doi:10.11959/j.issn.1000-436x.2017264

基于用户影响与兴趣的社交网信息传播模型

王瑞，刘勇，朱敬华，玄萍，李金宝
(1. 黑龙江大学计算机科学技术学院，黑龙江 哈尔滨 150080；2. 黑龙江省数据库与并行计算重点实验室，黑龙江 哈尔滨 150080)

摘 要：提出了一种新的无拓扑结构的社交网信息传播模型，简称 NT-II，并使用表达学习方式，构建了 2 个隐藏的空间：用户影响空间和用户兴趣空间，每个用户和每个传播项都映射成空间中的向量。模型在预测用户接收传播项的概率时，既考虑来自其他用户的影响程度，又考虑该用户对传播项的喜爱程度，分别根据 2 个用户向量之间的距离和用户向量和传播项向量之间的距离来推断。实验结果表明：NT-II 模型能更准确地模拟传播过程和预测传播结果。

关键词：传播模型；表达学习；用户影响空间；用户兴趣空间
中图分类号：TP311 **文献标识码**：A

Social network information diffusion model based on user's influence and interesting

WANG Rui, LIU Yong, ZHU Jing-hua, XUAN Ping, LI Jin-bao
(1. School of Computer Science and Technology, Heilongjiang University, Harbin 150080,China;
2. Key Laboratory of Database and Parallel Computing of Heilongjiang Province, Harbin 150080, China)

Abstract: A new non-topological information diffusion model of social network was proposed, called non-topological influence-interest diffusion model (NT-II). Representation learning was exploited to construct two hidden spaces for NT-II, called the user-influence space and the user-interest space, each user and each propagation item was mapped into a vector in space. The model predicted the probability of a user receiving a propagated item, considering not only the degree of influence from other users, but also the user's preference for propagated item. The experimental results show that the model can simulate the propagation process and predict the propagation results more accurately.

Key words: diffusion model, representation learning, user-influence space, user-interest space

1 引言

随着社会信息系统的广泛应用和数据量的激增，在海量的数据中提取隐藏有用信息的研究问题已经越来越普遍，其中拥有庞大数据量的在线社交网络也渐渐兴起，不断推陈出新，它们已经是许多人上网最主要的目的，与此同时，也带来了许多研究课题，在随处可见的社交网络中，如微博、论坛，用户的转发、评论、分享等行为已经成为企业和推广市场的重要信息源，所以正确地预测社交网中的信息传播过程和以后的传播趋势尤为重要。

社交网信息传播预测效果极大依赖于传播模型。自 2003 年独立级联(IC)模型和线性阈值(LT)模型[1]提出，以图结构为基本骨架的社交网络不断发

收稿日期：2017-09-14
通信作者：刘勇，acliuyong@sina.com
基金项目：国家自然科学基金资助项目（No.6137022, No.61602159, No.61300225）；黑龙江省自然科学基金资助项目（No.F201430, No.F2015013）；哈尔滨科技创新人才研究专项基金资助项目（No.2017RAQXJ094, No.2015RAQXJ004, No.2015RAXXJ004）；黑龙江省高校基本科研业务费基金资助项目（No.HDJCCX-201608）
Foundation Items: The National Natural Science Foundation of China (No.6137022, No.61602159, No.61300225), The Natural Science Foundation of Heilongjiang Province (No.F201430, No.F2015013), The Innovation Talents Project of Science and Technology Bureau of Harbin (No.2017RAQXJ094, No.2015RAQXJ004, No.2015RAXXJ004), The Fundamental Research Funds of Universities in Heilongjiang Province (No.HDJCCX-201608)

展，从最初的探索了解阶段，到对传播网络结构[2]和内容的继续深入，人们不断探索信息传播的过程和其应用价值，其中根据具体图结构预测社交网传播概率和信息传播路径的研究已日渐成熟，而基于无拓扑结构的社交网络模型的研究相对较少，本文构建一种有效的无拓扑结构信息传播模型。

随着数据时代的来临，拥有大量数据的社交网络飞速发展，结合现有社交网络的强大功能和信息传播方式，本文发现大部分现有模型并不适应现在很多真实的社交情况。例如，一个用户发布一条微博，很难知道其内容的源头在哪儿，呈现出来的通常是（用户，微博内容，时间）这样的一个三元组，即使微博有追溯信息从哪里转发而来的功能，直观上也只能观测到几个用户的部分连接，这些部分连接可能对应于有用的信息，但它们往往不能代表一个社交网络信息真正的扩散趋势。并且不只考虑微博这样的社交网络，真正的社交网络用户通常可以从不同渠道获取信息，大多数情况下并不知道具体是谁影响了某个用户。我们能捕捉到的通常是一个不完整的、无关紧要的甚至是未知（隐私）的部分[3]。在社交网的研究过程中，将其看作一个图结构 $G=\{V, E\}$，V 是图的顶点集合代表社交网中的用户，E 是图的边集合代表社交网中用户之间的连接（有连接代表一个用户可以向另一个用户传递信息），在之前的描述中，显然并不能完全知道社交网的传播过程，也就是某个社交网络的图结构，而大部分现有的传播模型的构建是以一些具有准确图结构的数据集为输入数据的，所以说它们不适用现在的社交网络。

为了解决上述问题，本文所采用的研究数据正是类似于（用户，微博内容，时间）形式的数据集，并且假定给定的数据中没有社交网用户之间的具体关系，而是通过构造的 2 个隐藏空间，判断其他用户和传播项对某个具体用户传播信息影响，第一个空间中，计算传播项发送方用户向量与接收方用户向量的距离，第二个空间中，计算传播项向量与接收方用户向量的距离，距离越近，该用户传播信息的概率就越大，结合这 2 方面因素去推断真正的影响关系，从而构建无拓扑结构影响—兴趣传播模型。

