doi:10.11959/j.issn.1000-436x.2015040

融合链接拓扑结构和用户兴趣的朋友推荐方法

尚燕敏[1]，张鹏[2]，曹亚男[2]
（1. 中国科学院 计算技术研究所，北京 100190；2. 中国科学院 信息工程研究所，北京 100093）

摘 要：提出一种新的朋友推荐方法，该方法同时使用用户兴趣和朋友关系这 2 种因素来为目标用户推荐朋友，对 PageRank 算法进行改进，提出一种能同时融合上述 2 种因素的 Topic_Friend_PageRank(TFPR)模型。首先，采用 LDA(latent Dirichlet allocation)分析用户发布的消息内容，将用户表示为若干主题上的分布，从而建模用户的兴趣。接下来，使用加权的 PageRank 算法建模用户在整个链接拓扑中的重要程度和用户之间朋友关系的相似性。最后根据主题感知的 PageRank 思想，将用户兴趣融入前面提到的加权 PageRank 中，形成同时融合用户兴趣和朋友关系的 TFPR 模型。采用新浪微博数据验证所提模型的性能，实验证明该模型能同时得到较高的准确率和召回率。
关键词：社交网络；朋友关系；主题模型；PageRank 算法
中图分类号：TP391 **文献标识码**：A

New interest-sensitive and network-sensitive method for user recommendation

SHANG Yan-min[1], ZHANG Peng[2], CAO Ya-nan[2]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)

Abstract: A new hybrid approach by incorporatin gusers' interests and users' friendships together to recommend new friends for target users is proposed. A variation of PageRank—Topic_Friend_PageRank(TFPR) is proposed, which can consider user interests and user friends at same time. Firstly, proposed method uses latent Dirichlet allocation (LDA) to model users' interests, and weighted-PageRank algorithm to model users' friendship network, and then merge these two factors into TFPR. This hybrid method models users' interests and users' friendships at the same time, and wedemonstrate the effectiveness of proposed hybrid model by using some social network datasets.
Key words: social network; friendship; topic model; PageRank algorithm

1 引言

近年来，随着社交网络用户的增多，社交网络中新的朋友链接不断添加，或者由于朋友之间取消关注而导致链接不断消失。朋友链接的消失或出现是社交网络的链接拓扑随时间动态变化的重要表现。预测未来时刻可能出现的新的朋友关系是研究社交网络演化过程的重要途径之一。

这里关注社交网络的基本功能—朋友推荐。如何为一个目标用户 A 推荐最有可能与其成为朋友的其他用户，提高朋友推荐的精度是社交网络运营商普遍关注的问题，也是学术界研究的热点问题。该问题本质上可归结为链接预测问题。链接预测问题的定义如下：给定 t_1 时刻社交网络的拓扑链接，准确预测从 t_1 到未来某一时刻 t_2 这段时间出现的新的朋友链接。例如，在 Facebook 中，各方面非常相似的 2 个用户 C 和 D 在当前时刻并不存在朋友链接，那么在未来时刻可以通过链接预测的方法将用

收稿日期：2013-10-16；**修回日期**：2014-02-22
基金项目：国家高技术研究发展计划（“863”计划）基金资助项目（2011AA010705）；先导专项基金资助项目（XDA 06030200）；国家自然科学基金资助项目（61003167）
Foundation Items: The National High Technology Research and Development Program of China (863 Program)(2011AA010705); Priority Research Program (XDA06030200); The National Natural Science Foundation of China (61003167)

户 C 作为候选朋友推荐给用户 D，或将用户 D 作为候选朋友推荐给 C。总之，链接预测的目的就是帮助用户找到新的朋友。

朋友推荐方法是机器学习在社交网络领域的一个应用，这些方法为目标用户 A 推荐与其最相似的用户作为他的候选朋友。如何找到与目标用户最相似的候选朋友，以及如何能最真实地衡量目标用户与候选朋友之间的相似性是本文关注的问题。目前，相似性度量方法主要分为 2 种，第一种方法将用户兴趣作为相似性度量标准，该方法认为如果 2 个用户 A 和 B 有相似的兴趣，或最近时刻他们关注的内容主题相似，就认为用户 A 和 B 可能成为朋友，从而将 A 推荐给 B 或者将 B 推荐给 A，此方法称为基于用户兴趣的度量方法；第二种方法将 2 个用户 A 和 B 的朋友关系相似性作为度量标准。例如，用户 A 在 Facebook 中有 10 个朋友，用户 B 有 8 个朋友，其中他们有 4 个共同的朋友，那么认为用户 A 和 B 在未来某时刻可能成为朋友。该相似度度量方法关注的是用户已有的朋友关系，即社交网络的拓扑链接，称为基于朋友关系的度量方法。

上述 2 种度量方法各有利弊，基于朋友的方法具有较高准确度但召回率较低，这种方法依赖朋友关系的传递找到在目标用户的新朋友，而对于不在目标用户的朋友关系可达范围内的潜在朋友，该方法不能将其推荐给目标用户。基于兴趣的方法根据用户之间的兴趣相似性推荐朋友，这种方法只关注用户的兴趣而不考虑用户在链接拓扑中的位置和用户之间的距离，因此该方法召回率较高，但准确率较低。产生这种情况的原因为：1) 兴趣相似的 2 个用户未必会成为朋友；2) 处于目标用户的朋友关系可达范围内的潜在朋友未必与目标用户具有相似的兴趣，对于这种情况的潜在朋友，基于兴趣的方法无法将其定位。

