

在线社交网络用户间影响量化的贝叶斯模型

戴云晶, 邓倩妮

(上海交通大学 电子信息与电气工程学院, 上海 200240)

摘 要: 传统的模型专注于合作网络等社交网络, 并不完全适用于在线社交网络. 而现有的针对在线社交网络的模型简单, 片面, 缺乏准确性. 因此我们提出了基于贝叶斯观点的模型, 通过集成在线社交网络提供的有用信息来量化用户间的影响. 通过分析对比贝叶斯模型与现有模型找到的影响者, 我们发现贝叶斯模型比现有的模型更准确和全面.

关键词: 在线社交网络; 贝叶斯模型; 影响者; 用户间影响量化

中图分类号: TP302

文献标识码: A

文章编号: 1000-7180(2013)03-0106-04

Pairwise Influence Quantification Bayes Model for Online Social Network

DAI Yun-jing, DENG Qian-ni

(School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: Traditional researches of pairwise influence quantification are focus on social network such as coauthor network. So they don't fully apply to current online social network. Furthermore, existing model which are proposed for online social network are simple, one-sided and lack of accuracy. So we propose a Bayes model which integrated all useful information online social network provides to quantify the pairwise influence between two users. Through analyzing and comparing the influential found by these models, we find that our model is more accurate and comprehensive than existing models.

Key words: online social network; Bayes model; influentials; pairwise influence quantification

1 引言

在线社交网络的爆炸式增长已经给数百万人提供了创建和分享内容的平台. 其中最显著的应用便是病毒营销, 在病毒营销中, 我们需要找到一些特别的人, 通过他们可以影响更多的人, 他们被称为影响者. 为了找出影响者, 其中最核心的问题是如何量化用户间的影响力. 这个影响力通常是以概率的形式表示. Goyal 等^[1]提出从用户的行为日志中学习用户间的影响概率. Tang 等^[2]则提出了 TAP 模型来构建社交网络的主题相关的影响. Weng 等^[3]进一步提出了吸收时间因素的动态因子图模型来量化用户的社交影响. 但是, 这些研究主要针对合作网络等

传统的社交网络. 可能并不完全适用于当前的在线社交网络. Kwak^[4]等利用推特用户间的关系信息来量化用户间的影响力. Weng^[5]提出了主题敏感的用户间影响量化模型. 然而, 他们的模型基于用户间存在着很高的相互性的假设. 但这个假设已经被许多之后的研究证明是错误的. Zhang 等^[6]则认为用户间的交互信息比关系信息更能够准确地量化用户间的影响力. 但在交互信息不充分时, 该模型存在过度拟合问题. 因此, 我们需要解决的问题是: 在交互信息不充分的情况下, 结合关系信息, 如何使用户间影响量化更准确.

针对这个问题, 文中推导了基于用户间关系信息和用户间交互信息的影响模型, 并指出它们存在

收稿日期: 2012-06-15; 修回日期: 2012-07-18

基金项目: 国家“八六三”计划项目(2009AA012201)

的问题.并针对这些模型存在的问题,提出了基于贝叶斯观点的用户间影响量化模型,通过贝叶斯方法结合用户间的关系信息和交互信息,提高用户间影响量化的准确性.

2 模型

对用户间的影响量化的模型建立在用户和他的所有朋友之间.每个用户是一个节点,每个用户的所有朋友节点都存在一条有向边指向该用户,边的权值就是用户间的影响概率.我们将推导现所有的模型,并提供它们的影响计算公式,其中 p_{ij} 代表用户 i 对用户 j 的影响.

2.1 基于关系信息的用户间影响量化模型

在基于关系信息的用户间影响量化模型中,只考虑用户间的关系信息.我们简单地将边的权值设置为用户朋友数的倒数,假设它为 K :

$$p_{ij} = \frac{1}{K} = \frac{1}{|\text{用户 } j \text{ 的朋友数}|}$$

2.2 基于交互信息的用户间影响量化模型

在基于交互信息的用户间影响量化模型中,只考虑用户间的交互信息.假设用户 j 有 K 个朋友,并使用一个 K 维向量 X 来表示微博转发自哪个朋友.如果用户 j 转发了用户 i 的一条微博,则 $x_{ij} = 1$,并且其它元素都为 0.假设 $x_{ij} = 1$ 的概率为 p_{ij} ,则 X 服从如下的分布:

$$p(x | p) = \prod_{i=1}^K p_{ij}^{x_{ij}}$$

p 满足 $\sum_i p_{ij} = 1$.现在假设我们有一个交互数据集, $D = \{x_1 \dots x_N\}$, 它的每个元素都是一条被观察到的转发微博.从而构造出极大似然函数

$$\begin{aligned} p(D | p) &= \prod_{n=1}^N \prod_{i=1}^K p_{ij}^{x_{ij}^{(n)}} \\ &= \prod_{i=1}^K p_{ij}^{\sum_n x_{ij}^{(n)}} = \prod_{i=1}^K p_{ij}^{m_{ij}} \end{aligned}$$

