文章编号：1004-4353(2019)01-0058-06

面向大数据的非结构化数据安全保障技术研究

陈志辉， 吴敏敏
（莆田学院 信息工程学院，福建 莆田 351100）

摘要：针对大数据的非结构化数据安全问题，基于数据类型和数据敏感度级别，提出了一种保障非结构化数据安全的方法. 首先，通过数据分析获取所需数据类型和敏感度级别，并构建数据库的数据节点. 其次，为数据节点提供安全算法，这些安全算法与数据节点交互形成安全套件. 再次，通过接口的调度算法为非结构化数据提供足够的安全性，以及降低系统的开销和提高访问效率. 最后，通过实验表明该方法在能够充分地保障大数据安全的前提下，系统的时间开销不超过传统方法的 52.85%.

关键词：非结构化数据；数据安全；数据分析；隐私保护；访问控制；安全套件
中图分类号：TP309 **文献标识码**：A

Research on security technology for big-data-oriented unstructured data

CHEN Zhihui， WU Minmin
（*College of Information Engineering*, *Putian University*, *Putian* 351100, *China*）

Abstract: Aiming at the problem of unstructured data security, a method to guarantee unstructured data security is proposed by considering types of data and their sensitivity levels. Firstly, types of data and their sensitivity levels are obtained through different analysis methods of data, and then the data nodes of the database are constructed. Secondly, security algorithms are designed for data nodes. These security algorithms interact with data nodes to form security suites. Thirdly, the security suite provides enough security for unstructured datathrough the scheduling algorithm of the interface, as well as reducing the overhead of the system and improving access efficiency. Finally, the experiment shows that this method can guarantee the security of large data fully, and the time cost of the system does not exceed 52.85% of the traditional processing time.

Keywords: unstructured data; data security; data analysis; privacy protection; access control; security suites

随着社会信息化和网络化的高速发展，大数据已成为目前继云计算之后信息技术领域的另一个信息产业增长点. 但随着大数据应用的不断扩展，其安全性问题也逐渐引起了人们的重视. 对此，许多学者对大数据的隐私管理、访问机制和数据加密等安全问题进行了研究. 例如，W. Itani 等提出了有关保护用户隐私的协议[1]，S. Creese 等提出了企业云部署中的隐私安全管理机制[2]，A. Parakh 等提出了共享隐式机制[3]，但是这 3 种方案都是基于加密机制的数据隐私性保护方案，缺乏动态的安全策略，不适合于大量非结构化数据的隐私保护. 学者们除了对数据的隐私和访问机制进行研究外，还对数据存储的加密算法进行了研究，如加法同态的 Paillier 算法[4]、加法和乘法的同态 IHC 和 MRS 算法[5]以及 C. Gentry 的理想格的加密算法[6]等. 这 3 种算法虽然都能满足聚合与计算的安全需求[7]，但无法对组合的数据集提供不同等级的安全保障，若直接将算法运用于大数据上，其效率必然低. 为此，本文提出一种基于熵值法赋权的非结构化数据敏感化模型，动

收稿日期：2019 - 01 - 13 **作者简介**：陈志辉(1977—)，男，副教授，研究方向为数据安全、软件工程等.
基金项目：福建省教育厅中青年教师教育科研项目(JK180480)；莆田市科技项目(2015G2011)

态地调度安全套件内的算法来保障非结构化的数据安全，并通过实验验证本文方法的有效性.

1　技术框架

本文的安全框架主要包括两个功能：一是应用数据分析（包括数据过滤、集群和分类）动态地识别出数据的类型和敏感度，该功能有助于为数据库构建数据节点. 数据库节点包含不同类型的数据，例如文本、XML、电子邮件、图像、视频和音频等. 二是将现有的安全服务标准或算法集成优化，构建安全套件，通过接口的调度保护数据的隐私性、完整性以及不可抵赖性. 本文的安全框架图如图 1 所示.