在现实中，社交网络中的用户结交朋友主要考虑 2 个因素：用户兴趣和朋友关系（如图 1 所示）。已有的朋友推荐方法仅考虑一种因素的影响，如何同时建模这 2 个因素对朋友推荐的影响是需要解决的问题。本文提出一种基于 PageRank 框架的朋友推荐方法——TFPR，该方法同时融合了用户兴趣和朋友关系这 2 种相似性度量因素，实验证明该混合模型能平衡准确率和召回率，得到更好的性能。

2 相关工作

本节将分别介绍前面提到的基于朋友关系和基于用户兴趣的朋友推荐方法。

2.1 基于朋友关系的方法

研究者将社交网络中朋友关系组成的链接拓扑看作一张图 G（如图 1 所示），朋友推荐问题就是预测图 G 中未来时刻可能出现的新链接。目前大部分方法通过为 G 中的任意一对节点(用户)$<x, y>$ 赋予一个链接得分 $score(x, y)$来描述这对节点之间存在链接的概率，并选择概率最大的 top k 个节点对作为未来可能出现的新连接。基于朋友关系的方法在计算 2 个节点之间的链接得分时主要考虑他们

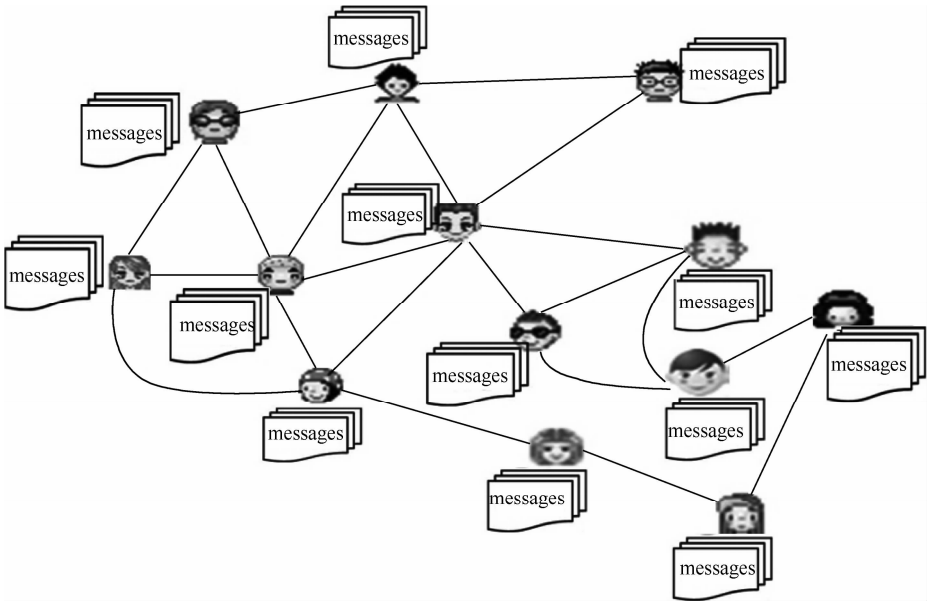

图 1 社交网络示意

之间的拓扑链接。文献[1]通过最短链的方法得到 2 个节点之间的链接相关性。还有方法通过计算 2 个节点的最短路径考察他们之间产生新链接的可能性。与此方法不同，文献[2]将 2 个节点(用户)之间所有路径的和作为衡量新连接产生的标准。除此之外，文献[3]通过 PageRank 算法计算 2 个节点之间的链接可能性。文献[4~9]通过度量 2 个节点的共同朋友比重得到他们成为朋友的概率。

总之，目前大部分朋友推荐方法依赖于社交网络的链接拓扑。因此，该类方法能将处在目标用户朋友关系可达范围内的潜在朋友全部找到，准确度高；而对于不在朋友关系可达范围内的潜在朋友，该方法无法将其推荐给目标用户，因此召回率较低。

2.2　基于兴趣的方法

朋友链接构成了社交网络的拓扑结构，用户所发的信息在该结构中流动，而流动的信息是用户之间结交朋友、交流互动的重要途径。这些信息反映了用户最近关注的事件或者兴趣，兴趣相同的 2 个用户很有可能成为朋友。基于此，许多研究者利用用户在社交网络上发布的信息预测未来可能产生的朋友关系。在文献[10,11]中，作者提出一个基于用户信息的朋友推荐系统，使用主题模型 LDA 分析用户发布的消息，并将用户表示为在一组混合主题上的分布，从而为目标用户推荐与其具有相似主题分布的用户作为朋友。该工作过度依赖用户发布的信息而忽略了用户已有的朋友关系，导致无法推荐那些与目标用户兴趣不同却与目标用户有着许多共同朋友的潜在朋友用户。除此之外，许多工作使用 LDA 的改进算法，如标签 LDA 来预测新的朋友关系。总之，这些工作从内容的层次上分析用户所发的消息、挖掘用户的兴趣，从而根据兴趣为用户推荐朋友。这类方法能找到所有与目标用户兴趣相似的用户，并将这些用户推荐给目标用户作为朋友，但现实中兴趣相似的用户未必就一定会成为朋友。比如在社交网络中距离很远的 2 个兴趣相似的用户，他们共同的朋友很少甚至不存在，那么他们成为朋友的概率就非常小。基于此，一些研究工作提出同时融合用户个人信息和朋友关系的统一预测模型。文献[12]考虑将用户的属性信息如性别、年龄等与朋友关系结合。文献[13]同时使用用户发布的信息和朋友关系来完成朋友推荐。该方法使用监督模型从用户发布的消息中提取代表用户兴趣的内容特征，同时从用户的朋友关系中抽取代表用户当前朋友链接的图特征，将上述 2 类特征作为监督模型的输入来预测2个用户之间是否会产生朋友关系。如何选择好的特征是此类方法的难点和不足。基于此，使用非监督的方法将用户消息信息和朋友关系信息融合，提出一种新的基于 PageRank 框架的混合朋友推荐模型——Topic_Friend_PageRank (TFPR)。

