ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2013/07(08)-0718-11
doi: 10.3778/j.issn.1673-9418.1305046

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

ACT-LDA:集成话题、社区和影响力分析的概率模型*

吴 良[1,2],黄威靖[1,2],陈 薇[1,2,3+],王腾蛟[1,2,3],雷 凯[3],刘月琴[4]
1. 高可信软件技术教育部重点实验室,北京 100871
2. 北京大学 信息科学技术学院,北京 100871
3. 北京大学 深圳研究生院 深圳市云计算关键技术与应用重点实验室,广东 深圳 518055
4. 国际关系学院 信息科技系,北京 100091

ACT-LDA: A Probabilistic Model of Topic, Community and User Influence*

WU Liang[1,2], HUANG Weijing[1,2], CHEN Wei[1,2,3+], WANG Tengjiao[1,2,3], LEI Kai[3], LIU Yueqin[4]
1. Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing 100871, China
2. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
3. The Shenzhen Key Lab for Cloud Computing Technology and Applications, Shenzhen Graduate School, Peking University, Shenzhen, Guangdong 518055, China
4. Department of Information Science and Technology, University of International Relations, Beijing 100091, China
+ Corresponding author: E-mail: pekingchenwei@pku.edu.cn

WU Liang, HUANG Weijing, CHEN Wei, et al. ACT-LDA: a probabilistic model of topic, community and user influence. Journal of Frontiers of Computer Science and Technology, 2013, 7(8): 718-728.

Abstract: With the development of social network, users in all kinds of social networks form a large user graph. The articles published by users and the links between articles form a large document graph. According to the user graph and document graph, how to mine the topics, communities and user influence is an important problem, but now this problem is handled independently. Considering the inter-dependence of the problem and making use of text content and coauthor relationship, this paper proposes a joint model of topic modeling, community discovery and user influence analysis based on LDA (latent Dirichlet allocation), called ACT-LDA (author-community-topic LDA),

* The National High Technology Research and Development Program of China under Grant Nos. 2012AA011002, 2011AA010706 (国家高技术研究发展计划(863计划)); the Science and Technology Development Fund Project of Shenzhen under Grant No. ZYA201106080025A (深圳市科技发展资金项目).
Received 2013-04, Accepted 2013-06.
CNKI网络优先出版:2013-06-25, http://www.cnki.net/kcms/detail/11.5602.TP.20130625.1615.001.html.

which uses variation method for inference. The experiments on DBLP data show very good results, which validates the proposed model.

Key words: social network; community discovery; topic modeling; user influence

摘 要:随着社交网络的发展,社交网络中的用户形成大规模的用户关系图,用户在社交网络中发表内容,这些内容及其链接关系形成大规模的文档图。如何根据用户关系图、文档图,挖掘出用户所形成的社区、社区用户的影响力以及各个社区的话题,是重要的问题,而目前这些工作相对独立。考虑了用户发表内容、用户之间的关系信息,利用话题传播、社区形成和用户影响力之间的关联性,提出了一个基于LDA(latent Dirichlet allocation)的集成话题发现、社区发现和用户影响力分析的统一模型ACT-LDA(author-community-topic LDA)。模型采用变分推理的方法解决推理问题。在DBLP数据上进行了实验,取得了非常好的结果,证明了模型的有效性。

关键词:社交网络;社区发现;话题模型;用户影响力

文献标志码:A **中图分类号:**TP311

1 引言

随着社交网络的发展,越来越多的人活跃其中,用户及用户之间的关系形成了一个大规模的用户关系图;社交网络也越来越融入到人们的日常生活之中,越来越多的用户将社交网络作为真实世界的一部分,在社交网络中发表关于自己日常生活的内容,内容及内容之间的链接等关系形成了一个大规模的文档图。

如何根据用户关系图、文档图,发现社交网络中存在的社区,分析社区中用户的影响力和话题等,是非常重要的问题。因为用户往往活跃于各个社区中,在不同的社区中关注着不同的话题。例如,在"脸谱网"(https://www.facebook.com/)或者"人人网"(http://www.renren.com/)等社交网络中,用户的邻居包括校友、朋友、老乡、同事和家人等,形成了不同的社区。不同社区谈论着不同的话题,用户之间通过发表内容、转发内容等方式传播话题,不同用户对社区中话题的传播造成不同的影响。需要从数据中挖掘出社区、话题、社区话题分布、用户影响力等信息。

目前的工作中,话题发现、社区发现、用户影响力分析这三方面的工作在相对独立地进行。但是这些问题是相互关联的:用户关注话题的差异和用户之间关系强弱的差异,导致了社区的形成;由于社区内用户互动以及用户的影响力,话题才能够快速传播。基于话题、社区和用户影响力之间的相互关联性,采用统一建模方法可以同时进行用户层次和话题层次上细致的分析。可利用话题发现、社区话题分布分析使社区发现结果具有直观语义,利用社区发现和社区用户影响力分析,使话题发现结果增加可验证性,因此提出了一个基于LDA(latent Dirichlet allocation)的集成话题发现、社区发现、用户影响力分析的统一模型ACT-LDA(author-community-topic LDA)。

