面向“人人网”的用户信息采集及拓扑结构测量研究

陈兴蜀，尹雅丽，李 卫，王文贤，王海舟

(四川大学计算机学院 成都 610065)

【摘要】以“人人网”为例，研究社交网站数据采集技术，并对其网络拓扑结构进行详细研究。结果表明：1) “人人网”的节点度分布不同于一般社交网络符合的幂律分布，更倾向于具有指数分布特征，且其度分布具有一定的重尾特性，在小范围内出现了类似小变量饱和现象，并且出现“双峰”现象；2) “人人网”符合小世界特性；3) “人人网”具有同配性，节点度高的节点倾向于与高度节点连接；4) 用户状态数、照片数和访客数没有明显的正相关特性。研究成果对于进一步了解社交网络的拓扑结构特征具有重要意义，为后续实现资源监管、跨社交网站的数据挖掘奠定了基础。

关 键 词 主动测量； 聚集系数； 网络拓扑结构； 小世界网络

中图分类号 TP393.08 文献标志码 A doi:10.3969/j.issn.1001-0548.2015.06.023

Measurement Study of Topologies Characteristics for “Renren” Social Networking System

CHEN Xing-shu, YIN Ya-li, LI Wei, WANG Wen-xian, and WANG Hai-zhou

(College of Computer, Sichuan University Chengdu 610065)

Abstract In this paper, taking as “Renren” for example, the social networking site’s data collection technology is studied. The collected data is used to study “Renren”s topological structure. The results show that, 1) different from general social networks’ power-law distribution, the node degree distribution of “Renren” tends to follow an exponential distribution; “Renren”s degree distribution has some heavy-tailed feature, and there is a saturation phenomenon of small variables on a small scale; it also presents the “double peak” phenomenon; 2) “Renren” has a smaller average shortest path length and a larger clustering coefficient, which means the small world characteristics; 3) “Renren” shows the assortativity, which means the node with high degrees is inclined to connection to the nodes with high degree; 4) No obvious positive correlation is found in status number, photos number and the visitors number of “Renren” users. The results are of great significance for the further understanding of the “Renren” and other social networks’ topology structure, and they will lay a foundation for resources supervision and cross-social network site’s data mining.

Key words active measurement; clustering coefficient; network topology; small-world networks

随着互联网技术的快速发展，以新浪微博、腾讯QQ空间、“人人网”等为代表的社交网络发展壮大，吸引着越来越多的用户。但是由于这一类“自媒体”普泛化、传播快等特点，使得一些不法分子有机可乘，利用社交网络发布不良信息，对网民造成不良的引导作用。其中以“人人网”为代表的社交网站具有以下特点：1) 用户群虽然在近年来扩展到每一个人，但是主要用户仍为大学生，由于这类群体的特殊性，容易被不法分子利用；2) 具有传播快、用户多等特点；3) 现有的社交网站只有针对浏览对象(包括隐私设置访问、浏览权限)的安全设置，没有针对内容安全的审查机制。因此对发布信息的采集以及其合法性检测对于舆情监控和信息安全等都具有十分重要的意义。本文以“人人网”为例，分析其网站结构特点，研究社交网站的数据采集技术，以网络爬虫为基础，通过设计主题网络爬虫，实现对特定网页的定向抓取，用正则表达式匹配出所需信息，存入数据库，用于后续分析，进而实现对网站的监管。本文实现了“人人网”数据采集系统，并分析了“人人网”的网络拓扑结构。

目前，国内外针对社交网站的研究主要集中在社交网络的拓扑分析[1]、用户行为特征分析[2-3]、社交网络中的信息传播、安全隐私问题、网络拓扑演化模型[4]、用户影响力度量以及社交网络盈利模式

收稿日期：2014 – 10 – 12；修回日期：2015 – 03 – 15

基金项目：国家科技支撑计划(2012BAH18B05)；国家自然科学基金(61272447)

作者简介：陈兴蜀(1969 –)，女，教授，博士生导师，主要从事信息安全、云计算安全等方面的研究.

