基于PCA和禁忌搜索的网络流量特征选择算法

冶晓隆 兰巨龙 郭 通
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要 针对网络流量特征属性选择的寻优和效率问题,提出了一种PCA结合禁忌搜索的网络流量特征选择方法。该方法通过PCA对高维特征属性空间进行特征约减,并利用禁忌搜索得到全局最优特征子集。实验证明,相比流行的遗传算法(GA)和粒子群寻优算法(PSO-SVM),PCA和禁忌搜索方法具有更好的处理效率和特征选择精度。

关键词 特征约减,特征选择,主成分分析,禁忌搜索
中图法分类号 TP393 **文献标识码** A

Algorithm of Network Traffic Feature Selection Based on PCA and Tabu Search

YE Xiao-long LAN Ju-long GUO Tong
(National Digital Switching System Engineering Technological R&D Center, Zhengzhou 450002, China)

Abstract A network traffic feature selection method using principal component analysis and tabu search(PCA-TS) was proposed for the purpose of the efficiency and quality when using fecture selection. This approach reduces high-dimensional features using PCA and gets the optimal feature subset on the basis of tabu search. Experiment shows that PCA-TS method has better efficiency and selection accuracy compared with GA and PSO-SVM.

Keywords Feature reduction, Feature selection, Principal component analysis(PCA), Tabu search

1 引言

各种新型网络业务的不断出现及网络异常的频繁发生,给网络安全、网络管理和网络服务质量带来巨大挑战。许多新技术如动态端口和端口伪装等的大量应用,使得传统基于流量特性和载荷分析进行流量分析的方法已经难以满足需求。基于特征分析的机器学习方法成为网络流量研究的热点,然而网络流量特征空间的"维数灾难"给网络分析带来困难。因此基于特征降维的最优特征子集选择具有十分重要的意义。

特征选择通过将原始流量特征空间进行降维,提取有用的特征组成一个新的特征空间来实现[1]。传统特征选择方法假设特征之间满足条件独立性或某种特定概率分布,这与实际网络流量特性不符。特征选择中寻找最小特征子集是NP完全问题[2],如何在特征子集质量和运算效率之间寻找一个平衡点即最优解,是特征选择的首要问题。目前特征选择方法主要包括两类,一类是转换法,如主成分分析(Principal Component Analysis, PCA)、独立成分分析(Independent Component Analysis, ICA)和流行学习算法等,这些方法都是将高维数据简单压缩到低维空间,无法提取最优特征子集;另一类是组合优化方法,主要有基于信息熵、遗传算法、粗糙集和神经网络等。文献[3]应用信息熵来度量特征贡献度,通过选择若干信息熵值较大的特征实现特征选择,但这种方法提取的特征无法精确表征网络流量。Berg等[4]使用贪心策略选择对数似然增益最大的特征获取特征几何,这种方法只能找到局部最优解并且计算量太大。Li等[5]利用遗传算法进行特征选则,此方法适合于大规模问题解决,但容易过早收敛且计算量较大。文献[6]利用粒子群算法实现高效特征搜索,但该方法容易陷入局部最优使得处理能力有限。文献[7]提出了一种基于互信息的主成分分析特征选择算法,将特征之间互信息矩阵的特征值作为评价准则,降低了线性降维对非线性特征关系的影响,但该方法对稀疏特征空间选择的精度较低。

真实网络环境中网络流量的高维特征属性空间中存在大量冗余和弱属性,并且高维数据对网络流量的分析带来巨大挑战,而现有特征提取方法存在特征提取质量和效率不高等局限。基于此,本文通过分析网络流量特征对流量表征的贡献度,提出了一种基于PCA和禁忌搜索(Principal Component Analysis and Tabu Search, PCA-TS)的网络流量特征提取方法,即通过PCA对高维特征进行约减,利用禁忌搜索得到全局最优特征子集。通过实验分析,验证了该方法的有效性。

