

基于关联规则挖掘的口令字典生成技术

陈君山,刘京菊,李振汉
(国防科技大学,安徽 合肥 230031)

摘 要: 为提高口令破解效率,在获取个人信息后构造针对性强的口令破解字典,采用关联规则挖掘的方法获取口令设置规则. 针对网络上搜集获取的具有详细用户信息的数据进行拆分,使用 Apriori 算法挖掘拆分结果得到用户的口令设置规则,根据挖掘得到的规则针对测试集用户信息生成相应的口令字典,测试结果表明了该方法的有效性.

关键词: 口令破解,关联规则,apriori 算法,用户数据,属性拆分

中图分类号: TP309

文献标识码: A

文章编号: 1000—7180(2018)12—0110—05

Research on Password Dictionary Generation Based on Association Rule Mining

CHEN Jun-shan, LIU Jing-ju, LI Zhen-han

(National University of Defense Technology, Hefei 230031, China)

Abstract: To improve password cracking efficiency, construct targeted password cracking dictionary when getting personal information, the method of association rule mining is adopted to obtain the password setting rules. Split the data collected on the network with detailed user information, use the apriori algorithm to mini the split result to get the user password setting rules. Generate the corresponding password dictionary for the test set user information according to the rules obtained by mining, the test results show the effectiveness of the method.

Key words: password cracking; association rules; apriori algorithm; user data; property split

1 引言

互联网用户数据一直以来都是各个互联网公司重点防护的对象,但由于其背后强大的利益驱动,用户数据经常成为攻击者的目标,因此仍有较多的互联网用户数据安全事件被频频曝出. 大量用户数据都是以明文的方式存储的,因此通过对这些信息的搜集整理和关联分析得到较多潜在的有用信息. 作为一种方便、有效的认证方式,口令认证被广泛使用,虽然登陆令牌、生物特征识别等认证方式逐渐开始应用,但由于方便性、软硬件支持等方面的原因,口令认证在较长一段时间内,仍将是主流的认证方式. 一个安全性高的口令应当足够长、足够复杂、没有规律、难以猜测破解. 然而在现实生活中,人们为了便于记忆,通常将姓名、生日、手机号等个人信息

直接作为口令,或者将这些信息进行简单的排列组合作为口令,也有用户直接使用由纯数字或纯字符组成的弱口令作口令. 因此通过挖掘用户个人信息与口令密码设置之间的潜在关联关系,从而得到口令设置的规则习惯对于生成针对性较强的口令字典,从而进行口令安全性测试或口令恢复都有着重要意义和研究价值.

目前网络上提供大量口令字典生成工具,如“真空密码生成器”、“骨灰字典”等,这些工具通过枚举的方式根据输入的日期、字符集、身份证、拼音、电话、英文、前缀后缀等信息生成包含所有符合规则要求的口令字典. 这些口令字典生成工具生成的口令虽然与用户信息相关,但会枚举所有结果,使得字典容量很大且针对性不强,直接导致口令破解效率较低. 文献[1]基于数据挖掘技术和口令因子的概念对

原始用户口令进行数据挖掘,得到了相应的关联规则,将关联规则应用在口令字典生成中,并在 WinRAR 文件破解中得到了很好的效果. 文献[2]基于密码的可记忆性和统计特性提出了基于语义的概率字典破解,通过统计密码训练集的各种语义组合生成概率规则,以此生成猜测字典. 文献[3]提出结合标准马尔可夫模型与词剩余空间的有效枚举算法,以减少搜索空间实现时间空间的平衡,并在真实口令哈希值数据库上对该方法进行了评估,在 2×10^9 的空间中成功恢复了 67.6% 的密码. 当前的口令字典生成方式还存在一些不足:采用穷举方式生成的字典容量大,针对性不强;利用个人信息生成的字典涵盖内容不全面.

针对现有的字典生成方式的不足,本文提出一种新的口令字典生成方法,通过对搜集到的包含相对完整用户信息的数据进行属性拆分,对拆分结果与相应口令匹配得到每名用户的口令属性向量,利用关联规则挖掘的方法得到口令设置规则,并由规则生成口令字典,验证规则的有效性. 利用此方法生成的口令字典能有效解决字典针对性不强的问题,且涵盖所有常用信息,能在较小的字典容量下包含较为完整的用户口令,并且能够根据新增信息灵活添加.

2 基于关联规则挖掘生成口令字典流程

基于关联规则挖掘生成口令字典流程包括数据准备、数据预处理、关联规则挖掘、口令字典生成及验证等,整体流程如图 1 所示.

