论文题目：

利用社交数据评估分享者声誉

部　　院　系：　信息科学与技术学院

专　　业：　计算机科学与技术

学　　号：　201521210022

学 生 姓 名：　黄勇

2015 年 12 月 10 日

利用社交数据度量评估分享者的声誉

Jaewon Yang. Bee-Chung Chen. Deepak Agarwa

在线社交网络变成了一个非常重要的通道，尤其对于用户想要分享他们的联系和社交数据来说。尽管不同的工作都涉及到发现分享者的影响力。然而对于发现“有声誉的”、分享好内容的分享者，很少有人关注。对于用户声誉的评价打分对各种应用都有重要作用，比如向用户推荐应该去关注的重要人物等、创造好的内容区鼓励高质量的分享。而且，为了估计分享者的声誉，很重要的就是根据内容分享后，看到内容的人对于相关内容的评价（通过点击流、点赞等）。然而这种数据通常是有偏差的，由于分享者的内容只能被连接到分享着的用户看到和评价到。为了纠正这种偏差，我们利用能提供无偏差的估计小规模的数据集，提出通过一种新的多层次的模型度量有偏差的社交数据，这种多层次的模型描述如何通过利用分享着的声誉得分获得有偏差的、无偏差的数据并生成联合概率分布，同时，无偏的数据也为不同方法定量评价提供了坚实的基础。基于这些真实数据的评价估计用户的社交影响力，展示了我们提出的模型相比于现存的方法有很大的性能提升。

关键词：分享者声誉、有影响力用户

1. 前　言

通过社交网络无缝的在相互关注的用户之间分享内容，简称为 connection。大量的内容生产者允许方便的通过社交网站来分享数据，比如 Facebook、twitter、LinkedIn 等多个社交平台都有分享按钮。通常这些内容转载事件，我们称之为 shares，是通过广播在相关网站上和自己相关联的用户，广播之后，相关的内容接收者会对内容进行评价，比如再分享、点赞、评论等方式，这样的社交传播方式又会传播相关的内容给其他的关注者。比如 linkedin 主页的网站的更新流（network update stream 提供了一系列用户好友分享内容，用户可以通过回复这些分享、点击查看这些分享(通过点赞或者分享按钮)来广播给自己的好友，或者仅仅评论内容。这样的数据我们称之为回复数据（response data）。

分享者声誉和内容的影响力。我们研究了如何通过分享者不同的分享内容来估计分享者的声誉。直觉上，分享者对于某些主题的声誉展示了他是否能分享好内容。更准确地来说，我们按照随机用户（不一定是直接是用户的好友）对于内容是否感兴趣定义了分享者的声誉。　对于浏览，我们我们将随机用户的分享趋势成为内容的吸引力。这种应用能撬动分享着的声誉并提高关注者的数量。

- 推荐有声誉的分享者给用户。有声誉的分享着能对不同的主题做信息的过滤者。关注这样的用户能帮助用户接受在大量潜在的内容中找到到最好的内容。
- 选择高质量内容的监护人：有声誉的分享者能帮忙定义高质量的内容。鉴于他们物质激励能获得更好的内容分享，这样更能有拓展性招聘编辑（编辑高质量内容）来做比机器学习方法更好的方式。
- 提供内容的特征进行排序。有声誉的分享者分享的内容（在调整偏差的特征后），比其他内容排序排的更好。

有偏差的社交回复数据。采用社交社交网站的回复数据估计无偏差的用户声誉非常困难，尽管这些数据允许我们测评一个用户对分享者的积极程度，但它是有偏差的。

- 选择性偏差。内容的分享通常最容易被关系最接近的好友分享。然而我们定义声誉的方法通常按照随机的用户（对相关主题感兴趣），为了确定有声誉的用户分享的内容使关注相关领域的用户感兴趣。这种选择性的偏差由于好友之间的关系而加强社交网络之间的联系，产生量回复内容的偏差。
- 回复的偏差。信息的接收者知道分享者的身份，从而决定分享内容。正如一些文献讨论过的，用户回复数据通常一句分享-接受者之间的关系。比如，通常好友之间的内容分享比随机用户之间的分享更积极。上级的内容更容易被下级分享。这样的回复偏差（response bias）也需要被纠正。