2 基于PCA的禁忌搜索特征选择算法

2.1 PCA降维

网络流量特征属性空间的"维数灾难"严重降低了基于特

到稿日期:2013-03-09 返修日期:2013-06-29 本文受国家科技支撑计划课题(2012BAH02B01,2012BAH02B03),国家高技术研究发展计划(863计划)课题(2011AA01A103,2011AA01A101),国家高技术研究发展计划(863计划)基金资助项目(2011BAH19B04)资助。
冶晓隆(1987—),男,硕士生,主要研究方向为网络异常流量检测,E-mail:yexiaolong_2854@163.com;兰巨龙(1962—),男,教授,博士生导师,主要研究方向为宽带信息网;郭 通(1984—),男,博士生,主要研究方向为网络流量测量。

征分析方法的效率，而这些特征中存在大量的冗余和弱特征属性，需要通过特征约减来去除冗余和弱属性，得到精简特征属性空间。PCA 作为一种高效的降维算法，已经广泛应用于网络流量特征分析。

PCA 是统计学中数据分析的一种有效方法，主要用于特性抽取和数据降维。其思想是利用数据集统计性质的特征空间变换，将一个数据维数较高且互相关联的数据集进行降维处理。PCA 降维后，将原始空间转换为新的主成分空间，且各主成分互不相关。

假设含有 N 个样本的网络流量数据集 $X=\{x_1,x_2,\cdots,x_m\}\in R^n$，其中：$R^n$ 为特征空间，m 为特征维数。求得变量空间 $Z=\{z_1,z_2,\cdots,z_k\}$，满足 $k<m$ 且 $\mathrm{cov}(z_i,z_j)=0$，通过变换求得 k 个新变量 Z 可以代表 m 个原始变量 X 的大部分信息，即：

$$Z=\Sigma^{\mathrm{T}}X \tag{1}$$

式中，Σ 为一个 $m\times m$ 的正交矩阵，它是数据样本协方差矩阵 $C=\frac{1}{N}\sum_{i=1}^{N}(x_i-u)(x_i-u)^{\mathrm{T}}$ 的特征值矩阵，其中 $u=\frac{1}{N}\sum_{i=1}^{N}x_i$。因此转化为求解如下本征问题：

$$\lambda_i P=CP \tag{2}$$

式中，λ_i 为 C 的特征值，P 为相应的特征向量。

主成分分析通过选择贡献率较大的几个特征值 λ_i 对应的特征向量 P 作为主成分，以达到降维的目的。特征贡献率依下式计算：

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i}=\frac{\sum_{i=1}^{m}\lambda_i}{n}\geqslant R \tag{3}$$

式中，R 为特征贡献率阈值，特征维数 m 的选择根据 R 来确定，一般选择 R 为 85%～95%。

在使用 PCA 进行分析时，由于数据中不同的变量往往有不同的量纲，直接分析会引起各变量取值的分散程度差异较大，从而影响计算精度。为了消除量纲的不同可能带来的影响，首先需要对变量进行标准化处理，然后利用 PCA 进行降维。

2.2 禁忌搜索

禁忌搜索(Tabu Search，TS)最早由 Glover 于 1986 年提出，作为对局部搜索的扩展，它是一种启发式全局寻优搜索方法，其通过标记已搜索局部最优解和避免迭代计算中重复搜索来获得全局最优解[8]。TS 主要思想是：首先确定一个初始有效解 z，对每个解 z 定义一个邻域 $Y(z)$，从当前解的邻域中确定若干的候选解，从中选出最佳候选解。选择最佳候选解是一个搜索过程，为了避免搜索过程限于循环，TS 通过构造禁忌表和定义停止规则避免了搜索算法的局部最优。其中禁忌表存入前 n 次禁忌长度，避免了回到原先的解，从而提高了解空间的搜索能力；停止规则定义在若干迭代次数内最优解无法改进时算法停止。另外禁忌搜索算法中涉及邻域、禁忌表，禁忌长度、特赦规则和初始解等都会直接影响搜索优化结果[9]。

