doi:10.14132/j.cnki.1673-5439.2018.03.009

# 大规模网络中基于LDA模型的重叠社区发现

张 伟,祁德昊,陈云芳

(南京邮电大学 计算机学院,江苏 南京 210023)

**摘要**:传统的重叠社区发现基于网络的结构信息,具体依靠节点之间的连接关系,由于没有使用节点的内容信息,难以反映网络社区的语义。文中提出了一种大规模网络中基于节点属性的重叠社区发现算法(Overlapping Community Detection algorithm based on LDA,OCD_LDA),该算法使用LDA主题模型对节点内容进行多维属性建模,将网络节点看作文章,节点所携带的多维属性值看作文章中的单词,因此网络中的社区对应了主题模型中的主题,节点的多重社区归属对应于文章的多个主题。算法进一步考虑到网络中节点内容短小在主题建模过程中导致的数据稀疏问题,在LDA主题模型中引入Spike and Slab prior方法辅助实现变量选择和参数估计,有效地解决节点上社区分布的稀疏性和平滑性问题。实验使用DBLP文献数据集对算法进行了验证,结果表明,OCD_LDA算法能够更加有效地发现大规模网络中的重叠社区分布,揭示出复杂数据的内在特性。

**关键词**:社会网络;LDA;社区发现;重叠社区

**中图分类号**:TP391.1 **文献标志码**:A **文章编号**:1673-5439(2018)03-0054-11

# Overlapping community detection algorithm based on LDA in large scale networks

ZHANG Wei, QI Dehao, CHEN Yunfang

(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

**Abstract**: The traditional overlapping community detection is based on the network structure information, and depends on the connection relationship between the nodes. Without the content information of the nodes, it is difficult to reveal the semantics of the network community. An overlapping community detection algorithm based on node attributes in largescale networks, overlapping community detection algorithm based on LDA( OCD_LDA), is proposed. The LDA topic model is used to model the multi-dimensional attributes of the node content in the algorithm, while a network node is regarded as an article and the multi-dimensional attribute value carried by the node is regarded as the words in the article. Therefore, the community in the network corresponds to the theme in the topic model, and the multiple community attribution of nodes corresponds to multiple themes of the article. Moreover, the data sparsity caused by short content of the nodes in the topic modeling process is considered, and then the Spike and Slab prior method is introduced in the LDA topic model to help implement variable selection and parameter estimation to solve the sparsity and smoothness issues of community distribution on nodes. The experimental result in the DBLP bibliographic data set shows that the OCD_LDA can more effectively detect the distribution of overlapping communities in large-scale networks and reveal the intrinsic properties of complex data.

**Keywords**: social networks; LDA; community detection; overlapping communities

**收稿日期**:2018-04-03 **本刊网址**:http://nyzr.njupt.edu.cn

**基金项目**:国家自然科学基金(61272422,61672297)资助项目

**通讯作者**:张 伟,男,教授,E-mail:zhangw@njupt.edu.cn

**引用本文**:张伟,祁德昊,陈云芳. 大规模网络中基于LDA模型的重叠社区发现[J]. 南京邮电大学学报(自然科学版),2018,38(3):54-64.

信息社会中的诸多关系都可以使用网络结构图来表示,例如社会关系网络、作者合作网络和交通网络等。这些复杂网络蕴含着大量的信息,分析这些信息对了解网络拓扑结构、发现网络组织规律、预测网络演化以及进行个性化推荐等具有重要意义[1]。

一个社会网络表示成一张结构拓扑图,其中每个节点表示一个个体,节点之间的边表示个体之间的关系[2]。例如,在作者合作网中,一个节点表示一个作者,节点之间的边表示作者撰写文章中的合作关系。在对社会网络的研究过程中,研究者发现大多数复杂网络具有社区结构特性,即复杂网络是由若干个社区组成的,处于同一个社区的内部节点连接比较紧密,而不同社区之间的节点连接相对稀疏。作者合作网中,同一个社区的个体之间可能研究方向比较相似,沟通交流较多从而连接紧密;而不同社区间的个体可能研究方向存在较大差异,所以连接较为稀疏。并且,不同社区可能会具有不同的特性[3]。此外,研究者进一步发现网络中的节点有时候不仅仅属于唯一的一个社区,还有可能同时隶属于多个社区,即网络社区存在重叠性[4]。同样以作者合作网络为例,一个作者可能会因具有多个研究方向而同时隶属于多个社区。网络中同时隶属于多个社区的节点被称为重叠节点,包含重叠节点的社区被称为重叠社区[4]。

目前,社区发现的经典方法包括:分割方法[5-6]、GN 算法[7]、基于随机游走策略方法[8]、基于模块度方法[9]等等。这些方法解决的是非重叠社区发现,其主要思想是社区内的节点连接紧密,社区之间的连接相对稀疏,一个节点只能属于单个社区。考虑节点多重社区归属的重叠社区发现算法典型的有:基于派系过滤方法[10]、基于局部扩展算法[11]以及基于 Agent 方法[12]等,其思想是认为社区之间存在重叠性,重叠节点在其中起着桥接作用。上述非重叠社区和重叠社区发现算法都是基于网络的拓扑结构,即根据节点之间的连接关系来划分社区,这些算法难以解决高维网络空间的重叠社区划分问题。此外,由于网络节点的内容信息提供了不同于节点连接关系的语义信息,基于节点的内容属性的社区划分也开始得到关注。

基于内容的社区发现方法根据节点的语义信息来划分社区,即内容相似的节点属于同一个语义社区[13]。传统内容挖掘使用的主题模型主要用于文本分类、情感分析、推荐系统和学术文献挖掘等,随着主题模型在文本信息挖掘中体现出显著的优越性,有研究者提出将主题模型的概念应用到社区挖掘中[14]。文本信息对于研究网络的结构具有重要作用,通过对节点的文本信息进行主题建模,不但可以挖掘出节点背后隐含的文本主题,还可以进一步根据不同的主题信息来划分重叠社区。

本文提出了一种基于 LDA 主题模型的重叠社区发现方法,即对节点的内容信息进行 LDA 主题建模,根据每个节点内容的主题分布划分对应的社区。对比分析社区发现与 LDA 主题建模的目标和具体过程可以发现,社区发现的目的是对给定的一个网络结构,找到网络图中节点关系,这也可以看成是一种聚类,即将同一类的节点聚集到同一个社区。主题建模则是对于一个语料库中的每篇文章进行语义分析,从而得到每篇文章的主题概率分布。一般情况下,每篇文章会有一个或几个主题概率偏高,那么便可以根据这几个主题概率对文章进行分类,将同种主题的文章归为一类。在本文提出的方法中,我们将网络节点的集合看成是 LDA 主题模型中的语料库,一个节点看作是一篇文章,每个节点携带的内容属性值看作是每篇文章中的单词,网络中的社区则对应主题建模中的主题。通过对网络中节点的属性值进行建模采样,得到每个节点上的社区概率分布。在实际应用中,我们发现 LDA 主题模型在文本内容过小时会发生数据稀疏问题[15-16],而通常社会网络中节点的内容属性相对来说比较短小,从而造成数据稀疏,进一步导致节点的社区分布结果不精确。对此,我们引入了 Spike and Slab prior 方法[22]加以解决。本文主要创新点如下:

(1) 构建了 LDA 主题模型和社区发现之间的桥梁,可以实现更贴近语义、更加细致的社区划分;

(2) 传统的重叠社区发现算法得到的是节点属于哪几个社区的明确结果,挖掘出的重叠节点一方面数量较少,另一方面是其归属的重叠社区数目存在较大限制,而本文得到的是节点的重叠社区分布概率,这两方面问题都得到了解决;

(3) 针对网络数据稀疏性问题,我们在主题模型中运用 Spike and Slab prior 方法进行了有效解决。

## 1　相关工作

复杂网络上的社区发现,对于分析理解网络中的拓扑结构具有重要意义。研究者们提出了大量的重叠社区发现算法。代表算法包括派系过滤方法[10]、基于局部扩展算法[11]以及基于 Agent 方

法[12]等。CPM(Clique Percolation Method)[10]算法是派系过滤经典算法之一,该算法认为社区内部节点间连接紧密,边密度高,可以通过寻找网络中的完全子图(派系)来发现社区,但是 CPM 算法不适合处理稀疏矩阵。基于局部扩展算法中比较有代表性的是 Lancichinetti 等[11]提出的 LFM(Local Fitness Maximization)算法,该方法基于节点适应度的思想来进行重叠社区划分,从任意一个种子节点出发来扩展社区,合并邻居节点或删除当前社区内部节点直至社区适应度值不再增长,此时停止本轮扩展;然后再随机从另外一个尚未被划分的节点不断扩展,当所有节点都被划分社区后算法终止。这种方法比较灵活,但由于种子节点的选择比较随机,可能会造成不稳定的结果。基于 Agent 算法主要采用多标签传播方式进行社区发现,该算法赋予节点多个标签,从而每个节点携带多个社区的信息。Gregory[12]提出的 COPRA(Community Overlap Propagation Algorithm)算法即属于此类,该算法每步迭代过程中通过计算节点对不同标签的隶属系数来得到每个节点的标签与隶属程度的关系对。但是这种算法有时候准确度不是很高,容易产生较小的社区结构。

主题模型是一种挖掘文本背后隐含主题的统计模型,通过对目标数据降维,寻找出数据中的语义信息,帮助快速理解文本中的内容。Blei 等[17]提出的 LDA(Latent Dirichlet Allocation)是一种经典的无监督主题模型,只要事先指定好主题的数量,便能自动对文本进行主题分类。在 LDA 主题模型之后,很多基于 LDA 思想的改进模型被提出。例如 Steyvers 等[14]提出的 AT(Author-Topic Model)模型,该模型主要适用于文献挖掘中文献作者的建模,它提出将作者的兴趣与文本内容结合在一起。作者有哪些兴趣通常就会使用某些词,每个作者对应一个主题分布,这是对作者兴趣的一种建模方法。Blei 等[18]提出的 CTM(Correlated Topic Model)模型,主要用于文本信息中主题相关性建模,它指出在实际应用中主题之间并不是毫无联系,而是相关的。Iwata 等[19]提出的 TTM(Topic Tracking Model)模型,主要用于跟踪消费者购物行为,分析用户消费兴趣从而方便对消费者进行商品推荐。Zuo 等[20]提出的 PTM(Pseudo-document-based Topic Model)模型,该模型提出伪文档的概念,将几个短文本聚合成一个伪文档来解决数据稀疏问题。Wang 等[21]提出的 TTM(Targeted Topic Model)模型,该模型给定一个语料库,语料库中包含大量文档,能够根据指定的关键词发现用户感兴趣的相关主题。

从上述分析中可以看出,目前主题模型应用在社区发现方面的研究较少,而主题模型中文本的多主题性质和网络节点的多社区归属性质具有相似性。本文结合主题模型特点以及数据稀疏处理方法,提出了基于节点内容的重叠社区发现方法 OCD_LDA(Overlapping Community Detection algorithm based on LDA),有效地解决了传统社区发现算法中忽略文本内容信息的缺陷以及重叠社区归属分布问题。

## 2 基础知识

### 2.1 社区发现

复杂网络作为一种具有特定拓扑的结构图,大都具有社区结构的特性。社区可以看成是网络图中的特殊子图,也常称为社群、社团或者凝聚子群,是复杂网络中常见的一种深层次特性。令图 $G=\langle V,E\rangle$ 表示一个特定的社会网络,其中,$V=\{v_1,v_2,\cdots,v_n\}$ 表示网络中节点的集合,每个节点 $v_i$ 代表一个个体,$E=\{e_{ij}\mid i,j\in V,i\neq j\}$ 则是网络中的边集。节点数目为 $n$,边数为 $m$,即 $|V|=n$,$|E|=m$。

传统的非重叠社区发现是对 $G$ 中所有节点的一个划分,其定义如下:

若节点集合 $C_1,C_2,\cdots,C_k$ 满足下面的几个条件:

$$C_1\cup C_2\cup\cdots\cup C_k=G$$

$$C_i\cap C_j=\varnothing,\forall i\neq j$$

$$Pr(V_{C_i},V_{C_i})>Pr(V_{C_i},V_{C_j}),\forall i\neq j$$

则称 $C=\{C_1,C_2,\cdots,C_K\}$ 是网络 $G$ 的一个划分,每个集合 $C_i$ 代表一个社区,$Pr(V_{C_i},V_{C_i})$ 表示 $C_i$ 内部两个节点存在边的概率,$Pr(V_{C_i},V_{C_j})$ 表示 $C_i$ 和 $C_j$ 之间节点存在边的概率。

### 2.2 重叠社区发现

网络中不仅存在节点与社区一一对应的情况,还有可能一个节点同时出现在多个社区中,这种社区结构称作重叠社区。复杂网络图 $G=\langle V,E\rangle$ 中,若节点 $v_i$ 满足以下条件:

$$\exists m\neq n,v_i\in C_m\wedge v_i\in C_n$$

则称 $v_i$ 节点为重叠节点,社区 $C_m$ 和 $C_n$ 为重叠社区。

图 1 给出了包含 2 个重叠节点的 3 个社区结构图,其中节点 $A$ 是重叠社区 1 和社区 2 之间共享的重叠节点,节点 $B$ 是重叠社区 1 和社区 3 之间共享的重叠节点。复杂网络中的重叠社区可以更好地表

