一种基于主题相似性和网络拓扑的微博社区发现方法①

王卫平, 范 田

(中国科学技术大学 管理学院, 合肥 230026)

摘 要: 随着微博的迅速发展和大量普及, 微博社区发现已经成为新兴的研究热点. 发现网络社区有助于运营商理解网络结构和用户特征, 为用户提供个性化服务. 目前有关社区挖掘的研究大多只关注于网络结构, 忽略节点内容. 本文综合考虑网络结构和节点内容, 提出一种基于用户主题相似性和网络拓扑结构的微博社区发现方法. 首先从微博文本中抽取用户主题, 然后结合用户之间的链接关系, 对它们进行基于相似性的聚类, 最终获得社区结构. 在真实数据集上的实验证明: 所提出的方法不但能够发现潜在社区, 而且还能获知社区主题.

关键词: 社会网络; 微博; 社区发现; 聚类; LDA

Community Discovery Method Based on Users' Interest Similarity and Social Network Structure

WANG Wei-Ping, FAN Tian

(School of Management, University of Science and Technology of China, Hefei 230026, China)

Abstract: With the rapid development and a large popularity of microblogging, community discovery has become the current new research focuses, which could help the operators understand network structure and the characteristics of users, then provide users with personalized services. Most of previous study only emphasized the network structure without considering the content. The paper provides a community discovery method based on the users' theme similarity and network structure. Firstly, retrieve users' theme from their microblogging; then cluster the similar users based on the links among them and users' similarity; finally gain the communities. The experiments on the big data show that the method can not only find potential community, but also gain its theme.

Key words: social network; Microblogging; community mining; clustering; LDA

微博(MicroBlogging)又称"微博客", 是一个基于用户关系的信息分享、传播以及获取平台. 用户可以通过 WEB、WAP 以及各种客户端组建个人社区, 以 140 字左右的文字更新信息, 并实现即时分享[1]. 据中国互联网信息中心统计, 到 2011 年 12 月, 中国微博用户总数达到 2.498 亿, 居于全球第一.

微博中, 用户建立关系, 可关注其它用户或被其它用户关注, 这些关系形成了显性社交网络. 用户发表微博, 可原创微博或转发(评论)其它用户的微博. 微博之间的关系使用户形成了隐性社交网络. 网络结构如图 1 所示, U_i 表示用户, B_i 表示微博, R 表示转发. 研究发现: 用户的行为由其兴趣驱动, 拥有相似兴趣的用户更有可能成为朋友, 而成为朋友后也促使它们分享更相似的兴趣[2]. 网络社区逐渐形成, 社区内部用户之间的联系比较紧密, 社区之间的联系相对比较稀疏. 社区挖掘可以帮助用户在茫茫人海中发现相似用户, 有助于运营商提供个性化服务、企业实行精确性营销.

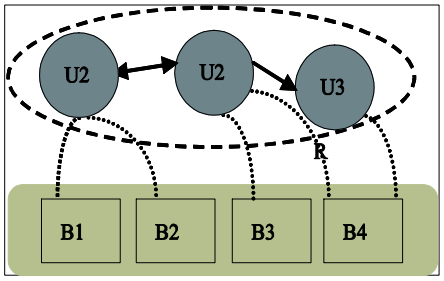

图 1 微博网络结构图

① 收稿时间:2012-11-22;收到修改稿时间:2012-12-29

目前, 网络社区发现已经成为数据挖掘领域的热点之一. 主要方法有两种: 一种基于用户间的关系, 即图的拓扑结构, 使用复杂网络社区划分的方法来发现社区, 包括基于分级聚类的 GN 算法和基于图形分割的 KL 算法、谱聚类算法等[3]. 另一种基于用户主题特性, 将拥有相似兴趣的用户划分到同一社区中, 包括 CUT (community user topic)模型和 CAT (community author topic)模型等. 无论哪种方法都只关注于社交网络的一个特性. 事实上, 即便用户之间没有显式链接, 如果它们兴趣相似, 也倾向于加入同一社区, 而且用户之间的链接关系也能在社区发现中起到指引作用.

本文提出了一种综合考虑节点内容(即用户主题)和(显性)社交网络拓扑结构(即用户间的关系)的社区发现方法. 该方法先从用户的微博中抽取出用户主题, 再结合用户之间的链接关系, 进行基于相似性的用户聚类, 从而获得社区结构. 并在真实数据集上进行实验, 验证了算法的准确性和性能.