研究等方面。文献[5]从测量角度对在线社会网络的拓扑结构、用户行为和网络演化等方面进行了综述，总结了常见的测量方法和典型的网络拓扑参数，着重介绍了用户行为特征、用户行为对网络拓扑的影响以及网络的演化。文献[2]基于“人人网”用户主页的行为记录数据，对个体行为和群体互动行为的时间统计特性进行实证研究；并针对“人人网”群体互动行为设计了社交驱动系数影响下的兴趣驱动模型。文献[3]分析社交网络中的用户行为，总结出了SNS中的用户行为图谱，研究了社交网络中的用户影响力模型。文献[6]把OSN的聚类系数与用不同算法生成的网络聚类系数进行对比，发现OSN的聚类系数要远大于理论模型的聚类系数。文献[7]通过对Facebook的用户交互，提出社会关系加强模型来量化人际关系指标。

相比以上的研究工作，本文采用申请应用的方式获得了“人人网”提供的API，通过调用API接口快速、高效地获取用户的完整的好友关系，该方法能有效解决通用网络爬虫抓取信息时存在的数据采集不完整等问题。基于采集的数据，对“人人网”网络拓扑结构进行详细研究，包括了“人人网”网络拓扑的聚集系数、同配系数、平均最短路径长度、平均度和度分布和小世界特性。本文的研究成果对进一步分析社交网络的用户行为、网络拓扑结构具有重要意义，为跨社交网站的数据挖掘研究奠定了良好基础。

1　“人人网”数据采集系统

通过对“人人网”网站结构分析，发现该网站用户个人资料(包括基本信息、学校信息、联系方式)，用户好友关系及用户状态等数据具有重要价值。而获取这些信息首先需要用户ID，然后根据ID采集每个用户的信息。其次，针对话题，一般以话题标题的小写字母表示，并作为该话题的唯一标识，因此采集前需要采集话题的名字，然后根据该名字(话题ID)采集该话题的具体内容和评论。该系统由用户ID和用户好友关系采集模块、用户个人资料和状态采集模块、话题ID采集模块、话题评论内容采集模块和数据存储模块5部分组成。

2　用户行为分析

2.1　用户主页信息统计分析

本文统计了两个数据集的好友数和访客数的关系，数据集1(data1)：目前采集到的所有数据中好友数在1～1 000的107 567个用户；数据集2(data2)：不限制好友数的112 454个用户。统计其好友数和访客数的关系，数据集2的统计结果如图1所示。

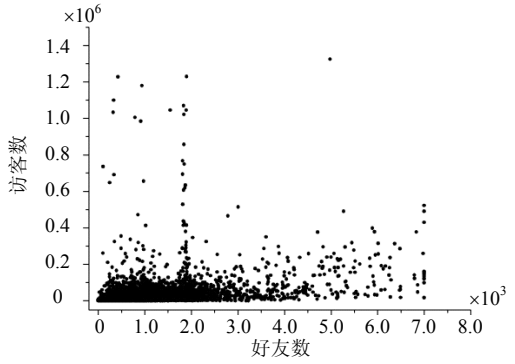


图1　用户好友数和访客数关系

从图1可以看出用户好友数在0～2 500时，访客数主要集中在10 000以下，当用户好友数大于2 500时，访客数分布无特定规律。数据集1的统计结果和图1类似，该数据集主要用于与文献[2]的测量结果进行对比。分析图1可知，用户好友数和访客数没有明显的正相关特性，而文献[2]通过统计272个好友数在1～1 000的用户的好友数和访客数的关系，发现其存在一定的正相关特性。本文得出的结论与文献[2]不同，可能的原因是文献[2]的数据集太小，导致得到了完全不同的相关特征。另外，对用户状态数、照片数和访客数也进行了统计，发现统计结果都集中在一个范围内，没有如文献[2]所显示的明显的正相关特性。

2.2　用户行为特征分析

对爬取的114 034个用户ID进行统计，其中由于用户设置了权限，或者账号已被注销等因素无法获取主页数据有的ID有1 088个，占总用户的0.954%。采集用户个人资料时，对5 612个数据集进行统计，发现填写了个人资料的用户只有237个，只占总数的4.223%，说明大多数用户不愿意公开自己的隐私信息，而这部分信息恰好是利用价值最高的，对于这部分数据，“人人网”可以采取一定的激励机制，促使用户完善信息，用户个人资料的完整性是后续数据挖掘中的关键。

