

基于霍克斯过程的社交网络用户关系强度模型

于 岩,陈鸿昶,于洪涛

(国家数字交换系统工程技术研究中心,河南郑州 450002)

摘 要: 社交网络节点之间的关系强度建模是研究信息传播、实现推荐服务等社交网络服务的关键. 传统关系强度模型主要研究简单二元关系与静态关系,未考虑用户交互影响及其动态衰减. 本文提出一种基于霍克斯过程的社交网络用户关系强度模型,将用户关系强度视为潜在因子,用户相似性与历史交互行为分别视为潜在因子诱因与表象,并使用霍克斯过程刻画历史交互行为与用户关系强度之间的关系,解决了已有模型未考虑用户历史交互影响及其动态衰减的问题. 采用微博社交网络数据对模型进行的评估表明,本模型可以提高用户关系强度预测精度以及基于关系强度排序 Top- N 邻居节点的覆盖率.

关键词: 社交网络; 霍克斯过程; 关系强度预测; 微博

中图分类号: TP393

文献标识码: A

文章编号: 0372-2112 (2016)06-1362-07

电子学报 URL: <http://www.ejournal.org.cn>

DOI: 10.3969/j.issn.0372-2112.2016.06.015

A Social Networks User Relationship Strength Model Based on Hawkes Process

YU Yan, CHEN Hong-chang, YU Hong-tao

(China National Digital Switching System Engineering & Technological R&D Center, Zhengzhou, Henan 450002, China)

Abstract: The relationship strength model between social network nodes is the key of social networks service such as information dissemination researches and recommendation system. The traditional researches focus on modeling simple binary relations and static relations, without considering dynamic attenuation of user interaction effects. Aiming at this problem, this paper proposes a social networks user relationship strength model based on Hawkes process (HP-URS), which takes the relationship strength, similarity and history interaction behavior between users as a latent factor, latent factor incentive and presentation respectively. This model uses Hawkes process to characterize relationship between history interaction behavior and user relationship strength. This model provides a solution of the disadvantages of the original model without considering user history interaction effects and their attenuation. This paper uses the data from microblog social networks evaluating HP-URS model, and the experimental results show that this model can improve relationship strength prediction accuracy and coverage rate of the Top- N neighbor nodes based on relationship strength.

Key words: social networks; Hawkes process; relationship strength prediction; microblog

1 引言

人际关系将人们现实生活连成一个真实的社交网络,同时也限定着我们每天信息交互范围. 而在线社交网络中的朋友关系、关注关系等也限定着我们信息交互对象. 现有研究表明利用用户关系强度可以提高对用户社交结构和用户行为预测效果. 文献[1,2]研究用户属性相似性与交互相似性,对用户之间的关系进行建模,并在真实网络中进行推荐服务实践,获得了较高

的推荐效率与用户满意度. 文献[3]提出一种潜在变量模型,将用户之间的关系强度视为引起交互行为的潜在变量,利用概率生成模型与判别式模型得出交互行为与用户关系强度之间的数学关系,并在 Facebook 与 LinkedIn 数据集进行算法评估,取得了较高的分类准确率与自相关系数. 文献[4]提出一种隐含变量混合模型,检测用户之间的弱关系强度,并分析关系强度对网络结构以及信息传播的影响.

但是,现有的研究主要集中二元关系^[5,6](如:是否

是朋友关系,关系强弱等等)和静态交互^[2,7](未考虑历史交互事件影响).简单的二元关系模型已不能较好满足针对社交网络的应用:商品与服务的推荐、诈骗的防范与追踪、谣言的预防与管控等.这些社交网络应用都需要准确、动态地量化用户之间关系强度,以便保证信息传播效果的准确性与时效性.

在社交网络中,用户的一些属性信息与交互信息是公开的,为研究用户之间的关系强度提供了便利.本文提出一种基于霍克斯过程社交网络用户关系强度模型,以便更为准确地预测用户关系强度.首先,模型将用户之间的关系强度视为潜在因子,用户之间的相似性与历史交互事件是潜在因子的决定因素,而用户之间发生的交互事件是潜在因子的表象;其次,利用判别式模型与生成模型定量分析三者之间的数学关系;最后,在利用腾讯微博公开 API 接口爬取的数据进行了模型的评估.

2 霍克斯模型分析

该部分主要介绍霍克斯模型应用范围,阐述其物理含义以及解释其对用户关系强度的适应性,并分析用户关系强度的影响因素,提出计算用户关系强度公式.