本文提出的模型适用于任何包含用户关系和用户发表内容的网络。其中,用户发表的内容称为文档,用户称为文档作者。本文在实验中以论文网络为例进行分析,实验结果说明本文模型具有如下特点:

(1)改进了话题生成的结果,使发现的话题之间区别度增加。

(2)改进了只基于用户关系图结构信息的方法的社区划分结果。

(3)社区发现结果与社区的话题分布结果完全对应,使社区发现的结果具有正确而又直观的语义。

(4)挖掘社区最具影响力作者的结果非常准确,也从侧面验证了社区发现结果、话题发现结果的正确性。

本文组织结构如下:第2章主要介绍相关工作;第3章详细介绍ACT-LDA模型及其贡献;第4章描述实验过程,并给出实验结果分析,验证ACT-LDA

模型的有效性;第5章对全文进行总结,同时对未来工作进行展望。

2 相关工作

2.1 社区发现

对于社区而言,不同社区用户之间关系较少,社区内用户之间关系较多。基于社区的定义,社区发现有很多种方法,包括:基于聚类的方法,例如K-means聚类[1]、垂直聚类[2-3]、谱聚类[4-5];基于图分割的方法,例如最小切算法[6]、最小最大切算法[7]等;还有基于中心度[8]的算法等其他算法。

这些算法都是只使用了图的结构信息,为了优化目标函数(例如减少不同社区用户之间关系的权重和)而进行的图划分。单一的目标函数往往导致图划分结果严重偏斜。另外,只基于图结构信息得到的社区发现结果,需要进行人工的分析、定义才能确定其每个子图的意义。

2.2 话题模型

话题模型是在文档集合中自动发现话题的一种统计模型。话题模型有LSA(latent semantic analysis)模型[9]、PLSA(probabilistic latent semantic analysis)模型[10]和LDA模型[11]。其中LDA模型是目前使用最广泛的话题模型,是PLSA模型的泛化。话题模型通过得到每个话题生成词的概率,从而得到每个话题的抽象描述。

本文采用的是社区发现与话题生成统一建模的概率模型方法,生成的每个社区都有相应的话题分布,每个社区都由其相应领域的作者组成,因此具有自然直观的划分语义。本文模型基于LDA[11],由于LDA平滑的性质,划分的社区规模相对平均。

2.3 混合模型

话题模型与社区发现相结合的工作包括:

(1)完全基于话题模型的社区发现方法

①SSN-LDA(simple social network LDA)模型[12],它主要利用了作者之间的共现关系,类似于Blei的LDA模型利用词的共现关系,但是社区发现的结果缺乏可解释的语义。

②AT(author-topic)模型[13],在LDA模型基础上添加作者参数,由文档作者生成话题,可以进行话题发现,并分析每个作者的话题关注,但是没有引入社区的概念,只能间接地基于话题划分社区。

③CUT(community-user-topic)模型[14],CUT模型分为CUT1模型和CUT2模型。CUT1可进行话题发现、社区发现和作者话题关注分析,CUT2可进行语义社区发现与用户用词习惯分析。CUT1、CUT2是两个独立的模型,它并没有把话题发现、社区发现以及社区语义解释集成到统一的框架中。

(2)话题模型与网络结构相结合的社区发现方法

该方法包括NetPLSA(PLSA with network regularization)模型[15]和Topic-link模型[16],它们基于假设引入正则化约束,以修正话题模型的结果。这些模型的推导过程复杂,实现复杂度过高。

针对以上问题,将话题发现、社区发现、社区话题分析、用户影响力分析集成于统一的概率模型之中,可以快速得到更为精确、细致的分析结果,并且通过话题分布对生成的社区进行描述,使社区发现的结果具有直观的语义。

2.4 用户影响力分析

给定社区及社区内用户关系图,用户影响力分析方法有:

(1)基于用户度数、用户在内容中被提及次数、用户内容被转发或引用次数等指标分析用户影响力,如在Twitter(https://twitter.com/)大数据集上的研究[17]。但是这些指标基于直观经验,过于简单。

(2)基于用户的PageRank[18]值分析用户影响力。PageRank算法的应用广泛,可以利用用户间关系对用户进行PageRank排序,分析用户影响力。但是PageRank算法不能用于社区发现,社区发现过程与社区影响力分析过程相互独立。