示出社区与社区之间的关系,但是,传统的重叠社区发现算法都是基于网络的拓扑结构,即根据节点之间的连接关系来划分社区,这些算法形成的社区形状往往具有非凹性,因此难以解决高维网络空间的重叠社区划分问题,不能细致地刻画出重叠节点在多个重叠社区中的不同归属程度。

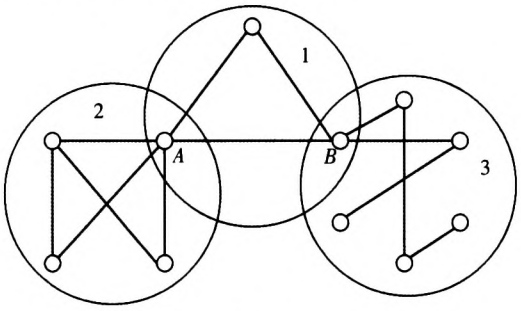


**图 1　重叠社区图**

### 2.3　LDA 模型

在语义关联挖掘过程中,两篇文档是否相关通常不仅仅取决于字面上的词语重复,还涉及到文字背后的语义关联性。挖掘文字背后的语义关联可以使搜索更加智能化,而主题模型则是用于解决此类问题的重要模型。主题模型最早源于 Deerwester 等人提出的 LSI(Latent Semantic Indexing)主题模型概念,Hofmann 对此模型进行改进,提出了 PLSI(Probabilistic Latent Semantic Indexing)主题模型。不过,LSI 和 PLSI 概率模型不够完备,缺乏统计基础,并且计算量较大。因此,Blei 等人在此基础上提出了 LDA(Latent Dirichlet Allocation)主题模型,LDA 是一种无监督模型,具有很好的扩展性,目前许多主题模型方法都是基于 LDA 模型扩展而来,图 2 是 LDA 模型生成过程。

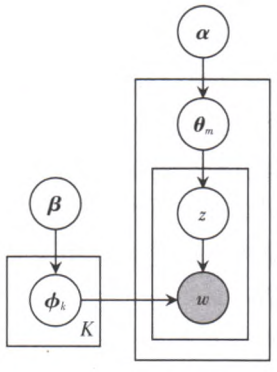


**图 2　LDA 模型**

(1) $\boldsymbol{\alpha}\rightarrow\boldsymbol{\theta}_m\rightarrow z_{m,n}$这一过程表示对于语料库中每一篇文档 $m\in M$,首先由公式 $\boldsymbol{\theta}_m\sim Dir(\boldsymbol{\alpha})$,得到每篇文档 $m$ 上的主题多项式分布,该分布的先验分布是参数为 $\boldsymbol{\alpha}$ 的 Dirichlet 分布,其中参数 $\boldsymbol{\alpha}$ 也被称为超参数;然后再根据 $z_{m,n}\sim Multinomial(\boldsymbol{\theta}_m)$,生成文档 $m$ 中词 $n$ 的主题 $z_{m,n}$。

(2) $\boldsymbol{\beta}\rightarrow\boldsymbol{\phi}_k\rightarrow w_{m,n}\mid k=z_{m,n}$这一过程表示对于文档 $m$ 中的每一个主题 $k\in K$,由公式 $\boldsymbol{\phi}_m\sim Dir(\boldsymbol{\beta})$,得到主题 $k$ 上的词多项式分布,该分布的先验分布是参数为 $\boldsymbol{\beta}$ 的 Dirichlet 分布,其中 $\boldsymbol{\beta}$ 同样被称为超参数;最后由公式 $w_{m,n}\sim Multinomial(\phi_{z_{m,n}})$ 生成文档 $m$ 中主题 $k$ 下的词 $n$。

因为 LDA 的目标是对语料库中的每个词进行判断,分析其潜在的主题,所以需要计算后验概率:

$$p(z\mid w)=\frac{p(z,w)}{p(w)}$$

本文使用 Gibbs Sampling 推理方法,在 Gibbs Sampling 算法中,假设已知词 $w_i=t$,根据贝叶斯法可以得到:

$$p(z_i=k\mid z_{-i},w)\propto$$

$$p(z_i=k,w_i=t\mid z_{-i},w_{-i})=\hat{\theta}_{mk}\cdot\hat{\phi}_{kt}$$

根据 Dirichlet 分布参数估计公式得到:

$$\hat{\theta}_{mk}=\frac{n_{m,-i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_{m,-i}^{(t)}+\alpha_k)} \tag{1}$$

$$\hat{\phi}_{kt}=\frac{n_{k,-i}^{(t)}+\beta_t}{\sum_{t=1}^{V}(n_{k,-i}^{(t)}+\beta_i)} \tag{2}$$

有了式(1)、式(2)之后,便可得到最终的 Gibbs sampling 公式。

在 LDA 模型中,利用 Gibbs sampling 可以在语料库中训练 LDA 模型,同时将其运用到主题语义分析中。首先,对于语料库中每个词随机赋予一个主题编号 $z$;然后遍历整个语料库,对其中的每个词通过 Gibbs sampling 公式重新计算,得到新的主题编号;重复上述过程,直至 Gibbs sampling 收敛。此时,即可得到最终的词与主题对应的概率矩阵。

## 3　基于 LDA 主题模型的重叠社区发现算法

### 3.1　问题描述

我们将重叠社区发现问题进一步细化为重叠社区归属分布问题,给定大小为 $M$ 的网络节点集合 $\{a_i\}$,社区个数 $K$,每个节点 $a_i$ 上的内容属性值集合 $\{w_i\}$ 以及超参数 $\boldsymbol{\beta}$、$\bar{\boldsymbol{\alpha}}$、$\gamma_0$,节点的内容属性值可以来自节点的标签或者节点与节点间交流的内容信息。通过对每个节点 $a_i$ 上的内容属性值集合 $\{w_i\}$

进行建模,从而得到节点 $i$ 上的社区归属分布向量 $\boldsymbol{\theta}_i$。定义概率分布矩阵 $\boldsymbol{\theta}=\{\boldsymbol{\theta}_i \mid i=1,2,\cdots,M\}$,其中节点 $i$ 的社区归属概率为 $\boldsymbol{\theta}_i=(p_1,p_2,\cdots,p_j,\cdots,p_k)$,$p_j$ 表示节点 $i$ 归属社区 $j$ 的概率,满足 $\sum_{j=1}^{k} p_j=1$。得到概率分布矩阵 $\boldsymbol{\theta}$ 后,便可获知每个节点上的重叠社区归属分布。

### 3.2 基于 LDA 的重叠社区发现算法

OCD_LDA 算法由两部分组成,如图 3 所示。模型左半边部分为基本的 LDA 主题模型,主要是对网络中节点的内容属性信息建模,得到节点在重叠社区上的归属分布概率。考虑到网络中节点的内容较为短小,在文本建模过程中易造成数据稀疏问题,使得社区归属分布结果不精确,因此模型的右半边部分通过引入 Spike and Slab prior 方法来帮助进行变量选择和参数估计,解决节点上社区分布的稀疏性和平滑性问题。其中,社区选择器 $b$ 是一个二进制变量,服从伯努利分布,它能够反映社区 $c$ 是否和个体 $x$ 相关,其详细说明见 3.3 节。