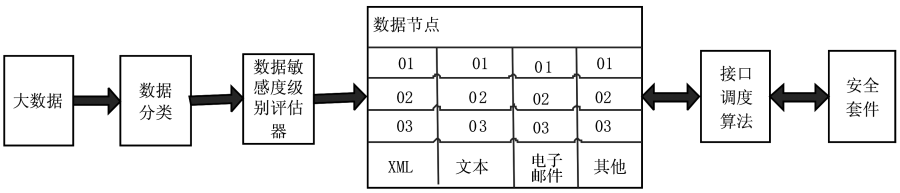

图 1　大数据安全框架

1.1　数据敏感度级别

目前，非结构化数据普遍采用 KerBeros 协议，并结合访问控制表对用户进行授权访问[7]. 该访问模式的数据敏感性只能通过人工识别后标识，因此文件只能是静态的访问控制方式，容易导致数据的泄露. 因此，有必要设计一个数据敏感度级别评估器，自动地识别出数据的敏感度级别. 本文设计的数据敏感度级别识别模型如图 2 所示. 图 2 中，U_i 表示用户；D_i 表示数据集；R_i 表示数据项的集合；$R_{i,j}$ 表示集合 R_i 指向一个具体数据项 j 的边；C_{U_i,D_j} 表示用户 U_i 对数据集 D_i 的访问次数；$d_{R_{i,j},R_{m,n}}$ 表示数据集 R_i 中的一个具体数据项 j 指向集合 R_m 中的一个具体数据项 n 的边.

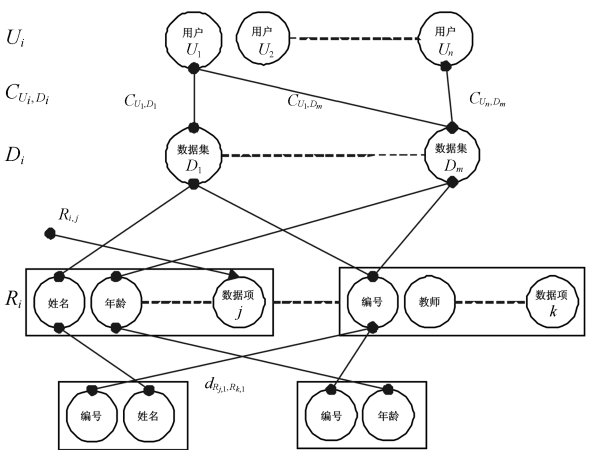

图 2　数据敏感度级别识别模型

信息熵可以解决对数据敏感度级别的量化度量问题，即用熵的变化来识别数据敏感度.

定义 1　某一数据集的敏感度由全部数据集的熵 $\sum_{i=1}^{n}H(x_i)$ 减去该数据集的熵 $H(x_i)$ 后所得值表示. 数据集 i 的敏感度 $C(x_i)$ 表示为

$$C(x_i)=\sum_{i=1}^{n}H(x_i)-H(x_i), \tag{1}$$

信息熵表示为

$$H(x)=-\sum_{i=1}^{n}p(x_i)\log x_i, \tag{2}$$

其中 $p(x_i)$ 表示数据集 i 的概率函数，它主要受数据使用率、连接度和数据质量 3 个因素的影响，即：数据项使用频率越高，敏感度越高；数据项连接越多，数据敏感度越高；数据的损坏或缺失越大，数据敏感度越高.

定义 2　数据使用率函数

$$p(x'_i)=\sum_{i=1}^{m}C_{U_i,D_j}\Big/\sum_{j=1}^{k}\sum_{i=1}^{m}C_{U_i,D_j}, \tag{3}$$

式中 $\sum_{j=1}^{k}\sum_{i=1}^{m}C_{U_i,D_j}$ 和 $\sum_{i=1}^{m}C_{U_i,D_j}$ 分别表示所有数据集和数据集 D_i 的访问次数.

定义 3　连接度概率函数

$$p(x''_i)=\sum_{j,m,n=1,1,1}^{j,m,n=a,b,c}d_{R_{i,j},R_{m,n}}\Big/\sum_{i,j,m,n=1,1,1,1}^{i,j,m,n=d,a,b,c}d_{R_{i,j},R_{m,n}}, \tag{4}$$

式中 $\sum_{i,j,m,n=1,1,1,1}^{i,j,m,n=d,a,b,c}d_{R_{i,j},R_{m,n}}$ 和 $\sum_{j,m,n=1,1,1}^{j,m,n=a,b,c}d_{R_{i,j},R_{m,n}}$ 分别表示连接点所有的边数和连接 R_i 点的边数.