本文模型基于话题、社区和用户影响力之间的相互关联性,使用户影响力分析在社区层次得到更为细致、准确的结果。

3 ACT-LDA模型

ACT-LDA模型基于LDA模型,集成了话题发现、社区发现、用户影响力分析功能,能够同时进行

话题发现分析、社区发现分析、社区话题分布分析、社区内用户影响力分析,从而能够发现话题,划分社区,得到每个社区的话题描述,以及计算每个作者在社区中的影响力权值。

社区中文档的话题与社区中的话题分布紧密相关,话题形成与社区形成相互影响,文档中作者间关系的形成与社区的形成相互影响,在社区中具有较大影响力的用户对于文档的生成同样具有较大影响,这些内在的联系在ACT-LDA的概率生成过程中得到了充分体现。基于话题、社区和用户影响力之间的相互关联性,ACT-LDA能够对话题发现、社区发现、用户影响力统一建模,可以同时进行用户层次、话题层次和社区层次上的分析,得到更为准确和细致的分析结果。实验部分的结果也很好地验证了模型的有效性。

3.1 术语解释以及符号说明

ACT-LDA模型适用于任何包含用户关系和用户发表内容的网络。

定义1(文档)用户发表的内容称为文档,用户称为文档作者。在ACT-LDA模型中,文档包含3个要素:文档作者,文档内容,文档中作者之间形成的关系。所有文档d组成的集合称为文档集合D。

定义2(用户关系图)用户关系图$G_S=(S,E)$。其中,S为文档集D中所有作者的集合,E为文档集中作者之间形成关系的集合。

以论文网络为例,D为论文的集合,G_S中点集S为论文作者集合,边集E为论文中作者之间所有合作关系的集合。以微博网络为例,D为微博的集合,G_S由微博用户以及用户之间的转发关系、关注关系组成。

ACT-LDA模型中使用的符号说明见表1。

3.2 生成过程

ACT-LDA模型的概率生成图如图1所示。

假设1 每个文档都属于一个社区。

假设2 每个社区之间的话题分布相互独立。

假设3 每个社区生成作者的过程相互独立。

假设4 作者在文档中的共现关系说明其位于一个社区的概率值增加。

基于以上假设,ACT-LDA模型对每篇文档独立

Table 1 Notations in ACT-LDA model

表1 ACT-LDA模型的符号说明

符号	说明
D	所有文档集合
S	所有作者集合
Q	社区的数目
K	话题的数目
V	词表中词的数目
N_d	文档d中包含词的数目
H_d	文档d中包含作者的数目
γ_d	文档d所属的社区
θ_{dn}	第d个文档中第n个词所属的话题
w_{dn}	第d个文档中第n个词
a_{dh}	第d个文档的第h个作者
C_q^1	社区q生成话题服从的多项式分布参数
C_q^2	社区q生成作者服从的多项式分布参数
ϕ_k	话题k生成词服从的多项式分布参数
α	γ_d服从的多项式分布先验参数
β	ϕ_k服从的狄利克雷分布先验参数
κ^1	C_q^1服从的狄利克雷分布先验参数
κ^2	C_q^2服从的狄利克雷分布先验参数

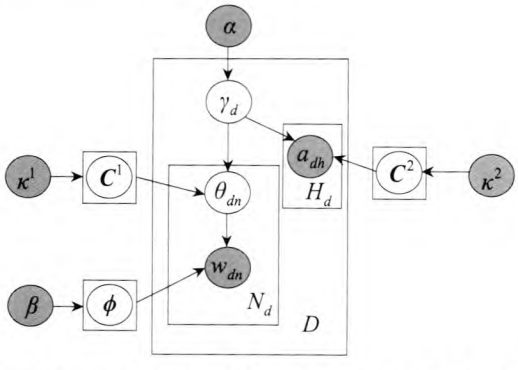

Fig.1 The graphical model representation of ACT-LDA

图1 ACT-LDA模型概率生成图

生成此文档所属社区,然后根据作者在此社区中的生成概率(即社区内影响力)生成文档的作者集合;对文档中每个词,先根据社区中话题分布选择话题,再根据话题生成词的概率最终生成词。

生成过程利用了社区话题与文档话题间、话题与社区间、用户影响力与文档生成过程间、社区形成与文档作者集合形成间的内在联系。生成文档集合D和用户关系图G_S的过程形式化说明如下。

对于每篇文档$d\in D$:

(1)选取社区 γ_d~$Multinomial(\alpha)$;

(2)选取 N_d~$Poisson(\xi_1)$;

(3)选取 $C^1_{\gamma_d}$~$Dir(\kappa^1)$;

(4)对于文档 d 中的 N_d 个词 $w_{dn}(1\leqslant n\leqslant N_d)$:

①选取话题 θ_{dn}~$Multinomial(C^1_{\gamma_d})$;

②选取词 w_{dn}~$Multinomial(\phi_{\theta_{dn}})$;

(5)选取 H_d~$Poisson(\xi_2)$;

