计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(12):227−233 [doi: 10.15888/j.cnki.csa.006662]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于地理标签的 LBSN 链接预测模型①

王　勇[1], 王　超[2], 程　凯[3]

[1](铜陵有色金属集团股份有限公司金冠铜业分公司, 铜陵 244000)
[2](金诚信矿业管理股份有限公司, 北京 100044)
[3](北京宸控科技有限公司, 北京 102200)
通讯作者: 程　凯, E-mail: chengkai@bjkcst.com

摘　要: 为更深入挖掘用户位置信息, 本文从位置语义相似性角度挖掘用户特征. 利用 LDA 算法对用户签到信息进行位置主题建模, 采用 Gibbs 采样算法计算 LDA 模型中的分布函数, 并根据这些分布提出了基于签到地点语义的用户相似性特征向量. 利用有监督的机器学习算法,综合 LBSN 的网络结构信息、签到地点信息、地点语义信息得到多维相似性特征向量来进行链接预测. 在 Gowalla 数据集上的实验结果表明, 相较于传统的链接预测算法, 将基于签到信息的多个相似性特征作为辅助信息的链接预测算法显著提高了 LBSN 链接预测的性能.

关键词: 链接预测; 机器学习; 主题模型; 地理标签; 数据挖掘

引用格式: 王勇,王超,程凯.基于地理标签的 LBSN 链接预测模型.计算机系统应用,2018,27(12):227–233. http://www.c-s-a.org.cn/1003-3254/6662.html

LBSN Link Prediction Model Based on Geographic Tag

WANG Yong[1], WANG Chao[2], CHENG Kai[3]

[1](Gold Crown Copper Branch, Tongling Nonferrous Metal Group Co. Ltd., Tongling 244000, China)
[2](JCHX Mining Management Co. Ltd., Beijing 100044, China)
[3](Beijing KingKong Science & Technology Co. Ltd., Beijing 102200, China)

Abstract: In order to find out more location information from users, the user feature is dug out from location semantic similarity. The location topic model is built for user sign-in information by using LDA algorithm, and distribution functions can be calculated by the Gibbs sampling algorithm. The user similarity feature vector based on sign-in location semantic is put forward by these distribution functions. Then, the supervised machine learning algorithm is put forward to make link prediction by multi-dimensional similarity feature vector from fusing LBSN network structure information, sign-in location information and location semantic similarity. The experiments result on Gowalla databases shows that the link prediction algorithm using more similarity feature as subsidiary information can improve performance of LBSN link prediction significantly comparing with the traditional algorithm.

Key words: link prediction; machine learning; topic model; geographic tag; data mining

随着智能手机的爆炸式发展, 用户可随时随地在各大社交平台分享自己的地理位置信息. 无论是视频、图片或文本信息, 都可轻易地嵌入当前用户的地理位置标签 (Location Tag). 大量的位置 Tag 构成了基于地理位置的社交网络[1](Location-Based Social Networks, LBSN), 结合现有的各种定位系统, 可以为用户提供一些个性化服务. 根据 LSBN 中已经存在的链接和节点信息, 可以预测出用户位置网络中遗失的 Tag 或即将出现的 Tag 链接, 该方法称之为链接预测[2]. 例如, 在微博、微信等社交平

① 收稿时间: 2018-04-22; 修改时间: 2018-05-14; 采用时间: 2018-05-24; csa 在线出版时间: 2018-12-03

台中, 用户等同于节点, 链接预测可用于建立新的好友关系.

