文章编号: 1003-5850(2009)01-0001-03

基于概率后缀树的宏观网络报警事件序列分析*

A PST based Alert Event Sequence Analysis Method for Large Scale Network

师鸣若[1] 姜中华[2] 赵明茹[1]

([1] 北京物资学院 北京 101149) ([2] 中国科学院软件研究所信息安全国家重点实验室 北京 100080)

【摘 要】提出一种基于概率后缀树的宏观网络报警事件序列分析框架, 能够分析报警事件序列中存在着某种规律性, 检测出存在大规模宏观网络异常的序列, 通过网络报警事件数据集, 既可以得到网络异常信息, 又能够分析报警事件间的关联。

【关键词】概率后缀树 (PST), 序列分析, 关联分析, 聚类

中图分类号: TP309 文献标识码: A

ABSTRACT Large amount of alert events will be produced for a large scale network. These alert events imply some laws, by which abnormal sequences can be detected. Therefore, we propose a probability suffix tree based alert event sequence analysis method for large scale network. By the method to analyze network alert event sets, we can not only get network abnormal information, but also can analyze the associations among alert events.

KEYWORDS probability suffix tree, sequence analysis, association analyses, clustering

为了分析全国范围内的网络异常和关联报警事件, 在给定网络报警事件数据集的基础上, 我们提出了基于概率后缀树(PST)[3]的网络报警事件序列分析方法。每天网络都会产生大量的报警事件, 这些事件的时序关系和频繁程度在一定程度上反映了当前网络运行情况。我们把某一地区的网络报警事件看作一种网络行为, 并假设从较长的一段时间来看, 异常的行为在整体中只占很小的比例, 由此可构造出正常的行为模型, 并据此来分析网络异常行为。

序列分析[1,2]的关键是定义序列之间的相似度, 相似度可以刻画出序列之间的差别。曾有研究者利用两个序列之间的编辑距离(edit distance), 即根据一个序列需要作多少次插入、删除、替换操作才可以和另一个序列一致, 来衡量相似度。但是这种度量方法只能度量两条序列之间的相似度, 不能衡量一个序列和一个序列集合之间的相似度。也有研究者把序列映射到多维空间的点上, 利用小波分析的方法作研究。还有的通过提取频繁出现的子序列作为词, 利用文本分析的方法构造词-文档(word-document)矩阵来作研究。本文我们利用序列的概率性质来进行相似度研究。

1 PST 概率模型

网络报警事件的数据可以看作是一个按照时间排序的事件序列。定义事件发生的条件概率为在前n个事件已发生的条件下第$n+1$个事件发生的概率。不同的n个事件对应不同的条件概率分布, 并且不同的n也对应不同的条件概率分布。因为正常的网络行为有其固有的规律, 因而从宏观角度分析, 某一地区的条件概率分布也应该是稳定的。

为了描述这种条件概率的稳定性, 我们引入一种称为概率后缀树(PST, probability suffix tree)[3]的模型。一个PST 节点对应一个事件序列。而一个事件序列对应一个d维条件概率矢量, 其中d是事件个数。以三个事件a, b, a 为例, 对于aba 这个事件序列来说, 就对应一个三维条件概率矢量, 其中x是第一个事件a 发生的条件概率, y是b 发生的条件概率, z是第二个事件a 发生的条件概率。在一个PST 节点上还记录了相应事件序列在数据集中的发生次数。整个PST 是树形的层次结构, 父节点是子节点的最长后缀。一个具体的PST 模型如图1 所示。从一个或多个事件序列都可以构造出相应的PST。

在得到PST 模型之后, 一个事件序列发生后下一个事件的条件概率可以逆序从顶向下搜索PST 得到。以图1 中的PST 为例, 如果想知道ab 事件发生后a 发生的条件概率, 可以从根节点开始搜索b, a, 之后就可以到达ab 节点, 而此时a 发生的条件概率为0.606。当事件序列长度n很大时, 为限制PST 的深度, 可以用最近发生的k个事件来代替这n个事件来

* 2008-08-12 收到, 2008-11-22 改回

** 基金项目: 国家863 高技术研究发展计划(2007AA012447); 北京市教育委员会科技计划(KM200810037001); 北京市属高等学校人才强教计划(PHR-IHLB); 北京市教育委员会科研基地建设项目; 北京物资学院青年基金资助项目。

*** 师鸣若, 女, 1975 年生, 硕士, 讲师, 研究方向: 软件设计, 网络安全和数据挖掘。

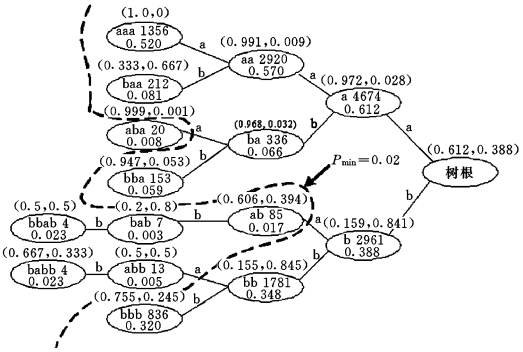

图1 一个具体的PST 模型

计算条件概率。图1 中有一条虚线, 它是概率为 P_{min} = 0.02 的分解线, 小于该概率的事件都在虚线的左边, 否则在虚线的右边。在实际计算中, 可以忽略概率太小(如小于0.02)的子序列, 提高计算效率。