(6)选取 $C^2_{\gamma_d}$~$Dir(\kappa^2)$;

(7)对于文档 d 中 H_d 位作者 $a_{dh}(1\leqslant h\leqslant H_d)$,选取 a_{dh}~$Multinomial(C^2_{\gamma_d})$。

其中,$\boldsymbol{C}^1$ 是 $Q\times K$ 的矩阵,C^1_q 为对社区 q 的条件概率分布;$\boldsymbol{C}^2$ 是 $Q\times S$ 的矩阵,C^2_q 为对社区 q 的条件概率分布;$\boldsymbol{\phi}$ 是 $K\times V$ 的矩阵,ϕ_k 为对话题 k 的条件概率分布。$\boldsymbol{C}^1$、$\boldsymbol{C}^2$ 和 $\boldsymbol{\phi}$ 均服从狄利克雷分布。

3.3 模型推理

本文采用变分推理[19]的方法进行推理。估计ACT-LDA后验概率分布的变分图如图2所示。

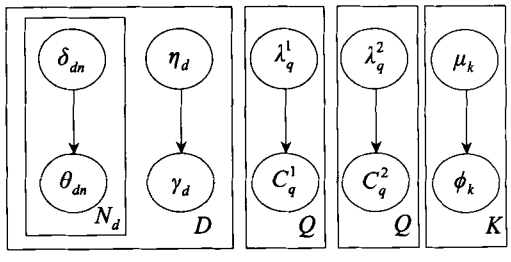

Fig.2 The graphical representation of the variational distribution to approximate the posterior probability in ACT-LDA

图2 估计ACT-LDA后验概率的变分图示

条件分布满足式(1)~式(5)。

$$\theta_{dn}|\delta_{dn}\sim Multinomial(\delta_{dn}) \tag{1}$$

$$\gamma_d|\eta_d\sim Multinomial(\eta_d) \tag{2}$$

$$C^1_q|\lambda^1_q\sim Dir(\lambda^1_q) \tag{3}$$

$$C^2_q|\lambda^2_q\sim Dir(\lambda^2_q) \tag{4}$$

$$\phi_k|\mu_k\sim Dir(\mu_k) \tag{5}$$

通过使用变分推理对对数似然函数差求导,得到式(6)~式(10)。

$$\eta^q_d\propto\alpha_q\exp\{\sum_{n=1}^{N_d}\sum_{k=1}^{K}\delta^k_{dn}(\Psi(\lambda^1_{qk})-\Psi(\sum_{k=1}^{K}\lambda^1_{qk}))+\sum_{h=1}^{H_d}\sum_{s=1}^{S}a^s_{dh}(\Psi(\lambda^2_{qs})-\Psi(\sum_{s=1}^{S}\lambda^2_{qs}))\} \tag{6}$$

$$\lambda^1_{qk}=\kappa^1_k+\sum_{d=1}^{D}\sum_{n=1}^{N_d}\delta^k_{dn}\eta^q_d \tag{7}$$

$$\lambda^2_{qs}=\kappa^2_s+\sum_{d=1}^{D}\sum_{h=1}^{H_d}a^s_{dh}\eta^q_d \tag{8}$$

$$\delta^k_{dn}\propto\exp\{\sum_{q=1}^{Q}\eta^q_d\Psi(\lambda^1_{qk})+\sum_{v=1}^{V}w^v_{dn}(\Psi(\mu_{kv})-\Psi(\sum_{v=1}^{V}\mu_{kv}))\} \tag{9}$$

$$\mu_{kv}=\beta_v+\sum_{d=1}^{D}\sum_{n=1}^{N_d}w^v_{dn}\delta^k_{dn} \tag{10}$$

模型使用式(6)~式(10)进行迭代计算,直到收敛。

3.4 模型变种ACT-LDA$^+$

本文还提出了ACT-LDA$^+$模型。ACT-LDA$^+$模型在ACT-LDA模型基础上,将每篇论文的会议信息作为论文已知的社区信息,在变分推理的迭代中提升其在相应会议中的概率,用于有指导的社区划分。即相对于ACT-LDA模型,ACT-LDA$^+$概率生成模型中每篇论文所属社区由隐变量变为可观察变量。

在已知论文的社区信息时,ACT-LDA$^+$能够提高社区划分的结果。由于论文所属社区信息作为已知输入,ACT-LDA$^+$增强了实验结果的可解释性以及与真实数据的联系。这在实验部分将得到验证。

4 实验结果与分析

本文以论文网络为例进行实验分析。使用的数据集为DBLP(The Digital Bibliography and Library Project)数据。DBLP是一个计算机科学出版物的数据库网络资源,本文采用了其中4个会议(SIGIR、NIPS、WWW、KDD)的论文数据,提取出有标题、作者信息的论文,使用标题作为该论文的文本信息,去除了停用词,最后总计得到9 010篇论文、13 398位作者、单词数为6 667的词表。根据合作关系建立作者间关系图,总计有31 816条边。

