

# 基于关联规则挖掘的社区网络数据爬虫算法

穆 俊

(滇西科技师范学院 信息工程学院, 云南 临沧 677000)

**摘 要:** 为了提高社区网络的联合推荐性能, 需要进行网络数据爬虫设计, 提出一种基于关联规则挖掘的社区网络数据爬虫算法. 构建社区网络的信息传递模型, 挖掘社区网络用户行为信息特征量, 根据数据的属性特征进行关联规则合并, 采用模糊指向性聚类方法进行社区网络的用户行为属性特征聚类处理, 对聚类输出的特征量采用自相关模板匹配方法实现信息融合, 实现社交网络数据的关联规则挖掘, 结合数据聚类分布属性进行网络信息爬虫, 实现社区属性混合推荐. 仿真结果表明, 采用该算法进行社区网络数据爬虫处理的准确度较高, 个性化匹配程度较好, 对社区网络联合推荐结果的置信度水平较高, 提高了社区发现能力.

**关键词:** 关联规则挖掘; 社区网络; 数据爬虫; 推荐

**中图分类号:** TP391

**文献标识码:** A

**文章编号:** 1000—7180(2018)08—0105—04

## A Hybrid recommendation Model for Community Attributes of Social Networks Based on Association Rule Mining

MU Jun

(School of Information Science & Engineering, West Yunnan University, Lincang 677000 China)

**Abstract:** In order to improve the joint recommendation performance of the community network, we need to design the network data crawler, and propose a community network data crawler algorithm based on association rule mining. Constructing the information transfer model of the community network, mining the information characteristic quantity of the user behavior of the community network, combining the association rules according to the attribute characteristic of the data. The fuzzy directivity clustering method is used to cluster the user behavior attributes of the community network. The autocorrelation template matching method is used to realize the information fusion and the association rules mining of the social network data. The network information crawler is implemented by combining the data clustering and distribution attributes, and the community attribute mixed recommendation is realized. The simulation results show that the proposed algorithm has higher accuracy, better personalized matching degree and higher confidence level to the community network joint recommendation results, which improves the community discovery ability.

**Key words:** association rule mining; community network; data crawler; recommendation

### 1 引言

目前已经有多种算法用于挖掘社区网络中的数据结构, 结合网络排重方法, 将社交网络的社区发现问题转化为单目标优化问题. 传统方法中, 对社区网络的数据爬虫方法主要有模糊 C 均值方法<sup>[1]</sup>、PSO 网络爬虫方法<sup>[2]</sup>、层次分割聚类爬虫挖掘方法等<sup>[3]</sup>,

采用自适应学习方法进行数据关联性搜索和聚类, 结合层次分割方法实现对社区网络的用户行为特征数据爬虫, 文献[4]中提出一种基于用户位置轨迹跟踪预测的社区网络数据爬虫方法, 构建社交网络用户位置分布的关联拓扑结构模型, 采用分段特征提取方法进行社交网络用户位置的关联特征提取, 提高网络的联合推荐和数据爬虫能力, 但该方法在进

行社区网络的属性关联挖掘中存在计算开销过大的问题;KernighanLin 算法则需以网络社区数作为先验知识<sup>[5]</sup>,结合先验特征分布式挖掘进行社区属性特征推荐,提高了社区发现能力,该方法在进行数据爬虫中抗干扰能力不强;文献<sup>[6]</sup>中提出一种基于谱平分的爬虫方法,以用户行为、社区属性以及场景信息进行谱特征提取,提高社区网络的行为发掘能力,该方法在数据爬虫中具有大规模的先验样本作为测试集,在样本知识缺失的情况下爬虫输出的准确性不好。

针对上述问题,本文提出一种基于关联规则挖掘的社区网络数据爬虫算法。

## 2 数据爬虫算法优化

### 2.1 关联规则特征量挖掘

在构建社区网络的信息传递模型,挖掘社区网络用户行为信息特征量的基础上,进行数据爬虫算分的优化设计,本文提出一种基于关联规则挖掘的社区网络数据爬虫算法,根据初始网络拓扑结构,构建数据爬虫的轨迹图如图 1 所示。

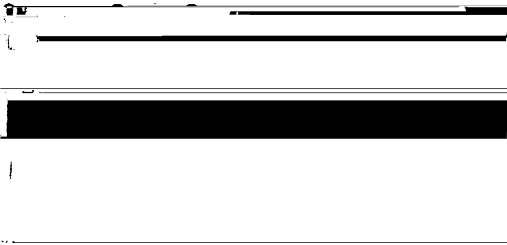


图 1 数据爬虫轨迹图

采用模糊指向性聚类方法进行社区网络的用户行为属性特征聚类处理,得到社区网络用户行为特征爬虫的偏好属性值记为  $C_{uw}^*$ :

$$C_{uw}^* = \sqrt{\frac{d_{in}(v)}{d_{out}(u) + d_{in}(v)}} \times C_{uw} \quad (1)$$

