利用社交数据评估分享者声誉

北京师范大学信息科学与技术学院　　黄勇

利用社交数据评估分享者声誉

- 社交网络分享者声誉的问题引入
- PageRank、HITS等当前研究方法
- 作者提出新的分享者声誉评估模型
- 实验数据收集、建模、结果分析评价

社交网络分享者声誉的问题引入

- 社交网络能非常方便地和好友之间相互分享内容、好内容能获得更好赞誉
- 能为用户推荐高质量的内容阅读，分享者能对内容进行过滤，读者也能从高质量的内容中受益
- 发现好的分享者、为好内容条目排序
- 当前社交网络数据是充满偏差，不利于评估分享者的声誉。
- 声誉定义基于一个随机用户对于话题兴趣，社交网络内容是选择性地分享给好友等相关用户，产生选择偏差
- 用户知道分享者的身份，由于上下级、亲密度等因素进行回复，产生回复偏差

利用社交数据评估分享者声誉

- 社交网络分享者声誉的问题引入
- PageRank、HITS等当前研究方法
- 作者提出新的分享者声誉评估模型
- 实验数据收集、建模、结果分析评价

当前的研究方法

- PageRank-V算法

  分享内容回复的总数作为有向边的权重

- PageRank-R算法

  回复占浏览的比例作为有向边的权重

- HITS算法

- 问答系统和答案排序

利用社交数据评估分享者声誉

- 社交网络分享者声誉的问题引入
- PageRank、HITS等当前研究方法
- 作者提出新的分享者声誉评估模型
- 实验数据收集、建模、结果分析评价

作者提出新的方法

基本假设

- 阅读分享内容的用户随机的
- 分享者的身份被隐藏
- 保证对内容的评价是无偏的

数据选择

- LT:匿名推荐的数据
- UPS:好友的信息流更新
- 个人页面中分享过的信息

挑战和优势

- 用户行为数据是聚集的
- 用户回复数据是分散的
- 结合两者构建了层次模型

模型的参数和模型的概括

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
$\boldsymbol{x}_s$	Feature vector for sharer s
$\boldsymbol{x}_{si}$	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
$\boldsymbol{\phi}_k, \boldsymbol{\theta}_k$	Topic-specific regression coefficients between μ_{sk} and α_{sk}
$\boldsymbol{\beta}$	Regression coefficients for a bias term for z_{sij}
b	Bias for y_{ij}

Table 1: Definitions of the symbols.

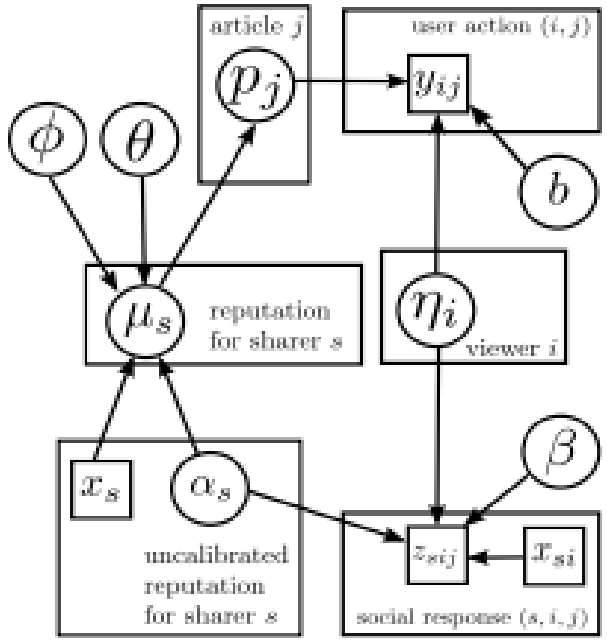


Figure 2: Graphical Representation of our model (variance components are not shown)

无偏数据的建模过程

- 无偏的用户行为建模
- 用户回复内容服从二项分布
- 用户的声誉服从正态分布
- 稀疏数据采用先验概率补全

User action model. For the unbiased user action data, we assume the mean of the binary response y_{ij} for user i on item j is a function of user i's interest vector η_{ik} for different topics k, and the attractiveness p_{jk} of item j for users interested in different topics k. More specifically,

$$y_{ij} \sim \text{Bernoulli}(\text{probability} = \sigma(\textstyle\sum_k \eta_{ik} p_{jk} + b)), \quad (1)$$