先验的工作。大部分关注用户声誉的文章并没有纠正社交数据的偏差。一个常用的方式就是构建用户属性为节点的图，然后用 pageRank 或者 HIT 等方法来确定图中有

影响力的节点。然而这样的方法被有偏差的选择数据和回复偏差所影响。因此，有影响力的节点常常源于高的指向节点（in-degree）。或者附近对于相关工作的关注度。我们实验证明论这种方法没有很好地估计分享者的声誉。Section2 讨论了相关的工作。有偏差的移除工作已经在问答网站、评论排序环境中；在这种环境下，少量数据集被人工的编辑人员提供。这篇文章，我们研究了一种不同的问题设定和开发了数据驱动方式而不再需要人的监督。

我们的方法。我们提出了有偏差的用户回复数据和无偏差的用户行为数据（点击、点赞等），并通过一种新的多层次模型描述无偏和有偏的数据联合生成，估计了无偏的信誉得分。无偏的用户行为数据记录了用户如何通过回复分享的他们感兴趣的内容。

- **随机用户:**分享内容应该被随机采样分配给用户，而非只有好友之间。
- **隐藏身份：**当用户决定要对某个内容采取行动（如点击、点赞）的时候，他应该不知道分享这个内容的分享者的身份。

可用无偏差的数据：我们认为大部分社交网站能通过积极的实验获取无偏差的数据。大部分社交网站能拥有或者创造内容推荐模块。比如 LinkedIn 的 LinkedIn Today 模块，这个模块可以给用户推荐相关的内容并不向用户展示分享者的身份信息。一种获取信息的方式就是通过随机展示部分分享的内容并记录他们的回复数据。我们发现，尽管没有随机化，只要推荐内容覆盖一系列典型的内容，数据直接通过这样的模块手机通常有很低的选择和回复偏差。我们也发现无偏的数据通常不需要太大，只需要覆盖到典型的分享内容即可，合适的性能能通过数千个典型的分享内容进行分析。

模型挑战：考虑到这些事情，由于用户对一个内容任务是无偏的，有无偏的用户行动数据是聚集的，这样聚集的评价是所有的分享者对于某个内容的评价，然而另一方面，因为每个回复所以个人的数据，所以有偏差的社交回复数据是聚集的。因此，主要的挑战是小规模的无偏的用户响应内容和大规模有偏差的用户对于所有分享内容如何构建联合的模型。

贡献：我们通过结合有偏差的社交回复数据和无偏差的用户活动数据，提出了新的解决方法来估计无偏差的用户声誉得分，并构建一个新的多层次模型。我们的模型联合了无偏差的数据来估计用户潜在的无偏差的声誉、有偏差的数据通过分享者相关的线性聚类。为了提高模型的性能，我们用少量无偏差的数据，提出了马氏随机游走先验来估计声誉得分。我们也严格的评价了我们的无偏差数据获得的模型，并且我们的方法比现存的方法有巨大的提升。

2. 相关工作

估计在信息传播中重要个人用户已经被很多主题进行了研究，重要的人物通常能快速的影

响相关的邻居。按照种子理论的提出者，发现有影响力的用户在很多地方都被研究，在讨论我们的工作如何和其他工作不同前，我们先发现，有影响力的用户通常指那些能影响好友的人，而有声誉的用户在我们的定义里指那些对不知名的听众能收到好的评价。

一种研究成熟的方案是通过设置一个种子节点，通过种子节点设置最大的在网络中的到达区间来查找有影响力的用户。我们期望估计所有节点的声誉，影响最大化的节点主要关注少部分重要的节点，而且，影响最大化的节点被证明是合成的数据，而我们采用了真实的无偏数据进行训练和测试。

