粗糙集理论在网络安全管理中的挖掘与应用*

张烈超

（武汉交通职业学院，湖北 武汉 430065）

摘 要：针对网络管理中的弊端及潜在的安全性问题，将粗糙集理论应用于系统日志分析中，在开放数据挖掘平台下实现粗糙集相关算法，并将其应用于网络安全管理历史数据分析中，挖掘出其中隐含的有效知识，并给出该算法与其他算法相融合后在云计算平台下进行较大日志数据挖掘的模型，为优化网络管理提供辅助决策。

关键词：粗糙集；网络管理；数据挖掘；知识提取

DOI：10.3969/j.issn.1672－9846.2015.01.019

中图分类号：TP393.08 文献标志码：A 文章编号：1672－9846(2015)01－0081－05

随着计算机网络、移动通信、移动互联网应用渗透至各行各业，人们所拥有的数据面临着前所未有的爆炸式增长，进入“大数据”时代。网络安全已经成为一个潜在的巨大问题。相关媒体经常报道用户信息被窃取、服务器系统被恶意病毒攻击等，对企业及用户造成了很大的损失。针对这些问题，计算机专家及网络安全管理者采取各种应对措施，如通过增加软硬件防火墙技术、入侵检测技术、上网行为分析技术等手段，但只能部分解决问题。对于网络安全系统日志、上网行为数据中蕴含与安全相关的其他知识却无法获取，数据挖掘相关算法，正是为处理此类数据应运而生相关技术。本文在相关平台下进行了二次开发，利用粗糙集算法处理网络安全信息（如系统日志，上网行为数据等），并将挖掘处理的结果作为网络安全管理策略的依据，进一步改善安全防范措施。在此基础上，给出粗糙集与其他算法相融合及处理“大数据”的思路。

1 网络安全管理中的一般问题及防范措施

一般来说，网络所面临的威胁包括如下几个方面：

（1）由于在设计及编程时缺乏周全考虑等原因，网络软件或系统软件难免留有漏洞，成了黑客攻击的对象。如通过漏洞植入木马病毒，使得计算机网络安全面临巨大的威胁。

（2）防火墙的入口及出口机制配置不当，给恶意攻击者留有机会。同时用户安全防范意识不强，口令选择不慎，或者不及时地更新防护系统，易造成网络安全的威胁。

（3）人为恶意攻击，包括：主动攻击及被动攻击等。

（4）管理不当，造成网络设施无意或恶意的损坏。

针对以上网络安全方面的内容，一般网络安全的防御措施主要包括以下内容：

（1）利用防火墙技术，有效地利用防火墙技术可提高内外网络的正常运行；但该技术无法阻止来自内网的攻击。

（2）网络防病毒技术，当前，根据计算机网络采取的客户机——服务器工作模式，故需要从服

*收稿日期：2015－01－13

基金项目：武汉交通职业学院校级项目“粗糙集在信息安全管理中的挖掘应用研究”（编号：2013y13）。

作者简介：张烈超（1973－），男，湖北武穴人，武汉交通职业学院实践教学中心实验师，主要从事数字化校园优化管理研究。

务器和工作站二者结合来解决防范病毒的问题,一般而言,适时升级客户机及服务器的系统补丁、更新病毒库,建立全方位、多层次的防病毒系统,以确保网络免于病毒的侵袭。

(3)数据加密及虚拟专用网(VPN)技术,虚拟专用网主要为企业用户远程安全访问服务,因而更多地对关注用户数据的安全性,利用隧道技术、加解密技术、密钥管理技术、使用者与设备身份认证技术。采用这些技术能有力地抵制不法分子篡改、偷听、拦截信息的能力,保证数据的机密性。

(4)入侵检测及访问控制技术,首先通过入侵检测技术对网络或者计算机系统中若干关键节点收集如:病毒入侵信息记录、主机运行日志、网络连接流量等信息并加以分析,检测其中是否有违反安全策略的行为;再利用访问技术设置网络安全防范规则及保护策略。