文献[3]表明, 同一时间出现在同一位置的用户成为好友的概率要远高于处于不同地理位置的用户. 因此, 挖掘 LBSN 中潜在的标签信息对实现链接预测具有重大意义. 目前, 国内外已有很多科研工作者专注于基于地理位置标签的推荐算法研究, 判断用户地理位置的途径主要有两种: 第一, 挖掘用户发布到互联网中的内容信息可推断出用户的地理位置信息[4]. 第二, 通过社交平台中好友的地理位置推测用户的位置[5]. 近年来也有很多学者研究基于 LDA 主题建模的层次聚类[6]、无监督学习[7]、标签关联[8]推荐算法. 为提高位置预测的准确性, 可对用户位置信息进行筛选, 过滤掉无用的信息. 还可对用户签到信息建立 LDA 主题生成模型, 分析地理位置标签的特征, 设计出基于地理位置的推荐系统[9].

从用户签到信息中提取出时间特征和位置特征对于链接预测算法至关重要, 因为这些特征可用于评估用户之间的相似度, 进而提高预测的准确度. 然而实际的 LBSN 签到信息中, 地理位置的分布十分稀疏, 想要挖掘出位置和时间信息相当困难. 基于用户地理位置标签, 本文建立了新的 LBSN 链接预测模型, 提高了链接预测的准确度. 首先, 本文对 Gowalla 数据集进行聚类分析, 改善了地理位置分布的稀疏性问题. 其次, 本文对用户地理位置标签进行语义分析, 建立基于用户地理标签的 LDA 主题模型, 采用 Gibbs 抽样算法进行参数估算, 分析出用户的地理位置标签的相似性特征. 最后, 本文综合网络结构相似性特征和基于用户地理位置信息的相似度特征, 采用有监督策略的链接预测在 Gowalla 数据集上进行实验. 实验结果表明, 本文提出的模型能有效提高 LBSN 的链接预测准确度.

1 基于地理位置的 LDA 主题模型

本文对 LBSN 中的用户地理位置标签建立 LDA 主题模型, 以便挖掘出用户的行为偏好. 用户的位置标签集合可当作一篇文档, 位置标签集合中的某条具体位置相当于构成文档的词汇. 对该主题模型进行求解, 可得出用户地理位置标签中隐藏的主题分布和地理标签主题下的位置分布.

假定用户 u 对应的地理位置标签集合为 $\vec{D}_u=[w_{u,1},w_{u,2},\cdots,w_{u,m}]^{\mathrm{T}}$, $\vec{D}_u$ 相当于一篇文档, 其中 m 代表用户 u 的位置标签条数, $w_{u,i}$ 代表用户 u 的第 $i(1\leqslant i\leqslant m)$ 条位置标签信息, 相当于构成文档 $\vec{D}_u$ 的某个词汇. 地理位置文档集合为 $D=[\vec{D}_1,\vec{D}_2,\cdots,\vec{D}_M]^{\mathrm{T}}$, 其中 M 代表用户数量. 假定具体的位置数量为 V, 则可建立基于地理位置的 LDA 主题模型, 如图 1 所示.

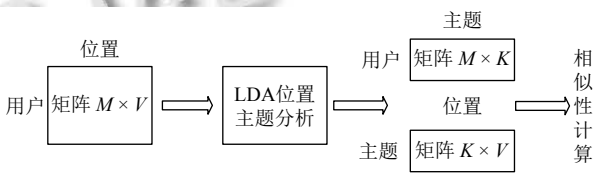

图 1 基于地理位置的 LDA 主题模型

模型中的位置主题概率分布可用 Gibbs[10]采样算法估算得出, 该分布可用一个 doc-topic 矩阵 $\Theta=[\vec{\theta}_1,\cdots,\vec{\theta}_M]^{\mathrm{T}}$ 来描述, 用户 u 在地点主题 t_k 下出现的概率分布可表示为 $\vec{\theta}_u=[p_u(t_1),\cdots,p_u(t_K)]_u=[p_u(t_1),\cdots,p_u(t_K)]$. 同理, 每个位置主题下对应的位置概率分布可用一个 topic-word 矩阵 $\vec{\varphi}_k=[p_k(\varphi_1),p_k(\varphi_2),\cdots,\mathrm{p}_k(\varphi_V)]$ 来描述, 其中 $\mathrm{p}_k(\varphi_v)$ 代表位置 v 在主题 k 下出现的概率.

