一种基于用户距离改进的线性影响力传播模型

蔡国永 裴广战
(桂林电子科技大学计算机与信息安全学院 桂林 541004)

摘 要 根据在线社交网络中用户的历史行为进行信息传播的预测是当前研究的热点之一,然而传统的传播模型仅解释了信息在社交网络中的传播规律,不具备信息传播预测能力。Jaewan Yang 和 Jwe Leskovec 根据未激活的用户会受到激活用户的影响,提出了线性影响力模型 LIM(Linear Influence Model),但是 LIM 模型在信息传播的过程中只考虑了时间因素,忽略了信息在传播过程中的空间因素,即用户间的相互关系。首先引入社交网络中用户间距离的度量,并结合距离的度量对 LIM 模型进行了改进,提出了基于距离正则化的 LIM 模型,即 d-LIM 模型。真实数据集上的对比实验表明,d-LIM 模型能获得更准确的预测结果。

关键词 社交网络,影响力,传播预测
中图法分类号 TP311.13 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.013

Improved Linear Influence Diffusion Model Based on Users' Distance

CAI Guo-yong PEI Guang-zhan
(School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract The prediction of information diffusion based on historical behavior of users in on-line social network is one of the hot spots of current research. However, the traditional propagation models can only explain the diffusion regular pattern of information in social networks, and it cannot predict the dissemination of information. Since uninfected users will be affected by infected users, Jaewan Yang and Jwe Leskovec proposed a linear influence model (LIM), but LIM model only considers the time factor in the process of information dissemination, and it ignores the spatial information, namely the relationship of users. Therefore, firstly, a measure of users' distance in social network was introduced in this paper. Combining the distance measure with LIM model, we proposed an improved LIM model based on distance regularization, namely d-LIM model. Through experiments on real data sets, the result shows that d-LIM model can achieve better prediction accuracy than the compared methods.

Keywords Social network, Influence, Diffusion prediction

1 引言

在 Web2.0 时代,受到社交网络中一些实际应用(如病毒式营销、个性化推荐等)的推动,国内外学者对社交网络中信息传播的研究产生了极大的兴趣。早期的研究工作主要集中于信息传播过程的理论分析和小规模的实验仿真,在该方面的成果中出现了两个比较经典的模型:独立级联模型(Independent Cascade, IC)[1]和线性阈值模型(Linear Threshold, LT)[2]。IC 模型和 LT 模型揭示了信息在传播过程中的影响因素和影响规律,但并不具备信息传播的预测能力[3]。对社交网络中信息传播的研究不仅要揭示影响信息传播的因素和规律,更要对信息在社交网络中的传播范围进行预测。

为此,研究者们相继提出了一些预测方法,包括回归分析[4,5]、机器学习[6,7]和概率预测[8-10]等。Jaewan Yang 和 Jwe Leskovec 根据未激活的用户会受到激活用户的影响,提出了线性影响力模型 LIM 来预测信息在网络中的传播[11]。作者认为每一个用户都有一个非负的影响力函数,用户影响力函数被认为是当用户创建(转发)一条消息后在一段时间内激活其他用户的数量,而激活的用户数量与影响力函数之间存在线性关系。LIM 模型在预测信息的传播过程中只考虑了消息传播的空间因素,而忽略了消息在网络中传播时用户间影响力的时态关系,即用户行为间的相互关系。本文首先引入了社交网络中用户与用户行为间的距离度量,并利用用户行为距离来对 LIM 模型进行正则化,提出了基于行为距离正则化的 LIM 模型,称为 d-LIM 模型。实验结果表明,d-LIM模型能获得比 LIM 模型更准确的传播预测能力。

2 LIM 模型基础

设 N 表示社交网络中用户的个数,K 表示传播信息的条数。令 $V_k(t)$ 表示第 k 条信息在 t 时间单元内传播的用户数,$M_{uk}(t)$ 表示信息 k 在 t 时间单元内是否激活了用户 u,若成功激活 u,则 $M_{uk}(t)=1$,否则为 0。LIM 模型假定 $V_k(t)$ 和 $M_{uk}(t)$ 之间存在线性关系,则第 k 条信息在 $t+1$ 时刻激活用户的个数可以表示为:

