软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(9):2468−2480 [doi: 10.13328/j.cnki.jos.005267]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于文本与社交信息的用户群组识别*

王中卿, 李寿山, 周国栋

(苏州大学 计算机科学与技术学院 自然语言处理实验室,江苏 苏州 215006)

通讯作者: 周国栋, Email: gdzhou@suda.edu.cn

摘 要: 社交媒体上的个人群体信息对于理解社交网络结构非常有用,现有研究主要基于用户之间的链接和显式社交信息识别用户的个人群体,很少考虑使用文本信息与隐含社交信息.在显式社交信息缺乏时,隐含社交信息以及文本信息对于识别用户的群体是非常有帮助的.提出一种隐含因子图模型,有效地利用各种隐含与显式的社交与文本信息对用户的群组进行识别.其中,显式的文本与社交信息是通过用户发表的文本与个人关系生成的.同时,利用矩阵分解模型自动生成隐含的文本与社交信息.最后,利用因子图模型与置信传播算法对显式与隐含的文本与社交信息进行集成,并对用户群组识别模型进行学习与预测.实验结果表明,该方法能够有效地对用户群组进行识别.

关键词: 群组推荐;社交网络;隐含信息;矩阵分解;因子图模型

中图法分类号: TP311

中文引用格式: 王中卿,李寿山,周国栋.基于文本与社交信息的用户群组识别.软件学报,2017,28(9):2468−2480. http://www.jos.org.cn/1000-9825/5267.htm

英文引用格式: Wang ZQ, Li SS, Zhou GD. Personal group recommendation via textual and social information. Ruan Jian Xue Bao/Journal of Software, 2017,28(9):2468−2480 (in Chinese). http://www.jos.org.cn/1000-9825/5267.htm

Personal Group Recommendation via Textual and Social Information

WANG Zhong-Qing, LI Shou-Shan, ZHOU Guo-Dong

(Natural Language Processing Laboratory, School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Personal group information on social media is useful for understanding social structures. Existing studies mainly focus on detecting personal groups using explicit social information between users, but few pay attention on using implicit social information and textual information. In this paper, a latent factor graph model (LFGM) is proposed to recommend personal groups for each person with both explicit and implicit information from textual content and social context. Especially, while explicit textual and social contents can be easily extracted from user generated content and personal friendship information, a matrix factorization approach is applied to generate both implicit textual and social information. Evaluation on a large-scale dataset validates the effectiveness of the proposed approach.

Key words: group recommendation; social network; implicit information; matrix factorization; factor graph model

社交网络服务(social networking services,简称 SNS),是指以一定社会关系或共同兴趣为纽带,以各种形式为在线用户提供沟通、交互服务的互联网应用.这种以人与人关系为核心的方式建立的社会关系网络在互联网上就形成了以用户为中心、以人为本的互联网应用.随着社交网络的发展,人们按照兴趣组成了一个个群体.通常来说,个人群体(personal group)是指一种基于兴趣的自组织(self-organized)社区.群体信息在发现和认识社交结构上存在广泛应用,比如可以发现具有相同观点或者相同兴趣的人群,也能发现人与人之间的关联.因此,个人群体推荐(personal group recommendation)作为一个新的任务,已经得到了越来越多的重视[1].以往关于群体分

* 基金项目: 国家自然科学基金(61331011, 61375073, 61402314)
Foundation item: National Natural Science Foundation of China (61331011, 61375073, 61402314)
收稿时间: 2016-01-29; 修改时间: 2016-05-25; 采用时间: 2017-02-17; jos 在线出版时间: 2017-03-17
CNKI 网络优先出版: 2017-03-17 14:37:28, http://kns.cnki.net/kcms/detail/11.2560.TP.20170317.1437.004.html

析的研究主要针对实体群组分析,比如根据标签(tag)将图片聚合为不同的群组[2],或者利用用户之间的链接信息分析用户之间的群组关系[3,4].但是,上述方法的问题在于很多用户的社交网络只存在有限的好友和内容.因此在很多情况下,上述基于用户之间联系与用户行为的方法可能会失效.而用户发表的文本在很多情况下可以帮助分析用户间的社交关系,比如从文本中抽取的用户兴趣点和写作风格等都能够有效地将用户联系起来,从而帮助识别用户所属的群组,因此在本文中,我们主要研究同时利用文本与社交信息识别用户所属的群组.需要注意的是:识别的群组主要是用户的兴趣群组,比如"阅读"或"跳舞"群组等.

在很多时候,除了显式的文本信息与社交信息之外,潜在的文本与社交信息也能有效地帮助识别用户的群组.比如:如果两个人有相近的好友群,或者存在相近的兴趣,那么他们非常有可能存在相近的群组.基于这一点,我们需要去发现用户之间的隐含社交信息(implicit social information)用来推荐群组.另外,隐含的文本信息(implicit text information),比如隐含的主题信息,也是非常有用的,并可以用来进行群组推荐.利用隐含文本信息的原因在于,这样的信息更能体现出同一群体之间的联系.比如:都喜欢发表与书籍评论相关文本的人,更可能都属于"阅读"群组.

因此,我们提出一种隐含因子图模型(latent factor graph model,简称 LFGM),有效地集成显式与隐含的文本与社交信息,从而将用户与兴趣集成在一起,并对于用户所属的群组进行识别.首先,从用户发表的文本与社交内容中抽取显式的文本与社交信息;其次,我们使用矩阵分解(matrix factorization,简称 MF)模型来发现隐含的文本和社交信息;再次,我们将显式的文本信息构建为属性函数,而将隐含的文本信息与社交信息构建为因子函数;最后,使用概率图模型融合上述信息,并利用置信传播算法对于模型进行学习与预测.通过实验验证了 LFGM 方法能够有效地对于用户群组信息进行识别.本文的主要贡献为:

1) 同时结合文本与社交信息对于用户群组进行识别;
2) 由于显式与隐含的信息能够反映不同方面的信息,分别抽取了显式和隐含信息,其中,隐含信息是通过矩阵分解模型获得的;
3) 提出一种隐含因子图模型 LFGM.