式中,  $d_{out}(u)$  表示以  $u$  为起点爬虫链路轨迹;  $u$  为社区属性混合推荐关联信息集合,即  $u$  的出度集合,采用模糊判定决策方法<sup>[7]</sup>,计算起始时刻  $T_0$  到爬虫终点的用户位置轨迹集合,得到用户行为属性的关联规则挖掘的特征值为:

$$\begin{aligned} CW_{\min}^{l_M} &= CW_{\min}^{l_{M-1}} \times (1 + \bar{D}_{l_{M-1}})^x \\ &= CW_{\min}^0 \times (1 + \bar{D}_{l_0})^x \times (1 + \bar{D}_{l_1})^x \times \cdots \times (1 + \bar{D}_{l_{M-1}})^x \\ &= CW_{\min}^0 \times [(1 + \bar{D}_{l_0}) \times (1 + \bar{D}_{l_1}) \times \cdots \times (1 + \end{aligned}$$

$$\bar{D}_{l_{M-1}})]^x \quad (2)$$

其中

$$\begin{aligned} &[(1 + \bar{D}_{l_0}) \times (1 + \bar{D}_{l_1}) \times \cdots \times (1 + \bar{D}_{l_{M-1}})] \\ &\leq ((1 + \bar{D}_{l_0} + 1 + \bar{D}_{l_1} + \cdots + 1 + \bar{D}_{l_{M-1}})/M)^M \\ &= (1 + \bar{D})^M \end{aligned} \quad (3)$$

对聚类输出的特征量采用自相关特模板匹配方法实现信息融合<sup>[8]</sup>,信息融合的适应度函数为:

$$\beta_d = (MPDist - d + 1)/MPDist, d \in [2, MPDist] \quad (4)$$

式中,  $adj(a, c)$  表示  $a \rightarrow c$  路径的个数,  $\beta_d \in (0, 1]$ , 采用模糊指向性约束控制方法进行社区网络的用户行为属性特征聚类处理,提高社区发现能力。

### 2.2 网络爬虫及社区发现

采用 Parallel Sets 变元轴排序方法进行社交网络的用户行为特征数据爬虫,结合模糊指向性约束控制方法进行最优爬虫路径设计,得到优化路径为  $A \rightarrow B, B \rightarrow C$ ,采用异步迭代方法,推出社区网络数据爬虫的联合关联分布满足:

$$\begin{aligned} MSD_{a \rightarrow b} &= \\ &1 - \frac{\sum_{i=1}^{|I_{a,b}|} \sqrt{(d_{a,i} - \bar{d}_a)^2 + (d_{b,i} - \bar{d}_b)^2}}{|I_{a,b}| \times \sum_{i=1}^{|I_{a,b}|} [\sqrt{(d_{a,i} - \bar{d}_a)^2} + \sqrt{(d_{b,i} - \bar{d}_b)^2}]} \end{aligned} \quad (5)$$

为了提高爬虫过程中的关联匹配性,结合互信息分析方法,进行全局位置寻优,输出社区网络爬虫数据的属性分布互信息为:

$$I(Q, S) = H(Q) - H(Q|S) \quad (6)$$

其中

$$H(Q|s_i) = - \sum_j \left[ \frac{p_{sj}(s_i, q_j)}{p_s(s_i)} \right] \log_2 \left[ \frac{p_{sj}(s_i, q_j)}{p_s(s_i)} \right] \quad (7)$$

根据社区网络的用户行为属性爬虫结果进行社交群体的混合推荐,得到混合推荐准则为:

准则(1):

$$\sqrt{\frac{R_{(m+1)n}^2 - R_{nn}^2}{R_{(m+1)n}^2}} = \frac{|x_{n(n)+m\pi} - x_{n+m\pi}|}{R_{(m+1)n}} \geq R_{tol} \quad (8)$$

准则(2):

$$\frac{R_{(m+1)n}}{\sqrt{\frac{1}{N} \sum_{k=1}^N \left[ x_k - \frac{1}{N} \sum_{k=1}^N x_k \right]^2}} > A_{tol} \quad (9)$$

综上分析,实现社交网络数据的关联规则挖掘,

结合数据聚类分布属性进行网络信息爬虫,并跟踪根据结果实现社区发现和网络推荐.改进算法的实现流程图如图 2 所示.

