基于半监督聚类的Web流量分类

陆伟宙 余顺争
(中山大学电子与通信工程系 广州 510275)

摘 要 提出了一种基于半监督学习的方法对Web流量进行聚类分析,使用隐马尔可夫模型对用户流量进行描述和聚类分析。该方法通过对少量数据进行人工标识,利用已标识数据对无监督聚类结果进行调整,以得到与人工分类匹配的聚类结果。使用真实的Web流量对提出的方法进行验证,实验结果表明该方法能有效地对Web流量进行分类,并得到相应的描述模型。

关键词 半监督聚类,隐马尔可夫模型,Web流量

Web Traffic Classification Based on Semi-supervised Clustering

LU Wei-zhou YU Shun-zheng
(Department of Electronic and Communication Engineering, Sun Yat-Sen University, Guangzhou 510275, China)

Abstract This paper presented a Web traffic classification method based on semi-supervised clustering, which uses HMM (Hidden Markov Model) to model and analyze Web client traffic. The method first runs an unsupervised clustering process on the whole data set, and then uses the pre-labeled data to adjust the result clusters. The paper also presented the experiment result on real network data to validate the purposed method.

Keywords Semi-supervised clustering, Hidden markov model, Web traffic

1 引言

Web流量是目前互联网上的主要流量之一。对于同一个网站,用户往往呈现多样性:根据访问方式,可以分为直接访问和通过代理服务器访问;根据用户浏览模式,可以分为逐次点击浏览和多次点击再浏览等方式。用户的多样性使得其产生的流量也具有多样性的特点,对不同类型的用户的Web数据流进行区分和描述,对于优化Web服务,抵御网络攻击都有重要的作用。例如,Web用户中可能存在多种类型用户,包括普通用户、代理用户、搜索引擎(Web Robot,例如Google等公司的自动访问网页程序Googlebot)等,不同类型的用户对服务器的响应要求也不一样,通过对Web用户进行区分,可以根据用户的类型把用户的请求调度到不同的服务队列,以提高服务器的总体性能。另一方面,目前有不少攻击通过泛洪HTTP请求对网站进行攻击,通过对用户数据流进行分类建模,精细描述正常Web用户流量,有助于检测和过滤攻击流量。传统的Web流量分析[1],往往着重于总体流量的描述或者是流量的产生,对个体流量的描述的文献还比较缺乏。本文使用隐马尔可夫模型(Hidden Markov Model, HMM)[2]对Web用户的请求到达过程进行描述。HMM在模型训练、状态估计与或然概率计算等理论方面已经相当成熟,在语音识别、手写体识别、数字通信编解码、DNA序列分类等许多重要领域获得了广泛和成功的应用。我们之前的研究[3]表明,使用HMM可以对正常用户流量进行建模,并能对异常用户进行检测。

研究人员关注的是同类型用户的共同特性,这需要对大量用户的流量进行分类研究。一般而言,对于大规模样本分类,有3种不同的方法:第一种方法是基于全监督学习的方法,对所有样本记录进行人工分类标识再进行建模,这种方法不仅工作量巨大,而且依赖于分析人员对样本特性的理解;第二种方法是基于无监督学习的方法,使用聚类对样本进行分类,其结果依赖于算法的优劣程度,我们之前的研究使用HMM用户数据流进行分类[3]也属于无监督学习;第三种方法是基于半监督学习的方法[4-10],根据少量已分类标识的样本,对其余未标识的样本进行分类,这种方法一方面仅对部分样本进行标识减低了人工分类需要的工作量,另一方面利用部分有标识数据提高了分类的效率。对于Web用户流量分类而言,实际上难以对所有用户的流量进行人工分类,但对少量用户进行人工分类是可能的,因此半监督学习的方法较为适用。本文尝试使用半监督学习方法对Web用户数据流进行分类,通过对少量的Web用户数据流进行人工分类建立模型,再使用基于HMM的半监督聚类的方法对其余Web用户数据流进行分类,并得到各个类别的描述模型。

已标识数据在半监督聚类中的作用,可以分为作为聚类种子、作为聚类限制和作为聚类反馈3种。在已标签数据作为聚类种子的方法[5]中,初始聚类根据已标签的数据的分类结果生成,再进行标准的聚类操作;在已标签数据作为聚类限制的方法[6]中,修改聚类结果中已标识数据与聚类的对应关

到稿日期:2008-04-01 本文受国家高技术研究发展计划(863)资助项目(批准号:2007AA01Z449),国家自然科学基金-广东联合基金重点项目(U0735002),国家自然科学基金项目(90304011)资助。