文章的第 2 部分对用户主题模型进行了介绍, 第 3 部分对提出的社区发现方法进行了描述, 第 4 部分给出了实验研究结果, 最后是对本文的总结.

1 微博用户主题

1.1 LDA 主题模型

为了从用户的微博文本中自动识别出用户主题, 本文使用 LDA(Latent Dirichlet Allocation)模型[4], 它假设文本是主题的概率分布, 主题是词汇的的概率分布, 且认为文本集中的文本及文本中的词都是可交换的.

假设 D 个文本包含 W 个词汇, 涉及 K 个主题, 则所给文本 d 中的词汇 w_i 的概率如公式(1)下所示:

$$p(w_i \mid d)=\sum_{j=1}^{K} p(w_i \mid z_i=j)p(z_i=j) \qquad (1)$$

其中, w_i 是观测变量, z_i 是潜在变量, 表明词汇 w_i 取自该主题. 令 θ_d 表示文本 d 在 T 个主题上的多项分布, 假设服从 $Dir(\alpha)$ 先验概率. 令 φ_z 表示主题 z 在 W 个词汇上的多项分布, 假设服从 $Dir(\beta)$ 先验概率.

LDA 生成方式可用一个“文本-主题-词汇”三层的贝叶斯模型表示, 如图 2 所示. 它首先从 $Dir(\beta)$ 分布中抽取主题与词汇的关系 φ_z. 当生成一个文本时, 先从 $Dir(\alpha)$ 分布中抽样出该文本和各个主题的关系 θ_d. 对文本中的每一个词, 从 $Mult(\theta_d)$ 分布中抽样出所属的主题 $z_{d,n}$, 再从 $Mult(\theta_{z_{d,n}})$ 分布中抽取出具体的词汇 $w_{d,n}$. 生成过程如图 3 所示.

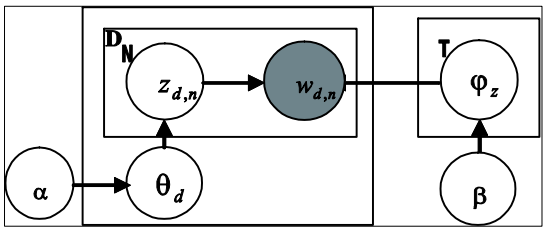

图 2 LDA 模型的贝叶斯网络图

1.for each topic $k \in \{1 \cdots k\}$ draw $\varphi_z \sim Dir(\beta)$
2.for each document $d \in \{1 \cdots D\}$
a) draw $\theta_d \sim Dir(\alpha)$
b) for each word $w \in \{1 \cdots N_d\}$
i. draw $z_{d,n} \sim Mult(\theta_d)$
ii. draw $w_{d,n} \sim Mult(\varphi_{z_{d,n}})$

图 3 文本的生成过程

一般采用 Gibbs Sampling[5]方法求解, 它通过迭代采样来逼近真实的概率分布, 具体通过估计当前采样词 w_i 的主题 z_i 的后验概率 $p(z_i=j \mid w_i, z_{-i})$, 估算出模型 θ 的参数和 φ[6]. 计算如公式(2)、(3)和(4)所示:

$$p(z_i=j \mid w_i, z_{-i}) \propto \frac{n_{-i,j}^{(w_i)}+\beta}{n_{-i,j}^{(\cdot)}+N\beta} \bullet \frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+K\alpha} \qquad (2)$$

舍弃词汇记号, 以 w 表示唯一性词, 对每个单一样本, 可按下式估算.

$$\tilde{\varphi}_w^{z_i=j}=\frac{n_j^{(w)}+\beta}{n_j^{(\cdot)}+N\beta} \qquad (3)$$

$$\tilde{\theta}_{z_i=j}^{d}=\frac{n_j^{(d)}+\alpha}{n_{\cdot}^{(d)}+K\alpha} \qquad (4)$$

其中, $n_j^{(w)}$ 表示词汇 w 被分配给主题 j 的频数; $n_j^{(\cdot)}$ 表示分配给主题 j 的所有词数; $n_j^{(d)}$ 表示文本 d 中分配给主题 j 的词数; $n_{\cdot}^{(d)}$ 表示文本 d 中所有被分配了主题的词数.