由于网络安全的动态性发展模式及复杂性,对网管安全人员提出了更高的要求,应改变过去那种被动式查漏补缺形式,根据系统安全管理的特点,采取主动防御及事先规划的方式。由于对网络设备及系统的操作均会产生相应的日志文件,并且日志文件随着网络的使用呈海量增加趋势,因此,本文从网络安全入侵检测及访问控制技术入手(对网络操作日志数据进行收集,不定时进行分析),将数据挖掘相关技术粗糙集(Rough Set)应用于网络安全管理中,从数据分析的角度对安全检测对象的行为数据或审计记录进行分析,找出其中蕴藏的规律,并设定置信度(可信度),给出相应的规则,将发现的知识(规则)反馈给网络安全管理者。管理者根据获取的可信(有效)规则对网络或核心设备的安全性策略予以改进或重新规划。

2 粗糙集及知识约简规则提取算法

2.1 粗糙集提取知识的优势与步骤

数据挖掘过程中涉及到的数据类型比较复杂,数据类型不同,挖掘处理的方式也不同,如采用与模糊逻辑、神经网络、云模型、遗传算法等处理方式。粗糙集(Rough Set)是由 Pawlak 提出的理论,能有效分析不确定、不精确、不一致、不完整等各种不完备的信息[1],主要应用于知识约简、知识依赖性及属性重要性估计等方面,且已成为信息分析和数据挖掘的重要方法。需引起注意的是:多数作者使用该理论处理数据时,在其文献中为方便表述,可能采用类数字数据,但实际程序中无法处理该类型数据。因该理论在进行知识表达、决策处理时,参与处理的信息各属性都必须转换成离散值(即非数值型数据)。

与模糊逻辑、神经网络、云模型、遗传算法等不确定性软计算理论相比,粗糙集(Rough Set)优势在于:它具有很强的定性分析能力,不需要预先给定某些特征或者参数,可直接从给定的问题描述集合出发,运用不可分辨关系(Indiscernibility Relation)和不可分辨类理论来确定给定问题的近似域(Approximate Space),从而得到该问题的内在规律。

粗糙集(Rough Set)处理问题信息(问题)主要步骤如下[1-2]:

①针对各种完整或者不完整及众多变量的数据;运用不可分辨关系(Indiscernibility Relation)和不可分辨类理论来来确定数据的精确性或非确性;

②运用粗糙集(Rough Set)知识约简方法对①的内容进一步求核(约简);

③由②得到的约简,合并属性相同的实例,得到简化的模型(决策表);

④由③产生的模型(决策表)产生精确而又易于检查各证实的规则。

2.2 粗糙集知识约简规则提取算法

知识约简是粗糙集理论的核心内容之一。而所有约简的计算是 NP－Hard 问题,因此运用启发信息来简化计算以找出最优或次优约简是常用较好的方法。求约简算法一般使用核作为计算约简的出发点,利用一些启发式规则,按照属性的重要性从大到小逐个加入属性,直到该集合是一个约简为止。以下给出计算最佳归约集或用户定义最小属性集算法[3]。

输入:①信息系统,其中为非空有限论域,C为条件属性集,D为决策属性集,令;②用户定义属性集(UA 可以为空)。

输出:一个最佳归约集或是用户定义最小属性集 REDU。

步骤:

①计算信息表的核心集 $CORE$;$REDU = CORE \cup UA$;

②AR′＝AR－REDU;

③对于每个属性 a∈AR′,计算属性的重要性(即

SGF)，并按由大到小排序；

④While$\gamma(REDU,D)\neq\gamma(AR,D)$DO

{选择属性 $a\in AR'$，使得

$SGF(a,REDU,D)=MAX(SGF(a_i,REDU,D)\ a_i\in AR',(i=1,2,\cdots,m)$（如果有几个属性 $a_i\in AR'$ 具有相同的最大重要性，则选择具有最少复合属性值的属性）$REDU=CORE\cup\{a_i\}$，$AR'=AR-\{a_i\}$，$(i=1,2,\cdots,m)$；

