

文章编号:1001-9081(2007)08-1862-03

## 面向校园网的 IP 地址逐步优化层次聚类算法

楼若岩<sup>1</sup>,许晓东<sup>1,2</sup>,朱士瑞<sup>1</sup>

(1. 江苏大学 计算机科学与通信工程学院, 江苏 镇江 212013;

2. 江苏大学 现代教育技术中心, 江苏 镇江 212013)

(louruoyan@yahoo.com.cn)

**摘 要:**对校园网主干数据流中 IP 地址进行聚类,可以得到网络用户访问地址的分布概况从而了解用户行为特征。已有聚类算法大都将 IP 地址作为普通数字考虑,忽略了其特征属性以致聚类结果不合理。为此提出一种改进算法:首先基于最长前缀匹配和改进的最近邻规则算法得到初始聚类,然后运用逐步优化层次聚类的思想进一步聚合最靠近子类,最终得到基于 IP 地址特征属性的聚类。实验结果表明该算法与以往算法相比,提高了聚类效果,具有较好的准确性和可行性。

**关键词:**IP 地址聚类;最近邻规则;最长前缀匹配;逐步优化的层次聚类

**中图分类号:** TP393.07 **文献标志码:** A

## Campus-oriented stepwise-optimal hierarchical clustering algorithm of IP address

LOU Ruoyan<sup>1</sup>, XU Xiaodong<sup>1,2</sup>, ZHU Shirui<sup>1</sup>

(1. School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang Jiangsu 212013, China;

2. Modern Education and Technology Center, Jiangsu University, Zhenjiang Jiangsu 212013, China)

**Abstract:** The cluster analysis of IP addresses can reveal useful knowledge for profiling of traffic flows and user behavior. However, the popular clustering algorithms were not applicable directly to IP addresses of the campus network traffic flows. The clusters which were generated by generic algorithms were inconsistent with the IP addresses partition and difficult to interpret. To overcome the shortcoming of the current algorithms which neglect the characteristics of IP addresses, a new algorithm which could effectively improve IP addresses clustering was proposed. Firstly, the initial clusters were got by adopting the longest prefix algorithm and the nearest neighbor clustering algorithm. Then the thought of stepwise-optimal hierarchical clustering was applied to merge the nearest groups of initial clusters. The similarity between initial clusters was determined by the longest prefix of IP addresses contained in these clusters. Finally, the algorithm automatically and meaningfully yielded clusters which were in accord with the characteristics of IP addresses on traffic flows. The results show that the proposed algorithm is accurate and effective in clustering IP addresses and robust to the input sequence of data.

**Key words:** IP address clustering; nearest neighbor rule; Longest Prefix Match (LPM); stepwise-optimal hierarchical clustering

## 0 引言

在网络工程中,对 IP 地址进行聚类分析可以得到网络用户的访问行为特征和基于访问地址的主干流量分布情况,这将为网络管理员了解用户网络行为及流量概况,规划设计网络和管理网络资源等提供有用信息。

文献[1]提出使用 K-Means 和 K-Medoids 算法对 IP 地址进行聚类分析,但这两种算法都受到 IP 地址特征属性的约束,而且都需要输入最终的聚类数目,得到的 IP 地址的聚类与网段划分不一致。文献[2]提出把进化算法应用到聚类能够更好地处理属性间的交互。文献[3]也证明进化算法应用在聚类上比 K-Means 算法好,而且随着聚类数目的增加其效果更优。但进化算法比传统算法慢一到两个数量级,所以不适合使用在基于大规模数据集的 IP 地址上。基于密度的算法如 DBSCAN<sup>[1]</sup>只能应用到可数的数据集中,数据对象被看作是  $n$  维空间中被分到各个聚类中的密度相关的点,由于 IP 地址不能像普通的数字一样定义,所以该算法也不可行。文

献[4-6]描述了最长前缀匹配的概念并用于 IP 地址聚类上,但是没有对含有 IP 地址属性的数据集提供一个通用的聚类算法。综上所述,以往研究存在的主要问题有:算法中没有考虑 IP 地址的特征属性;不能用于大规模数据集的处理;没有对包含 IP 地址属性的聚类数据集提供一个通用的算法。