到稿日期:2015-10-13 返修日期:2016-01-26 本文受国家自然科学基金(61540053),广西研究生教育创新计划资助项目(YCSZ2015147)资助。
蔡国永(1971-),男,博士,教授,主要研究领域为社会媒体挖掘、可信网络计算;裴广战(1987-),男,硕士生,主要研究领域为社交网络信息传播、机器学习。

$$V_k(t+1)=\sum_{u=1}^{u=N}\sum_{l=0}^{l=L-1}M_{uk}(t-l)I_u(l+1) \quad (1)$$

其中，$I_u(l)$表示在 l 时间单元内用户 u 的影响力，向量 I_u 的长度为 L，表示 u 的影响力在 L 时间单元后降为 0，即 u 不能再激活其他的用户。对当前 t 时刻，若 t 时刻之前的 $l(l\leqslant t)$ 个时间单元内第 k 条信息成功激活用户 u，则 $M_{uk}(t-l)=1$。根据文献[11]，$V_k(t)$ 和 $M_{uk}(t)$组成的向量和矩阵可用图 1 表示。

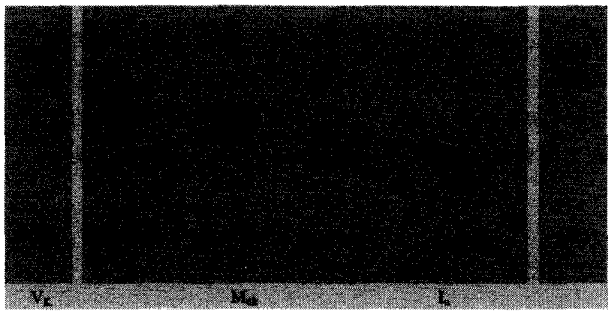

图 1 V_k，M_{uk}和 I_u 的矩阵表示

其中，向量 $V_k=(V_k(1),V_k(2),\cdots,V_k(T))$，$T$ 表示信息 k 传播过程结束的时刻；向量 $I_u=(I_u(1),I_u(2),\cdots,I_u(L))$ 表示用户 u 的影响力向量；由式(1)可知，矩阵 M_{uk} 的行表示信息 k 传播的总时间为 T，列表示结点 u 在激活之后影响力随时间变化的关系，总时间为 L。

令 $V=(V_1,V_2,\cdots,V_K)$，则向量 V 表示 K 条信息在 T 时间单元内激活用户的总数量，矩阵 $M=(M_{uk})_{K\times N}$ 表示影响力指示矩阵，向量 $I=(I_1,I_2,\cdots,I_N)$ 表示影响力向量。图 2 示出了 K 条信息激活用户的个数(V)等于参与消息传播的 N 个用户(M)与用户影响力(I)的线性组合关系。