通过实验,选取了经验值,固定社区数目 Q=4,话

题数目 $K=4$,$\alpha=0.25$,$\beta=0.95$,$\kappa^1=0.95$,$\kappa^2=0.95$。

设计了以下4个实验从不同角度验证ACT-LDA模型的有效性:话题发现、社区发现、社区话题分布、社区用户影响力分析。

4.1 话题发现

对比了LDA模型(http://www.cs.princeton.edu/~blei/lda-c/index.html)与ACT-LDA模型的话题发现结果,计算每个话题下生成概率最大的8个词作为该话题的描述,实验结果如表2和表3所示。

Table 2 Topics extracted with LDA

表2 LDA话题发现结果

话题1	话题2	话题3	话题4
data	learning	retrieval	web
mining	neural	learning	search
web	networks	clustering	text
retrieval	analysis	models	based
modeling	network	model	semantic
relevance	detection	document	query
support	dynamic	search	retrieval
models	data	image	classification

Table 3 Topic extracted with ACT-LDA

表3 ACT-LDA话题发现结果

话题1	话题2	话题3	话题4
data	learning	retrieval	web
mining	networks	model	search
modeling	neural	models	classification
visual	analysis	clustering	based
selection	network	approach	text
relevance	detection	document	query
support	efficient	algorithm	system
distributed	social	image	semantic

从表中可以看到:话题1对应于数据挖掘(data mining),即对应于KDD会议的主要话题;话题2对应于机器学习(machine learning)、神经网络(neural networks),即对应于NIPS会议的主要话题;话题3对应于信息检索(information retrieval)、文本聚类(document clustering)等,即对应于SIGIR会议的主要话题;话题4对应于WWW会议的主要话题,它包括了一些WWW会议中关注的话题,如语义网(semantic web)、系统(system)等。话题发现结果与论文数据中4个会议的主要话题一一对应。

LDA生成的话题中(见表2),"learning"出现在话题2和话题3中,"web"出现在话题1和话题4中,"retrieval"出现在话题1、话题3和话题4中,话题间区别度较小,话题间平均Jaccard相似度为0.094。ACT-LDA生成的话题间没有出现共同词描述,话题平均Jaccard相似度为0。ACT-LDA生成的话题和4个会议主要话题的对应程度以及话题之间的区别度更大,因此生成的话题就结果而言更优。这说明话题模型与社区发现的统一建模有助于改进话题生成的结果,从而验证了ACT-LDA模型在话题发现方面的有效性。

4.2 社区发现

本文在社区间边权重和、Ncut[4]值、Rcut[5]值和最后划分的社区大小4个方面与SSN-LDA模型和谱聚类方法划分的结果进行了对比,还使用了ACT-LDA$^+$模型进行有社区指导信息的自对比实验。各个模型生成社区并不一一对应,实验结果如表4所示。

需要说明的是,谱聚类实际划分的社区大小是346、11 316、290、52,因为存在1 394个孤立点,将其平均划分到4个社区,所以结果为694、11 664、639、401。若不考虑孤立点,谱聚类的Rcut会更大,得到

Table 4 The comparison of community discovery with different models

表4 使用不同模型的社区发现结果对比

模型	社区间边权重和	Ncut	Rcut	社区1大小	社区2大小	社区3大小	社区4大小
SSN-LDA	8 717	1.512	2.623	3 900	3 106	3 106	3 286
ACT-LDA	4 845	0.723	1.453	4 031	3 354	2 982	3 031
ACT-LDA$^+$	3 952	0.615	1.335	3 335	2 103	5 171	2 789
谱聚类	512	0.261	0.207	694	11 664	639	401

的社区划分偏斜会更严重(按照Rcut计算公式,孤立点不影响Rcut结果,这也是这里谱聚类的Rcut值小于Ncut值的原因之一)。但是,谱聚类方法所得社区间边权重和远远小于基于LDA的模型所得社区间边权重和。这是由于谱聚类以减少Ncut和Rcut值为目标,基于LDA的模型以发现话题、社区为目标。因为LDA平滑的性质,所以基于LDA的模型得到的社区大小会更加平均。

SSN-LDA、ACT-LDA、ACT-LDA$^+$迭代都很快达到收敛。ACT-LDA、ACT-LDA$^+$模型以LDA模型生成的话题作为话题发现部分的初始值,使得这两个模型的迭代次数小于10。根据迭代算法,算法复杂度为 $O((KVM_1+QSM_2)\times I)$。其中 K 为话题数目,V 为词表大小,M_1 为文档集合大小(即所有文档中词数量之和),Q 为社区数目,S 为作者数目,M_2 为所有文档中作者数量之和,I 为迭代次数。谱聚类算法复杂度高达 $O(S^3)$。考虑到每篇文档的作者数目往往不超过5,Q、K、I 值也较小,V、S 值接近,ACT-LDA模型的复杂度要远远小于谱聚类算法,这也符合本文在实验中对运行时间观察的结果。