1.2 用户主题模型

一条微博一般为 140 字左右, 这个“短文本”特性给传统的文本分析处理方法带来了严重的数据稀疏问题. 若把用户的微博文本单独直接应用到 LDA 模型中, 获取的主题会比较稀疏, 不利于后序相似用户的发现. 而且本文关注的是每个用户的主题, 并不是每条微博的主题. 针对这个问题, 文献[7]把用户的全部微博合

成一个文本，挖掘的文本主题就是用户主题，其本质是对 author-topic model[8]的应用. 但是这种处理方式忽略了微博文本的时间特性. 事实上，受外界环境或自身状况的影响，用户会不断地产生新兴趣，放弃旧兴趣，话题内容和偏好强度会随时间而动态变化[9]. 忽略时间特性就忽略了这些变化. 用户的兴趣是有有效期的，过去和现在的兴趣对于用户具有不同意义，其重要程度随着时间而衰减.

于是，本文提出一个适用于微博用户的主题模型. 首先，根据时间信息把用户的微博文本离散到时间序列上对应的时间窗口上，把每个时间窗口内的文本聚合，形成时间片文本集. 然后把它们应用于 LDA 模型中，估计出模型的参数θ_d,φ_k,z，并考虑受时间影响的主题有效性，获得用户的主题概率表示.

另外，不同于一般文本，微博文本除了具有时间特性外，还有一些特殊标识(@、//和#). 其中，“@”表示微博之间的联系人关系，“//”表示微博之间的文本关系，“#***#”间的内容表示微博的标签. 例如，用户“范小田田爱读书”的一条微博——“好消息！//@读书微吧: #莫言获诺贝尔文学奖#莫言的写作能力令人难以置信！”中，“//”之前为该用户的原创内容，“//”之后为转发内容，“@读书微吧”表示转发部分的作者是用户“读书微吧”，“莫言获诺贝尔文学奖”则是这条微博的标签. 显然，用户名称和微博标签往往比其它一般内容更有助于主题识别. 然而 LDA 主题模型采用词袋形式表示文本，同等对待每个词而不考虑词的权重. 本文通过改变求解 LDA 模型的求解方法，以区别对待它们.

本文设定主题由一组语义上相关的词及词与主题相关的权重的向量表示. 时间片主题由时间片文本的主题概率分布表示. 用户主题则由不同时间片主题加权构成. 定义如公式(5)、(6)和(7)所示:

$$Z_j^t=\varphi_j=\{(w_1,p(w_1\mid z_j)),\cdots,(w_n,p(w_n\mid z_j))\} \quad (5)$$

$$D^t=\theta_d=\{(z_1,p(z_1\mid d_t)),\cdots,(z_k,p(z_k\mid d_t))\} \quad (6)$$

$$U_i=\varpi_t\cdot D^t=\{(z_1,p(z_1\mid u_i)),\cdots,(z_k,p(z_k\mid u_i))\} \quad (7)$$

其中，w_i是与主题相关z_j的词，$p(w_i\mid z_j)$是两者的相关性度量，Z_j^t为时间片t内第j个主题表示. z_i是与时间片文本d_t相关的主题，$p(z_i\mid d_t)$是两者的相关性度量，D^t为第t个时间片文本主题表示. ϖ_t为d_t的时间片权重，为一个随时间而衰减的函数，U_i为该用户的主题表示.

将处理后的文本数据分别应用于 LDA 模型之中. 因为需要考虑不同位置词的权重，所以使用改进的 Gibbs Sampling 算法[10]求解，即当把文本d_i中的一个词w_i分配给主题j时，公式(3)和(4)中的$n_j^{(w)}$、$n_j^{(\cdot)}$、$n_j^{(d)}$和$n_{\cdot}^{(d)}$的值不再累加 1，而是累加其权重.