2 基于位置信息的相似度分析

2.1 位置信息语义化

本文实验使用的数据集是 Gowalla 的地理位置标签数据集, 可从 SNAP 官网直接下载得到, 用户地理位置标签的存储格式为:

$$Checkin_{i,j}=\{user\ ID,\ latitude,\ longitude,\ timestamp,\ location\ ID\}$$

$Checkin_{i,j}$ 表示用户 i 的第 j 条位置信息. 此外, 该数据集还记录了用户之间的好友关系. 具体示例见表 1 和表 2.

表 1 地理位置标签格式

用户 ID	签到时间	纬度	经度	位置 ID
96249	2010-08-16T09:42:54Z	58.403 760 216 7	15.580 063 266 7	350417
12121	2010-04-09T13:49:00Z	30.281 982 052 5	–97.740 436 792 4	28193
12121	2010-04-06T23:27:17Z	30.473 056 018 3	–97.801 076 173 8	14627
12121	2010-03-16T19:05:27Z	30.264 460 124 8	–97.738 412 282 2	722978

表 2 用户好友关系存储格式

用户 1 ID	用户 2 ID
90	1900
1910	1920
1930	1940

由于数据集并没有具体的地理语义信息, 故需要先对其进行语义化. 本文采用百度公司免费的 Place API 和 Geocoding API 进行语义转换, 可将数据集中的地理坐标转换为具体的地址以及附近的 POI (Point Of Interest) 信息.

由于很多用户只在某一个地点签到过, 或跟其他用户没有共同的签到地点, 这类用户称之为孤点用户. 大量的孤点用户造成了数据的稀疏性, 严重影响链接预测的准确性. 为解决该问题, 本文降低了具体地点的限制, 对地点标签进行层次聚类, 以签到区域来构建用户关系网络. 设定一个距离阈值δ, 若不同签到地点的距离不超过该值, 则认为两个地点属于一个区域, 本文在实验章节会对该参数进行调优试验. 然后利用该距离阈值对签到数据集进行聚类, 可得到区域集合$D=\{d_1,d_2,\cdots,d_n\}$, 由此可得到区域矩阵:

$$CH=\begin{bmatrix} d_{1,1},\cdots,d_{1,n} \\ d_{2,1},\cdots,d_{2,n} \\ \vdots\ \vdots\ \vdots \\ d_{m,1},\cdots,d_{m,n} \end{bmatrix} \tag{1}$$

式 (1) 中的$d_{i,j}$代表第 i 个用户在区域 j 处的签到. 显然, 利用区域矩阵来构建用户网络关系可极大地减小孤点用户的数量, 对签到地点标签信息的挖掘也更充分, 因此可降低数据稀疏性的影响.

2.2 基于位置信息的相似度分析

上文提到, 采用 Gibbs 采样算法可估算出 LDA 主题模型中的两个概率分布: 位置主题分布Θ和所有主题下的地理位置分布Ψ. 每个用户的签到主题分布可以表示成 K 个位置主题的概率组合, 所有用户的签到主题分布构成矩阵Θ. 每个主题下的位置分布可以表示为签到位置的概率组合, 所有主题下的签到位置分布构成矩阵Ψ. 利用本文的签到语义数据集, Gibbs 采样可输出矩阵Θ和Ψ.

本文首先筛选出位置主题概率最大的 K 个主题来表达用户的位置主题. K 的取值可按照经验预设, 剩下的主题概率可先置 0. 本文先设定 K=5, 在模型学习的过程中会对 K 值进行不断的修正. 然后对这 K 个地理位置主题分布函数进行归一化处理, 如公式 (2) 所示:

$$p'_u(t_k)=\frac{p_u(t_k)}{\sum_{k\in Top-K}p_u(t_k)} \tag{2}$$

其中, $p'_u(t_k)$代表归一化的结果, 即用户u的地理位置出现在主题k下的概率, $p_u(t_k)$代表归一化之前的概率分布函数. 因此能够得到全新的概率分布矩阵Θ'和Ψ'.