Aggregation of user reputation. We connect attractiveness of items to user reputation through modeling the attractiveness p_{jk} of item j for users interested in different topics k as the average of reputation scores μ_{sk} of the sharers $s \in \mathcal{S}_j$ of item j; i.e.,

$$p_{jk} \sim \mathcal{N}(\text{mean} = \frac{1}{|\mathcal{S}_j|} \sum_{s \in \mathcal{S}_j} \mu_{sk}, \ \text{var} = \frac{1}{\lambda_1 |\mathcal{S}_j|}) \quad (2)$$

有偏数据的建模过程

- 未调整偏差的回复数据的声誉模型
- 社交回复数据服从二项分布
- 基于线性回归，利用有偏差数据和无偏差数据，调整模型偏差

Social response model. In the biased social response data, we assume that each response z_{sij} represents whether user i would respond positively to item j shared by sharer s, and it is modeled as a function of user i's interest vector η_{ik} and the "uncalibrated reputation score" α_{sk} of sharer s on different topics k; i.e.,

$$z_{sij} \sim \text{Bernoulli}(\text{probability} = \sigma(\textstyle\sum_k \eta_{ik}\alpha_{sk} + \boldsymbol{\beta}'\boldsymbol{x}_{si})), \quad (5)$$

Regression-based calibration. We model the relationship between α_{sk} and μ_{sk} through a linear regression, where the regression coefficients depend on user features; i.e.,

$$\mu_{sk} \sim \mathcal{N}(\text{mean} = (\boldsymbol{\phi}_k'\boldsymbol{x}_s)\alpha_{sk} + \boldsymbol{\theta}_k'\boldsymbol{x}_s,\ \text{var} = 1/\lambda_3), \quad (6)$$

构建联合的概率模型

- 构建联合模型
- Y和Z分别代表有偏差和无偏差的模型
- 采用先验概率消除稀疏的数据

- $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are conditionally independent:

$$\Pr(\boldsymbol{Y}, \boldsymbol{Z}|\Theta) = \Pr(\boldsymbol{Y} \mid \{p_{jk}\}, \{\eta_{ik}\}) \cdot \Pr(\boldsymbol{Z} \mid \{\alpha_{sk}\})$$

- Joint prior on latent variables:

$$\Pr(\{p_{jk}\}, \{\mu_{sk}\} \mid \{\alpha_{sk}\}) = \Pr(\{p_{jk}\}|\{\mu_{sk}\}) \cdot \Pr^{\text{CSH-MRF}}(\{\mu_{sk}\}) \cdot \Pr(\{\mu_{sk}\}|\{\alpha_{sk}\}, \phi_k, \theta_k)$$

where $\Pr^{\text{CSH-MRF}}(\{\mu_{sk}\})$ is the co-sharing Markov random field prior. Note that the prior on $[\{\mu_{sk}\}]$ is proportional to

$$\Pr^{\text{CSH-MRF}}(\{\mu_{sk}\}) \cdot \Pr(\{\mu_{sk}\}|\{\alpha_{sk}\}, \phi_k, \theta_k).$$

利用社交数据评估分享者声誉

- 社交网络分享者声誉的问题引入
- PageRank、HITS等当前研究方法
- 作者提出新的分享者声誉评估模型
- 实验数据收集、建模、结果分析评价

数据集的收集

- 选用LinkedIn在2012年5月到8月的数据
- 选择LinkedIn Today模块和Network Update Stream模块数据
- 将5月的数据作为训练集合、其他月份的数据作为测试集合

选用数据集和度量方法

- 选用LinkedIn在2012年5月到8月的数据
- 选择LinkedIn Today模块和Network Update Stream模块数据的CTR和日志
- 将5月的数据作为训练集合、其他月份的数据作为测试集合
- 基于Kendall秩和检验：计算分享者声誉排序和LT CTR是否有一致性
- 计算平均点击率最高的k个人和所有分享者内容的点击率是否有差别