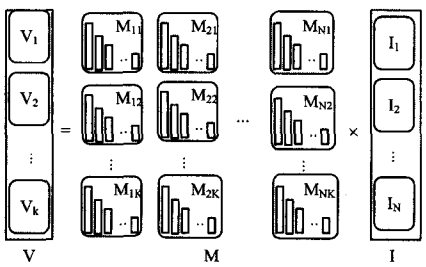

图 2 矩阵的关系等式 $V=M\cdot I$

给定 V，M 和 I，式(1)表示成矩阵的形式为：

$$V=M\cdot I \quad (2)$$

根据式(2)，LIM 模型可以通过求解一个非负的最小二乘问题(Non-negative Least Square Problem)来学习影响力向量 I：

$$\min_I \frac{1}{2}\|V-M\cdot I\|_F^2$$

$$\text{subject to } I\geqslant 0 \quad (3)$$

其中，$\|\cdot\|_F$ 表示 Frobenius 范数。

3 改进的 LIM 模型

LIM 模型在信息的传播过程中只考虑了信息评论次数与用户影响力随时间变化的关系，忽略了信息传播过程中用户行为间的相互联系所起的作用。在实际情况中，信息很大程度上是在具有一定相似关系的用户间传播。针对这种情况，可以对 LIM 模型进行改进以获得更好的预测结果。首先，根据网络中用户兴趣的不同，引入了用户间的距离度量，然后结合距离度量对 LIM 模型进行了扩展，提出了基于用户间距离的 LIM 模型(d-LIM)。

3.1 距离度量

对于社交网络中的两个用户，如果他们的行为是相似的，则信息很可能在这个两个用户间相互传播，于是根据用户行为的不同来衡量用户间的距离。考虑社交网络用户 u 和 v，令 C_u 和 C_v 分别表示 u 和 v 评论的信息集，则用户 u 和 v 之间的距离定义为：

$$dis(u,v)=1-\frac{C_u\cap C_v}{C_u\cup C_v} \quad (4)$$

其中，$C_u\cap C_v$ 表示 u 和 v 共同评论的信息，$C_u\cup C_v$ 表示 u 或 v 评论的信息。从式(4)可以看出，u 和 v 共同评论的信息越多，两个用户间的距离也就越小，即信息在他们之间传播的可能性越大。

3.2 d-LIM 模型

令 $F^+(u)$表示用户 u 的出度邻居集合，I_u 表示用户 u 的影响力向量(与 LIM 模型一致)。对于给定的影响力向量 I_u 和 I_v 以及 $dis(u,v)$，两个用户间的影响力与距离应满足式(5)的约束关系：

$$\frac{\beta}{2}\sum_{u=1}^{N}\sum_{v\in F^+(u)}dis(u,v)\|I_u-I_v\|_F^2 \quad (5)$$

式(5)表示用户影响力 I_u 和 I_v 间的范数距离应该受到行为距离的调控，用户行为间距离 $dis(u,v)$的值越小，在正则化时用户 u 的影响力与用户 v 的影响力也应更接近。因此，结合式(3)和式(5)，d-LIM 模型采用如式(6)所示的优化目标函数：

$$\min_{I_u} L(V,M,I)=\frac{1}{2}\sum_{u=1}^{N}\sum_{k=1}^{K}\|V_k-M_{uk}\cdot I_u\|_F^2+\frac{\beta}{2}\sum_{u=1}^{N}\sum_{v\in F^+(u)}dis(u,v)\|I_u-I_v\|_F^2+\frac{\lambda}{2}\sum_{u=1}^{N}\|I_u\|_F^2$$

$$\text{subject to } I_u\geqslant 0 \quad (6)$$

其中，第 1 项与 LIM 模型类似，第 2 项与第 3 项为正则约束项，即对用户的邻居间的行为相似关系进行了显式的约束。正则化后的目标函数即式(6)的局部最小值可以通过梯度下降的方法进行求解。对影响力向量 I_u 求偏导数得：

$$\frac{\partial L}{\partial I_u}=\sum_{k=1}^{K}(V_k-M_{uk}\cdot I_u)\cdot M_{uk}+\lambda I_u+\beta\sum_{v\in F^+(u)}dis(u,v)(I_u-I_v) \quad (7)$$

根据式(7)中对影响力向量 I_u 的更新，可以写出目标函数即式(6)的梯度下降算法，如算法 1 所示。

算法 1 目标函数(式(6))的梯度下降算法

输入：V_k，M_{uk}

输出：I_u

```
1. τ←0, I_u←0
2. for u=1 to N do
3.     for v in F+(u)
4.     d+=dis(u,v)//记录用户 u 和他邻居用户的距离
5.   end for
6. end for
7. while τ<count do  //count 为迭代次数
8.     δe←Muk · Iu−Vk//计算误差
```

9.　$I_u \leftarrow I_u - \alpha(M_{uk}^T \times \delta_e + \beta d(I_u - I_v) + \lambda I_u)$

10.　$\tau \leftarrow \tau + 1$

11. end while

12. return I_u

算法 1 给出了近似求解影响力 I_u 的方法。首先通过 d 记录了一个用户和他的出度邻居用户之间的距离；其次算法的 5－8 行给出了梯下降算法更新 I_u 的计算过程，*count* 表示求解过中迭代的次数[1]，δ_e 表示每次更新过后的误差值，M_{uk}^T 表示 M_{uk} 的转置矩阵。当 while 循环结束后，算法 1 返回影响力向量 I_u。

计算出 I_u 后，可以通过 M_{uk} 与 I_u 的乘积求出第 k 条信息激活的用户数$\hat{V}_k$，与真实的 V_k 相比，$\hat{V}_k$ 和 V_k 的值越近，表明求出的影响力向量 I_u 越精确。令 $m = |F^+(u)|$ 表示用户 u 的出度用户的个数，N 表示用户的个数，T 为算法迭代的次数，则算法 1 的时间复杂度为 $O(T+Nm)$。

4　实验与分析

本节在真实数据集中对比了 LIM 模型和 d-LIM，证实了 d-LIM 模型在预测信息传播时能获得更好的结果。

4.1　数据集

本文实验中用到的数据集来自著名的新闻分享网站 Digg[2]。Digg 中的用户可以对新闻进行投票（即 Digg）、分享和评论等，同时也可以关注其他用户从而形成好友关系。在本文的实验中，从 Digg 网站中 2009 年 6 月 3553 条最受欢迎的新闻中抽取了约 7000 个用户的评论信息，这 3553 条信息总共收到了 7000 个用户大约 30 万的次投票，以及 7000 个用户包含了 14 万多的好友关系。数据集的详细信息如表 1 和表 2 所列。

表 1　Friends 关系表数据信息

Tag	Time	user1	user2
0 或 1	时间	用户 ID1	用户 ID2

表 2　News 传播表数据信息

Time	User	News
时间	用户 ID	新闻 ID

表 1 中的 Time 表示了 user1 和 user2 建立朋友关系的时刻，Tag 表明了两个用户的关系是否是相互的，若 Tag＝1，则 user1 和 user2 的关系是相互的，否则，user1 和 user2 的关系是单向的。表 2 中的 Time 信息表明了用户（user）在这一时刻对新闻（news）做出了评论。

4.2　评价方法

采用两种比较流行的度量方法：平均绝对误差（MAE）和均方根误差（RMSE）。MAE 的表达式定义为：

$$MAE = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{T}\left(\sum_{t=1}^{T}|\hat{V}_k(t) - V_k(t)|\right)\right) \tag{8}$$

其中，$\hat{V}_k(t)$ 表示第 k 条信息在 t 时间段内预测感染用户的数量，$V_k(t)$ 表示相应的真实值，K 表示新闻的总数量，T 表示信息传播的总时间。RMSE 的表达式定义为：

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{T}\left(\sum_{t=1}^{T}(\hat{V}_k(t) - V_k(t))^2\right)\right)} \tag{9}$$

从 MAE 和 RMSE 的定义可以看出，两者的值越小则预测的结果越准确。

4.3　实验结果

在算法 1 中，需要设置一些参数（如时间标签 L，正则化的系数 α，β 和 λ），在所有的实验中，α，β 和 λ 的值分别设置为 0.0002，0.001 和 0.001，实验参数的选取依据参考文献[12]。参数 L 表明一个用户的影响力在多长时间内降为 0，在 Digg 数据集中，一条新闻在发布之后，传播速度一般在 1 天后开始迅速下降，因此，在实验过程中选取了 L＝24（小时），即文中的实验预测的是新闻在发布 24 小时内激活用户的个数。