定义 4　数据质量函数

$$p(x'''_i)=\Big(\sum_{S_r}\mathrm{co}(R_j)/S_r\Big)\Big/\sum_{i=1}^{n}\Big(\sum_{S_i}\mathrm{co}(R_i)/S_i\Big), \tag{5}$$

式中 $\mathrm{co}(R_i)$ 表示正确数据项的数目，S_r 表示数据项总数目，$\sum_{S_r}\mathrm{co}(R_j)/S_r$ 表示数据集 x'''_i 中数据项的正确率；$\sum_{i=1}^{n}\left(\sum_{S_i}\mathrm{co}(R_i)/S_i\right)$ 表示全部数据集的正确率.

定义 5 组合熵为数据使用率 $H(x'_i)$、连接度 $H(x''_i)$ 和数据质量 $H(x'''_i)$ 的熵之积，即：

$$H(x_i)=H(x'_i)\cdot H(x''_i)\cdot H(x'''_i). \tag{6}$$

将式(3)—(5)分别代入式(2)中，得：

$$H(x'_i)=-\sum_{i=1}^{n}p(x'_i)\log x'_i, \tag{7}$$

$$H(x''_i)=-\sum_{i=1}^{n}p(x''_i)\log x''_i, \tag{8}$$

$$H(x'''_i)=-\sum_{i=1}^{n}p(x'''_i)\log x'''_i. \tag{9}$$

将式(7)—(9)分别代入式(6)中，得：

$$H(x_i)=-\sum_{i=1}^{n}p(x'_i)\log x'_i\cdot \sum_{i=1}^{n}p(x''_i)\log x''_i\cdot \sum_{i=1}^{n}p(x'''_i)\log x'''_i. \tag{10}$$

将式(10)代入式(1)，计算所得的取值范围为 0～1. 本文为数据集设计一个基于代码的场景. 所有数据被定义为 3 个级别，其中敏感度 $C(x_i)$ 趋近 0 是受敏感协议保护的数据集，设定其敏感度级别代码为“01”，这类数据必须实施最强的安全标准或算法，以确保最高级别的安全性；敏感度 $C(x_i)$ 趋近 0.5 的数据集，设定其敏感度级别代码为“02”，这类数据需要处理速度快的安全算法，但其安全标准可低于“01”标准；敏感度 $C(x_i)$ 趋近 1 是对访问和存储数据的环境风险控制要求较低的数据集，设定其敏感度级别代码为“03”，这类数据仅使用身份验证即可访问.

1.2 安全套件

安全套件是数据文件安全防护的重要模块之一，它包含访问控制、加密和签名等功能组件[8]. 本文中的安全套件是指在现有安全标准或算法的基础上，按照不同的安全需求变换组合后形成的安全算法库. 基于文本 TX 的安全套件(见表 1)，本文从 3 个方面考虑安全问题：一是隐私性(CS). 隐私性首选的数据加密标准为 DES(Data Encryption Standard)，3DES(3 次 DES 算法)用于敏感度级别更高的数据集. 二是完整性(HF). 散列算法 Snefu-256 和 Tiger 都能保证数据的完整性，但蛮力攻击测试结果表明 Snefu-256 算法更适合于敏感度级别高的数据集. 三是真实性(MC). AES-CCM 和 HMAC-SHA1 算法都能保护数据的真实性，但由于 AES-CCM 是加密算法和认证算法的混合模式，执行时间相对长，因此更适合于敏感度级别高的数据集. 当用户访问或存储数据时，系统依据代码调度安全套件中的相应算法，并激活适合的安全服务. 例如，对于代码为 TXCS01，系统选取 3DES 算法为数据提供加密服务.