陆伟宙(1980—),男,博士生,主要研究领域为计算机网络与通信;余顺争(1958—),男,教授,博士生导师,主要研究领域为计算机网络与通信。

系，以保证在聚类过程中已标识数据的标签不变；在已标签数据作为聚类反馈的方法[7]中，调整聚类结果，使得每个预设的标签对应一个聚类。Basu 等[5]使用前两种半监督聚类方法对文本进行聚类并进行了比较，认为已标签数据作为聚类限制的半监督聚类方法优于已标签数据作为聚类种子的方法。Zhong 等人[8]对 3 种方法都做了比较，认为当已标签数据中包含了总体数据所有类别的情况下，已标签数据作为聚类限制的半监督聚类方法是最优的；而当已标签数据中缺少总体数据中的某些类别的情况下，已标签数据作为聚类反馈的方法较为优胜。考虑到在对 Web 用户数据流进行人工标识时并不知道类别数目，本文使用已标签数据作为聚类反馈的半监督聚类方法[7,8]，对基于 HMM 的聚类算法和参数重估算法进行修改，以实现对 Web 用户数据流的分类。

本文第 2 部分首先对 Web 用户数据流的流量模型进行介绍，包括模型的建立以及参数的更新等，然后简要回顾无监督聚类方法，最后提出已标签数据作为聚类反馈的半监督聚类方法；第 3 部分给出使用本文的半监督聚类方法对实际流量进行聚类的实验结果；最后小结全文。

2 流量模型

2.1 HTTP 流量的批到达过程

在 Web 用户和服务器的交互过程中，用户发送一系列的 HTTP 请求并从 Web 服务器获取响应，如果将连续的时间划分成一个个小的时间单元（例如 1 秒），每个时间单元内到达的请求的个数可能是 0，也可能是大于等于 1 的整数，我们把在同一个时间单元内到达一个或多个请求的事件称为一个批到达事件，记录批到达事件之间的空闲时间，由此得到批到达事件的观测序列，如图 1 所示。假设从任意时刻开始，首个批到达事件开始的时间记录为 T_0，记录此后第 t 个批到达事件开始的时间为 T_t，该批到达事件中到达的包的个数为 r_t，则第 $(t-1)$ 个批到达事件开始的时间和第 t 个批到达事件开始的时间差 w_t 可以通过 $w_t=T_t-T_{t-1}$ $(t\geqslant 1)$ 计算。对于每一个访问网站的用户，其数据包批到达过程可以通过二维观测序列 $\{o_t=(w_t,r_t),t\geqslant 1\}$ 进行描述。

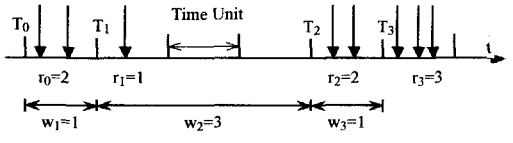


图 1 批到达与观测序列

实际网络中的 Web 服务器会对用户的每一个请求的到达时间、请求内容、处理情况等信息进行记录，生成相应的日志文件，因此上述的观测序列可以通过对服务器进行简单修改在线得到，也可以从 Web 服务器的日志文件中获取观测序列进行离线的模型训练。

2.2 描述模型

对于用户的包的批到达事件目前已经有一定的研究，可以认为它具有马尔可夫性，即将来的状态只与当前状态有关，与之前的状态无关。在这里，我们使用隐马尔可夫模型（Hidden Markov Model，HMM）[2]对批到达事件进行描述。我们假设每一个 Web 用户（包括单个用户和代理服务器）的到达流具有 M 个离散状态，分别表示为 $1,2,\cdots,M$，并记这些状态的集合为 S。状态转移关系可以用具有 M 状态的马尔可夫链来描述。令 $A=\{a_{mn}\}$ 是状态转移概率矩阵，它的元素 a_{mn} 代表从状态 m 到状态 n 的转移概率，$m,n\in S$。流的不同状态对应着不同的发包模式，产生的批到达事件有所不同，具体体现在批到达事件产生的时间间隔以及每个批到达事件中到达的包的个数上。给定状态 m，下一个批到达事件包含的包的个数为 q，与之前的批到达事件的时间间隔为 d 的概率为 $b(q,d|m)$，即