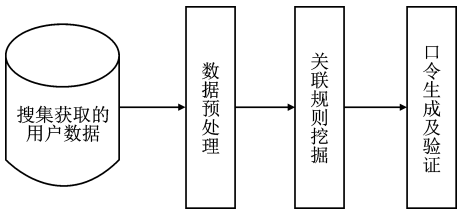


图 1 基于关联规则生成口令字典流程

2.1 数据准备

本文研究使用从网上搜集到的某网站上泄露的包含完整用户信息的 131 020 条数据,包括用户名、姓名、邮箱、手机号、身份证号、口令等,表 1 为完整数据包含内容.

2.2 数据预处理

直接获取到的用户信息无法直接挖掘获取信息与口令间的关联,需要将数据进行拆分,并将拆分结果与口令进行匹配来获得每条数据的口令

向量.

表 1 完整数据包含内容

字段	描述
email	邮箱
username	用户名
name	姓名(中文)
id	身份证号(18 位)
tel	电话号码(11 位)
password	口令

2.2.1 用户信息属性拆分

用户信息的属性拆分是从原始数据中分析获得量化数据的关键步骤,为之后的规则挖掘提供数据基础. 用户信息属性拆分的定义如下:输入属性集合 $A = \{a_1, a_2, \dots, a_n\}$ 是已经获取到的各项用户信息,包含姓名、用户名、生日、邮箱、身份证号等;输入的口令集合 $P = \{p_1, p_2, \dots, p_n\}$ 为 a_i 所对应的将要分析其组成的口令. 对于 P 中的每一个口令 p_i ,分析其包含的属性,得到 p_i 拥有的属性集合 $pA_i = \{a_{i1}, a_{i2}, \dots, a_{ij}\}$,其中属性 $a_{ij} \in A$. 口令属性拆分的结果为每条口令所对应属性集合,即 $PASet = \{(p_1, pA_1), (p_2, pA_2), \dots, (p_n, pA_n)\}$.

用户设置的与个人信息相关的口令,通常是多条个人信息的组合,最终获取的口令模式也将是个人信息属性的组合,如:姓首字母+名首字母+生日,因此需要将完整的个人属性按照需求进行拆分.

2.2.2 建立匹配

根据拆分结果,与对应的口令进行匹配,将口令中存在的项进行标记,形成表示该条口令组成的向量. 对获取到的用户信息进行拆分,将拆分得到的内容与对应的口令进行比对,当口令中存在该项时,将对应的项置 1,不存在则置 0,例如某条口令由姓首字母+名首字母+生日构成,则其对应的向量为 $(0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$,数据拆分项与口令匹配结果如图 2 所示,由于属性较多,图中仅展示部分属性与口令的匹配结果.

3 关联规则及 apriori 算法

关联规则是形如 $X \Rightarrow Y$ 的蕴含式,其中 X 和 Y 分别称为关联规则的先导和后继,关联规则存在支持度和置信度,关联规则 $X \Rightarrow Y$ 的支持度 s 是数据库中包含 $\text{support}(X \Rightarrow Y)$ 的事务占全部事务的百分比,它是概率 $P(X \cup Y)$,记作 $\text{support}(X \Rightarrow Y) =$

$P(X \cup Y)$. 关联规则 $X \Rightarrow Y$ 的置信度 c 是包含 $X \cup Y$ 的事务数与包含 X 的事务数的比值,它是条件

概率 $P(Y | X)$, 记作 $\text{confidence}(X \Rightarrow Y) = P(Y | X)$ [4].

email	ming_szm	xing_lower	xing_upper	ming_lower	ming_upper	xszm_lower	mszm_lower
254829462@qq.com	0	0	0	0	0	1	1
27229431@qq.com	0	1	0	1	0	0	0
411625357@qq.com	0	0	0	0	0	1	1
chengweipeng123@163.com	0	1	0	1	0	0	0
gxno@163.com	0	0	0	0	0	0	1
steven_jiang@yeah.net	0	1	0	0	0	0	0
ismezy@163.com	0	0	0	0	0	1	1
188232673@qq.com	0	0	0	0	0	0	0
stonedy@yahoo.cn	0	0	0	0	0	0	0
yiguang123abcd@163.com	0	0	0	0	0	0	0
407160239@qq.com	0	0	0	0	0	0	0
119444785@qq.com	0	0	0	0	0	1	1

图 2 数据拆分项与口令匹配结果

Apriori 算法基于先验知识,使用逐层搜索的迭代方法. 主要的一句有两个先验原理: 如果 X 是频繁项集,则它的所有非空子集都是频繁项集;如果 X

不是频繁项集,则它的所有超集都是非频繁项集. 频繁项集生成的主要步骤为[5]:

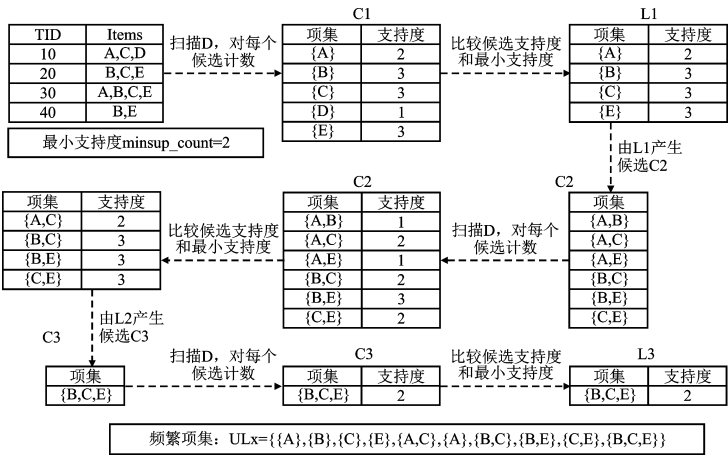


图 3 频繁项集生成过程

- (1)扫描数据库统计每个项的支持度,收集满足最小支持度的项以形成频繁 1—项集 L_1 ;
- (2)利用 L_2 产生频繁 2—项集 L_2 ,再利用 L_2 产生 L_3 ,如此下去,直到不能再找到频繁 k -项集为止. 频繁项集生成过程如图 3 所示:
- 在得到所有频繁项目集后,按以下步骤生成关联规则:
- (1)对每一个频繁项目集 L 生成其所有非空子集;
- (2)对 L 的每一个非空子集 x 计算其置信度,若置信度不小于最小置信度,则规则 $x \rightarrow (L - x)$ 成立. 规则产生过程示意如图 4 所示(以最小置信度为 75%为例).

根据已经获取到的口令组成向量,使用 apriori 算法来预测属性间的关联关系,在用户使用某一个人信息时,哪些信息会被一起使用,从而得到属性间潜在的关联关系,获取用户的口令设置规则. 在关联关系挖掘过程中,所有的口令属性集合作为挖掘的资料库,属性集合 A 对应于所有项集,每条口令的属性 PA_i 对应于一条记录,将 Apriori 直接应用于 PASET 来获取口令属性 a_i 之间的关联关系,输出的挖掘结果作为口令设置规则. 由于支持度反映属性在总体中的占比,置信度反映属性间的关联程度,可以通过支持度和置信度来决定输出规模的大小,以此来改变生成口令规则的规模大小. 挖掘得到的口令设置规则见图 5.



图 4 规则产生过程

```
frozenset(['mm', 'ming_lower']),frozenset(['dd', 'xing_lower']),0.636695018226
frozenset(['xszm_lower', 'dd', 'email_a']),frozenset(['mszm_lower', 'email_d']),0.629007633588
frozenset(['mm', 'xszm_lower', 'email_a']),frozenset(['mszm_lower', 'email_d']),0.618131868132
frozenset(['xszm_lower']),frozenset(['mszm_lower']),0.617508282301
frozenset(['mm', 'mszm_lower', 'email_a']),frozenset(['xszm_lower', 'dd']),0.609
frozenset(['xing_lower', 'm']),frozenset(['dd', 'ming_lower']),0.608582574772
frozenset(['mm']),frozenset(['dd']),0.602614828832
frozenset(['xing_lower']),frozenset(['ming_lower']),0.602519813046
frozenset(['mm', 'mszm_lower', 'username_a']),frozenset(['xszm_lower', 'dd']),0.597461468722
frozenset(['mm', 'mszm_lower']),frozenset(['xszm_lower', 'dd']),0.587903225806
```

图 5 挖掘得到的口令设置规则

4 口令字典生成与验证

实验的基础为收集到的 131 020 条带有完整用户信息的数据,选取其中的一部分作为训练集,利用算法挖掘得到口令设置规则,剩余部分作为测试集,利用挖掘得到的规则生成口令字典来验证规则的有效性. 选用不同数量训练集与测试集的实验结果如表 2.