本文模型对比其他无拓扑结构社交网络信息传播模型的进步之处在于考虑了传播项内容对于传播概率的影响，对于其他的信息预测问题，比较新也比较好的方法是通过挖掘用户和用户相互影响的隐藏特征来预测真实的信息传播过程[4]，但是此方法没有考虑传播项内容对于传播概率的影响。事实上判断一个用户是否转发一个传播项，不仅取决于他能否看见其他用户的转发，也取决于他对传播内容本身是否感兴趣[5]，所以传播项对于用户是否转发信息同样存在决定性作用。本文中，利用表达学习思想[6]，使用空间向量构建概率模型，该模型不仅考虑用户之间的相互影响，也考虑传播项对用户的影响，使用梯度上升算法学习模型参数。本文的主要贡献如下。

1) 提出了无拓扑结构影响—兴趣传播模型 NT-II。其优势是可以直接从观察数据集入手而不用假设未知的图结构潜在的各种影响关系，并且使用表达学习方法可以学习更少的参数，适用于更大的社交网络。

2) 现有的传播模型中，预测传播过程通常使用训练集中传播项的特征向量，但是测试集中的传播项还存在一些新的传播项，为了提高预测的准确性，也提出了学习新传播项的特征向量的学习算法。

3) 实验结果表明，相比之前的社交网信息传播模型，本文可以更准确地模拟传播过程和预测传播结果。

2 相关工作

本节主要介绍社交网信息传播模型的相关工作，主要分为两大类：1) 有拓扑结构的社交网信息传播模型；2) 无拓扑结构的社交网信息传播预测。

有拓扑结构的社交网信息传播模型。2003 年提出的 IC 模型是最传统的传播模型，由于 IC 模型非常依赖前项用户，Saito 等[7]根据这一特性，在 2008 年成功预测出此模型的传播概率。随后，Saito 团队又进一步考虑了节点的属性提出了改进的 IC 模型并对信息传播过程进行预测[8]。近几年中，数据挖掘技术越来越成熟，研究者可以从传播项文本内容中提取重要信息并构建传播模型[9]。还有很多其他的方法分别从用户和传播项 2 个方面入手来构建传播模型，这里不一一列举。研究人员也在不断尝试多种因素相结合的社交网信息传播模型。例如，2016 年 Lagnier 成功地提出了一种用户和文本内容相结合的传播模型[10]。

无拓扑结构的社交网信息传播模型。目前，对于无图结构的社交网信息传播的研究工作还比较少。文献[11]中 Rodriguez 通过新的新闻媒体信息和博客信息构建传播模型，并成功地追溯用户的传播路径。但其实验的复杂度和结果的准确性还有待提高。还有利用泊松过程去推测一段单一文本的事件级联的传播模型[12]。同样构建了连续隐藏空间的传播模型，如将时序信息映射到一个连续的空间中并结合扩散核函数预测信息传播[13]。最后介绍一种初步使用用户隐藏空间特征预测传播概率的模型[4]，它使用了嵌入式传播的概念，将每个用户表示为空间中的特征向量，通过计算向量中的距离来分析 2 个用户影响关系的可能性大小，也就是一个用户影响另一个用户转发某个传播项可能性的大小，从而来推测信息的传播过程，构建传播模型。

在众多的预测信息传播的方法中，既利用以往的有效经验和手段，也要总结其中的不足，例如有的算法并不适用于较大的数据，有的实验只对特定的数据集来说准确性较高，有的算法只考虑用户对于用户的重要影响，忽略传播项本身的重要作用等。基于以上问题，提出了本文模型 NT-II 模型，该模型与之前模型相比较的优势为：1) 更适用于较大的数据；2) 对于现实的无拓扑结构的数据集来说准确性较高；3) 在构建社交网传播模型的过程中，对于一个用户是否转发某个传播项的原因，不只考虑用户之间的相互影响，还考虑了该传播项本身在传播过程对用户是否传播它的影响程度（如用户对该传播项的喜爱程度）。

3　相关概念

提到在社交网络的研究中，通常将其看作一个图结构 $G=\{V, E\}$，集合 V 相当于社交网中的用户集合，本文提出的是无拓扑结构的社交网传播模型，对于集合 E（用户的连接）并不过多关注。

本文将社交网中可观察的传播片段定义 D。D_i 代表某一条信息 i 在社交网中的传播轨迹，数学符号定义为 $D_i=\{\{(u, t_i(u))|u\in V\wedge 0\leqslant t_i(u)<+\infty\}$，其中，$t_i(u)$表示用户 u 接收事件 i 的时间，或是用户 u 接收 D 中事件的时间。举例说明，假设观察到微博中某一个话题 i“我爱吃甜食”，那么在一段时间中所有转发此话题的用户就可以叫作一个传播片段 D_i，代表这段时间涉及相同事件的用户集合。本文将这些传播片段的集合定义为 Ω，因此 $\Omega=\{D_1, D_2,\cdots, D_K\}$表示社交网上的多条传播轨迹的集合。

在实际运算中，考虑不同时间段的同一传播片段上的用户集合。$D(t)$表示在时间 t 前接收传播轨迹 D 中的事件的用户集合，$\overline{D}(t)$表示时间 t 前未接收传播轨迹 D 中的事件的用户集合。同理，本文以 $D(\infty)$表示接收传播轨迹 D 中事件的全体用户。$\overline{D}(\infty)$表示没有接收传播轨迹 D 中事件的全体用户。最后补充说明一个用户集合 $A(A\in V)$表示的是同一传播片段上被影响的用户集合。