$$b(q,d|m)=P(r_t=q,w_t=d|s_t=m),d,q\in \mathrm{N},m\in S,\forall t \tag{1}$$

其中 N 为自然数集合，s_t 为在产生第 t 个批到达事件时该发送源所处的状态。显然 $b(q,d|m)$ 需要满足 $\Sigma_{q,d}b(q,d|m)=1$。定义给定状态 m 在单位时间内到达 q 个包的概率为 $b_r(q|m)$，其中 $q\in$ N，且满足 $\Sigma_q b_r(q|m)=1$。同时定义给定状态 m 相邻两个批到达事件的时间间隔 d 的概率为 $b_w(d|m)$，其中 $d\in$ N，且满足 $\Sigma_d b_w(d|m)=1$。假定给定状态下批到达包含的包的个数（主要决定于页面内嵌对象的个数与各级 cache 的缓存作用）和批到达时间间隔（主要决定于用户点击页面的时间间隔与网络时延）相互独立，即

$$\begin{aligned} b(q,d|m)&=P(r_t=q,w_t=d|s_t=m)\\ &=P(r_t=q|s_t=m)\cdot P(w_t=d|s_t=m)\\ &=b_r(q|m)\cdot b_w(d|m),d,q\in \mathrm{N},m\in S,\forall t \end{aligned} \tag{2}$$

为了减少整个模型的参数数量，我们进一步假定 $b_r(q|m)$ 和 $b_w(d|m)$ 为参数化的分布，根据我们对实际数据的观测，可以假定它们分别为 Poisson 和 Pareto 分布：

$$b_r(q|m)=P(r_t=q|s_t=m)=\frac{\mu_m^{q-1}}{(q-1)!}e^{-\mu_m},q\in \mathrm{N},$$
$$\mu_m>0,m\in S,\forall t \tag{3}$$

$$\begin{aligned} b_w(d|m)&=P(w_t<d+1|s_t=m)-P(w_t<d|s_t=m)\\ &=d^{-\lambda_m}-(d+1)^{-\lambda_m} \end{aligned}$$
$$d\in \mathrm{N},\lambda_m>0,m\in S,\forall t \tag{4}$$

其中 (μ_m+1) 是给定状态 m 时批到达包数的均值，λ_m 是 Pareto 分布的形状参数，且假定 $\mu_1\leqslant\mu_2\leqslant\cdots\leqslant\mu_M$，即 m 较大的状态，q 取较大值的概率大一些。由此，我们得到一个参数化的二维观测向量 HMM，并用它的参数的集合 $\Omega=\{A,\pi,\lambda,\mu\}$ 来代表该模型，其中 $\pi=(\pi_1,\pi_2,\cdots,\pi_M)'$ 是初始状态概率分布向量，$\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_M)'$，$\mu=(\mu_1,\mu_2,\cdots,\mu_M)'$。

2.3 半监督聚类

2.3.1 无监督聚类

在介绍半监督聚类之前，我们简要回顾基于模型的无监督聚类方法。假设待分类的用户数据流观测序列集合为 $O=\{O^{(i)},1\leqslant i\leqslant N\}$，需要通过基于 HMM 的聚类算法得到 K 个聚类及其描述模型，第 k 聚类对应的描述模型为 $\Omega^{(k)}=\{A^{(k)},\pi^{(k)},\lambda^{(k)},\mu^{(k)}\}$，其中 $1\leqslant k\leqslant K$。

如果指定聚类数目 K，基于 k-means 算法的无监督聚类算法可以描述如下：随机初始化 K 个聚类中心，计算样本到聚类中心的距离并把样本分配到最近的聚类中去，再根据各聚类中的样本重新计算聚类中心，反复调整各样本的归属聚类和聚类中心，直到聚类结果不发生变化为止。

在基于模型的聚类方法中，使用模型作为聚类中心，而序列相对于模型的或然概率体现了序列与模型匹配程度，可以作为序列与聚类中心距离的度量。或然概率可以通过前向-

后向算法(Forward-backward algorithm)[2]求得,或然概率越大表示对应序列与模型的匹配程度越好。使用矩阵 $Z=\{z_{lk}\}$,$1\leqslant l\leqslant N$,$1\leqslant k\leqslant K$ 标识训练序列属于的聚类,其中,$z_{lk}=1$ 表示第 l 个序列属于第 k 个聚类,$z_{lk}=0$ 表示第 l 个序列不属于第 k 个聚类。如果采用硬聚类的方法,只有产生最高或然概率的模型对应的 z_{lk} 才会等于 1,即

$$z_{lk}=\begin{cases}1, & \text{if } k=\text{argmax}_j(P(O^{(l)}|\Omega^{(j)}))\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