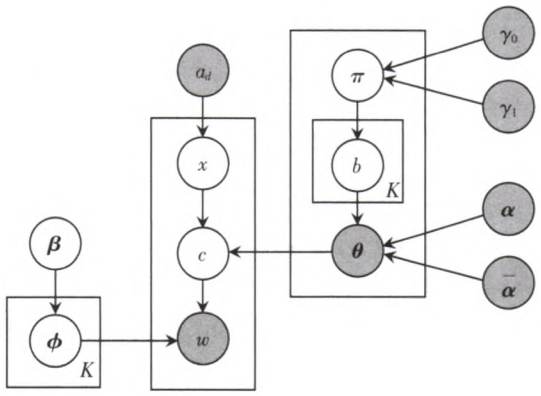


**图 3 OCD_LDA 算法**

表 1 是图 3 中的符号定义,无填充色的圆形内表示的符号为随机变量,填充灰色的圆形内符号则代表可以观测到的变量。

该模型执行流程首先从节点集合 $\{a_i\}$ 中选取一节点 $x$,然后根据节点上的社区分布 $\boldsymbol{\theta}$ 给节点 $x$ 分配社区 $c$。Spike and Slab prior 方法通过二进制社区选择器变量 $b$ 来反映社区 $c$ 是否和节点 $x$ 相关。完成社区分配后,便可以根据社区上的词分布,生成内容属性值 $w$。有了此生成模型后,即可通过 Gibbs sampling 公式不断地迭代抽样出 $\boldsymbol{\theta}$,从而进一步分析得出节点在重叠社区上的归属分布。接下来,3.3 节将给出模型的详细生成过程,3.4 节为模型中公式的推理过程,最后给出算法整体流程。

**表 1 符号定义说明**

| 参数 | 定义 |
|---|---|
| $\{a_i\}$ | 节点集合 |
| $\boldsymbol{\alpha},\boldsymbol{\beta}$ | 超参 |
| $K$ | 社区个数 |
| $x$ | 集合 $\{a_i\}$ 中一个节点 |
| $c$ | 社区 |
| $w$ | 节点中的属性词 |
| $\boldsymbol{\theta}$ | 节点上的社区多项式分布 |
| $\boldsymbol{\phi}$ | 社区上的词多项式分布 |
| $\gamma_0,\gamma_1$ | Beta 分布超参数 |
| $\pi$ | 伯努利分布参数 |
| $b$ | 社区选择器 |

### 3.3 生成过程

假设共有 $K$ 个社区 $\{\boldsymbol{\phi}_c\}_{c=1}^{K}$ 和 $D$ 个节点,用多项式分布 $\boldsymbol{\theta}$ 来模拟社区在节点上的分布,用多项式分布 $\boldsymbol{\phi}$ 来模拟属性值在社区上的分布。节点中的每个属性值生成过程首先根据节点上的社区分布 $\boldsymbol{\theta}$ 生成社区 $c$,然后再由社区上的属性值分布 $\boldsymbol{\phi}$ 生成属性值 $w$。

本文提出的模型符号图见图 3 所示,算法 1 是 OCD_LDA 模型左半边部分的生成过程:

**算法 1** 模型左半边部分生成过程

输入:节点,$\boldsymbol{\alpha},\boldsymbol{\beta}$

输出:节点属性值

1. 对于每个社区:采样 $\boldsymbol{\phi}_c \sim Dir(\boldsymbol{\beta})$
2. 对于每个节点:采样 $\boldsymbol{\theta}_x \sim Dir(\boldsymbol{\alpha})$
3. 对于 $\{a_i\}$:
4. (a)从 $\{a_i\}$ 中选择一个体 $x$
5. (b)对于集合中的每个属性值 $w_i$:
6. i. 采样一个社区

$$c \sim Multi(\boldsymbol{\theta}_x)$$

7. ii. 采样第 $i$ 个属性值

$$w_i \sim Multi(\boldsymbol{\phi}_c)$$

由于相比于主题建模的文档规模,网络中节点所携带的内容信息比较少,在建模的过程中可能会产生数据稀疏的缺陷,因此,引入了 Spike and Slab prior[22] 方法。结合图 3 中模型图,下面给出社区选择器、平滑先验和弱平滑先验相关概念以便更好地理解我们的模型。

**定义 1** 社区选择器:对于每一个个体 $x$,社区选择器 $b_{x,c}, c\in\{1,2,\cdots,K\}$ 是一个二进制变量,其反映社区 $c$ 是否和个体 $x$ 相关。$b_{x,c}$ 由分布 $Bernoulli(\pi_x)$ 采样生成,$\pi_x$ 是一个伯努利参数。

**定义 2** 平滑先验:平滑先验是一个狄利克雷超参数 $\boldsymbol{\alpha}$,用来平滑化社区选择器选中的社区划分。

**定义 3**　弱平滑先验:弱平滑先验是另外一个狄利克雷超参数 $\bar{\alpha}$,它被用作平滑未被社区选择器选中的社区。由于 $\bar{\alpha} \ll \alpha$,所以超参数 $\bar{\alpha}$ 被称为弱平滑先验。

在 OCD_LDA 算法中,Spike and Slab prior 方法中 Spikes 通常指的是社区选择器,而平滑先验和弱平滑先验则与 Slabs 相关,社区选择器 $b_x = \{b_{x,c}\}_{c=0}^{K}$,节点上的社区分布由 $Dir(\alpha b_x + \bar{\alpha} x)$ 采样生成。

引入 Spike and Slab prior 后,OCD_LDA 模型右半边部分生成过程如算法 2:

**算法 2**　模型右半边部分生成过程

输入:节点,$\gamma_0$,$\gamma_1$,$\alpha$

输出:$b$,$\theta$

1. 采样 $\pi_x \sim Beta(\gamma_0, \gamma_1)$
2. 对于每个社区 $c$:
3. i. 采样社区选择器,

$$b_{x,c} \sim Bernoulli(\pi_x), b_x = \{b_{x,c}\}_{c=0}^{K}$$

4. 采样 $\theta_x \sim Dir(\alpha b_x + \bar{\alpha} x)$

### 3.4　推理过程

由于精确的后验推断计算量过大,因此我们采用了 collapsed Gibbs sampling 算法来推理 OCD_LDA 算法中的参数。Collapsed Gibbs sampling 是 MCMC (Markov Chain Monte Carlo) 方法的一种,它需要完成的是对 $P(z|w, alpha, beta)$ 的抽样,然后利用抽样结果通过简单的似然估计方法求得 $\theta$ 和 $\phi$。Collapsed Gibbs sampling 方法虽然推理起来比较复杂,但是实现简单。我们将直接给出本文使用的采样公式,对推理的细节不做具体阐述,具体可参见文献[24]。