基于禁忌搜索的特征选择是通过目标函数进行约束的最优化问题，合适的目标函数提高了搜索和最优特征选择的质量[10]。一个好的特征解应在最少的特征数量上保证尽可能多的分类信息。在信息论理论中，一个属性的信息增益越大，其包含的信息量也越大。基于信息增益可以有效评估特征向量的分类信息，因此本文选择信息增益作为目标函数，定义目标函数如下：

$$G_T=\sum_{i=1}^{m}C(i)\times\frac{\sum_{i=1}^{n}G(A_j)}{n} \tag{4}$$

式中，$C(i)$ 为第 i 个样本是否被正确分类，m 为样本数目，$G(A_i)$ 为第 i 个特征的信息增益。通过式(4)确保特征以较少的特征数量保证最大的分类信息，选择除以 n 可以确保更快的禁忌搜索速度和避免过拟合。

禁忌搜索中初始解的选择对禁忌搜索的效果影响很大，在基于网络流量特征的最优特征选择中，由于实际网络流量特征维数较大，会影响禁忌搜索算法的效率，同时网络流量特征的冗余也对最优特征子集的选择产生影响，因此本文充分利用 PCA 可以对高维数据进行快速有效的特征约减的特点，通过剔除特征冗余和降低空间维数，提高禁忌搜索方法的最优解求解效率。

2.3 基于 PCA 的禁忌搜索特征选择算法

特征选择是从特征集 $C=\{c_1,c_2,\cdots,c_n\}$ 中选择一个子集 $C'=\{c_1',c_2',\cdots,c_n'\}$，$s'\leqslant s$。其中：$s$ 为原始特征空间大小，s' 为特性选择后新特征空间大小。即：从原始特征空间中选择部分有效特征，组成新的低维特征空间，其本质为一个寻优过程。为了获取最优特征子集，必须耗费大量搜索和计算时间进行穷尽搜索。因此需要利用寻优方法获取近优特征子集来减少计算时间。

网络流量的统计特征指的是在报文(packets)和流(Flow)的属性中，抽取和端口及协议无关的特征，如报文长度、报文到达间隔时间、报文数量、流的持续时间、流中报文个数等，这些统计特征用特征矢量来表示。如一条网络流 F，基于该流的特征描述可表示为 $F=\{y_1,y_2,\cdots,y_n\}$，其中 y_i 代表特征的取值。流的特征集合可能包含多达几百个特征，寻找少量最优特征子集来近似描述流量对提高学习效率等具有重要意义。

在基于网络流量特征的流量分析中，一般情况下，特征数量越大，会产生更高的分析精度。但实际中，过大的特征空间会产生两个问题：(1)巨大的特征空间不仅需占用更大的存储空间，而且增加了测量时间，难以应用于实时流量分析；(2)在部分应用中，如异常检测、业务分类等，不同的网络业务的表征需不同的特征属性向量，如果用全部特征去表征不同业务流，不仅降低了学习效果，而且增加了学习时间。所以特征选择就是挖掘最能有效刻画网络流量的最优特征集，禁忌搜索提供了一个近优解决方法。基于 PCA-TS 特征选择模型如图 1 所示。

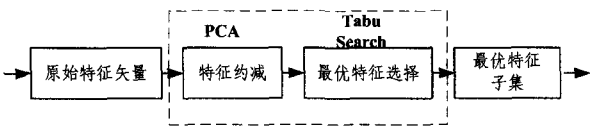

图 1 PCA-TS 特征选择模型

由于网络流量特征数量巨大且存在特征冗余，不能作为禁忌搜索的初始解，因此首先采用 PCA 对特征进行约减，去

除特征属性冗余，然后利用禁忌搜索得到最优特征子集。具体步骤如下：

(1)禁忌表置空，设置初始化参数：禁忌长度 $L_J=13$，最大迭代次数 $D_{max}=600$，最大改进次数 $\overline{D}_{max}=100$。