总或然概率可以通过式(6)进行计算

$$P(O|\Omega)=\prod_{l=1}^{L}P(O^{(l)}|\Omega^{(z_l)}) \tag{6}$$

其中,$\Omega^{(z_l)}$ 表示使得 z_{lk} 取值为 1 的模型。根据前向-后向算法,我们使用(7)－(12)式对模型进行更新,其中 $m,n\in S$,$1\leqslant k\leqslant K$,$\gamma_t^{(l)}(m)=\Pr[s_t=m|O^{(l)},\Omega]$ 和 $\xi_t^{(l)}(m,n)=\Pr[s_t=m, s_{t+1}=n|O^{(l)},\Omega]$ 都可以通过前向-后向算法[2,3]求得。

$$\hat{a}_{mn}^{(k)}=\frac{\sum_{l=1}^{L}z_{lk}\sum_{t=1}^{T-1}\xi_t^{(l)}(m,n)}{\sum_{l=1}^{L}z_{lk}\sum_{t=1}^{T-1}\gamma_t^{(l)}(m)} \tag{7}$$

$$\hat{\pi}_m^{(k)}=\frac{\sum_{l=1}^{L}z_{lk}\gamma_1^{(l)}(m)}{\sum_{l=1}^{L}z_{lk}\sum_{m=1}^{M}\gamma_1^{(l)}(m)} \tag{8}$$

$$\hat{b}_r^{(k)}(q|m)=\frac{\sum_{l=1}^{L}z_{lk}\sum_{t=1\ s.t.\ r_t=q}^{T}\gamma_t^{(l)}(m)}{\sum_{l=1}^{L}z_{lk}\sum_{t=1}^{T}\gamma_t^{(l)}(m)} \tag{9}$$

$$\hat{b}_w^{(k)}(d|m)=\frac{\sum_{l=1}^{L}z_{lk}\sum_{t=1\ s.t.\ w_t=d}^{T}\gamma_t^{(l)}(m)}{\sum_{l=1}^{L}z_{lk}\sum_{t=1}^{T}\gamma_t^{(l)}(m)} \tag{10}$$

$$\hat{\mu}_m^{(k)}=\sum_{q\geqslant 1}\hat{b}_r^{(k)}(q|m)(q-1) \tag{11}$$

$$\hat{\lambda}_m^{(k)}\approx 2*\left[\sum_{d\geqslant 1}\hat{b}_w^{(k)}(d|m)(\ln d+\ln(d+1))\right]^{-1} \tag{12}$$

综上所述,一种基于模型的无监督聚类算法可以用图 2 表示,其中,收敛条件可以是总或然概率变化很少或者达到一定的迭代次数。

无监督聚类算法

输入:序列集合 O,聚类个数 K,初始模型

输出:最优模型 $\Omega=\{\Omega^{(1)},\Omega^{(2)},\cdots,\Omega^{(K)}\}$,序列和模型对应关系矩阵 Z

算法:

1. 随机初始化 Z
2. 根据初始模型和 Z 初始化 K 个 HMM $\Omega=\{\Omega^{(1)},\Omega^{(2)},\cdots,\Omega^{(K)}\}$
3. 计算每个序列相对于各个 HMM 描述模型的或然概率,根据(5)式更新 Z,计算总或然概率
4. 根据 Z 和(7)－(12)式更新 K 个 HMM $\Omega=\{\Omega^{(1)},\Omega^{(2)},\cdots,\Omega^{(K)}\}$
5. 重复步骤 3 和步骤 4,直到 Z 不再发生变化或者达到收敛条件

图 2　无监督聚类算法

2.3.2　半监督聚类

在半监督聚类中,观测序列集合被分为已标识序列集合和未标识序列集合,其中前 N_L 个序列使用标签集合$\{1,2,\cdots,K_L\}$进行标识,这些序列构成已标签序列集合 $O_L=\{O^{(i)},1\leqslant i\leqslant N_L\}$,对应的标签构成标签序列 $Y_L=\{y_i,1\leqslant i\leqslant N_L\}$,其中第 i 个序列的标签为 y_i。对于已标签序列集合 O_L,根据其标签将已标签序列分为 K_L 个组,标签 k 对应的组记作 $G_k=\{O^{(i)}|y_i=k\}$,其中 $1\leqslant k\leqslant K_L$。观测序列集合 O 中其余 $N_U=N-N_L$ 个未标识序列构成未标签序列集合 $O_U=\{O^{N_L+i}, 1\leqslant i\leqslant N_U\}$,显然有 $O=O_L\cup O_U$。