3　网络拓扑结构分析

3.1　节点度和度分布

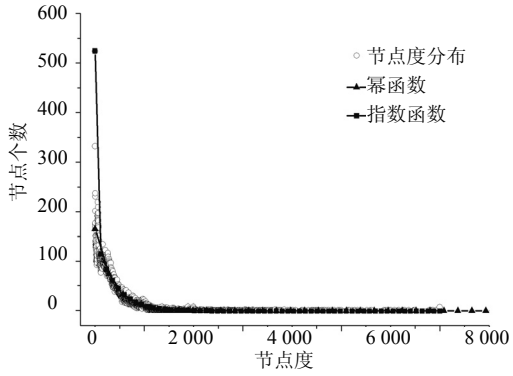
节点度是指与该节点相关联的边的条数[4]。在现实网络中，两种节点度分布比较常见：一种是指数分布，另一种是幂律分布[1]。“人人网”中节点度是指某一用户的好友个数。为描述“人人网”的度

分布，本文使用指数函数和幂律函数对几组数据集进行拟合。拟合函数为：

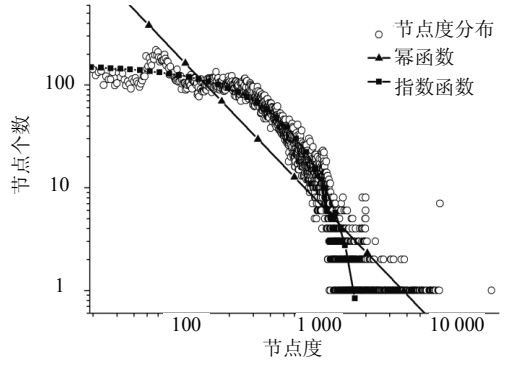
$$y = ax^b \tag{1}$$

$$y = A_1 e^{-\frac{x}{t_1}} + A_2 e^{-\frac{x}{t_2}} + A_3 e^{-\frac{x}{t_3}} + y_0 \tag{2}$$

式中，式(1)为幂函数，式(2)为指数函数。对4组数据(分别含有31 746个节点、57 733个节点、79 594个节点和85 010个节点)进行拟合，选取其中两组拟合结果如图2、图3所示。

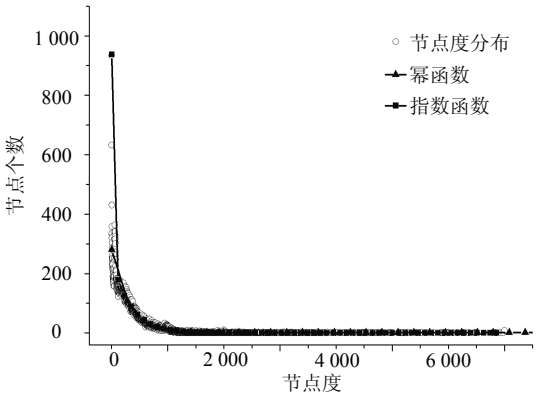


a. 线性坐标

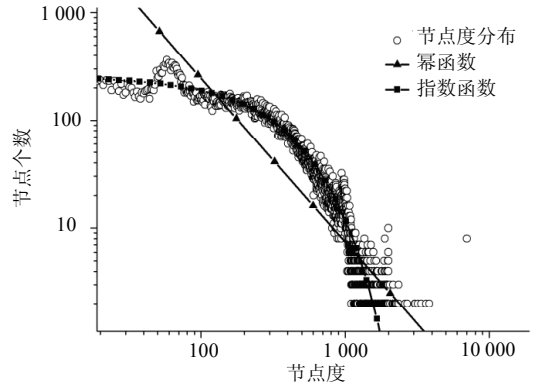


b. 双对数坐标

图2 57 733个节点的度分布图



a. 线性坐标



b. 双对数坐标

图3 85 010个节点的度分布图

对4组数据进行幂函数和指数函数拟合得到的参数分别如表1、表2所示。

表1 幂函数拟合效果和参数表

数据组别	节点个数	幂函数参数值($y = ax^b$)		
		R^2	a	b
data1-1	31 746	0.854 97	1.303 41	106.907 33
data1-2	57 733	0.904 2	–0.371 21	165.975 34
data1-3	79 594	0.885 77	1.905 41	268.087 37
data1-4	85 010	0.890 63	1.612 62	280.155 61