值得注意的是，本文关注的是某人在某个时间做了什么，并不关注这件事情是如何发生的。在本文的模型中，一个用户传播信息的概率取决于同一传播片段 D 上该用户之前的用户和 D 上的信息 i 对其的影响。

4　无拓扑结构的影响—兴趣传播模型 NT-II

4.1　无拓扑结构的影响—兴趣传播模型 NT-II 的定义

本节分析用户和传播项对信息传播的双重影响继而构建本文提出的新的社交网信息传播模型 NT-II。

本模型中 $P_{u, v}$ 为用户 u 对用户 v 的影响程度，即影响空间形成的信息传播概率。$P_{v, i}$ 为用户 v 对传播项 i 的喜欢程度，也就是兴趣空间形成的信息传播概率，然后本文介绍如何用适合的函数来求解 2 种传播概率。最后，本文结合概率 $P_{u, v}$ 和 $P_{v, i}$ 的值预测出更接近真实传播概率的预测概率，称为无拓扑结构的影响—兴趣传播模型传播概率。对此概率和无拓扑结构的影响—兴趣传播模型的具体定义如下。

无拓扑结构的影响—兴趣传播模型传播概率 $P(v|u, i)$：用户 v 在看到用户 u 接收传播项 i 之后，v 也接收传播项 i 的概率。因为用户 v 是否接收传播项 i 既与用户 v 对传播项 i 的喜欢程度有关，又与用户 u 对用户 v 的影响程度有关。如式(1)所示（α 代表 2 种传播概率的权重）。

$$P(v|u,i)=\alpha P_{u,v}+(1-\alpha)P_{v,i} \tag{1}$$

无拓扑结构的兴趣—影响传播模型（简称 NT-II）：在信息传播的过程中，当用户 v 看到用户 u 接收传播项 i 时，用户 v 有且只有一次机会接收传播项 i。v 接收传播项 i 的概率为 $P(v|u, i)$，v 不

接收传播项 i 的概率为 $1-P(v|u, i)$。假定 v 接收传播项 i 后不会再次接收传播项 i。因为没有图的拓扑结构，假定在传播项 i 上活跃的用户对其他用户都可能产生影响。因此，在用户集合 A 接收传播项 i 后用户 v 接收传播项 i 的概率 $P(v|A, i)$如式(2)所示。基于这样的传播过程，本文构建的传播概率模型称为无拓扑结构的兴趣—影响传播模型。

$$P(v \mid A,i) = 1 - \prod_{u \in A}(1 - P(v \mid u,i)) \tag{2}$$

这里介绍影响空间的传播概率 $P_{u,v}$ 和兴趣空间的传播概率 $P_{v,i}$ 的具体表示方法。假设 2 个空间都是 d 维的隐藏空间，本文用 2 个函数 f 和 g 分别表示 2 个概率，如下所示

$$P_{u,v} = f(\boldsymbol{z}_u, \boldsymbol{w}_v) \tag{3}$$

$$P_{v,i} = g(\boldsymbol{x}_v, \boldsymbol{y}_i) \tag{4}$$

其中，函数的参数 $\boldsymbol{z}_u$ 为影响空间中用户 u 的 d 维的隐藏空间（影响其他用户的向量，信息发送方），$\boldsymbol{w}_v$ 为影响空间中用户 v 的被影响的 d 维特征向量(被其他用户影响的向量，信息接收方)。$\boldsymbol{x}_v$ 为兴趣空间中用户 v 的 d 维特征向量，$\boldsymbol{y}_i$ 为兴趣空间中传播项 i 的 d 维特征向量。

特征向量的表示方法是人为设置初始向量每一维的值，然后使用本文给出参数学习算法对向量每一维的数值不断进行迭代，通过训练集的训练最后得到每个用户的最终近似向量，再通过这些用户向量在测试集中去预测社交网的传播过程。算法 1 中的输出（NT-II 模型的参数集合 $Z=\{\boldsymbol{z}_u\}$，$W=\{\boldsymbol{w}_u\}$，$X=\{\boldsymbol{x}_u\}$，$Y=\{\boldsymbol{y}_i\}$）就是用户向量和传播项向量中的每一维的具体数值。向量中的每一维代表的含义在刚开始都是隐含的，但是它们都是用户向量和传播项向量的一个特征，具体信息在初始时是未知的，所以本文选择表达学习算法来进行模型的构建，因为表达学习算法正是一种能够发现隐含解释特征的学习方法。它捕捉观测输入数据的隐含解释性因素的来构建传播模型比直接输入大量的具体特征来构建传播模型效率要高很多。

算法 1 求无拓扑结构的兴趣—影响传播模型 NT-II 参数的学习算法

输入 用户集合 U；传播项集合 I；传播轨迹集合 $\Omega = \{D_1, D_2, \cdots, D_k\}$；学习率 ε

输出 NT-II 模型的参数 $Z = \{\boldsymbol{z}_u\}$、$W = \{\boldsymbol{w}_u\}$、$X = \{\boldsymbol{x}_u\}$、$Y = \{\boldsymbol{y}_i\}$