本文采用的方法属于使用已标签序列作为聚类反馈的半监督聚类方法,在文献[8]中的方法的基础上发展而来,考虑了序列在聚类中移动的情况,采用了不同的聚类调整方法。这种方法首先使用无监督聚类的方法对整个观测序列集合 O 进行聚类,再使用已标签数据对无监督聚类的结果进行调整;利用调整后的聚类结果作为初始值进行聚类,在聚类过程中保持已标签数据的聚类不变,直到聚类结果不再发生变化或者达到其他收敛条件(例如总的或然概率变化很小)为止。在这种方法中,无监督聚类的作用是为了发现并未在已标识序列集合中出现的聚类,而使用已标签数据对无监督聚类的结果进行调整是为了充分利用人工分类的结果。图 3 是根据标签数据对聚类结果进行调整的例子,图中的小圆点表示待分类的样本,其中有 3 组共 9 个点已经被预先标识(分别用黑点、单阴影和双重阴影标识)。图 3(a)是使用无监督聚类得到的结果,由于无监督聚类的结果与聚类算法有很大的关系,聚类结果可能与预标签样本分布不匹配,如图 3(a)中单阴影点被分到两个不同的聚类中去,我们希望通过一定的调整,使得聚类的结果与预标签数据分布匹配,如图 3(b)所示。调整后的聚类应当满足:第一,已标签序列中同一个标签的序列应当属于同一个聚类;第二,同一个聚类中的已标签数据应当只有一个标签。我们称使用已标签数据对无监督聚类结果进行的调整称为标签-聚类绑定。

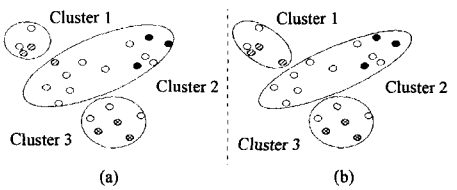


图 3　根据标签数据调整聚类

假设通过无监督聚类产生 $K(K\geqslant K_L)$ 个聚类,其中第 k 个聚类记作 C_k,并根据或然概率的大小把 N_L 个已标签序列划分到不同的聚类中去。为了得到从 K_L 个预设标签到 K 个聚类之间的映射,我们使用图 4 所示的方法进行标签-聚类绑定。

标签-聚类绑定算法

输入:已标签序列组集合$\{G_i\}$,其中 $1\leqslant i\leqslant K_L$ 无监督聚类得到的聚类集合$\{C_k\}$,其中 $1\leqslant k\leqslant K$

输出:每个已标签序列组对应的聚类 map(G_i),其中 $1\leqslant i\leqslant K_L$ 以及更新后的聚类集合$\{C_k\}$,其中 $1\leqslant k\leqslant K$

算法:

1. 计算各个标签组序列落在不同聚类中的频数,第 i 个标签组 G_i 的序列落在第 k 个聚类 C_k 的频数为 $\text{Freq}(C_k,G_i)=\sum_{l\in G_i}z_{lk}/|G_i|$,其中 $|G_i|$ 为 G_i 中序列个数
2. 对得到的频数进行降序排列,得到一个频数数列{Freq}。
3. 当频数数列{Freq}非空
   i. 提取频数数列中第一个元素 $\text{Freq}(C_{j'},G_{k'})$
   ii. 如果 $\text{Freq}(C_{j'},G_{k'})>1/K_L$,绑定 $C_{j'}$ 和 $G_{k'}$,将 $G_{k'}$ 中得序列移动到 $C_{j'}$ 中,记为 $\text{map}(G_{k'})=C_{j'}$,删除频数数列{Freq}中的 $\text{Freq}(C_{j'},*)$ 和 $\text{Freq}(*,G_{k'})$;否则,新建一个聚类 G_{K+1},把 $G_{k'}$ 中的序列都移动到聚类 C_{K+1} 中,记为 $\text{map}(G_{k'})=C_{K+1}$,并删除频数数列{Freq}中的 $\text{Freq}(*,G_{k'})$,修改 K=K+1
4. 检查所有聚类,如果有空聚类,删除该聚类,并令 K=K－1