计算新的依赖程度：$\gamma(REDU,D)$，}

End while；

⑤$|REDU|\rightarrow N$；

⑥FOR　i=0　to　N−1　DO

{IF　　$\notin CORE\cup UA$　Then

$REDU=REDU-\{a_i\}$

End if；

计算　$\gamma(REDU)$；

IF　$\gamma(REDU,D)\neq\gamma(AR,D)$ Then

$REDU=REDU\cup\{a_i\}$

End if

End for

}

该算法根据每个属性的重要性排序，每次选择重要性最大的属性加入属性集 REDU，在步骤④的前后选择结束后，属性集 REDU 已包含了一些起重要作用的属性，且不改变原始属性集与决策属性之间的依赖程度。最后的反馈过程是，从属性集 REDU 中按重要性由小到大、一个一个地去掉属性，如果去掉该属性会造成依赖度变化，则恢复该属性，否则剔除该属性。最后剩下的属性集就是最佳归约或用户定义最小属性集。

3　网络安全管理规则的机器实现及决策生成

由于粗糙集(Rough Set)处理数据时具有很强的定性分析能力及不需要预先给定某些特征或者参数等特性，故可直接从给定的问题描述集合出发，但应对所处理的信息系统定义有较为严格的约束。而且现实中收集的数据并不完备，所以首先必对原始数据进行数据预处理，再利用粗糙集(Rough Set)约简算法进行求核处理，最后得出决策规则。本文利用开放源码数据挖掘平台 WEKA[4] 进行二次开发，有效利用平台中的相关功能，将粗糙集算法嵌入到 WEKA 中，再按如下步骤进行相关处理。

3.1　数据采集

对网络设备及系统的操作产生的日志文件大部分为文本类型，包含若干数据项，根据相关数据项的特征建立数据库表字段类型，并建立相应数据表，通过程序及数据库的触发机制定期将日志文件通过上述特征进行分类写入数据库中，为后续数据挖掘处理提供数据源。

3.2　数据预处理

针对信息安全管理中经常出现的安全管理问题——病毒相关信息及特征，从相关的网站收集计算机病毒相关信息(包括名字、时间、病毒扩展名、数据信息、适用系统、类型)，根据数据的相关特征，将其转存于数据库中。

我们借助开放平台 Weka 相应的功能，先对数据采集阶段采集的数据进行预处理，如去过滤、去噪处理等，由于 WEKA 中已经有相关处理缺失数据项的函数，故预处理时，首先将病毒信息数据先去除不必要的信息(如名字等)，由于粗糙集算法无法处理除字符类型之外的数据，故将涉及类数字的时间字段按时间段进行分类：before(1980－1991 年)，middle(1992－1995 年)，current(1995 年至今)，将其作为决策属性；对于病毒发生的系统、类型、及病毒扩展名分别用 sys，type 及 extend 作属性标题，并将其作为条件属性，此外，由于粗糙集算法可以处理包含缺失项的内容，对于本信息数据中缺失的数据项，此处用“?”代替。当然，WEKA 也能处理扩展名为 ARFF、CSV、C4.5 和 BSI 的数据，故可将已处理过的数据信息利用 WEAK 软件转存扩展名为 ARFF 格式的文件。

3.3　粗糙集算法的实现

在 WEKA 平台下利用相关的开源代码进行二次开发并嵌入 WEKA 平台，针对粗糙集理论所涉及的内容，实现相应的约简算法。主要内容包括：缺失属性类处理函数，即利用 Weka 平台内置过滤函数对缺失数据进行处理；正区域 POS()相关函数，即根据粗糙集理集(Rough Set)理论求解可以确定的属性；属性类函数 Attr 函数；衡量相关知识依赖度的函数 Importance()函数，即通过粗糙集理论计算各属性之间的定量依赖程度；计算最佳归约集及最小属性集算法，即根据依赖度函数求解的定量结果，在不改变最终知识决策前提下，保留能决定其他属性的最核心或者最重要的属性，剔除属性中依赖程度值较小的属性；决策规则生成函数，即根据最佳归约集是最小属性集算法得到的核心属性来求解相应的决策规则。