本文第 1 节对用户群组识别以及其他社交网络分析的相关工作进行总结.第 2 节主要介绍收集的数据,并针对相关数据进行统计与分析.第 3 节主要介绍基于显式和隐含的文本信息和社交内容,提出隐含因子图模型.第 4 节进行相关的实验,并对 LFGM 模型性能进行验证.第 5 节给出全文总结和对未来工作的展望.

1 相关工作

1.1 用户属性标签抽取

在社交媒体上,用户的属性可以从很多维度进行度量.比较常见的分析用户属性的方法是抽取用户的标签.在某种程度上,群组分析也可以认为是一种用户属性标签抽取任务,即,抽取用户所属的群组标签.基于社交网络的标签推荐(social tag suggestion)的目的是抽取社交网络上的适当的标签,这样的标签可以更好地帮助人们在无结构的数据上组织他们的信息[5,6].Ohkura 等人[6]提出了一种自动标签抽取方法,为每篇博客打上一个特别的标签,用来代表该博客.Lappas 等人[7]提出了基于社会关系网络和担保确认(endorsement-based)的标签生成算法,同时,他们使用了很多种从推荐和评论中抽取出来的特征.Liu 等人[8]提出了一种利用概率模型提取微博上链接词与标签之间的语义联系的方法,该方法把社会关系网络结构作为正则化因子,从而帮助标签抽取的学习.Li 等人[9]提出了一种基于上下文相关(context-aware)的标签关系抽取算法,用来从相关社交网络资源中抽取标签.Zhang 等人[10]提出了一种基于 LDA 模型(latent dirichlet allocation)的融合表示对象间关系与资源内容的标签系统 TSM/Forc,从而在社交媒体上抽取标签信息.

1.2 用户关系分析

朋友关系是社交网络的基本构成.用户关系分析是群组分析的前提,只有对于用户之间的关系做了深入的分析,才能将用户联系起来,更好地对用户的群组进行抽取.已有的研究主要关注于无监督和有监督方法两方

面.大部分无监督的链接预测(link prediction)算法都是基于图中节点之间相似度关系.开创性的工作是由Liben-Nowell 和 Kleinberg[11]提出的无监督学习方法,将问题从算法节点视图转变为如何通过邻接节点特征来预测社会关系网络上的新建连接.最近,研究者主要研究将监督学习应用到链接预测中.Tang 等人[12]将概率图算法应用到学习大规模网络的社会关系中.Tang 等人[13]进一步扩展了这个工作,在异构网络中通过平衡社会关系理论来关联不同的网络结构. Hasnan 等人[14]比较了若干监督学习算法(决策树、SVM、*k* 近邻、RBF 网络、朴素贝叶斯等)的预测性能.在 BioBase 和 DBLP 网络库中进行的实验表明,SVM 方法要优于其他算法.He 等人[15]将关联规则挖掘应用到多关系抽取中.此外,通过对特征的排序,发现小集合的特征对于链接预测总能起到重要的作用.

目前,只有很少的关于从文本中抽取用户之间社交联系的工作.Elson 等人[16]提出一种从 19 世纪英国小说中抽取社会关系网络的方法,他们是基于将两个角色是否有对话将两个人联系起来的.McCallum 等人[17]探索了将结构化信息比如邮件头用来构建社会关系网络.Qiu 等人[18]探索了在论坛上利用情感信息和交互信息来发现人与人之间关系.

1.3 群体关系分析

群组关系分析是社会网络分析的一个基本任务.在社会网络中发现一个群组,就是识别一个结点的集合,使得集合内结点之间的相互作用比它们与集合外结点的相互作用更强.群组关系分析有助于其他社会计算任务的实现,并被应用于许多实际问题的求解.比如:根据相似的兴趣对社会网络中的客户进行划分,就可以给客户推荐一系列相关的产品,从而提高交易的成功率.群组也可以用来压缩巨大的网络,从而降低网络的规模.换句话说,问题的求解可以在群组级,而不是在节点级来完成.在以往的关于群组推荐的工作中,主要可以分为两类:基于内容和基于链接[1,2].

基于内容的方法主要利用用户生成内容来探索群组信息、发现实体的群组,比如发现图片的群组.通常的做法是,利用图像信息发现图像的群组.Zha 等人[19]利用多种图像特征来抽取图像的标签和群组.Yu 等人[1]利用标签传播算法来推荐图片内容的群组.Zhang 等人[20]提出了一种利用特征项的方法,从用户行为信息中进行群组推荐.

基于链接方法主要利用用户行为(user's behavior)信息来检测群组.比如:Lerman 和 Jones[2]利用图片之间流行度的联系和相应的群组编号来发现群组;Negoescu 和 Gatica-Perez[3]分析了用户标记和群组标记之间的联系,从而提出一个主题模型用来表示基于标记的用户和群体关系;Zheng 等人[4]基于张量分解算法(tensor decomposition),利用用户和标记信息来推荐用户群组.

与上述研究相比,很少有研究关注于集成内容和用户行为这两种信息.Kubica 等人[21]利用链接和实体的内容信息来检测群组,但是作为一个流水线系统,他们利用内容信息和链接信息是分开的.Wang 等人[22]提出了一种基于张量模型的方法,基于图像内容,用户信息和标记信息来推荐群组,但是他们利用的用户行为信息只是用户的标识.

总体来说,虽然已经有了一些相关的工作,但是大部分针对群组分析的工作都是基于检测实体,比如图像的群组,很少有研究是针对个人群体进行推荐的.另外,我们对于群组分析这个任务进行了细化,主要的研究对象是基于兴趣的用户群组分析;为了发现群组之间的隐含联系,提出了概率图模型与矩阵分解模型进行群组关系分析.最后,通过概率图模型方法提供了一种全局优化方式来同时利用显式的和隐含的文本和社交信息.

2 数据收集和分析

首先介绍我们收集的数据并针对相关数据进行统计与分析.

2.1 数据收集

我们用到的数据集收集自豆瓣网(www.douban.com),豆瓣网是国内流行的一个社交网站,已有很多针对这个网站的相关研究[23,24].在豆瓣网上包含了大量的用户生成的信息,比如个人的日志、对于电影和书籍的评论

和大量的社交内容信息,包括用户喜欢的电影、书籍列表和用户的好友列表.我们删去用户名,从而保证用户的隐私.