图 4　标签-聚类绑定算法

图 4 所示的标签-聚类绑定算法实际上完成了已标签数

据中的 K_L 个标签和调整后的 K 个聚类的映射，并确保了每一个标签对应一个聚类，同时每一个聚类最多对应一个标签。此后，我们以调整后的聚类结果作为初始值进行聚类，聚类方法与无监督聚类方法类似，但在每一次的迭代过程中，保持标签与聚类的绑定关系，即根据绑定关系，把具有同一标签的已标签序列归于绑定的对应聚类。具体算法如图 5 所示（其中收敛条件可以是总或然概率变化很小或者达到一定的迭代次数）：

半监督聚类算法

输入：已标签序列集合 O_L，O_L 对应的标签序列 Y_L，标签集合{1，2，…，K_L}，未标签序列集合 O_U，初始模型，预设聚类个数 K

输出：最优模型 $\Omega=\{\Omega^{(1)},\Omega^{(2)},\cdots,\Omega^{(K)}\}$，序列和模型对应关系矩阵 Z

算法：

1. 使用无监督聚类的方法对所有序列进行聚类，得到 K 个聚类，其中第 k 个聚类记作 C_k，$1\leqslant k\leqslant K$，K 个描述模型 $\Omega=\{\Omega^{(1)},\Omega^{(2)},\cdots,\Omega^{(K)}\}$ 和序列与聚类的对应关系 Z
2. 执行标签－聚类绑定算法，得到更新后的聚类集合{C_k}，其中 $1\leqslant k\leqslant K$
3. 根据 Z 和(7)－(12)式更新 K 个 HMM $\Omega=\{\Omega^{(1)},\Omega^{(2)},\cdots,\Omega^{(K)}\}$
4. 计算每个序列相对于各个 HMM 描述模型的或然概率，根据(5)式更新 Z，计算总或然概率
5. 标签 k 从 1 到 K_L，将 G_k 中所有序列转移到聚类 map(G_k) 中去，修改 Z
6. 重复步骤 3－5，直到 Z 不再发生变化或达到收敛条件为止。

图 5　半监督聚类算法

3　实验结果

我们使用真实数据对本文提出的算法进行验证。我们利用中山大学官方网站（http://www.sysu.edu.cn）的日志文件作为实验数据。通常 Web 日志文件中包含每个请求的 IP 地址和请求时间，我们提取这些信息用于建模和聚类。实际上，该网站的日志文件采用了扩展日志格式，能提供用户请求的详细信息，包括请求时间、请求对象、请求方法、使用的用户代理等。我们通过这些额外信息对聚类结果进行分析。

3.1　实验数据的提取

我们提取了该网站从 2005 年 5 月 11 日零点到 2005 年 5 月 17 日中午 12 时共 6.5 天的数据，期间总的请求数目为 19560311 个。我们根据日志文件中的 IP 地址对用户进行区分，假定每个 IP 对应一个用户（尽管有可能多个实际用户利用同一个 IP 地址访问网络，我们把这些用户看作是代理用户），整理出每个用户对应的请求。我们过滤了部分访问量过少的用户，只保留了整个观测期间请求数超过 1000 个的用户，从中得到 2112 个会话的数据。我们以 1 秒为时间单元长度，根据日志文件中的到达时间，提取每个批到达事件中请求的个数以及每个批到达之间的间隔，得到了 2.1 节中所描述的观测序列，这 2112 个序列构成了总序列集合。

我们对其中的约十分之一的数据（217 个序列）进行分析，并对其进行分类，这 217 个序列构成了已标签序列集合，它们的一些统计信息如表 1 所列。我们把来自私有地址（172.16.*.*-172.31.*.*）的 100 个用户序列归为一类，这些序列都来自中山大学内部用户，标识为类别 1；把来自 A 类地址的 100 个用户序列归为一类，这些序列一般来自校外用户，标识为类别 2；把 IP 为国内一个搜索引擎的 17 个用户归为一类，它们的请求都带有同样的用户代理信息“sohu-search”，标识为类别 3。需要说明的是，来自中山大学内部的用户并不局限于已标签序列中的地址段；已标签的 A 类地址也没有包含所有的校外用户，实际上在总序列集合中来自中山大学以外的用户超过一半。

表 1　已标签序列分类

	描述	数目	批到达请求数均值(个)	批到达时间间隔均值(秒)
标签 1	来自 172.16.0.0/12	100	7.7924	56.3631
标签 2	来自 A 类地址	100	2.8342	23.6368
标签 3	来自同一搜索引擎	17	1.1972	44.7817