本文针对聚类算法中没有考虑 IP 地址的特征属性这一问题进行改进,提出一种新的聚类算法,并对校园网主干流量中 IP 地址进行分析。

## 1 算法的基本思想

### 1.1 最近邻规则算法

最近邻规则算法<sup>[1]</sup>不需要计算中间值或者均值,它把对象反复合并到已存在且最接近该对象的某个聚类中<sup>[1]</sup>。算法开始执行时需要设定一个阈值来测量对象间的最小相似度。如果某个对象达不到这个阈值(比如它更加相似),那么就新建一个类,将该对象包含其中。通常,考虑把邻近的对象的最大数目作为在最小相似处理上除了阈值之外的一个

收稿日期:2007-02-09;修回日期:2007-04-05。 基金项目:江苏省教育厅高校科学研究基金资助项目(03KJD520073)。

作者简介:楼若岩(1982-),女,山东济南人,硕士研究生,主要研究方向:网络管理;许晓东(1965-),男,福建漳州人,副教授,主要研究方向:网络管理和系统集成;朱士瑞(1984-),男,江苏泗洪人,硕士研究生,主要研究方向:网络安全。

约束。

### 1.2 最长前缀匹配

通常在TCP/IP通信中使用使用最长前缀匹配(Longest Prefix Match, LPM)来查看路由表。图1说明了LPM的概念,显示了两个16字节字符串,从最左边开始,依次对两个字符串每个字节进行匹配,前10个字节都相同,直到第11个字节出现不同。所以LPM值是10。

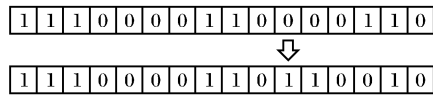


图1 最长前缀匹配

### 1.3 逐步优化的层次聚类

逐步优化的层次聚类是通过合并最靠近的两个子类来实现聚类过程,体现了一种最小方差的思想<sup>[7]</sup>,算法描述如下:

初始化:  $c$  (期望的最后聚类的数目)

$c' = n$ ,  $D_i = \{x_i\}$ ,

$i = 1, 2, \dots, n$ ,

do  $c' = c' - 1$

    寻找其合并类,例如  $D_i$  和  $D_j$

    合并  $D_i$  和  $D_j$

while ( $c = c'$ )

return  $c$  个聚类

## 2 算法实现

本文提出了一种新的基于数据集中IP地址的聚类算法,考虑到IP地址的属性特征,因此应用LPM和最近邻规则算法。算法采用两层聚类,通过比较两个地址中最长的前缀匹配值来决定它们的相似性,把改进后的最近邻规则算法用在划分相似的IP地址上,并对第二层聚类结果采用逐步优化的层次聚类思想,合并最靠近子类。

算法描述如下:第一层聚类基于IP地址的网段划分来建立,产生的每一个聚类对应一个网段(网段A,B,C,D,E)。这一层可以跳过,但是使用它可以提高算法的聚类效果。第二层聚类对每一个网段中的地址分别聚类,首先计算IP地址之间的LPM值并存储在建立的邻接矩阵中,然后,依次考虑每一个IP地址,基于最近邻规则的思想依次对所有拥有LPM值的IP地址建立起聚类。与最初始的最近邻规则算法不同,本文的算法为每一个拥有LPM值的IP地址建立一个新的聚类,当某个IP地址与当前IP地址间的LPM值比它之前某个IP地址的LPM值更大时,这个IP地址可以从LPM值较小的聚类转到较大的聚类,聚类被迭代修改。以往设定阈值的算法有可能把某些属于相同网段,但具有较低LPM值的地址分到不同的类中。改进后的最近邻规则算法不需要输入两个IP地址间相似度的阈值,实现了算法自动化。

文献[8]提出的混合聚类算法使用了LPM算法,但是由于仅考虑把最近邻的两个IP地址聚类,使得许多应该聚到一个类中的地址被分割到几个子类中,破坏了聚类的完整性。这样不仅给网络管理的分析工作带来了庞大的工作量,更没有达到使不同类之间具有最大的相异度这一聚类目的。

