DOI: 10.7652/xjtuxb201608017

面向用户话题相似性特征的链路预测方法

王菲菲[1], 杨扬[2], 蒋飞[2], 许进[2]
(1.北京大学光华管理学院, 100871, 北京; 2.北京大学高可信软件教育部重点实验室, 100871, 北京)

摘要: 针对线上用户间的链路预测对用户文本内容特征的挖掘不够充分的现象,提出了面向用户兴趣话题相似性的二次特征抽取方法。该方法应用主题模型得到任意用户的主题分布,利用用户在主题上相异的分布比例提取各自的兴趣话题集合,基于兴趣话题集合构造了一组话题相似性特征用于链路预测。不同于传统方法中对用户主题分布的直接利用,该方法对用户文本内容的相似性特征进行了再次挖掘,使得文本特征具有等同于结构特征的预测能力,并能够作为结构预测特征的有效补充。实验结果表明,内容特征的独立预测效果普遍优于结构特征,并且在联合预测中将结构特征的预测效果提高了3%。

关键词: 链路预测;用户内容;主题模型;相似性特征
中图分类号: TP391.9 **文献标志码**: A **文章编号**: 0253-987X(2016)08-0103-07

A Link Prediction Method Based on Similarity of User's Topics

WANG Feifei[1], YANG Yang[2], JIANG Fei[2], XU Jin[2]
(1. Guanghua School of Management, Peking University, Beijing 100871, China;
2. Key Lab of High Confidence Software Technologies, Peking University, Beijing 100871, China)

Abstract: A new topical feature extraction method based on similarities of user's topics is proposed to solve the insufficiency of topical feature mining of link predictions in social networks. The topic distributions of social network users are firstly obtained using a topic model and then topic groups of interests for each user are extracted for further similarity-based feature extractions. The proposed topical features exhibit comparable performance of structural features and is efficiently combined with structural features to achieve better results in link predictions. Experimental results based on the dataset collected from Sina Microblog show that independent prediction of topical features is better than that of structural features and the F-measure of structural features is improved by up to 3% with joint predictions.

Keywords: link prediction; user generated content; topic model; feature similarity

社交服务提供商为了保持用户黏性,需培养用户从事在线社交活动的行为习惯,维持其较高的参与度和存在感,一个重要的方法就是通过链路预测来进行推荐,引导用户关注具有共同爱好的线上朋友或发现尚未建立关注关系的线下好友。社交网络中好友间可被观测到的相互关注关系通常被简称为链路。链路预测问题可以概括为:对于目标网络中每一对未观测到存在链路的节点对,预测它们之间是否真的存在链路,或者是否会在将来建立新的链路。

链路预测中一类重要方法是基于节点相似性假

收稿日期: 2016-03-12。 作者简介: 王菲菲(1988—),女,博士生;许进(通信作者),男,教授,博士生导师。 基金项目: 国家重点基础研究发展计划资助项目(2013cb329600);国家自然科学基金资助项目(61372191,61572492,61472433)。
网络出版时间: 2016-06-27 网络出版地址: http://www.cnki.net/kcms/detail/61.1069.T.20160627.1231.008.html

设进行预测,即相似性越强的节点对之间存在链路的概率越高。早期研究中对于节点相似性的刻画,主要体现在提取有效的结构特征来刻画节点对的相似性,例如以社交行为为背景的共同好友数[1]、局部结构相似性和社团贡献度[2]等特征。这类研究普遍具有数据采集便捷、特征计算复杂度低、解释性强和精度良好等优势,并常在后续工作中作为预测效果的评价基准。在近期研究中,人们开始引入多元和异质的信息来提高原有结构信息的预测精度,即除了原有结构信息外,还引入用户在同一社交网络中产生的文本内容信息、地理位置信息或在其他平行社交网络中产生的结构、文本、地理位置信息等。

在社交网络所包含的多元异质信息中,文本信息占据着主导地位,其大量形成于网络中的节点,并时刻借助网络中的链路进行传播。相比于结构特征对用户关注关系直接刻画,文本特征从信息传播角度间接体现了用户间潜在的好友关系。对于文本信息的充分挖掘能够有效改善结构特征的预测效果,并体现在基于文本相似性假设的链路预测[3]等研究工作中。