表2 指数函数拟合效果和参数表

数据组别	节点个数	指数函数参数值($y = A_1 e^{-\frac{x}{t_1}} + A_2 e^{-\frac{x}{t_2}} + A_3 e^{-\frac{x}{t_3}} + y_0$)							
		R^2	y_0	A_1	t_1	A_2	t_2	A_3	t_3
data1-1	31 746	0.928 7	0.606 2	–7.897 7	21.600 5	100.66	305.055 1	470.582 8	1.741 32
data1-2	57 733	0.953 1	–1.205 0	797.559 2	0.623 6	314.676 1	2.420 8	157.545 8	382.614 7
data1-3	79 594	0.956 8	0.278 7	108.455 3	316.332 9	137.778 4	316.392 0	1 015.381 0	2.068 8
data1-4	85 010	0.959 8	–0.173 6	1178.778 7	0.829 3	431.733 0	3.671 5	256.138 1	328.372 7

由表1、表2可以看出：对数据进行幂函数拟合得到的拟合优度分别为：0.854 97、0.904 2、0.885 77、0.890 63；对数据进行指数函数拟合得到的拟合优度分别为：0.928 65、0.953 12、0.956 75、0.959 84。从两个表的数据可以看出，随着节点数的增多，拟合优度值R^2有增长的趋势。但是数据经过幂函数拟

合得到的拟合优度在0.9附近，最高的也只有0.904 2；而经过指数函数拟合得到的拟合优度R^2都大于0.9，且随着节点数的增多，越接近1。说明“人人网”的节点度分布符合幂律分布的程度比较低，它更倾向于符合指数分布，且呈指数衰减趋势。

从图中可以看出，在双对数坐标系下，“人人网”节点度分布具有幂律分布的重尾特征，但是幂律分布的程度比较低。且图中出现了类似小变量饱和现象，即网络中较小强度节点的强度分布是接近饱和的。并且在小范围内形成了类似文献[4,8]所发现的“双峰”现象。文献[8]提出，这种新的多峰分布对网络可靠性有一定的影响，更统一的连接分布可能会保存网络处理随机节点故障的能力，减少对高度连接节点的依赖性。

3.2　聚集系数

集聚系数(clustering coefficient)用于描述一个节点邻居之间的相互连接的紧密程度，即网络的集团化程度，是网络拓扑的另一个重要参数[4]。节点i的簇系数c_i描述的是网络中与该节点直接相连的节点之间的连接关系，即与该节点直接相邻的节点间实际存在的边数目占最大可能存在的边数的比例，C_i的表达式为$C_i=\frac{2e_i}{k_i(k_i-1)}$，式中$k_i$表示节点$i$的度，$e_i$表示节点$i$的邻接点之间实际存在的边数，网络的聚集系数$C$为所有节点聚集系数的算术平均值，计算公式为[1]：

$$C=\frac{1}{N}\sum_{i=1}^{N}C_i \tag{3}$$

式中，N为网络节点数。对“人人网”的聚集系数进行计算，得到表3所示的结果，节点数131、281、526、1 078、2 383的聚集系数分别为：0.689、0.649、0.513、0.313、0.259。可以看出，随着数据集的增大，聚集系数有降低的趋势，但是整体水平仍然比较高。文献[9]指出，网络同时具有较小的平均路径长度和较大的集聚系数，这类网络称为小世界网络。因此，聚集系数也是体现小世界特性的一个参数。

表3　“人人网”用户平均最短路径长和聚集系数比较

数据集	节点个数	边数	平均最短路径长	聚集系数
data2-1	131	1 657	2.032	0.689
data2-2	281	7 004	2.046	0.649
data2-3	526	7 643	3.622	0.513
data2-4	1 078	8 884	7.041	0.313
data2-5	2 383	10 058	4.946	0.259

3.3　小世界特性

小世界特性是指一个网络如果它具有较短的平均路径长度(有文献指出同时具有较大聚集系数[9])，那么这个网络称为小世界网络[10]。平均最短路径长度是指网络中所有节点对之间最短路径的平均值，通常以节点间的跳数作为度量来计算，平均最短路径长度的计算公式为[4]：

$$l=\frac{1}{N(N-1)}\sum_{i\neq j}d_{ij} \tag{4}$$

式中，N为网络中节点个数；d_{ij}为节点i和节点j之间的最短路径长度。

对“人人网”的几组数据的平均最短路径长度进行计算，得到表3所示结果。

表4　小世界特性对照表

不同网络	参数	节点个数				
		131	281	526	1 078	2 383
“人人网”	聚集系数	0.689	0.649	0.513	0.313	0.259
	平均最短路径	2.032	2.046	3.622	7.041	4.946
	边数	1657	7 004	7 643	8 884	10 058
ER网络	聚集系数	0.102 2	0.088 5	0.028 2	0.008 9	0.001 6
	平均最短路径	2.153 1	2.004 9	2.645 7	3.525 1	5.503 3
	边数	840	3 512	3 767	4 521	5 098