对于两个不同用户产生的概率分布函数, 需要计算出二者之间的距离. 统计学中的 KL 散度 (KL-Divergence) 可用于测量不同概率分布的差异, 被广泛应用于基于 LDA 的推荐算法. 然而 KL 散度并不适用于本文基于地理位置标签的链接预测算法, 因为该方法具备非对称性特征. 如果两个用户对某主题都无兴趣, KL 散度得出的结论是这两个用户具有很高的相似性. 同理, 如果两个用户都没有在某个地点签到, 那么 KL 散度会认为他们具有很高的相似度, 这显然会造成极大的误差. 因此, 本文采用一种新的方法来评估用户之间地理位置主题的差异性. 用户 i 在k个地理位置主题下的位置总数设为$N(i,t_k)$, 则不同用户x,y 之间的相似度可用公式 (3) 计算得到:

$$f_W(x,y)=\frac{\sum\limits_{k\in Top-K}\min(N(x,t_k),N(y,t_k))}{\sqrt{\sum\limits_{k\in Top-K}N(x,t_k)}\cdot\sqrt{\sum\limits_{k\in Top-K}N(y,t_k)}} \tag{3}$$

其中, 分子代表用户 x 和 y 在 k 个主题下的位置总数最小值之和, 该值越大, 说明用户 x 和 y 在同一区域签到的数量越大, 二者的相似度越高. 式 (3) 进行了归一化处理, 最终结果可用于计算用户之间的相似度.

最后, 为了验证$f_W(x,y)$的有效性, 基于从当前数据集中获取的网络关系, 本文将对比分析$f'_P(x,y)$和$f_W(x,y)$的性能. LBSN 链接网络中基于$f_W(x,y)$的好友用户对与非好友用户对的累积分布图 (CDF) 如图 2 所示.

从图 2 可看出, 基于位置标签语义分析的用户相似性特征$f_W(x,y)$能够有效地识别好友与非好友的区别. 因此可得出结论, $f_W(x,y)$对于分析用户之间的链接预测具有重要意义.

3 基于地理标签的 LBSN 链接预测模型

3.1 实现方法

本文通过对用户地理位置信息的充分挖掘, 得出了基于地理位置语义分析的相似性特征. 这是本文所提出的链接预测模型的基础. 机器学习中的监督式学

习算法经常被用于推荐系统的设计, 将收集到的海量训练数据集作为先验知识, 建立一个模型, 并根据输入的标签不断修正该模型, 最终该模型可针对新的输入预测出相应结果. 本文的链接预测算法基于有监督学习的思想, 输入为 Gowalla 数据集中的位置标签, 建立用户特征向量函数, 对其进行模型训练, 最终可用于链接预测.

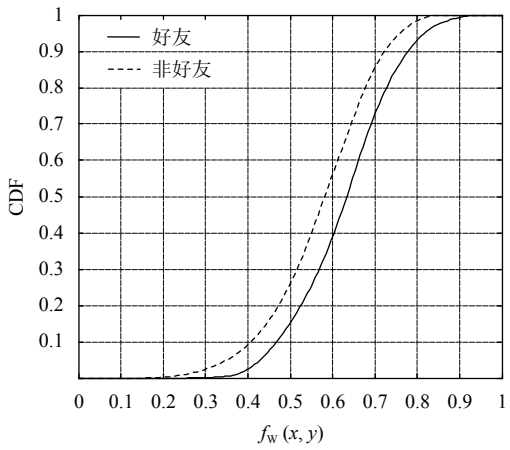

图 2 $f_W(x,y)$好友与非好友用户的 CDF 曲线

接下来, 本文将采用有监督学习的策略对其进行链接预测. 实验中, LBSN 链接预测采用 Gowalla 数据集进行仿真, 使用 LBSN 基于地理位置语义分析的相似性特征进行辅助实验. 本文实施的链接预测实验步骤如下:

(1) 筛选原始数据集, 过滤掉其中无用的冗余信息和独立用户 (即无任何好友关系的用户), 最总得到一个可用的 LBSN 社交关系网络图$G=(V,E)$.