初始数据集中包含 4 379 个用户,其中,只有 1 584 个用户包含了群组信息.在我们的研究中,只选择了有群组信息的用户.在数据集中,我们一共获得了 36 147 个群组,其中,32 021 个群组出现的次数是少于 10 次.在剩下的 3 496 个群组中,我们选择了出现频率最高的 10 个群组作为候选群组进行分析.表 1 为这 10 个群组的统计分布情况.从表中我们能够发现:出现频率最高的 10 个群组都是与生活与工作相关的,这应该与豆瓣网本身的定位有关.需要注意的是,所有候选群组占比之和大于 100%.这是由于一个人可能有多个感兴趣的群组造成的.

Table 1 Distribution of candidate personal groups

表 **1** 候选个人群组分布情况

群组	用户数目	频率(%)
笑话	535	0.338
阅读	516	0.326
电影	507	0.320
网站推荐	465	0.294
美食	405	0.256
工作	341	0.215
生活	335	0.211
创新	295	0.186
星座	286	0.181
百科	279	0.176

2.2 统计分析

本节给出一些统计分析,并对用户的群组与用户之间的显式与隐含社交联系的关联程度进行分析与验证.

- 显式社交联系

如果两个人被显式的社会关系所联系,那么他们是否会存在相同的群组.图 1 显示的是相应的统计,即:基于两个人被联系的情况,他们存在 n 个相同群组的概率.其中,Y 轴表示当两个人有显式的社会联系即好友关系的情况下,两个用户存在 n 个相同群组.从图 1 中我们发现:相比随机的情况,如果两个用户是好友,那么他们倾向于有更多的相同群组.这和我们的直观认识是接近的:如果两个人是好友,那么他们很可能有相近的兴趣,自然会趋向于有很多相同的兴趣小组.

- 隐含社交联系

如果两个人之间存在隐含的社交联系,比如,两个人存在相同的好友群或者相近喜欢的产品,那么他们也倾向于属于相同类型的群组.我们从用户之间分别抽取了相近好友群与相近兴趣两种联系.分别分析:(1) 如果两个人之间存在 n 个共同的好友,那么他们有相同的群组的概率;(2) 如果两个人之间存在 n 个相同的喜欢的产品,那么他们之间存在相同群组的概率.需要注意的是:在图 1 中,我们的验证是在两个人是好友的情况下;而在图 2 中,我们验证的是两个人有相同的好友群的情况下,即使两个人有很多相同的好友,他们本身也不一定是好友.

图 2 与图 3 分别描述了如果两个人存在 n 个相近好友群或相近感兴趣产品的情况下,他们至少存在一个相同群组的概率.Y 轴是两个人存在至少一个相同群组的概率.我们有如下有趣的发现:相比显式的联系,用户存在隐含联系的情况下,更可能存在相近的群组.这可能是由于这样的隐含联系能够更紧密地联系人们.因为我们分析的是用户的兴趣群组,而相近的好友群与感兴趣的商品就更能体现出用户之前的基于兴趣的群组联系.

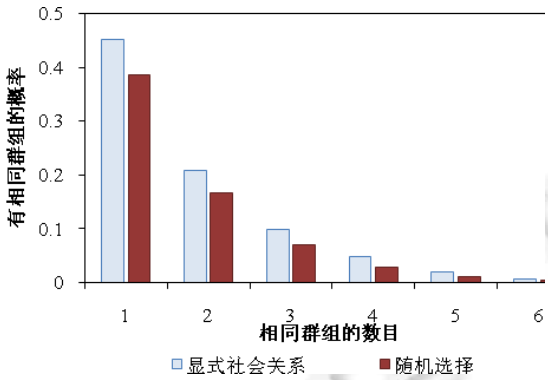


Fig.1 Probabilities that two connected users have *n* same groups according to explicit social connections (aka friendship)

图 1 当两个人存在显式社交联系(好友关系)时,他们有 *n* 个相同群组的概率

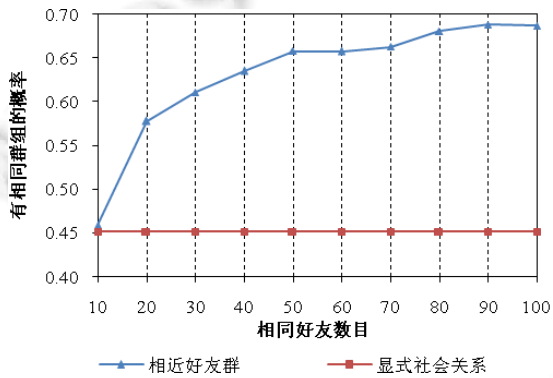


Fig.2 Probabilities that two connected users have *n* same groups according to similr friend group

图 2 在两个人有相同好友群的情况下,他们有 *n* 个相同群组的概率

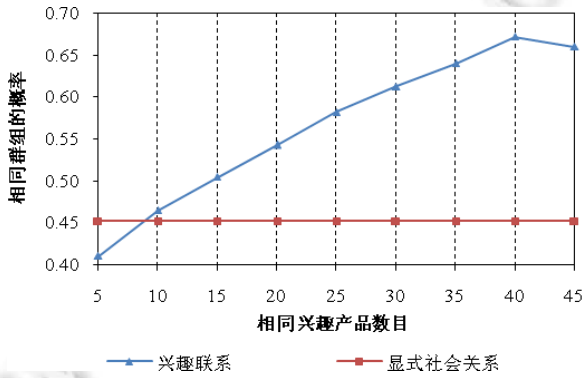


Fig.3 Probabilities that two connected users have *n* same groups according to similr favorite products

图 3 在两个人有相同兴趣的情况下,他们有 *n* 个相同群组的概率

3 隐含因子图模型

本节基于显式和隐含的文本信息和社交内容提出了隐含因子图模型(latent factor graph model,简称

LFGM),用来学习和预测群组.图 4 表述 LFGM 模型的整体框架.