3.4 粗糙集决策规则生成

利用在 Weka 平台中实现的粗糙集(Rough Set)算法对上述经过预处理的数据进行挖掘分析,经过对属性约简处理,得出相关的决策分析规则,如下所示(部分):

RoughSet rules

If extend = EXE ==> Before(5)

If extend = EXE or COM ==> Before(8)

If extend = DOS scrambler ==> Before(1)

If extend = COM ==> Before(2)

If extend = DOS file infector ==> Before(9)

If extend = Boot_sector ==> Before(1)

If extend = Win32_worm ==> Middle(2)

其中,规则集左边信息为条件属性,右边为决策属性。数字代表规则出现的次数,此处引入可信度衡量评判标准。对于出现次数较多的规则,可考虑将其作为可信规则并予以采用;对于出现次数较少的规则,将其剔除。根据可信度等评判标准,可得出如下重要规则:

If extend = EXE ==> Before(5)

If extend = EXE or COM ==> Before(8)

If extend = DOS file infector ==> Before(9)

经过分析,通过该算法处理之后得到的规则与现实中病毒相关信息一致,验证了粗糙集提取规则的有效性。因此,信息安全管理人员可考虑为网络安全要求较高的核心设备或者主机及应用系统设置访问日志、网络连接流量、病毒检测、用户使用行为分析等信息定期收集机制,再利用粗糙集理论进行挖掘处理,找出其中蕴藏的规律,信息安全管理者可根据获取的可信(有效)规则对网络或核心设备的安全性策略予以改进或重新规划以增强安全防范规则。

4 粗糙集挖掘应用扩展

4.1 粗糙集与其他算法的融合应用

综合以上研究,可知粗糙集理论能直接操纵数据来提取知识,特别是概念(对象子集合)可用属性值来表达,为知识发现提供了有力的工具。且提取的知识方便理解。但粗糙集理论在数据挖掘过程中仅适合处理非数字类型的数据,而现实应用中可能包含多种类型的数据,因此,可与适合处理数字型的数据算法结合[5](如聚类算法等),有望取得更优的结论。在后期将对粗糙集理论与其它算法融合作深入研究。

4.2 利用粗糙集处理海量网络安全日志的挖掘思路

由于单机系统的局限性,在所处理的网络安全日志及系统数据量不是超大的情况下,对于适合类型数据使用以上在 Weka 平台下开发的粗糙集算法可取得较好效果及高效的挖掘效率。而对于从核心系统或者设备中收集的海量数据日志文件(大数据,如单个日志文件可能有 1GB 甚至更大)中获取"数据石油"(有价值信息)是当前大数据挖掘技术研究与应用的热门问题。

大数据挖掘的主流方式是基于云计算平台框架模式,即利用互联网、开放的云计算平台(如 Hadoop)框架及并行编程模式 Map/Reduce,通过相应的分配机制将海量数据分发至多台普通计算机组成的集群(Cluster),每台计算机分别运行数据挖掘的算法,将结果集中并进行后续处理,快速得到结果。下面分别给出云计算平台模型[6](如图 1)及在云计算法平台下利用粗糙集数据挖掘约简算法[7]的模型图(如图 2)。