2　基于主题相似性和网络拓扑的社区发现

2.1 基本思想

本文目的在于发现社区内部节点相似度高(链接较多、主题相似)，而社区间的相似性低的社区结构. 社交网络中的社区往往是由大量普通节点围绕着少量“中心节点”建立起来的. 本文先根据最大最小距离原则选择初始聚类中心，然后对剩余节点进行基于相似性的聚类，从而获得社区结构. 在将节点加入到相关社区时，若存在节点与已知社区都不大相关，则把它作为一个新的社区中心. 若存在节点与若干社区都很相关，则把它加入到多个社区之中，这样获得的社区结构是重叠的，符合用户的“多样性”特点. 聚类结束后，再对社区结构进行调整，合并高度重叠的社区，以获得较大规模较完整的社区结构.

2.2 用户相似性

基于以上考虑，本文从主题相似度和路径距离两个方面度量用户相似性.

在隐性社交网络层面，节点之间的距离由其内容相似性决定. 显然，两个节点的主题越相似，它们之间的联系就越紧密，相似性就越高.

一般地，最常用 KL 距离衡量两个概率密度之间的相似度，KL 距离也被称为相对熵、交叉熵，是样本概率为 P 的信息使用概率编码后期望额外增加的比特数，用来衡量概率密度之间的差距.

用户主题概率密度之间的差距越小，用户越相似，本文使用 JS 距离的倒数来度量用户之间的主题相似度，定义如公式(8)所示:

$$sim(v_i,v_j)=\frac{1}{JS(v_i\parallel v_j)} \quad (8)$$

$$JS(v_i\parallel v_j)=\frac{1}{2}D_{KL}(v_i\parallel v_k)+\frac{1}{2}D_{KL}(v_j\parallel v_k)$$

$$D_{KL}(v_i\parallel v_j)=\sum_{k=1}^{K}p(z_k\mid v_i)\log_2\frac{p(z_k\mid v_i)}{p(z_k\mid v_j)}$$

$$v_k=\frac{1}{2}(v_i+v_j)$$

其中，v_i, v_j 为用户主题概率密度，$JS(v_i \| v_j)$ 为两者的 JS 距离，$D_{KL}(v_i \| v_j)$ 为两者的 KL 距离，v_k 为两者的均值.

在显性社交网络层面，节点之间的距离由它们之间的最短路径长度决定. 显然，两个节点之间的路径越长，相似性越低. 节点之间的最短路径定义为含有最少边的一条路径，可以通过 Dijkstra 算法或 Floyd 算法[12]求得.

设网络中非相邻节点 v_i 和 v_j 节点之间的最短路径 $link(v_i, v_j)$ 为 $\{(v_i, v_k), (v_k, v_m), \cdots, (v_n, v_j)\}$，定义它们之间的相似度为最短路径上所有节点对的相似度之积，存在多条最短路径时，则取最大值，如公式(9)所示:

$$rel(v_i, v_j) = \max_i \prod_{(v_m, v_n) \in link_i} sim(v_m, v_n) \tag{9}$$

其中，$sim(v_m, v_n)$ 为节点 v_i, v_j 的主题相似度，$link_i$ 为节点 v_i, v_j 的最短路径.

2.3 算法描述

社区发现算法分为三个步骤: 首先选择初始聚类中心，然后挖掘网络社区，最后调整社区结构.

2.3.1 初始中心选择

传统的聚类方法中，处理大数据集时，*K*-均值算法[12]是相对可伸缩和高效的，但是它的执行效率和准确率对于初始中心的选取有严重的依赖性，不同的选择会产生不同的聚类结果，选择不当时，更极易陷入局部最优. 一般希望找到散布较大的且具有一定代表性的点作为初始中心点. 本文根据最大最小距离[13](即最小最大相似度)原则选择初始中心. 具体方法如下:

输入: 节点集合 $V = \{v_1, v_2, \cdots, v_n\}$，簇个数 k

输出: 初始中心节点集合 $C = \{c_1, c_2, \cdots, c_k\}$

算法步骤:

① 选取集合 V 中度最大的节点作为第一个聚类中心 c_1，加入到集合 C 中;

② 若 $|C| \neq k$，则执行步骤③，否则，结束算法;

③ 对集合 V 中的剩余节点，计算与集合 C 中各节点的相似度，取较大值 $D_i = \max_{i \in C} rel(v_i, c_j)$，然后将该值最小 $\min_{i \in V} D$ 对应的节点 v_i 作为新的中心加入到集合 C 中，然后转至步骤②.

