计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20170746
56(2):306-318, 2019

基于信息融合的概率矩阵分解链路预测方法

王智强[1] 梁吉业[1,2] 李 茹[1,2]
[1](山西大学计算机与信息技术学院 太原 030006)
[2](计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)
(zhiq.wang@163.com)

Probability Matrix Factorization for Link Prediction Based on Information Fusion

Wang Zhiqiang[1], Liang Jiye[1,2], and Li Ru[1,2]
[1](*School of Computer & Information Technology, Shanxi University, Taiyuan* 030006)
[2](*Key Laboratory of Computation Intelligence & Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan* 030006)

Abstract As one kind of typical network big data, social-information networks such as Weibo and Twitter include both the complex network structure among users and rich microblog/Tweet information published by users. It is notable that most of the existing methods only make use of the network topological information or the non-topological information for link prediction, but there is still a lack of effective methods by fusing the topological information or non-topological information in social-information networks. A link prediction method is proposed from the perspective of users' topic by fusing users' topic similarity in social-information networks. The method goes in accordance with the following sequence: firstly, a topic similarity between users based on users' topic representation is defined, followed by which a topic similarity-based sparse network is constructed; secondly, the information of the following/followed network and the topic similarity-based network are fused into a unified framework of probabilistic matrix factorization, based on which the latent-feature representation of the network nodes and the linking relation parameters are obtained; finally, the linking probability between network nodes is calculated based on the obtained latent-feature representation and linking relation parameters. The proposed approach provides a general modeling strategy fusing multi-network information and a learning-based solution. Link prediction experiments are conducted on four real network datasets, i.e. Twitter and Weibo. The experimental results demonstrate that the proposed method is more effective than others.

Key words social-information network; link prediction; probability matrix factorization; fusion model; network data analysis

---

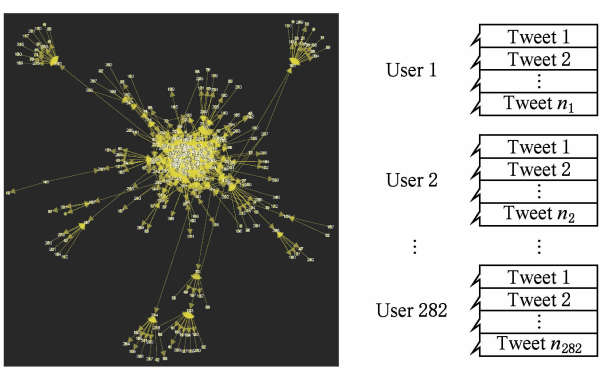
收稿日期:2017-09-30;**修回日期**:2018-03-20
基金项目:国家自然科学基金项目(U1435212,61432011,61876103);山西省重点研发计划项目(201603D111014);山西省 1331 工程项目;山西省青年科技基金项目(201701D221098)
This work was supported by the National Natural Science Foundation of China (U1435212, 61432011, 61876103), the Project of Key Research and Development Plan of Shanxi Province (201603D111014), the 1331 Engineering Project of Shanxi Province, and the Youth Technology Fundation of Shanxi Province (201701D221098).
通信作者:梁吉业(ljy@sxu.edu.cn)

摘 要 作为一种典型的网络大数据,社交信息网络如微博、Tweeter 等,不仅包含用户间复杂的网络结构,而且包含大量用户所发表的微博/Tweet 信息.现有链路预测算法大多只利用单方面的网络拓扑信息或非拓扑信息,仍然缺乏有效融合社交信息网络中拓扑与非拓扑信息的链路预测方法.为此,从社交信息网络中用户的主题角度出发,提出一种融合主题相似信息的链路预测方法.首先基于用户文本内容抽取用户的主题表示,并定义用户间的主题相似度;然后基于用户主题相似度,构建了一种用户主题相似稀疏网络;进一步将用户主题相似网络与用户间关注/被关注网络融合在统一的概率矩阵分解框架下,通过学习获得用户的潜在特征表示和网络链路参数;最终在此概率矩阵分解框架下,基于用户的潜在特征表示和链路参数计算得到用户间的链路可能性.所提出的模型提供了一种融合多种网络信息的通用策略和学习方法.实验在包含网络结构与文本信息的 4 组微博与推特数据集中显示,所提出的融合概率矩阵分解链路方法相比其他链路预测方法更有效.

关键词 社交信息网络;链路预测;概率矩阵分解;融合模型;网络数据分析

中图法分类号 TP399

社交信息(social-information)网络[1](如微博,Twitter 等),不仅包含用户间的网络结构,同时存在大量用户所分享的文本信息,具有规模大、动态变化、信息混杂等大数据通常所具有的特点,属于典型的网络大数据[2].图 1 为一个带有文本信息的 Twitter 数据集可视化示例,图 1(a)中的网络代表用户间关注/被关注(follwing/followed)的有向网络,图 1(b)分别对应网络中每个用户所分享的 Tweet 文本信息.



(a) The following/followed network (b) Users' Tweet

Fig. 1 A Twitter example of social-information network with text information

图 1 带有文本信息的社交信息网络 Twitter 示例

链路预测作为网络数据挖掘中一项基本挖掘任务[3],它在许多研究及应用领域发挥着重要作用.例如对于网络演化[4-5]、网络数据补齐[6-7]等问题的研究具有重要意义,同时在推荐[8-9]、生物信息[10]等领域发挥着重要应用价值.链路预测的研究目标[11-13]是预测当前网络中节点之间是否存在缺失的连边或是未来网络中是否会产生新的连边.本文将面向社交信息网络,针对其信息混杂特点,从信息融合角度研究社交信息网络中的链路预测问题.

目前针对链路预测的研究已有来自计算机、物理、生物等许多不同领域的研究者,针对不同的网络提出针对性的链路预测方法.现有链路预测方法通常从网络的拓扑结构出发,通过定义网络中节点间的拓扑度量或通过建立网络的统计学习模型来获得网络中节点间链路的可能性.已有基于度量的方法中,如基于邻居信息[14-16]、路径信息[17-18]、随机游走[19-20]等度量方法考虑了网络中节点间的拓扑相似度\距离.基于学习的方法如层次网络模型[6]、随机块模型[21-22]、潜在特征模型[23-25]及矩阵分解模型[26-27]等从网络结构全局出发来学习节点间的链路可能性,相比基于度量的方法具有更好的适用范围.

尽管当前针对不同网络的链路预测方法研究已经取得了许多进展,但面向信息混杂的社交信息网络的链路预测研究仍然有待深入.用户所分享的文本信息作为众多形式信息中广泛存在的数据载体,如何充分利用用户所分享的文本信息来提高面向社交信网络的链路预测是一个挑战.社交信息网络中,信息的传播依赖于网络中节点间的关注/被关注关系,1 条 Tweet 会从其发布者传递到其跟随者(关注者),同时在 Tweet 的传播过程中也会影响到网络中链路关系的形成.如当用户对某条 Tweet 十分感兴趣时,该用户将有一定可能性成为此信息发布者的关注者.当然除此之外,社交信息网络中链路的变化与信息传播间相互影响的情况还有许多.从建模角度来讲,如何将拓扑的网络结构与非拓扑的文本信息进行融合建模是面向社交信息网络链路预测的关键.

本文试图从信息融合的角度,在抽取用户主题信息的基础上,构建用户间的主题相似网络,并提出

融合用户间关注/被关注网络与主题相似网络的概率矩阵分解模型;进一步基于模型学习到的用户潜在特征表示与用户间的链路参数计算用户间的链路可能性,实现面向社交信息网络的链路预测.

1 相关研究

目前针对网络中链路预测问题所提出的方法主要包括:基于度量的方法、基于分类的方法、基于概率图的方法及基于矩阵分解方法等.

1.1 基于度量的方法

通过定义网络节点间拓扑或非拓扑上的度量来对网络节点间的连边可能性进行打分是链路预测中最为常见的方法.拓扑度量是利用网络中节点对间的各种拓扑信息来定义,如基于邻居信息的度量有公共邻居(common neighbors, CN)[15]、Jaccard[28]、大度节点有利(hub promoted, HP)[29]、大度节点不利(hub depressed, HD)[13]、邻居贡献(adamic-adar, AA)[14]、偏好性(preferential attachment, PA)[16]以及资源分配(resource allocation, RA)[7]等.Katz[17],Local Path[18],FriendLink[30]等方法利用了网络中节点间路径信息;平均通勤时间(average commute time, ACT)[13]、余弦相似度时间(cosine similarity time, CST)[19]、重启随机游走(random walk with restart, RWR)[31]、SimRank[20]、局部随机游走(local random walk, LRW)[32]等则是基于网络随机游走过程来定义.社交网络中常见的非拓扑信息包括用户的简介、标签、文本信息、关键词等[25,33-35].Matthew 等人[33]基于用户标签进行用户相似度度量.Prantik 等人[35]基于关键词分析社交网络用户的相似度.已有基于拓扑或非拓扑信息度量的链路方法通常与具体网络的性质、领域等密切相关,单一度量方法往往容易领域受限.

1.2 基于分类的方法

基于分类的方法是将链路预测看作二分类问题,其中将网络中存在连边的节点对看作正例,没有连边的节点对看作负例.在建立分类模型时,许多分类器如支撑向量机(support vector machine, SVM)、逻辑斯蒂回归(logistic regression, LR)、k 近邻(k-nearest neighbor, kNN)、朴素贝叶斯(Naive Bayesian, NB)等被用在链路预测问题中,而影响预测结果的关键因素之一是特征选择问题.Sa 等人[36]利用节点对间的公共邻居、Jaccard、邻居贡献、偏好性及局部路径等来作为分类的特征;Lichtenwalter 等人[37]提出了一种基于节点搭配的局部拓扑特征;Leskovec 等人[38]利用节点的度和三元组特征来实现链路预测;Chiang 等人[39]扩展了 Leskovec 等人[38]的工作,提出从更长的网络环路中来抽取拓扑特征.同 1.1 节中的度量方法类似,分类模型的特征选择同样容易受网络性质、领域等影响,且类的非平衡性问题也是制约基于分类模型进行链路预测的一个瓶颈.

1.3 基于概率图模型的方法

概率图模型是对网络数据进行建模的一类重要方法,它能够从网络全局出发揭示深层次的网络结构与性质.Clauset 等人[6]提出一种层次网络模型,将网络建模为一种层次结构,其中叶子节点对应网络节点,中心节点对应不同叶子节点间的链路概率.随机块模型[21-22]假设网络节点可以被划分为不同的组,节点间的链路可能性由节点所在的组决定.潜在特征模型[23-24]是一种产生式模型,它假设网络中的节点可以用潜在特征向量来表示,网络中连边的产生式基于节点的潜在特征表示和连边产生的分布假设.已有网络概率图模型通常能够获得较好的链路预测效果,然而此类模型具有较高的计算复杂度,难以适用于较大规模的网络,且难以扩展到具有混杂信息的网络中.

1.4 基于矩阵分解的方法

矩阵分解方法是通过对网络的邻接矩阵进行低秩近似来解决链路预测问题.目前,矩阵分解方法如奇异值分解(singular value decomposition, SVD)[40]、非负矩阵分解(non-negative matrix factorization, NNMF)[41]及概率矩阵分解(probabilistic matrix factorization, PMF)[42]等已被广泛用于推荐系统.相比基于推荐方法,基于矩阵分解的网络链路预测方法研究仍有待深入.Menon 等人[26]通过在矩阵分解时加入排序损失来克服链路预测中的非平衡问题,在链路预测准确率与效率 2 方面都获得了较好的结果.Zhai 等人[27]提出一种结合矩阵分解和自编码的链路预测方法,获得了较好的链路预测效果.矩阵分解方法通常可以采用随机梯度下降等快速算法来实现,容易适用于大规模的网络数据[26].总体来讲,矩阵分解方法在解决网络链路预测问题时通常具有时间与效果两方面的综合优势,但目前仍缺乏面向带有混杂信息网络的矩阵分解链路预测方法.

本文试图从矩阵分解方法入手,通过结合社交信息网络中用户的文本内容信息,建立一种能够融合用户网络结构和文本信息的矩阵分解方法,在此基础上实现面向社交信息网络的链路预测.

2 融合的概率矩阵分解模型

社交信息网络既包含用户间的网络拓扑结构信息,同时还包括丰富的用户文本信息.本文基于概率矩阵分解模型来建模的目标是为了将2种信息融合在统一的概率分解框架下,以获得新空间下融合2种信息的用户表示,将此模型称为:融合概率矩阵分解模型(fusion probability matrix factorization, FPMF).以下分3小节来介绍FPMF模型.为便于理解,在此将文中涉及的主要符号汇总为表1:

Table 1 The List of the Main Notations

表1 主要符号列表

Notation	Description
$n\in\mathbb{R}$	The number of network nodes
$N_{\boldsymbol{A}}$	The following/followed network
$\boldsymbol{A}\in\mathbb{R}^{n\times n}$	The adjacency matrix of $N_{\boldsymbol{A}}$
$N_{\boldsymbol{S}}$	The topic similarity based network
$\boldsymbol{S}\in\mathbb{R}^{n\times n}$	The adjacency matrix of $N_{\boldsymbol{S}}$
$\boldsymbol{U}\in\mathbb{R}^{n\times l}$	Users' latent-feature matrix
$g_{\boldsymbol{A}}$	The relation function between nodes in $N_{\boldsymbol{A}}$
$g_{\boldsymbol{S}}$	The relation function between nodes in $N_{\boldsymbol{S}}$
$\boldsymbol{W}$	The link parameter matrix of $N_{\boldsymbol{A}}$

2.1 用户主题相似网络构建

同质性[43-44]是社交网络中普遍存在的一种特性,它指社交网络中具有相似性的人之间容易产生链接,而存在链接的人之间也会因为相互影响而具有一定相似性.同质性说明社交网络中链路的产生与用户相似性之间具有一定相关性.本节试图从用户的主题相似度出发,利用社交信息网络中用户分享的文本信息抽取用户主题表示,并定义用户间的主题相似度,进一步基于主题相似度构建一种用户间的主题相似网络.

为获取用户主题信息,我们将每个用户历史所分享的所有微博信息看作一个文档,n个用户组成包含n个文档的文档集合D.基于潜在狄利克雷(latent Dirichlet allocation, LDA)主题模型[45],抽取每个用户的主题分布表示(见4.1节关于数据预处理的内容).

由于LDA模型所抽取的用户主题向量是一种多项分布表示,而KL(Kullback-Leibler)[46]散度适合于度量分布之间的距离,因此首先采用了KL散度来度量用户间的主题分布距离;然后基于指数函数$\exp(-x)$将此距离映射到(0,1)之间作为相似度度量.考虑到KL度量是一种非对称的距离度量,对其进行了对称化处理.具体地,用户主题相似度定义如下:

定义1. 用户主题相似度.给定任意用户i与用户j的主题分布表示$\boldsymbol{t}_i=(t_{i,1},t_{i,2},\cdots,t_{i,K})$($t_{i,k}\in(0,1)$,$k\in(1,K)$,$\sum_{k=1}^{K}t_{i,k}=1$)和$\boldsymbol{t}_j=(t_{j,1},t_{j,2},\cdots,t_{j,K})$($t_{j,k}\in(0,1)$,$k\in(1,K)$,$\sum_{k=1}^{K}t_{j,k}=1$),用户$i$与用户$j$之间的主题相似度定义如下

$$S_{ij}=\exp\left(-\frac{D_{\mathrm{KL}}(\boldsymbol{t}_i\|\boldsymbol{t}_j)+D_{\mathrm{KL}}(\boldsymbol{t}_j\|\boldsymbol{t}_i)}{2}\right),\quad(1)$$

其中,$D_{\mathrm{KL}}(\boldsymbol{t}_i\|\boldsymbol{t}_j)$代表从用户$i$到用户$j$之间的主题距离,$\frac{D_{\mathrm{KL}}(\boldsymbol{t}_i\|\boldsymbol{t}_j)+D_{\mathrm{KL}}(\boldsymbol{t}_j\|\boldsymbol{t}_i)}{2}$则是将主题距离进行对称化,进一步通过指数函数将其映射到(0,1),作为最终的主题相似度.

基于主题相似度(式(1)),定义用户之间的主题相似网络如下:

定义2. 用户主题相似网络.用四元组$N_{\boldsymbol{S}}=(V,E,\boldsymbol{S},S_{\mathrm{thr}})$表示用户主题相似度网络.$V$表示此网络的节点集合;$E$表示网络中边的集合;$\boldsymbol{S}$表示该网络的邻接矩阵,其中矩阵$\boldsymbol{S}$中每个元素的取值$S_{ij}$代表用户$i$与用户$j$之间的主题相似度;$S_{\mathrm{thr}}$表示构建主题相似网络的稀疏化参数,即当用户间的主题相似度值大于S_{thr}时,为其建立一条无向的边,反之不建立边,因此,S_{thr}的取值直接影响主题相似网络的稀疏性,S_{thr}越大网络越稀疏.

如图2所示,一个包含282个用户的Twitter数据集的主题相似网络构建过程.图2中从左至右分别表示用户的Tweets信息、主题向量以及用户间的主题相似网络,其中网络稀疏化参数$S_{\mathrm{thr}}=0.7$.需要提及的是,对用户主题相似网络进行稀疏化,主要基于2方面考虑:一方面在利用用户大量的文本信息来计算用户间主题相似度时不可避免会有噪声,稀疏化设置使后续建模中只关注相似性较大的连边,缓解噪声影响;另一方面,网络稀疏化能够减小模型的计算复杂度.稀疏化参数S_{thr}的具体选取将在实验部分(4.4.3节)进行详细实验分析与讨论.

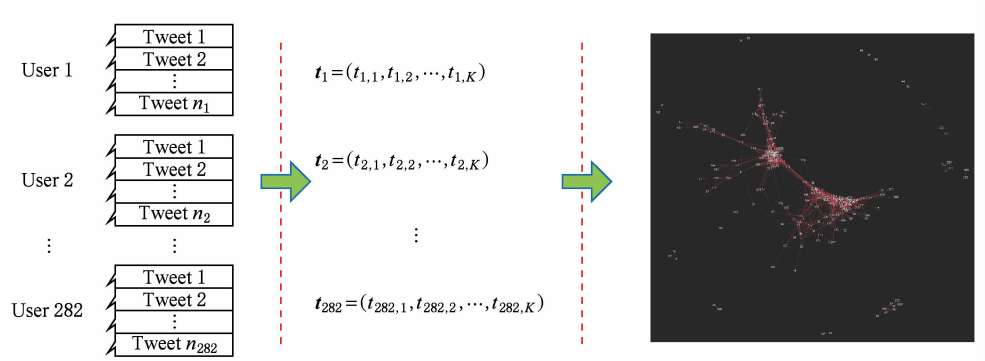


Fig. 2 The construction process of the users' topic similarity-based network

图 2 用户主题相似网络构建

2.2 FPMF 模型

社交信息网络中用户间的关系是一种有向的关注/被关注的关系,而 2.1 节所构建的主题相似网络则刻画了用户间的主题相似性. 2 种网络代表了 2 种不同语义,本节试图建立一种融合的概率矩阵分解模型将 2 种不同语义的网络融合在统一的概率矩阵分解框架下.

给定用户之间关注/被关注网络 N_A 和用户主题相似网络 N_S. 融合概率矩阵分解模型是基于概率产生式框架,此框架下 2 种网络在满足一定概率分布假设下被同时产生. 模型描述之前需要引入 2 组假设:一是假设网络中每个节点 i 用潜在特征向量 $\boldsymbol{U}_i\in\mathbb{R}^{1\times L}$ 来表示;二是引入参数矩阵 $\boldsymbol{W}\in\mathbb{R}^{L\times L}$ 度量网络中节点间关注/被关注的链路关系,其具体度量方式见后续模型介绍.

从概率产生式角度,网络 N_A 和 N_S 的产生式过程可以用 3 个步骤来描述:

第 1 步. 对于任意 $i\in\{1,2,\cdots,n\}$,有 $\boldsymbol{U}_i\sim N(0,\sigma_U^2\boldsymbol{I})$;

第 2 步. 对于每个 $l\in\{1,2,\cdots,L\}$,有 $\boldsymbol{W}_l\sim N(0,\sigma_W^2\boldsymbol{I})$;

第 3 步. 对于任意 $i\neq j\in\{1,2,\cdots,n\}$,有 $A_{ij}\sim N(g_A(\boldsymbol{U}_i,\boldsymbol{U}_j),\sigma_A^2)$,$S_{ij}\sim N(g_S(\boldsymbol{U}_i,\boldsymbol{U}_j),\sigma_S^2)$.

其对应的概率图如图 3 所示.

在网络产生式过程的第 1 步,网络中每个节点 i 的潜在向量 $\boldsymbol{U}_i\in\mathbb{R}^{1\times L}$ 从高斯先验分布 $N(0,\sigma_U^2\boldsymbol{I})$ 中产生,网络中所有节点的潜在特征矩阵 $\boldsymbol{U}$ 的先验分布满足

$$P(\boldsymbol{U}\mid\sigma_U^2)=\prod_{i=1}^{N}N(\boldsymbol{U}_i\mid\boldsymbol{0},\sigma_U^2\boldsymbol{I}). \tag{2}$$

然后模型中的链路关系矩阵参数 $\boldsymbol{W}$ 同样从先验分布为 $N(\boldsymbol{0},\sigma_W^2\boldsymbol{I})$ 的高斯分布中产生(见网络产生式过程的第 2 步),$\boldsymbol{W}$ 产生的先验分布满足

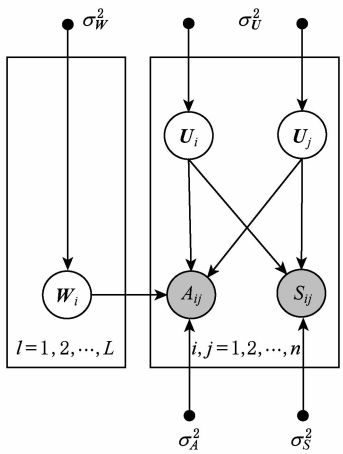


Fig. 3 The probabilistic graph model of FPMF

图 3 FPMF 模型概率图表示

$$P(\boldsymbol{W}\mid\sigma_W^2)=\prod_{l=1}^{L}N(W_l\mid\boldsymbol{0},\sigma_W^2\boldsymbol{I}). \tag{3}$$

在生成网络节点的潜在特征表示矩阵 $\boldsymbol{U}$ 和网络 N_A 的链路关系矩阵参数 $\boldsymbol{W}$ 的基础上,进一步产生网络 N_A 和 N_S 中的连边与非连边. 见网络产生式过程的第 3 步:网络 N_A 中任意节点对 $\langle i,j\rangle$ 间的连边 $A_{ij}=1$ 或非连边 $A_{ij}=0$ 是基于高斯分布 $N(g_A(\boldsymbol{U}_i,\boldsymbol{U}_j),\sigma_A^2)$ 产生的,即

$$P(A_{ij}\mid\sigma_A^2)=N(A_{ij}\mid g_A(\boldsymbol{U}_i,\boldsymbol{U}_j),\sigma_A^2), \tag{4}$$

其中,$g_A(\boldsymbol{U}_i,\boldsymbol{U}_j)$ 是为了度量网络 N_A 中任意节点对 $\langle i,j\rangle$ 间链路关系而定义的关系函数:

$$g_A(\boldsymbol{U}_i,\boldsymbol{U}_j)=\boldsymbol{U}_i\boldsymbol{W}\boldsymbol{U}_j^{\mathrm{T}}. \tag{5}$$

之所以采用 $\boldsymbol{U}_i\boldsymbol{W}\boldsymbol{U}_j^{\mathrm{T}}$ 来定义节点间的关系,是由于 $\boldsymbol{U}_i\boldsymbol{W}\boldsymbol{U}_j^{\mathrm{T}}$ 代表一种泛化的节点间关系度量,而当 $\boldsymbol{W}$ 取值为单位矩阵时,$\boldsymbol{U}_i\boldsymbol{W}\boldsymbol{U}_j^{\mathrm{T}}$ 等同于欧氏空间的内积 $\boldsymbol{U}_i\boldsymbol{U}_j^{\mathrm{T}}$,属于 $\boldsymbol{U}_i\boldsymbol{W}\boldsymbol{U}_j^{\mathrm{T}}$ 的一种特例. 针对社交网络中用户间链路语义的复杂性,用 $\boldsymbol{U}_i\boldsymbol{W}\boldsymbol{U}_j^{\mathrm{T}}$ 更适合捕捉其中复杂的链路语义. 在模型求解时,参数 $\boldsymbol{W}$ 需要从具体的网络数据中学习出来,以适应具体网络的链路

语义复杂性.此外,Zhu 等人[47]验证了在有向网络中用 $\boldsymbol{A}\approx\boldsymbol{UWU}^{\mathrm{T}}$ 的近似方式相比在推荐系统中经常使用的 $\boldsymbol{R}\approx\boldsymbol{UZ}^{\mathrm{T}}$($\boldsymbol{R}$ 表示推荐系统中的评分矩阵,$\boldsymbol{Z}$ 表示产品的潜在特征矩阵)更适合于表示网络中节点语义间的传递性.

基于网络 N_A 中每对节点连边与非连边的产生过程,其概率分布可以记为

$$P(\boldsymbol{A}\mid\sigma_A^2)=\prod_{i=1}^{N}\prod_{j=1}^{N}N(A_{ij}\mid g_A(\boldsymbol{U}_i,\boldsymbol{U}_j),\sigma_A^2). \quad (6)$$

在此统一的概率产生式框架下,主题相似网络 N_S 中任意节点对〈i,j〉连边与不连边同样基于一个高斯分布 $N(g_S(\boldsymbol{U}_i,\boldsymbol{U}_j),\sigma_S^2)$而产生,即

$$P(S_{ij}\mid\sigma_S^2)=N(S_{ij}\mid g_S(\boldsymbol{U}_i,\boldsymbol{U}_j),\sigma_S^2), \quad (7)$$

其中,$g_S(\boldsymbol{U}_i,\boldsymbol{U}_j)$是度量主题相似网络 N_S 中任意节点对〈i,j〉之间的链路关系度量函数,定义如下:

$$g_S(\boldsymbol{U}_i,\boldsymbol{U}_j)=\boldsymbol{U}_i\boldsymbol{U}_j^{\mathrm{T}}. \quad (8)$$

不同于网络 N_A,主题相似网络 N_S 是从主题角度刻画节点间相似性的网络,属于一种相似语义网络.为刻画 N_S 中节点间的相似语义特性,这里 g_S 中采用内积 $\boldsymbol{U}_i\boldsymbol{U}_j^{\mathrm{T}}$ 来定义节点潜在特征间的语义相似性.

依据第 3 步中主题相似网络 N_S 的产生过程,其概率分布表示为

$$P(\boldsymbol{S}\mid\sigma_S^2)=\prod_{i=1}^{N}\prod_{j=1}^{N}N(S_{ij}\mid g_S(\boldsymbol{U}_i,\boldsymbol{U}_j),\sigma_S^2). \quad (9)$$

基于网络 N_A 和网络 N_S 的整个产生过程,融合概率矩阵分解的目标就是通过极大化网络中节点的潜在特征表示 $\boldsymbol{U}$ 和链路参数矩阵 $\boldsymbol{W}$ 的后验分布,来估计极大后验下的 $\boldsymbol{U}$ 和 $\boldsymbol{W}$,即

$$\arg\max_{\boldsymbol{U},\boldsymbol{W}}P(\boldsymbol{U},\boldsymbol{W}\mid\boldsymbol{A},\boldsymbol{S},\sigma_A^2,\sigma_S^2,\sigma_W^2). \quad (10)$$

由于网络 N_A 和 N_S 是已知可观察的,式(10)中的后验分布正比于如下联合概率分布:

$$P(\boldsymbol{U},\boldsymbol{W}\mid\boldsymbol{A},\boldsymbol{S},\sigma_A^2,\sigma_S^2,\sigma_W^2)\propto P(\boldsymbol{U},\boldsymbol{W},\boldsymbol{A},\boldsymbol{S},\sigma_A^2,\sigma_S^2,\sigma_W^2). \quad (11)$$

依据图模型的产生式过程描述,联合概率分布 $P(\boldsymbol{U},\boldsymbol{W},\boldsymbol{A},\boldsymbol{S},\sigma_A^2,\sigma_S^2,\sigma_W^2)$可以分解为

$$P(\boldsymbol{U},\boldsymbol{W},\boldsymbol{A},\boldsymbol{S},\sigma_A^2,\sigma_S^2,\sigma_W^2)=P(\boldsymbol{A}\mid\boldsymbol{U},\boldsymbol{W},\sigma_A^2)P(\boldsymbol{S}\mid\boldsymbol{U},\sigma_S^2)P(\boldsymbol{W}\mid\sigma_W^2)P(\boldsymbol{U}\mid\sigma_U^2). \quad (12)$$

因此,式(10)等价于

$$\arg\max_{\boldsymbol{U},\boldsymbol{W}}P(\boldsymbol{A}\mid\boldsymbol{U},\boldsymbol{W},\sigma_A^2)P(\boldsymbol{S}\mid\boldsymbol{U},\sigma_S^2)\times P(\boldsymbol{W}\mid\sigma_W^2)P(\boldsymbol{U}\mid\sigma_U^2). \quad (13)$$

将式(2)(3)(6)(9)带入式(13),并取对数形式,可以推导出式(13)的优化目标等价于最小化如下目标函数:

$$\ell=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(A_{ij}-\boldsymbol{U}_i\boldsymbol{W}\boldsymbol{U}_j^{\mathrm{T}})^2+\frac{\lambda_S}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(S_{ij}-\boldsymbol{U}_i\boldsymbol{U}_j^{\mathrm{T}})^2+\frac{\lambda_U}{2}\sum_{i=1}^{N}\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}}+\frac{\lambda_W}{2}\sum_{l=1}^{L}\boldsymbol{W}_l\boldsymbol{W}_l^{\mathrm{T}}+M, \quad (14)$$

其中,M 是常量,$\lambda_S=\sigma_S^2/\sigma_A^2$ 可以看作模型中信息融合时的权重参数.$\lambda_U=\sigma_U^2/\sigma_A^2$ 与 $\lambda_W=\sigma_W^2/\sigma_A^2$ 可以看作模型中的正则化参数.更简洁地,式(14)记作:

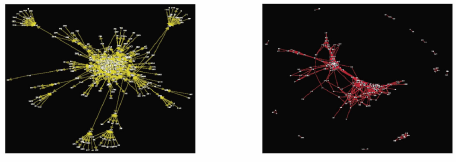
$$\ell=\|\boldsymbol{A}-\boldsymbol{UWU}^{\mathrm{T}}\|_{\mathrm{F}}^2+\frac{\lambda_S}{2}\|\boldsymbol{S}-\boldsymbol{UU}^{\mathrm{T}}\|_{\mathrm{F}}^2+\frac{\lambda_U}{2}\|\boldsymbol{UU}^{\mathrm{T}}\|_{\mathrm{F}}^2+\frac{\lambda_W}{2}\|\boldsymbol{WW}^{\mathrm{T}}\|_{\mathrm{F}}^2+M. \quad (15)$$

目标函数的局部最小值可以采用梯度下降方法进行求解[47],参数 $\boldsymbol{U}$ 和 $\boldsymbol{W}$ 的梯度下降公式为

$$\frac{\partial\ell}{\partial\boldsymbol{U}}=(\boldsymbol{UW}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{UW}+\boldsymbol{UW}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{UW}^{\mathrm{T}}-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{UW}-\boldsymbol{AUW}^{\mathrm{T}})+\lambda_S(2\boldsymbol{UU}^{\mathrm{T}}\boldsymbol{U}-\boldsymbol{SU}-\boldsymbol{S}^{\mathrm{T}}\boldsymbol{U})+\lambda_U\boldsymbol{U}, \quad (16)$$

$$\frac{\partial\ell}{\partial\boldsymbol{W}}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{UW}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}-\boldsymbol{U}^{\mathrm{T}}\boldsymbol{AU}+\lambda_W\boldsymbol{W}. \quad (17)$$

以包含有 282 个用户的社交信息网络 Twitter 为例(如图 4),图 4(a)为用户之间的关注/被关注网络 N_A,图 4(b)为构建的用户主题相似网络 N_S.模型的意义在于:通过共享同一个网络节点的潜在特征表示 $\boldsymbol{U}$,将网络 N_A 邻接矩阵 $\boldsymbol{A}$ 的近似 $\boldsymbol{A}\approx\boldsymbol{UWU}^{\mathrm{T}}$ 和网络 N_S 邻接矩阵 $\boldsymbol{S}$ 的近似 $\boldsymbol{S}\approx\boldsymbol{UU}^{\mathrm{T}}$ 融合在一个矩阵分解模型中.



$\boldsymbol{N}\approx\boldsymbol{UWU}^{\mathrm{T}}$ (a) N_A　　$\boldsymbol{S}\approx\boldsymbol{UU}^{\mathrm{T}}$ (b) N_S

Fig. 4 Sketch of the FPMF model

图 4 FPMF 模型示意图

2.3 模型复杂度分析

由于矩阵 $\boldsymbol{A}$ 与矩阵 $\boldsymbol{S}$ 的稀疏性,$\boldsymbol{A}$ 与 $\boldsymbol{U}$ 乘积的计算复杂度为 $O(\eta_A L)$,其中 η_A 为稀疏网络 $\boldsymbol{A}$ 中非 0 项个数;类似地,$\boldsymbol{SU}$ 的乘积复杂度为 $O(\eta_S L)$,η_S 为矩阵 $\boldsymbol{S}$ 中的非零项个数;剩余部分乘积的计算复杂度为 $O(nL^2)$.因此,模型求解时每次迭代的总时间复杂度为 $O(\eta_A L+\eta_S L+nL^2)$.由于网络的稀疏性,网络中边的个数几乎是网络节点数的倍数,则有 $\eta_A=O(n)$ 与 $\eta_S=O(n)$,因此模型每次迭代的复杂度理论上与网络节点个数成近似的线性关系.

3 基于 FPMF 的链路预测

基于 FPMF 模型,能够学习到网络 N_A 和网络 N_S 中节点共同的潜在特征表示 $\boldsymbol{U}$ 以及网络 N_A 中的链路参数矩阵 $\boldsymbol{W}$. 依据模型中所描述的网络 N_A 的产生式过程,对于任意节点对$\langle i,j\rangle$间产生连边的概率可以计算如下:

$$P(\langle i,j\rangle=1)=P(A_{ij}=1|\boldsymbol{U}_i,\boldsymbol{U}_j,\boldsymbol{W},\sigma_A^2)$$
$$=\frac{1}{\sqrt{2\pi}\sigma_A^2}\exp\left(-\frac{(1-\boldsymbol{U}_i\boldsymbol{W}\boldsymbol{U}_j^{\mathrm{T}})^2}{2\sigma_A^2}\right). \quad (18)$$

基于任意节点对$\langle i,j\rangle$间的连边可能性度量能够实现链路预测,算法 1 为最终基于 FPMF 模型的链路预测算法.

算法 1. 基于 FPMF 模型的链路预测.

输入:关注网络邻接矩阵 $\boldsymbol{A}$、主题相似网络邻接矩阵 $\boldsymbol{S}$;

输出:链路预测的 *AUC*(area under the receiver operating characteristic)值和 *Accuracy* 值.

① 初始化:网络节点的潜在特征矩阵 $\boldsymbol{U}_{N\times L}$,网络 $\boldsymbol{A}$ 的链路参数矩阵 $\boldsymbol{U}_{L\times L}$,$L=integer\leqslant n$;

② REPEAT

$$\boldsymbol{U}^{\mathrm{new}}:=\boldsymbol{U}^{\mathrm{old}}-\gamma\frac{\partial\ell}{\ell\boldsymbol{U}},\text{见式(16)};$$

$$\boldsymbol{W}^{\mathrm{new}}:=\boldsymbol{W}^{\mathrm{old}}-\gamma\frac{\partial\ell}{\ell\boldsymbol{W}},\text{见式(17)};$$

③ UNTIL CONVERGENCE

④ FOR 所有未知节点对$\langle i,j\rangle$ do

$P(A_{ij}=1|\boldsymbol{U}_i,\boldsymbol{U}_j,\boldsymbol{W},\sigma_A^2)$;

END FOR

⑤ 基于 P 值对所有未知节点对进行排序;

⑥ 基于排序计算链路预测结果的 *AUC* 值和 *Accuracy* 值,见式(19)(20);

⑦ RETRUN: *AUC* 和 *Accuracy*.

4 实验及结果分析

4.1 实验数据

验证本文方法所需的数据集需同时包含用户间的网络结构和用户的文本信息,本文采用了两种社交信息网络数据集,Weibo 和 Twitter,均来源于唐杰团队在文献[48]中所共享的数据. 由于原始数据集中有许多用户没有发表过或者很少发表微博/Tweets,难以满足用户主题抽取需求,因此实验中仅抽取了网络中发表或转发微博/Tweets 数量超过 100 条的用户. 最终在 2 个原始的数据集中分别抽取了 4 组子数据集:Weibo 1,Weibo 2,Twitter 1,Twitter 2,如表 2 所示:

Table 2 Datasets

表 2 数据集

Datasets	Tweets	Links	Nodes	Density%
Twitter 1	13 792	1 475	282	1.84
Twitter 2	16 187	1 551	337	1.37
Weibo 1	51 597	3 182	977	0.33
Weibo 2	70 654	5 877	1 378	0.31

针对中文数据集 Weibo 1 和 Weibo 2 的微博文本预处理主要包括:首先去除掉其中的非文本字符,如数字、表情符号等;然后进行分词和停用词去除处理,并对用户发表或转发微博时经常会出现的"分享"、"转发"、"微博"等频繁词也进行去除处理. 针对 Twitter 数据集,由于文献[48]中的原始数据集已对Tweets进行了预处理,本实验中不作进一步处理.

在数据集预处理基础上,为获得用户主题表示,我们将网络中每个用户所有发表或转发微博/Tweets 看作一个文档,然后采用 Scikit-learn[49] 工具包所提供的 LDA 主题抽取工具对网络用户的主题进行抽取,获得用户的分布式主题向量表示.

4.2 实验设置与评价指标

依据链路预测相关文献通常采用的实验设置[6,13,27],实验中将数据集中的网络划分为训练集和测试集. 具体地,本实验将网络划分为 90%∶10%,其中 90%的部分作为训练集,10%作为测试集,且划分时分别执行独立随机划分 5 次,每次随机划分保证训练集网络的连通性.

结果评价时,采用了 2 种评价指标:一是链路预测中最为常用的 *AUC* 评价指标[13];另一种是链路预测结果的准确率指标 *Accuracy*.

1) *AUC* 指标

AUC 是指 ROC 曲线(receiver operating characteristic curve)下方的面积. 实际计算时可以采用抽样的方法来计算 *AUC* 值[13,50],即在测试集中随机选择一条边的分值比在不存在的边集合中随机选择一条边的分值高的概率. 具体计算公式为

$$AUC=\frac{n'+0.5n''}{n}, \quad (19)$$

其中,n'表示随机从测试集与不存在边的集合中抽取 n 次出现 n'次抽取的测试集边的分值大于没有

边的分值,而 n'' 是分值相等的次数.

2) *Accuracy* 指标

Accuracy 是指预测的准确率,其计算公式为:

$$Accuracy=\frac{k}{L}\times 100\%, \tag{20}$$

其中,L 表示测试集中共有 L 条边,k 表示在排序前 L 个节点对中正确的个数.

4.3 比较方法

为验证本文方法在社交信息网络中链路预测的有效性,设置了 2 部分比较实验:一是本文方法与 2 种基线(baseline)方法作比较;二是本文与已有相关文献中常见的链路预测方法进行比较.

4.3.1 基线方法

基线方法的设定是为了验证本文融合模型将 2 种信息(网络和主题)在融合后相比融合前的优势.具体设定了 2 种基线方法:

1) 基本概率矩阵分解的方法

基本概率矩阵分解的方法是只针对网络结构进行概率矩阵分解,即只针对网络 $\boldsymbol{A}$ 进行概率矩阵分解,其形式化表示为:

$$\arg\max_{\boldsymbol{U},\boldsymbol{W}}\ell=\|\boldsymbol{A}-\boldsymbol{U}\boldsymbol{W}\boldsymbol{U}^{\mathrm{T}}\|_{\mathrm{F}}^{2}+\frac{\lambda_U}{2}\|\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\|_{\mathrm{F}}^{2}+\frac{\lambda_W}{2}\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_{\mathrm{F}}^{2}, \tag{21}$$

其中,$\lambda_U=\sigma_U^2/\sigma_A^2$,$\lambda_W=\sigma_W^2/\sigma_A^2$,可以看作模型中的正则化参数.为便于描述,将此模型记为 BPMF(basic probabilistic matrix factorization).

2) 基于主题相似度量的方法

基于主题相似度量的方法是利用用户的主题信息,通过度量用户间主题相似度的大小进行链路预测.用户间主题相似度的计算见式(1),此方法记为 TS(topic similarity).

4.3.2 已有的链路预测方法

1) Katz

Katz 是在已有文献[13]中被验证表现最好的方法之一.对于给定网络中的节点对$\langle i,j\rangle$,Katz 方法定义如下:

$$Katz_{i,j}=\sum_{l=1}^{m}\beta^{l}\,|\,paths_{i,j}^{l}\,|. \tag{22}$$

它是将网络中节点对$\langle i,j\rangle$间所有路径的条数求和来为节点对之间是否连边进行打分.$Katz_{i,j}$ 值越大说明$\langle i,j\rangle$间连边可能性越大.式(22)中 $paths_{i,j}^{l}$ 表示节点对$\langle i,j\rangle$间所有路径长度为 l 的路径集合,β^l 对应路径长度为 l 的权重(本文实验中设置 $\beta=0.001$).

2) 公共邻居(CN)[13]

CN 指标是链路预测中非常广泛使用的指标,它在许多网络也被验证通常具有良好的表现[13].由于社交信息网络中用户之间是关注与被关注的关系,用户之间具有不同类型的邻居,即关注者或被关注者.因此,在定义用户间 CN 指标时利用了不同类型的邻居,最终采用 4 种 CN 指标:CN1,CN2,CN3,CN4,如表 3 所示:

Table 3 CN Indexes

表 3 CN 指标

CN Indexes	Formulas
CN1	$\lvert\Gamma^{-}(i)\cap\Gamma^{-}(j)\rvert$
CN2	$\lvert\Gamma^{+}(i)\cap\Gamma^{+}(j)\rvert$
CN3	$\lvert(\Gamma^{-}(i)\cap\Gamma^{+}(i))\cap(\Gamma^{-}(j)\cap\Gamma^{+}(j))\rvert$
CN4	$\lvert(\Gamma^{-}(i)\cup\Gamma^{+}(i))\cap(\Gamma^{-}(j)\cup\Gamma^{+}(j))\rvert$

表 3 中 $\Gamma^{+}(i)$表示社交信息网络中关注用户 i 的用户集合,$\Gamma^{-}(i)$表示被用户 i 关注的用户集合,$|\Gamma^{+}(i)|$和$|\Gamma^{-}(i)|$分别代表集合 $\Gamma^{+}(i)$与 $\Gamma^{-}(i)$的用户数量.

3) 偏好性(PA)

偏好性指标 PA 也是在社交网络中常用的链路预测方法[13].给定网络中用户 i 与用户 j 的邻居数量,PA 指标等于 2 个用户邻居数据的乘积.同 CN 类似,依据社交信息网络中不同类型的邻居,采用了 4 组不同的 PA 指标:PA1,PA2,PA3,PA4,如表 4 所示:

Table 4 PA Indexes

表 4 PA 指标

PA Indexes	Formulas
PA1	$\lvert\Gamma^{-}(i)\rVert\Gamma^{-}(j)\rvert$
PA2	$\lvert\Gamma^{-}(i)\rVert\Gamma^{-}(j)\rvert$
PA3	$\lvert\Gamma^{-}(i)\cap\Gamma^{+}(i)\rVert\Gamma^{-}(j)\cap\Gamma^{+}(j)\rvert$
PA4	$\lvert\Gamma^{-}(i)\cup\Gamma^{+}(i)\rVert\Gamma^{-}(j)\cup\Gamma^{+}(j)\rvert$

4) 低秩近似(low-rank approximation, LAR)

低秩近似方法的目的是寻找一个低秩矩阵用来近似原始的矩阵.在链路预测中,低秩近似方法通过计算得到一个秩为 k 的最优低秩矩阵 $\boldsymbol{A}_k$ 用来近似原始的网络邻接矩阵 $\boldsymbol{A}$.进行链路预测时,将矩阵 $\boldsymbol{A}_k$ 中的第 i 行和第 j 列对应位置取值作为节点对$\langle i,j\rangle$间链路的打分.文献[12]采用了奇异值分解 SVD 方法对原始网络矩阵进行低秩近似表示.本文实验中不仅采用了 SVD 方法,同时也采用了非负

矩阵分解方法 NNMF,将 2 种基于低秩近似表示的链路预测方法分别记为:LAR-SVD 和 LAR-NNMF.

4.4 实验结果及分析

实验中经过反复测试发现,参数分别进行设置时结果最优:

Twitter 1 中,$L=80$,$S_{thr}=0.9$,$\lambda_S=0.8$,

Twitter 2 中,$L=40$,$S_{thr}=0.97$,$\lambda_S=0.6$,

Weibo 1 中,$L=80$,$S_{thr}=0.9$,$\lambda_S=1.6$,

Weibo 2 中,$L=60$,$S_{thr}=0.9$,$\lambda_S=0.6$.

除上述参数外,模型中的正则化参数设置为 $\lambda_U=\lambda_W=0.05$. 以下实验中若非特别说明,上述参数均为最优结果时的参数.

4.4.1 基线方法比较

表 5 为本文方法 FPMF 与基线方法 BPMF 和 TS 在 4 组数据集中的比较结果.

Table 5 Comparison of Baseline Methods

表 5 基线方法比较(*AUC*±*std*)

Datasets	Baseline Methods		Our Method
	BPMF	TS	FPMF
Twitter 1	0.7921±0.1436	0.6978±0.0188	**0.9518±0.0057**
Twitter 2	0.8414±0.1437	0.6466±0.0128	**0.9501±0.0009**
Weibo 1	0.9635±0.0195	0.5857±0.0130	**0.9894±0.0025**
Weibo 2	0.9527±0.0118	0.5202±0.0086	**0.9743±0.0116**

从表 5 中可以看出 FPMF 方法在 4 组数据集中比 2 个基线方法获得了更好的 *AUC* 结果. 2 种基线方法 BPMF 和 TS 分别是基于用户网络结构和主题相似信息进行链路预测,而 FPMF 是将 2 种信息融合在统一的概率产生式框架下进行链路预测. 实验中融合后的方法能够优于融合前的 2 种方法,验证了本文融合概率矩阵分解的有效性.

4.4.2 与已有方法比较

图 5 与图 6 分别显示了本文方法与已有 4 类共 11 种常见链路预测方法比较的 *AUC* 和 *Accuracy* 结果. 从图 5,6 中可看到,FPMF 方法在 4 组数据集中大多数情况下获得了最好的 *Accuracy* 和 *AUC*,并在不同的数据集中 FPMF 方法能够获得较为稳定的结果. 比较方法中,基于公共邻居的 4 种方法(CN1,CN2,CN3,CN4)在 2 组 Weibo 数据集上的表现都要差于 2 组 Twitter 数据集,说明其预测能力的不稳定. 在基于偏好指标的方法(PA1,PA2,PA3,PA4)中,PA1 与 PA2 表现较好,且能够获得相比公共邻居更好的 *AUC* 结果,同时在 2 类数据集中保持了较好的稳定性,但就 *Accuracy* 结果而言明显低于本文 FPMF 方法. 同偏好指标类似,基于 Katz 的方法能够获得相对不错的 *AUC* 结果,但 *Accuracy* 结果也明显低于本文方法. 这说明在链路预测排序精度方面,本文 FPMF 方法能够获得更好的链路排序结果.

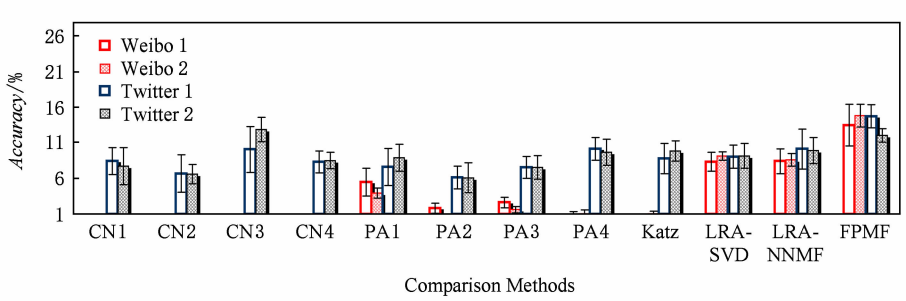


Fig. 5 *Accuracy* results of comparison methods

图 5 比较方法的 *Accuracy* 结果

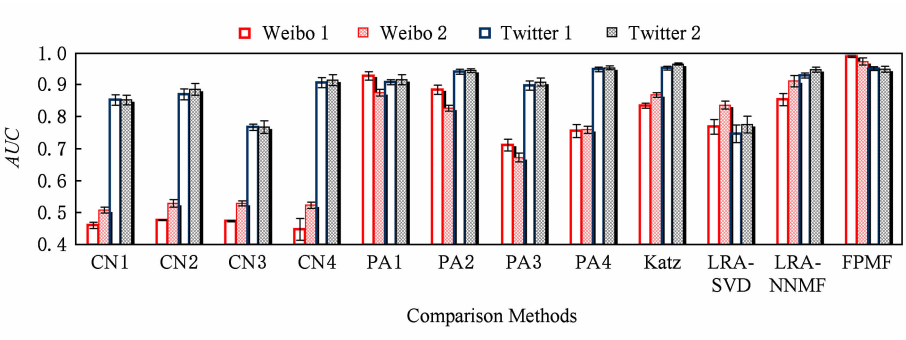


Fig. 6 *AUC* results of comparison methods

图 6 比较方法的 *AUC* 结果

相比基于公共邻居、偏好性和 Katz 方法,基于低秩近似的方法(LAR-SVD,LAR-NNMF)在不同数据集中获得了更稳定的 *AUC* 和 *Accuracy* 结果,但这 2 种方法仍然低于本文 FPMF 方法. 总的来说,本文方法既继承了低秩近似方法的稳定性,同时由于融合了社交信息网络中拓扑和非拓扑 2 方面信息,使得该方法与已有链路预测方法相比有明显提升.

4.4.3 稀疏参数 S_{thr} 分析

本文 FPMF 方法中,参数 S_{thr} 用来控制用户主题相似网络 S 的稀疏化程度,S_{thr} 取值越大主题相似网络越稀疏. 图 7 显示了稀疏化参数 S_{thr} 的不同取值在 4 组数据集中对链路预测结果和算法性能的影响. 每个图中横轴代表稀疏化参数 S_{thr} 的取值(从0.1～0.99),2 个纵轴分别代表模型算法迭代 1000 次所花费的时间和链路预测的 *AUC* 结果. 从图 7(a)～(d)中时间曲线可以看出,随着稀疏化参数 S_{thr} 取值变大,模型算法迭代所花费的时间会明显减小. 这说明用户主题相似网络越稀疏,算法运行所需时间越少. 观察图 7(a)～(d)中的 *AUC* 曲线,可以发现随着稀疏化参数 S_{thr} 取值增大,链路预测 *AUC* 的结果也在提高,而当稀疏化参数 S_{thr} 达到一定阈值时预测结果将会停止提高甚至减小. 以上现象说明用户主题相似网络的稀疏化程度在一定程度影响最终链路预测的结果与性能,合理选择 S_{thr} 取值能够有效提升链路预测结果. 分析其原因,我们认为当 S_{thr} 取值很大时,用户主题相似网络会十分稀疏,FPMF 模型中只能考虑少量的用户对之间具有很高主题相似度的信息,而当 S_{thr} 取值很小时,模型中将会考虑大量用户对之间的主题相似度信息,从而增加噪声,一定程度影响链路预测结果.

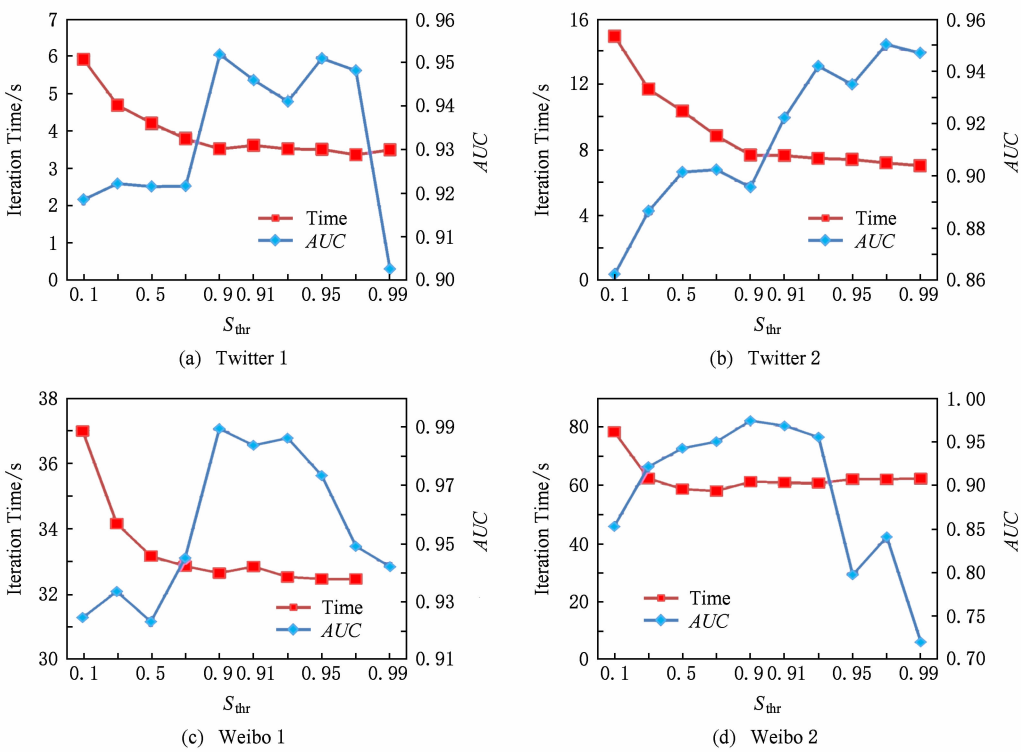


Fig. 7 Analysis of the sparse parameters

图 7 稀疏化参数分析

4.4.4 参数 λ_s 分析

λ_s 参数是平衡 FPMF 模型中用户关注/被关注网络和主题相似网络 2 种信息的. 如果模型中 $\lambda_s=0$,等同于只考虑用户间关注网络的信息(如基线中的 BPMF 方法);如果 λ_s 取无穷大,模型只考虑用户间主题相似网络的信息.

图 8(a)～(d)显示了该模型在 4 组数据集中随着 λ_s 参数取值变化时 *AUC* 结果的变化. 可以看到随着 λ_s 参数取值增大,*AUC* 结果开始提升;当 λ_s 参数增加到一定阈值时,*AUC* 结果会停止提升甚至开始下降. 这说明在模型中应选择适中的 λ_s 参数来平衡用户关注/被关注网络和用户主题相似网络 2 种信息.

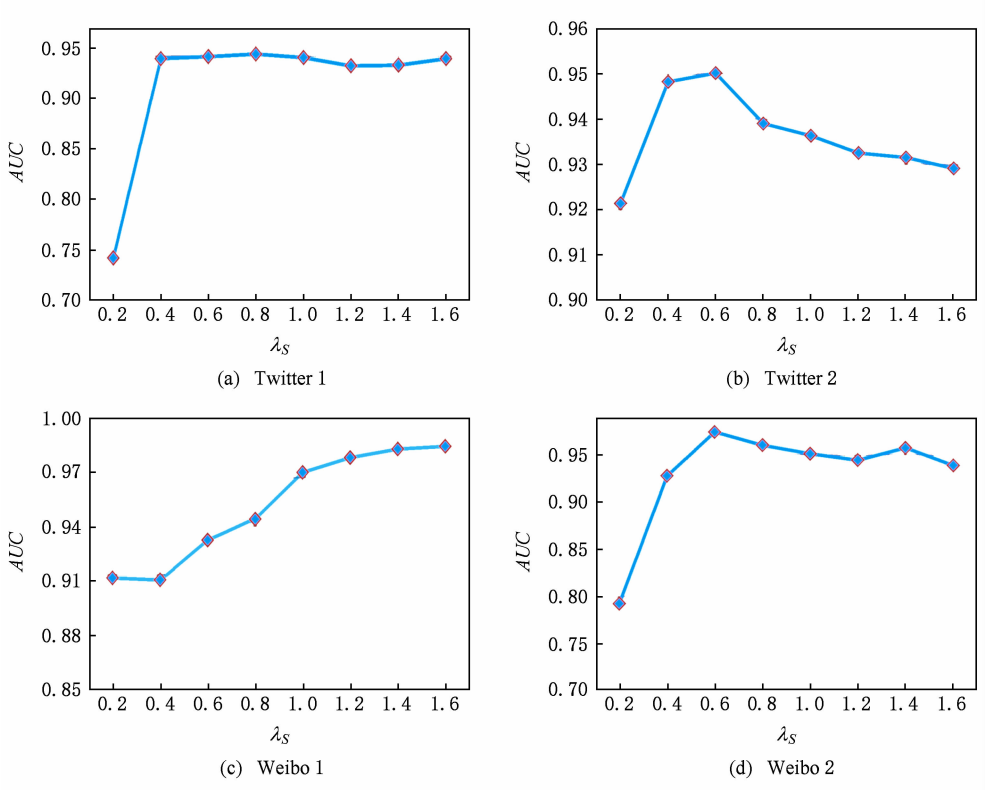


Fig. 8 Analysis of the weight parameters

图 8 权重参数分析

5 结论与展望

社交信息网络中既包含用户间关注/被关注关系的拓扑网络信息又包含大量非拓扑信息,如何充分利用 2 方面信息来提高面向社交信息网络的链路预测仍然是一个具有挑战性的任务. 本文从信息融合角度,提出一种 FPMF 链路预测模型. 该模型首先利用社交信息网络中用户的文本信息抽取用户主题表示,并基于用户主题分布表示构建用户间的主题相似网络;然后在统一的概率产生式框架下,建立融合用户间关注/被关注网络和主题相似网络的 FPMF 模型;最终基于 FPMF 模型学习的结果实现面向社交信息网络的链路预测. 下一步研究工作中,将一方面关注多种非拓扑信息的融合;另一方面关注网络的动态性,研究面向社交信息网络动态演化的增量式学习模型.

参 考 文 献

[1] Romero D M, Kleinberg J. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter [C] //Proc of the 4th AAAI Int Conf on Weblogs and Social Media. Menlo Park, CA: AAAI, 2010: 138-145

[2] Wang Yuanzhuo, Jin Xiaolong, Cheng Xueqi. Network big data: Present and future [J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138 (in Chinese)
(王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望[J]. 计算机学报, 2013, 36(6): 1125-1138)

[3] Zhao Shu, Liu Xiaoman, Duan Zhen, et al. A survey on social ties mining [J]. Chinese Journal of Computers, 2017, 40(3): 535-555 (in Chinese)
(赵姝, 刘晓曼, 段震, 等. 社交关系挖掘综述[J]. 计算机学报, 2017, 40(3): 535-555)

[4] Zhao Zeya, Jia Yantao, Wang Yuanzhuo, et al. Link inference in large scale evolutionable knowledge network [J]. Journal of Computer Research and Development, 2016, 53(2): 492-502 (in Chinese)
(赵泽亚, 贾岩涛, 王元卓, 等. 大规模演化知识网络中的关联推理[J]. 计算机研究与发展, 2016, 53(2): 492-502)

[5] Wang Li, Cheng Suqi, Shen Huawei, et al. Structure inference and prediction in the co-evolution of social networks [J]. Journal of Computer Research and Development, 2013, 50(12): 2492-2503 (in Chinese)
(王莉, 程苏琦, 沈华伟, 等. 在线社会网络共演化的结构推断与预测[J]. 计算机研究与发展, 2013, 50(12): 2492-2503)

[6] Clauset A, Moore C, Newman M E. Hierarchical structure and the prediction of missing links in networks [J]. Nature, 2008, 453(7191): 98-101

[7] Zhou Tao, Lü Linyuan, Zhang Yicheng. Predicting missing links via local information [J]. The European Physical Journal B, 2009, 71(4): 623-630

[8] Wu Sen, Sun Jimeng, Tang Jie. Patent partner recommendation in enterprise social networks [C] //Proc of the 6th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2013: 43-52

[9] Aiello L M, Barrat A, Schifanella R, et al. Friendship prediction and homophily in social media [J]. ACM Transactions on the Web, 2012, 6(2): No. 9

[10] Jansen R, Yu Haiyuan, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data [J]. Science, 2003, 302(5644): 449-453

[11] Wang Peng, Xu Baowen, Wu Yurong, et al. Link prediction in social networks: The state-of-the-art [J]. Science China Information Sciences, 2015, 58(1): 011101

[12] Liben-Nowell D, Kleinberg J. The link prediction problem for social networks [C] //Proc of the 12th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2003: 556-559

[13] Lü Linyuan, Zhou Tao. Link Prediction [M]. Beijing: Higher Education Press, 2013 (in Chinese)
(吕琳媛, 周涛. 链路预测[M]. 北京: 高等教育出版社, 2013)

[14] Adamic L A, Adar E. Friends and neighbors on the Web [J]. Social Networks, 2003, 25(3): 211-230

[15] Lorrain F, White H C. Structural equivalence of individuals in social networks [J]. The Journal of Mathematical Sociology, 1971, 1(1): 67-98

[16] Barabasi A L, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286(5439): 509-512

[17] Katz L. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18(1): 39-43

[18] Lü Linyuan, Jin Cihang, Zhou Tao. Similarity index based on local paths for link prediction of complex networks [J]. Physical Review E, 2009, 80(2): 046122

[19] Fouss F, Pirotte A, Renders J M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation [J]. IEEE Transactions on Knowledge & Data Engineering, 2007, 19(3): 355-369

[20] Jeh G, Widom J. SimRank: A measure of structural-context similarity [C] //Proc of the 8th Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2002: 538-543

[21] Edoardo M A, David M B, Stephen E F, et al. Mixed membership stochastic blockmodels [J]. Journal of Machine Learning Research, 2008, 9(5): 1981-2014

[22] Holland P W, Laskey K B, Leinhardt S. Stochastic blockmodels: First steps [J]. Social Networks, 1983, 5(2): 109-137

[23] Zhu Jun. Max-margin nonparametric latent feature models for link prediction [C] //Proc of the 29th Int Conf on Machine Learning. New York: IEEE Communications Society, 2012: 1179-1186

[24] Palla K, Knowles D, Ghahramani Z. An infinite latent attribute model for network data [C] //Proc of the 29th Int Conf on Machine Learning. New York: IEEE Communications Society, 2012: 1179-1186

[25] Wang Zhiqiang, Liang Jiye, Li Ru, et al. An approach to cold-start link prediction: Establishing connections between non-topological and topological information [J]. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(11): 2857-2870

[26] Menon A K, Elkan C. Link prediction via matrix factorization [C] //Proc of the 22nd European Conf on Machine Learning and the 15th Principles and Practice of Knowledge Discovery in Databases. Berlin: Springer, 2011: 437-452

[27] Zhai Shuangfei, Zhang Zhongfei. Dropout training of matrix factorization and autoencoder for link prediction in sparse graphs [C] //Proc of the 15th Int Conf on Data Mining. Philadelphia, PA: SIAM, 2015: 451-459

[28] Jaccard P. Etude de la distribution florale dans une portion des Alpes et du Jura [J]. Bulletin De La Societe Vaudoise Des Sciences Naturelles, 1901, 37(142): 547-579

[29] Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks [J]. Science, 2002, 297(5586): 1551-1555

[30] Papadimitriou A, Symeonidis P, Manolopoulos Y. Fast and accurate link prediction in social networking systems [J]. Journal of Systems & Software, 2012, 85(9): 2119-2132

[31] Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks, 2012, 56(18): 3825-3833

[32] Liu Weiping, Lü Linyuan. Link prediction based on local random walk [J]. Europhysics Letters, 2010, 89(5): 58007

[33] Rowe M, Stankovic M, Alani H. Who will follow whom? Exploiting semantics for link prediction in attention-information networks [C] //Proc of the 11th Int Semantic Web Conf. Berlin: Springer, 2012: 476-491

[34] Wang Zhongqing, Li Shoushan, Zhou Guodong. Personal group recommendation via textual and social information [J]. Journal of Software, 2017, 28(9): 2468-2480 (in Chinese)
(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[J]. 软件学报, 2017, 28(9): 2468-2480)

[35] Bhattacharyya P, Garg A, Wu S F. Analysis of user keyword similarity in online social networks [J]. Social Network Analysis and Mining, 2011, 1(3): 143-158

[36] Sa H R D, Prudencio R B C. Supervised link prediction in weighted networks [C] //Proc of the Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2011: 2281-2288

[37] Lichtenwalter R N, Chawla N V. Vertex collocation profiles: Subgraph counting for link analysis and prediction [C] //Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 1019-1028

[38] Leskovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks [C] //Proc of the 19th Int Conf on World Wide Web. New York: ACM, 2010: 641-650

[39] Chiang K Y, Natarajan N, Tewari A, et al. Exploiting longer cycles for link prediction in signed networks [C] // Proc of the 20th Int Conf on Information and Knowledge Management. New York: ACM, 2011: 1157-1162

[40] Golub G, Kahan W. Calculating the singular values and pseudo-inverse of a matrix [J]. Journal of the Society for Industrial & Applied Mathematics, 1965, 2(2): 205-224

[41] Lee D D, Seung H S. Algorithms for non-negative matrix factorization [C] //Proc of the 7th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000: 556-562

[42] Salakhutdinov R, Mnih A. Probabilistic matrix factorization [C] //Proc of the 14th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2007: 1257-1264

[43] Aral S, Muchnik L, Sundararajan A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(51): 21544-21549

[44] Wu Xindong, Li Yi, Li Lei. Influence analysis of online social networks [J]. Chinese Journal of Computers, 2014, 37(6): 735-752 (in Chinese)
(吴信东, 李毅, 李磊. 在线社交网络影响力分析[J]. 计算机学报, 2014, 37(4): 735-752)

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

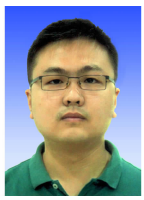
[46] Kullback S. The Kullback-Leibler distance [J]. American Statistician, 1987, 41(4): 340-341

[47] Zhu Shenghuo, Yu Kai, Chi Yun, et al. Combining content and link for classification using matrix factorization [C] // Proc of the 30th Annual Int Conf on Research and Development in Information Retrieval. New York: ACM, 2007: 487-494

[48] Zhang Jing, Liu Biao, Tang Jie, et al. Social influence locality for modeling retweeting behaviors [C] //Proc of the 22nd Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2013: 2761-2767

[49] Pedregosa F, Gramfort A, Michel V, et al. Scikit-learn: Machine learning in python [J]. Journal of Machine Learning Research, 2011, 12(10): 2825-2830

[50] Fawcett T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874



Wang Zhiqiang, born in 1987. PhD, lecturer. Member of CCF. His main research interests include social network data mining and machine learning.

Liang Jiye, born in 1962. Professor and PhD supervisor. Distinguished member of CCF. His research interests include artificial intelligence, granular computing, data mining and machine learning.

Li Ru, born in 1963. Professor and PhD supervisor. Senoir member of CCF. Her research interests include computational linguistics and Chinese information processing.