(2) 对集合E进行随机采样, 其中 2/3 的数据作为训练集E^T, 余下 1/3 的链接数据作为测试数据集E^P, 显然$E=E^T+E^P$且$E^T\cap E^P=\emptyset$. 从集合E^T中取出现有链接中所有的用户集合$V'\in V$且$V'\neq V$, 则子网络图$G'=(V',E')$. 由V'可将测试数据集修正为$E^P=E'\cap E^P$, 所以空间集合$(V'\times V')-E^T$可用来表示一切隐含节点对的集合.

(3) 由V'可得出所有用户的位置标签列表, 获取这个列表集合中的地理位置信息, 对其进行聚类, 最终得出一个新的用户-位置矩阵.

(4) 分析求解隐藏的用户节点信息中基于地理位置的相似性特征$f_{UP}(x,y)$.

(5) 同理, 分析求解隐藏的用户节点对基于时间戳的相似性特征$f_T(x,y)$.

(6) 采用 Gibbs 抽样算法估算出用户基于地理位置主题的概率分布函数, 然后进一步求解隐藏用户节点对基于地理位置信息的用户相似度特征$f_W(x,y)$.

(7) 对社交网络中所有隐藏用户节点之间的相似度进行分析计算, 此处主要记录预测性能最佳的 Resource allocation (RA) 系数指标$S_{x,y}^{RA}$.

(8) 利用有监督学习策略的算法, 对上文计算得出的各类相似性特征做链接预测, 最终可得出特征向量$\vec{feature}(x,y)=\left(f_{UP}(x,y),f_T(x,y),f_W(x,y),S_{x,y}^{RA}\right)$, 然后对其进行模型训练, 最后再对测试数据集进行链接预测, 得出基于地理标签的 LBSN 链接预测结果. 最终得到的结果集中, 用 1 标注确实存在的链接节点信息, 0 标注不存在链接的节点信息.

3.2 评估方法

将上文求解得到的结果集与测试数据集做对比, 可分析出本文链接预测算法的性能. 由于本文采用有监督的链接预测算法, 故采用信息检索算法中常用的四大性能评估指标来衡量本文算法的性能优劣: 精度 (Accuracy)、准确率 (Precision)、召回率 (Recall) 以及综合 Accuracy 和 Precision 的加权调和平均 (F-measure).

由于实验数据中链接分布不均匀, 本文还采用了一个新的评估指标 AUC(area under the receive operating characteristic curve)[11], 如公式 (4) 所示:

$$AUC=\frac{n'+0.5n''}{n} \tag{4}$$

其中, n代表测试数据集中所有标签对被随机独立抽样的次数, 对于链接节点对而言, 包含了存在和不存在两种情况. n'表示链接节点对存在时的相似度分数大于不存在时的次数, n''则表示两种情况相似度分数相等的次数. 从上式可看出, 若存在链接的相似度值大于不存在时, 则相似度值加 1, 若相等则加 0.5. 因此, AUC 指标能够整体地评估链接预测模型的准确度, 其值得取值范围是(0.5, 1), AUC 的值越大, 表示链接预测模型的精准度越好.

3.3 参数调优

本文采用有监督学习的方式对样本进行分类学习, 基于前人对的研究, 我们可获取关于类别特征的先验

知识. 基于已有的类别特征信息, 可对模型进行训练并构造相应的分类器. 由于本文采用的是真实的 Gowalla 数据集, 故存在样本数据分布不均匀的情况. 为深入挖掘该数据集中的隐藏信息, 可采用机器学习中常用的 k-折交叉验证 (k-fold cross Validation) 法, 该方法得到的实验结果更加真实. 实验证明, 当 k 值取 10 的时候可得到最佳的实验效果[12], 故本文采取 10 倍交叉验证来评估模型的性能.