2 异常分析

PST 模型的一个主要用途是可以用来计算事件序列之间的相似度。这里定义的相似度为一个事件序列与一个事件序列集合之间的相似度。首先我们计算一个事件序列 σ 在某个事件序列集合 S 中出现的概率。计算公式如下:

$$\sigma = s_1, s_2, \cdots, s_l$$

$$P_s(\sigma) = P_s(s_1) \times P(s_2 | s_1) \times \cdots \times P_s(s_l / s_1, \cdots, s_{l-1}) = \prod_{i=1}^{l} P_s(s_i | s_1, \cdots, s_{i-1})$$

其中 $P_s(\sigma)$ 表示事件序列 σ 在事件序列集合 S 的条件概率, 而这些条件概率可以通过构造PST 模型得到。显然, 如果 $P_s(\sigma)$ 很大时, 这表示 σ 在事件序列集合 S 中出现的概率很大。从聚类的角度看, 这也表示 σ 与这一事件序列集合的相似度很大。我们进一步定义事件序列 σ 与事件序列集合 S 的相似度为:

$$sim_s(\sigma) = \frac{P_s(\sigma)}{P_r(\sigma)} = \frac{\prod_{i=1}^{l} P_s(s_i | s_1, \cdots, s_{i-1})}{\prod_{i=1}^{l} p(s_i)}$$

其中 $P_r(\sigma)$ 表示 σ 中所有事件 s_i 独立随机发生的概率, 这也表示 σ 对于事件序列集合 S 完全是一个随机序列的概率。显然如果 $sim_s(\sigma)$ 大于1, 则表示 σ 有可能在 S 中发生; 如小于1, 则表示 σ 不太可能在 S 中发生, 说明有新的聚类产生。所以1 对于相似度来说是一个天然的阈值。

在定义了相似度以后, 就可以对相似的事件序列进行聚类。首先我们提取某一地区某一时间段的事件序列。因为计算相似度时, 序列的长度可能对计算结果有影响, 所以我们选取固定长度的事件序列进行分析。事件序列就被分割为固定长度的事件序列集合。同时, 因为连续出现的某一事件往往是同一事件的多次报警, 为处理简便起见, 我们把他们合并为一个事件进行处理。聚类的步骤如图2 所示。

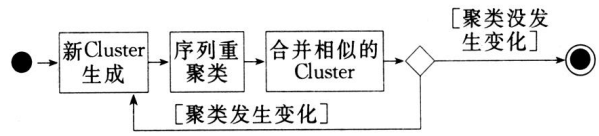

图 2 聚类算法

聚类过程是一个反复的迭代过程, 为减少计算时间, 我们在循环20 次以后终止聚类算法。因为假设异常的事件序列在整体中只有很小的比例, 所以对数据进行聚类以后可以预见到那些比较大的cluster(聚类)所代表的是正常的事件序列。为了得到某地区一定时间内的正常事件序列模型, 取该地区一个月内的事件序列进行聚类。图3 是对北京地区2007 年7 月的事件序列进行聚类后的最大的10 个聚类的大小。

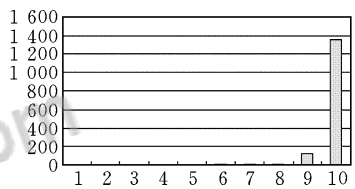

图3 北京2007 年7 月的事件序列的聚类结果

可以看出较大的cluster 与小cluster 之间的差别非常明显。经过实验, 发现每月的事件序列聚类中大于10 的那些cluster 的总大小几乎占全部cluster 大小总和的90% 以上, 而且大于10 的cluster 往往不超过5 个。所以取包含序列个数大于10 的那些cluster 作为基准cluster。也就是说这些基准cluster 代表了正常的事件序列。

因为一个地区的网络行为应该是相对稳定的, 所以相对上一个月而言, 该地区本月的正常事件序列应该有很大的相似性。所以可以用上个月的基准事件序列聚类来衡量当月的网络异常行为。因为每天都会有一些事件序列相对任何一个基准cluster 的相似度都小于1, 也就是说这些事件序列无法聚类到基准cluster 中。而这些事件序列在每天的数据中占一定比例, 所以可以计算出上个月的平均比例, 以此来判断每天的事件序列与上月的差异程度。在实际实现中, 用每天的比例减去上月的平均比例然后除以上月比例的标准差, 把它定义为每天数据的异常度。

3 事件关联分析

异常度的定义提供给用户一种分析事件序列的手段。在对数据建立了PST 模型之后, 我们还可以从中分析事件之间的关联。如果我们认为两个事件之间有

(下转第11 页)