本文提出改进,运用逐步优化的层次聚类的思想,合并最靠近的两个子类。算法中对第二层聚类的初次结果进行优化,即对类进行聚类。类间相似度取值大小多少会影响聚类结果的完整性,所以,确定类间相似度的多少也是决定算法有效性的关键参数之一。类间相似度过小会使聚类无法形成完

整结果,结果支离破碎,不易于分析网络行为概况;类间相似度过大有可能合并多个单位或组织的IP地址到一个类中,同样使算法生成的聚类结果不易分析。所以要想获得较准确的基于单位或组织的划分和对网络分析有帮助的类,就要确定合适的类间相似度。经过大量的试验数据测试发现,在校园网环境中,类间相似度设置为4,可以得到类的最大相异度,所得聚类能够基于单位或组织进行划分,达到预期的希望了解用户访问地址分布概况的目的。因此设定类间相似度4以内的子类聚合为同一聚类,达到有效聚类的目的。

校园网之外的网络情况可能会有所不同,需要根据具体情况调整类间相似度。如果网络管理者希望以更宏观的角度了解访问分布,可以将类间相似度取更大值或通过神经网络自学习确定其值。

改进后算法复杂度  $O(n^2/2)$ , 优于最初的最近邻规则算法复杂度  $O(n^2)$ 。算法相关定义及实现如下:

**定义1** LPM[ $i$ ][ $j$ ] 表示IP[ $i$ ]和IP[ $j$ ]的LPM值,其中IP[ $i$ ]和IP[ $j$ ]分别表示第 $i$ 个IP地址和第 $j$ 个IP地址( $i, j \in \{1, 2, \dots, n\}$ )。

**定义2** 函数LPM( $K[c]$ , IP[ $i$ ])表示在第 $c$ 个聚类集合中,IP[ $i$ ]与该集合内其他元素的LPM值,其中  $K[c] = \{IP[i], IP[j], \dots\}$  ( $i, j \in \{1, 2, \dots, n\}$ ),  $K[c][i]$  是第 $c$ 个聚类中的第 $i$ 个IP地址;

第二层聚类算法实现:

```
getLPM(); /* 初始化,得到LPM矩阵 */
K[1][1] = IP[1]; c = 1;
for (i = 1; i <= n; i++)
{
    for (max(LPM[i][j]))
    {
        K[c] = K[c] ∪ IP[i]; /* 将IP[i]添加到K[c]中 */
        for (m = 1; m < c; m++)
        {
            if (j在其他集合K[m]中)
            {
                if (abs(LPM(K[c], IP[j]) - LPM(K[m], IP[j])) < 4)
                {
                    K[c] = K[c] ∪ K[m];
                    /* 对初次聚类结果逐步优化,合并为一类 */
                }
            }
            else
            {
                if (LPM(K[c], IP[j]) < LPM(K[m], IP[j]))
                {
                    K[c] = K[c] - IP[i];
                    /* 从K[c]中删除IP[i] */
                    K[m] = K[m] ∪ IP[i];
                }
            }
            else
            {
                K[m] = K[m] - IP[i];
                /* 从K[m]中删除IP[i] */
                K[c] = K[c] ∪ IP[i];
            }
        }
    }
}
c ++;
```

3 算法应用及实验结果

我们采集了校园网主干数据流中属于 C 类的 16 个不同的目的 IP 地址(假设第一层聚类已完成)。选取目的 IP 地

址,是为了得到基于目的地址的主干流量分布和用户的访问倾向(该算法也同样适用于源地址分析)。IP 地址的二进制形式及其 PLM 值如表 1 所示。该邻接矩阵对称,为避免重复运算,左下三角区域及对角线元素值都置为零。

