基于双层结构聚类的分析网络行为①

周广新[1]，王传安[1]，赵海燕[1,2]

[1](安徽科技学院 理学院，凤阳 233100)
[2](江苏大学 计算机与通信工程学院，镇江 212013)

摘 要：首先对网络服务器监测到的流数据进行采集，提出将双层结构聚类算法应用于流数据的聚类分析，进而得到校园网用户网络行为的特征，该特征对于进一步优化校园网络建设具有重要意义。

关键词：网络行为；聚类算法；双层结构

Analysis of Network Behavior Based on Two-Tier Structure Clustering

ZHOU Guang-Xin[1], WANG Chuan-An[1], ZHAO Hai-Yan[1,2]

[1](College of sciences, Anhui Science and Technology University, Fengyang 233100, China)
[2](College of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: First, this paper puts forward an algorithm of two-level aggregate, which can be applied to the analysis of flow data collected from a remote on-line monitoring system. Then, with the result of the algorithm we can get the characteristics of students' network behavior, which will have a substantial influence on advancing the construction of campus network.

Key words: network behavior; clustering; two-tier structure

网络用户行为指用户在使用网络资源所呈现出的规律，可以用某些特征量的统计特征或特征量的关联关系定量或定性的表示[1]。传统的网络行为研究主要是针对性能而言，在这方面比较有代表性的研究成果是 IETF 的 IPPM（IP performance metrics）工作组在测度定义方面的一些工作和 CAIDA（cooperative association for internet data analysis）所做的一些针对网络行为的测量研究。互联网中心实验室给了一种基于 web 文档内容的网民行为模型，它将用户行为分为信息查询、沟通交流、休闲娱乐、电子服务、电子商务五大类，并对用户群体进行分类，得出十类特色网民群体[2]。这是从应用层对网络行为的一种表示。

由于网络监测信息数据以流的方式出现，数据流的广泛应用,使得对它的研究已经成为一个热点问题，流环境下的流聚类问题也成为热点方向之一。对流数据的研究首先是由 Munro 和 Patterson 提出的；Henzinger 等人定义了流数据模型；Alon 等人证明了单路访问算法计算数据流统计摘要所需内存大小的上限和下限[3]。以往的聚类数据流算法只是简单地将数据流聚类问题看作一次扫描聚类算法的变种.随着数据流聚类问题从理论研究转向应用研究, Aggarwal 等人认为，需要能够只使用新数据就能够追踪聚类变化的算法，这就要求算法必须是增量式的，对聚类表示要简洁，对新数据的处理要快速，对噪音和异常数据是稳健的。近年来，有学者提出了针对流数据的聚类算法，典型的有 STREAM[4,5]算法和 CluStream[6]算法.

本文通过提出将双层结构聚类算法应用于校园网络采集到的流数据分析上，对采集到的 NetFlow 数据流记录进行聚类分析，来获取用户访问行为，为了进一步抽取用户感兴趣的潜在有用模式与信息，对用户访问路径进行优化，得到用户的频繁行为组合模式。该算法能够满足处理速度快，准确率高的优点，通过对聚类结果的分析可以得到网络行为的特征，这将对优化网络拓扑结构、改善负载均衡，进行网络应用方

① 基金项目:安徽科技学院教研项目(X201106)
收稿时间:2010-12-18;收到修改稿时间:2011-01-20

面的调整做准备，是有效的改善学校网络拥塞状况的前提。

2 双层结构聚类

流数据是快速变化连续并且是海量而有序的，因而要有效地利用有限的空间与时间。流数据本身所具有的特征使得传统的聚类算法不可能直接应用于(甚至不能应用于) 流数据聚类。本文采用一种双层的结构用于解决流数据中的聚类问题,分别是快速计算层和精确分析层。

2.1 快速计算层

该层主要目的是产生反映当前块中相似性较高的数据区域特征点,由于其计算简单, 处理快速，且输出的这些特征点的数量远远小于最初流数据的数量，通常采用快速但粗糙的聚类方法。本文快速计算层采用经典的 k-means 算法，并根据分析网络行为的需要进行了改进。