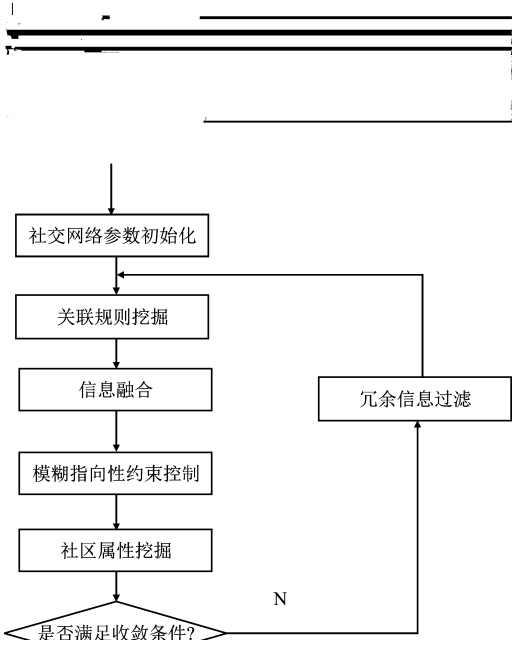


图 2 爬虫算法实现流程

3 仿真实验与结果分析

为了验证本文方法在实现社区网络用户行为属性特征数据爬虫中的应用性能,进行仿真实验,实验采用 Matlab 7 平台进行算法设计,社区网络平台的构造采用 Myeclipse 8.0 平台,采用网络爬虫抓取方法进行数据抓取,抓取了社区网络中包括了 83 432 个用户与 1 000 235 条边的用户评论信息,行为特征的属性量化集为 24,数据的离散采样率为  $f_s = 10 * f_0 \text{ Hz} = 10 \text{ KHz}$ ,其它关键参数设定为  $\rho_1 = \rho_2 = 0.41, \delta = 0.8, Q = 40, c_1 = 120, c_2 = 350, c_r = 26, \mu_1 = \mu_2 = 0.131$ ,根据上述仿真环境和参数设定,进行社区网络用户行为属性数据爬虫和社区混合推荐仿真,测试数据集采集的样本规模为 2 000,训练样本规模为 1 000,学习迭代步数为 50,得到数据爬虫结果如图 3 所示.

以图 3 的爬虫数据为测试样本集,进行关联规则挖掘,得到关联规则特征提取输出如图 4 所示.

分析图 4 得知,采用本文方法进行社区网络数据爬虫和特征提取,输出的关联规则的特征匹配性较好,说明数据爬虫的准确度较高,从而提高了社区发现能力.为了对比性能,采用不同爬虫方法进行社区发现,得到社区发现的置信度水平,得到结果如图

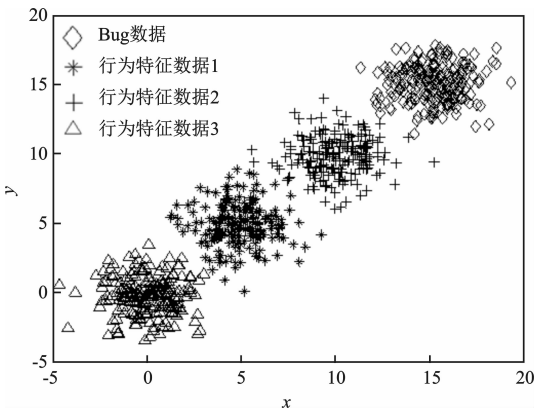


图 3 数据爬虫结果

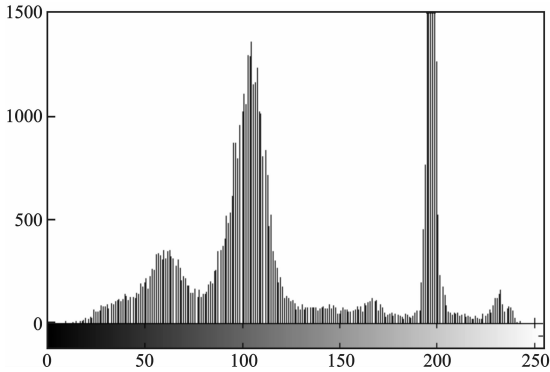


图 4 社区网络的用户行为属性关联规则挖掘

5 所示.