表 2 不同数量训练集与测试集的实验结果

训练数 据量	测试数 据量	生成规 则数量	破解口 令数量	破解率(%)
111 020	20 000	894	5 032	25.16
121 020	10 000	2 921	3 539	35.39

从实验结果来看,挖掘得到的规则随训练集的增加而增加,相应生成口令字典的规模也会随之增大,训练集为 111 020 条数据时,破解率为 25.13%,训练集为 121 020 条数据时,破解率为 35.39%,因而在获取到更多数据时能够得到更加完整的口令设置规则,进一步提升口令破解的成功率.

5 结束语

本文根据真实的用户信息拆分得到的属性信息,采用关联规则挖掘的方法,分析属性与口令设置之间的潜在关联,得到口令设置规则,通过规则生成的口令字典在测试集上有较好的覆盖,说明了该方法的有效性. 用户的口令设置不仅与本人信息相关,而且与配偶、子女等信息有较大的关联,考虑到规则的完整性以及生成口令字典的有效性,在后期的研究中加入更多的相关信息及其对应的属性信息,以提高规则的覆盖率. 同时也表明了,在安全的前提下,用户在设置口令时,应避免使用与本人相关的信息,以此来提高设置口令的强度.

参考文献:

[1] 刘玮文. 基于语义的概率字典破解技术研究[D]. 武汉:华中科技大学,2013.

[2] 郭奕东,邱卫东,刘伯仲. 基于属性拆分与数据挖掘的真实口令分析[J]. 计算机工程,2014,40(7):254-257.

[3] Weir M, Aggarwal S, De Medeiros B, et al. Password

cracking using probabilistic context-free grammars [C]. Security and Privacy, 2009 30th IEEE Symposium on. IEEE,2009(5): 391-405.

[4] 徐华. 数据挖掘:方法与应用[M]. 北京:清华大学出版社,2014.

[5] 毛国君,段立娟,王实,等. 数据挖掘原理与算法(第二版)[M]. 北京:清华大学出版社,2007.

作者简介:
陈君山 男,(1993-),硕士研究生. 研究方向为网络安全、数据挖掘. E-mail:634073516@qq. com.
刘京菊 女,(1974-),硕士,副教授. 研究方向为网络安全、态势感知.
李振汉 男,(1989-),硕士,讲师. 研究方向为网络安全、数据挖掘.

(上接第 109 页)

由图所示,随着节点数的增加,加速比的增加逐渐缓慢,这是因为节点数的增多同时造成数据的节点网络传输开销的增加. 同时,随着数据量的增加,算法的加速比仍然呈现线性增加的态势,不同大小的数据集均呈现较好的加速比,表明本文算法在大数据集上具有较好的扩展性.

5 结束语

本文以海量数据聚类为背景,改进了传统 Canopy 算法,克服了其随机选择中心点带来的结果不稳定问题,并引入了稳定性判定,增加了信息熵加权,提高了聚类结果的准确性和稳定性,减少了聚类比较次数,同时结合马氏距离度量,并利用 MapReduce 框架设计算法的并行化实现. 实验验证了本文算法具有良好的有效性和扩展性.

参考文献:

[1] Hadian A, Shahrivari S. High performance parallel k-means clustering for disk-resident datasets on multi-core CPUs[J]. The Journal of Supercomputing, 2014, 69(2):845-863.

[2] Wu G, Lin H, Fu E, et al. An Improved K-means Algorithm for Document Clustering[J]. Journal of Shandong University, 2008, 43(1):60-64.

[3] Frey B J, Dueck D. Clustering by passing messages between data points[J]. science, 2007, 315(5814):

972-976.

[4] Niknam T, Amiri B. An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis [J]. Applied Soft Computing, 2010, 10(1): 183-197.

[5] Liu J, Han L, Hou L. Cluster Analysis Based on Particle Swarm Optimization Algorithm [J]. Systems Engineering-theory & Practice, 2005(6): 54-58.

[6] Lee G, Tolia N, Ranganathan P. Computing cluster performance simulation using a genetic algorithm solution; U. S. Patent 9,785,472[P]. 2017-10-10.

[7] Elhoseny M, Yuan X, Yu Z, et al. Balancing energy-consumption in heterogeneous wireless sensor networks . using genetic algorithm[J]. IEEE Communications Letters,2015, 19(12): 2194-2197.

[8] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.

[9] Shang T, Zhao Z, Guan Z, et al. A DP Canopy K-Means Algorithm for Privacy Preservation of Hadoop Platform[C]//International Symposium on Cyberspace Safety and Security. Springer, Cham, 2017: 189-198.

作者简介:
王万良 男,(1957-),博士,教授. 研究方向为智能调度等.
胡 禹(通讯作者) 男,(1992-),硕士研究生. 研究方向为聚类算法.
E-mail:huyuchina@163. com.