其中充足统计量 m_{ij} 满足 $m_{ij} = \sum_n x_{ij}^{(n)}$, 它代表了 $x_{ij} = 1$ 的总次数,也就是用户 j 转发用户 i 的微博的总次数.最大化上述似然函数的对数并同时保证 p_{ij} 的和为 1 的限制来量化 p_{ij} 的值.这可以通过一个拉格朗日乘子 λ 实现,最大化下式:

$$\sum_{i=1}^K m_{ij} \ln(p_{ij}) + \lambda (\sum_{i=1}^K p_{ij} - 1)$$

结果将获得概率:

$$p_{ij} = \frac{m_{ij}}{\sum_{i=1}^K m_{ij}}$$

$$= \frac{|\text{用户 } j \text{ 转发用户 } i \text{ 的微博数}|}{|\text{用户 } j \text{ 转发的微博总数}|}$$

2.3 贝叶斯观点

到目前为止,存在着 2 种概率观点.其中之一将概率视为随机、可重复事件发生的频率.称之为经典概率学派或频率派.另一个观点认为概率提供了一种不确定性的量化.称它为贝叶斯观点.

在频率派的观点中,使用似然函数 $p(D | \omega)$ 来表示对于不同假设 ω 被观察到的数据为 D 的可能性.我们选择能最大化似然函数的假设 ω .上述的模型就是基于该观点.当数据量 D 充分的时候,结果假设 ω 将收敛于最终的真实值.因此我们可以获得准确的概率量化.然而,当数据量不充足时,极大似然函数存在着过分拟合的问题.

贝叶斯观点通过引入先验概率,从而克服了这个问题.先验概率通常是经验或常识,当我们不断地获取新的数据后,我们可以用来修正这个先验概率,并将它转化为后验概率.公式如下:

$$p(\omega | D) = \frac{p(D | \omega) \cdot p(\omega)}{p(D)}$$

ω 是我们的假设, $p(\omega)$ 是观察到数据之前的先验概率, $p(D | \omega)$ 是似然函数, $p(\omega | D)$ 是在观察到数据后的后验概率. $p(D)$ 则只是一个正则化系数.

2.4 基于贝叶斯观点的用户间影响量化模型

正如我们所看到的那样,现有模型存在着过度拟合的问题,因而是 inaccurate 的.但我们可以通过使用贝叶斯模型来解决所有的这些问题.贝叶斯模型不但解决了过度拟合问题,同时关系信息和交互信息可以很自然地集成到一起.在贝叶斯的设置中,我们假设用户间的关系信息表示用户间影响概率的先验分布,并使用交互信息将这个先验分布转换成后验分布.

为了方便起见,我们根据共轭先验理论,使用迪拉克分布作为先验分布.该分布的正则化形式如下:

$$Dir(p | \alpha) = \frac{\Gamma(\alpha_{0j})}{\Gamma(\alpha_{1j}) \dots \Gamma(\alpha_{Kj})} \prod_{i=1}^K p_{ij}^{\alpha_{ij}-1}$$

其中 $\Gamma(x)$ 是伽玛函数, $\alpha = (\alpha_1 \dots \alpha_K)$ 是先验参数并且满足 $\alpha_{0j} = \sum_{i=1}^K \alpha_{ij}$.

将先验分布和似然函数相乘,获得后验分布,这同样是一个迪拉克分布:

$$\begin{aligned} p(p | D, \alpha) &= Dir(p | \alpha + m) \\ &= \frac{\Gamma(\alpha_{0j} + \sum_{i=1}^K m_{ij})}{\Gamma(\alpha_{1j} + m_{1j}) \dots \Gamma(\alpha_{Kj} + m_{Kj})} \prod_{i=1}^K p_{ij}^{\alpha_{ij}-1+m_{ij}} \end{aligned}$$

之后可以得到用户间的影响概率:

$$p_{ij} = \frac{m_{ij} + \alpha_{ij}}{\sum_{i=1}^K m_{ij} + \alpha_{0j}}$$

实际上,基于关系信息的用户间影响量化模型和基于交互信息的用户间影响量化模型都是贝叶斯模型的特例.如果将所有的先验参数 α_{ij} 的值都设置为无穷大,则 $p_{ij} = 1/K$.与基于关系信息的用户间影响量化模型相同.如果我们将所有的 α_{ij} 的值都设为 0,则 $p_{ij} = m_{ij} / \sum_{i=1}^K m_{ij}$.与基于交互信息的用户间影响量化模型相同.