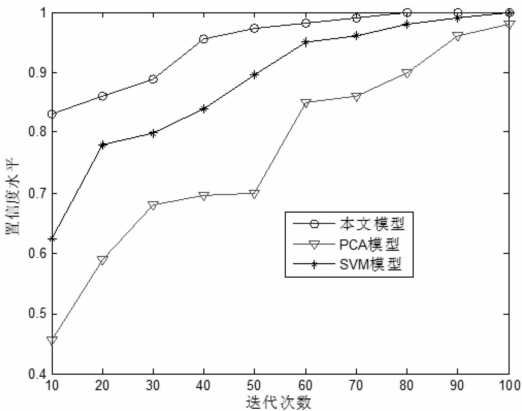


图 5 性能对比测试

分析图 5 得知,随着迭代步数的增大,数据爬虫的准确度提高,当迭代步数达到 80 次,置信度水平达到  $P=0.95$ ,具有显著性,社区发现能力比传统方法平均提升 12.67%,计算开销比传统方法降低 2.45s,仿真结果说明,本文方法进行社区网络的用户行为属性数据爬虫,能有效提高社区发现能力,置信度水平得到有效提升,整个算法的收敛性较好,对用户和网络的个性化匹配能力较强.

4 结束语

本文提出一种基于关联规则挖掘的社区网络数据爬虫算法。挖掘社区网络用户行为信息特征量,根据数据的属性特征进行关联规则合并,采用模糊指向性聚类方法进行社区网络的用户行为属性特征聚类处理,采用自相关特模板匹配方法实现信息融合,实现社交网络数据的关联规则挖掘,结合用户行为特征分布属性进行网络信息爬虫,实现社区属性混合推荐。研究表明,本文方法进行社区网络的用户行为属性特征数据爬虫挖掘的准确性较高,特征匹配能力较好,提高社区发现能力。

参考文献:

[1] 刘东,吴泉源,韩伟红,等. 基于用户名特征的用户身份同一性判定方法[J]. 计算机学报, 2015, 38(10): 2028-2040.

[2] 罗磊,陈恳,杜峰坡,等. 基于改进型粒子群算法的曲面匹配与位姿获取[J]. 清华大学学报(自然科学版), 2015, 55(10): 1061-1066.

[3] DURAO F, DOLOG P. Improving tag-based recommendation with the collaborative value of wiki pages for knowledge sharing[J]. Journal of Ambient Intelli-

gence Humanized Computing, 2014, 5(1):21-38

[4] JIA D Y, ZHANG F Z. A collaborative filtering recommendation algorithm based on double neighbor choosing strategy[J]. Journal of Computer Research and Development, 2013, 50(5):1076-1084.

[5] POLATIDIS N, GEORGIADIS C K. A multi-level collaborative filtering method that improves recommendations[J]. Expert Systems with Applications, 2016, 48:100-110.

[6] GUO Huaping, ZHOU Jun, WU Chang'an, FAN Ming. k-nearest neighbor classification method for class-imbalanced problem. Journal of Computer Applications, 2018, 38(4): 955-959.

[7] 孟祥武,刘树栋,张玉洁,等. 社会化推荐系统研究[J]. 软件学报, 2015, 26(6):1356-1372.

[8] 吴铮,于洪涛,刘树新,等. 基于信息熵的跨社区网络用户身份识别方法[J]. 计算机应用, 2017, 37(8): 2374-2380.

作者简介:

穆俊男, (1979-), 硕士, 副教授. 研究方向为计算机软件与理论、计算机应用、数据挖掘.  
E-mail: mujun511@126.com.

(上接第 104 页)

[3] Wu M T, Zhan J, Lin C W. Mining of High-Utility Itemsets by ACO Algorithm[C]// The Multidisciplinary International Social Networks Conference. 2016:44.

[4] Liu Y, Diao X, Cao J, et al. Local Enhanced Catfish Bat Algorithm[C]// International Conference on Robots & Intelligent System. 2016:238-245.

[5] Król A. Application of the Genetic Algorithm for Optimization of the Public Transportation Lines[M]// Intelligent Transport Systems and Travel Behavior. Springer International Publishing, 2017.

[6] Wang B, Zheng X, Zhou S, et al. Constructing DNA Barcode Sets Based on Particle Swarm Optimization.

[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2017, PP(99): 1-1.

[7] Meng Y S, Enriquez A C, Hsiao T Y, et al. Enhanced Differential Evolution Algorithm for Coordination of Directional Overcurrent Relays[J]. Electric Power Systems Research, 2017, 143:365-375.

作者简介:

梁豪默男, (1988-), 博士研究生. 研究方向为系统工程.

王智学男, (1961-), 硕士, 教授. 研究方向为需求工程、系统工程.

刘艺(通讯作者)男, (1990-), 博士研究生. 研究方向为进化算法. E-mail: albetliu20th@163.com.