本文把数据分为训练集和测试集，训练集的大小分别为 80%，60%，40%，例如，80% 的训练集表示将从 3553 条新闻中随机选取 2800 条新闻在 24 小时内激活的用户数作为训练集，剩余的 753 条新闻作为测试集。两个模型的 MAE 值和 RMSE 值如表 3 所列。

表 3　LIM 模型和 d-LIM 模型的 MAE 值和 RMSE 值

Training(%)	Metrics	LIM	d-LIM
40	MAE	2.1979	1.9625
	RMSE	4.7745	4.5297
60	MAE	1.9883	1.7552
	RMSE	4.4384	4.2344
80	MAE	1.6995	1.4919
	RMSE	4.0479	3.8964

从表 3 的实验结果可知，d-LIM 模型在相应的训练集中比 LIM 模型具有更低的 RMSE 值和 MAE 值，从而在预测信息传播时能获得更好的结果。

本文同时列出了 d-LIM 模型与 LIM 模型的算法运行时间，如表 4 所列。

表 4　LIM 模型和 d-LIM 模型的算法运行时间（RMSE）

Training(%)	LIM(seconds)	d-LIM(seconds)
40	112	144
60	166	196
80	221	252

从表 4 中可以看出，与 LIM 模型相比，d-LIM 模型的算法运行时间增加得很少，但是从表 3 中可知，d-LIM 模型预测的准确率有很大的提高。

结束语　本文在 LIM 模型的基础上，结合社交网络用户的行为距离度量，提出了 d-LIM 模型。实验表明，d-LIM 模型在预测信息传播时可以获得更高的准确率。同时从实验中可以看出，用户间的行为距离对用户的影响力有着重要的作用，在实验中只考虑了用户对新闻评论的次数来计算用户行为间的距离。然而，社交网络中用户具有很多不同的属性，如何根据用户的历史交互数据和属性提出用户间更准确的距离度量，并在预测传播时把这些距离度量显示地嵌入到预测模型的学习中，是未来工作中需要考虑的方向。

参 考 文 献

[1]　SAITO K, NAKANO R, KIMURA M. Prediction of information diffusion probabilities for independent cascade model [M] // Knowledge-based Intelligent Information and Engineering Sys-

[1] 实验过程中选取 count 的迭代次数为 5000

[2] http://www.isi.edu/~lerman/downloads/digg2009.html

tems. Springer Berlin Heidelberg, 2008: 67-75.

[2] CHEN W, WANG C, WANG Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010: 1029-1038.

[3] LI Dong, XU Zhi-ming, LI Sheng, et al. A survey on information diffusion in online social networks [J]. Chinese Journal of Computers, 2014, 37(1): 189-206. (in Chinese)
李栋, 徐志明, 李生, 等. 在线社会网络中信息扩散[J]. 计算机学报, 2014, 37(1): 189-206.

[4] YANG J, COUNTS S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter[J]. ICWSM, 2010, 10: 355-358.

[5] WANG F, WANG H, XU K. Diffusive logistic model towards predicting information diffusion in online social networks[C]// 2012 32nd International Conference on Distributed Computing Systems Workshops (ICDCSW). IEEE, 2012: 133-139.

[6] GUILLE A, HACID H. A predictive model for the temporal dynamics of information diffusion in online social networks[C]// Proceedings of the 21st International Conference Companion on World Wide Web. ACM, 2012: 1145-1152.

[7] BOURIGAULT S, LAGNIER C, LAMPRIER S, et al. Learning social network embeddings for predicting information diffusion [C]// Proceedings of the 7th ACM International Conference on Web Search and Data Mining. ACM, 2014: 393-402.

[8] WANG Y, XIANG G, CHANG S K. Sparse Multi - Task Learning for Detecting Influential Nodes in an Implicit Diffusion Network[C]// AAAI. 2013.

[9] LIN Y, RAZA A A, LEE J Y, et al. Influence propagation: patterns, model and a case study[M]// Advances in Knowledge Discovery and Data Mining. Springer International Publishing, 2014: 386-397.