为了得到 $\theta$,$\phi$ 和 $\pi$ 的采样公式,分析本文模型可知采样算法需要的潜在变量是社区分配器 $c$ 和社区选择器 $b$。此外,还需要采样 Dirichlet 超参 $\alpha$ 和 Beta 超参 $\gamma_1$,我们设置 $\bar{\alpha}$ 值为 $10^{-12}$,$\gamma_0$ 值为 1。

首先采样 Community assignments $c$。Community assignments $c$ 的采样方法和潜在狄利克雷分配有些类似,不同之处在于 $\theta$ 不再属于原始的短文本,而是属于网络中的节点。并且 $\theta$ 也不再直接由超参 $\alpha$ 采样所得,而是从 Spike and Slab prior 中采样。采样公式如下:

$$p(c_{m,i} = c \mid rest) \propto (N_m^c + b_{m,c}\alpha + \bar{\alpha}) \frac{N_c^{w_{m,i}} + \beta}{N_c + V\beta} \tag{3}$$

其中,$N_c^w$ 表示词 $w$ 在社区 $c$ 中出现的次数,并且 $N_c = \sum_{w=0}^{V} N_c^w$;$b_{m,c}$ 表示社区选择器,值为 1 即意味着社区 $c$ 被节点 $m$ 选择。

其次,采样社区选择器 $b$,为了采样 $b_m$ 向量,我们参照了 Wang 等在文献[25]中提出的抽样方法,即引入辅助变量 $\pi_m$,通过聚合变量 $\pi_m$ 实现对 $b_m$ 的采样。定义在节点 $m$ 中出现的社区集合为 $B_m = \{c: N_m^c > 0, c \in \{1, \cdot, \cdot K\}\}$,接下来给出 $\pi_m$ 和 $b_m$ 的联合条件分布:

$$p(\pi_m, b_m \mid rest) \propto \prod_c p(b_{m,c} \mid \pi_m) p(\pi_m \mid \gamma_0, \gamma_1) \frac{I[B_m \in A_m] \Gamma(|A_m|\alpha + K\bar{\alpha})}{\Gamma(N_m + |A_m|\alpha + K\bar{\alpha})} \tag{4}$$

其中,$I[B_m \in A_m]$ 表示指示函数,若 $B_m \in A_m$ 函数返回 1,否则函数返回 0。$A_m = \{c: b_{m,c} = 1, c \in \{1, 2, \cdots, K\}\}$ 表示向量 $b_m$ 值为 1 的集合,$|A_m|$ 即为集合 $A_m$ 的大小。

有了该联合条件分布,便可以迭代地对 $b_m$ 进行采样,以 $\pi_m$ 为条件,最终从 $\pi_m$ 得到 $b_m$ 的采样。对于 Dirichlet 超参 $\alpha$,我们使用 Metropolis-Hastings 算法对其进行采样;对于 Beta 超参 $\gamma_1$,使用 Gamma priors[23] 方法进行采样。

至此,我们已经基本阐明了本文模型的 collapsed Gibbs sampling 算法。有了这些潜在变量之后,就可以估计出 $\theta$ 和 $\phi$ 计算公式:

$$\theta_{m,c} = \frac{N_m^c + b_{m,c}\alpha + \bar{\alpha}}{N_m + |A_m|\alpha + K\bar{\alpha}} \tag{5}$$

$$\phi_{c,w} = \frac{N_c^w + \beta}{N_c + V\beta} \tag{6}$$

最后,给出 OCD_LDA 算法整体流程。

**算法 3**　OCD_LDA 算法整体流程

输入:节点属性集,$K$,$\beta$,$\bar{\alpha}$,$\gamma_0$,Niter

输出:$\theta$ 和 $\phi$

1. 为所有节点随机初始化社区分配
2. 执行迭代 Niter 次:
3. 对于每一节点:
4. 对于每一个属性:
5. 根据式(2)计算 $b_{m,c}$
6. 根据式(1)计算 $c_{m,i}$
7. 更新 $N_m^c$,$N_c^w$
8. 计算式(3)、式(4)更新 $\theta$ 和 $\phi$
9. 输出 $\theta$ 和 $\phi$ 的最终后验估计

## 4 实验结果分析

考虑数据集采集的便利性,实验采用 DBLP 文献数据集中的一个子集,数据覆盖 4 个研究领域:数据库(SIGMOD, ICDE, VLDB, EDBT, PODS, ICDT, DASFAA, SSDBM, CIKM)、数据挖掘(KDD, ICDM, SDM, PKDD, PAKDD)、人工智能(IJCAI, AAAI, NIPS, ICML, ECML, ACML, IJCNN, UAI, ECAI, COLT, ACL, KR)和计算机视觉(CVPR, ICCV, ECCV, ACCV, MM, ICPR, ICIP, ICME),共 30 422 篇文章。将每篇文章看作一个节点,文章的摘要作为节点的属性信息,因此,30 422 篇文章构成 30 422 个节点组成的网络图。文章一般存在主题交叉和领域交叉现象,我们希望发现其中的重叠社区情况。首先对原始数据集进行预处理,即对所有的文档进行分词处理,去除停用词,包括语气助词和代词等出现频率很高,但是对社区挖掘没有帮助的词。

实验采取 F-measure 作为综合评价指标,该值是准确率(Precision)和召回率(Recall)的加权调和平均:

$$F=\frac{(\alpha^2+1)P*R}{\alpha^2(P+R)}$$

其中,$P$ 指标和 $R$ 指标分别代表准确率和召回率,即:

$$P=\frac{\text{提取出的正确信息条数}}{\text{提取出的信息条数}}$$

$$R=\frac{\text{提取出的正确信息条数}}{\text{样本中的信息条数}}$$

当参数 $\alpha=1$ 时,就是最常见的 $F1$ 指标,即 $F1=\frac{2*P*R}{P+R}$,本文采取的也是 $F1$ 指标,通常 $F1$ 较高时说明算法性能比较优越。

### 4.1 算法评估

#### 4.1.1 算法效果对比

为了对比 OCD_LDA 算法的性能,选择与 K-means 算法进行对比分析。由于 K-means 算法作为数据挖掘领域重要算法之一,可运用在多种情景,并且在社区挖掘中也有良好的性能,因此 K-means 算法是一个合适的对比算法。

将 OCD_LDA 算法与 K-means 算法分别运用在 DBLP 文献数据集上,记录在设置不同社区个数的情况下迭代 1 000 次所得的社区挖掘结果的 $F$ 值变化情况,如图 4 所示。我们可以看出两种算法的 $F1$ 指标都是先增加,然后达到最大值后减小,均在社区个数为 11 的时候取得最大值,即两种方法都可以发现最佳社区数 11。本文提出的 OCD_LDA 算法 $F1$ 曲线位于上方,表现效果优于 K-means 算法。在社区个数为 11 时,$F1$ 指标为 0.46,相对于 K-means 算法的 $F1$ 指标值 0.41,提高将近 12%。