对比SSN-LDA模型与ACT-LDA模型,ACT-LDA模型得到的社区间边权重和接近于SSN-LDA模型的1/2,Ncut和Rcut值也为SSN-LDA模型相应值的1/2。可见相对于单独使用LDA模型进行社区发现,基于LDA统一建模话题发现与社区发现的方法可以取得更好的社区发现结果。

从ACT-LDA与ACT-LDA$^+$模型的对比中可以看到,将文档所属社区作为已知信息输入时,即对于每篇文档 $d\in D$,在模型的 γ_d 部分,迭代计算时提升 γ_d 等于 d 所在会议的概率,可以得到更好的社区发现结果,这是符合经验的。并且在已知这一额外信息时可以采用ACT-LDA$^+$模型改进模型最终结果,但是改进的幅度不是很大,不同会议的论文数目差异也导致最终社区划分结果的相对偏斜。这说明ACT-LDA模型在不利用已有文档所在社区信息时,也能接近达到利用文档所属社区信息时的结果,验证了ACT-LDA模型在社区发现方面的有效性。

4.3 社区话题分布

ACT-LDA最终生成的话题可以很好地对应于各个会议的主要话题,但是其社区的概念不是直接对应于会议,即社区划分不是通过会议划分。因此,没有直接的基准用来衡量模型所得社区话题分布的结果。

ACT-LDA$^+$中将会议作为额外的指导信息帮助划分社区,ACT-LDA$^+$与ACT-LDA模型在其他方面没有差异。并且ACT-LDA与ACT-LDA$^+$模型之间最终生成的话题差异极小,因此通过ACT-LDA$^+$模型,查看会议与生成的会议话题分布是否正确对应。将KDD会议编号为社区1,NIPS会议编号为社区2,SIGIR会议编号为社区3,WWW会议编号为社区4,每篇论文的会议信息作为已知信息输入到ACT-LDA$^+$模型。

利用ACT-LDA$^+$模型获得了每个社区的话题分布(话题与话题发现实验的结果一一对应),实验结果如表5所示。

Table 5 The topic distribution of 4 communities with ACT-LDA$^+$

表5 ACT-LDA$^+$社区话题分布结果

社区	话题1	话题2	话题3	话题4
社区1(KDD)	**0.294**	0.272	0.226	0.208
社区2(NIPS)	0.220	**0.385**	0.237	0.158
社区3(SIGIR)	0.236	0.136	**0.323**	0.305
社区4(WWW)	0.210	0.197	0.201	**0.391**

将表5中每行以及每列概率值最大的位置用粗体显示,将在4.1节实验中根据经验得到的正确的会议-主要话题对应关系用下划线显示。

实验结果显示,粗体部分与下划线部分完全重合。若对每个话题,取生成其概率最大的社区作为该话题所流行的社区,那么对应结果完全正确。如果对于每个社区,取它生成概率最大的话题作为其主要话题,也完全正确。实验说明,ACT-LDA$^+$模型进行社区发现的结果具有直观而且正确的语义,也验证了ACT-LDA模型在社区发现方面的有效性。

4.4 社区用户影响力分析

首先使用著名的PageRank[18]算法根据合作链接关系计算所有作者的排名,PageRank参数α取常用值

Table 6 The Top-K authors and the corresponding PageRank values by PageRank

表6 PageRank模型作者排名及相应PageRank值

作者	PageRank值	作者	PageRank值
W. Bruce Croft	3.787E-4	Geoffrey E. Hinton	2.659E-4
C. Lee Giles	3.134E-4	Jiawei Han	2.628E-4
Michael I. Jordan	2.896E-4	Zheng Chen	2.623E-4
Bernhard Sch	2.796E-4	Wei-Ying Ma	2.539E-4
Terrence J. Sejnowski	2.714E-4	Christos Faloutsos	2.463E-4
Christof Koch	2.713E-4	Yoshua Bengio	2.290E-4