表 1 文本 TX 的安全套件

数据代码	安全访问代码	敏感度级别代码	算法
TX	CS	01	3DES
		02	DSDES
		03	UD
	HF	01	Snefu-256
		02	Tiger
		03	UD
	MC	01	AES-CCM
		02	HMAC-SHA1
		03	UD

基于 XML 文档的安全套件，本文主要考虑加密、数字签名、身份验证和访问控制等服务，这些服务的代码分别定义为 EC、DS、AC 和 AP. XML Enc 用于对 XML 文档实施加密，维护 XML 文档的隐私性；XML_DSig 用于提供 XML 文档数字签名，确保消息的不可抵赖性等；安全断言标记语言 SAML 用于提供身份验证和信息授权等；可扩展访问控制标记语言 XACML 用于访问控制策略. 部分 XML 文档安全套件如表 2 所示，表中 XX 表示通用服务代码. 因电子邮件、图像、视频和音频等数据的安全套件创建办法与文本 TX 和 XML 文档类似，故在此省略.

表 2 XML 文档的安全套件

数据代码	安全访问代码	敏感度级别代码	算法
XM	EC	XX	XMLEnc
	DS	XX	XML_DSig
	AC	XX	SAML
	AP	XX	XACML

1.3 性能分析

为了分析系统的性能，本文构建一个评估函数对数据安全性能进行评估.

定义 6　某一数据集的系统开销（处理时间）由该数据集内不同敏感度级别的安全性权值与概率的各乘积之和来表示，即：

$$O(S)=\sum_{k=1}^{3}V_kP_k. \tag{11}$$

其中：$O(S)$ 表示开销（处理时间）的函数，若 $O(S)=1$，则该套件将承担所需的全部开销；S 表示不同数据类型的安全套件. V_k 表示敏感度级别为 k 的数据其安全性所需的值. 对于敏感度级别“01”的数据，使用安全度最高的服务，设定 $V_1=1$；对于敏感度级别“02”的数据，提供必要的安全性，设定 $V_2=0.6$；对于敏感度级别“03”的公共数据，则设定 $V_3=0.1$. P_k 表示敏感度级别为 k 的数据的概率，$k=1,2,3$.

为了研究数据敏感度问题，对某一地区机构组织的数据进行分析，结果见表 3. 从表 3 数据可知，教育机构的敏感性相对最低. 若单独对教育机构提供安全套件的安全保障，可为其节省 59.5% 的系统开销.

教育机构 $O(S)$ 的计算方法如下：

$$O(S)=V_1P_1+V_2P_2+V_3P_3=1\times0.2+0.6\times0.25+0.1\times0.55=0.405.$$

若为教育、医疗机构和研究所等 6 个机构组织提供安全套件的安全保障，则可为其节省 44% 的系统开销. 由此可知，本文提出的方法不仅具有安全性，还能显著提升系统的性能.

表 3　某地区不同机构的数据敏感度级别

机构组织	各敏感度级别数据所占百分比/%			开销率/%	节省率/%
	代码“01”	代码“02”	代码“03”		
教育	20	25	55	0.405	59.5
医疗	65	20	15	0.785	21.5
研究所	25	70	5	0.675	32.5
房地产	55	25	20	0.720	28.0
软件公司 developer	40	45	15	0.685	31.5
金融/银行	55	25	20	0.725	27.5
平均	43.34	35	21.66	0.660	44.0

2　实验结果与分析

使用 Java 编程语言支持的包访问不同的数据源，并将数据存储到 MongoDB 数据库中，然后利用实验验证本文方法的有效性.

2.1　数据源与检索

实验数据源选取于维基百科数据和百度，因为二者都有单独的数据检索，可检索到大量的图片和文本等数据. 首先，通过 Java 支持的包对维基百科和百度数据源进行数据检索，并确定文件的数据类型；其次，通过数据敏感度级别评估器对检索到的数据集进行评估，并生成对应的数据敏感度级别代码；最后，利用 Java 程序将每个数据文件（包括数据类型和敏感度级别）存储到 MongoDB 数据节点中.