3.2　聚类分析

如 2.2 节所述，我们使用一个二维观测向量的 HMM 对流量进行描述，所有的模型的状态数都取 3，初始值 A 采用均匀分布，即 A 中的所有元素都取 0.33；初始的 π 为(0.33，0.33，0.33)′，初始的 $\lambda=(1,2,3)'$，初始的 $\mu=(1.5,1.5,1.5)'$。通过不同训练序列对初始模型进行训练，最终得到描述训练序列的相应模型。

我们预设聚类数目 K 为 3，并使用 2.3 节中的半监督聚类算法对序列集合进行聚类，最终得到 3 个聚类及其对应的聚类模型。其中聚类 1 对应原始分类的 2 号标签组，包含大部分的中山大学校内用户序列；聚类 2 对应原始分类中的 1 号标签组，集中了大部分校外用户序列，聚类 3 对应原始分类中的 3 号标签组，包含大量的搜索引擎产生序列。各个聚类的一些统计信息如表 2 所列，其中校外用户是指非搜索引擎的用户；各个聚类模型的参数如表 3 所列。从表 2 的统计信息和表 3 的模型参数可以看到，不同聚类的序列批到达请求个数均值有明显的差别；另一方面，尽管不同聚类序列的批到达时间间隔均值相近，但聚类模型的参数 λ 具有较大的差别。

表 2　半监督聚类结果

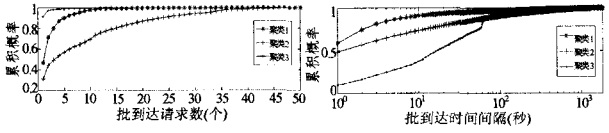
项目	聚类 1	聚类 2	聚类 3
对应标签组	标签组 2	标签组 1	标签组 3
序列数目	867	1047	198
批到达请求数均值(个)	2.6547	8.1697	1.6275
批到达时间间隔均值(秒)	60.0776	64.6126	60.0776
校内用户	19	822	58
校外用户	801	221	21
搜索引擎	47	4	119

表 3　聚类模型参数

	参数		
聚类模型 1 $\Omega^{(1)}$	$A=\begin{pmatrix}0.87 & 0.08 & 0.05\\ 0.12 & 0.82 & 0.06\\ 0.25 & 0.27 & 0.48\end{pmatrix}$	$\mu=(0.31\ \ 1.75\ \ 8.14)'$	$\lambda=(0.84\ \ 1.65\ \ 0.88)'$	$\pi=(0.96\ \ 0.02\ \ 0.02)'$
聚类模型 2 $\Omega^{(2)}$	$A=\begin{pmatrix}0.59 & 0.23 & 0.18\\ 0.46 & 0.35 & 0.20\\ 0.39 & 0.45 & 0.16\end{pmatrix}$	$\mu=(0.63\ \ 8.81\ \ 26.3)'$	$\lambda=(0.61\ \ 0.66\ \ 0.61)'$	$\pi=(0.81\ \ 0.10\ \ 0.09)'$
聚类模型 3 $\Omega^{(3)}$	$A=\begin{pmatrix}0.99 & 0.01 & 0.00\\ 0.51 & 0.40 & 0.09\\ 0.31 & 0.52 & 0.17\end{pmatrix}$	$\mu=(0.07\ \ 4.26\ \ 19.7)'$	$\lambda=(0.38\ \ 0.40\ \ 0.44)'$	$\pi=(0.90\ \ 0.06\ \ 0.04)'$

为了更好地对各个聚类进行分析，我们对各个聚类中的

序列的批到达请求个数和批到达之间的时间间隔进行概率统计，其结果如图 6 所示。其中，图 6(a)是批到达请求个数的累积分布，横坐标是请求个数；图 6(b)是批到达时间间隔的累积分布，其横坐标是批到达的请求间隔，单位是秒，以对数坐标显示。从图 6 可以看出，不同聚类的序列的批到达请求个数和批到达时间间隔具有不同的概率分布，而不同聚类的模型参数的差异也反映了这一点。



(a)各聚类序列批到达请求数的统计　(b)各聚类序列批到达时间间隔的统计

图 6　不同聚类序列的统计特性

聚类 1 的主体为非中山大学的用户，这一类用户的一个特点是批到达请求个数和批到达间隔偏小，根据进一步统计，聚类 1 的序列 90.1%的批到达只包含 5 个以下的请求，86.5%的批到达时间间隔少于 5 秒。从聚类模型参数来看，该聚类模型的 λ 值较大，说明批到达更可能在短时间内到达，而 μ 值较小表示每个批到达包含的请求也较小。