基于文本相似性假设的链路预测是随着文本建模研究的发展而出现的。文本模型经历了由 TF-IDF(text frequency-inverse document frequency)模型到 LDA (latent dirichlet allocation)模型[4]的重要发展阶段,其中 LDA 模型是一类非常重要的概率主题模型。它在传统的空间向量模型和语言模型的基础上引入了主题空间,将每篇文档看成是在主题空间上的概率分布,将每个主题看成是在整个词典空间上的概率分布,从而构建了“文档-主题-词典”的多层次概率模型。该模型具有较好的泛化能力和较强的可扩展性,因而被广泛地应用于包括社交网络文本内容建模在内的研究工作中。

在社交网络的链路预测中,对于用户的文本内容的利用主要通过以下两种方式:

(1)对用户的内容进行文本建模,并将得到的内容特征应用于预测模型,或对初步得到的内容特征进行进一步抽取,用新得到的用户内容特征进行预测;

(2)扩展原有的文本模型,对用户内容和网络结构进行联合建模,并通过模型拟合后的相应参数直接或间接推导出链路的存在概率。

在第一种方式中,Puniyani 等在回归模型中将每对文档在每个主题上分布概率之差的平方值作为自变量,研究了每个特定主题与文档间链路的存在性是否显著相关[5]。相似地,Quercia 等则证明了改进的监督型主题模型 Labeled-LDA 能够取得与 LDA 相近的用户相似性评价效果[3]。

第二种方式中建立的联合模型是对 LDA 主题模型的进一步扩展,能够建模网络中所有文本与结构信息的生成过程。Erosheva 等提出的 Link-LDA 模型假设一篇文档中的每一个词以及该文档的引用连接服从同一个主题分布,并且由该引用连接的主题归属来分配被引文档的编号[6],从而模拟了文档间的引用与被引用的关系以及文档内容的生成过程,然而该模型并不服从文本相似性假设。Ramage 等提出了 Link-PLSA-LDA 模型[7]。该模型对 Link-LDA 模型进行了扩展,将 PLSA(probabilistic latent semantic analysis)模型和 LDA 模型统一在一个模型框架中,并显式地刻画了主引文档和被引文档之间的拓扑结构关系,即文档链路关系,进而经过模型推导,给出了主引文档与被引文档之间主题相关性的条件概率,并通过该条件概率取值来进行链路预测。Chang 等提出了相关性主题模型,从而对文档网络的内容和链路的生成过程进行建模[8]。该模型用一个二元指示变量表示一对文档之间链路的存在性,并且假定该二元变量与其所对应文档的主题相关性系数具有逻辑回归关系。

本文属于第一种方式,研究了基于 LDA 主题模型提取的用户内容特征,在社交网络中进行链路预测的效果。本文首先应用 LDA 主题模型对用户微博内容进行主题建模,得到了每个用户内容的主题分布,进而创新性地提出了直接内容特征和间接内容特征两类相似性指标。其中在直接内容特征中,提出将用户主题熵之间的差异性与主题分布的相似性进行联合,来评价不同用户在话题爱好上的差异性;在间接内容特征中,利用用户内容在主题分布上的差异性,设定不同阈值筛选出任意一个用户的话题优先性集合,进而基于不同用户的话题集合,提出一组相似性指标来评价他们的话题集合相似性。对于以上所有内容特征,首先分别将其与常见结构特征的独立预测效果进行了对比,证实了直接内容特征和间接内容特征在链路预测中的有效性;接着,在不同监督型算法中,对不同特征集合的预测效果进行了比较。实验结果表明,内容特征集合能够独立地取得较高的预测准确率,并且能够有效提高传统的基于结构特征的方法的预测效果。

本文的贡献主要表现在:

(1)对 LDA 文本建模得到的用户主题分布进

行二次特征抽取，得到了直接内容特征和间接内容特征两类内容特征；

(2)通过充分的实验，验证了间接内容特征在独立预测中优于结构特征，并能在联合预测中改善结构特征的预测效果。

1　数据采集和预处理

1.1　数据采集

本文所用的原始数据采集自新浪微博，采集时间段为 2014 年 8 月 31 日 8 时至 2014 年 8 月 31 日 18 时，其中包含用户 54 119 人，有向链接 754 149 条，以及所有用户在近一周内原创和转发的微博文本内容共 1 114 693 条。

1.2　数据预处理

由于本文关注的是无向网络中的链路预测问题，因此需要把有向链路转换成无向链路，并去除在此过程中产生的重边。最终得到的网络具有无向链路 621 875 条，其中用户节点的最大度数为 3 219，最小度数为 4。

本文所用到的文本内容数据包括用户原创和转发的微博，每条长度不超过 140 个中文汉字。在采集到的数据中，绝大多数用户发表的微博数都在 20 到 60 条之间，占总人数的 78.3%。为了取得更好的主题建模效果，将属于同一个用户的所有微博内容进行拼接，以避免在主题建模的过程中由于文本稀疏性而引发的词频共现性较差等问题，因而最终得到了对应于每个用户的 54 119 条长度不同的文本内容。

对于每条长文本，首先去掉微博内容中的标点符号、特殊字符以及外文字符，进而去除停用词并进行中文分词。本文所使用的分词工具是北京理工大学发布的 NLPIR 中文文本分词系统[9]。

2　预测特征

本节中分别给出本文在进行链路预测时用到的 9 个基准结构预测特征、2 个由用户主题分布直接计算得到的内容相似性特征以及 6 个由对主题分布进行进一步挖掘得到的用户内容特征。

2.1　结构预测特征

结构预测特征是基于网络结构数据计算得到的用于表示社交网络中节点对的结构相似性。本文用做基准的结构特征包括 CN、LHN、Salton、Sonrensen、Jaccard、Adamic-Adar(AA)、PA、HPI 和 HDI 指标[10]，分别设为 S_1、S_2、S_3、S_4、S_5、S_6、S_7、S_8 和 S_9。

2.2　直接内容特征

文本内容特征的提出是建立在相邻用户具有更高内容相似性的假设之上的。对于数据预处理之后得到的中文文本分词结果，应用 LDA 主题模型进行文本建模，得到每一个用户的主题分布，主题个数为 100。

基于得到的用户主题分布，本文给出 2 个直接计算得到的相似性度量指标：

(1)主题相关性系数 R_{xy}。相关性系数能够度量两个变量之间线性相关关系的密切程度，因此本文通过计算不同用户的微博文本内容在主题分布上的相关性系数，来度量用户之间的主题相似程度。已知用户 x 和用户 y 的文本内容在 n 个主题上的分布 θ_x 和 θ_y，则相关性系数的计算方法为

$$R_{xy}=\frac{\sum_{i=1}^{n}(\theta_{xi}-\overline{\theta_x})(\theta_{yi}-\overline{\theta_y})}{\left(\sum_{i=1}^{n}(\theta_{xi}-\overline{\theta_x})^2\cdot\sum_{i=1}^{n}(\theta_{yi}-\overline{\theta_y})^2\right)^{1/2}} \tag{1}$$

式中：θ_{xi} 表示用户 x 的内容在第 i 个主题上的概率，$\overline{\theta_x}$ 表示 θ_x 的均值；θ_y、$\overline{\theta_y}$ 的含义与此类似。如果一对用户具有更多的共同爱好，则他们在各个主题上的概率应该越接近，相关性系数也会越高。反之，如果一对用户之间缺少共同语言，则他们的相关性系数较低。