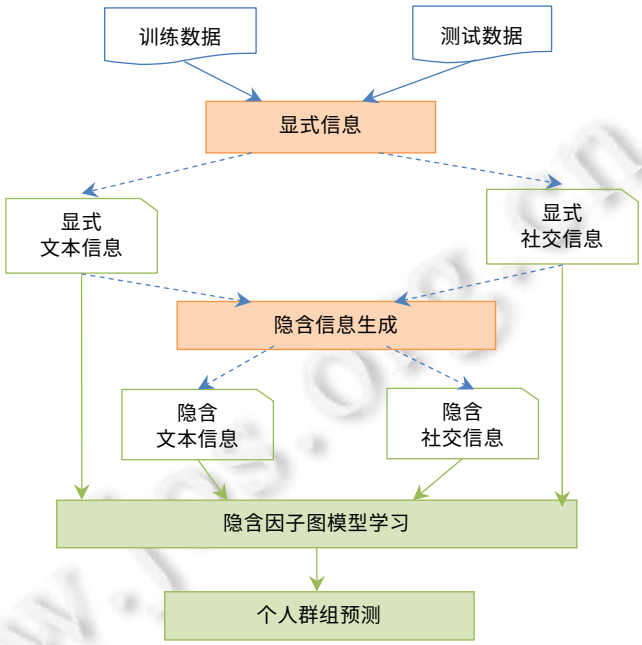


Fig.4 Overview of our proposed framework

图 4 我们提出的框架的描述

我们的整体方法是一种基于监督的迁移学习方法.

- 首先,分别从文本和社交内容中抽取显式的文本和社交信息.我们从用户发表的日志和评论中抽取显式的文本信息,并从用户之间的社交内容中抽取两个用户之间的好友关系等社交联系作为显式的社交内容;
- 其次,通过矩阵分解模型[25]的方式抽取用户之间的基于主题的隐含文本联系.同样的,隐含社交联系也是通过类似的方式从好友和喜好的产品中抽取获得的;
- 最后,集成了上述显式的和隐含的信息用来构建因子图模型,并利用环置信传播算法学习和预测整个模型.

总的来说,框架的核心为:(1) 如何利用矩阵分解方法生成隐含信息;(2) 如何同时基于显式信息和隐含信息来构建因子图模型;(3) 如何学习我们提出的因子图模型,并利用学习到的模型进行预测.

3.1 隐含信息的生成

与显式信息不同,隐含信息能够反映用户之间的潜在联系.比如一个用户喜欢读“C++”相关的书,而另外一个用户喜欢阅读“Java”相关的书,虽然这两个用户在显式的特征上没有相同,因为他们喜欢阅读的书不同,但是通过隐含信息抽取,可以抽取出他们都喜欢读与“计算机”相关的书.“计算机”即是他们的隐含兴趣特征.这两个用户也能过通过“计算机”隐含特征联系起来.在我们的研究中,矩阵分解被用来生成隐含的文本和社交信息.对于一个 $m\times n$ 的属性——个人矩阵 M,我们需要分解生成它的 3 个子矩阵[25],用来获得属性和个人的主题信息:

$$\left.\begin{aligned}&\min_{Q,S,G}\|M-FST^T\|\\&\text{s.t. } S\geq 0,F\geq 0,T\geq 0\end{aligned}\right\} \tag{1}$$

当生成隐含的文本联系时,M 是文本-用户矩阵;而当生成隐含的社交联系时,M 是好友/爱好-用户矩阵;$F\in R^{m\times K}$ 代表了属性聚合到 K 个主题之后的后验概率;$S\in R^{K\times K}$ 用来构建包含 K 维的主题矩阵;另外,$T\in R^{n\times K}$ 代表

了一个用户 u 对应于 K 个主题的后验概率.

在获得矩阵分解的结果之后,我们可以决定用户属于的主题:从用户-主题矩阵 T 中,我们能够直观地获得用户属于每一个主题的概率.为了推导用户-主题概率矩阵 T,优化问题可能通过如下的更新公式来完成.

$$T_{jk} \leftarrow T_{jk} \frac{(M^T FS)_{jk}}{(TT^T M^T FS)_{jk}} \tag{2}$$

$$S_{ik} \leftarrow S_{ik} \frac{(F^T MT)_{ik}}{(F^T FST^T T)_{ik}} \tag{3}$$

$$F_{ik} \leftarrow F_{ik} \frac{(MTS^T)_{ik}}{(FF^T MTS^T)_{ik}} \tag{4}$$

对于整个算法的准确性和收敛性分析,请参考 Ding 等人的工作[25].通过如图 5 所示的迭代算法,获得最终收敛的用户-主题概率矩阵 T.

开始
1. 初始化
   初始化 $S=(F^TF)^{-1}F^TMT(T^TT)^{-1}$
2. 迭代
   更新 T:固定 F,S 的值,更新 T 的值
   更新 F:固定 T,S 的值,更新 F 的值
   更新 S:固定 F,T 的值,更新 S 的值
结束

Fig.5 Matrix factorization algorithm

图 5 矩阵分解算法

在获得用户-主题矩阵 T 后,我们能够通过公式(5)衡量用户之间的隐含联系程度:

$$sim(i,j)=\sum_{k=1}^{K} T_{ik}T_{jk} \tag{5}$$

其中,当 M 为文本-用户矩阵时,$sim(i,j)$用来衡量用户 i 和用户 j 之间的隐含文本联系程度;当 M 为好友-用户矩阵时,$sim(i,j)$用来衡量用户 i 和用户 j 之间的隐含社交联系程度.

3.2 集成隐含信息的因子图模型

在抽取了显式信息和基于矩阵分解模型生成了隐含信息后,我们利用因子图模型来构建整个模型,并集成显式的与隐含的文本与社交信息.