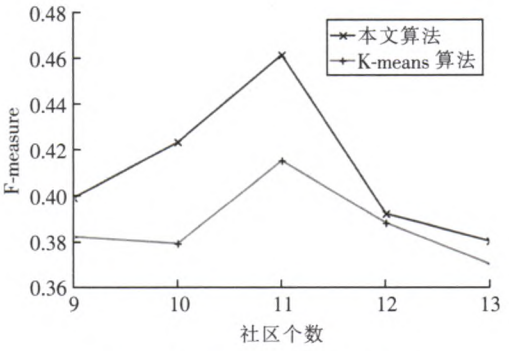


图 4 不同算法的 $F1$ 值对比

#### 4.1.2 算法的收敛性

为了评估本文算法的收敛性,实验设置了不同聚类个数,并分别在不同迭代次数下运行,记录所得结果的 $F1$ 值。从图 5 可以看出,OCD_LDA 算法迭代次数达到 700 次后,$F1$ 值达到最大值并开始趋于平稳,并且社区个数为 11 的效果最佳。所以,在后续实验中,设置迭代次数为 1 000 次,社区个数为 11,以保证划分结果的精准性。需要强调的是,和 K-means 算法不同,OCD_LDA 方法发现的社区是重叠社区,社区的重叠情况还需要进一步分析。

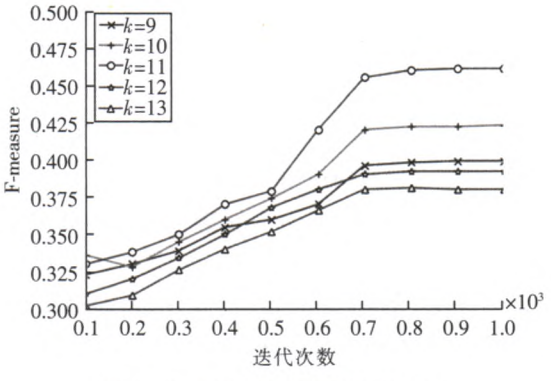


图 5 不同迭代次数下 $F1$ 值曲线

### 4.2 结果分析

OCD_LDA 算法运用在 DBLP 文献数据集中得到每篇文章的社区归属分布情况,节点在每个归属社区上都有一个 0 到 1 之间的归属概率。如果考虑节点在社区中归属概率只要大于 0,节点就有效归属该社区,那么每个社区拥有节点数量占总节点比例的情况如图 6 所示。

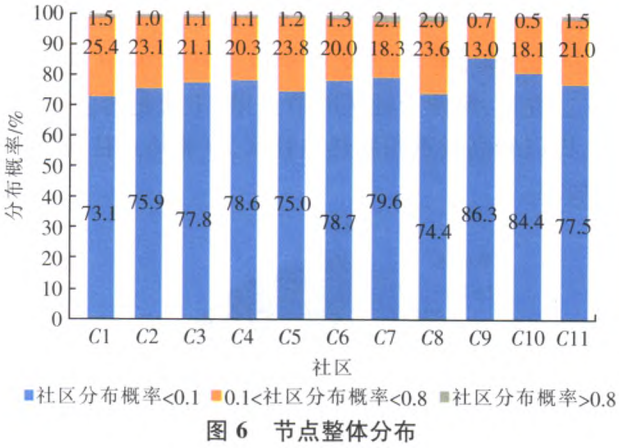


图 6　节点整体分布

图 6 显示所有社区都拥有几乎所有的节点,或者说几乎所有的节点都不同程度地归属于所有的社区,这个显然是不合理的。分析其原因,在主题模型中太小的归属概率是没有实际意义的,因此需要设置一个合理的阈值进行过滤。下面我们通过对节点的社区分布概率来设置一个合理的归属概率阈值。

我们随机选取了 4 个节点,将其归属的所有 11 个社区归属程度按照从大到小排序,如表 2 所示,可以发现,*N*1 节点社区归属概率最大的前三分别为 0.440 16、0.418 22、0.132 82,而从下一个社区开始归属概率仅为 0.001 10,相对来说很小。并且其他几个节点都在社区归属概率为 0.1 时发生较大落差,社区归属分布中概率小于 0.1 的值均很小,基本可以忽略不计。因此,本文实验取 0.1 为社区归属分布界值,即设定当节点在某个社区中的归属概率大于 0.1 时,才认为节点有效归属该社区。此外,当节点在某个社区中归属概率小于等于 0.1 时,该节点归属该社区的概率相对来说非常小,可以忽略不计,视为无效归属。此外,设定当节点在某个社区中归属概率大于 0.8 时,认为节点只属于该社区,即该节点不是重叠节点。

**表 2　抽样节点社区归属分布情况**

| 社区节点 | 11 社区的归属概率分布(按照归属概率大小) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| *N*1 | 0.440 16 | 0.418 22 | 0.132 82 | 0.001 10 | 0.001 10 | 0.001 10 | 0.001 10 | 0.001 10 | 0.001 10 | 0.001 10 | 0.001 10 |
| *N*2 | 0.429 15 | 0.215 09 | 0.164 12 | 0.153 92 | 0.031 60 | 0.001 02 | 0.001 02 | 0.001 02 | 0.001 02 | 0.001 02 | 0.001 02 |
| *N*3 | 0.500 92 | 0.242 14 | 0.131 24 | 0.057 30 | 0.057 30 | 0.001 85 | 0.001 85 | 0.001 85 | 0.001 85 | 0.001 85 | 0.001 85 |
| *N*4 | 0.877 11 | 0.091 21 | 0.003 52 | 0.003 52 | 0.003 52 | 0.003 52 | 0.003 52 | 0.003 52 | 0.003 52 | 0.003 52 | 0.003 52 |

按照分析得到的有效社区归属度 0.1 和单社区归属度 0.8 的设置阈值进行统计。从图 6 可以看出,各个社区中节点归属概率在 0.1 ~0.8 间的比例分别为 25.4%、23.1%、21.1%、20.3%、23.8%、20%、18.3%、23.6%、13%、18.1%、21%。依据有效社区归属度 0.1 进行筛选去除了大量噪声,有效归属社区所含节点比例共为 227.7%,重叠率为 147.1%,这是由于存在重叠节点同时归属于不同社区,但是这样的重叠现象是少数节点造成的还是大部分节点造成的,需要我们进一步分析。