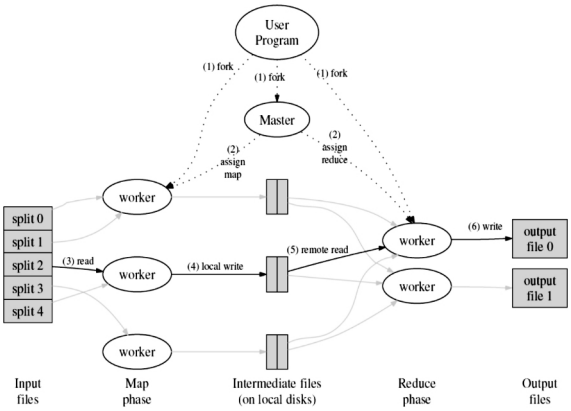

图 1 云计算平台并行计算的示意图

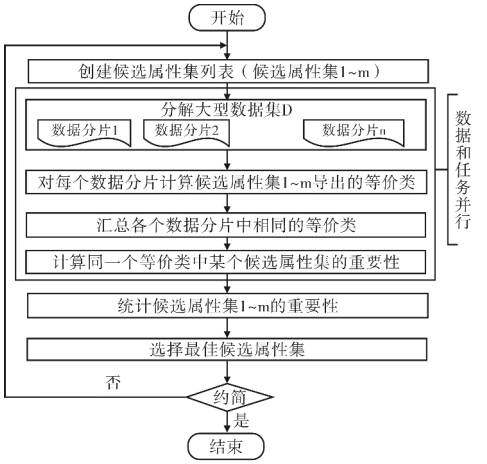

图 2 云计算环境下粗糙集约简算法的流程图

基于粗糙集相关的热点，拟在下一步继续探索不同挖掘算法融合及基于云平台的粗糙集改进，设计实现并行数据挖掘算法，处理大数据集且提高执行效率，将其应用于系统管理中，为信息安全管理提供辅助决策。

参考文献：

[1]李剑，范小军，黄沛. 基于粗糙集的知识理论及其应用[J]. 系统工程理论方法应用，2001，(1)：184－188.

[2]李雄飞，董元方，李军，等. 数据挖掘与知识发现：第2版[M]. 北京：高等教育出版社，2010：44－65.

[3]刘同明. 数据挖掘技术及应用[M]. 北京：国防工业出版社，2001：15－35.

[4]陈慧萍，林莉莉，王建东，等. WEKA 数据挖掘平台及其二次开发[J]. 计算机工程与应用，2008，(19)：76－79.

[5]于达仁，胡清华，鲍文. 融合粗糙集和模糊聚类的连续数据知识发现[J]. 中国电机工程学报，2004，(6)：205－210.

[6]Jeffrey Dean，Sanjay Ghemawat. MapReduce：Simplified Data Processing on Large Clusters[EB/OL]. [2015－01－02]. http://wenku.baidu.com/link?url=S6ENu2gjG5q72dea6s3giLQfnU7aYh3Y_wULHkHZ5Fr5IIbQYXA63O_qddu16tN_q96GKfJV6kEeKmy4Dkl5fGe7DTgNj_X38AZ9HnmtElW.

[7]钱进，苗夺谦，张泽华，等. MapReduce 框架下并行知识约简算法模型研究[J]. 计算机科学与探索，2013，(1)：35－45.

（上接第 64 页）增删还能算是翻译吗？笔者认为答案是肯定的。因为译文版本还是要依托原中文版本，只不过译者发挥了很强的主体性，使得译文有很多译者创作的成分，这种创作虽然显得不忠实于原文，但它能更好地服务于旅游翻译的目的，激发国外游客来中国观光旅游的兴趣。

另外，要做好旅游翻译，从事旅游翻译的工作人员也要加强自身的修养和责任感，面对一篇旅游文本要仔细分析译文读者更需要的是哪些信息，必要时还要做一些文献考察、资料搜集等工作。译者担任着做好对外宣传，让世界了解中国的重任，要有较强的民族自豪感、工作责任感和扎实的专业功底。

参考文献：

[1]陈刚. 旅游翻译与涉外导游[M]. 北京：中国对外翻译出版公司，2004：11－12.

[2]Nida E. A. Towards a Science of Translating[M]. Leiden：E. J. Brill，1964：166－167.

[3]曾诚. 实用汉英翻译教程[M]. 北京：外语教学与研究出版社，2002：44－45.

[4]孟庆升. 新编实用汉英翻译教程[M]. 天津：天津大学出版社，2009：84，93，192.

[5]段连城. 呼吁：请译界同仁都来关心对外宣传[J]. 中国翻译，1990，(5)：2－10.

[6]叶涛. 关于民俗旅游的思考[J]. 东岳论丛，2003，(5)：137－139.

[7][晋]葛洪. 西京杂记：卷一[M]. 北京：中华书局，1985：3－4.

[8]张勃. 从乞巧节到中国情人节——七夕节的当代重构及意义[J]. 文化遗产，2014，(1)：34－40.