(2)使用 PCA 对原始网络流量特征进行约减，得到约减特征集 $T=\{T_1,T_2,\cdots,T_P\}$，p 为约减后的特征集数量。

(3)对特征集 T 进行二进制编码，得到初始解 R_{initN}。

(4)设置终止条件，当达到 D_{max} 时，搜索停止；当通过 $\overline{D}_{max}$ 寻找最优解无改进时，停止搜索。

(5)判断是否满足终止条件，如果满足终止条件，结束运算，输出最优特征子集，否则转到下一步。

(6)初始解 R_{initN} 带入邻域结构计算邻域解，通过目标函数选择最佳候选解。

(7)判断候选解是否满足特赦规则，如果满足，则更新禁忌表中最优解，转入步骤(4)；否则转到下一步。

(8)计算候选解的禁忌属性，选择非禁忌对象的最优值替换禁忌表的最初值，转入步骤(4)。

(9)结束，输出最优特征子集。

基于 PCA 的禁忌搜索方法在寻找最优特征集时，通过 PCA 对初始特征属性进行约减，提高了禁忌搜索的效率。同时利用禁忌搜索可以在全局特征空间寻找最优特征子集，避免了局部最优。该方法降低了特征冗余，提高了收敛速度，可以有效地进行特征选择。

3 实验分析

3.1 实验数据和环境

本文采用 Moore 等人在文献[11]使用的网络流量实验数据 Moore_Set 作为测试数据。Moore_Set 数据集是采用 3 个生物研究所共享的 1 条千兆全双工网络出口，其原始流量的记录集合包含了 2003 年 8 月 20 日 24 小时内流经该链路的双向网络流量，由于数据量庞大，Moore 等人对原始流量数据进行抽样，分别提取了 10 个流量子集，这 10 个子集的平均持续时间大约是 1680s，并且在抽取过程中只选取了语义完整的 TCP 双向流作为流量样本。

Moore_Set 数据集共包含 10 种类型的 377526 个网络流样本，其中每种网络流的数量和所占的比例见表 1。Moore_Set 数据集中每条网络流样本都包含 249 项属性，最后一项指明了该流的类型。

表 1 Moore_Set 数据集统计信息

流量类型	业务名称	流个数	比例%
WWW	Http, Https	328091	86.91
MAIL	Imap, Pop2/3, Smtp	28567	7.569
BULK	Ftp	11539	3.056
DB	Postgres, Sqlnet, Oracle, Ingres	2648	0.701
SER	X11, Dns, Ident, Ldap, Ntp	2099	0.556
P2P	Kazaa, Bit torrent, Gnutella	2094	0.555
ATT	Worm, Virus Attacks	1793	0.475
MULT	Windows Media Player, Real	1152	0.305
INT	SSH, Klogin, Rlogin, Telnet	110	0.029
GAMES	Half-life	8	0.002
共计	26 种业务类型	377526	100

本文采用的实验仿真硬件平台为普通 PC 机，具体配置为：CPU 为 Intel Core2 1.86GHz；内存为 2GB；操作系统为 Windows XP Professional SP3。实验仿真软件工具包括 Matlab 2008 和 Weka-3.6.8。

3.2 评价指标

为了分析所提方法的性能并与其他特征选择方法进行比较，使用 3 个评价指标：准确率(precision)、召回率(recall)和整体精度(O-accuracy)。其中准确率为网络流量实际分类数量中分类正确的数量；召回率为网络流量应有的分类数量中正确分类的数量；O-accuracy 表示被正确分类样本的比例，考虑的是算法的整体分类精度。具体定义如下：

$$presicion=\frac{N_{tp}}{N_{tp}+N_{fp}}\times 100\% \tag{5}$$

$$recall=\frac{N_{tp}}{N_{tp}+N_{fn}}\times 100\% \tag{6}$$

$$O\text{-}accuracy=\frac{\sum_{i=1}^{n}N_{tpi}}{\sum_{i=1}^{n}(N_{tpi}+N_{fpi})} \tag{7}$$

式中，N_{tp} 为类型为 A 的网络流量样本被分类模型正确分类的数量；N_{fn} 为类型为 A 的网络流量样本被分类模型错误分类的数量；N_{fp} 为类型为非 A 的网络流量样本被分类模型分类为类型 A 的数量。

3.3 结果分析

3.3.1 特征选择及必要性分析

为了研究网络流量的特征对流量分析的影响，本文对 Moore_Set 数据集(样本个数为 22932)分别取不同数量的特征进行特征分类，通过比较分类精度来分析特征数量对流量表征的影响，结果如图 2 和图 3 所示。