K-means 算法基于使聚类性能指标最小化的原则，通常使用的聚类准则函数是聚类集中的每个样本点(数据或对象)到该类中心的误差平方和,并使它最小化。K．means 算法有很多的变种，基本的思想是不变的：(1)随机的确定 K 个聚类中心；(2)再根据欧氏距离把每个随机点分配到最接近其均值的聚类中；(3)计算分配到每个聚类的点的均值向量，并作为新的中心进行递归。

由于网络数据流数据是大规模数据，为了提高聚类的速度和精确度，本文对 K-means 算法进行了改进：

(1) 确定数据流的关键属性字段。

(2) 对所有记录按照关键属性字段从小到大进行排序，设为全集 A。

(3) 假设分成层：令 m=total／K。

求从第 0～第 m 条记录的关键属性字段的平均值 a0；第 m～第 2m 条记录的关键属性字段的平均值 a1.；第 2m～第 3m 条记录的金额字段的平均值 a2:；……；求第(k 一 1)m～最后一条记录值 ak-1。将这些值放在一个集合 R={ a0,a1,a2,…,ak-1}之内。

(4) 对 A 内的每一条记录的关键属性值，进行 abs(x-ai)的运算，取得最小的 i，将{ x,ai }归为一类。对所有记录做完之后，整个关键属性字段共分成类：{ a0，…}；{ a1，…}；{ a2，…}；…；{ ak-1，…}。

(5) 对上面每个类分别求平均值，得出的集合 R′={a′ 0，a′ 1，…，a′ k-1}。

(6) 最后对每个类内的 a′ i 值去掉(i=0,1,…,k)，最后得出的个类为所分出的层为所要的结果。

上面的算法中，对于初始中心点的选取中，在排序的基础上做了两次算术平均值的运算，使得中心点更具有代表性，而且收敛程度更快。利用此算法之前，可以事先观察数据流样本，然后选取初始的值。利用图标工具观察分出的层是否每层都近似属于正态分布，如果效果不理想，可以增减的值。数据划分也许不能够达到绝对满意，只要能够达到相对符合正态分布，目的就达到了。

2.2 精确分析层

该层对快速计算层得到的数据上进行精确的分析, 此时数据对算法时间的要求远远小于最初的流数据, 所以可以采用复杂的 BIRCH 聚类方法进行分析, 为了抽取用户感兴趣的潜在有用模式与信息，采用关联规则对路径进行优化。

BIRCH 算法可用于动态聚类,用它进行聚类时需要 2 个参数:类的最大个数 K,类的最大半径 ε .每一个类在内存中用二元组{Ci,Ri}来表示,Ci,Ri 分别表示第 i 个类的中心与半径. BIRCH 算法保证在类的数目不大于 K 时,各个类都是紧密的,即 Ri≤ ε ;当得到的类多于 K 时,需要适当的放大阈值 ε [7]。

算法步骤如下:

1) 初步确定聚类中心:随机选取几个会话记录向量,满足条件:彼此之间距离大于 2 ε ,r=1;

2) 读入第 r 条会话记录,计算它与各个聚类中心的距离 d,设它与第 i 个聚类中心 Ci 距离最小;

3) 假设已经把会话 r 归于第 i 类,计算第 i 个聚类新半径 R′ i,如果 R′ i≤ ε ,这说明第 i 个类仍然紧密,就把会话 r 归于类 i,把它的半径更新为 R′ i;如果 R′ i> ε ,这说明若把会话 r 归于第 i 类,第 i 个类就不在是紧密的:若此时聚类数小于 K,就把会话 r 单独归为一类,该类的中心 Ci=r,Ri=0;若此时聚类数不小于 K,则应适当的放大 ε ,返回步骤 1)重新计算;

4) r=r+1;若 r≤绘画记录总数,转向步骤 2),否则聚类结束。

由于 Web 站点在初始设计时不可能得到用户频繁访问路径的信息,且动态聚类后的数据往往不能准确地与用户频繁访问路径相吻合，需要进一步对聚类后的路径进行优化。路径优化算法可表示如下:

首先根据频繁访问路径算法得到频繁访问路径FP,对应记录为X(fp, np),其中fp表示URL组成的序列,np表示浏览路径集合中FP出现的次数。

for all X∈FP

for allY∈F Pand Y≠X

if X•fp∈Y•fp•sub then X•np=X?np-Y•np //检查X•fp是否为另一记录Y•fp的子序列

ifX•np>n //n为预先设定的次数阈值

setSuplink(X•fp) //设计新的从X•fp起点指向终点的超链。

3.1 采用双层结构聚类进行网络行为分析

本文采用双层结构聚类模型进行网络行为分析，主要功能模块如图1所示。

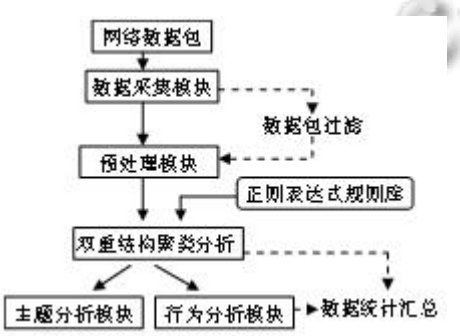

图1　网络行为分析模型图

3.2 分析策略定制

系统可以提供灵活、动态的策略定制机制，用户可根据使用网络范围和用户行为分析目标，自定义分析策略。从而可以动态、全面地了解网络流量的分布，完成网络行为的分析。

流量数据是多维的格式数据,网络数据来源,流经设备,链路类型比较复杂.通过对网络元素进行分组,将同类元素作为一个分析对象进行管理,能帮助用户进行网络资源组合,同时将 NetFlow 输出到不同的类别,以便进行针对性类聚分析。

网络元素由网络设备、端口、AS 号、IP 段组成,元素属性应该包含 ID、名称、元素组成、链接类型(专线、拨号、pppoe、其它)等信息.对象是针对某个用户、单位、地区或某类应用而聚合起来的元素组合。对象属性包括 ID、对象名称、元素列表等。行为分析策略包含内容如表1所示。

3.3 数据采集及预处理

实验开发平台及工具包括：Linux 9.0，Perl，PHP，Flow-tools，Cflow，FLowScan，RRDtools，CUFlow，Oracle 9i。现已在安徽科技学院网络中心搭建平台，通过 Cisco6509 定期将经过的流数据抛到指定服务器的指定端口，该服务器上的 Netherlandsflow 流量工具和采集器采集到流数据，并每五分钟生成流报表文件。

表1　行为分析策略

序号	字段	说明
1	策略名称	标识策略
2	元素类型	设备、端口、IP 段、AS 号
3	监考内容	流量、包、会话、协议类型、源地址、目的地址、应用
4	汇总方式	按流入、流出汇总或按双向汇总
5	汇总周期	日、周、月
6	TOPN	降序排列显示前 N 个

实验采用的是2010年6月某一周的NetFlow数据,记录每5 min输出1次,即全天被分成288个统计时段。在开发的过程中, 根据测控点形态的不同采取不同的清理方法，过滤掉噪声点和孤立点。然后将数据转换成适合于机器挖掘的格式。

3.4 网络行为分析

采用双层结构聚类法对预处理后的 NetFlow 数据流记录进行聚类分析，如下表 2 所示，其中 Src_host 为源主机，dst_host 为目标主机， service 为连接协议，duration 为平均持续时间，Host_count 为源/目标主机出现次数。（注意：in 为校内用户 IP，out 为校外用户 IP）。

表2　聚类分析后的数据

NO	Src_host	Dst_host	Sevice	Duration (s)	Bytes (M)	Host_count
1	out	211.70.51.14	http	9503 (招生)	0.75	12432
2	In	219.133.40.91	udp	432348 (qq)	31.34	4321
3	in	222.186.10.71	smtp	4626 (126)	1.45M	3346
4	out	211.70.48.17	tcp	9835 (教学处)	1.21M	1945
5	in	211.70.51.33	ftp	324980 (影视)	642M	1897
6	in	211.70.50.138	http	363301 (图书)	21.78 M	1689

对表 2 分析可得，校园网招生办主页（211.70.51.14）访问最频繁，因为此时正值高考结束及专升本考试招生期间，每年可以在此期间对教育网（www.edu.cn 等网站做镜像以提高访问速度）。同时可以看出校园网用户对学校影视服务器和空间中转服务器（211.70.51.33）访问量较大，需要完善这些服务