2.1 霍克斯模型

霍克斯模型是一种特殊的线性自激模型,1972年由 Hawkes 提出一种特殊的线性自激模型^[8].在该模型提出后被广泛应用于经济分析预测领域^[9,10]以及地震预测^[12]等等,并取得良好的应用效果.近年来,霍克斯过程逐渐在社交网络中应用、发展,文献[12]中利用霍克斯过程对用户之间的关系、网络结构进行研究分析,通过实验验证,都取得了较好的实验效果.目前,求解霍克斯过程最常用的方法为最大似然法^[13,14],本文也采用该方法对问题进行求解.本文将霍克斯过程引入到社交网络分析中,具体公式如下,式(1)为线性自激过程,其中 N 为点过程, $v(t)$ 为自激过程的核函数, $\lambda_0(t)$ 为该自激过程的基础强度, t_i 为过去事件发生的时间;式(2)为霍克斯过程函数,即将核函数确定为 $v(t)$

$$= \sum_{j=1}^P \alpha_j e^{-\beta_j t} \mathbf{1}_{R_i}, \text{ 且因为双重求和序列皆为有限项,故求和号可调换.}$$

$$\begin{aligned} \lambda(t) &= \lambda_0(t) + \int_{-\infty}^t v(t-s) dN_s \\ &= \lambda_0(t) + \sum_{t_i < t} v(t-t_i) \end{aligned} \quad (1)$$

$$\begin{aligned} \lambda(t) &= \lambda_0(t) + \int_0^t \sum_{j=1}^P \alpha_j e^{-\beta_j(t-s)} dN_s \\ &= \lambda_0(t) + \sum_{t_i < t} \sum_{j=1}^P \alpha_j e^{-\beta_j(t-t_i)} \end{aligned} \quad (2)$$

霍克斯过程可以表示为如图 1 所示.

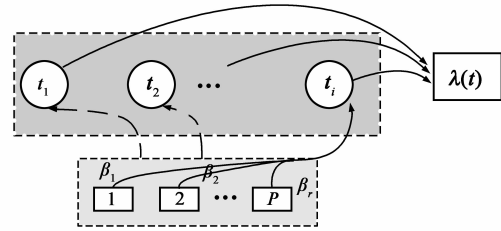


图1 霍克斯过程

其中序列 $\{t_i\}$ 为顺序事件集合,序列中每一项是一个事件集合,由 P 个事件组成.

2.2 霍克斯模型的适应性分析

由式(2)可知,霍克斯过程是 I 维点过程经过线性叠加构成的一种特殊的线性自激过程,其物理含义可以理解为:考察当前事件(点过程)的发生概率(强度),认为当前事件发生概率受历史该类事件的影响,因而可以通过建立历史事件与当前事件的某种关系,进而预测当前事件发生的概率(强度).

对于本文研究的用户关系强度,可理解为用户之间历史交互强度与属性信息相似度的加权和,概率形式表示为式(3),其中 a, b 分别为用户 i 与 j 的交互强度和其属性相似度的权重.

$$\begin{aligned} p(z(i,j)) &= abp(y(i,j))p(s(i,j)) \\ &\propto p(y(i,j))p(s(i,j)) \end{aligned} \quad (3)$$

用户之间的发生交互行为可以理解成点过程,交互强度即为点过程发生强度,故使用霍克斯过程刻画用户关系强度模型,进行关系强度的预测以及后续的应用.

3 基于霍克斯模型的用户关系强度模型

在社会学中,存在同质相似性原理,在社交网络中该原理同样适用,本模型正是基于这样的假设.该模型包括三部分:属性相似度、用户交互行为与用户关系强度.

3.1 用户相似度

用户之间的相似度可以分成文本相似度与背景相似度.

基于同质相似性的假设,得出以下两个结论:

(1) 如果两个用户具有相似的属性信息,那么他们具有较强的关系强度,且属性信息越相似,属于同一“圈子”的概率越大,关系强度越强;

(2) 如果两个用户发布相似的微博内容,那么他们具有较强的关系强度,且内容越相似,关系强度越强;

基于以上两条结论,本文使用文本相似与属性相似作为影响用户关系强度的两个因素.其中用户的属性信息相对稳定,而微博内容主题变化较大.

3.2 用户交互行为

用户之间的交互行为可分为转发、回复与提及。

用户之间的交互行为是用户关系强度的表象,用户之间的关系强度越大,用户之间的交互行为越频繁.同时,用户的历史交互行为也增强了用户之间的关系强度.故用户之间的交互行为即是用户关系强度的表象,也是影响用户关系强度的重要因素.

用户之间交互行为不同于相对静态的属性信息,与文本相似同属于动态影响因素,故本文将微博内容的相似也归为用户交互行为的一种,同样使用霍克斯过程刻画其与用户关系强度之间的关系.