上文提到, 本文利用有监督的学习思想对模型中需要输入的参数根据经验预先给出, 然后通过实验对其不断修正. 本文模型中有三个输入参数需要进行调优: 对地理位置信息聚类处理时的距离阈值δ, 基于用户地理位置标签的 LDA 主题模型中的主题 K 的取值, 以及分析用户相似度时$TOP\text{-}K$中 K 的取值.

对距离阈值δ分别取不同的值, 可得到用户基于地理位置标签的相似度特征函数$f_{UP}(x,y)$, 现将该特征函数导入样本分类器进行链接预测, 实验得出的距离阈值δ与加权调和平均 F 值的关系如图 3 所示.

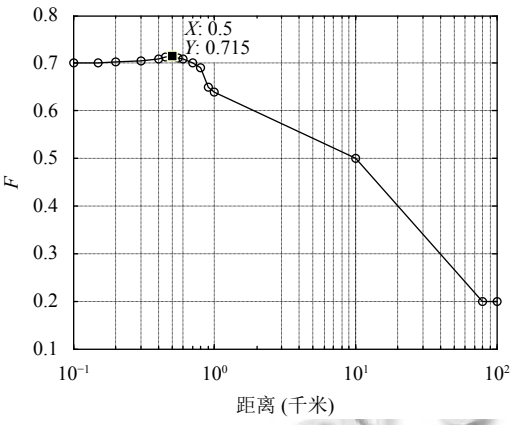

图 3 距离阈值δ对链接预测性能影响曲线

由图 3 可知, 当链接预测距离阈值δ=500 m 时, 链接预测的结果最优. 当$\delta \in [400,700]$时, 该算法的链接预测效果较为良好. 此外, 随着δ的不断增大, 加权调和平均 F 值不断减小, 即距离越大, 链接预测效果越差. 当$\delta > 1km$时, F 值的值显著降低, 说明人与人之间的地理位置距离越远, 两者之间的关系也会越生疏, 该结论显然符合人类社会客观事实. 基于以上分析, 本文对地理位置标签进行层次聚类时的最优距离阈值设定为 500 m.

上文提到, 基于地理位置的 LDA 主题模型可采用 Gibb 采样算法进行分析求解, 最后得出用户地理位置主题分布Θ和每个主题下的具体地点分布Ψ. 根据Θ和Ψ这两个概率分布函数计算出基于地理位置标签信息的相似性特征函数$f_W(x,y)$. Gibbs 采样算法需要预设的参数是 α, β 和K. 其中, α和β是 Dirchlet 先验分布的经验参数, 由于在对数据进行抽样的过程中会不断地更新α和β, 因此这两个值先根据经验预设即可, 本文设定$\alpha = 0.1, \beta = 0.01$. LDA 主题模型中主题个数 K 值的选取十分重要, 故本文对其进行参数调优实验. 分别对不同的 K 值计算基于地理位置标签信息的相似性特征函数$f_W(x,y)$, 然后根据$f_W(x,y)$进行链接预测实验. 实验结果如图 4 所示, 该图展示了主题个数 K 和加权调和平均 F 值的关系. 由图可知, 当主题个数为 13 时链接预测性能最优, 因此本文实验设定$K = 13$.