其中，步骤③最为关键，将 $\min_{i \in V} D \max_{i \in C} rel(v_i, c_j)$ 对应的节点作为新的聚类中心的候选对象，从而避免了将与已有中心很相似而与其它中心不相似的节点作为候选对象的可能. 事实上，社交网络中的社区往往是由大量普通节点围绕着少量“中心节点”建立起来的. 这些“中心节点”通常与其它节点联系很多，度比较大. 因此把集合 V 中度较小的节点过滤掉，避免它们成为初始中心. 这样一方面降低了该算法的开销，另一方面也使所发现的初始节点中心性更强.

2.3.2 社区发现

获取初始聚类中心集合后，对剩余节点进行基于用户相似性的聚类. 具体方法如下:

输入: 节点集合 $V = \{v_1, v_2, \cdots, v_n\}$，初始中心节点集合 $C = \{c_1, c_2, \cdots, c_k\}$

输出: 社区集合 $\{C_1, C_2, \cdots, C_k\}$

算法步骤:

① 根据与中心点的相似度，将剩余节点指派到相应的簇中;

② 更新各子簇的中心点;

③ 如果所有的子簇中心均不发生变化，则结束，并获得社区结构，否则转至步骤①继续执行.

其中，步骤①在指派节点到相应子簇时，根据公式(8)计算剩余节点与各个簇中心的相似度，将其指派到相似度大于阈值 ε 的簇中. 若存在与已知簇中心都不大相关的节点，则形成一个新的簇. 若存在与若干簇都很相关的节点，则把它们同时指派到多个簇之中. 步骤②更新中心点时，选择子簇中与所有其它节点的平均相似度最高的节点作为新的中心点.

2.3.3 社区合并

由于通过相似度阈值来判断节点是否加入某一社区，所以存在一些边界节点属于多个社区，即社区内容相互重叠. 社区重叠度定义为子社区共同节点的个数占全部节点个数的百分比.

$$o(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \tag{10}$$

其中，$|C_i \cap C_j|$ 为社区 C_i, C_j 共同节点的个数，$|C_i \cup C_j|$ 为社区 C_i, C_j 所有节点的个数.

那么，当社区 C_i, C_j 的重叠度达到阈值 τ 时，表明它们联系相当紧密，可以合并成一个社区. 步骤如下:

① 根据公式(10)计算社区 C_i, C_j 的重叠度;

② 当 $o(C_i, C_j) > \tau$ 时，将 C_i, C_j 并为一个社区;

③ 当任意两个社区的重叠度均小于阈值 τ 时，则结束，获得最终社区结构，否则，转至步骤返回①继续执行.

社区合并后，网络最终被划分成若干个社区，社区内部的节点相似度高(链接稠密，主题相似)，社区之间相似度低，而且一个节点可以同时属于多个社区.

3 实验结果与分析

3.1 实验准备

3.1.1 数据集

本文采用的数据集的原始数据来源于新浪微博. 该数据集收集了 2012 年 8 月至 10 月期间的微博内容和用户间的链接关系. 过滤掉发微博数少于 50 的用户和关注数/被关注数少于 5 的用户. 最终数据共涉及 671 个用户、5877 条链接和 105434 条微博.

对于文本数据进行分词去停用词等处理. 其中，使用汉语词法分析系统 ICTCLAS 进行分词，并采用停用词字典的方法，过滤掉代词、语气助词、微博中的特殊字符(如表情)等出现频率很高却对于主题挖掘没有帮助的词汇.

3.1.2 实验环境

本文的实验环境为 Intel Core i5 2.40GHz 的 CPU, 2GB 的内存, 500G 硬盘的 PC 机，操作系统为 Windows 7 旗舰版，实验工具为 Eclipse.

3.2 实验方法和评价指标

本文方法本质上是基于聚类的社区划分，类似地，也可以采用平方误差和来衡量社区划分的效果. 定义如公式(11)所示:

$$E=\sum_{i=1}^{k}\sum_{v\in C_i}\left|v-o_i\right|^2 \tag{11}$$

其中，k 为社区个数，v 为社区 C_i 的点，o_i 为社区 C_i 的中心点.