2.2　实验结果与分析

为了证明系统的性能，使用 2 组数据集（数据均从 MongoDB 中读取）进行实验. 其中一组读取原始数据集，以最高安全 3DS 算法、敏感度级别为“01”和“02”的安全标准对其进行实验，执行时间如表 4—表 6 所示. 另一组读取含有数据类型和敏感度级别的数据集，应用安全套件对其进行实验，执行时间如表 7 所示. 从表 4—表 7 中的数据可知，应用安全套件的执行时间均低于上述各级别算法的执行时间.

表 4　应用 3DES 算法的执行时间

代码	算法	执行时间/ms
TXCSXX	3DES	234
TXCSXX	3DES	239
TXCSXX	3DES	254
TXHFXX	3DES	3 781
TXHFXX	3DES	3 512
TXHFXX	3DES	4 230
TXMCXX	3DES	4 289
TXMCXX	3DES	3 566
TXMCXX	3DES	3 478
XMEXXX	3DES	1 438
XMEXXX	3DES	1 322
XMEXXX	3DES	1 305
XMEXXX	3DES	1 293
总计		28 941

表 5 应用敏感度级别“01”算法的执行时间

代码	算法	执行时间/ms
TXCS01	3DES	290
TXCS01	3DES	255
TXCS01	3DES	220
TXHF01	Snefu-256	3 900
TXHF01	Snefu-256	3 500
TXHF01	Snefu-256	3 751
TXMC01	CCM	3 000
TXMC01	CCM	3 312
TXMC01	CCM	2 985
XMEX01	XML Enc	1 312
XMEX01	XML Enc	875
XMEX01	XML Enc	1 110
XMEX01	XML Enc	950
总计		25 460

表 6 应用敏感度级别“02”算法的执行时间

代码	算法	执行时间/ms
TXCS02	DES	41
TXCS02	DES	36
TXCS02	DES	32
TXHF02	Tiger	4 119
TXHF02	Tiger	3 100
TXHF02	Tiger	3 075
TXMC02	HMAC-Sha1	95
TXMC02	HMAC-Sha1	89
TXMC02	HMAC-Sha1	91
XMEX02	XML Dsig	900
XMEX02	XML Dsig	855
XMEX02	XML Dsig	796
XMEX02	XML Dsig	812
总计		14 041

表 7 应用安全套件的执行时间

代码	算法	执行时间/ms
TXCS01	3DES	252.5
TXCS02	DES	32.5
TXCS03	None	0
TXHF01	Snefu-256	3 654.0
TXHF02	Tiger	3 042.0
TXHF03	None	0
TXMC01	AES-CCM	3 149.0
TXMC02	HMACSha1	88.0
TXMC03	None	0
XMEXXX	XML Enc	1 303.5
XMEXXX	XML Dsig	667.0
XMEXXX	SAML	639.0
XMEXXX	XACML	628.0
总计		13 455.5

各算法处理数据集的时间开销如图 3 所示. 从图 3 中可以看出,应用安全套件处理数据的时间开销均小于应用其他安全算法处理数据的时间开销. 这是因为安全套件是动态的安全标准,它根据数据的敏感度选取最适宜的安全标准,因此处理速度较快. 图 4 为应用安全套件的时间开销与其他算法的时间开销的百分比. 从图 4 中可以看出,若为其提供安全级别高或较高的算法,如 3DS 算法和敏感度级别“01”算法,其处理时间大幅增加;若为其提供安全级别低的算法,如敏感度级别“02”算法,虽然处理时间会相应减少,但其数据安全性较低. 而应用安全套件算法,不仅能保证数据的安全性,而且还能明显降低时间开销(仅占 3DS 算法时间开销的 46.49%和敏感度级别“01”算法时间开销的 52.85%).