新模型和基准模型的比较

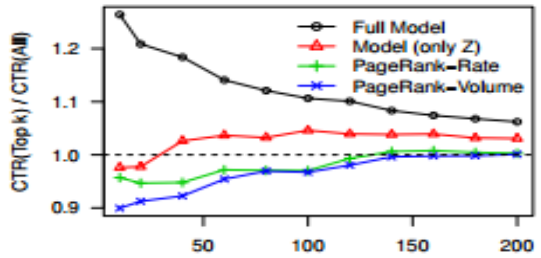
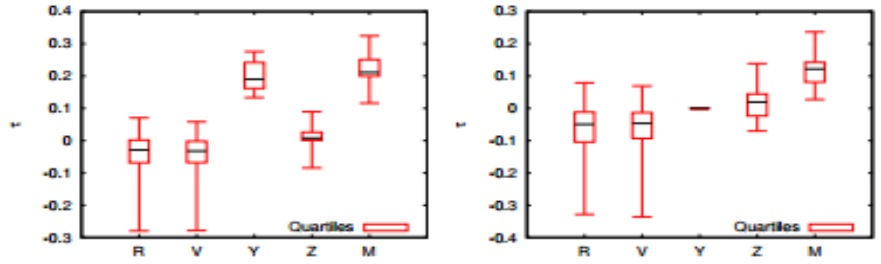
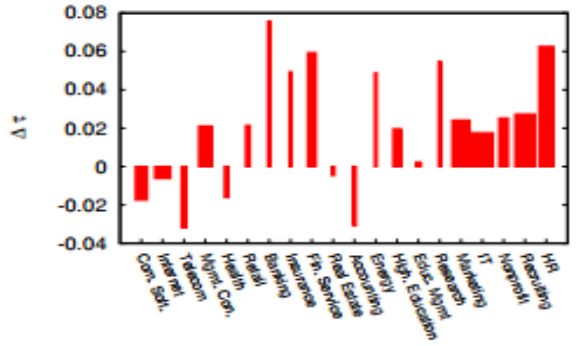


Figure 3: CTR of top k sharers for different models as a function of k, normalized by the average CTR of all items

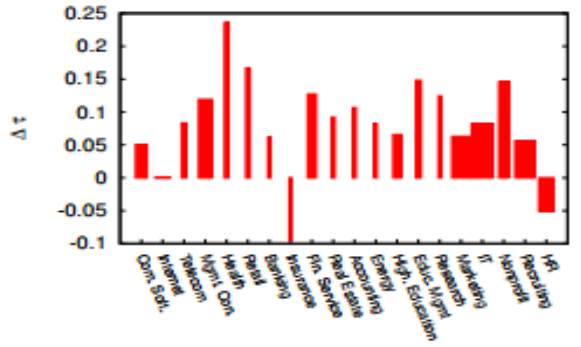


(a) All users (b) Users with cold-start

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

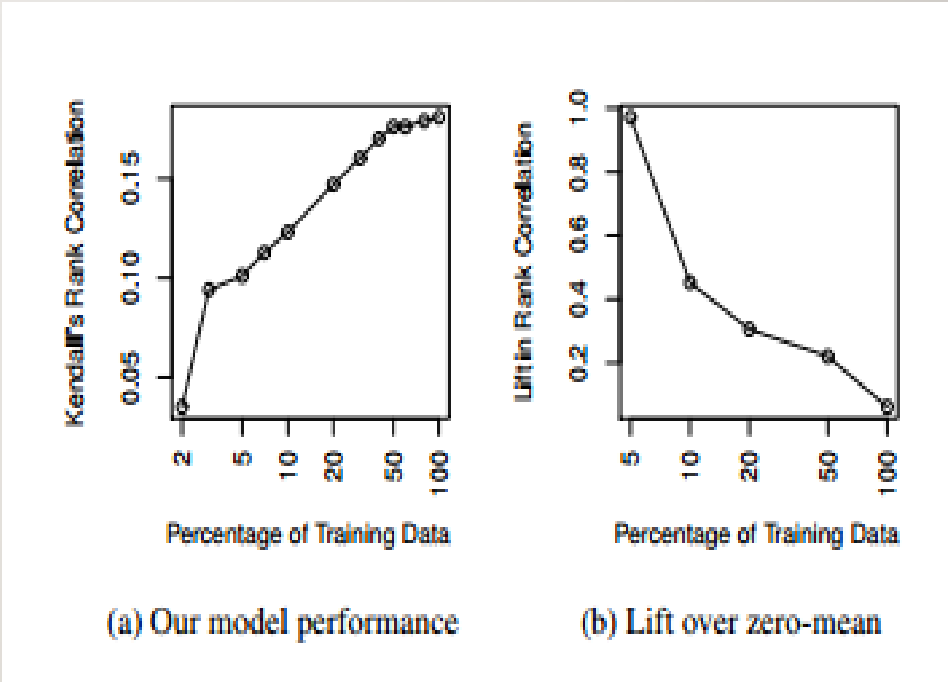


(b) Users with cold-start

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

结果分析和评价

- 模型比PageRank方法有巨大提升，考虑到数据分为有偏差的和无偏差的两种类型
- 考虑到数据的冷启动问题，适于发现高质量分享内容少的用户、扩大模型的覆盖面
- 基于先验概率处理缺失值，在没有大量数据情况下，能更好估计分享者声誉



(a) Our model performance (b) Lift over zero-mean

利用社交数据评估分享者声誉

北京师范大学信息科学与技术学院　　黄勇