表 1 测试的 16 个目的 IP 地址的不同进制表示及最长前缀匹配邻接矩阵

序号	IP 地址的 十进制形式	IP 地址的 32 位二进制形式																最长前缀匹配矩阵															
1	202.195.164.141	1100	1010	1100	0011	1010	0100	1000	1101	0	22	8	12	12	3	3	1	1	1	21	8	12	12	1	12								
2	202.195.167.22	1100	1010	1100	0011	1010	0111	0001	0110	0	0	8	12	12	3	3	1	1	1	21	8	12	12	1	12								
3	202.119.35.198	1100	1010	0111	0111	0010	0011	1100	0110	0	0	0	8	8	3	3	1	1	1	8	22	8	8	1	8								
4	202.205.14.71	1100	1010	1100	1101	0000	1110	0100	0111	0	0	0	0	22	3	3	1	1	1	12	8	21	16	1	23								
5	202.205.12.121	1100	1010	1100	1101	0000	1100	0111	1001	0	0	0	0	0	3	3	1	1	1	12	8	21	16	1	22								
6	211.65.82.87	1101	0011	0100	0001	0101	0010	0101	0111	0	0	0	0	0	0	20	1	1	1	3	3	3	3	1	3								
7	211.65.95.119	1101	0011	0100	0001	0101	1111	0111	0111	0	0	0	0	0	0	0	1	1	1	3	3	3	3	1	1								
8	162.105.129.12	1010	0010	0110	1001	1000	0001	0000	1100	0	0	0	0	0	0	0	0	0	16	16	1	1	1	1	16	1							
9	162.105.127.128	1010	0010	0110	1001	0111	1111	1000	0000	0	0	0	0	0	0	0	0	0	0	17	1	1	1	1	21	1							
10	162.105.15.221	1010	0010	0110	1001	0000	1111	1101	1101	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	17	1							
11	202.195.163.209	1100	1010	1100	0011	1010	0011	1101	0001	0	0	0	0	0	0	0	0	0	0	0	8	12	12	1	12								
12	202.119.32.7	1100	1010	0111	0111	0010	0000	0000	0111	0	0	0	0	0	0	0	0	0	0	0	0	8	8	1	18								
13	202.205.11.139	1100	1010	1100	1101	0000	1011	1000	1011	0	0	0	0	0	0	0	0	0	0	0	0	0	16	1	21								
14	202.205.195.162	1100	1010	1100	1101	1100	0011	1010	0010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	16								
15	162.105.120.1	1010	0010	0110	1001	0111	1000	0000	0001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1								
16	202.205.15.178	1100	1010	1100	1101	0000	1111	1011	0010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0								

表 2 IP 地址最终聚类结果比较

Asim Karim 的算法聚类结果		改进后的聚类结果	
Cluster1	202.195.164.141 202.195.167.22	Cluster1	202.195.163.209 202.195.164.141 202.195.167.22
Cluster2	202.195.163.209	Cluster2	202.119.32.7 202.119.35.198
Cluster3	202.119.32.7 202.119.35.198	Cluster3	202.205.14.71 202.205.11.139 202.205.15.178 202.205.12.121
Cluster4	202.205.14.71 202.205.15.178	Cluster4	202.205.195.162
Cluster5	202.205.12.121	Cluster5	211.65.82.87 211.65.95.119
Cluster6	211.65.82.87 211.65.95.119	Cluster6	162.105.129.12 162.105.120.1 162.105.127.128 162.105.15.221
Cluster7	162.105.129.12		
Cluster8	162.105.127.128 162.105.120.1		
Cluster9	162.105.15.221		
Cluster10	202.205.11.139		
Cluster11	202.205.195.162		

对使用文献[8]提出的混合聚类算法和本文改进后算法的最终聚类结果进行了比较,如表 2 所示。16 个地址使用 Asim Karim 的算法聚为 12 类,使用本文改进算法聚为 6 类,可以看出改进后的算法达到有效聚类的目的。聚类结果的准确性使用 nslookup 进行了验证:聚类 1 和聚类 5 属于江苏大学,聚类 2 属于南京大学,聚类 3 属于中国教育和科研计算机网网络中心,聚类 4 属于北京市行政学院,聚类 6 属于北京大学计算中心/北京市北京大学。