3　背景知识

3.1　主题模型(LDA)

在文档分析领域，主题模型是典型的用来挖掘文档主题的统计模型，该模型将语料库中的每个文档看作在不同主题上的分布，将每个主题看作词典中词的分布。主题模型是生成模型，它以概率描述一个文档的生成过程，其中使用最广泛的主题生成模型是由 Blei 等提出的 LDA[14]。图 2 给出了 LDA 的图模型表示。首先，假设语料库 D 由 M 个文档构成，$D=\{d_1,d_2,\cdots,d_m\}$；每个文档 d_i 由 N_i 个单词构成 $d_i=\{w_{i1},w_{i2},\cdots,w_{iN_i}\}$，这里每个单词 $w_{ij}\in V$，V 是词典库。那么文档的生成过程如下。

1) 对于每个文档 d_i：

生成它的主题分布 $\theta_i\sim \mathrm{Dir}(\alpha)$

2) 对于每个词 w_{ij}：

选择一个主题 $z_{ij}\sim \mathrm{Mul}(z_{ij}|\theta_i)$

选择一个词 $w_{ij}\sim \mathrm{Mul}(z_{ij})$

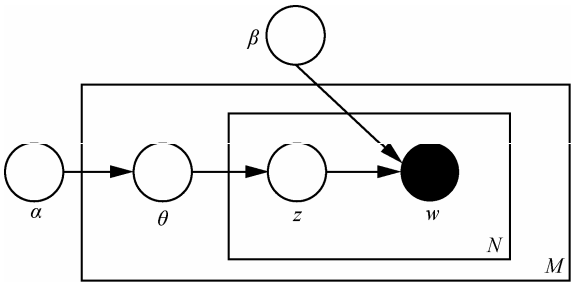

图 2　LDA 图模型

LDA 不仅用于文档分析领域，在广告预测、社交网络、个性化推荐[10,11,14,15]等领域也有广泛应用。本文使用 LDA 来预测社交网络中的朋友关系。

3.2　主题感知的 PageRank

最初，PageRank 算法用来排序搜索引擎返回的搜索结果。这项由 Google 提出的技术旨在使用网页之间的链入链出关系计算每个网页在整个链接拓扑中的重要性。近年来，有些工作将 Pagerank 算法及其变形应用在社交网络中[3,16,17]。将主题感知

的 PageRank[18]算法思想引入社交网络的朋友关系预测问题中。该算法仍是计算一个网页的重要性得分，与最初的 PageRank 算法不同，主题感知的 PageRank 算法在考虑网页之间链接结构的同时，还考虑了每个网页内容的语义信息。该算法认为每个网页侧重的主题不同，对同一个主题而言，不同网页在该主题上所占的比重影响了网页的排序。基于此，主题感知的 PageRank 算法为每个网页计算不同主题下的 PageRank 值，该值代表网页在其所处的链接结构中关于某个主题的重要性得分。每个网页在所有主题下的重要程度可由一个向量表示，该向量的每个元素即前面提到的 PageRank 值。当用户在搜索引擎上输入查询词时，主题相关的 PageRank 算法将与此查询词语义最相关的网页推荐给用户。该算法模型如下。

$$\overrightarrow{\boldsymbol{Rank}} = d\boldsymbol{M}\overrightarrow{\boldsymbol{Rank}} + (1-d)\vec{\boldsymbol{p}} \tag{1}$$

这里，用 G 表示网页链接构成的图，$\boldsymbol{M}$ 是该图的平方随机矩阵。$\vec{\boldsymbol{p}}$ 为 $N\times 1$ 维的个性化向量。在文献[17]中 $\vec{\boldsymbol{p}} = \left[v_{ji}\right]_{N\times 1}, \forall i \in web\left[v_{ji} = \dfrac{\omega_{ij}}{\sum_k \omega_{kj}}\right]$，其中 v_{ji} 描述了网页 i 中 topic j 的权重与其他网页相比所占的比重，ω_{ij} 代表网页 i 中 topic j 所占的权重。本文将该算法思想引入社交网络的朋友推荐问题中，每个用户对应一个网页，使用 LDA 从用户发布的消息中挖掘用户兴趣，构建个性化向量 $\vec{\boldsymbol{p}}$。