对于一个网络结构 $G=(V,\Psi^L,\Psi^U,X)$,其中,V 为用户集合,Ψ^L 代表了已经标记的训练数据,而 Ψ^U 代表了未标记的测试数据,X 为用户属性集合.每一个群组 $\psi_i \in \Psi$ 被一个用户的属性 x_i 和一个标签 y_i 所联系,标签是用来识别一个用户是否包含在对应的群组中.其中,y_i 的值是二元的:当值为 1,说明一个用户属于对应的群组;反之亦然.当 $X=\{x_i\}$ 且 $Y=\{y_i\}$ 时,我们有如下公式:

$$P(Y \mid X,G)=\frac{P(X,G \mid Y)P(Y)}{Y(X,G)} \tag{6}$$

其中,$P(Y|G)$代表了整个网络结构中标签的分布概率,而 $P(X|Y)$代表了基于标签 Y 所对应的属性 X 的概率分布情况.简单起见,我们假设属性的概率分布是条件独立,那么有如下公式:

$$P(Y \mid X,G) \propto P(Y \mid G)\prod_i P(x_i \mid y_i) \tag{7}$$

其中,$P(x_i|y_i)$是在给定标签 y_i 的情况下,对于整体的属性 x_i 的概率.因此,我们的问题转变为如何估计 $P(Y|G)$和 $P(x_i|y_i)$.在实践中,它们可以通过很多种方式来估计.在我们的研究中,通过马尔可夫随机场(Markov random field)来进行模型的构建.基于 Hammersley-Clifford 理论[26],$P(Y|G)$和 $P(x_i|y_i)$能够被如下的公式来实现:

$$P(x_i \mid y_i)=\frac{1}{Z_1}\exp\left\{\sum_{j=1}^{d}\alpha_j f_j(x_{ij},y_i)\right\} \tag{8}$$

$$P(Y \mid G)=\frac{1}{Z_2}\exp\left\{\sum_i\sum_{j\in TR(i)}t(i,j)\right\}+\frac{1}{Z_3}\exp\left\{\sum_i\sum_{j\in NB(i)}g(i,j)\right\}+\frac{1}{Z_4}\exp\left\{\sum_i\sum_{j\in SR(i)}h(i,j)\right\} \tag{9}$$

其中,(i,j)是作为输入网络的一组对象,$TR(i)$代表了节点 i 的隐含文本联系,$NB(i)$代表了节点 i 的显式社交关系邻居,$SR(i)$代表了节点 i 的隐含社交联系.

如何定义属性函数$\{f(x_{ij},y_i)\}_j$和 3 个因子函数 $t(i,j)$,$g(i,j)$与 $h(i,j)$.

本地文本属性函数$\{\boldsymbol{f}(\boldsymbol{x}_{ij},\boldsymbol{y}_i)\}_j$. 这是用来定义每个用户 i 的文本属性函数.我们将本地的文本属性作为一组特征[27].计算一个人的所有属性函数的本地熵:

$$\frac{1}{Z_1}\exp\left(\sum_i\sum_k\alpha_k f_k(x_{ik},y_i)\right) \tag{10}$$

其中,α_k是函数的权值,代表了属性 k 的影响程度.我们使用文本的一元语法(unigram)作为基本的文本特征.如前所述,对于每个用户来说,文本信息是从用户发表的日志和评论文本中抽取.

隐含文本因子函数 $\boldsymbol{t}(\boldsymbol{y}_i,\boldsymbol{y}_j)$. 这是用来定义用户之间的隐含文本关联程度的因子函数.如果用户 i 和另外一个用户 j 之间的隐含文本相似度大于阈值,定义如下的隐含文本因子函数:

$$t(y_i,y_j)=\exp\{\beta_{ij}(y_i-y_j)^2\} \tag{11}$$

其中,β_{ij}是函数的权值,代表用户 i 到用户 j 的关联程度.用户之间的隐含文本相似度 $sim(i,j)$是通过第 3.1 节的矩阵分解模型计算获得的.

显式社交联系因子函数 $\boldsymbol{g}(\boldsymbol{y}_i,\boldsymbol{y}_j)$. 这是用来定义用户之间显式社交联系的因子函数.当用户 i 和用户 j 是好友,那么我们定义如下显式社交联系因子函数:

$$g(y_i,y_j)=\exp\{\gamma_{ij}(y_i-y_j)^2\} \tag{12}$$

其中,γ_{ij}是函数的权值,代表 i 到 j 的关联程度.

隐含社交联系函数 $\boldsymbol{h}(\boldsymbol{y}_i,\boldsymbol{y}_j)$. 这是用来定义用户之间的隐含社交联系的因子函数.与 $t(y_i,y_j)$类似,如果用户 i 和另外一个用户 j 之间的隐含社交相似度大于阈值,定义如下的隐含社交因子函数:

$$h(y_i,y_j)=\exp\{\delta_{ij}(y_i-y_j)^2\} \tag{13}$$

其中,δ_{ij}是函数的权值,代表 i 到 j 的关联程度.用户之间的隐含社交关联相似度 $sim(i,j)$是通过第 3.1 节的矩阵分解模型计算获得的.

图 6 给出了整个隐含因子图模型的描述.

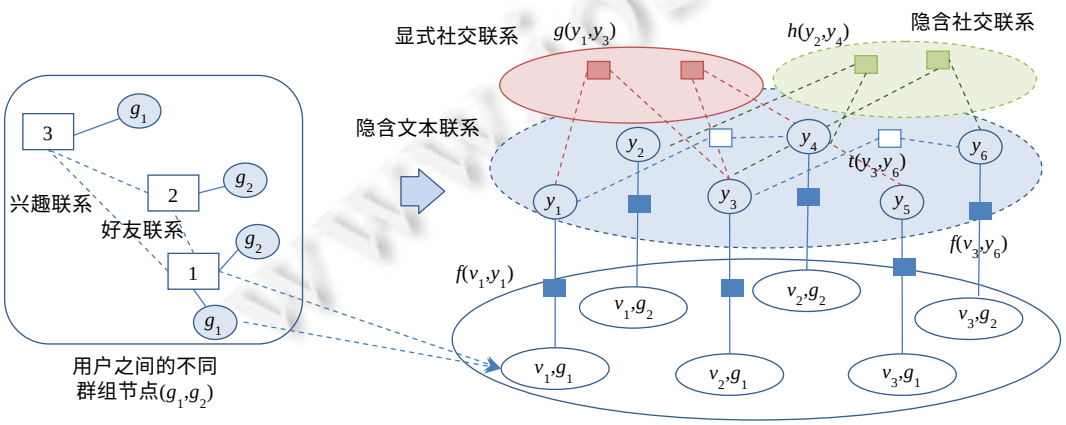


Fig.6 An example of latent facot graph model

图 6 隐含因子图模型示例

从图 6 中能够发现:每个用户与每个候选群组组成一个向量(比如 v_1),向量的属性即为本地的显式文本信息(比如 $f(v_1,y_1)$),而向量之间通过 3 种联系进行连接,即,显式社交联系函数(比如 $g(y_1,y_3)$)、隐含社交联系函数(比如 $h(y_2,y_4)$)与隐含文本联系函数(比如 $t(y_3,y_6)$).