1)for all the $u \in U$ do
2) $\boldsymbol{z}_u \leftarrow$ random values in$[-1,1]^d$;
3) $\boldsymbol{w}_u \leftarrow$ random values in$[-1,1]^d$;
4) $\boldsymbol{x}_u \leftarrow$ random values in$[-1,1]^d$;
5) end
6) for all the $i \in I$ do
7) $\boldsymbol{y}_i \leftarrow$ random values in$[-1,1]^d$;
8) end
9) repeat
10) for each $D \in \Omega$ do
11) for each $v \in \overline{D}(1)$;
12) if $t_D(v) < \infty$ then根据式(1)～式(3)计算P_v^D
13) end;
14) for $u \in D(t_D(v))$ do
15) $\lambda_v^+ = \dfrac{\partial \mathrm{lb} P(v \mid u, i_D)}{\partial x_v}$;
$\lambda_i^+ = \dfrac{\partial \mathrm{lb} P(v \mid u, i_D)}{\partial y_i}$
16) $\lambda_v^- = \dfrac{\partial \mathrm{lb}(1 - P(v \mid u, i_D))}{\partial x_v}$;
$\lambda_i^- = \dfrac{\partial \mathrm{lb}(1 - P(v \mid u, i_D))}{\partial y_i}$
17) if $t_D(v) < \infty$ then
18) $x_v \leftarrow x_v + \varepsilon(\dfrac{P(v \mid u, i)}{P_v^D}\lambda_v^+ + (1 - \dfrac{P(v \mid u, i)}{P_v^D})\lambda_v^-)$;
19) $y_i \leftarrow y_i + \varepsilon(\dfrac{P(v \mid u, i)}{P_v^D}\lambda_i^+ + (1 - \dfrac{P(v \mid u, i)}{P_v^D})\lambda_i^-)$;
20) else
21) $x_v \leftarrow x_v + \varepsilon\lambda_v^-$;
$y_i \leftarrow y_i + \varepsilon\lambda_i^-$
22) end
23) end
24) end
25) end
26) for each $D \in \Omega$ do
27) for each $v \in \overline{D}(1)$;
28) if $t_D(v) < \infty$ then根据式(1)～式(3)

计算P_v^D

29) end;

30) for $u \in D(t_D(v))$ do

31) $\delta_u^+ = \dfrac{\partial \log P(v|u,i_D)}{\partial \boldsymbol{z}_u}$;

$\delta_v^+ = \dfrac{\partial \log P(v|u,i_D)}{\partial \boldsymbol{w}_v}$

32) $\delta_u^- = \dfrac{\partial \log(1-P(v|u,i_D))}{\partial \boldsymbol{z}_u}$;

$\delta_v^- = \dfrac{\partial \log(1-P(v|u,i_D))}{\partial \boldsymbol{w}_v}$

33) if $t_D(v) < \infty$ then

34) $z_u \leftarrow z_u + \varepsilon \times (\dfrac{P(v|u,i)}{P_v^D}\delta_u^+ + (1-\dfrac{P(v|u,i)}{P_v^D})\delta_u^-)$;

35) $w_v \leftarrow w_v + \varepsilon \times (\dfrac{P(v|u,i)}{P_v^D}\delta_v^+ + (1-\dfrac{P(v|u,i)}{P_v^D})\delta_v^-)$;

36) else

37) $\boldsymbol{z}_u \leftarrow \boldsymbol{z}_u + \varepsilon\delta_u^-$;

$\boldsymbol{w}_v \leftarrow \boldsymbol{w}_v + \varepsilon\delta_v^-$

38) end

39) end

40) end

41) end

42) until convergence

43)return $Z=\{\boldsymbol{z}_u\}, W=\{\boldsymbol{w}_u\}, X=\{\boldsymbol{x}_u\}, Y=\{\boldsymbol{y}_i\}$

影响空间中 2 个用户向量之间的距离越近，表示用户 v 越有可能受用户 u 的影响而传播此片段上的信息，兴趣空间中用户向量和传播项向量之间的距离越近，表示用户 v 越有可能因为对传播项 i 感兴趣而传播此片段上的信息。所以构建函数 f 和 g 时既要考虑概率的特征，又要充分利用向量的距离公式。但是函数 f 和 g 仍有多种选择方式，本文通过多次测试最终采用 *sigmoid* 函数，它不仅连续平滑，还可以很好地返回[0,1]的真实值，作为本模型所求的 2 种概率值。具体计算式如下所示。

$$f(z_u, w_v) = \frac{1}{1+\exp(\sum_{j=0}^{d-1}(z_u^{(j)} - w_v^{(j)})^2)} \tag{5}$$

$$g(x_v, y_i) = \frac{1}{1+\exp(\sum_{j=0}^{d-1}(x_v^{(j)} - y_i^{(j)})^2)} \tag{6}$$

4.2 无拓扑结构的影响—兴趣传播模型 NT-II 学习算法

本文观察一个特定的传播片段，其传播概率如式(7)所示，也就是用户 v 接收传播轨迹 D 中事件的概率[4]为

$$P(D) = \prod_{v\in D(\infty)} P_v^D \prod_{v\in \overline{D}(\infty)} (1-P_v^D) \tag{7}$$

其中，$P_v{}^D$ 表示在传播轨迹 D 中用户 v 变成活跃用户的概率。实际上，本文所研究模型是一个概率模型，本文通常通过对数似然函数来求解模型的参数，也就是似然函数的参数估计问题，通过对相关概念的了解后，本文定义在传播轨迹集合 $\Omega=\{D_1, D_2,\cdots,D_K\}$ 上的对数似然函数，如下所示。

$$\begin{aligned} L(P;\Omega) &= \sum_{D\in\Omega}(\sum_{v\in D(\infty)} \mathrm{lb} P_v^D + \sum_{v\in\overline{D}(\infty)} \mathrm{lb}(1-P_v^D)) \\ &= \sum_{D\in\Omega}\{\sum_{v\in D(\infty)} \mathrm{lb}(1-\prod_{u\in D(t_D(v))}(1-P(v|u,i_D))) + \\ &\quad \sum_{v\in\overline{D}(\infty)}\sum_{u\in D(\infty)} \mathrm{lb}(1-P(v|u,i_D))\} \end{aligned} \tag{8}$$