4 本文模型—Topic_Friend_PageRank

所提的混合模型包含 3 个子模块（如图 3 所示）。模块 1 关注用户的兴趣，该模块的主要功能是通过分析用户所发布的消息挖掘用户的兴趣。这里采用主题模型将每个用户表示为在若干主题上的一个分布。模块 2 关注用户的朋友关系。该模块的主要功能是分析用户之间已有朋友中存在的共同朋友的比例，进而得到 2 个用户朋友关系的相似程度。模块 3 以 PageRank 算法为基本框架，将前 2 个模块的成果融入其中，使每个用户对应一个 PageRank 向量，此向量中的每个元素表示该用户在现有的朋友关系基础上对某一个主题的偏好得分。最后，通过计算 2 个用户的 PageRank 向量之间的距离得到这 2 个用户成为朋友的概率。

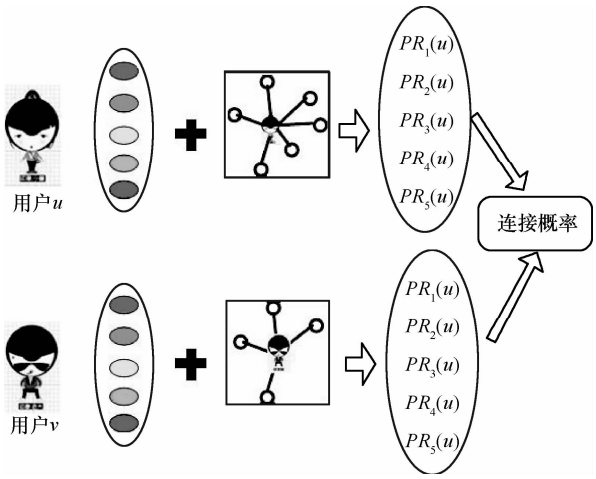

图 3 混合模型示意

4.1 抽取用户的主题分布

用户在社交网络上发布的消息反映了他们最近关注的内容，这些内容从一定程度上反应了用户的兴趣，而兴趣相同的 2 个用户很可能成为朋友，因此用户发布的消息蕴含的语义信息可以用来推荐朋友。这里，使用主题模型——LDA(狄利克雷分配)来挖掘用户消息中隐含的语义信息。

与文档建模中的“词袋子”模型类似，将每个用户发布的消息集合看作一个文档，消息中出现的词作为文档中的词。这样社交网络中的多个用户构成一个文档集，使用主题模型分析该文档集中每个文档的主题分布情况。

在图 2 中，采用吉布斯采样得到每个用户 u 在不同主题上的分布情况，并将其记为 θ_u。那么每个用户在已有朋友网络的基础上对某个主题的偏好得分可根据 3.2 节中提到的主题感知的 PageRank 算法得到，如式(2)所示。

$$PR_j(u) = (1-d)\frac{\theta_{uj}}{\sum_{u^*}\theta_{u^*j}} + d\sum_{v\in M(u)}\frac{PR_j(v)}{L(v)} \tag{2}$$

其中，$PR_j(u)$ 表示用户 u 在已有朋友关系基础上对 topic j 的偏好得分。θ_{uj} 表示从用户 u 的历史消息中挖掘出的 topic j 所占的比重。$M(u)$表示用户 u 的朋友集合，而 U 表示所有用户的集合，任意用户 $u^*\in U$。

4.2 计算朋友关系的链接强度

早前有关社交网络链接预测的大部分工作只关注用户之间是否存在链接，而忽略了链接由于用户之间熟悉程度的不同而产生的强弱之分。一对亲密朋友之间的链接关系强于普通朋友之间的

链接。本文中，考虑朋友链接的强弱差别，并将这种差别加入 PageRank 算法来辅助预测新的朋友关系。

1) 无向图网络

对于 Facebook 等由无向链接构成的社交网络，如果用户 u 在用户 v 的朋友列表中，那么用户 v 必在用户 u 的朋友列表中。假定用户 u 和 v 是朋友，那么必有一条链接 (u,v)，该链接的强弱可通过 u 和 v 的共同朋友数目来表示。具体细节如下。

对于任意一个用户 u，用 $Friend(u)$ 表示他的所有朋友集合，$Friend(u)=\{f_{u1},f_{u2},\cdots,f_{um}\}$；$Friend(v)$ 代表用户 v 的朋友集合，$Friend(v)=\{f_{v1},f_{v2},\cdots,f_{vm}\}$。链接 (u,v) 的强弱记为 $stren(u,v)$，其计算过程如下

$$Stren(u,v)=\frac{|Friend(u)\bigcap Friend(v)|}{|Friend(u)|+|Friend(v)|} \quad (3)$$

其中，$|Friend(u)\bigcap Friend(v)|$ 代表用户 u 和 v 共同朋友的数目，$|Friend(u)+Friend(v)|$ 表示 2 个用户总的朋友数目。对于用户 u 的任意一个朋友 $v^*\in Friend(u)$，链接 (u,v^*) 的强度为 $stren(u,v^*)$，链接 (u,v) 与其他链接 (u,v^*) 相比，其强度比重为

$$\pi(u,v)=\frac{Stren(u,v)}{\sum_{v^*\in Friend(u)}Stren(u,v^*)} \quad (4)$$

这里，$\sum_{v^*\in Friend(u)}\pi(u,v^*)=1$。

2) 有向图网络

对于 Twitter 等有向链接构成的社交网络，链接分为 2 种：射入链接和射出链接。例如，用户 u 关注用户 v，那么有一条链接从 u 射入 v，该链接对于用户 u 来说是一条射出链接；而对用户 v 来说是一条射入链接。对于此链接，u 关注 v，是 v 的粉丝。