算法进一步应用在所收集的校园网 7 天的数据流上,去除了重复地址,数据清理后得到 21 050 个数据流。为了了解网络用户访问地址概况,对数据流中目的 IP 地址进行聚类,结果如表 3 所示。第一层聚类产生的网段 D 和 E 中的 IP 地址多用作组播或保留,不再进一步聚类。第二层聚类共产生 1 877 个聚类结果,对应了原始集大小的 9.72%。使用 nslookup 工具对结果进行验证,共识别了 1 120 个类(大约 60%),其中有 1 003 个类(占可识别类的 89.6%)能正确的代表 IP 地址基于单位或组织的自然划分。其他研究人员也面临着不能识别 IP 地址域名和验证聚类结果的问题<sup>[9]</sup>。文献[9]中只有 50%的地址通过使用 nslookup 识别。但是本文提出的算法考虑了 IP 地址的特征属性,大大提高了聚类正确率。

表 3 数据集聚类的结果

	ClassA	ClassB	ClassC	Others	Total
第一层聚类	3 579	1 746	6 560	7 415	19 300
第二层聚类	583	357	937	—	1 877

实验最后对算法在读取数据顺序上的鲁棒性进行了测试。把相同的数据按相反顺序或任意排序读取,不同的读取顺序得到了相同的聚类结果,这表明算法不受读取顺序的影响。

4 结语

对数据流中 IP 地址的聚类分析是了解主干流量地址分布、用户访问行为特征的基础,寻找适用于 IP 地址的算法对于提高聚类质量和准确性有至关重要的意义。实验证明,本文算法能正确地形成约 89.6%的聚类,且生成的聚类与基于单位或组织的自然划分相一致,在算法复杂度上优于最初的最近邻规则算法,在聚类的完整性上优于 Asim Karim 的混合聚类算法,并大大提高了准确性。目前正在把该算法扩展应用到数据流的其他属性,如源端口和目的端口,发送和收到的

(下转第 1867 页)

## 2.4 仿真参数讨论

由于分组到达服从泊松分布的特点,所以在仿真对比实验中时将时隙设为10 ms,对1 s内的分组进行调度仿真试验,统计结果按照泊松均值50递增的顺序进行。

由于在一个时隙内四种业务流的到达比率是一个随机的过程,但是结合实际业务模型来看,UGS业务流的业务请求会多于rtPs业务流,rtPs业务流的业务请求会多于nrtPs业务流,nrtPs业务流的业务请求会多于BE业务流,所以在仿真对比实验中将四种业务流产生的比例按照4:3:2:1进行设定。

预留带宽和定制类用户所占比例这两个参数和网络的实际部署特点以及应用场景密切相关。针对不同的实际应用场景,上述两个参数可能会存在较大的差异,所以在仿真对比实验中暂且将预留带宽设置为系统整个带宽的1/3,定制类用户所占比例和非定制类用户所占比例设置为1:1。

从图2的仿真结果来看,本文所提的分组调度算法在对定制类业务的服务率上较文献[4]中的分组调度算法有较大的提高。在实际应用中,由于定制类用户对应的用户优先级更高,其业务请求应该尽可能的给予满足,所以本文所提算法较文献[4]中的算法具有更大的实际意义。