另外，本文发现的社区结构是综合考虑网络结构和节点内容两个方面的. 一个好的划分，社区内部的节点应该链接稠密且主题相似. 这里把社交网络视作一个加权网络，节点之间的主题相似度即为边的权重. 因此可以通过比较各社区内外部包含边的权重和来评价结果，定义如公式(12)所示:

$$F=\frac{1}{k}\sum_{i=1}^{k}\frac{\boldsymbol{W}_{in}^{C_i}}{\boldsymbol{W}_{in}^{C_i}+\boldsymbol{W}_{out}^{C_i}} \tag{12}$$

其中，k 为社区个数，$\boldsymbol{W}_{in}^{C_i}$ 表示社区 C_i 内部边的加权和，$\boldsymbol{W}_{out}^{C_i}$ 表示社区 C_i 外部边的加权和. 边 (v_i,v_j) 的权值是根据公式(8)计算的主题相似度 $sim(v_i,v_j)$. 可知，F 值越大，社区内部节点链接越紧密，主题越相似.

实验内容包括 3 部分: 首先，使用本文提出的方法对处理后的微博数据集进行社区划分，展示挖掘结果；然后，设置不同的社区个数，分析该算法迭代不同次数所得的平方误差和 E，研究其收敛性；最后，以 K-均值算法为基准，设置不同的社区个数，以 F 值为评价指标，进行对比实验.

3.3 实验准备

3.3.1 社区挖掘结果

主题模型的参数设置为 $\alpha=50/T,\beta=0.01,T=50$.

表格 1 列出了该算法所发现的一些比较有代表性的主题，并给出了每一个主题出现的概率，以及在该主题下出现概率最高的前 10 个词汇. 可以看出，在每一个主题下出现概率最高的词汇可以很好地描述该主题的内容. 分析得: T5 是关于钓鱼岛的主题, T8 是关于“中国好声音”的主题, T13 是关于财经的主题， T22 是关于生活感悟的主题，T24 是关于社会的主题，而 T37 是关于手机的主题. T5 和 T8 是这段时间的热门话题.

表 1　主题挖掘结果

主题	主要词汇
T5 0.02279	日本 0.04232,钓鱼岛 0.0267,报道 0.01788,美国 0.01318,政府 0.00893,航母 0.00814,总统 0.00717,海军 0.00617,海域 0.00601,领土 0.00549
T8 0.01297	卫视 0.02137,浙江 0.01901,视频 0.01218,学员 0.01125,导师 0.01041,播出 0.00975,直播 0.00971,互动 0.00765,声音 0.07398,考核 0.00663
T13 0.01916	经济 0.03096,银行 0.03012,上市 0.01018,金融 0.00886,资产 0.00878,基金 0.00853,股市 0.00810,政府 0.00734,证券 0.00733,人民币 0.00690
T22 0.03626	希望 0.00944,永远 0.00882,心情 0.00726,努力 0.00706,朋友 0.00692,梦想 0.00682,美好 0.00680,事情 0.00632,珍惜 0.00611,经历 0.00564
T24 0.02541	社会 0.01724,政府 0.01224,国家 0.01171,改革 0.00933,政治 0.00846,制度 0.00747,官员 0.00651,历史 0.00628,教育 0.00582,自由 0.00571
T37 0.02570	手机 0.03350,苹果 0.02757,用户 0.01873,三星 0.01168,产品 0.00888,游戏 0.00856,智能 0.00825,诺基亚 0.0062,服务 0.00546,地图 0.00528

然后，进一步挖掘出每一个社区中的成员及所涉及的主题.

表格 2 展示了社区 3 和社区 5 中的 5 个核心作者和涉及的主要主题. 其中，社区 3 涉及的主要为生活感悟(T22)、日常唠叨 (T33)、健康(T27)、美食(T26)和星座 (T43)等方面的主题. 社区 5 涉及的主要为政治(T24)、钓鱼岛(T5)、新闻事件(T7)、日常唠叨(T33)

和生活感悟(T22)等方面的主题. 全面分析可得, 大部分社区都会涉及一些类似的主题, 如日常唠叨(T33)、生活感悟(T22)和一些热门话题(T8、T5)等. 除此之外, 每个社区仍存在一些与其他社区相区别且出现概率很高的主题, 如社区 3 包含美食(T26)和星座 (T43)等主题, 而财经(T13)和社会(T24)主题在社区 9 中占有很高的比例. 这符合研究结果“微博用户更多的是用简短信息来记录心情、寻找兴趣相同的群体、讨论共同兴趣的话题等”[13].