(2)主题熵 H_x。在本文中，熵被用于度量用户主题分布的无序性，对于用户 x 的内容在 n 个主题上的分布，其主题熵的计算公式为

$$H_x=-\sum_{i=1}^{n}\theta_{xi}\log\theta_{xi} \tag{2}$$

当用户内容在主题上不是服从均匀分布，而是集中分布在一个或少数几个主题上时，用户的主题熵较小，这反映了用户爱好比较集中的现象；相反，当用户的主题分布较为均匀，没有明显的倾向性时，用户的主题熵较大，这说明了用户的爱好比较广泛。为了评价任意一对用户 x 和 y 在兴趣爱好的集中程度上的差异性，本文采用主题熵之差的绝对值作为度量指标，其计算公式为

$$D_{xy}=\mathrm{abs}(H_x-H_y) \tag{3}$$

如果该指标的取值较小，则说明用户 x 和 y 的兴趣爱好的集中程度比较接近，反之则说明集中程度的差异性较大。

2.3　间接内容特征

由于用户微博内容的主题是分布在给定的 100

个主题上的，因而对每一个用户的主题分布概率进行降序排列，可以得到该用户微博内容的主题优先性排序。按比例依序选取概率之和不大于阈值 $p(0<p<1)$ 的所有主题，可以为每个用户 x 分配一个新的优先性主题集合 T_x，在本文中称为用户 x 的兴趣主题集合。该集合是用户主题集合的子集，反映了用户兴趣爱好集中在部分主题上的特性。用 C_k^{T} 表示话题 k 出现在所有用户的兴趣主题集合中的计数，该指标刻画了话题 k 在所有用户中的受欢迎程度。进而可以得到一组如表 1 所示的间接内容特征。

(1)共同兴趣主题数设为 F_1。对于用户 x 和 y，该特征是他们主题集合 T_x 和 T_y 的交集，反映了两人兴趣主题集合的重叠数目。具有越多共同话题的用户之间的共同兴趣主题数越高。

(2)共同兴趣主题占比设为 F_2。对于用户 x 和 y，该特征是他们兴趣话题集合 T_x 和 T_y 的交集与并集之比，反映了共同兴趣主题占两人所有爱好主题的比例。若两人具有较多的共同兴趣主题数和较少的总爱好主题，则该特征取值较高。该特征刻画了两人兴趣主题集合的重叠程度。

(3)以主题分布概率赋权的共同兴趣主题数和共同兴趣主题占比分别设为 F_3 和 F_4。

(4)共同兴趣主题数和共同兴趣主题熵的 Adamic-Adar 测度设为 F_5 和 F_6。它们分别是对用户 x 和 y 的共同兴趣主题数的倒数和主题数的倒数熵进行求和。若二人的所有共同主题较少，并且每一个共同主题也较少成为其他用户的爱好主题，即 C_k^{T} 取值较小，则 F_5 取值较高，这反映了二人共同兴趣的特殊性较强；同理，F_6 从熵的角度反映了二人兴趣爱好的特殊性。

3 实验对象和算法

本文所采用的数据集为 2014 年 8 月 31 日当天采集得到的静态数据，通过按相同比例随机抽取该数据集中存在链接与不存在链接的节点对，构造出无重复的训练数据与测试数据。

由于实验是判断给定的静态网络中任意两点之间的链路是否存在，因此用 1 表示链路存在，用 0 表示链路不存在，可以将该问题转化为二分类问题。在此基础上，在监督学习框架下对结构特征、直接内容特征和间接内容特征的预测能力进行评价。首先通过每个特征的独立预测效果对结构特征和内容特征的预测能力进行比较，进而在多个分类器中对这些特征的整体预测效果进行评价。

表 1 内容特征

预测特征	计算公式
F_1	$\lvert T_x \cap T_y \rvert$
F_2	$\lvert T_x \cap T_y \rvert / \lvert T_x \cup T_y \rvert$
F_3	$\sum_{t_k \in T_x \cap T_y} \theta_{xk}\theta_{yk}$
F_4	$\dfrac{\sum_{t_k \in T_x \cap T_y} \theta_{xk}\theta_{yk}}{(\sum_{t_k \in T_x \cap T_y} \theta_{xk}^2)^{-1/2} \times (\sum_{t_k \in T_x \cap T_y} \theta_{yk}^2)^{-1/2}}$
F_5	$\sum_{t_k \in T_x \cap T_y} 1/\log C_k^{\mathrm{T}}$
F_6	$\sum_{t_k \in T_x \cap T_y} \dfrac{1}{E_k^{\mathrm{T}}}, E_k^{\mathrm{T}} = -\dfrac{1}{C_k^{\mathrm{T}}} \log \dfrac{1}{C_k^{\mathrm{T}}}$