器的服务质量，扩充空间，高峰访问期放宽下载上传速度限制，以尽可能满足校园网用户需求。

为了进一步抽取用户感兴趣的潜在有用模式与信息，采用精确分析层的路径优化算法对行为路径进行优化，结果如图 2 所示：

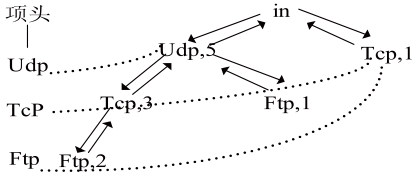

图 2 路径优化结果

从图 2 可以看出，路径优化算法得到的频繁行为组合模式为：{Udp,Tcp,Ftp|2}，由频繁组合模式可知校内各个网段的用户在上网的时候都使用聊天软件（如 QQ），其中学生区网段至少三分之二用户访问学校影视服务器，而教职工网段用户几乎全部访问了校图书馆。

4 结论

本文对网络服务器监测到的流数据进行采集，提出一种将双层结构聚类算法应用于流数据的聚类分析，该算法能够满足处理速度快，准确率高的优点，进而分析得到校园网用户网络行为的特征，这将对优化网络拓扑结构、改善负载均衡，进行网络应用方面的调整做准备，是有效的改善学校网络拥塞状况的前提。

参考文献

1 马力,焦李成,董富强.一种 Internet 的网络用户行为分析方法的研究.微电子学与计算机,2005,22(7).

2 李冠强,陈雅,李强.中国互联网用户网络使用行为分析.中国图书馆学报,2004,23(5).

3 蒋盛益,李庆华,李新.数据流挖掘算法研究综述.计算机工程与设计,2005,26(5).

4 Guha S, Mishra N, Motwani R, et al. Clustering data streams. Proc. of IEEE Symposiumon Foundations of Computer Science(FOCS00). 2000, 71−80.

5 Guha S, Meyerson A, Mishra N, et al. Clustering data streams: Theory and practice. Knowledge and Data Engineering. IEEE Transactions, 2003,15(3).

6 Aggarwal C, Han J, Wang J, et al. A framework for clustering evolving data streams. Berlin, Germany: Procof Int Confon Very Large Data Bases (VLDB03), 2003.

7 Ramakrishnan R. Database management systems (2nd ed).北京:清华大学出版社,2001.726−729.

(上接第 164 页)

持终端，操作者可以在手持终端选择测试点，也可按测试点的序号连续采集，接收到的温湿度数据经过处理在手持终端的显示屏上显示。经过模块化的电路测试、软件调试和系统组装，测温精度可达到±1℃，测湿精度为士 2%RH，通信距离可达 300 米，可广泛应用于温室和大坝粮仓等领域的温度、湿度监测中。

参考文献

1 项新建.基于多传感器数据融合的粮食仓库温度监测系统.仪器仪表学报,2003,24(5):525−527.

2 Dorf RC. Modern conctrol systerm. Beijing: Science Publishing House, 2002.20−160.

3 李朝青.单片机原理及接口技术.北京:北京航空航天大学出版社,2009.180−182.

4 王松武,于鑫.电子创新设计与实践.北京:国防工业出版社,2005.212−214.

5 张志伟.基于 PTR2000 的电力无线手持抄表系统.电测与仪表,2004,41(7):56−58.

6 Sensirion Company. Application Note SHTxx Humidity & Temperature Sensor, Shenzhen: SUNSTAR, 2008.1−10.

7 冯显英,葛荣雨.基于数字温湿度传感器 SHT11 的温湿度测控系统.自动化仪表,2006,21(1):59−61.

8 贺桂芳.基于 SHT11 的温湿度无线测控系统设计.微计算机信息,2007,23(8):307−309.

9 卢超.基于 PC 机与单片机分布式温度采集系统的设计.仪表技术与传感器,2007,(6):38−40.

10 薛瑞,适用于 51 单片机的 CRC 算法研究.北华航天工业学院学报,2007,17(1):12−14.