[10] WANG F, WANG H, XU K, et al. Characterizing information diffusion in online social networks with linear diffusive model [C]// 2013 IEEE 33rd International Conference on Distributed Computing Systems (ICDCS). IEEE, 2013: 307-316.

[11] YANG J, LESKOVEC J. Modeling information diffusion in implicit networks[C]// 2010 IEEE 10th International Conference on Data Mining (ICDM). IEEE, 2010: 599-608.

[12] MA H, ZHOU D, LIU C, et al. Recommender systems with social regularization[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, 2011: 287-296.

(上接第 47 页)

[13] HE Feng-ying. Orientation analysis for Chinese blog text based on semantic comprehension [J]. Journal of Computer Applications, 2011, 31(8): 2130-2133. (in Chinese)
何凤英. 基于语义理解的中文博文倾向性分析[J]. 计算机应用, 2011, 31(8): 2130-2133.

[14] LI Rong-jun, WANG Xiao-jie, ZHOU Yan-quan. Semantic Orientation Computing Using PageRank Model [J]. Journal of Beijing University of Posts and Telecommunications, 2010, 5(5): 141-144. (in Chinese)
李荣军, 王小捷, 周延泉. PageRank 模型在中文情感词极性判别中的应用[J]. 北京邮电大学学报, 2010, 5(5): 141-144.

[15] COLACE F, SANTO M D, GRECO L. SAFE: A Sentiment Analysis Framework for E-Learning[J]. International Journal of Emerging Technologies in Learning, 2014, 9(6): 37-41.

[16] MUKKAMALA R R, HUSSAIN A, VATRAPU R. Fuzzy-Set Based Sentiment Analysis of Big Social Data[C]// IEEE 18th International Enterprise Distributed Object Computing Conference (EDOC), 2014. IEEE, 2014: 71-80.

[17] TURNEY P D, LITTMAN M L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association[J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[18] CHEN Lu, WANG Wen-bo, NAGARAJAN M, et al. Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter[C]// The Sixth International AAAI Conference on Weblogs and Social Media (ICWSM). 2012.

[19] JO Y, OH A H. Aspect and sentiment unification model for online review analysis[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, 2011: 815-824.

[20] NEVIAROUSKAYA A, PRENDINGER H, ISHIZUKA M. SentiFul: Generating a reliable lexicon for sentiment analysis[C]// 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009 (ACII 2009). IEEE, 2009: 1-6.

[21] SAIF M, CODY D, Bonnie D. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus [C]// Proc. of 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09). 2009: 599-608.

[22] CONTE H R, PLUTCHIK R. A circumplex model for interpersonal personality traits[J]. Journal of Personality & Social Psychology, 1981(4): 701-711.

[23] TOMÀŠ M. Statistical Language Models based on Neural Networks[D]. Brno University of Technology, 2012.

[24] TOMÀŠ M, KARAFIÁ T M, BURGET L, et al. Recurrent neural network based language model[C]// Conference of the International Speech Communication Association, 2010. Makuhari, Chiba, Japan, 2010: 1045-1048.

[25] CHEN Jian-mei, LIN Hong-fei, YANG Zhi-hao. Word Emotion Disambiguation Based on Bayesian Model[C]// The Ninth China National Conference on Computational Linguistics, 2007. (in Chinese)
陈建美, 林鸿飞, 杨志豪. 基于贝叶斯模型的词汇情感消歧[C]// 内容计算的研究与应用前沿——第九届全国计算语言学学术会议论文集. 2007.

[26] DING Ru-yi, ZHOU Hui, LIN Ma. Cognitive Appraisal Basis of Gratitude. [J]. Acta Psychologica Sinica, 2014, 46(10): 1463-1475. (in Chinese)
丁如一, 周晖, 林玛. 感激情绪的认知评估体系[J]. 心理学报, 2014, 46(10): 1463-1475.