粉丝　对于用户 u，他所有粉丝的集合记为 $Follower(u)$，$Follower(u)=\{fer_{u1},fer_{u2},\cdots,fer_{um}\}$；对于用户 v，他所有粉丝的集合记为 $Follower(v)$，$Follower(v)=\{fer_{v1},fer_{v2},\cdots,fer_{vn}\}$，那么有向链接 (u,v)（$(u\to v)$ 或 $(v\to u)$）的强度可通过他们的粉丝用户衡量，记为 $stren\_er(u,v)$，其计算过程如下

$$Stren\_er(u,v)=\frac{|Follower(u)\bigcap Follower(v)|}{|Follower(u)+Follower(v)|} \quad (5)$$

对于用户 u 的任意粉丝 $v^*\in Follower(u)$，用 $stren\_er(u,v^*)$ 表示有向链接 (u,v^*)（$u\to v^*$ 或 $v^*\to u$）的强度。链接 (u,v) 与其他链接 (u,v^*) 相比，其粉丝链接的强度比重为

$$\pi\_er(u,v)=\frac{Stren\_er(u,v)}{\sum_{v^*\in Follower(u)}Stren\_er(u,v^*)} \quad (6)$$

这里，$\sum_{v^*\in Follower(u)}\pi\_er(u,v^*)=1$。

关注用户　对于用户 u，他关注用户集合记为 $Following(u)$，$Following(u)=\{fing_{u1},fing_{u2},\cdots,fing_{um}\}$；对于用户 v，他关注的用户集合记为 $Following(v)$，$Following(v)=\{fing_{v1},fing_{v2},\cdots,fing_{vn}\}$，那么有向链接 (u,v)（$(u\to v)$ 或 $(v\to u)$）的强度可通过他们关注的用户衡量，记为 $stren\_ing(u,v)$，其计算过程如式(7)所示。

$$Stren\_ing(u,v)=\frac{|Following(u)\bigcap Following(v)|}{|Following(u)+Following(v)|} \quad (7)$$

对于用户 u 关注的任意用户 $v^*\in Following(u)$，用 u 表表示有向链接 (u,v^*)（$u\to v^*$ 或 $v^*\to u$）的强度。链接 (u,v) 与其他链接 (u,v^*) 相比，其关注用户的链接强度比重为

$$\pi\_ing(u,v)=\frac{Stren\_ing(u,v)}{\sum_{v^*\in Following(u)}Stren\_ing(u,v^*)} \quad (8)$$

这里，$\sum_{v^*\in Following(u)}\pi\_ing(u,v^*)=1$。

最后，同时考虑粉丝链接和关注用户链接来计算有向链接的完整强度，如式(9)所示。

$$\pi(u,v)=\gamma\pi\_er(u,v)+(1-\gamma)\pi\_ing(u,v) \quad (9)$$

这里，参数 γ 控制粉丝链接强度和关注用户链接强度之间的权重，本文中，γ 分别取值 0、1 和 0.5。

将链接强度加入 PageRank 算法中，称为加权的 PageRank 算法。例如，对于用户 u，如果存在一条有向链接 $(v\to u)$，那么根据 PageRank 算法思想，用户 v 分配给 u 的 PageRank 值如下

$$PR(u)=(1-d)\frac{1}{n}+d\sum_{v\in(Follower(u)}\pi(u,v)PR(v) \quad (10)$$

这里，$PR(u)$ 是用户 u 的 PageRank 值。

4.3 TFPR 模型框架

将前面两模块功能融合，提出 Topic_Friend_

PageRank 模型（式(11)）来预测朋友关系（如图 3 所示），由于无向朋友网较简单，这里给出有向朋友网下的 TFPR 模型。

$$PR_j(u)=(1-d)\frac{\theta_{uj}}{\sum_{u^*}\theta_{u^*j}}+d\sum_{v\in Follower(u)}\pi(u,v)PR_j(v) \tag{11}$$

本文认为朋友链接的生成有 2 个因素。第一个主要因素是考虑用户之间兴趣的相似性，如果一对用户有许多相似的爱好，那么他们成为朋友的可能性就很大。第二个因素是考虑用户之间朋友关系的相似性，如果一对用户有许多相同的朋友，那么他们成为朋友的可能性同样很大。将这 2 个因素融合，提出一种新的 PageRank 变种算法，如式(11)所示。在该模型中，从每个用户发布的消息集合中抽取该用户的主题分布，对每个主题，计算加入链接信息的 PageRank 值。式(11)表示用户 u 以概率 $1-d$ 按照某一主题 j 找到新的朋友，以概率 d 依循已有的朋友关系找到新的朋友。按照此思想，针对每个主题，都可以得到一个 PageRank 值，从而将用户表示为一个 PageRank 向量，而向量中的每个元素则对应一个主题。按照式(11)计算所有用户 $U=(u_1,u_2,\cdots,u_n)$ 的向量矩阵如下

$$\boldsymbol{TFP}=\begin{pmatrix}\boldsymbol{PR}(u_1)\\ \boldsymbol{PR}(u_2)\\ \vdots\\ \boldsymbol{PR}(u_n)\end{pmatrix}=\begin{pmatrix}PR_1(u_1) & PR_2(u_1) & \cdots & PR_K(u_1)\\ PR_1(u_2) & PR_2(u_2) & \cdots & PR_K(u_2)\\ \vdots & \vdots & \vdots & \vdots\\ PR_1(u_n) & PR_2(u_n) & \cdots & PR_K(u_n)\end{pmatrix} \tag{12}$$