4 实验评价

本节在监督学习框架下对提出的预测特征进行评价，以验证本文所提内容特征在链路预测应用中的有效性。

4.1 独立特征评价

分别应用 9 个基准结构特征、2 个直接内容特征和 6 个间接内容特征在新浪微博数据集上进行链路预测，并采用基于抽样方法实现的近似 AUC 得分进行效果评价[11]。

对于每个特征，在新浪微博数据中进行 30 次随机抽样，每次抽取 100 000 组样本，并计算其 AUC 得分，其中结构特征和直接内容特征如图 1 所示，间接内容特征分别在 p 值为 0.2、0.4、0.6 和 0.8 时进行计算，结果如图 2 所示。

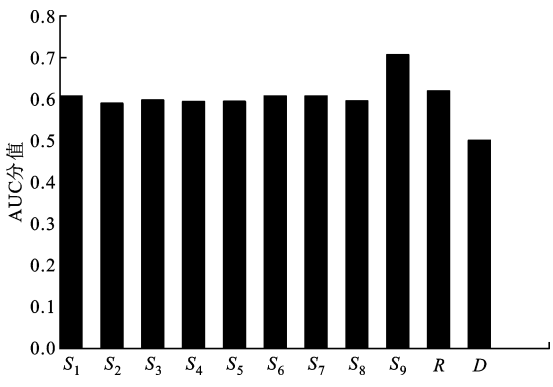

图 1 结构特征和直接内容特征的 AUC 得分

由图 1、图 2 可以看出，在独立特征预测中，PA 指标取得了结构特征中的最高得分。这说明该数据集中的链路通常存在于两个度数较高的节点之间，即新浪微博用户表现出较强的择优连接特性。相应

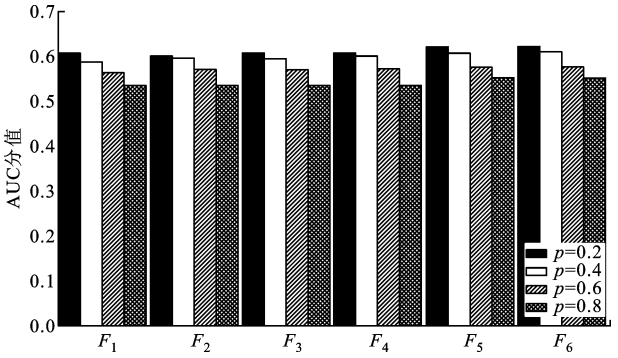

图 2　间接内容特征的 AUC 得分

地，其他与 PA 指数存在负相关的指标预测效果相对较差。在直接内容特征中，主题相关性系数 R_{xy} 取得了最好的预测效果，而间接内容特征中的 F_5 测度取得了更好的预测效果。

从整体上来看，内容特征取得了与结构特征相近的独立预测效果，并且大多数内容特征的单独预测能力都优于结构特征。间接内容特征在 p 值较小(p=0.2)时能够取得最好的预测结果，而随着 p 值增大，间接内容特征的预测效果随之下降。直接内容特征中的主题熵差绝对值 D_{xy} 取得了最差的预测效果，几乎与随机预测器效果相同。

4.2　监督预测评价

本节在监督学习的框架下，给出了一个同时基于内容特征和结构特征的链路预测模型，以评价所提出特征的预测效果。首先利用逻辑回归(LR)分类模型，对各个内容特征与链接存在与否的相关关系进行了检验；然后通过支撑向量机(SVM)、随机森林(RF)、决策树(DT)和朴素贝叶斯(NB)4 种常用分类器进一步验证同时基于内容特征和结构特征的链路预测模型的有效性。

逻辑回归分类器的优势在于不但能够完成分类工作，还能够在建模后检验自变量是否与因变量显著相关，并且自变量对应的回归系数可以显示自变量与因变量的相关关系是正相关还是负相关。本文用逻辑回归分类器判断基于内容特征的各项指标是否具有较强的链路预测能力。