表 2 社区 3 和社区 5 的核心用户和主要主题

	核心用户	主要主题
社区 3	阿狸心情日志、几米漫画集、好小孩健康日记、生活小智慧、健康美容美食	T22、T33、T27、T26、T43
社区 5	Vista 看天下、新周刊、闾丘露薇、郎咸平、任志强	T13、T24、T5、T22、T33

3.3.2 研究算法的收敛性

图 4 展示了本文方法在获取不同个数的社区时, 迭代不同次数产生的平方误差和的变化情况. 可以看出, 本文提出的方法能够快速收敛, 迭代 60 次以上时, 误差稳定.

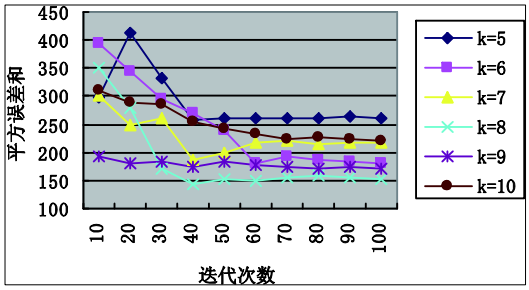

图 4 不同迭代次数下平方和误差的变化情况

3.3.3 对比评估社区效果

图 5 展示了本文方法与 *K*-均值算法在获得包含不同社区个数的划分结果的值的变化情况. 可以看出, 随着社区个数的增加, 它们都逐渐获得最优效果, 然后再变差. 在社区个数为 8 时, 本文方法表现最好.

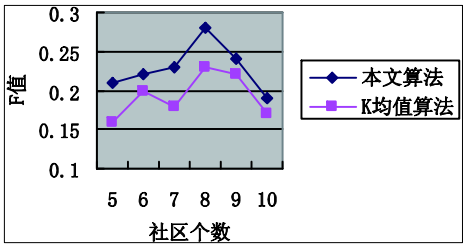

图 5 不同算法的 F 值对比

4 总结

本文综合考虑节点内容和网络拓扑结构, 提出了一种基于用户主题相似性和用户之间的关系的微博社区发现方法, 获得内部链接稠密且主题相似的社区结构. 算法首先通过对微博文本应用主题模型获得用户主题. 在这个过程中, 考虑了用户主题的动态变化和微博文本的特殊位置词汇问题. 然后基于最大最小距离原则选择初始聚类中心, 并在此基础上对剩余节点进行基于用户相似性的聚类, 随后合并高度重叠的社区, 最终获得社区结构. 在真实数据集上的实验证明, 本文提出的方法不但能够发现潜在社区, 而且还能获知社区主题, 解决了传统基于链接结构社区发现方法缺乏语义性解释的问题.

参考文献

1 熊小兵,周刚,黄永忠,等.新浪微博话题流行预测技术研究.信息工程大学学报,2012,13(4):496–502.

2 Lauw H, Shafer JC, Agrawal R. Homophily in the digital world: a live journal case study. IEEE Internet Computing, 2010,14(2):15–23.

3 解伍,汪小帆.复杂网络中的社团结构分析算法研究综述.复杂系统与复杂性科学,2005,2(3):1–12.

4 Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003(3):993–1022.

5 Griffiths TL, Steyvers M. Finding scientific topics. Proc. National Academy of Science of United States of America. 2004,(101):5228–5235.

6 张小平,周雪忠,黄厚宽,等.一种改进的 LDA 主题模型.北京交通大学学报,2010,34(2):111–114.

7 Weng J, Lim EP, Jiang J, He Q. Twitter Rank: Finding topic-sensitive influcential twitters. Proc.of the 3rd ACM WSDM., 2010.

8 Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T. Probabilistic author-topic models for information discovery.SIGKDD,2010.

9 Blei DM, Lafferty JD. Dynamic topic models. ICML,2006.

10 石晶,胡明,石鑫,等.基于 LDA 模型的文本分割.计算机学报,2008,31(10):1865–1873.

11 Han J, Kamber M.数据挖掘:概念与技术.范明,孟小峰,译.北京:机械工业出版社,2001.

12 李金宗.模式识别导论.北京:高等教育出版社,1994.

13 傅志华.数据:2010 微博与社区调查.北京: DCCI 互联网数据中心,2010.