另一种方案是对有影响力的节点进行排序，这种方法包括 PageRank 和 HITS，早期的研究主要采用 pagerank 来估计网络影响力，而最劲，研究人员发现信息传播过程的结构和网络的结构并不一样，信息传播中信息被网络的结构深深地影响。两个人在网络中没有链接，数据驱动的方法可能会假设这两个人没有相互影响，我们的方法能避免这样的问题，通过度量网络中独立不相关的数据，而且，我们的评价是量化的，而之前的方法采用非定量的方法进行分析。

另一个弱相关的实证研究分析了用户行为和声誉之间的关系，问答系统采用声誉排名方式，让用户有权按照得分进行排名。我们的方法估计了潜在声誉能对有效的问答网站的排名系统有帮助。

3. 详细方法

这一部分我们详细介绍自己的新方法，如何有效通过联合无偏差的聚集数据和有偏差的书来构建模型估计分享者的声誉。

3.1 问题设置

我们首先精确的定义了我们研究的论文中用户的声誉。

分享者的声誉：我们定义了分享者 s 在主题 k 的声誉，考虑到一个随机的用户会对一个分享者 s 的分享内容产生积极的回应。通常，一个用户对多种主题感兴趣，为了简单起见，我们假设用户感兴趣的主题都是知道的，否则，定义用户部分兴趣能被用来估计声誉，利用我们已有的数据，我们使用点击流作为我们主要的行为，而且，其他的行为方式都能被按照相似的方式进行处理。

我们定义声誉按照如下方式进行处理：（1）内容推荐对象能明确相连接来最大化在推荐内容上的用户行为，比如推荐高分的分享着能最大化点击量。（2）足够多的无偏差的无偏差的数据能被精确计算，并提供有价值的真实数据来评价不同的精确的方法。稍后我们会讨论无偏差的、丰富的数据集。

现在，我们提供更详细的两种数据集的描述。

有偏差的回复数据：在社交网站上，用户通过分享内容给自己相关的好友来传播自己的影响力，当接受者看到内容和分享者的身份时，他就会对于相关的内容进行点击、点赞、再分享等，或者由于内容原因假装自己没有看到。我们的试验中使用 LinkedIn 主页下网络更新流模块的日志。

假设用户 i 分享了用户 j 的分享内容 s，称之为z_{sij}，为了简化，我们不区别积极的回复类型的种类，同时，回复数据能被量化为分享者的声誉。然而选择和回复误差应该被纠正。

无偏差的用户行为数据：为了能移除社交网站回复数据中的偏差，我们收集了一些无偏差的用户行为数据，和回复数据一样，无偏差的行为数据记录用户对他们看到东西的行为，然而这种无偏差的数据需要选择随机的用户和隐藏的身份属性。

尽管无偏差的用户行为数据确实有回复和选择偏差，对分享者信息得分估计是可行的，系统知道分享者内容的集合，这样就能通过有偏差的数据用来估计分享者声誉。

无偏差的用户行为数据能通过实验或者收集推荐系统的日子获取，我们的实验采用日志数据获取方式，另外，我们使用了 LinkedIn 主页中 LinkedIn Today 模块，考虑到回复数据无偏差并且内容能被推荐到所有的用户里面（不仅仅是好友），但由于内容不能被推荐给所有的用户，估计分享者声誉时可能也会存在一定偏差。然而，由于文章流行度是典型的通过实验算法进行估计，也需要确定随机性，选择偏差很弱，在第四部分有我们详细的数据，正如第一部分讨论过的一样，这种无偏的用户行为数据也可以被其他网站使用。

定义y_{ij}为用户 i 对内容 j 的回复，同时 i 不知道分享 j 的人的信息。我们的数据集里面，每个点击都是积极的行为，其他类型的行为也可以采用这样的方式。非常重要的一点就是，用户行为数据在响应回复中没有偏差，分享信息能被聚集解决提出了新的挑战。