3 实验

在完成对所有用户间的影响量化之后,我们需要基于这些影响量化找出影响者.最常用的方法是马尔可夫模型,通过将用户间的影响概率作为转移概率,通过不断的迭代计算用户的全局影响力并进行排序,找出最有影响力的用户排名.为了方便起见,将模型中的排名分别标记为 PR、RR 和 SR.

3.1 数据

本文的数据集来自于新浪微博.整个数据集包含 11285 个用户以及这些用户之间的所有的粉丝关系和转发关系,包括直接(用于用户间影响量化)和间接的(用于模型的有效性评估)转发关系.在对数据集的 11285 个用户的分析中,我们发现只有 5048 个用户(不到一半)转发过至少一条微博,而且其中大部分用户的转发数都很少.这些都证明了实际的交互信息量确实不足,因而现有的使用交互信息的量化模型是存在问题的.

3.2 相关性

为了对这三个模型进行量化的比较,我们将使用 Fagin 等的泛化 Kendall's tau 来计算排名的关联性.在图 1 中,我们绘制了从 Top10 到 Top100 排名的相关系数.我们观察到,PR 及 RR 之间的相关系数非常低,几乎低于 0.4,远低于 PR 与 SR 以及 RR 与 SR 的相关系数.而 SR 与 PR 和 SR 都有较高的相关系数.这些都表明 SR 比 PR 以及 RR 考虑的更全面,也更综合.

到目前为止,我们都固定了先验参数 α 的值.实际上,它是一个非常重要的参数,因为它控制着关系和交互所占的比例.因此,我们通过调整 α 的值来测试它对相关系数的效果.为了实验的方便,我们这里统一设置 α 的值为数值 α ,但实际上它们可以设置为不同的值.图 2 刻画了相关系数与 α 的值之间的关系.当 α 较小时,PR 和 SR 之间的相关系数比 RR

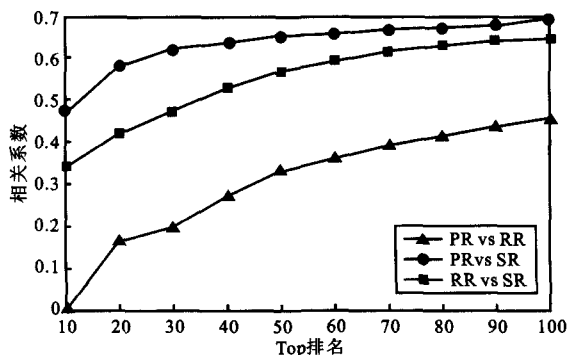


图1 三个排名间的相关系数

和 SR 之间的相关系数大,这表明此时用户间的交互信息比用户间的关系信息贡献更大.反之,当 α 变大时,情况则正好相反,用户间的关系信息占据了更重要的位置.当 α 约等于 0.6 时,我们发现 PR 和 SR 的相关系数与 RR 和 SR 之间的相关系数几乎相同,此时用户间的交互信息与关系信息同等重要.总之,交互信息的重要程度是随着 α 的值单调增加的,反之关系信息的重要性则随着 α 减少而减少.因此,我们可以通过 α 来控制关系信息和交互信息之间的相对重要性.

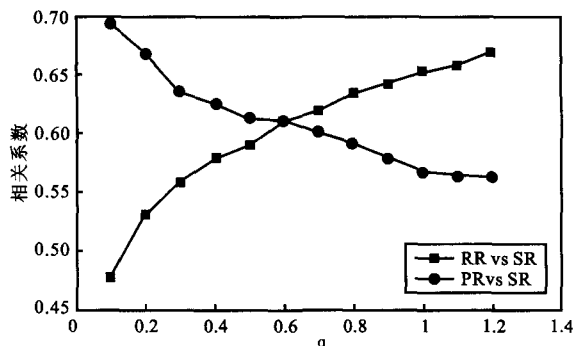


图2 相关系数与先验参数的 α 关系

3.3 有效性

到目前为止,我们的主要定性分析了三种模型的不同以及定量分析了三种模型之间的相关性.但我们仍然需要一个第三方的基准,用来评估这些模型的有效性.因此,我们提出了一个有效性评价方法,我们使用用户的影响力上限来评估用户的影响力.