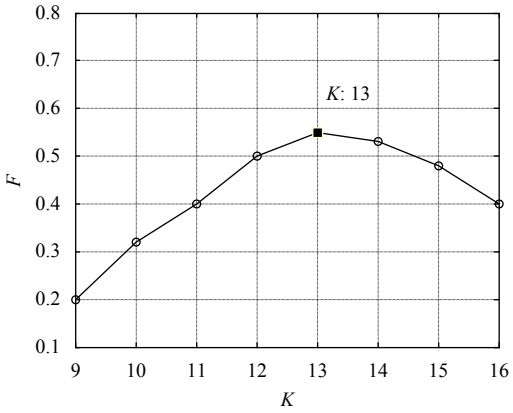

图 4 主题个数 K 选取对预测性能影响曲线

为了以最低的时间复杂度计算出特征函数$f_W(x,y)$, 并取得最优的链接预测效果, 本文将对地理位置标签主题进行$TOP\text{-}K$选择, 然后根据选取的$TOP\text{-}K$个主题计算基于地理位置标签信息的相似性特征函数$f_W(x,y)$. 实验结果如图 5 所示, 途中展示了$TOP\text{-}K$值与加权调和平均 F 值的关系. 从图中观察到当$TOP\text{-}K \geqslant 10$时, F 值较高且相对平稳, 而当$TOP\text{-}K$=5时 F 值达到巅峰, 故本文选取$TOP\text{-}K$=5进行链接预测实验.

3.4 实验结果与分析

本文在第 3.3 小节进行了相关参数调优, 接下来本文将使用开源的智能分析环境 WEKA 提供的几种分类器对 Gowalla 数据集进行链接预测实验: 朴素贝叶斯分类器 (NB)、随机森林分类器 (RF) 以及决策树分

类器 (J48). 分别对特征向量$\vec{feature}(x,y)=(f_{UP}(x,y),f_T(x,y),f_W(x,y),S_{x,y}^{RA})$和单一的用户网络特征$S_{x,y}^{RA}$进行实验, 并对比分析两种不同算法的性能. 采用的评估指标是 *Precision*、*Recall*、*F-measure* 和 AUC, 这些指标值是通过对比链接预测的实验结果和测试数据集中真是的数据作对比所得出的, 如表 3 所示.

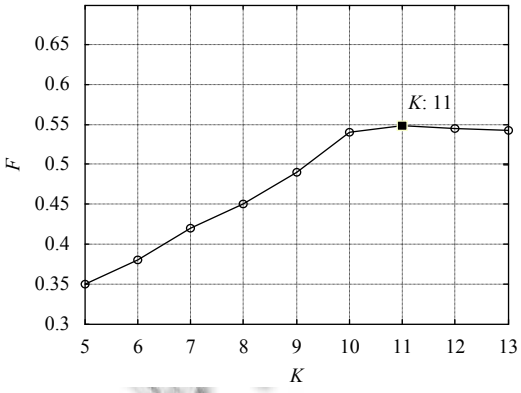

图 5 *Top–K* 主题个数对预测性能的影响曲线

从表 3 中可看出, 基于地理位置标签的特征向量$\vec{feature}(x,y)$在三种分类器上的链接预测效果均比基于单一用户网络特征$S_{x,y}^{RA}$取得的效果要好. 综合三个分类器的实验结果, $\vec{feature}(x,y)$算法的加权调和平均 $f-measure$ 值比$S_{x,y}^{RA}$平均高出 4.9%, $\vec{feature}(x,y)$算法的 AUC 比$S_{x,y}^{RA}$平均高 7.9%. 显然, 本文提出的基于地理标签的 LBSN 链接预测算法由于引入了地理标签语义特征, 其性能得到了明显的提升. 此外, $\vec{feature}(x,y)$的精度 *Precision* 平均值为 0.920, 比$S_{x,y}^{RA}$的 0.848 高出了 8.6%. 而$\vec{feature}(x,y)$的召回率 Recall 也比$S_{x,y}^{RA}$平均提高了 4.1%. 因此, 从用户地理位置标签信息中挖掘出来的信息纬度更加广阔, 能够得到更多的用户相似性特征, 从而降低了链接预测出错的概率.

从表 3 中还可看出, 以上两个实验中链接预测性能优劣为随机森林算法优于朴素贝叶斯算法, 而朴素贝叶斯算法又优于决策树算法, 单三者之间并无明显差异, 说明本文提出的算法具有良好的稳定性, 不同分类器对该算法的影响几乎可以忽略不计.