3.3 用户关系强度模型构建

文献[3]中提出的一种潜在因子关系强度模型(Latent Factor Relationship Strength Model, LRS)(图2右半部分),预测结果具有较高的自相关性与分类准确性,但是其未考虑用户历史交互行为对用户关系强度的影响,针对这一问题本文提出一种基于霍克斯过程的社交网络用户关系强度模型(A Social Networks User Relationship Strength Model Based on Hawkes Process, HP-URS),将影响用户之间关系强度的因素增加为两类:历史交互事件与属性相似度;模型使用霍克斯过程刻画历史交互事件对当前研究用户之间的关系强度的影响,高斯过程刻画属性相似度对用户之间关系强度的影响,并将历史事件与用户相似度看成相互独立影响用户关系强度的两个因素;将历史事件按时间段提取,而不是按传统方式,不区分历史时间内的事件,以此将时间处理粒度变小,增强模型的实时评估效果.

根据3.1节、3.2节的分析,将影响用户关系强度的因素分成两类,即用户之间的属性相似度与历史交互事件.故模型结构表示为图2所示.

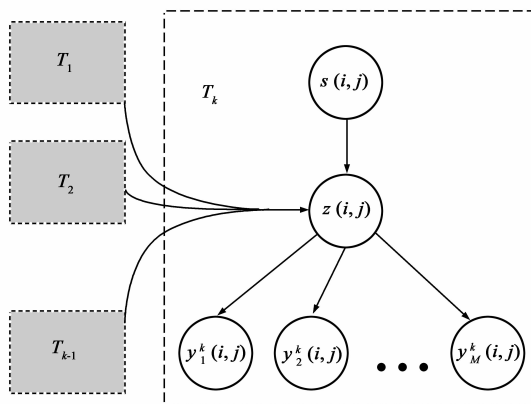


图2 关系强度模型结构图

其中 $s(i, j)$ 、 $z(i, j)$ 与 y^n 的集合 $\{y_m^n(i, j)\}$ (m 为特定交互事件, $m=1, \dots, M$; n 表示考察时间段, $n=1, \dots, k-1$) 分别表示用户 i 与 j 之间的属性相似度、关系强度

与在时间 n 的交互行为, T_1 至 T_{k-1} 部分表示 T_1 至 T_{k-1} 阶段的历史交互事件.

关系强度与交互行为之间表示为生成模型 $p(z(i, j), y(i, j))$. 故联合概率密度可写成

$$\begin{aligned} p(z(i, j), y^k(i, j) | s(i, j), \sum_{n=1}^{k-1} y^n(i, j)) \\ = p(z = z_1 + z_2 | s(i, j), \sum_{n=1}^{k-1} y^n(i, j)) \\ \cdot \prod_{m=1}^M p(y_m^k(i, j) | z(i, j)) \end{aligned} \quad (4)$$

其中,假设当前用户之间的属性相似性与历史交互事件无关,事实上,用户的属性一般是固定的,并不随着用户之间的交互行为而改变,例如用户住址、教育背景、标签等等.因此可以认为用户之间的属性相似性与历史交互事件无关,故式(4)可以写成

$$\begin{aligned} p(z(i, j), y^k(i, j) | s(i, j), \sum_{n=1}^{k-1} y^n(i, j)) \\ = \int_0^z [p(Z_1 = z_1 | s) p(Z_2 = z - z_1 | \sum_{n=1}^{k-1} y^n) \\ \cdot \prod_{m=1}^M p(y_m^k | z)] dz_1 \end{aligned} \quad (5)$$

在式(5)中,对于用户属性相似度与用户关系强度之间的概率模型采用普遍的条件概率模型:高斯分布,即

$$p(z_1(i, j) | s(i, j)) = N(\omega^T s(i, j), v) \quad (6)$$

其中,为了方便计算令 $v=0.5$, ω 为用户属性中各属性对相似度贡献的比重.

对于用户历史交互行为与用户关系强度之间的概率模型则使用霍克斯过程模型表示,即

$$p(z_2(i, j) | \sum_{n=1}^{k-1} y^n(i, j)) = \lambda_0(k) + \sum_{n=1}^{k-1} \sum_{m=1}^M \alpha_m e^{-\beta_m(k-n)} \quad (7)$$

其中, M 表示用户交互行为种类总数,参数 α_m 表示第 m 项交互行为的权重, β_m 表示第 m 项交互行为的时间调节因子 ($\alpha_m > 0, \beta_m > 0$).

在本模型中,关系强度视为用户之间发生交互行为的诱因.本文采用通用的预测方法多元非线性回归刻画用户交互行为与关系强度之间的关系.由于数据的稀疏性,本文采用二项分布来描述用户的交互行为(发生为1,未发生为0),故 $z(i, j)$ 与集合 $\{y_m^n(i, j)\}$ 之间的关系表示为式(8),

$$p(y_m^k(i, j) = 1 | z(i, j)) = \frac{1}{1 + e^{-\theta_m z(i, j)}} \quad (8)$$

其中, θ_m 表示关系强度对第 m 项交互事件的权重,为待估参数.由于 $\{y_m^k(i, j)\} \in B(n, p)$, 故

$$p(y_m^k(i, j) = 0 | z(i, j)) = \frac{e^{-\theta_m z(i, j)}}{1 + e^{-\theta_m z(i, j)}} \quad (9)$$

为了避免过拟合,本文采用高斯先验 $p(\boldsymbol{\omega}) \propto e^{-\lambda_\omega/2 \cdot \boldsymbol{\omega}^T \boldsymbol{\omega}}$ 与 $p(\theta_m^k) \propto e^{-\lambda_\theta/2 \cdot (\theta_m^k)^2}$ ($\lambda_\omega = \lambda_\theta = 1$), 为了表示方便,使用 z, α_m, β_m, s 表示 $z(i, j), \alpha_m^{(i, j)}, \beta_m^{(i, j)}, s(i, j)$.

在阶段 T_k 的取样空间为 $S = \{(i_1, j_1), (i_2, j_2), \dots, (i_N, j_N)\}$, $Z = \{z(i, j) | (i, j) \in S\}$, 首先得出联合分布为

$$\left(\prod_{(i, j) \in S} [p(z, \mathbf{y}^k | s(i, j), \sum_{n=1}^{k-1} \mathbf{y}^n(i, j), \boldsymbol{\omega}, \boldsymbol{\theta}^k)] \cdot p(\boldsymbol{\omega}) p(\boldsymbol{\theta}^k) \right) \quad (10)$$

$$\begin{aligned} &= \prod_{(i, j) \in S} \int_0^z [p(z_1 | s, \boldsymbol{\omega}) p(z_2 | \sum_{n=1}^{k-1} \mathbf{y}^n(i, j)) \\ &\quad \cdot \prod_{m=1}^M p(y_m^k | z, \theta_m^k) p(\boldsymbol{\omega}) p(\theta_m^k) dz_1] \\ &\propto \prod_{(i, j) \in S} \int_0^z [(e^{-1/(2\sigma)} \cdot (\boldsymbol{\omega}^T \mathbf{s} - z_1)^2} (\lambda_0 + \sum_{n=1}^{k-1} \sum_{m=1}^M \alpha_m^{(i, j)} e^{-\beta_m^{(i, j)}(k-n)}) \\ &\quad \cdot \prod_{m=1}^M \frac{e^{-\theta_m z(1-y_m^k)}}{1 + e^{-\theta_m z(i, j)}}) e^{-1/2 \cdot \lambda_\omega \boldsymbol{\omega}^T \boldsymbol{\omega}} \prod_{m=1}^M e^{-1/2 \cdot \lambda_\theta (\theta_m^k)^2} dz_1] \\ &\approx \prod_{(i, j) \in S} ((-\frac{z^3}{3} + z^2 \boldsymbol{\omega}^T \mathbf{s} + (1 - (\boldsymbol{\omega}^T \mathbf{s})^2) z) (\lambda_0(k) + \\ &\quad \sum_{n=1}^{k-1} \sum_{m=1}^M \alpha_m e^{-\beta_m(k-n)}) \prod_{m=1}^M \frac{e^{-\theta_m z(1-y_m^k)}}{1 + e^{-\theta_m z(i, j)}}) e^{-1/2 \cdot \lambda_\omega \boldsymbol{\omega}^T \boldsymbol{\omega}} \prod_{m=1}^M e^{-1/2 \cdot \lambda_\theta (\theta_m^k)^2} \end{aligned}$$

其次, 将 $z(i, j)$ 视为参数, 对式(10)求对数得 $\ell(z(i, j), \boldsymbol{\omega}, \boldsymbol{\theta}^k)$

$$\begin{aligned} &= \sum_{(i, j) \in S} \log(-\frac{z^3}{3} + z^2 \boldsymbol{\omega}^T \mathbf{s} + (1 - (\boldsymbol{\omega}^T \mathbf{s})^2) z) \\ &\quad + \sum_{(i, j) \in S} \log(\lambda_0(k) + \sum_{n=1}^{k-1} \sum_{m=1}^M \alpha_m^{(i, j)} e^{-\beta_m^{(i, j)}(k-n)}) \\ &\quad + \sum_{(i, j) \in S} \sum_{m=1}^M (-\theta_m z(i, j)(1 - y_m^k(i, j)) - \log(1 + e^{-\theta_m z(i, j)})) \\ &\quad - \frac{1}{2} \lambda_\omega \boldsymbol{\omega}^T \boldsymbol{\omega} - \sum_{m=1}^M \frac{1}{2} \lambda_\theta (\theta_m^k)^2 \end{aligned} \quad (11)$$

式(11)的立方项与对数项皆为凹函数, 故函数有极值, 使用最大似然函数方法, 得出如下式

$$\frac{\partial \ell}{\partial z} = \frac{-z^2 + 2\boldsymbol{\omega}^T \mathbf{s} z + 1 - (\boldsymbol{\omega}^T \mathbf{s})^2}{-\frac{z^3}{3} + z^2 \boldsymbol{\omega}^T \mathbf{s} + (1 - (\boldsymbol{\omega}^T \mathbf{s})^2) z} \quad (12)$$

$$\begin{aligned} &+ \sum_{m=1}^M (y_m^k \theta_m - \frac{\theta_m}{1 + e^{-\theta_m z}}) \\ \frac{\partial \ell}{\partial \boldsymbol{\omega}} &= \sum_{(i, j) \in S} \frac{z^2 \mathbf{s} - 2z \mathbf{s}^2 \boldsymbol{\omega}}{-\frac{z^3}{3} + z^2 \boldsymbol{\omega}^T \mathbf{s} + (1 - (\boldsymbol{\omega}^T \mathbf{s})^2) z} - \lambda_\omega \boldsymbol{\omega} \end{aligned} \quad (13)$$

$$\frac{\partial \ell}{\partial \theta_m} = \sum_{(i, j) \in S} (y_m^k - \frac{1}{1 + e^{-\theta_m z}}) z - \lambda_\theta \theta_m \quad (14)$$

$$\frac{\partial \ell}{\partial \alpha_m} = \frac{\sum_{n=1}^{k-1} e^{-\beta_m(k-n)}}{\lambda_0(k) + \sum_{n=1}^{k-1} \sum_{m=1}^M \alpha_m e^{-\beta_m(k-n)}} \quad (15)$$

$$\frac{\partial \ell}{\partial \beta_m} = \frac{\sum_{n=1}^{k-1} (n - k) \alpha_m e^{-\beta_m(k-n)}}{\lambda_0(k) + \sum_{n=1}^{k-1} \sum_{m=1}^M \alpha_m e^{-\beta_m(k-n)}} \quad (16)$$

对于 $z(i, j), \boldsymbol{\omega}, \boldsymbol{\theta}^k$ 直接采用最大似然估计方法, 令式(12) ~ (14) 为 0, 由式(15)、(16) 可以看出, 其无法直接为 0, 故考虑整个研究区间 $[0, T_k]$ 的最大似然估计. 发生强度为 λ 的点过程的 $N(t)$ 的最大似然估计可以写成

$$\ln \ell((N(t))_{t \in [0, T]}) = \int_0^{T_k} (1 - \lambda(s)) ds + \int_0^{T_k} \ln \lambda(s) dN(s) \quad (17)$$

故与此相对应的霍克斯过程的似然函数可以写成

$$\begin{aligned} \ln \ell((t_i)_{i=1}^k) &= t_k - \Lambda(0, t_k) + \sum_{i=1}^k \ln \lambda(t_i) \\ &= t_k - \Lambda(0, t_k) \\ &\quad + \sum_{i=1}^k \ln [\lambda_0(t_i) + \sum_{j=1}^P \sum_{n=1}^{i-1} \alpha_j e^{-\beta_j(t_i - t_n)}] \end{aligned} \quad (18)$$

其中

$$\Lambda(0, t_k) = \sum_{i=0}^k \lambda_0(s) ds + \sum_{j=1}^P \sum_{t_n \leq t_i} \frac{\alpha_j}{\beta_j} (1 - e^{-\beta_j(t_i - t_n)}) \quad (19)$$

假定 λ_0 为常数, 且该霍克斯过程是稳定的 ($\sum_{j=1}^P \frac{\alpha_j}{\beta_j} < 1$)^[9], 参数的最大似然估计需要满足的条件为 (将 $\ln \ell((t_i)_{i=1, \dots, k})$ 记为 $\ln T$)

$$\frac{\partial \ln T}{\partial \alpha_m} = \sum_{i=2}^k \frac{\sum_{n=1}^{i-1} e^{-\beta_m(t_i - t_n)}}{\lambda_0 + \sum_{m=1}^P \sum_{n=1}^{i-1} \alpha_m e^{-\beta_m(t_i - t_n)}} - \sum_{m'=1}^k \frac{1 - e^{-\beta_m(t_i - t_{m'})}}{\beta_m} \quad (20)$$

$$\begin{aligned} \frac{\partial \ln T}{\partial \beta_j} &= \sum_{i=2}^k \frac{\sum_{n=1}^{i-1} \alpha_j (t_n - t_i) e^{-\beta_j(t_i - t_n)}}{\lambda_0 + \sum_{m=1}^P \sum_{n=1}^{i-1} \alpha_m e^{-\beta_m(t_i - t_n)}} \\ &\quad - \sum_{m=1}^k \frac{e^{-\beta_j(t_i - t_{m'})} (1 + \beta_j(t_k - t_m)) - 1}{\beta_j^2} \end{aligned} \quad (21)$$

通过分析数据得出用户之间的交互行为与相似度, 并令上述一阶偏导为 0, 得出各个参数的最大似然估计值 $\hat{z}, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$, 并对 $\hat{z}$ 进行归一化处理, 确定关系强度预测模型, 利用式(6)、(7) 预测用户之间关系强度.

4 仿真实验

4.1 实验环境及数据

4.1.1 实验平台

实验硬件环境: Core (TM) i7 双核处理器, 主频 3.4GHz, 内存 16GB; 实验仿真软件: Matlab2008b, eclipse4.4.

4.1.2 实验数据获取及预处理

2013 年中国社交网络排名中, QQ 空间与腾讯微博用户分列 2、3 位, 总量超过第一的新浪微博. 用户通过该平台发表状态、联系好友、获取信息等等, 是目前活跃度较大的社交网络平台.

本文数据是利用腾讯微博提供的公开 API 接口, 通过网络爬虫获取, 主要步骤是:

- (1) 选取种子节点, 并获取种子节点微博信息;
- (2) 采用广度优先策略获取种子节点队列中节点的微博信息, 并将种子节点的关注列表中的用户作为待获取信息节点, 加入种子节点队列;
- (3) 待获取的节点数达到设定的阈值或者种子节点中无节点.

4.2 实验内容及结果分析

采用腾讯微博作为数据源, 使用 P@N 以折扣累计利润 (Discounted Cumulative Gain, DCG)^[15] 评价指标评价通过 HP-URS 模型得出的基于用户关系强度排序的用户 Top-N 邻居节点序列, 实验表明该序列相比其他方法得到的序列, 具有较高的覆盖率及排序准确率.

4.2.1 实验数据处理

实验数据来源 2014 年 06 月 28 日至 2014 年 10 月 28 日腾讯微博公开数据.

考虑降低实验的复杂度, 本文将分段时间长度定为 5 天, 并且最多考虑 10 段时间内的历史交互事件.

通过对微博网络进行分析, 确定用户之间的属性相似项为所在地、学校、简介、性别、教育信息与标签信息; 用户之间的交互行为包括转发、回复、提及以及微博文本相似性. 将这两种影响用户之间关系强度的因素及计算方法分别表示为表 1 与表 2.

表 1 用户属性相似度以及度量方法

s_1	省、市一致, $s_1 = 1$; 只有省一致 $s_1 = 2/3$, 其他 0
s_2	学校院系一致, $s_2 = 1$; 学校一致, 院系不一致 $s_2 = 2/3$; 其他 $s_2 = 0$
s_3	教育信息一致, $s_3 = 1$; 相差一个等级 $s_3 = 2/3$; 相差两个等级 $s_3 = 1/3$; 其他 0
s_4	性别一致 $s_4 = 1$; 不一致 $s_4 = 0$
s_5	使用余弦距离计算简介文本相似度
s_6	使用余弦距离计算标签文本相似度

本文选取 50 个种子用户节点, 考察其一跳范围内的邻居用户节点, 由于用户关系强度的直接体现为用户之间的交互行为, 故以用户之间的交互频次作为标准, 得出用户邻居节点的有序序列 $S(u)$, 作为基准序列, 将 HP-URS 模型、属性相似度、微博文本相似度以及 LRS 模型^[3] 得出的邻居节点序列作为对比序列, 其中, 属性相似度只考虑用户之间的属性相似性; 微博文本相似度只考虑用户之间的文本相似度, LRS 模型考虑了用户之间的微博文本相似度与属性相似度, 但未考虑用户之间的历史交互行为; HP-URS 考虑以上所述三种因素.

表 2 用户交互行为以及度量方法

y_1^k	k 阶段用户 i, j 用户相互转发 $y_1^k = 1$, 只有一个用户转发 $y_1^k = 1/2$, 否则为 0
y_2^k	k 阶段用户 i, j 用户相互回复 $y_2^k = 1$, 只有一个用户回复 $y_2^k = 1/2$, 否则为 0
y_3^k	k 阶段用户 i, j 用户相互提及 $y_3^k = 1$, 只有一个用户提及 $y_3^k = 1/2$, 否则为 0
y_4^k	y_4^k 为 k 阶段用户 i, j 用户微博文本相似度

为了进一步验证该模型的有效性, 采用[2]中相似性度量模型 (Similarity Measure between microblog Users, SMU) 作为对比算法, 该篇文献将用户关系强度定义为用户之间的相似度, 通过分析用户的各种信息, 最后得出能反映其社交信息的关注信息与粉丝信息作为度量用户关系强度指标效果最好. 本文采用 SMU 作为对比模型, 检验 HP-URS 模型的有效性.

4.2.2 实验结果评价

按照表 1 与表 2 的方法计算用户在各阶段的交互向量以及在所考察阶段的用户与其邻居节点之间的用户相似性, 利用 3.3 节中的最大似然公式计算各参数, 预测用户与邻居节点之间的关系强度, 并使用两种评价指标 P@N、DCG 对实验结果评价.

(1) P@N 评价方法

该评价方法是覆盖率评价指标, 评价对比序列中包含的基准序列中用户数量占基准序列中总用户数量的比重, 采用五种序列作为对比序列, 分别属性相似度序列、微博文本相似度序列、通过 LRS 得到的序列、通过 SMU 模型得到序列以及通过 HP-URS 得到的序列, 结果如图 3 所示 ($N = 10, 20, 30, 40, 50$)

由图 3 可以看出, 基于属性相似度得出的用户关系强度 Top-N 列表的 P@N 值最小, 而基于文本相似度、LRS 与 SMU 模型得出 P@N 值相差不大, HP-URS 的 P@N 值最大. 由除 SMU 曲线外, 四条曲线之间的关系可以得出影响用户关系强度的因素主要是用户之间的交互行为以及用户微博文本之间的相似度; 由 SMU 模型

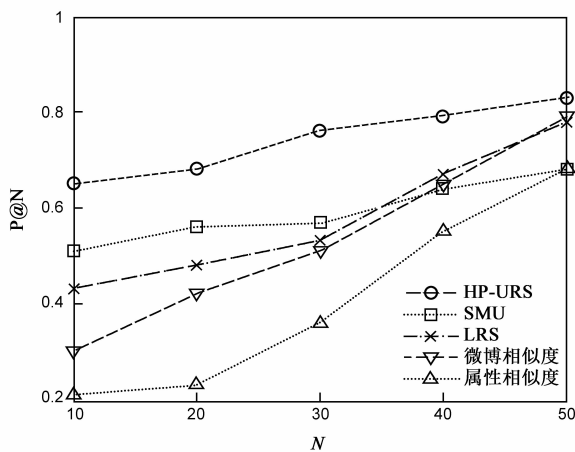


图3 各Top-N序列的P@N曲线

与 HP-URS 模型之间的对比证明与 SMU 模型相比,本模型在覆盖率上具有一定的优势,但稳定性略逊与 SMU 模型.该模型相比于实验表明 HP-URS 模型可以提高基于用户关系强度排序的 Top-N 序列的预测覆盖率.

(2) DCG 评价方法

针对 P@N 评价方法未能考虑对比序列元素之间的排序缺点,引入推荐系统中排序准确率评价方法: DCG,其主要思想是与关系强度用户喜欢的商品被排在推荐列表前面比排在后面会更大程度上增加用户体验,定义为:

$$DCG(b, L) = \sum_{i=1}^b r_i + \sum_{i=b+1}^L \frac{r_i}{\log_b i} \quad (22)$$

式中, r_i 表示排在对比序列第 i 位的邻居节点是否在基准序列中; $r_i = 1$ 表示在; $r_i = 0$ 表示不在; b 是自由参数多设为 2; L 为推荐列表长度.结果如图 4 所示 ($N = 10, 20, 30, 40, 50$)

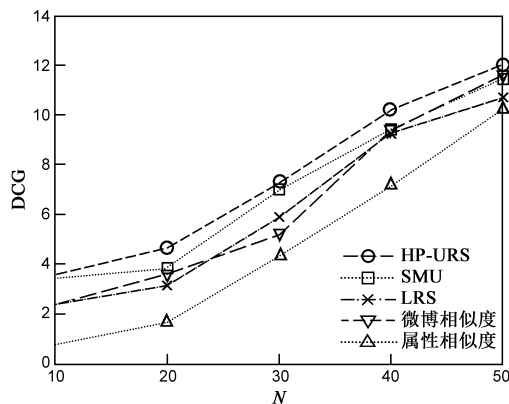


图4 各Top-N序列的DCG曲线

由图 4 可以看出,五种预测序列的 DCG 值排序与 P@N 值排序大体一致,HP-URS 模型与 SMU 模型在预测排序准确性效果最好,原因是在本质上,SMU 模型考虑的核心是用户之间的社交信息,是用户交互行为的一

种体现,但是由于其未考虑随着时间社交关系的变化,无论从覆盖率还是排序准确率上都略逊 HP-URS 模型.进一步证实影响用户关系强度的因素主要是用户之间交互行为与微博文本的相似度,而微博文本的本质即为用户的兴趣.

5 总结

本文利用腾讯微博的数据,提出基于一种基于霍克斯过程社交网络用户关系强度模型,考虑用户之间的历史交互行为、微博文本相似度与属性相似度对用户之间的关系强度进行建模,实验结果表明影响用户之间的关系强度的因素主要是用户的交互行为与微博文本相似度,属性相似度只能在一定程度上表征用户之间的关系强度.实验结果还表明,HP-URS 模型在预测用户关系强度上能提高预测 Top-N 序列覆盖率与排序准确性.

本文中关于高斯分布的参数选择了经典的、具有代表性的值,取得了显著的实验效果,但并不是最优解;模型对数据的完整性与有效性要求较高,鲁棒性与健壮性有待进一步加强.下一步的工作针对以上不足,对模型进一步优化.

参考文献

- [1] 荣辉桂,火生旭,胡春华,等. 基于用户相似度的协同过滤推荐算法[J]. 通信学报,2014,35(2):16-24.
Rong Hui-gui, Huo Sheng-xu, Hu Chun-hua, et al, User similarity based collaborating filtering recommendation algorithm[J]. Journal on Communications, 2014, 35(2): 16-24. (in Chinese)
- [2] 徐志明,李栋,刘挺,等. 微博用户的相似性度量及其应用[J]. 计算机学报,2014,37(1):207-218.
Xu Zhi-ming, Li Dong, Liu Ting, et al, Measuring similarity between microblog users and its application[J]. Chinese Journal of Computers, 2014, 37(1): 207-218. (in Chinese)
- [3] Xiang R, Neville J, Rogati M. Modeling relationship strength in online social networks[A]. Proceedings of the 19th International Conference on World Wide Web[C]. Raleigh, North Carolina, USA: ACM, 2010. 981-990.
- [4] Zhang S. Influence of relationship strengths to network structures in social network[A]. 2014 14th International Symposium on Communications and Information Technologies[C]. Incheon, South Korea: IEEE, 2014. 279-283.
- [5] Gilbert E, Karahalios K. Predicting tie strength with social media[A]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems[C]. Monterey, California, USA: ACM, 2009. 211-220.
- [6] Panovich K, Miller R, Karger D. Tie strength in question &

- answer on social network sites [A]. Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work [C]. Seattle, Washington, USA: ACM, 2012. 1057 – 1066.
- [7] He Y, Zhang C, Ji Y. Principle features for tie strength estimation in micro-blog social network [A]. 2012 IEEE 12th International Conference on Computer and Information Technology [C]. Chengdu, Sichuan, China: IEEE, 2012. 359 – 367.
- [8] Liniger TJ. Multivariate Hawkes Processes [D]. Zurich: Swiss Federal Institute of Technology Zurich, 2009.
- [9] Filimonov V, Bicchetti D, Maystre N, et al. Quantification of the high level of endogeneity and of structural regime shifts in commodity markets [J]. Journal of International Money and Finance, 2014, 42: 174 – 192.
- [10] Bowsher C G. Modelling security market events in continuous time: Intensity based, multivariate point process models [J]. Journal of Econometrics, 2007, 141 (2): 876 – 912.
- [11] Hawkes A G, Adamopoulos L. Cluster models for earthquakes: Regional comparisons [J]. Journal of the International Association for Mathematical Geology, 1976, 8 (4): 463 – 475.
- [12] Juncen L, Sheng G, Yu Z, et al. Inferring links in cascade through Hawkes process based diffusion model [A]. 2014 4th IEEE International Conference on Network Infrastructure and Digital Content [C]. Beijing, China: IEEE, 2014. 471 – 475.
- [13] Ozaki T. Maximum likelihood estimation of Hawkes' self-exciting point processes [J]. Annals of the Institute of Statistical Mathematics, 1979, 31 (1): 145 – 155.
- [14] Guo C, Luk W. Accelerating maximum likelihood estimation for Hawkes point processes [A]. 2013 23rd International Conference on Field Programmable Logic and Applications [C]. Porto, Portugal: IEEE, 2013. 1 – 6.
- [15] 朱郁筱, 吕琳媛. 推荐系统评价指标综述 [J]. 电子科技大学学报, 2012, 41 (2): 163 – 175.
- Zhu Yu-xiao, Lv Lin-yuan. Evaluation metrics for recommender systems [J]. Journal of Electronic Science and Technology. 2012, 41 (2): 163 – 175. (in Chinese)

作者简介



于 岩 男, 1989 年 6 月生于吉林白城. 现为国家数字交换系统工程技术研究中心硕士研究生. 主要研究方向为数据挖掘与电信网信息关防.

E-mail: yuynil@gmail.com



陈鸿昶 男, 1964 年 4 月生于河南新密. 现为国家数字交换系统工程技术研究中心教授、博士生导师. 主要研究方向为电信网信息关防与信息通信安全.

E-mail: chenhongchang@ndsc.com.cn