结合式(1)~式(6)，式(8)是关于参数 $\boldsymbol{z}_u$、$\boldsymbol{w}_v$、$\boldsymbol{x}_v$、$\boldsymbol{y}_i$ 的似然函数。所以最终优化目标是如何在影响空间(Z，W)和兴趣空间(X，Y)选择参数 $Z=\{\boldsymbol{z}_u\}$、$W=\{\boldsymbol{w}_v\}$、$X=\{\boldsymbol{x}_v\}$、$Y=\{\boldsymbol{y}_i\}$ 使传播轨迹集合上的对数似然 $L(P;\Omega)$ 最大。

具体如算法 1 所示，使用随机梯度上升法更新模型参数，使似然函数达到局部最大。传播概率依赖于参数在连续空间的相对位置。也就是说向量 $\boldsymbol{z}_u$ 和向量 $\boldsymbol{w}_v$ 的距离越近，用户影响空间的用户 v 就越有可能传播该片段的信息；同样，向量 $\boldsymbol{x}_v$ 和向量 $\boldsymbol{y}_i$ 的距离越近，用户兴趣空间的用户 v 就越有可能传播该片段的信息。算法的更新过程分为 2 个阶段。1) 先固定影响空间向量，更新兴趣空间向量（10)~25)行）；2) 固定兴趣空间向量，更新影响空间向量（26)~41)行）。这 2 个阶段交替进行，直到似然函数收敛为止。

以下对伪代码进行简单说明。

1) 第 10)行：从 Ω 中取出一条传播轨迹 D；

2) 第 11)行：v 为所在传播轨迹中还未被之前用户激活的点。

3) 第 12)行：如果在 D 中 v 被激活了，需要计

算 v 的 P_v^D 和 $P_{u,v}$ 的预测值。

4) 第 14)~23)行：梯度上升更新影响空间向量参数集合 Z 和 W

5) 第 30)~39)行：梯度上升更新兴趣空间向量参数集合 X 和 Y

6) 第 43)行：得到最后的参数结合，并返回，模型构建完成。

7) 代码中所涉及的 8 个导数的求解过程如下。以求解 lb(1- $P(v|u, i_D)$)的导数为例。

$$\delta_u^- = \frac{\partial \mathrm{lb}(1-P(v|u,i_D))}{\partial \boldsymbol{z}_u}$$

$$= \frac{-1}{1-P(v|u,i_D)} \cdot \frac{\partial P(v|u,i_D)}{\partial \boldsymbol{z}_u}$$

$$= \frac{-1}{1-P(v|u,i_D)} \cdot \frac{\partial(\alpha P_{u,v} + (1-\alpha)P_{v,i})}{\partial \boldsymbol{z}_u}$$

$$= \frac{-1}{1-P(v|u,i_D)} \cdot \frac{\partial(\alpha P_{u,v})}{\partial \boldsymbol{z}_u}$$

$$= \frac{-\alpha}{1-P(v|u,i_D)} \cdot \frac{\partial(\frac{1}{1+\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2)})}{\partial \boldsymbol{z}_u}$$

$$= \frac{-\alpha}{1-P(v|u,i_D)} \cdot \frac{-\exp(\|\boldsymbol{z}_u - w_v\|^2)}{(1+\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2))^2} \cdot \frac{\partial \|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2}{\partial \boldsymbol{z}_u}$$

$$= \frac{-\alpha}{1-P(v|u,i_D)} \cdot \frac{-\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2)}{(1+\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2))^2} \cdot 2(\boldsymbol{z}_u - \boldsymbol{w}_v)$$

算法 1 中所有导数的结果如下。

1) $$\delta_u^+ = \frac{\partial \log P(v|u,i_D)}{\partial \boldsymbol{z}_u}$$

$$= \frac{\alpha}{P(v|u,i_D)} \cdot \frac{-\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2)}{(1+\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2))^2} \cdot 2(\boldsymbol{z}_u - \boldsymbol{w}_v)$$

2) $$\delta_v^+ = \frac{\partial \log(P(v|u,i_D))}{\partial \boldsymbol{w}_v}$$

$$= \frac{\alpha}{1-P(v|u,i_D)} \cdot \frac{\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2)}{(1+\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2))^2} \cdot 2(\boldsymbol{z}_u - \boldsymbol{w}_v)$$

3) $$\lambda_v^+ = \frac{\partial \log P(v|u,i_D)}{\partial x_v}$$

$$= \frac{1-\alpha}{1-P(v|u,i_D)} \cdot \frac{-\exp(\|\boldsymbol{x}_v - \boldsymbol{y}_i\|^2)}{(1+\exp(\|\boldsymbol{x}_v - \boldsymbol{y}_i\|^2))^2} \cdot 2(x_v - y_i)$$

4) $$\lambda_i^+ = \frac{\partial \log(P(v|u,i_D))}{\partial y_i}$$

$$= \frac{1-\alpha}{P(v|u,i_D)} \cdot \frac{\exp(\|\boldsymbol{x}_v - \boldsymbol{y}_i\|^2)}{(1+\exp(\|\boldsymbol{x}_v - \boldsymbol{y}_i\|^2))^2} \cdot 2(\boldsymbol{x}_v - \boldsymbol{y}_i)$$

5) $$\delta_u^- = \frac{\partial \mathrm{lb}(1-P(v|u,i_D))}{\partial \boldsymbol{z}_u}$$

$$= \frac{-\alpha}{1-P(v|u,i_D)} \cdot \frac{-\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2)}{(1+\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2))^2} \cdot 2(\boldsymbol{z}_u - \boldsymbol{w}_v)$$

6) $$\delta_v^- = \frac{\partial \mathrm{lb}(1-P(v|u,i_D))}{\partial \boldsymbol{w}_v}$$

$$= \frac{\alpha}{1-P(v|u,i_D)} \cdot \frac{-\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2)}{(1+\exp(\|\boldsymbol{z}_u - \boldsymbol{w}_v\|^2))^2} \cdot 2(\boldsymbol{z}_u - \boldsymbol{w}_v)$$

7) $$\lambda_v^- = \frac{\partial \mathrm{lb}(1-P(v|u,i_D))}{\partial \boldsymbol{x}_v}$$

$$= \frac{\alpha-1}{1-P(v|u,i_D)} \cdot \frac{-\exp(\|\boldsymbol{x}_v - \boldsymbol{y}_i\|^2)}{(1+\exp(\|\boldsymbol{x}_v - \boldsymbol{y}_i\|^2))^2} \cdot 2(\boldsymbol{x}_v - \boldsymbol{y}_i)$$

8) $$\lambda_i^- = \frac{\partial \mathrm{lb}(1-P(v|u,i_D))}{\partial \boldsymbol{y}_i}$$

$$= \frac{\alpha-1}{1-P(v|u,i_D)} \cdot \frac{\exp(\|\boldsymbol{x}_v - \boldsymbol{y}_i\|^2)}{(1+\exp(\|\boldsymbol{x}_v - \boldsymbol{y}_i\|^2))^2} \cdot 2(\boldsymbol{x}_v - \boldsymbol{y}_i)$$

本文提到在测试集中，还存在着一些新的传播项，为了使本文的模型更适用实际情况，还需学习测试集中新的传播项的特征向量集合。假设一些用户传播的是新的传播项 h，新传播项的特征向量为 $\boldsymbol{y}_h$，而本文在算法 1 中得到这些用户传播项向量 $\boldsymbol{y}_i$，直觉上会与真实的 $\boldsymbol{y}_h$ 接近，为此，为了求得 $\boldsymbol{y}_h$，需要将下面的目标函数最小化。具体算法如算法 2 所示（n 为传播项个数）。

$$\arg\min \boldsymbol{y}_h \sum_i^n (1-\overline{\boldsymbol{y}}_i \overline{\boldsymbol{y}}_h)^2 \tag{9}$$

算法 2 求新传播项的特征向量的学习算法

输入 旧传播项的特征向量$\overline{\boldsymbol{y}}_i$,学习步长$\varepsilon$

输出 新传播项的特征向量$\boldsymbol{y}_h$

1) init $\boldsymbol{y}_h$

2) repeat

$$\boldsymbol{y}_h = \boldsymbol{y}_h + 2\,\varepsilon \boldsymbol{y}_h \sum_i^n (1-\overline{\boldsymbol{y}}_i \overline{\boldsymbol{y}}_h)^2$$

3) until convergence

5 实验与结果

5.1 数据集

对于本文的实验，本文在网上下载了 2 种数据集。预处理后的数据信息如表 1 所示。

表 1　　数据集信息

数据集	用户数	时间段	训练集传播片段	测试集传播片段
Digg	4 916	159	400	100
Flixster	109 816	155	800	200

1) Digg:中文称为“掘客网”，是一种由用户推动的新闻网站，在网站中，大量用户共同撰写、评论和提交来自网络各个角落的新闻故事，用户还可以对这些文字，视频进行收藏和转播。这里，本文使用 2016 年 11 月完整的 Digg 历史数据，作为一个数据集。

2) Flixster：这是一个电影社交网站，可以让用户分享电影的评分，讨论新的电影，也可以通过电影认

5.2　对比模型与评价指标

本文提出无拓扑结构的影响—兴趣传播模型来预测社交网信息传播，为了证明本文预测方法的有效性，本文介绍另外 2 个预测结果较好的传播模型，并使用和本文相同的数据集合，进行对比。

1) IC 模型：传统的独立级联模型，它可以用特定的数据集合预测出较好的传播概率，但它并不适用于无拓扑结构社交网络数据集合。因为在预测信息传播过程之前，IC 模型首先要根据数据推测图结构，而其推测的结果存在很大的不准确性。

2) Embedded IC：一种嵌入版本的独立级联模型，充分考虑用户之间的相互影响，推测传播概率，把用户嵌入隐藏的投影空间中，借助 EM 算法中 Q 函数的形式求出发送方和接收方的用户参数，进一步构建模型。该模型已经证明其性能优于 CTIC[14]模型、NetTate[15]模型，故本文不再与其比较。

本文通过以下几个评估指标来分析模型预测信息传播效果的准确性。

1) MSE。均方误差，通过测量值误差的平方和的平均数来统计整体误差。值越小，越准确。假设每个用户传播概率 $P(v\,|u,\,i)$为实际值为 P_i，预测值为 $\overline{P_i}$，第 i 个样本实际值和真实值的误差为 E_i，定义如下。

$$MSE=\frac{1}{N}\sum_{i=1}^{n}E_i^2=\frac{1}{N}\sum_{i=1}^{n}(P_i-\overline{P_i})^2$$

2) 准确度（accuracy）。对于给定的测试数据集，反映了模型对整个样本的判定能力。

$$A=\frac{TP+TN}{TP+FP+TN+FN}$$

其中，TP 为预测为正样本，实际也为正样本的特征数；FP 为预测为正样本，实际为负样本的特征数；TN 为预测为负样本，实际也为负样本的特征数；FN 为预测为负样本，实际为正样本的特征数