Table 7 The Top-K authors in every community ranked by ACT-LDA$^+$

表7 ACT-LDA$^+$模型社区作者影响力排名

社区1(KDD)	社区2(NIPS)	社区3(SIGIR)	社区4(WWW)
Jiawei Han	**Michael I. Jordan**	**W. Bruce Croft**	**C. Lee Giles**
Christos Faloutsos	**Terrence J. Sejnowski**	James Allan	**Zheng Chen**
Philip S. Yu	**Geoffrey E. Hinton**	**Zheng Chen**	Yong Yu
Jian Pei	**Bernhard Sch**	ChengXiang Zhai	**Wei-Ying Ma**
Jieping Ye	Peter Dayan	Mark Sanderson	Lei Zhang
Andrew W. Moore	**Yoshua Bengio**	Leif Azzopardi	Ravi Kumar
Mohammed Javeed Zaki	**Christof Koch**	Maarten de Rijke	Erik Wilde
Ravi Kumar	Klaus-Robert M	Ryen W. White	Katsumi Tanaka
Vipin Kumar	Zoubin Ghahramani	Susan T. Dumais	Andrew Tomkins
Srinivasan Parthasarathy	Tommi Jaakkola	Charles L. A. Clarke	Gui-Rong Xue

α=0.85。限于篇幅,选取了前12位排名最前的作者,结果如表6所示。

为了对应于会议信息,使用ACT-LDA$^+$模型,对于每个社区(会议),用户影响力权值为社区生成该作者的概率。取用户影响力权值最大的前10位作者,结果如表7所示。

把PageRank算法得到的排名前12位作者在表7中用粗体显示。从表7中可以看到,ACT-LDA$^+$模型挖掘社区最大影响力作者的效果非常好。

在KDD会议中,Jiawei Han(http://en.wikipedia.org/wiki/Jiawei_Han)教授是国际数据挖掘方面非常著名的专家,他的学生Jian Pei也是活跃的数据挖掘研究者;Christos Faloutsos(http://en.wikipedia.org/wiki/Christos_Faloutsos)是SIGKDD的执行委员会成员,1989年获美国国家自然科学基金会美国总统奖,7篇最佳论文奖,发表140多篇被引用论文。

在NIPS会议中,Michael I. Jordan(http://en.wikipedia.org/wiki/Michael_I._Jordan)是机器学习、人工智能方面的领导级研究者;Terrence J. Sejnowski(http://en.wikipedia.org/wiki/Terrence_J._Sejnowski)1984年获得美国国家自然科学基金会美国总统奖,1999年因为在学习算法中的卓越贡献获得赫步奖,他在神经网络、计算机神经科学方面的研究都是先驱性的。

KDD以及NIPS其他的生成概率最大的作者,以及SIGIR、WWW会议方面的结果,限于篇幅在此不进行详细的解释,可以自行验证。

从结果可以看到,每个社区影响力最大的前K个作者都是在其相应会议极具影响力的研究者,并且作者的研究领域和挖掘出社区的主要话题非常相关。本节进行的实验不仅说明了ACT-LDA$^+$模型在

挖掘社区最具影响力作者方面的准确度，也从另一方面验证了ACT-LDA+模型社区发现结果的正确性。

同时，PageRank算法计算得到的前12位排名最前的作者，皆处于ACT-LDA+模型得到的各个社区的前10位作者之中，排名皆位于前几位，这也验证了ACT-LDA+模型的有效性。由于数据中NIPS会议论文偏多(3 617/9 010篇)，PageRank排名前12位作者中参加NIPS会议的作者居多。另外PageRank算法只能基于链接关系计算整体排名，并没有进行社区划分和利用文本信息，因此不能分析得到作者的社区信息以及在各个社区中的影响力，更不能得到其在各个话题领域的影响力。而ACT-LDA+模型可以综合分析社区划分、社区中话题分布以及社区内不同用户的影响力，因此可以得到更为细致准确的分析结果。实验说明，ACT-LDA+模型在用户影响力分析方面非常有效，也验证了ACT-LDA模型在用户影响力分析方面的有效性。

5 总结与未来工作

随着社交网络的不断发展，对社交网络进行分析的工作越来越重要。本文提出了一个集成话题发现、社区发现、用户影响力分析的基于LDA的统一贝叶斯模型ACT-LDA，用于分析、挖掘包含文本内容的用户关系网络。ACT-LDA模型的优势在于利用统一的模型集成话题发现与社区发现、用户影响力分析这三个相互关联的模块，从而取得在话题层次、社区层次更为细致和准确的分析结果。同时使用了变分推理解决模型推理问题。在DBLP数据上的实验结果说明，ACT-LDA模型有助于改进话题发现、社区发现的结果，社区内用户影响力的分析结果也非常准确和有效，并且生成的社区具有直观、正确的语义，从而验证了模型假设的正确性以及模型的有效性。ACT-LDA模型适用于任何包含用户关系和用户发表内容的网络。

在将来的工作中计划引入时间参数，用于分析话题的演化以及社区的演化，并且在话题层次分析用户之间的信息传播过程。

References:

[1] Macqueen J B. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. [S.l.]: University of California Press, 1967: 281-297.

[2] Sibson R. SLINK: an optimally efficient algorithm for the single-link cluster method[J]. The Computer Journal, 1973, 16(1): 30-34.