由于分类问题可看成是二分类问题，因此在逻辑回归模型中，将节点对之间的链路是否存在作为因变量 Y(取值为 0 或 1)；将结构特征、直接内容特征和间接内容特征作为自变量，分别用 $X_1, X_2, \cdots, X_{17}$ 表示。设在 n 个自变量作用下链路存在的条件概率为 $P=P(Y=1\mid X_1, X_2, \cdots, X_n)$，则可将 LR 模型表示为

$$z_1 = a_{-1} + a_1 X_{i1} + a_2 X_{i2} + \cdots + a_{17} X_{i17} + \varepsilon_i$$

$$P_i = 1/(1+\mathrm{e}^{z_i})$$

式中：a_{-1} 和 a_j 分别表示回归常数和第 j 个自变量的回归系数；ε_i 为随机误差项。本文对数据集中可观测到的链路和未观测到的链路进行了 10 组随机抽样，每组的抽样规模为 100 000，并保证正负样本的比例相同。将其中的任意一组数据用于进行逻辑回归分类器的训练，就能够得到与每个自变量一一对应的回归系数及其显著性指标，下面根据实验数据进行详细说明。

表 2 中列出了其中一组逻辑回归分类器的训练结果。标准误差可用于评价回归系数是否显著不为 0；z 值是由回归系数与标准误差之比得到的，与 p 一起用于检验回归系数为 0 的零假设。一般来说，取显著性水平 α 为 0.05，当 p 小于 α 时，说明回归系数显著不为 0，即自变量与因变量显著相关。由表 2 可以看出，只有 D_{xy} 的 p 大于显著性水平 α($\alpha<0.166$)，因此除了直接预测特征中的主题熵之外，其他预测特征对链路存在与否均具有显著影响。

表 2　逻辑回归结果

预测特征	回归系数	标准误差	Z	p
S_1	−3.870	0.317	−12.191	0.000
S_2	−1 659.261	440.902	−3.763	0.000
S_3	−56.927	9.863	−5.772	0.000
S_4	661.698	89.063	7.430	0.000
S_5	−178.818	57.475	−3.111	**0.002**
S_6	117.643	24.558	4.790	0.000
S_7	−1 074.456	216.838	−4.955	0.000
S_8	0.001	2.48×10^5	28.410	0.000
S_9	16.353	1.591	10.279	0.000
R_{xy}	−1.009	0.072	−14.011	0.000
D_{xy}	−1.552	1.119	−1.387	**0.166**
F_1	−36.207	1.212	−29.868	0.000
F_2	9.282	0.669	13.878	0.000
F_3	−65.626	5.503	−11.926	0.000
F_4	82.227	7.180	11.452	0.000
F_5	-1.21×10^6	4.07×10^4	−29.748	0.000
F_6	1.71×10^5	5 725.000	29.803	0.000

在所有 10 组实验中，结构相似性指标一致地表现出显著性特征，其中 F_5 在少数几次实验中不具有显著性；直接内容特征中，主题相关性系数 R_{xy} 均表现出显著性，而主题熵之差的绝对值 D_{xy} 均不具

有显著性;在间接内容特征中,所有自变量均一致地表现出显著性。实验结果表明,在逻辑回归分类器中,除主题熵特征外,其余文本内容特征均能够有效地影响链路预测结果。

接下来,在支撑向量机(SVM)、随机森林(RF)、决策树(DT)和朴素贝叶斯(NB)4 个分类器中对模型的预测效果进行进一步测试。实验中训练数据与测试数据的规模和抽样方法与逻辑回归分类器相同。在每个分类器中分别对结构特征、内容特征以及全部特征进行了测试,内容特征包括直接内容特征与间接内容特征,其中间接内容特征分别在 p 为 0.2、0.4、0.6 和 0.8 时进行抽取。所有的参数取值都是对 10 次重复实验结果求得的平均值。模型的分类效果如图 3、图 4 和表 3 所示。