表3分别代表由131、281、526、1 078、2 383个节点的平均最短路径长度和聚集系数，其中平均最短路径长度分别为：2.032、2.046、3.622、7.041、4.946，最大为7，最小为2。由上述结果可看出，随着节点数的增多，平均最短路径长有升高的趋势，但是最高也在6～7左右，说明“人人网”在一定范围内符合六度分隔理论，且数据集越大，越接近真实水平。表4表示在相同节点数和平均顶点度的情况下随机网络(ER: random network)和“人人网”的聚集系数、平均最短路径长和边数对照情况。由表4中数据可以看出，同样情况下，“人人网”和随机网络(ER)的平均最短路径长差别不大，但是“人人网”的聚集系数要远远大于随机网络的聚集系数。由此说明，“人人网”具有较小平均最短路径长和较大聚

集系数，“人人网”符合小世界特性。

另外，从边数增长的速度来看，随着节点数的增多，边数增长越来越缓慢，说明“人人网”中用户呈一定的社团化。社团化是指一组节点，这组节点构成一个连通子图，它们之间的连接要密于它们与外界节点的连接[5]。即用户更倾向于与一个范围内的人联系，而范围之间联系就不那么密切。

3.4 同配性

无标度性质和同配性说明社交网络中有一些紧密连接的度较大的核心，它们把整个网络连接起来，度较小的节点分布在网络的边缘[5]。分析网络的同配性，对于揭示网络自身组织结构与形成机制有着重要意义，也可以进一步量化社交网络的度相关性。同配性的计算公式为：

$$r=\frac{M^{-1}\sum_i j_i k_i-\left[M^{-1}\sum_i \frac{1}{2}(j_i+k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2+k_i^2)-\left[M^{-1}\sum_i \frac{1}{2}(j_i+k_i)\right]^2} \quad (5)$$

式中，j_i 和 k_i 分别为第 i 条边的两个端点的度，$i=1,2,\cdots,M$，M 为网络边数；$-1\leqslant r\leqslant 1$。该系数描述网络中的节点和与其度相同的节点连接的倾向性；若 $r>0$，网络是同配的(assortative)，表示节点倾向于和与其度相同的节点连接；若 $r<0$，网络是异配的(disassortative)，表示节点倾向于和与其度相异的节点连接。

文献[11]测得“人人网”同配系数为0.15。本文通过4组数据集计算“人人网”的同配系数，结果如表5所示。

表5 同配系数

数据集	节点数	同配系数
data3-1	182	0.833 3
data3-2	669	0.666 7
data3-3	3 233	0.536 2
data3-4	5 985	0.541 9

由表5结果可以看出，“人人网”同配系数 $r>0$，最大为0.833 3，最小为0.541 9，说明“人人网”是同配的。也从另一个方面反映了“人人网”的无尺度特性。随着数据集越大，同配系数有减小的趋势，但是越来越接近真实水平。

4 总 结

本文分析了“从网”网站特点，探索了用户模拟登陆过程，为了解决好友关系获取的完整性，通过调用“人人网”API的方式获得完整好友关系；设计并实现了数据采集系统，采集用户主页数据、好友关系、用户状态、个人资料、话题ID和话题评论。详细研究了网络拓扑结构，包括“人人网”网络拓扑的聚集系数、同配系数、平均最短路径长度、平均度和度分布以及小世界特性。得出以下结论：

1) “人人网”节点度分布不同于一般社交网络服从幂律分布，而是更符合指数分布特点；且出现了类似小变量饱和现象，并且在小范围内形成了“双峰”现象；

2) “人人网”具有较小的平均最短路径长和较大的聚集系数，符合小世界特性；

3) 计算得出“人人网”同配系数大于0，说明“人人网”具有同配性，节点度高的节点倾向于与高度节点连接；

4) 通过分析“人人网”用户主页信息，发现其用户状态数、照片数和访客数主要集中在一个范围内，没有明显的正相关特性。

本文的研究成果对进一步分析社交网络的用户行为、网络拓扑结构具有重要意义，为跨社交网站的数据挖掘研究奠定了良好基础。后续的工作主要包括：首先，针对采集的大量数据，进行文本分析，挖掘关于用户信息的一些更深层的东西，如可以根据“人人网”用户信息的真实性和用户群的特殊性研究用户的专业和用户行为的特点，实现跨社交网站的数据挖掘；其次，可对该采集系统进行扩展，通过设置配置参数来实现针对不同社交网站的信息定向抓取，提高其通用性。