3.3 模型的学习与预测

对于根据上述方法构建的因子图模型,我们使用环置信传播算法(loopy belief propagation,简称 LBP)进行学习和预测.学习因子图模型的目的是为了寻找对于参数 $\theta=(\{\alpha\},\{\beta\},\{\gamma\},\{\delta\})$最好的设定值.其中,优化函数为最大化似然函数 $L(\theta)$:

$$\theta^*=\arg\max L(\theta) \tag{14}$$

由于整个社交网络结构是随意的,因此我们使用环置信传播算法来预测最大边际分布.为了解释如何学习这些参数,我们用基于目标函数来得到每个 γ_k 梯度的方式进行解释.对于 γ_k(显示社交联系因子函数的权重),我们的目标是要求:

$$\frac{L(\theta)}{\gamma_k}=E[g(i,j)]+E_{P_{\gamma_k}(Y|X,G)}[g(i,j)] \tag{15}$$

其中:$E[g(i,j)]$是在给定输入网络结构数据分布的基础上的因子函数 $g(i,j)$的期望;同时,$E_{P_{\gamma_k}(Y|X,G)}[g(i,j)]$ 是在整个模型分布上的期望,即 $P(y_i|X,G)$的期望.

基于边际概率,梯度计算的目的是求所有三元组的和(参数 α_k,β_{ij} 和 δ_{ij} 的计算方式是类似的).在我们的研究中,需要运行 LBP 算法两次,其中:第 1 次用来预测未知变量的边际概率,第 2 次用来预测整个网络所有对的边际概率.最终,当完成所有边际概率的预测之后,我们可以更新所有参数的值[28,29].整个置信传播算法的计算过程如图 7 所示.

输入:网络 G, 学习速率 η
输出:预测参数 θ
初始化 $\theta\leftarrow 0$
迭代
1) 通过执行 LBP 算法计算未知变量的边缘分布,即 $P(y_i|x_i,G)$
2) 通过执行 LBP 算法计算每个变量的边缘分布 $P(y_i,y_j|X_{(i,j)},G)$
3) 通过公式(10)计算 β_k 梯度(α也是通过类似公式计算)
4) 基于学习速率 η,更新参数 θ:

$$\theta_{new}=\theta_{old}+\eta\frac{L(\theta)}{\theta}.$$

直到收敛

Fig.7 Belief propagation algorithm

图 7 置信传播算法

从学习的过程中可以看出,额外的 LBP 是用来预测未知联系的标签.因此,在学习过程结束之后,所有的未知样本已经被最大的边际概率加上标签了,如下:

$$Y^*=\arg\max L(Y|X,G,\theta) \tag{16}$$

其中,$y_i=1$ 说明用户包含所对应的群组,反之亦然.

4 实验分析

本节首先对于用到的数据集进行统计与分析,并介绍实验设置;然后,将验证我们提出的基于文本和社交信息的隐含因子图模型的性能.

4.1 隐含与显式联系统计

我们首先给出隐含与显式联系的统计分析(见表 2).其中,当两个用户之间的隐含相似度大于 0.75 时,我们认为这两个用户存在隐含联系.我们能够发现:对于 1 584 个用户,一共有 12 586 条边,后者的数量要远远大于前

者.在所有边中,显式的社交联系是出现最多的边.因此我们认为,社交联系对于分析用户的兴趣小组是很有帮助的.

Table 2 Distribution of the edges on our data set

表 2 数据集上边的统计分布

	数量
用户	1 584
隐含文本联系	1 162
显式社交联系	6 963
隐含社交联系	4 461
总联系数	12 586

4.2 隐含文本联系的分析

在描述与分析实验结果之前,首先分析基于矩阵分解模型获得的文本隐含联系的主题分布情况.我们将相关的统计与实现呈现在图 8 中.矩阵分解模型一共从文本的 67 723 个单词中获得 50 个主题.

图 8 描述了 6 个基于矩阵分解模型学习得到的不同主题,我们为每个主题选择了出现频率最高的 6 个词作为示例,其中,Topic #25 主要包含年轻与旅游主题,Topic #28 主要包含找工作的主题,Topic #38 主要包含了回忆校园时光,Topic #4 主要包含了回忆过往时光,Topic #15 主要包含了工作经历.从图中能够发现,通过矩阵分解模型学习到的每个主题都比较独立以及有代表性.另外,通过矩阵分解得到的文本隐含主题与我们需要分析的用户兴趣小组也是有关联的.比如,Topic #28 与 Topic #15 都是与工作相关的主题,而我们的候选群组也有“工作”群组;又比如,Topic #38 与 Topic #25 也都是与“生活”群组相关的文本主题.

Topic #25	Topic #20	Topic #28
有趣	最初	职业
摇滚	她们	走向
团队	雨水	专业
旅游	愉快	专家
气质	历史	技术
明媚	成熟	毕业
Topic #38	**Topic #4**	**Topic #15**
资料	蓝色	房子
转眼	当时	去年
职业	原本	走向
以往	去年	意见
校园	自行车	文件
无聊	窗户	上班

Fig.8 Six examples of the topic model

图 8 主题模型中 6 个主题示例

4.3 实验设置

实验数据来自于豆瓣网.我们选择了出现频率最高的 10 个群组作为候选的群组.随机选择了 800 个用户作为训练样本,并把剩下的用户作为测试样本.我们使用 $F1$ 值($F1$-Measure)作为评价性能的指标.

我们使用一台 CPU 为 Intel Q8300、内存为 4G 的台式计算机进行实验.每次进行矩阵分解的实验时间为 3m,利用 LFGM 训练模型的时间为 30m.虽然我们训练模型的整体时间有些长,但是由于最终预测函数是一个线性函数,因此预测时间能在 1s 内完成,因此,在大规模数据集上也能迅速完成预测.

4.4 实验结果与分析

首先,将我们提出的方法和一些基准系统进行比较,并且分析了 LFGM 模型不同因素的贡献情况.