无偏的数据是稀疏的。比如我们的数据集里面，内容分享数据远大于行为数据，同时，许多用户分享内容永不回发生在无偏差的用户行为数据中。

由于一片内容会被多个分享者分享，很难弄确定分享者行为，由于数据稀疏性，我们面临着分享者的数量比分享内容的数量多的情况。问题的属性变得更难了。

问题定义:考虑一个z_{sij}的社交回复数据和一个y_{ij}无偏差的用户行为数据，用户兴趣数据α_i，对每个用户有一个特征向量x_i和一个特征矩阵w_{si}，表示分享者 s 和 i 互为好友，目标是估计无偏的声誉得分μ_{sk}，表示一个随机用户对于主题 k 感兴趣，对于分享者 s 分享的内容会采取积极的响应。

3.2 模型

我们定义了一个通用的生成模型来估计用户声誉得分。通过数据来学习未知的潜在因子，模型会在 3.3 节进行详细讲解。

用户行为建模：对于无偏的用户行为数据，我们假设二分的变量y_{ij}的均值表示用户 i 在内容 j 上的函数分布是服从于用户 i 的对不同主题兴趣的矩阵a_{ik}，影响力p_{jk}对于内容 k 的感兴趣用户 j，更精确地用公式表示y_{ij}~Bernoulli(probability $=$ $\sigma(\sum_k a_{ik}p_{jk} +$ b)) 而 $\sigma(x) = {1}/{(1+e^{-x})}$表示 sigmoid 函数，并且 b 是一个从数据中学习到的偏差，p_{jk}是一个潜藏的影子和内容 j、主题 k 相关联。我们观察到用户的兴趣矩阵a_{ik}被假设成为从用户个人介绍页面抽取出来的，内容消费模式在我们建模前是独立的过程。

累计的用户声誉：我们连接内容的吸引力和和用户的声誉通过模型的吸引力p_{jk}（不同主题 k 作为平均的声誉得分μ_{sk}），从而得到公式p_{jk}~$N(mean = \frac{1}{|s_j|}\sum_{s\in s_j}\mu_{sk}, var = \frac{1}{\gamma_i|s_i|})$，其中$\gamma_i$是一个调节的参数，展示了先验的信念，认为内容的吸引力和分享者的得分相关。

共同分享的先验概率：考虑到无偏的数据是稀疏的，item i 在无偏的数据中比分享者少，这是我们在 LT 中遇到的挑战。一个通用额方法是削减所有的空白的声誉得分到 0，然而这样的方法没有用，由于我们缺少我们分享者的数据，因此大部分分享者的声誉将会获得得分接近 0，我们提出了新的马尔科夫随机游走来确定先验的规则μ_{sk}。

基本的思想如下所示，在缺少用户行为的观察数据下，我们假设分享者的声誉和分享相同内容的人的声誉一样。这利用了先验的马尔科夫随机过程。

$$(\mu_{sk}|\mu_{sk}: all\ \ sharer\ \ t \neq s) \sim N(mean, var)$$

$$(\mu_{sk} \mid \{\mu_{tk} : \text{all sharer } t \neq s\}) \sim \text{Normal distribution with}$$

$$\text{mean} = \frac{\sum_{j\in\mathcal{J}_s}\frac{1}{|S_j|}\sum_{t\in S_j: t\neq s}\mu_{tk}}{\sum_{j\in\mathcal{J}_s}(1-\frac{1}{|S_j|})+\lambda_2/\lambda_1}$$

$$\text{var} = \frac{1}{\lambda_1\sum_{j\in\mathcal{J}_s}(1-\frac{1}{|S_j|})+\lambda_2}$$

为了理解先验的马氏概率，每个分享者 s 的邻居是其他分享了相同内容的分享者，我们看到权重不仅和总共享者的数量有关，也和每个分享者对每个分享数据的总数有关，如果，分享者分享了很多其他的内容，平均权重就会下降。

社交回复模型：在有偏差的数据中，我们假设每个回复z_{sij}表示用户 i 对分享者 s 的内容 j 回复，并用二项式分布进行建模：

$$z_{sij} \sim \text{Bernoulli}(\text{probability} = \sigma(\textstyle\sum_k \eta_{ik}\alpha_{sk} + \boldsymbol{\beta}'\boldsymbol{x}_{si})),$$

x_{si}表示一个特征矩阵，包括能潜在被说明在用户 s 和 i 之间偏差的特征值，β是一个聚类的矩阵相关因数。我们称a_{sk}为无度量的声誉得分。由于每个用户对于回复数据都不同，因此我们在无偏的用户行为数据中获得了无偏的声誉μ_{sk}。