我们首先根据数据集提供的信息构造两个 $N \times N$ 的矩阵 follow 和 retweet. 矩阵 follow 的元素 $\text{follow}[i][j]$ 等于 1 当且仅当用户 j 是用户 i 的粉丝. 同时矩阵 retweet 的元素 $\text{retweet}[i][j]$ 等于 1 当且仅当用户 j 至少转发过一条原始作者为用户 i 的微博. 在构造完这 2 个矩阵之后,利用这 2 个矩阵构造第三个矩阵, influence, 它的元素 $\text{influence}[i][j]$ 代

表用户 i 是否对用户 j 有影响. 在以下 2 种情景下, 可以认为用户 i 对用户 j 有影响: (1) 用户 j 是用户 i 的粉丝. (2) 存在用户 k 转发了至少一条用户 i 的原始微博并且用户 j 是用户 k 的粉丝. 如果满足上述两个条件中的任意一个, 那么 $\text{influence}[i][j] = 1$, 即用户 i 对用户 j 有影响. 最后, 在构造完 influence 之后, 就可以通过将 influence 矩阵的第 i 行累加来估算出用户 i 的影响力上限了.

当我们得到了用户的影响力上限后, 我们可以根据获得的上限对用户进行排名, 并用来评估这三个模型. 我们提出了三个指标来评估模型的有效性, 它们分别是覆盖率, 总影响力以及相关性. 覆盖度衡量对应模型的排名与基准排名有多少是共同的. 总影响力将排名中的用户的影响力上限累加起来. 相关系数的定义与 3.2 节的定义是一致的. 所有这些指标越高, 对应的模型就越准确. 表 1 列出了基于三个模型的 Top10 排名的指标值. 可以看到, 在所有这三个指标, 我们的模型都优于其他两个模型. 对于覆盖度和相关性, 我们的模型几乎是其他两个模型的两倍. 因此, 可以得出: 基于贝叶斯的用户间影响量化模型比基于关系和交互的用户间影响量化模型更准确.

表 1 三个模型的有效性评估指标值

	覆盖率(%)	总影响力	相关系数
PR	20	97032	0.21
RR	20	94242	0.25
SR	50	102986	0.55

4 结束语

文中研究了在线社交网络中用户间影响的量化问题. 并针对现有模型的不足, 我们提出了基于贝叶斯改进的模型. 我们应用该模型来寻找社交网络中的影响者, 并将结果和其它模型进行比较. 实验结果表明, 我们的模型比现有的模型更准确全面.

然而, 还有很多地方需要进一步研究. 首先, 我们并没有得到如何设置 α 的值的方案. 其次, 我们为了计算的方便, 使用迪拉克分布作为先验分布, 但共轭先验可能不是最优的先验分布. 最后, 在我们的模型中, 没有考虑时间因素的影响. 因此, 在未来

可能将进一步解决这些问题.

参考文献:

[1] Amit Goyal, Francesco Bonchi, Laks V S Lakshmanan. Learning influence probabilities in social networks[C]//proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2010.

[2] Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks[C]//KDD'09. Paris, France: ACM, 2009.

[3] Chi Wang, Jie Tang, Jimeng Sun, et al. Dynamic social influence analysis through time-dependent factor graphs[C]//ASONAM' 11. Kaohsiung, Taiwan: IEEE, 2011.

[4] Kwak H, Lee C, Park H, et al. What is twitter, a social network or a news media? [C]//Proceedings of the 19th International World Wide Web (WWW) Conference. Raleigh, NC, USA: ACM, 2010.

[5] Weng J, Lim E P, Jiang J, et al. Twiterrank: finding topic-sensitive influential twitterers[C]//proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2010.

[6] Meng Zhang, Caihong Sun, Wenhui Liu. Identifying influential users of micro-blogging services: a dynamic action-based network approach[C]//Pacific Asia Conference on Information Systems 2011. Hyderabad, India: IEEE, 2009.

作者简介:



戴云晶 男, (1988—), 硕士. 研究方向为社交网络.



邓倩妮 女, (1973—), 博士, 副教授, 硕士生导师. 研究方向为高性能计算及其应用、数据中心网络.