接下来，任意 2 个用户之间存在链接的概率可通过计算他们对应的 PageRank 向量之间的距离来描述，然后选择概率最大的用户对作为未来可能出现的新链接。

4.4 矩阵解性质证明

本节将证明提出的 TFPR 模型满足基本 PageRank 算法的矩阵解性质，即证明向量 ***TFP***(12)是一个矩阵的特征向量，下面来构造该矩阵。

将式(11)代入式(12)，得到

$$\boldsymbol{TFP}=d\begin{pmatrix}\pi(u_1,u_1),\pi(u_1,u_2),\cdots,\pi(u_1,u_n)\\ \pi(u_2,u_1),\pi(u_2,u_2),\cdots,\pi(u_2,u_n)\\ \cdots\\ \pi(u_n,u_1),\pi(u_n,u_2),\cdots,\pi(u_n,u_n)\end{pmatrix}\boldsymbol{TFP}+$$

$$(1-d)\begin{pmatrix}r(u_{1,1}),r(u_{1,2}),\cdots,r(u_{1,K})\\ r(u_{2,1}),r(u_{2,2}),\cdots,r(u_{2,K})\\ \cdots\\ r(u_{n,1}),r(u_{n,2}),\cdots,r(u_{n,K})\end{pmatrix} \tag{13}$$

这里，在 topic j 下，对于任意的用户 u，$r(u,j)=\dfrac{\theta_{uj}}{\sum_{u^*}\theta_{u^*j}}$

$$\boldsymbol{TFP}=d\begin{pmatrix}\pi(u_1,u_1),\pi(u_1,u_2),\cdots,\pi(u_1,u_n)\\ \pi(u_2,u_1),\pi(u_2,u_2),\cdots,\pi(u_2,u_n)\\ \cdots\\ \pi(u_n,u_1),\pi(u_n,u_2),\cdots,\pi(u_n,u_n)\end{pmatrix}\boldsymbol{TFP}+$$

$$(1-d)\begin{pmatrix}\frac{\theta_{u1,1}}{\sum_{u^*}\theta_{u^*1}},\frac{\theta_{u1,2}}{\sum_{u^*}\theta_{u^*2}},\cdots,\frac{\theta_{u1,K}}{\sum_{u^*}\theta_{u^*K}}\\ \frac{\theta_{u2,1}}{\sum_{u^*}\theta_{u^*1}},\frac{\theta_{u2,2}}{\sum_{u^*}\theta_{u^*2}},\cdots,\frac{\theta_{u2,K}}{\sum_{u^*}\theta_{u^*K}}\\ \cdots\\ \frac{\theta_{un,1}}{\sum_{u^*}\theta_{u^*1}},\frac{\theta_{un,2}}{\sum_{u^*}\theta_{u^*2}},\cdots,\frac{\theta_{un,K}}{\sum_{u^*}\theta_{u^*K}}\end{pmatrix} \tag{14}$$

这样，***TFP*** 就是上式的解，证明完成。

5 实验设计与结果分析

本节通过实验验证 TFPR 模型的性能，以新浪微博为数据平台，使用主题模型 LDA 分析用户的消息，挖掘用户的兴趣。

5.1 数据描述与预处理

在本文中，使用新浪微博数据验证所提模型的性能。首先，去掉一些不活跃用户，比如发布消息少于 5 条，朋友数少于 5 个的用户。接下来，对于剩下的用户，从中选出 3 个用户子集，这些子集分别包含 50、100 和 200 个用户，记为 $D1$、$D2$ 和 $D3$。对于这些用户子集，将每个用户发布的所有消息整合为一个文档，对每个文档使用 *ICTCLAS* 进行文档分词和去停用词，并使用 LDA 抽取用户的兴趣，将每个用户表示为不同的主题上的一个分布。另

外，在进行实验之前，还需计算用户之间的链接强度。

5.2　对比模型

本文使用 3 个对比模型来验证所提模型的性能。

1) 基于粉丝关系的模型：该模型为目标用户推荐与他拥有许多共同粉丝的用户作为他的朋友。

2) 基于主题的模型：此模型考虑使用用户主题分布的相似性来预测新的链接[9]。

3) 模型 TFPR：同时使用用户兴趣和朋友关系来预测新的链接。

5.3　结果分析

1) 参数 d 的学习。在模型中（式(11)），参数 d 控制着用户的主题和朋友关系对链接预测的影响。如果 d 很小，TFPR 模型偏重于使用用户的主题来预测新的朋友关系；如果 d 很大，TFPR 模型偏重于使用用户已有的朋友关系来预测新链接。总的来说，d 不应太大或太小。为了选择合适的参数 d，在数据集 $D1$、$D2$ 和 $D3$ 上展开实验，分别考虑 γ 为 0,1 和 0.5 时，该如何选择合适的 d。实验设置如下：分别从 3 个数据集 $D1$、$D2$ 和 $D3$ 中选择 30 个用户构造 3 个测试集 $D1\_30$、$D2\_30$ 和 $D3\_30$。对于每个测试集中的用户，根据 TFPR 计算用户之间的链接概率，并将概率最大的前 20 对用户作为最可能产生朋友关系的用户，从而计算准确率和召回率。