基于回归分析的度量：我们在a_{sk}和μ_{sk}建模了一个聚类的模型，聚类的相关因子依靠用户特征。 $\mu_{sk} \sim \mathcal{N}(\text{mean} = (\phi_k' x_s)\alpha_{sk} + \theta_k' x_s,\ \text{var} = 1/\lambda_3)$

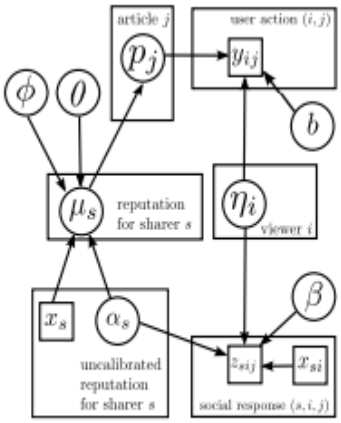


Figure 2: Graphical Representation of our model (variance components are not shown)

Symbol	Description
Observation	
z_{sij}	User i's response to item j shared by sharer s
y_{ij}	User i's response on item j
$\mathcal{J}_s$	The set of items shared by sharer s
$\mathcal{S}_j$	The set of sharers who shared item j
η_{ik}	User i's interest in topic k
x_s	Feature vector for sharer s
x_{si}	Feature vector between sharer s and user i
Variables to be learned	
μ_{sk}	Unbiased reputation score for sharer s on topic k
α_{sk}	Uncalibrated reputation score for sharer s on topic k
p_{jk}	Item j's attractiveness in topic k
ϕ_k, θ_k	Topic-specific regression coefficients between μ_{sk} and α_{sk}
β	Regression coefficients for a bias term for z_{sij}
b	Bias for y_{ij}

Table 1: Definitions of the symbols.

3.3 模型优化

我们构建了如下图所示的目标函数，为了优化目标函数，通过采用最大似然估计和不停迭代来获得P(θ|y, z)的最大化参数。 我们通过低于 10 次的迭代，学习了参数，估计了用户 s 在主题 k 上的声誉。同时，在附录中，我们解决了冷启动（在无偏差数据中从来没有分享过内容的用户）问题。

$$\begin{aligned} & \operatorname{argmax}_\Theta \log \Pr(\Theta|Y,Z) \\ = \ & \operatorname{argmax}_\Theta \log \Pr(Y,Z|\Theta) + \log \Pr(\Theta) \\ = \ & \operatorname{argmax}_\Theta \log \Pr(Y|\Theta) + \log \Pr(Z|\Theta) + \log \Pr(\Theta) \end{aligned}$$

为了能对概率模型进行优化，作者通过一系列的优化，模型通过构建最大似然估计，通过构建拉格朗日乘子，来最大化模型，估计模型的最优解，从而将数据的模型获得最优解。在建立最优化模型后，我们计算获得如下方式的模型最优解。

$$\begin{aligned} \log \Pr(\Theta) \ = \ & -\tfrac{\lambda_1}{2} \textstyle\sum_s \sum_{j \in \mathcal{J}_s} (p_{jk} - \mu_{sk})^2 - \tfrac{\lambda_2}{2} \sum_s \mu_{sk}^2 \\ & -\tfrac{\lambda_3}{2} \textstyle\sum_{s,k} (\mu_{sk} - (\phi_k' x_s)\alpha_{sk} - \theta_k' x_s)^2 \\ & + \text{constant} \end{aligned} \tag{9}$$

通过拉格朗日乘子，我们将最优化目标调整为如下函数。

$$\operatorname*{argmax}_{\{p_{jk}\},b} \log \Pr(Y|\{p_{jk}\}, b, \{\eta_{ik}\}) - \frac{\lambda_1}{2} \sum_{j,k} (p_{jk} - \frac{1}{|S_j|} \sum_{s \in S_j} \mu_{sk})^2 \tag{10}$$