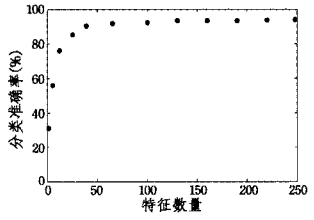

图 2 分类准确率和特征数量的关系

图 2 为 Moore_Set 数据集中样本个数最多的流量类型 WWW 的分类精度结果。由图中可以看出，随着特征数的增加，分类精度迅速提高，当特征数量增加到一定程度时，分类精度增加趋于平缓，说明网络流量可以通过较少的特征属性进行表征。

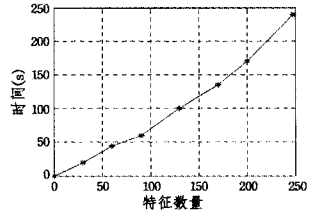

图 3 分类时间和特征数量的关系

图 3 表示取不同数量特征进行分类时计算时间的比较。为了避免分类算法的影响，本实验采用 Weka 工具集中的分类算法 NetBayes 进行分类，其中数据样本数量为 22932。由图 3 中可以看出，相同样本下分类的时间随着特征数量的增加而增加，而通过选择最优特征子集，不仅可以满足较高的分类精度，而且可以提高分类方法的效率。

本文使用 PCA-TS 方法对数据集特征选择后，从 248 个特征属性中提取出了包含 24 个特征属性的最优特征子集。如图 4 所示，其中横坐标为提取的特征属性，纵坐标为每个特征属性在数据集中所占的比例。

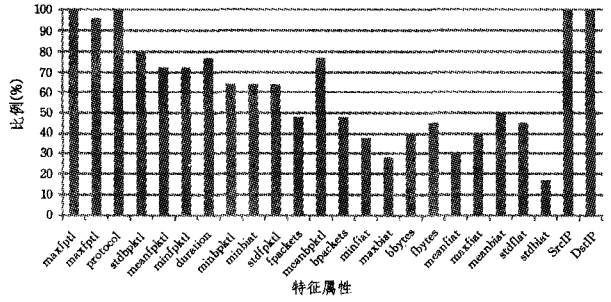

图 4　特征选择结果

3.3.2　特征选择结果分析

为了说明 PCA-TS 方法的有效性，本文选用流行的特征选择方法 GA[5]、PSO-SVM[6] 和本方法进行比较，同时为了说明 PCA-TS 的优势，引入 PCA 方法进行对比。首先分别采用上述 4 种方法对 Moore_Set 选择最优特征子集，采用 C4.5[12] 分类方法对最优特征子集的分类有效性进行比较，结果如图 5、图 6 所示。

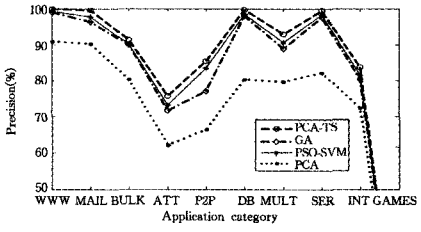

图 5　分类准确率比较

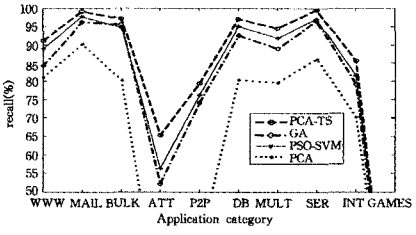

图 6　分类召回率比较

通过不同特征选择方法选择其最优特征集，使用 Moore_Set 对特征集进行分类并比较分类准确率和召回率。如图 5 和 6 所示，数据被分为 10 类（WWW，MAIL，BULK，ATT，P2P，DB，MULT，SER，INT 和 GAMES），从图中可以看出，PCA-TS 方法具有更高的分类准确率和召回率，相比 PCA 方法，其分类准确率和召回率更有大幅度的提升。分类整体精度如表 2 所列，PCA-TS 比传统 GA、PSO-SVM 等方法具有更高的分类整体精度，相对于 PCA 方法，通过禁忌搜索，其分类整体精度有 15% 的提升。这说明了 PCA-TS 方法能更好地选择最优特征子集。