[3] Szekely G J, Rizzo M L. Hierarchical clustering via joint between-within distances: extending Ward's minimum variance method[J]. Journal of Classification, 2005, 22(2): 151-183.

[4] Hagen L, Kahng A B. New spectral methods for ratio cut partitioning and clustering[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 1992, 11(9): 1074-1085.

[5] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.

[6] Chekuri C S, Goldberg A V, Karger D R, et al. Experimental study of minimum cut algorithms[C]//Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '97). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1997: 324-333.

[7] Ding C H Q, He Xiaofeng, Zha Hongyuan, et al. A min-max cut algorithm for graph partitioning and data clustering[C]// Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM '01). Washington, DC, USA: IEEE Computer Society, 2001: 107-114.

[8] Balakrishnan H, Deo N. Centrality based community discovery[J]. Congressus Numerantium, 2007, 188: 117-128.

[9] Papadimitriou C H, Tamaki H, Raghavan P, et al. Latent semantic indexing: a probabilistic analysis[C]//Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS '98). New York, NY, USA: ACM, 1998: 159-168.

[10] Hofmann T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99). New York, NY, USA: ACM, 1999: 50-57.

[11] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[12] Zhang Haizheng, Qiu Baojun, Giles C L, et al. An LDA-based community structure discovery approach for large-scale social networks[C]//Proceedings of the 2007 IEEE International Conference on Intelligence and Security Informatics (ISI '07), 2007: 200-207.

[13] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI '04). Arlington, Virginia, USA: AUAI Press, 2004: 487-494.

[14] Zhou Ding, Manavoglu E, Li Jia, et al. Probabilistic models for discovering e-communities[C]//Proceedings of the 15th International Conference on World Wide Web (WWW '06). New York, NY, USA: ACM, 2006: 173-182.

[15] Mei Qiaozhu, Cai Deng, Zhang Duo, et al. Topic modeling with network regularization[C]//Proceedings of the 17th International Conference on World Wide Web (WWW '08). New York, NY, USA: ACM, 2008: 101-110.

[16] Liu Yan, Niculescu-Mizil A, Gryc W. Topic-link LDA: joint models of topic and author community[C]//Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09). New York, NY, USA: ACM, 2009: 665-672.

[17] Cha M, Haddadi H, Benevenuto F, et al. Measuring user influence in Twitter: the million follower fallacy[C]//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM '10), 2010: 10-17.

[18] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the Web[R]. Stanford InfoLab, 1999.

[19] Wainwright M J, Jordan M I. Graphical models, exponential families, and variational inference[J]. Foundations and Trends® in Machine Learning, 2008, 1(1/2): 1-305.

WU Liang was born in 1990. He is a master candidate at Peking University. His research interests include machine learning, data mining and social networks, etc.

吴良(1990—),男,湖南永州人,北京大学硕士研究生,主要研究领域为机器学习,数据挖掘,社交网络等。

HUANG Weijing was born in 1986. He is a Ph.D. candidate at Peking University. His research interests include topic detection and tracking, machine learning and social networks, etc.

黄威靖(1986—),男,浙江温州人,北京大学博士研究生,主要研究领域为话题追踪检测,机器学习,社交网络等。

CHEN Wei was born in 1981. She received her Ph.D. degree from School of Electronics Engineering and Computer Science, Peking University in 2009. Now she is a research assistant at School of Electronics Engineering and Computer Science, Peking University. Her research interests include artificial intelligence, social networks analysis and Internet public opinion, etc.

陈薇(1981—),女,陕西人,2009 年于北京大学信息科学技术学院获得博士学位,现为北京大学信息科学技术学院助理研究员,主要研究领域为人工智能,社交网络分析,互联网舆情研究等。

WANG Tengjiao was born in 1973. He is a professor and Ph.D. supervisor at School of Electronics Engineering and Computer Science, Peking University. His research interests include data warehousing, data mining and Web information processing.

王腾蛟(1973—),男,北京大学信息科学技术学院教授、博士生导师,主要研究领域为数据仓库,数据挖掘,Web 信息处理。

LEI Kai was born in 1976. He received his M.S. degree from Department of Computer, Columbia University in 1999. Now he is the executive vice-director of Center for Internet Research and Engineering, Shenzhen Graduate School, Peking University, and the member of IEEE and ACM. His research interests include database, Internet information processing and mining, etc.

雷凯(1976—),男,1999年于美国哥伦比亚大学计算机系获得硕士学位,现为北京大学深圳研究生院互联网信息工程研发中心常务副主任,IEEE和ACM会员,主要研究领域为数据库,互联网信息处理和挖掘等。

LIU Yueqin was born in 1950. She is a professor at Department of Information Science and Technology, University of International Relations. Her research interests include information security and intelligent information processing.

刘月琴(1950—),女,江苏人,国际关系学院信息科技系教授,主要研究领域为信息安全,智能信息处理。