图 4 展示了参数 d 对链接预测准确率的影响。图 4(a)描述了在测试集 $D1\_30$ 上，当 γ 分别为 0,1 和 0.5 时，不同的 d 对应不同的预测结果。从图 4(a)中看到，随着参数 d 的增加，所提模型的准确率在最初也随之增加，但当 d 达到某一阈值时，准确率随着参数 d 增加而降低。类似，图 4(b)和图 4(c)分别描述了参数 d 在测试集 $D2\_30$ 和 $D3\_30$ 上对模型在预测朋友关系准确率的影响。从图中发现，参数 d 最合适的值为 $d=0.5$。除此之外，对比这 3 个图，本文模型的准确率也随着数据集规模增加而提高。

图 5 描述了在 3 个数据集上，参数 d 对本文模型在朋友预测召回率上的影响，其影响趋势与 d 对准确率的影响相同。

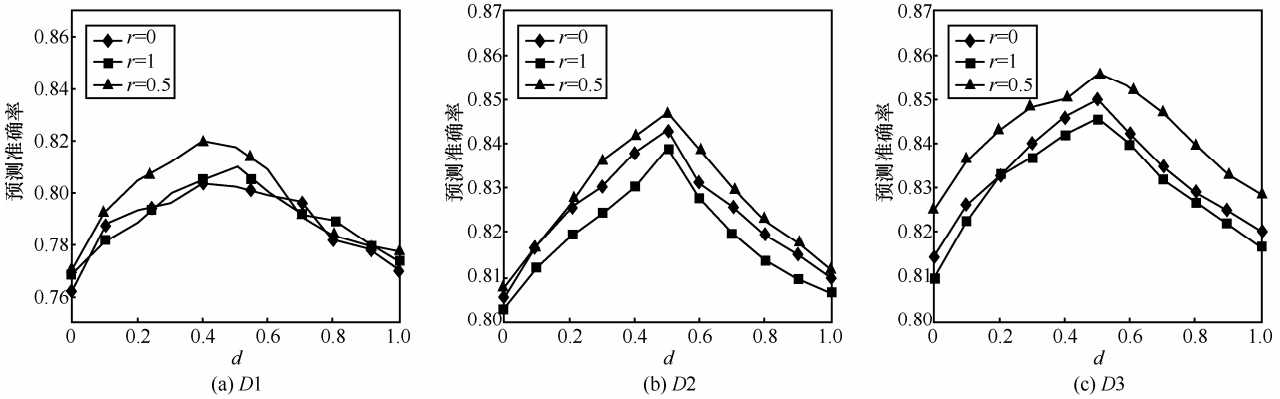

图 4　参数 d 对准确率的影响

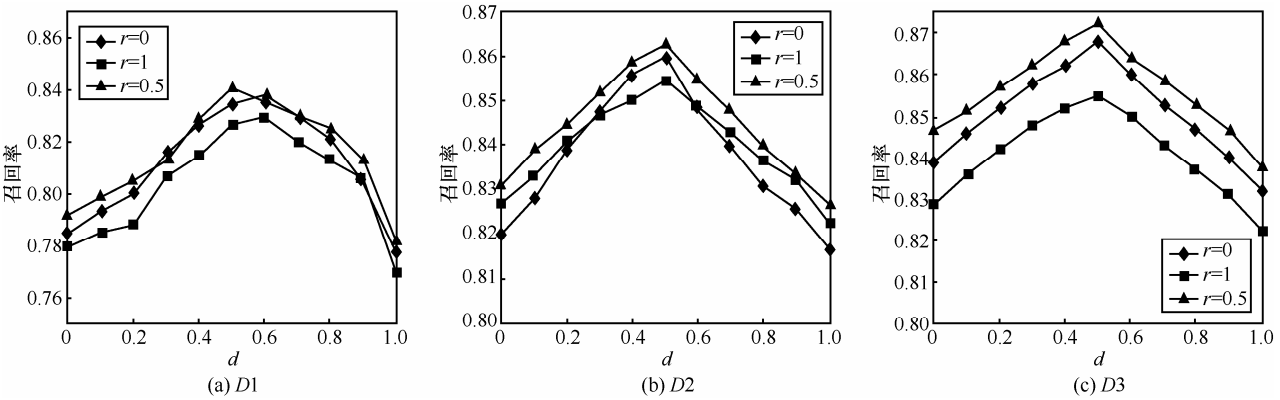

图 5　参数 d 对召回率的影响

2) 其他参数的学习。在数据集 *D*2 上学习合适的 LDA 参数，如主题数目和 LDA 迭代次数。表 1 说明了使本文的模型达到最好性能的 LDA 配置参数为：100 个主题，500 次迭代。另外，还测试了不同的向量距离计算方法对模型性能的影响：*KL* 偏差和 *Cosine* 距离。表 2 说明 *Cosine* 距离相对于 *KL* 偏差能使本文模型得到更好的性能。

表 1 LDA 参数对模型性能的影响

主题数目	迭代次数	预测准确率	召回率	*F*measure
20	100	0.690	0.813	0.746
	200	0.697	0.835	0.760
	500	0.724	0.833	0.775
50	100	0.785	0.843	0.813
	200	0.810	0.849	0.829
	500	0.821	0.857	0.839
100	100	0.834	0.855	0.844
	200	0.842	0.859	0.850
	500	**0.847**	**0.863**	**0.855**
200	100	0.820	0.851	0.835
	200	0.831	0.854	0.842
	500	0.828	0.848	0.838