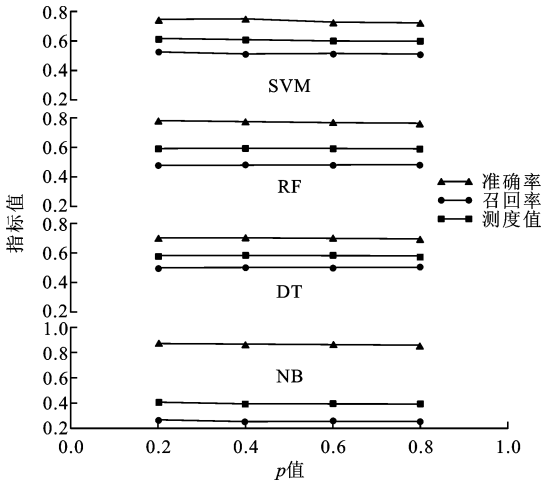

图 3 全部特征的分类效果

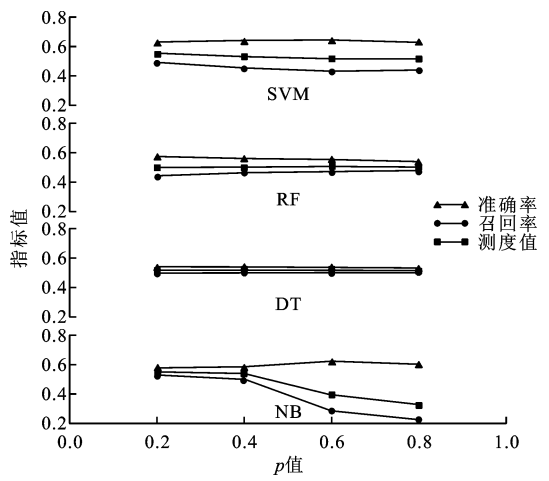

图 4 内容特征的分类效果

对比各个分类器的 F-测度可以看出,SVM 分类器在 3 组特征下都取得了最好的预测效果,其次分别为随机森林分类器、决策树分类器和朴素贝叶斯分类器。对比内容特征和结构特征各自的预测结果,可以看出内容特征具有接近结构特征的预测能力,其中对于 SVM 分类器、随机森林分类器和决策树分类器,内容特征的预测效果稍差于内容特征,而对于朴素贝叶斯分类器,应用内容特征分类的召回率和 F-测度明显高于结构特征的相应值。对比全部特征和结构特征的预测效果,可以看出内容特征分别将传统的结构特征预测效果(F-测度)提高了 1.2%至 3%,是对结构预测特征的有效补充。

表 3 结构特征的分类效果

分类器	准确率	召回率	F-测度
SVM	0.741	0.591	0.491
DF	0.762	0.578	0.465
DT	0.767	0.563	0.451
NB	0.874	0.372	0.231

实验中,内容特征和全部特征的预测效果均随着 p 的增大而保持稳定或表现出不明显的下降规律,但在 p 为 0.2 时均取得了最好的预测效果,这说明当主题占比 p 较小时,通过用户的兴趣主题集合得到的内容特征能够更加有效地区分存在链路与不存在链路的用户组合。

以上实验结果表明,在监督学习框架下,内容特征取得了与传统结构特征相近的预测效果,并且将两者结合在一起能够取得更好的链路预测效果。

5 结 论

在链路预测研究中,本文基于在线社交网络中的节点具有丰富的内容信息这一特点,提出了利用文本内容进行链路预测的研究思路。对于用户的文本内容,本文首先采用 LDA 主题模型进行文本建模,得到了所有用户内容在各个不同主题上的分布,进而提取出主题相关性系数和主题熵两个直接内容特征和其他间接内容特征。接着,通过独立特征预测,将两类内容特征与常用结构特征的预测效果进行了对比,发现除结构特征中的 PA 指标外,内容特征的独立预测能力普遍高于结构特征。最后,结合结构特征、直接和间接内容特征给出了一个监督学习框架下的链路预测模型。实验结果显示,通过准确率、召回率和测度值进行评价,内容特征具有接近结构特征的预测能力,并且将两者进行结合能够取得更好的预测效果。