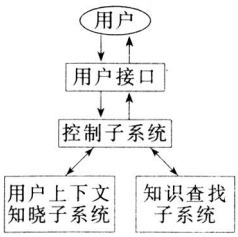

图2 MABCA系统体系结构

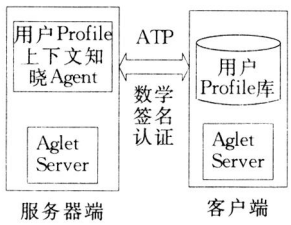

图3 MABCA实验原理图

以从地址簿插入; 实验中选用pmy 作为目的主机进行上下文知晓。点击"GO"按钮后, 系统便发送移动Agent"ContextAglet"到地址为Atp://pmy:4434 的主机上读取用户背景Profile 库, 并返回查询结果, 移动Agent 的后台运行情况如图5 所示。

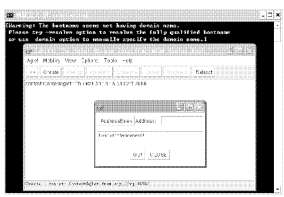

图4 ContextAglet 的运行界面

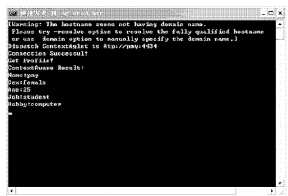

图5 上下文知晓移动Agent后台运行情况

由此我们可以读出用户pmy 的基本信息。Name: pmy; Sex: Female; Age: 25; Job: teacher; Hobby: Computer。根据这些用户上下文, 我们便可以提供给用户所需的各种知识。

5 结 论

上下文知晓是国内外一种新兴的研究领域, 目前国内在这一领域的研究还处于起步阶段。本文系统地介绍了移动Agent 上下文知晓模型的设计架构和实现技术, 在此基础上, 进行了MABCA 系统的设计与实验。该系统成功地实现了一个移动Agent 上下文知晓系统的基本功能, 对用户的背景(Profile)上下文进行了知晓。

参 考 文 献

[1] 张云勇, 刘锦德 移动Agent 技术(第二版)[M]. 北京: 清华大学出版社, 2003.

[2] 顾俊峰, 朱 亮 移动Agent 平台之分析与实现[J]. 计算机应用, 2000(4): 4-6.

[3] Chen G L, Kotz D. A Survey of Context-Aware Mobile Computing Research [M]. Department of Computer Science, Dartmouth College: Dartmouth Computer Science, Technical Report TR2000-381, 2000.

[4] 岳玮宁, 王 悦 基于上下文感知的智能交互系统模型 计算机辅助设计与图形学学报, 2005, 17: 74-79.

(上接第2 页)

较密切的关联, 那么他们往往存在密切的时序关系。因为PST 上的每个节点都记录了数据中事件相对于某个子序列发生的概率, 所以从分析数据的PST 模型入手可以找到那些联系密切的事件。首先我们确定一个概率阈值, 和一个支持度, 然后遍历PST, 如果找到一个节点, 它代表的事件序列到某一事件的条件转移概率大于概率阈值, 并且事件序列发生次数大于支持度, 就形成一条形如: $a_1, a_2, \cdots, a_n \rightarrow b$ 的规则。这表示如果事件序列 $a_1, a_2, \cdots, a_n$ 发生, 则下一个发生的事件是 b 的概率很大。这也表示事件序列 $a_1, a_2, \cdots, a_n$ 与事件 b 之间有密切的关联。

4 结束语

我们提出了一个基于概率后缀树的网络报警事件序列分析的框架, 在此基础上分析了网络报警事件数据集, 既可以得到网络异常的信息, 又能够分析报警事件之间的关联。因为数据中并没有明确界定异常与正常, 而且实际上二者也没有明确的界限。所以我们在实际的算法中并不判断异常与否, 而是给出异常的度量。用户可以根据实际情况来灵活的确定异常与正常的界限。同时网络报警事件之间可能存在某种因果联系, 这种联系可能涉及多个事件。假设这种联系在数据中体现为时序上的关联, 通过本文的方法, 可得到多个事件间的关联信息。

参 考 文 献

[1] Pei S, Sanjay C, Bavani A. Mining for Outlier in Sequential Database [C]. ACM SDM Conference, 2006.

[2] Wang J, Wang W. CLUSEQ: Efficient and Effective Sequential Clustering [C]. IEEE ICDE Conference, 2003.

[3] Cheung C, Yu J X, Lu H. Constructing Suffix Tree for Gigabyte Sequences with Megabyte Memory [J]. IEEE Trans on Knowl and Data Eng, 2005, 17(1): 90-105.

[4] Yuanyuan T. Practical Methods for Constructing Suffix Trees [J]. The VLDB Journal-The International Journal on Very Large Data Bases, 2005, 14(3): 281-299.

[5] Emilios C. All Maximal-pairs in Step-leap Representation of Melodic Sequence [J]. Information Sciences: an International Journal, 2007, 177 (9): 1 954-1 962.