在对节点的社区归属分布进一步分析过程中,发现绝大多数节点都是跨社区分布的,只有少部分节点属于单个归属社区。其次,在节点跨社区分布中,也存在多种情况。我们称归属权重大的社区为强社区,而权重较小的为弱社区。强社区也可以理解为热点社区,该社区中节点数量较多;弱社区则为非热点社区,该社区中节点数量较少。通过分析,发现节点社区归属分布情况可以分为四类:单社区、强-强社区、强-弱社区以及三社区分布,下面具体说明这四种类型。

从图 7 中可以清晰地看出 4 类节点社区归属分布。第一类节点只归属于一个社区,例如图 7(a)中节点在 *C*1 社区中的概率为 0.916 81,接近 1,那么其在另外几个社区中的概率基本为 0;第二类节点基本只归属于两个社区,并且权重相差不大,属于强-强社区。例如图 7(b)中节点属于 *C*2 社区的概率为 0.530 69,属于 *C*11 社区的概率为 0.398 32,而归属其他社区的概率非常小,可以忽略不计。因此,总的来说这类节点只归属于两个社区;第三类虽然也归属于两个社区,但在各自社区中的概率相差较大,属于强-弱社区。例如图 7(c)中节点归属于社区 *C*4 的概率为 0.577 73,而在社区 *C*5 中的概率只有 0.232 25;第四类节点主要归属 3 个社区,并且归属权重相似。比如图 7(d)中节点归属 *C*3 社区概率为 0.290 02,归属 *C*8 社区概率为 0.302 06,归属 *C*10 社区概率为 0.277 98。

根据统计结果,将单个社区归属以及跨社区分布情况绘制成饼状图以便更清晰地展示整体节点的归属分布。

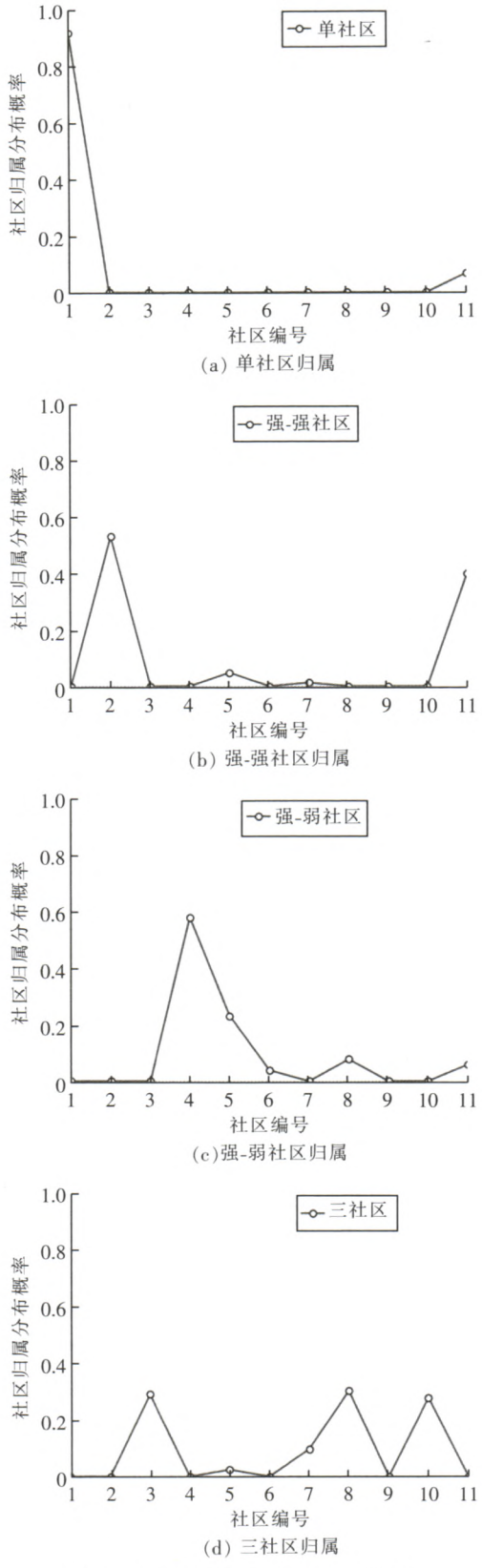


**图 7 节点社区归属分布的 4 种类型**

从图 8 中可以看出,仅有 14.1% 的节点归属单个社区,85.9% 的节点都是跨社区分布的,说明一些传统的社区划分方法中直接给节点贴上单个标签的方法是不合理的。而在跨社区分布中,跨 2 个社区的节点约占 66.3%(强强 + 强弱),跨 3 个社区的节点占 16.6%,跨 4 个及 4 个以上社区的节点仅有 3%。并且,跨两个社区的节点里,主流是跨强-弱社区,即以较高权重归属热点社区,在非热点社区中权重较小。

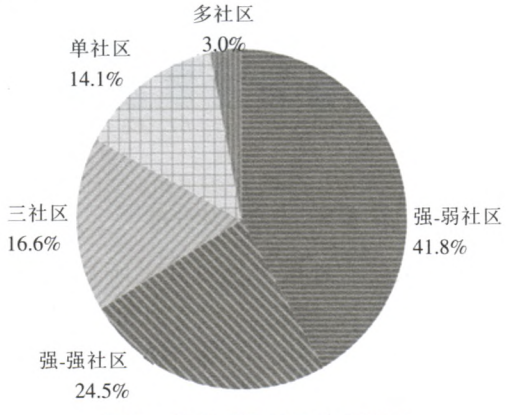


**图 8 节点重叠社区分布情况**

获取节点在各个社区中的归属分布概率之后,对每个节点内部分析,挖掘出每一个社区所涉及的主题信息。表 3 展示了每个社区内部涉及的关键属性词,而通过这些属性词即可大致判断出该社区的类别。

从表 3 中可以看出,社区 *C*1 涉及的主要为优化算法、人工智能领域的搜索,以及功能和模板;社区 *C*2 涉及的主要为基于 web 的用户数据的收集或分析等;社区 *C*3 涉及的主要为数据挖掘领域的聚类算法;社区 *C*4 涉及的主要为数据库领域中的数据库性能以及管理等问题;社区 *C*5 涉及的主要为人工智能领域的知识库系统和推理系统;社区 *C*6 涉及的主要为数据库领域的各类查询问题。社区 *C*7 涉及的主要为图像处理和检索等问题;社区 *C*8 涉及的主要为数据挖掘领域中的文本分类问题;社区 *C*9 涉及的主要为神经网络问题,如社交网络、贝叶斯网络和神经网络等;社区 *C*10 涉及的主要为计算机视觉领域中的视频图像认知等问题;社区 *C*11 涉及的主要为数据挖掘领域中的特征选择和特征向量等问题。

结合图 7、8 中节点归属分布情况,我们得出以下结论:跨社区中的交叉领域主要涉及同一大类中的研究问题,人工智能领域中的搜索、优化算法和知识库系统、推理系统易产生交叉;数据挖掘领域中的聚类算法和文本分类易产生交叉;数据库领域中的数据库管理和数据库查询易产生交叉。其次,不同领域间的某些研究方向也会产生交叉。人工智能领域中的搜索优化算法和数据挖掘领域中的文本分