在实际应用中,常需要对用户间链路的存在概率进行估计,而非简单判断链路是否存在,尤其是进行保守预测时。这时,可利用能够给出概率估值的

分类器或排序算法，例如本文所用的逻辑回归分类器能够给出样本点的概率估值，Ranking SVM[13]算法能够对样本进行排序。在得到概率估值或样本排序后，概率值更大或序数更小的样本（未观测到存在链路的用户组合）更可能被预测为存在链路。

本文的工作可以在模型参数优化和内容特征挖掘等方面进行改进。例如：通过优化主题模型在短文本上的建模效果，可获得更好的主题含义，进而得到解释性更强的用户内容特征；挖掘用户微博的时序性和频次性等信息，从而可提取更为丰富且合理的内容相似性特征等。

参考文献：

[1] 吕琳媛. 复杂网络链路预测 [J]. 电子科技大学学报，2010，39(5)：651-661.
LÜ Lin-yuan. Link prediction on complex networks [J]. Journal of University of Electronic Science and Technology of China，2010，39(5)：651-661.

[2] 伍杰华. 基于划分社区和差分共邻节点贡献的链路预测 [J]. 计算机应用研究，2013，30(10)：2954-2957.
WU Jiehua. Link prediction based on partitioning community and differentiating role of common neighbors [J]. Application Research of Computers，2013，30(10)：2954-2957.

[3] QUERCIA D，ASKHAM H，CROWCROFT J. TweetLDA：supervised topic classification and link prediction in Twitter [C]// Proceedings of the 4th Annual ACM Web Science Conference. New York，USA：ACM，2012：247-250.

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

[5] PUNIYANI K，EISENSTEIN J，COHEN S，et al. Social links from latent topics in Microblogs [C]// Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media. New York，USA：ACM，2010：19-20.

[6] EROSHEVA E，FIENBERG S，LAFFERTY J. Mixed-membership models of scientific publications [J]. Proceedings of the National Academy of Sciences，2004，101(S1)：5220-5227.

[7] RAMAGE D，HALL D，NALLAPATI R，et al. Labeled LDA：a supervised topic model for credit attribution in multi-labeled corpora [C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. New York，USA：ACM，2009：248-256.

[8] CHANG J，BLEI D M. Relational topic models for document networks [J]. Journal of Machine Learning Research，2009，9：81-88.

[9] ZHANG Huaping，YU Hongkui，XIONG Deyi，et al. HHMM-based Chinese lexical analyzer ICTCLAS [C]// Proceedings of the Second SIGHAN Workshop on Chinese Language Processing. New York，USA：ACM，2003：184-187.

[10] DONG Yuxiao，KE Qing，WANG Bai，et al. Link prediction based on local information [C]// Proceedings of the 2012 IEEE International Conference on Advances in Social Networks Analysis and Mining. Piscataway，NJ，USA：IEEE Computer Society，2011：382-386.

[11] 吕琳媛，周涛. 链路预测 [M]. 北京：高等教育出版社，2013：82-87.

[12] KOSSINETS G. Effects of missing data in social networks [J]. Social Networks，2006，28(3)：247-268.

[13] JOACHIMS T. Optimizing search engines using click through data [C]// Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York，USA：ACM，2002：133-142.

[本刊相关文献链接]

刘兆丽，秦涛，管晓宏，等. 采用用户名相似度传播模型的线上用户身份属性关联方法. 2016，50(4)：1-6. [doi：10.7652/xjtuxb201604001]

孟宪佳，马建峰，王一川，等. 面向社交网络中多背景的信任评估模型. 2015，49(4)：73-77. [doi：10.7652/xjtuxb201504012]

叶娜，赵银亮，边根庆，等. 模式无关的社交网络用户识别算法. 2013，47(12)：19-25. [doi：10.7652/xjtuxb201312004]

张赛，徐恪，李海涛，等. 微博类社交网络中信息传播的测量与分析. 2013，47(2)：124-130. [doi：10.7652/xjtuxb201302021]

孙艳，周学广，付伟. 无监督的主题情感混合模型研究. 2013，47(1)：120-125. [doi：10.7652/xjtuxb201301023]

莫同，褚伟杰，李伟平，等. 采用超图的微博群落感知方法. 2012，46(11)：120-126. [doi：10.7652/xjtuxb201211022]

（编辑　武红江）