参 考 文 献

[1] 陈兴蜀, 郝正鸿, 王海舟, 等. P2P网络电视拓扑测量方法研究与特性分析[J]. 四川大学学报: 工程科学版, 2012, 44(3): 86-94.
CHEN Xing-shu, HAO Zheng-hong, WANG Hai-zhou, et al. Measuring and characterizing topologies of P2P IPTV[J]. Journal of SiChuan University (Engineering Science Edition), 2012, 44(3): 86-94.

[2] 尤婷. 社交网站用户行为特征及其内在机制研究——以“人人网”为例[D]. 北京: 北京邮电大学, 2012.
YOU Ting. The research on social-networking users' behavior characteristics and interior mechanism: Take renren.com for example[D]. Beijing: University of Posts and Telecommunications, 2012.

[3] 邓夏伟. 基于社交网络的用户行为研究——用户行为分析与用户影响力建模[D]. 北京: 北京交通大学, 2012.
DENG Xia-wei. User behavior analysis based on social network service-user behavior analysis and user influence modeling[D]. Beijing: Beijing Jiao tong University, 2012.

[4] 姜志宏. 大规模P2PTV系统测量与建模研究[D]. 长沙: 国防科学技术大学, 2011.
JIANG Zhi-hong. Research on modeling and measurement of large scale P2P TV systems[D]. Changsha: National University of Defense Technology, 2011.
[5] 徐恪, 张赛, 陈昊, 等. 在线社会网络的测量与分析[J].计算机学报, 2014, 37(1): 165-188.
XU Ke, ZHANG Sai, CHEN Hao, et al. Measurement and analysis of online social networks[J]. Chinese Journal of Computers, 2014, 37(1): 165-188.
[6] MISLOVE A, MARCON M, GUMMADI K P, et al. Measurement and analysis of online social networks[C]// Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. [s.l.]: ACM, 2007: 29-42.
[7] WILSON C, BOE B, SALA A, et al. User interactions in social networks and their implications[C]//Proceedings of the 4th ACM European Conference on Computer Systems. [s.l.]: ACM, 2009: 205-218.
[8] MATEI R, IAMNITCHI A, FOSTER I. Mapping the Gnutella network[J]. Internet Computing, 2002, 6(1): 50-57.
[9] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用[M]. 北京: 清华大学出版社有限公司, 2006.
WANG Xiao-fan, LI Xiang, CHEN Guan-rong. Complex networks theory and its application[M]. Beijing: Tsinghua university press co, LTD, 2006.
[10] NEWMAN, MARK E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(2): 167-256.
[11] JIANG J, WILSON C, WANG X, et al. Understanding latent interactions in online social networks[J]. ACM Transactions on the Web (TWEB), 2013, 7(4): 18.

编 辑 蒋 晓

(上接第920页)

[10] KABBUR S, HAN E H, KARYPIS G. Content-based methods for predicting web-site demographic attributes [C]//2010 IEEE 10th International Conference on Data Mining (ICDM). Sydney: IEEE Press, 2010: 863-868.
[11] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 1998: 43-52.
[12] SU X, KHOSHGOFTAAR T M. A survey of collaborative filtering techniques[EB/OL]. [2014-01-15]. http://www.hindawi.com/journals/aai/2009/4214251.
[13] SARWAR B, KARYPIS G, KONSTAN J, et al. Application of dimensionality reduction in recommender system-a case study[R]. Minneapolis: Dept of Computer Science Univ of Minnesota, 2000.
[14] KOREN Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2008: 426-434.
[15] PRYOR M H. The effects of singular value decomposition on collaborative filtering[R]. Hanover: Dartmouth College, 1998.
[16] GOLUB G H, VAN LOAN C F. Matrix computations[M]. Maryland: Johns Hopkins University Press, 2012.
[17] JOACHIMS T. Making large scale SVM learning practical[R]. Dortmund: Universitat Dortmund,1999.
[18] LECHEVALLIER Y, SAPORTA G. Blum MGB choosing the summary statistics and the acceptance rate in approximate Bayesian computation[C]//Proceedings of Computational Statistics. Herdelberg: Springer, Physica Verlag, 2010: 47-56.

编 辑 蒋 晓