表 2　特征分类的整体精度

Method	O-accuracy
PCA	83.66%
GA	96.72%
PSO-SVM	97.24%
PCA-TS	99.38%

在特征选择中，TS 和 GA 方法都是利用搜索来选择最优特征子集，其相对于特征维度的处理时间如图 7 所示。从图中可以看出，随着特征维度的增加，两种方法的分析时间都成倍数的增长，但当特征维度较低时，TS 方法具有明显的处理效率。而结合 PCA 首先对网络流量高维特征进行降维，不仅去除了大量特征冗余，而且 PCA 降维后使 TS 方法具有更好的处理效率，PCA-TS 方法相对于 GA 方法针对不同特征的处理时间变化如图 8 所示。

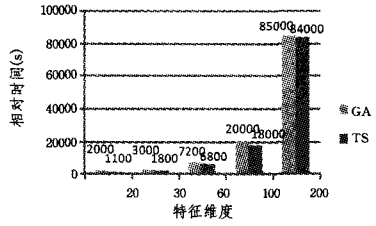

图 7　TS/GA 分析计算比较

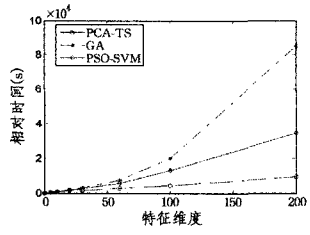

图 8　PCA-TS/GA 分析时间比较

从图 8 中可以看到，相比 GA 和 PSO-SVM 方法，PCA-TS 方法相对于特征维数变化时其处理时间增加较为缓慢，主要在于 PCA-TS 方法首先对高维数据进行特征约减，为禁忌搜索提供有效的低维初始特征。相比禁忌搜索，PCA 约减所耗时间可以忽略，因此该方法具有更高的效率，可以有效地应用于网络流量的高维特征选择。PCA-TS 方法结合了主成分分析和禁忌搜索的优点，首先通过 PCA 可以对高维特征空间进行有效的约减，去除了大量冗余，为禁忌搜索提供了较优的初始解，同时结合禁忌搜索避免“局部最优”的特点，确保了全局最优解的选择。

结束语　针对网络流量特征属性的“维数灾难”给流量分析带来的处理效率的降低，通过选择近优特征子集对高维网络流量特征属性空间进行降维。而现有特征选择方法具有较高的时间复杂度和较低的提取精度，本文提出了一种基于 PCA 和禁忌搜索的特征选择方法，首先利用 PCA 进行高维特征约减，为禁忌搜索提供了低冗余和低维的特征矢量，提高了特征选择的质量和处理效率。实验表明，该方法具有较高的特征选择精度，并且对于高维数据具有更好的处理效率。而禁忌搜索的相关参数会直接影响其搜索效率，因此分析网络流量特征提取中禁忌搜索参数的自适应选择对寻优搜索的影响，将是下一步研究的重点。

参 考 文 献

[1]　Faivishevsky L, Gold berger J. Unsupervised feature selection based on non-parametric mutual information [C] // Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on. IEEE, 2012: 1-6

[2] Davies S, Russl S. NP-completeness of searches for smallest possible feature sets [C]// Proceedings of the AAAI Fall 94 Symposiums on Relevance. Menlo Park, 1994: 37-39

[3] Lakhina A, Crovella M, Diot C. Mining anomalies using traffic feature distributions [J]. ACM SIGCOMM Computer Communication Review, ACM, 2005, 35(4): 217-228

[4] Berge A L, Pietra S D, Pietra V D. A maximum entropy approach to natural language processing[J]. Computational Linguistic, 1996, 22(1): 39-71

[5] Li Yong-ming, Zhang Su-juan, Zeng Xiao-ping. Research of multi-population agent genetic algorithm for feature selection [J]. Expert Systems with Applications, 2009, 36(7): 11570-11581

[6] Huang Cheng-long, Dun Jian-fan. A distributed PSO-SVM hybrid system with feature selection and parameter optimization [J]. Applied Soft Computing Journal, 2008, 8(4): 1381-1391