通过上面几个步骤的迭代，我们获得了估计分享者声誉最大化的模型，通过多步迭代，我们就能够学习分享者声誉的参数，同时，我们也可以获得冷启动用户的参数。

4. 实验和结论

我们采用 LinkedIn 数据集说明我们的方法，同时，将这种方法和 PageRank 以及其他相似的模型进行比较，我们发现无偏差的数据集为我们提供了坚实的数据分析性能。

数据集描述：LinkedIn 是世界上最大的专业社交网站，截至 2012 年 12 月 31 日，有超过 2 亿用户，无偏差的用户行为数据从 LT 模块获得，而有偏差的数据从网络更新流中获得，这些数据用了四个月获得，从 21 年 5 月到 8 月。

LinkedIn Today 数据集：大部分 LT 数据从一个随机的用户采样获得，LT 随机推荐了顶级算法获取的匹配用户过去个人主页 profile 的文章。在数据收集期间，LinkedIn Today 推荐了三个内容给每个浏览的用户，因此一个用户选择偏差很小，为了集成在展示相同内容的偏差，我们回复y_{ij}被认为是浏览事件，我们也只考虑用户至少在这四个月点过的数据。

网络更新流数据集：在 LinkedIn 中，文章被分享者在传播者中进行传播，在 LinkedIn 页面中是 Network Update Stream。相互关注的分享者能相互通过点击或者忽视分享内容，和 LT 数据相似，我们只考虑用户第一次浏览的时间。

用户的主题兴趣：在 LinkedIn 中，每个用户至少有一个所属的事业，从而用户主题兴趣可以按照用户的个人主页中的介绍来获得，一个用户也可以关注一系列的文章来获得文章推荐，关注 industry 的兴趣对于 LinkedIn 十分重要，由于这是一个专业的求职社交网站，因此，我们通过分享着在每个领域内的声誉获得总得分。分析其他的主题相关的也很相似，实际上，我们的模型可以拓展到其他相同的分层模型上。

分享者的声誉：为了测量分享者 s 对于主题 k 的声誉，我们使用了平均的 LT CTR 作为评价指标，CTR 是总的点击数目与总的浏览数目之比，我们统计了 LinkedIn Today 的数据集中的无偏差的 LT 数据。

度量方式：为了评价模型的性能，我们将获得的数据集分为两个部分，其中一部分作为训练集，另外一部分作为测试集，利用训练集进行训练，利用测试集评价模型训练的好坏。为了减少噪声，我们度量测试集按照如下两个方面进行选择：（1）一个分享着的测试声誉在自己的主页中至少有 10 个分享的无偏的 LT 数据，否则这样的集合就会太小而不能计算。（2）测试的集合中，至少有超过 100 的浏览量，否则过低的点击率会很难估计分享者在内容分享中的声誉。

由于我们的度量方式是基于声誉的，我们采用了两种度量方式。考虑到排序度量并不依赖于度量者的声誉分数。在不同规模的数据集中产生分数十分重要。我们的方法和 PageRank 在不同规模的数据集中进行了排序，但是我们的方法对于有好内容的分享者，排

序的得分更好。

Kendall 系数：为了发现分享着在模型中的得分和他们的平均的 LT CTR 得分是否一致，我们采用了 Kendall 秩和检验系数连检验两者的相关度。

点击率最高的 K 个分享者：相比于较差的模型，最佳的 k 个分享者能被更好的模型区别，分享者分享更好的内容应该有更高的 LT CTR，我们定义 top k 个分享者的作为平均的 LT CTR，并且用不同的 k 值（从 20 到 100）比较模型。

4.1 比较不同的方法

这一部分将我们的数据集分为训练集和测试集，我们用 5 月的数据集进行训练，以后三个月的数据集作为测试集，用训练数据估计分享者的声誉，用测试数据评价获得的平均 LT CTR，我们的测试集合远大于训练集，主要是为了减少度量的方差。