聚类 2 的主体为来自中山大学内部的用户，他们的流量特性与校外用户有明显的不同：一方面，聚类 2 的序列批到达中请求均值最大，有大量的批到达包含较多的请求，从描述模型的参数上和图 2 的统计可以看到这一点；另一方面，从表 2 对应模型不同的 λ 值相近，说明不同状态下批到达间的时间间隔的分布是类似的，对聚类 2 的序列的批到达时间间隔的统计也表明该类中的批到达时间间隔符合 Pareto 分布，Pareto 分布使得该聚类批到达时间间隔的均值偏大。中山大学官方网站校内用户与校外用户访问特性的明显差异，可能是两者对网站的使用习惯、内外网网络差异所致。在人工标识阶段，我们对使用私有地址(172.16.*.*-172.31.*.*)的用户进行标识，实际上来自中山大学内部的用户 IP 地址并不局限于这些地址，还包括另外的一些私有 IP 地址(如 192.168.*.*)和一些公有 IP 地址，通过半监督聚类之后，这些用户绝大部分(91.43%)都被归到聚类 2 中，说明这些用户的流量具有相似的性质。

聚类 3 包含了大量来自搜索引擎的用户，图 6(a)的统计可以看出这些用户的数据流批到达中的请求数最少，对应聚类描述模型的状态 2 和状态 3 的 μ 值偏大，但通过对状态跳转概率矩阵 A 分析可知出现状态 1 的概率远高于其他两个状态，即这些用户绝大多数时间都是低速发送请求。从图 6(b)可以看出，聚类 3 的序列的批到达时间间隔分布与其他聚类明显不同，在 20～55 秒内相对均匀，在 60 秒附近出现一个小峰，这些都与搜索引擎的行为类似。由于夹杂了非搜索引擎流量，批到达时间间隔在 10 秒内的概率偏大，但仍然比其他聚类的序列相应概率小很多。尽管我们只是使用其中一个搜索引擎的用户序列进行标识，但最终大部分的搜索引擎流量(70%)都被归入这一个聚类，说明这些搜索引擎虽然各自的实现方法不同，但对网站的访问模式还是有一定的相同点。搜索引擎的流量偏小，可能是因为该网站更新的页面不多，大多数搜索引擎在做完网页索引之后只会查询页面是否有更新，对更新的网页进行请求。

从上述对各个聚类的分析可以看出，对少量序列进行手工标识，通过半监督聚类的方法，可以实现对整个序列集合的分类，这些聚类都具有一定的物理意义。

结束语　Web 用户数据流的分类对刻画流量特性、提供分类服务和防御网络攻击都有重要的作用。本文提出了一种基于半监督聚类的 Web 流量分类方法，使用隐马尔可夫模型对 Web 用户数据流进行建模，通过对少量数据进行标识，实现对大量的用户数据流的分类，并得到相应的描述模型。本文使用该方法对中山大学官方网站的用户流量进行分类，实验结果表明该方法能有效地对流量进行区分并得到描述模型，不同类别的流量具有显著差异。

参 考 文 献

[1] Mah B. An empirical model of http network traffic // Proceedings of the INFOCOM'97. Vol 2. IEEE Computer Society Press, 1997:592-600

[2] Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition // Proc. of IEEE. 1989,77 (2): 257-286

[3] Lu wz, Yu sz. Clustering Web Traffic of Request Bursts // Proceedings of the IEEE TENCON. 2006:1-4

[4] 李和平，胡占义，吴毅红，等. 基于半监督学习的行为建模和异常检测. 软件学报，2004,18(3): 527-537

[5] Basu S, Banerjee A, Mooney R. Semi-supervised clustering by seeding // Proc. 19th Int. Conf. Machine Learning. 2002:19-26

[6] Wagstaff K, et al. Constrained k-means clustering with background knowledge // Proc. 18th Int. Conf. Machine Learning. 2001:577-584

[7] Cohn D, Caruana R, McCallum A. Semi-supervised clustering with user feedback. Technical Report TR2003-1892. Cornell University, 2003

[8] Shi Z. Semi-supervised model — based document clustering: A comparative study. Machine Learning, 2006, 65(1) 3-29

[9] 孙广玲，唐降龙. 基于分层高斯混合模型的半监督学习算法. 计算机研究与发展，2004,41(1):156-161

[10] Frigui H, Mahdi R. Semi-Supervised Clustering and Feature Discrimination with Instance-Level Constraints // Proceedings of IEEE International Fuzzy Systems Conference 2007. 2007:1-6