3) 召回率（recall）。对于给定的测试数据集，反映了被正确判定的正确传播路径所占总传播路径的比重。

$$R=\frac{TP}{TP+FN}$$

4) ROC 曲线。在 ROC 空间中，每个点的横坐标是 FPR，代表将负例错分为正例的概率，纵坐标是 TPR，代表能将正例分对的概率。曲线下面积越大，模型的学习效果越好。

5.3　维数 d 和权重 α 的选择

通过各种评价指标对多种模型的有效性分析对比之前，还需明确实验空间维数。

算法 1 中，本文已经说明实验是在 d 维空间中进行取值和计算的，d 的取值对实验结果也存在影响，本文需要找到最适合的维度来提高模型的精准度，本文将权重 α 的取值先固定 0.5。图 1、图 2 说明在对数似然函数中，不同维数的选择对于时间和在测试片段上平均对数似然值的影响，可以观察到随着维数的增长，实验的收敛时间增长迅速，这是因为特征向量的每一维都是要求的参数，维数越多，所进行的计算就越多，耗费的时间也更多，所以维数与收敛时间成正比。关于维数对本文所构建的平均对数似然函数的影响，该函数概率都是有用户特征向量和传播项特征向量作为 sigmoid 函数的变量与计算实现的，在求解过程中，向量的维数越多，也就代表着选择的特征越多，对概率的预测会相对准确一些，所以平均对数似然函数的绝对值会随着维数的增多逐渐变小，效率会越来越好些。结合对时间和效率的考虑，选取 d=20 作为空间维数，这时模型的运行时间和结果精确度都较好。

维数 d 固定后，对于权重 α 的选择如图 3 所示，当 α=0.65 时，平均似然函数的值较高，实验结果的准确性也教高。所以式(1)最终形式为

$$P(v\,|\,u,i)=0.65P_{u,v}+0.35P_{v,i}$$

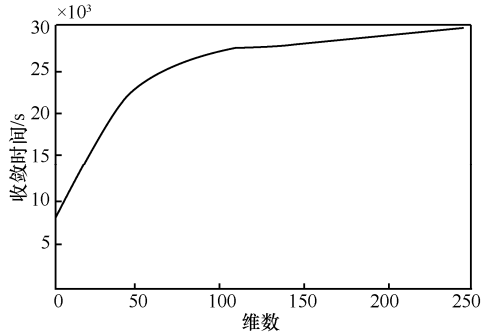

图 1　Digg 测试集中维数和时间的关系

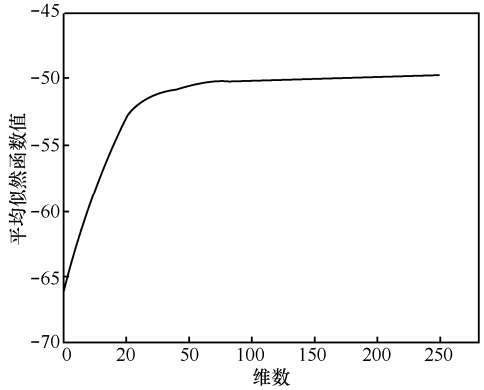

图 2　Digg 测试集中维数和平均似然函数值的关系

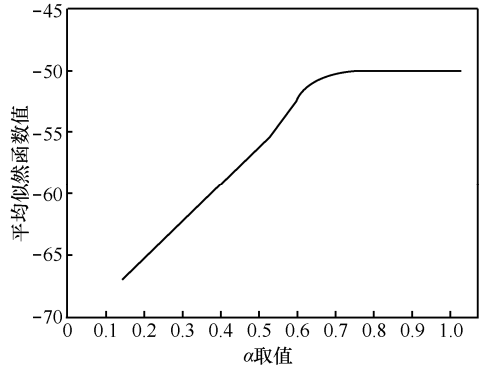

图 3　Digg 测试集中 α 和平均似然函数值的关系

5.4　用户传播概率预测

选择合适的空间维数和权重后，现在可以正式通过之前介绍的一些评价指标，对模型之间效率进行对比分析。表 2 是在多种数据集下根据之前介绍的 *MSE*、模型（model）、准确度（accuracy）、召回率（recall）和 ROC 曲线面积的评估方法对不同模型所预测的传播概率的准确性进行分析以及对比。表格中加粗的数字代表最好的结果。可见，在大部分情况下模型的评估情况都是最佳的。本实验的迭代次数为 30 万次，运行时间为 86 321 s（24 h 左右）。在大部分情况下本文模型的评估情况都是最佳的。这是因为在其他模型中，由于没有用户之间的拓扑关系，模型无法判断谁影响谁，简单来说，当 2 个用户共同喜欢或转发同一传播项时，并不意味着是他们 2 个中的一个影响另一个，但是 IC 模型会错误地认为某些先转发信息的用户就是后转发信息用户的前项，事实上他们可能只是来自同一个“源”，并无太大的影响关系。而 NT-II 模型并不会陷入这样的误区中，因为本文是通过隐藏空间中用户向量之间的距离来判断用户之间的影响关系的，通过计算向量的距离就可以判断出每队用户之间是否有影响关系。这就可以使 MSE 的值降低。

NT-II 的 MAP 值在不同的数据集中相比其他的模型也更好一些。在这个 2 个数据集中有个共同的特点，某一个传播项越受欢迎，就会推荐给越来越多的新用户。而这种现象更容易被本文的模型捕捉。