1) 与基准系统比较

为了验证基于全局最优的和同时利用隐含和显示文本和社交信息的 LFGM 模型的性能,将本文的方法和

如下的基准系统进行比较:

- Content-based,通过文本内容(textual content)来推荐邻接的群组,其中,用户之间的相近程度通过文本的余弦相似度进行计算;
- Item-based,通过社交内容来推荐邻接的群组,其中,用户之间的相近程度通过他喜欢的产品的余弦相似度来计算[19];
- MF-based,基于矩阵分解的协同过滤(collective filtering)算法用来推荐群组.协同过滤算法主要的思想是:通过好友与其喜好的产品关系构建矩阵,并基于矩阵分解得到隐含变量,从而进行群组推荐;
- MaxEnt-C,利用本地的文本内容信息作为特征来训练最大熵分类模型.本文使用 Mallet 工具包(http://mallet.cs.umass.edu)来构建最大熵分类器;
- MaxEnt-SC,利用文本内容和社交信息作为特征来训练最大熵分类模型;
- LFGM,本文提出的模型,基于隐含因子图模型来推荐个人群组,并同时利用隐含和显式的文本和社交信息.

图 9 是我们的方法和一些基准系统比较的结果,从图中发现:(1) 3 种无监督学习算法(content-based, item-based,MF-based)的效果是差不多的,这是由于单独考虑社交与文本信息对于群组分析都是有所缺失的;(2) 基于分类器的方法,比如 MaxEnt-C 和 MaxEnt-CS,可以显著地好于基于无监督学习算法,这是由于监督学习能够有指导、更好地集成文本或者社交信息;(3) 由于同时考虑了文本内容和社交内容,MaxEnt-CS 要好于单独考虑文本内容的 MaxEnt-C;(4) 由于同时考虑了隐含和显式的两方面信息,并考虑了全局最优,我们提出的 LFGM 模型能够获得最好的分类效果,比起 MaxEnt-CS 模型,在 $F1$ 值上能够提高 6.0%.

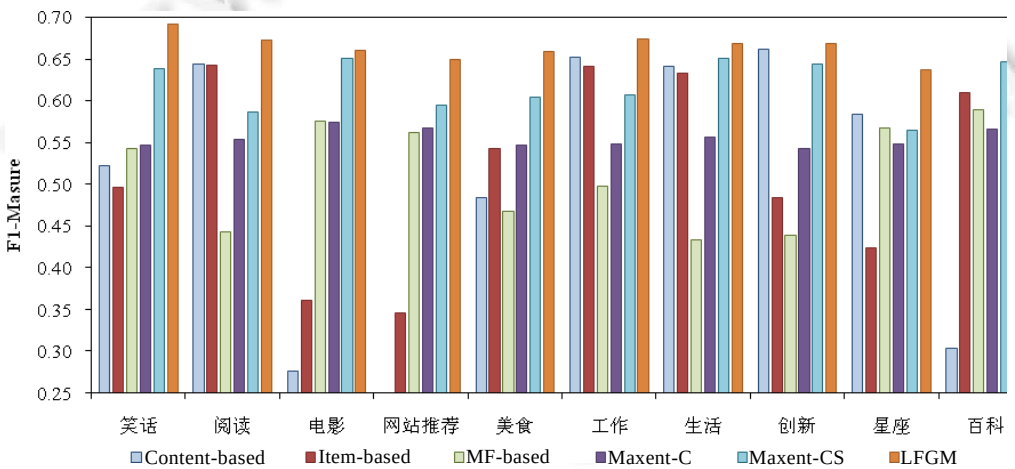


Fig.9 Comparison with different models

图 9 不同模型在群组推荐任务上的效果

2) 不同因素的贡献分析

表 3 表述了不同因素对于 LFGM 模型的贡献情况,其中,FGM-Content 是 LFGM 模型的简化版本,仅仅考虑基于文本属性函数和隐含文本联系因子函数来构建因子图模型.同样的,FGM-ExpSocial 仅仅考虑基于文本属性函数和显式社交联系函数来构建因子图模型.另外,FGM-ImpSocial 是利用文本属性函数和隐含社交联系函数来构建因子图模型.

从表 3 中我们能看出:(1) 所有的因子图模型方法都好于基准的 MaxEnt-CS 方法,这表明了基于全局优化的因子图模型相对于最大熵模型更能适合群组分析任务,这也是由于群组之间存在着相互的联系,从而导致了图模型能够获得更好的效果;(2) 由于考虑了隐含的社交联系,FGM-ImpSocial 比只考虑显式联系的 FGM-ExpSocial 的性能要好;(3) LFGM 方法比所有的其他因子图模型方法都要好,这说明了在群组推荐的时候考虑

不同因素的影响,才能获得最优的结果.

Table 3 Contribution with different factors

表 3 不同因素的贡献情况

模型	*F*1
MaxEnt-CS	0.619
FGM-Content	0.654
FGM-ExpSocial	0.667
FGM-ImpSocial	0.672
LFGM	0.679

5 总结与展望

本文主要研究基于兴趣的用户群组分析.我们基于以下两个假设:一方面,如果两个人是好友关系,或者存在相近的兴趣,那么他们非常有可能存在相近的群组;另一方面,文本信息也可以用来完善社交信息的不足,从而帮助进行群组推荐.另外,在显式的文本信息之外,潜在的文本信息,比如隐性的主题信息,也可以用来进行好友推荐.在此基础上,我们提出了一种新的隐含因子图模型,用来有效地集成上述显式和隐含信息.首先,我们使用矩阵分解模型发现并获得隐含的文本和社交信息;其次,利用显式的文本信息构建属性函数,而隐含的文本信息与社交信息一起构建因子函数;最终,利用上述信息构建一个基于全局最优的隐含因子图模型,从而利用置信传播算法分析用户感兴趣的群组.在豆瓣网上的实验证明,我们提出的隐含因子图模型能够有效地发现用户群组信息.

References:

[1] Yu J, Jin X, Han J, Luo J. Social group suggestion from user image collections. In: Proc. of the WWW 2010. 2010. 1215−1216. [doi: 10.1145/1772690.1772881]

[2] Lerman K, Jones L. Social browsing on flickr. In: Proc. of the ICWSM 2007. 2007.