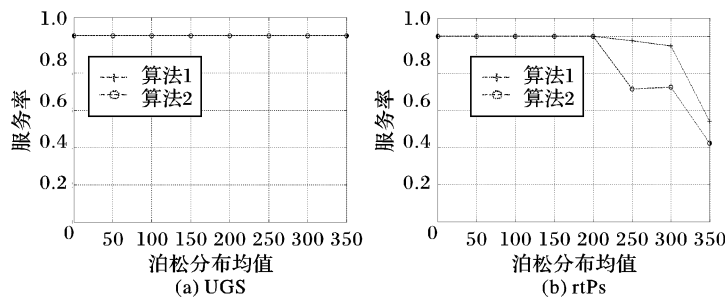


图3 定制类UGS业务和定制类rtPs业务服务率对比

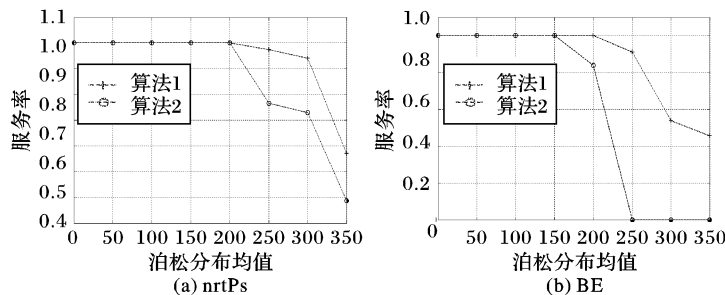


图4 定制类nrtPs和定制类BE业务服务率对比

图3、4为算法1、2在对定制类UGS业务流、定制类rtPs业务流、定制类nrtPs业务流和定制类BE业务流的仿真对比。

针对UGS业务流的特点,两种算法都给予了较高的优先级,所以在图3中,算法1、2对此类业务的服务率都为1。

对于rtPs和nrtPs业务流的特点,算法1对其采用的是将定制类业务和非定制类区分开,再分别采用WFQ算法进行调度;而算法2直接采用WFQ算法调度。由于处理方式不同,算法1能够更好地提供rtPs和nrtPs业务的服务率。

由于BE业务是一种尽力而为的处理方式,这类业务由于优先级低,在业务量大的情况下,会长期处于饥饿状态,算法1通过采用资源预留和定制类业务的优先处理,能够有效地缓解饥饿状态,提高BE业务的服务率;而算法2对BE业务只是采用传统的处理方式,故其BE业务服务率在负载比较大的情况下降到了0。

## 3 结语

基于对802.16-2004标准四种业务流的分析,提出了一种基于业务流均衡的分组调度算法,并进行了仿真对比。实验表明,本文所提算法能够更好地解决分组调度的优先权问题和“饥饿”情况,更适合实际系统的运营。

### 参考文献:

- [1] CHU G S, WANG D, MEI S L. A QoS architecture for the MAC protocol of IEEE 802.16 BWA system[C]// IEEE International Conference on Communications, Circuits and Systems and West Sino Expositions (ICCCAS'02). [S. l.]: IEEE Press, 2002, 1: 435 - 439.
- [2] HAWA M. Stochastic evaluation of fair scheduling with applications to Quality-of-Service in broadband wireless access networks[D]. Kansas: University of Kansas, 2003.
- [3] CHEN T. QoS issues in wireless packet network: term paper for research methodology[Z]. Italy: University of Trento, ICT School, 2006.
- [4] IEEE 802.16d. Draft IEEE standard for local and metropolitan area networks-part 16: air interface for fixed broadband wireless access systems[S].
- [5] VAARANDI R. A data clustering algorithm for mining patterns from event logs[C]// Proceedings of the 2003 IEEE Workshop on IP Operations and Management. [S. l.]: IEEE Press, 2003: 119 - 126.
- [6] WALDVOG M. Fast longest prefix matching: algorithms analysis and applications[D]. Zurich: ETH, Department of Electrical Engineering, 2002.
- [7] DUDA R O, HART P E, STORK D G. Pattern classification[M]. 2nd ed. Beijing: China Machine Press, 2004.
- [8] KARIM A, AHMAD I, JAMI S I. Cluster analysis of traffic flows on a campus network[C]// Proceedings of the 24th IASTED International Multi-Conference, Artificial Intelligence and Applications. Austria: ACTA Press, 2006: 416 - 421.
- [9] KRISHNAMURTHY B, WANG J. On network aware clustering of web clients[C]// Proceedings of the ACM SIGCOMM 2000. [S. l.]: ACM Press, 2000.
- [10] DUNHAM M H. Data mining: introductory and advanced topics[M]. Beijing: Tsinghua University Press, 2003.
- [11] FREITAS A A. Data mining and knowledge discovery with evolutionary algorithms[M]. Berlin: Springer-Verlag, 2002.
- [12] KRISHNA K, MURTY M N. Genetic k means algorithm[J]. IEEE Transactions on Systems, Man and Cybernetics, Part B, 1999, 29(3): 433 - 439.
- [13] ESTAN C, SAVAGE S, VARGHESE G. Automatically inferring patterns of resource consumption in network traffic[C]// Proceedings of the ACM SIGCOMM 2003. [S. l.]: ACM Press, 2003: 137 - 148.

(上接第1864页)

字节数等,这些属性的聚类同样可以为网络工程提供有意义的信息。

### 参考文献: