):151-158
ZHU Xudong, LI Hui, GUO Zhen. Privacy-preserving query over the encrypted image in cloud computing[J]. Journal of Xidian University (Natural Science), 2014, 41(2):151-158

[12] Samanthula B K, Rao F Y, Bertino E, et al. Privacy-preserving protocols for shortest path discovery over outsourced encrypted graph data[C]// IEEE International Conference on Information Reuse and Integration, 2015: 427-434

[13] Lewi K, Wu D J. Order-revealing encryption: New constructions, applications, and lower bounds[C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016:1167-1178

[14] Katz J, Lindell Y. Introduction to modern cryptography [M]. London: CRC Press, 2007

[15] Akiba T, Iwata Y, Yoshida Y. Fast exact shortest-path distance queries on large networks by pruned landmark labeling[C]// Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, 2013: 349-360

Efficient graph encryption mechanism for approximate shortest distance search based on cloud

SHEN Meng[1] ZHAO Mengjiao[1] ZHU Liehuang[1] MA Baoli[1]
1 School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081

Abstract Approximate shortest distance query is the basic pattern of graph search. The graph data is usually encrypted in order to protect the security of the outsourced data. The existing encryption schemes use the two-hop labeling model to construct the encryption index, which leads to the high-complexity of index structure and the reduction of query efficiency. This paper proposed an algorithm based on graph compression, which can improve the efficiency of the graph query and support the approximate shortest distance query of encrypted graph. The algorithm uses *K*-mediods clustering so that the nodes in the graph are divided into *K* clusters according to the distance. The nodes in each cluster use their central node as agent node. When querying the shortest distance between two points, the point in different clusters uses the proxy node to query distance. The experimental results show that the algorithm can effectively reduce the query time and improve the query efficiency, and the deviation rate of the query result is acceptable.

Key words approximate shortest distance; *K*-mediods clustering; graph compression