类、数据挖掘领域中的聚类算法和计算机视觉领域中的图像认知以及数据库领域中聚类算法和社会神经网络间都易产生交叉研究。

总体上,整体节点均匀分布在各个社区中,每个社区内部节点数量相差不大,但节点在各个社区中归属权重不同。通过 DBLP 文献数据集的分析结果,可以发现当前科研论文研究方向以及交叉领域,这对于发现新的研究热点具有指导意义。

**表 3　社区主要属性词**

| 社区 | 主要属性词/权重 | 主题 |
|---|---|---|
| C1 | algorithm 0.155 74,learning 0.011 69 , search 0.007 08,optimal 0.005 89, function 0.004 93, pattern 0.004 67 | 学习算法优化 |
| C2 | data 0.022 00,web 0.015 90, user 0.010 74,users 0.007 27, semantic 0.004 07,content 0.003 92 | 基于 web 的用户数据分析 |
| C3 | data 0.038 75,clustering 0.014 99, mining 0.014 88,algorithm 0.0106 02, clusters 0.008 25, model 0.007 55 | 聚类算法 |
| C4 | data 0.027 37,database 0.012 56, system 0.010 61,distributed 0.008 60,performance 0.007 60, management 0.005 42 | 数据库性能以及管理 |
| C5 | system 0.013 88, knowledge 0.011 06, reasoning 0.005 88,systems 0.005 50, intelligent 0.004 73, domain 0.004 14 | 知识库系统和推理系统 |
| C6 | query 0.031 06, data 0.018 77, queries 0.016 71,xml 0.010 15, database 0.006 85, efficient 0.006 75 | 数据库查询 |
| C7 | image 0.020 82, images 0.011 03, retrieval 0.007 41,analysis 0.006 86, shape 0.006 54, database 0.005 96 | 图像处理和检索 |
| C8 | classification 0.011 87, method 0.009 92, model 0.008 45,training 0.008 34, text 0.008 09, data 0.006 76 | 文本分类 |
| C9 | network 0.025 67, bayesian 0.012 38, markov 0.012 11,neural 0.009 59,social 0.009 06, networks 0.008 41 | 神经网络 |
| C10 | recognition 0.013 72, object 0.012 64, features 0.009 19,image 0.008 33, video 0.008 32, detection 0.008 22 | 视频图像认知 |
| C11 | feature 0.015 22, classification 0.010 66, logic 0.006 69, selection 0.006 67, structure 0.006 58, vector 0.005 58 | 特征选择、特征向量 |

## 5　结束语

在本文中,我们提出了大规模网络中基于 LDA 主题模型的重叠社区发现算法,算法通过对网络中节点的内容属性进行主题建模,得到了更加细致的节点在各个社区上的归属分布概率矩阵,更好地揭示了数据内部组织特征。考虑网络的动态特性以及将网络的结构特性融合内容特征进行社区发现是我们未来的研究方向。

**参考文献:**


[1] FANI H, ZARRINKALAM F, BAGHERI E, et al. Time-sensitive topic-based communities on twitter[C]// Canadian Conference on Artificial Intelligence on Advances in Artificial Intelligence. 2016:192 - 204.

[2] AYNAUD T, FLEURY E, GUILAUME J L, et al. Communities in Evolving Networks: Definitions, Detection, and Analysis Techniques [M]. New York: Springer, 2013: 159 - 200.

[3] NEWMAN M E J. Communities, modules and large-scale structure in networks [J]. Nature Physics, 2012, 8(8): 25 - 31.

[4] FORTUNATO S. Community detection in graphs[J]. Physics Reports, 2010, 486(3/4/5): 75 - 174.

[5] KERNIGHAN B W, LIN S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970, 49(2): 291 - 307.

[6] FIEDLER M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal, 1973, 23(23): 298 - 305.

[7] GRIVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2001(99): 7821 - 7826.

[8] PONS P, LATAPY M. Computing communities in large networks using random walks[C]// International Symposium on Computer and Information Sciences. 2005: 284 - 293.

[9] SHANG Ronghua, BAI Jing, JIAO Licheng, et al. Community detection based on modularity and an improved genetic algorithm[J]. Physica A Statistical Mechanics & Its Applications, 2013, 392(5): 1215 - 1231.

[10] PALLA G, DERÉNYI I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814.

[11] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure of complex networks[J]. New Journal of Physics, 2008, 11(3): 19 - 44.

[12] GREGORY S. Finding overlapping communities in networks by label propagation[J]. New Journal of Physics, 2010, 12(10): 2011 - 2024.

[13] DING Y. Community detection: topological vs. topical[J]. Journal of Informetrics, 2011, 5(4): 498 - 514.

[14] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, et al. The author-topic model for authors and documents[C]// Conference on Uncertainty in Artificial Intelligence. 2004: 487 - 494.

[15] HONG L, DAVISON B D. Empirical study of topic model-

ing in Twitter[C]//SIGKDD Workshop on Social Media Analytics. 2010:80 - 88.

[16] TANG J, MENG Z, NGUYEN X L, et al. Understanding the limiting factors of topic modeling via posterior contraction analysis[C]//International Conference on Machine Learning. 2014:185 - 190.

[17] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993 - 1022.

[18] LAFFERTY J D, BLEI D M. Correlated topic models [C]//Advances in Neural Information Processing Systems. 2006:147 - 154.

[19] IWATA T, WATANABE S, YAMADA T, et al. Topic tracking model for analyzing consumer purchase behavior [C]//International Joint Conference on Artificial Intelligence. 2009:1427 - 1432.

[20] ZUO Y, WU J, ZHANG H, et al. Topic modeling of short texts: a pseudo-document view[C]//International Conference on Knowledge Discovery and Data Mining. 2016: 2105 - 2114.

[21] WANG S, CHEN Z, FEI G, et al. Targeted topic modeling for focused analysis[C]//International Conference on Knowledge Discovery and Data Minin. 2016: 1235 - 1244.

[22] ISHWARAN H, RAO J S. Spike and slab variable selection: frequentist and bayesian strategies[J]. Annals of Statistics, 2005, 33(2):730 - 773.

[23] TEH Y W, JORDAN M I, BEAL M J, et al. Hierarchical dirichlet processes[J]. Journal of the American Statistical Association, 2006, 101(476):1566 - 1581.

[24] GRIFFITHS T L, STETVERS M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(1):5228.

[25] WANG C, BLEI D M. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process[C]//International Conference on Neural Information Processing Systems. 2009:1982 - 1989.

## 作者简介:

**张 伟**(1973 - ),男,江苏泰兴人。南京邮电大学计算机学院副院长,教授。研究方向为网络信息安全、恶意代码分析和社会网络分析。