基准方法：PageRank 方法在最近被用来作为用户在社交网络中的影响力排序指标，我们在试验中实现了两种增强的方法进行比较。这两种方法比较模型的依据在于好的内容通常获得更多的关注。两种方法不同点在于 PageRank-Volume 定义边的权重是：回复的总数作为从回复者到分享者的权重，PageRank-Rate 定义边的权重是回复的比例作为回复者到分享者的比重。两种方法对于边的权重都进行了标准化。

我们模型的不同：为了测量不同的组件对我们模型的影响和效果，我们考虑两种不同版本的的模型：模型 y 和模型 z 假设分享者的声誉在某个主题上的度量得分是基于有偏差的数据，而模型 z 是基于 UTS 数据集合进行估计，模型 y 是基于 LT 数据进行估计。需要考虑到模型 y 不能估计冷启动的分享者声誉。

不同模型性能的比较：表二的第一列表明我们的模型有最好的性能，我们的全模型获得最高的排序相关度，这说明将有偏差的社交回复数据和无偏差的用户行为数据混合是很适合提高模型性能的，模型 y 只获得了第二高的排序，这是由于我们的模型将用户声誉定义为无偏差的 LT 数据类型，模型 z 只获得了第三高的排序，这是由于我们定义声誉是基于 LT 数据的，对于有丰富 LT 数据的用户，那些 NUS 数据集合并不充分的需要。然而，对于没有很多 LT 数据的用户，他们的 NUS 数据非常有帮助，模型 PageRank 表现比较差，他们几乎呈现负相关。

在第三部分讨论过，我们采用无偏差的回复数据（NUS data Z）用来估计冷启动的用户，这些用户并没分享很多内容，为了测量这些冷启动的用户，我们限制了我们的测试集，只关注那些没有分享内容的用户，为了减少模型的方差，我们增加了模型测试集合冷启动的用户数量，这些模型中模型 Y 不能估计冷启动用户，我们的 full 模型能对性能有极大的提升，相比于第二高的模型 Z，提升度时 241%，相关度达到 0.1124。

表 3 展示了前 k 个分享者的平均分享率和所有人的平均分享率比值，由于冷启动，模型 Y 不能进行分析这个比例，同时，考虑到当所有的曲线最终趋近于 1（当 k 接近于总数的时候）。此外，full 模型的性能最好，模型 Z 其次，除了模型 z 其他的模型模型计算声誉都是有偏差的数据，当 k 很小时，top k 的平均 CTR 要比所有人的页面的点击率高，这是符

合我们的认知的。

按不同产业分解：表 5 展示了不同模型的性能，将模型分不同的行业进行计算他们的 Kendall 系数相比于基准方法的提升度，在表 5 中，有 14 类产业的模型构建中，全模型超过了基准模型的 Kendall 系数，全模型在冷启动、或者书用户很少分享内容的产业性能较差，在有冷启动的用户的模型中，full 模型在大部分产业中都很强，性能超过了 18 个产业中的基准线模型，达到总数的 90%。

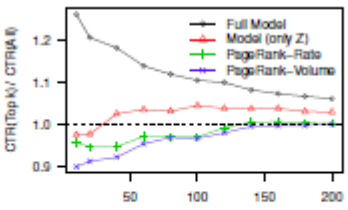
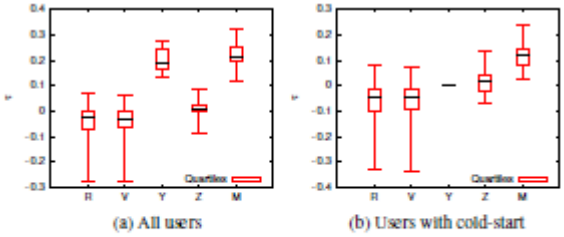
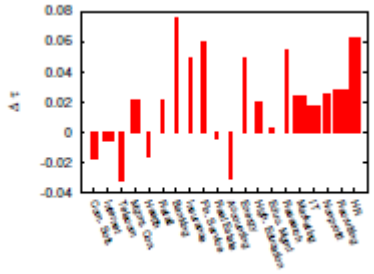