[7] 范雪莉,冯海泓,原猛. 基于互信息的主成分分析特征选择算法[J]. 控制与决策,2012,28(6):915-919

[8] Wang Yong, Li Lin, Ni Jun, et al. Feature selection using tabu search with long-term memories and probabilistic neural networks [J]. Pattern Recognition Letters, 2009, 30(7): 661-670

[9] Marinake M, Marinakis Y, Doumpos M, et al. A comparison of several nearest neighbor classifier metrics using Tabu Search algorithm for the feature selection problem [J]. Optimization Letters, 2008, 2(3): 299-308

[10] Tahir M A, Smith J. Improving nearest neighbor classifier using Tabu Search and ensemble distance metrics [C]// Proceedings of the Sixth International Conference on Data Ming (ICDM'06). Hong Kong, China: IEEE Computer Society, 2006: 1086-1090

[11] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques [J]. ACM SIGMETRICS Performance Evaluation Review, ACM, 2005, 33(1): 50-60

[12] 徐鹏,林森. 基于 C4.5 决策树的流量分类方法[J]. 软件学报,2009,20(10):2692-2704

(上接第 167 页)

[2] Bisgin H, Agarwal N, Xu X. Investigating homophily in online social networks[C]// Proceedings of 2010 IEEE/WIC/ACM International Conference on Web Intelligence. WI, 2010: 533-536

[3] Canny J. Collaborative filtering with privacy via factor analysis [C]// Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland, 2002: 238-245

[4] Caverlee J, Liu L, Webb S. Socialtrust: tamper-resilient trust establishment in online communities[C]// Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries. Pittsburgh PA, PA, USA, 2008: 104-114

[5] Deshpande M, Karypis G. Item-based top-N recommendation algorithms[J]. ACM Transactions on Information Systems, 2004, 22(1): 143-177

[6] Golbeck J. Computer science-Weaving a Web of trust [J]. Science, 2008, 321(5896): 1640-1641

[7] Golbeck J. Computing And Applying Trust In Web-Based Social Networks[D]. University of Maryland, USA, 2005

[8] Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks[C]// Proceedings of the Fourth ACM Conference on Recommender Systems. Barcelona, Spain, 2010: 135-142

[9] Jamali M, Ester M. TrustWalker: a random walk model for combining trust-based and item-based recommendation[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 397-406

[10] Karypis G. Evaluation of Item-Based Top-N Recommendation Algorithms[C]// Proceedings of the tenth international conference on Information and knowledge management. Atlanta, Georgia, USA, 2001: 247-254

[11] Kuter U, Golbeck J. SUNNY: a new algorithm for trust inference in social networks using probabilistic confidence models[C]// Proceedings of the 22nd national conference on Artificial intelligence. Columbia, Canada, 2007: 1377-1382

[12] Ma H, King I, Lyu M R. Learning to recommend with social trust ensemble[C]// Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. Boston, MA, USA, 2009: 203-210

[13] Massa P, Avesani P. Trust-aware recommender systems [C]// Proceedings of the 2007 ACM conference on Recommender systems. Minneapolis, MN, USA, 2007: 17-24

[14] McPherson M, Smith-Lovin L, Cook J M. Birds of a Feather: Homophily in Social Networks[J]. Annual Review of Sociology, 2001, 27(1): 415-444

[15] Moghaddam S, Jamali M, Ester M, et al. FeedbackTrust: using feedback effects in trust-based recommendation systems[C]// Proceedings of the third ACM conference on Recommender systems. New York, New York, USA, 2009: 269-272

[16] Niu S, Guo J, Lan Y, et al. Top-k learning to rank: labeling, ranking and evaluation[C]// Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. Portland, Oregon, USA, 2012: 751-760

[17] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]// Proceedings of the 10th International Conference on World Wide Web. Hong Kong, China, 2001: 285-295

[18] Sztompka P. Trust: A Sociological Theory [M]. Cambridge: Cambridge Univ. Press, 1999

[19] Xing X, Zhang W, Jia Z, et al. Learning to recommend top-k items in online social networks[C]// Proceedings of the 2012 World Congress on Information and Communication Technologies. WICT 2012, India, 2012: 1171-1176

[20] 周涛. 个性化推荐的十大挑战[J]. 计算机协会通讯,2012,8(7): 48-61