表 2 向量距离度量方法对模型性能的影响

迭代次数	向量距离度量方法	*F*measure
100	*Cosine*	**0.844**
	KL	0.787
200	*Cosine*	**0.850**
	KL	0.798
500	*Cosine*	0.855
	KL	**0.796**

3) 模型对比。本实验对比 TFPR 模型与其他模型的性能。所提模型的配置：数据集 *D*2，100 个主题，500 次迭代，$d=0.5$。对比模型为基于粉丝关系的模型和基于主题的模型。表 3 展示了对比结果，说明基于粉丝关系的模型具有较高的准确率，但是丢失了一大部分内容相似的潜在朋友；基于主题的预测模型能找到内容相似的潜在朋友，但是对于兴趣不同的潜在朋友却无法定位，所以此模型具有较高的召回率，但是准确率较低。而本文模型将上述 2 种思想结合，因而得到的准确率和召回率都比较高。

表 3 不同模型性能对比

方法	预测准确率	召回率	*F*measure
LDA-based method[9]	0.642	0.870	0.739
Graph-follower method	0.902	0.54	0.690
本文模型 r=0, d=0.5	0.843	0.860	0.851
本文模型 r=1, d=0.5	0.839	0.855	0.850
本文模型 r=0.5, d=0.5	**0.847**	**0.863**	**0.855**

6 结束语

本文提出了一种新的混合朋友推荐方法，该方法突破性的同时使用用户的消息和朋友链接来预测新的朋友。提出一种新的 PageRank 变种算法—TFPR，该算法满足 PageRank 算法的基本性质，同时融合了用户内容信息和链接拓扑信息，是一种统一的算法框架。实验证明从消息中提取的用户兴趣与用户的朋友链接一起能提高预测的性能。

在社交网络中，用户的兴趣不断变化。如何捕捉兴趣的动态变化过程，使用最新的兴趣来结合朋友关系是未来的工作重点。

参考文献：

[1] NEWMAN M E J. The structure of scientific collaborationnetworks[A]. Proceedings of the National Academy of Sciences[C]. USA, 2001.404-409.

[2] LEO K. A new status index derived from sociometricanalysis[J]. Psychometrika,1953,18(1):39-43.

[3] GLEN J, JENNIFER W. SimRank: a measure of structural-context similarity[A]. Proceedings of the ACMSIGKDD International Conference on Knowledge Discoveryand Data Mining[C]. 2002.538-543.

[4] LADA A A, EYTAN A. Friends and neighbors on the web[J]. Social Networks, 2003, 25(3): 211-230.

[5] JORN D, HOLGER E, STEFAN B. Emergence of a small world from local interactions: modeling acquaintance networks[J]. Physical Review Letters, 2002, 88:128701.

[6] EMILY M J, MICHELLE G, NEWMAN M E J. The structure of growing social networks[J]. Physical Review Letters E, 2001, 64:046132.

[7] NEWMAN M E J. Clustering and preferential attachment in growing networks[J]. Physical Review Letters E, 2001, 64:025102.

[8] GERARD S, MICHAEL J. Introduction to Modern Information Retrieval[M]. McGraw-Hill, 1983.

[9] GRANOVETTER M. The strength of weak ties: a network theoryrevisited[J]. Sociological Theory, 1983, 1:201-233.

[10] MARCO P, SIVA G. Investigating topic models for social media user recommendation[A]. WWW2011[C]. Hyderabad,India,2011.101-102.

[11] KRITI P, JACOB E, SHAY C, *et al*. Social links from latent topics in microblogs[A]. Proceedings of the NAACL HLT 2010 Workshop on-Computational Linguistics in a World of Social Media[C]. Los Angeles, California, 2010.19-20.

[12] YIN Z J, MANISH G, TIM W, *et al*. LINKREC:A unified framework for link recommendation with user attributes and graph structure[A]. WWW 2010[C]. North Carolina, USA, 2000.1211-1212.

[13] ROHIT P, DOINA C. Predicting friendship linksin social networks using a topic modeling approach[A]. PAKDD2011[C]. 2011.75-86.

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

[15] DANIELE Q, HARRY A, JON C. TweetLDA: supervised topic classification and link prediction in Twitter[A]. WebSci2012[C]. Evaston, Illinois USA, 2012.22-24.

[16] LAWRENCE P. The PageRank citation ranking: bring order to Web[J]. Stanford Inforlob, 1998.

[17] LARS B, JURE L. Supervised random walks: predicting and recommending links in social networks[A]. WSDM'11[C]. Hong Kong, China, 2011.9-12.

[18] TAHER H, HAVELIWALA. Topic sensitive PageRank[A]. WWW 2002[C]. Hawaii, USA, 2011.7-11.

作者简介：

尚燕敏（1982-），女，河北定州人，中国科学院博士生，主要研究方向为用户行为挖掘。

张鹏（1981-），男，江西南昌人，博士，中国科学院副研究员，主要研究方向为数据流挖掘。

曹亚男[通信作者]（1985-），女，山东德州人，博士，中国科学院助理研究员，主要研究方向为知识发现。E-mail: caoyanan@iie.ac.cn。