Figure 3: CTR of top k sharers for different models as a function of k, normalized by the average CTR of all items

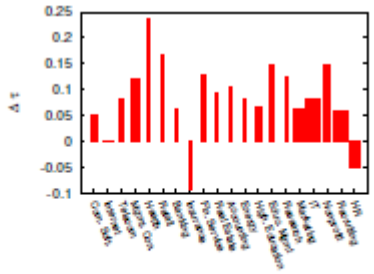


(a) All users (b) Users with cold-start

Figure 4: Box plots of the distributions of Kendall's τ for the top 20 industries. R: *PageRank-Rate*, V: *PageRank-Volume*, Y: Model (only Y), Z: Model (only Z), M: Full Model.



(a) All users



(b) Users with cold-start

Figure 5: Improvement in Kendall's τ for each industry by Full Model compared to the best baseline. The bar width is proportional to the number of test users in the industry

4.2 模型的特征

不同总数的无偏数据：首先研究在训练模型中使用不同的无偏差的用户行为数据（LT data）。除了按照时间来选择训练集合和测试集合，我们随机地选择了 40%的数据作为测试数据，60%的数据作为训练数据，其他的影响因素都不考虑。为了考虑冷启动问题，我们随机的选择了 50%的分享者，这些分享者满足之前提到了两个满足条件，此外，为了模拟不同总数的无偏差数据，我们设置了 P%的随机内容 作为训练集，剩下的作为测试集，我们考虑到不同性能的模型在不同子集中，模型的性能相互之间并不和之前的子集不兼容。

表 6a 展示了 Kendall 系数相关性，随着无偏数据集合中训练数据集合的提升，无偏差的数据集合能被减到 50%而不用非常大的影响模型的性能，从 5%-50%，模型的性能获得基本上服从 log 线性规律。表 6b 展示了 Kendall 系数的相关性提升，我们将共享的随机先验概率放在分享内容μ_{sk}。让μ_{sk}服从均值为 0，方差为$\frac{1}{\gamma}$的正态分布。从图 6b 中可以看出，数据集合越小，数据的 Kendall 排序系数的提升度越大。

4.3 讨论

我们的模型相比于其他的模型有以下几个优点：

- 传统的估计社交影响力的方式，比如 PageRank 方法效果相比于我们的方法较差，我们的方法通过构建一个多层的模型，纠正了社交网络中有偏差的用户回复模型和无偏差的用户响应模型，获得了巨大的性能提升。
- 我们的方法结合了两个部分的内容，考虑到数据的冷启动问题，对于预测有影响力的用户、但是这种用户没有分享很多无偏差的内容，我们的方法非常有创新的覆盖了更多的有声誉的用户，这样能获得更大的覆盖面积。
- 在预测中，非常有趣地看到马尔科夫随机过程先验起了非常重要的作用，当无偏差的用户行动数据非常少的时候，通过先验概率增加平滑性能非常好的提高模型的性能。实际上并不容易获得大量的数据，采用先验概率预测能有效处理缺失值。

5 结论

我们通过构建多层次的模型获得了评价无偏的用户声誉，这些模型结合了选择和回复偏差内容，以及无偏差的用户行为模型，我们的模型相比于现在的模型，对于估计声誉有很大的提升。我们发现，通过有标签的基准线计算，在社交数据是很有可能纠正偏差的，此外，通过马尔科夫随机过程先验概率帮忙提供了估计小部分的无偏差的数据，我们的方法也打开了以后需要的工作，我们说明了自己的想法并使用两种数据证明它，然而这种方法可以拓展到多种数据中，但是对于多种有偏差的数据，这些数据可以被调整不同的聚集方式，同时结合无偏差的数据，这样就可以获得很多有质量的高素质的内容文章进行阅读，为了能处理更多的有偏差的数据，我们需要在将来设计更多多层级的聚集方法，调整有偏差的数据，从而能更好的获得分享者的声誉。

参考文献略。