NT-II 的 F1 值更直观地说明了实验方法的有效性。ROC 曲线面积的值也明确地说明了本模型对于传播概率预测的准确率。

总之，本文的模型对比其他的模型有多种优势：预测的准确率更高；模型的泛化能力更强，在多种数据集上都适用；不依赖于图结构等。对于传播预测的分析使我们明白信息传播是一个复杂的现象，采取不同的模型，不同的数据集对于结果的影响也不同，这也是为什么本文在 2 个数据集中运用多种评价指标来评估模型的有效性。

5.5　用户影响关系预测

在得到社交网中用户的传播概率之后，本文还可以进一步地分析用户的影响关系。简单来说，就是预测某个时间点之前的信息传播结构和未来信息的传播序列。这进一步提高了本文预测方法的能力，当给定传播片段的数据集后，本文还可以反推出用户之间的影响关系（用户的连接，信息的真实转播轨迹）。这样的研究问题同样十分重要。对于预测每一对有连接的用户（u, v），本文学习到的传播概率 $P(v\ |u,\ i)$可以代表他们之间存在连接的可能性。本文可以通过模型以降序的形式为他们排序。

6　结束语

本文提出了一种新的社交网传播模型 NT-II。该模型既考虑了影响空间中传播项发送方用户对接收方用户的影响程度，又考虑了兴趣空间中用户对传播项的偏爱程度。在模型学习过程中，本文还

表 2　　不同模型的有效性对比

数据集	模型	*MSE*	准确率	召回率	ROC 曲线面积
Digg	IC	0.684	0.154	0.877	0.492
	Embedded IC	0.705	0.163	0.899	0.515
	NT-II	**0.546**	**0.163**	**0.900**	**0.537**
Flixster	IC	0.764	0.076	0.908	0.478
	Embedded IC	0.874	0.080	0.954	0.503
	NT- II	**0.644**	**0.079**	**0.970**	**0.505**

考虑了新传播项特征向量对于模型的影响，并给出了新传播项特征向量的学习算法。与现有的传播模型相比，本文可以得到更准确的信息传播预测结果。在今后的研究中，本文拟利用该模型预测信息传播轨迹。

参考文献：

[1] 海沫, 郭庆. 在线社交网络信息传播模型研究[J]. 小型微型计算机系统, 2016, 37(8):1672-1679.
HAI M, GUO Q. Research on online social network information transmission model[J]. Small Microcomputer System, 2016, 37(8): 1672-1679.

[2] GOMEZ RODRIGUEZ M, LESKOVEC J, LKOPF B. Structure and dynamics of information pathways in online media[C]// ACM International Conference on Web Search and Data Mining. ACM, 2013: 23-32.

[3] YANG J, LESKOVEC J. Modeling information diffusion in implicit networks[C]// ICDM 2010, the IEEE International Conference on Data Mining. 2011: 599-608.

[4] BOURIGAULT S, LAMPRIER S, GALLINARI P. Representation learning for information diffusion through social networks: an embedded cascade model[C]// ACM International Conference on Web Search and Data Mining. 2016:573-582.

[5] FENG S, LI X, ZENG Y, et al. Personalized ranking metric embedding for next new POI recommendation[C]// International Conference on Artificial Intelligence. 2015:2069-2075.

[6] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(8):1798-1828.

[7] SAITO K, NAKANO R, KIMURA M. Prediction of information diffusion probabilities for independent cascade model[C]// International Conference on Knowledge-Based Intelligent Information and Engineering Systems. Springer-Verlag, 2008:67-75.

[8] SAITO K, OHARA K, YAMAGISHI Y, et al. Learning diffusion probability based on node attributes in social networks[C]// Foundations of Intelligent Systems, International Symposium. 2011. 2011: 153-162.

[9] GOMEZ R M, LESKOVEC J, LKOPF B. Structure and dynamics of information pathways in online media[C]//ACM International Conference on Web Search and Data Mining. 2013:23-32.

[10] LAGNIER C, DENOYER L, GAUSSIER E, et al. Predicting information diffusion in social networks using content and user's profiles[J]. Lecture Notes in Computer Science, 2016, 7814:74-85.

[11] GOMEZ-RODRIGUEZ M, LESKOVEC J, KRAUSE A. Inferring networks of diffusion and Influence[C]//ACM Knowledge Discovery and Data Mining. 2011:1019-1028.

[12] SIMMA A, JORDAN M I. Modeling events with cascades of poisson processes[J]. Computer Science Learning, 2012:546-555.

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

[14] SAITO K, KIMURA M, OHARA K, et al. Learning continuous-time information diffusion model for social behavioral data analysis[C]// Asian Conference on Machine Learning: Advances in Machine Learning. Springer-Verlag. 2009: 322-337.

[15] RODRIGUEZ M G, BALDUZZI D, SCHÖLKOPF B. Uncovering the temporal dynamics of diffusion networks[C]// International Conference on Machine Learning. 2011:561-568.

作者简介：

王瑞（1993-），女，黑龙江绥滨人，黑龙江大学硕士生，主要研究方向为社交网络分析。

刘勇（1975-），男，河北昌黎人，博士，黑龙江大学副教授，主要研究方向为数据挖掘和社交网络分析。

朱敬华（1976-），女，黑龙江齐齐哈尔人，博士，黑龙江大学教授，主要研究方向为传感器网络与数据挖掘。

玄萍（1979-），女，黑龙江五常人，博士，黑龙江大学教授，主要研究方向为机器学习和生物信息学。

李金宝（1969-），男，黑龙江庆安人，博士，黑龙江大学教授，主要研究方向为传感器网络与大数据管理。