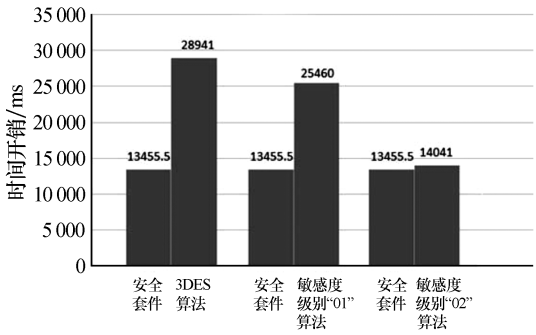

图 3 各算法的时间开销

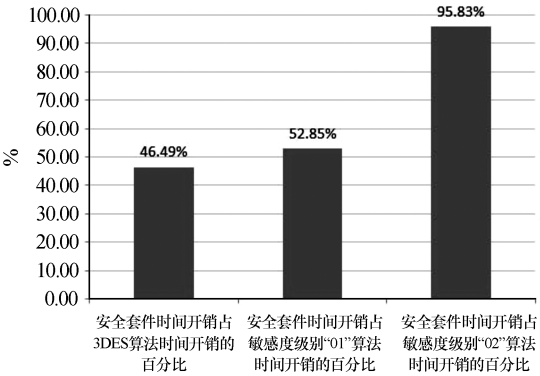

图 4 安全套件算法的时间开销与其他算法的时间开销的百分比

3 结论

实验表明,本文提出的基于熵值法赋权的非结构化数据敏感化模型,在能够充分保障大数据安全的前提下,其系统的时间开销不超过传统方法的 52.85%,因此本文模型有助于用户安全、快

速地访问非结构化数据.在研究中，本文仅对非结构化数据进行了动态访问策略的研究，而对于有实时性要求的场景未能进行研究，因此今后我们将考虑基于时间自动机的实时系统应用的研究.

参考文献：

[1] ITANI W，KAYSSI A，CHEHAB A. Privacy as a Service：Privacy-Aware Data Storage and Processing in Cloud Computing Architectures[C]//The 8th IEEE International Symposium on Dependable，Autonomic and Secure Computing. Washington DC，USA：IEEE Computer Society，2009：711-716.

[2] CREESE S，HOPKINIS P，PEARSON S. Data Protection-Aware Design for Cloud Services[M]. Germany：Springer Berlin Heidelberg，2014.

[3] PARAKH A，KAK S. Space efficient secret sharing for implicit data security[J]. Information Science，2011，181(2)：335-341.

[4] CATELANO D. Paillier's Cryptosystem Revisited [C]//In Proceedings of the 8th ACM conference on Computer and Communications Security. Philadelphia，PA，USA：Association for Computing Machinery，2001：206-214.

[5] BENDLIN R，DAMGARD I. Semi-Homomorphic Encryption and Multiparty Computation[M]. Mrmany：Springer Berlin Heidelberg，2011：302-310.

[6] GENTRY C. A Fully Homomorphic Encryption Scheme[D]. Virginia：Standford University，2009：120-131.

[7] 王杰，陈志刚，钱漫匀，等. 面向云隐私保护的 5A 问责制协议设计[J]. 南京邮电大学学报(自然科学版)，2018，38(6)：68-76.

[8] 刘莎，谭良. Hadoop 云平台中基于信任的访问控制模型[J]. 计算机科学，2014，41(5)：155-163.

[9] 张敬伦，张永生，高丽琴. 基于内网数据安全防护引擎的安全架构设计[J]. 通信技术，2017，50(1)：158-161.

(上接第 14 页)

参考文献：

[1] ADAMS J F. On the structure and application of the Steenrod algebra[J]. Math Helv，1958，32：180-247.

[2] ZHENG Qibing. Graphs and the (co)homology of Lie algebras[EB/OL]. (2011-07-01)[2012-12-11]. http：arxiv.org/abs/1107.0235.

[3] MAY J P. The cohomology of restricted Lie algebras and of Hopf algebras[J]. Journal of Algebra，1966(3)：123-145.

[4] 陆俊杰. 关于 May 谱序列 E_2 项的若干结果[D]. 天津：南开大学，2007.

[5] DWYER W G. Homology of integral upper-triangular matrices[J]. Proceedings of the American Mathematical Society，1985，94：523-528.

[6] 高姗. May 谱序列某些直和项上同调的计算[J]. 南开大学学报(自然科学版)，2013，46(5)：29-36.