表 3 $\vec{feature}(x,y)$与$S_{x,y}^{RA}$预测结果对比

	分类器	朴素贝叶斯 (NB)		随机森林 (RF)		决策树 (J48)	
			提升 (%)		提升 (%)		提升 (%)
$\vec{feature}(x,y)$	*Precision*	0.920	8.6	0.921	8.6	0.918	8.5
	Recall	0.759	4.1	0.762	4.2	0.758	4.1
	F-measure	0.840	4.9	0.834	4.9	0.833	4.9
	AUC	0.89	7.5	0.915	8.4	0.914	7.9
$S_{x,y}^{RA}$	*Precision*	0.847		0.849		0.848	
	Recall	0.729		0.730		0.729	
	F-measure	0.782		0.785		0.784	
	AUC	0.833		0.830		0.836	

4 结论与展望

为解决地理位置分布的稀疏性问题, 本文对测试数据集进行聚类分析, 并建立了基于用户地理标签的 LDA 主题模型, 分析出用户的地理位置标签的相似性特征. 最后, 本文综合了网络结构相似性特征和基于用户地理位置信息的相似度特征, 采用有监督策略的链接预测在 Gowalla 数据集上进行实验. 实验结果表明, 本文提出的模型能有效提高 LBSN 的链接预测准确度, 且具有良好的稳定性. 然而随着互联网的发展, LBSN 的数据规模呈现指数级的增长, 未来可进一步研究基于大数据分布式平台的链接预测算法.

参考文献

1 Fisher A, Gilat D, Nadler S, *et al*. Device, system, and method of generating location-based social networks: US, US 20090215469A1. 2009-08-27.

2 Srinivas V, Mitra P. Link Prediction in Social Networks: Role of Power Law Distribution. Cham: Springer, 2016.

3 Scellato S, Noulas A, Lambiotte R, *et al*. Socio-spatial properties of online location-based social networks. Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain. 2011. 2011.

4 朱荣鑫. 基于地理位置的社交网络潜在用户和位置推荐模型研究[硕士学位论文]. 南京: 南京邮电大学, 2013.

5 Jurgens D, Finethy T, McCorriston J, *et al*. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. Proceedings of the 9th International AAAI Conference on Web and Social Media. Oxford, England. 2015. 188–197.

6 卢文羊, 徐佳一, 杨育彬. 基于 LDA 主题模型的社会网络链接预测. 山东大学学报 (工学版), 2014, 44(6): 26–31. [doi: 10.6040/j.issn.1672-3961.1.2014.116]

7 刘红兵, 李文坤, 张仰森. 基于 LDA 模型和多层聚类的微博话题检测. 计算机技术与发展, 2016, 26(6): 25–30. [doi: 10.3969/j.issn.1673-629X.2016.06.006]

8 张佳明, 王波, 唐浩浩, 等. 基于 Biterm 主题模型的无监督微博情感倾向性分析. 计算机工程, 2015, 41(7): 219–223. [doi: 10.3969/j.issn.1000-3428.2015.07.042]

9 马慧芳, 贾美惠子, 李晓红, 等. 一种基于标签关联关系的微博推荐方法. 计算机工程, 2016, 42(4): 197–201. [doi: 10.3969/j.issn.1000-3428.2016.04.035]

10 Porteous I, Newman D, Ihler A, *et al*. Fast collapsed gibbs sampling for latent dirichlet allocation. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, NV, USA. 2008. 569–577.

11 Lobo JM, Jiménez-Valverde A, Real R. AUC: A misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography, 2008, 17(2): 145–151. [doi: 10.1111/geb.2008.17.issue-2]

12 Ying JJC, Lu EHC, Lee WC, *et al*. Mining user similarity from semantic trajectories. Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks. San Jose, CA, USA. 2010. 19–26.