[3] Negoescu R, Topickr DGP. Flickr groups and users reloaded. In: Proc. of the MM 2008. 2008. 857−860. [doi: 10.1145/1459359.1459505]

[4] Zheng N, Li Q, Liao S, Zhang L. Which photo groups should I choose? A comparative study of recommendation algorithms in flickr. Journal of Information Science, 2010,36(6):733−750. [doi: 10.1177/0165551510386164]

[5] Ohkura T, Kiyota Y, Nakagawa H. Browsing system for weblog articles based on automated folksonomy. In: Proc. of the WWW 2006. 2006.

[6] Si X, Liu Z, Sun M. Explore the structure of social tags by subsumption relations. In: Proc. of the COLING 2010. 2010. 1011−1019.

[7] Lappas T, Punera K, Sarlos T. Mining tags using social endorsement networks. In: Proc. of the SIGIR 2011. 2011. 195−204. [doi: 10.1145/2009916.2009946]

[8] Liu Z, Chen X, Sun M. A simple word trigger method for social tag suggestion. In: Proc. of the EMNLP 2011. 2011. 1577−1588.

[9] Li H, Liu Z, Sun M. Random walks on context-aware relation graphs for ranking social tags. In: Proc. of the COLING 2012. 2012. 653−662. [doi: 10.1145/2043932.2043952]

[10] Zhang B, Zhang Y, Gao K, Guo P, Sun D. Combining relation and content analysis for social tagging recommendation. Ruan Jian Xue Bao/Journal of Software, 2012,23(3):476−488 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4001.htm [doi: 10.3724/SP.J.1001.2012.04001]

[11] Liben-Nowell D, Kleinberg J. The link prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2007,58(7):1019−1031. [doi: 10.1002/asi.20591]

[12] Tang W, Zhuang H, Tang J. Learning to infer social ties in large networks. In: Proc. of the ECML/PKDD 2011. 2011. 381−397. [doi: 10.1007/978-3-642-23808-6_25]

[13] Tang J, Zhang Y, Sun J, Rao J, Yu W, Chen Y, Fong A. Quantitative study of individual emotional states in social networks. IEEE Trans. on Affective Computing, 2011,3(2):132−144. [doi: 10.1109/T-AFFC.2011.23]

[14] Al Hasan M, Chaoji V, Salem S. Link prediction using supervised learning. In: Proc. of the Workshop on Link Analysis, Counter-Terrorism and Security (SDM 2006). 2006.

[15] He J, Liu H, Du X. Mining of multi-relational association rules. Ruan Jian Xue Bao/Journal of Software, 2007,18(11):2752−2765 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/2752.htm [doi: 10.1360/jos182752]

[16] Elson D, Dames N, McKeown K. Extracting social networks from literary fiction. In: Proc. of the ACL 2010. 2010. 138−147.

[17] McCallum A, Wang X, Corrada-Emmanuel A. Topic and role discovery in social networks with experiments on enron and academic email. Journal of Artificial Intelligence Research, 2007,30:249−272. [doi: 10.1613/jair.2229]

[18] Qiu M, Yang L, Jiang J. Mining user relations from online discussions using sentiment analysis and probabilistic matrix factorization. In: Proc. of the NAACL 2013. 2013. 401−410.

[19] Zha Z, Tian Q, Cai J, Wang Z. Interactive social group recommendation for flickr photos. Neurocomputing, 2013,105(3):30−37. [doi: 10.1016/j.neucom.2012.06.039]

[20] Zhang J, Guo Y, Zhong Y. Item-Based collaborative information recommendation algorithm. Computer Engineering and Applications, 2004,15:4−6 (in Chinese with English abstract). [doi: 10.3321/j.issn:1002-8331.2004.15.002]

[21] Kubica J, Moore A, Schneider J, Yang Y. Stochastic link and group detection. In: Proc. of the AAAI 2002. 2002.

[22] Wang X, Ma J, Cui C, Gao S. Flickr group recommendation based on quaternary semantic analysis. Journal of Computational Information Systems, 2013,9(6):2235−2242.

[23] Liu Z., Chen X, Sun M. A simple word trigger method for social tag suggestion. In: Proc. of the EMNLP 2011. 2011. 1577−1588.

[24] Koha N, Hua N, Clemons E. Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. Electronic Commerce Research and Applications, 2010,9(5):374−385. [doi: 10.1109/HICSS.2010.154]

[25] Ding C, Li T, Peng W, Park H. Orthogonal nonnegative matrix tri-factorizations for clustering. In: Proc. of the SIGKDD 2006. 2006. 126−135. [doi: 10.1145/1150402.1150420]

[26] Hammersley J, Clifford P. Markov Field on Finite Graphs and Lattices. Unpublished Manuscript, 1971.

[27] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the ICML 2001. 2001. 282−289.

[28] Tang W, Zhuang H, Tang J. Learning to infer social ties in large networks. In: Proc. of the ECML/PKDD 2011. 2011. 381−397. [doi: 10.1007/978-3-642-23808-6_25]

[29] Zhuang H, Tang J, Tang W, Lou T, Chin A, Wang X. Actively learning to infer social ties. Data Mining and Knowledge Discovery. 2012,25(2):270−297. [doi: 10.1007/s10618-012-0274-x]

附中文参考文献:

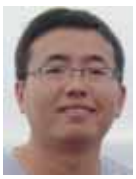
[10] 张斌,张引,高克宁,郭朋伟,孙达明.融合关系与内容分析的社会标签推荐.软件学报,2012,23(3):476−488. http://www.jos.org.cn/1000-9825/4001.htm [doi: 10.3724/SP.J.1001.2012.04001]

[15] 何军,刘红岩,杜小勇.挖掘多关系关联规则.软件学报,2007,18(11):2752−2765. http://www.jos.org.cn/1000-9825/18/2752.htm [doi: 10.1360/jos182752]

[20] 张剑,郭燕慧,钟义信.基于特征项的群组信息推荐算法.计算机工程与应用,2004,15:4−6. [doi: 10.3321/j.issn:1002-8331.2004.15.002]

王中卿(1987－),男,江苏苏州人,博士,主要研究领域为自然语言处理.

周国栋(1967－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理.



李寿山(1980－),男,博士,教授,CCF 专业会员,主要研究领域为自然语言处理.