

基于社交网络的问答专家推荐技术

周光有

课题基本信息

- ◆ 课题名称：基于社交网络的问答专家推荐技术
- ◆ 所属项目：基于自然语言处理的智能社会网络计算
- ◆ 课题编号：CCF2013-01-01
- ◆ 课题时间：2013年9月-2014年9月
- ◆ 课题经费：3万元

主要研究内容

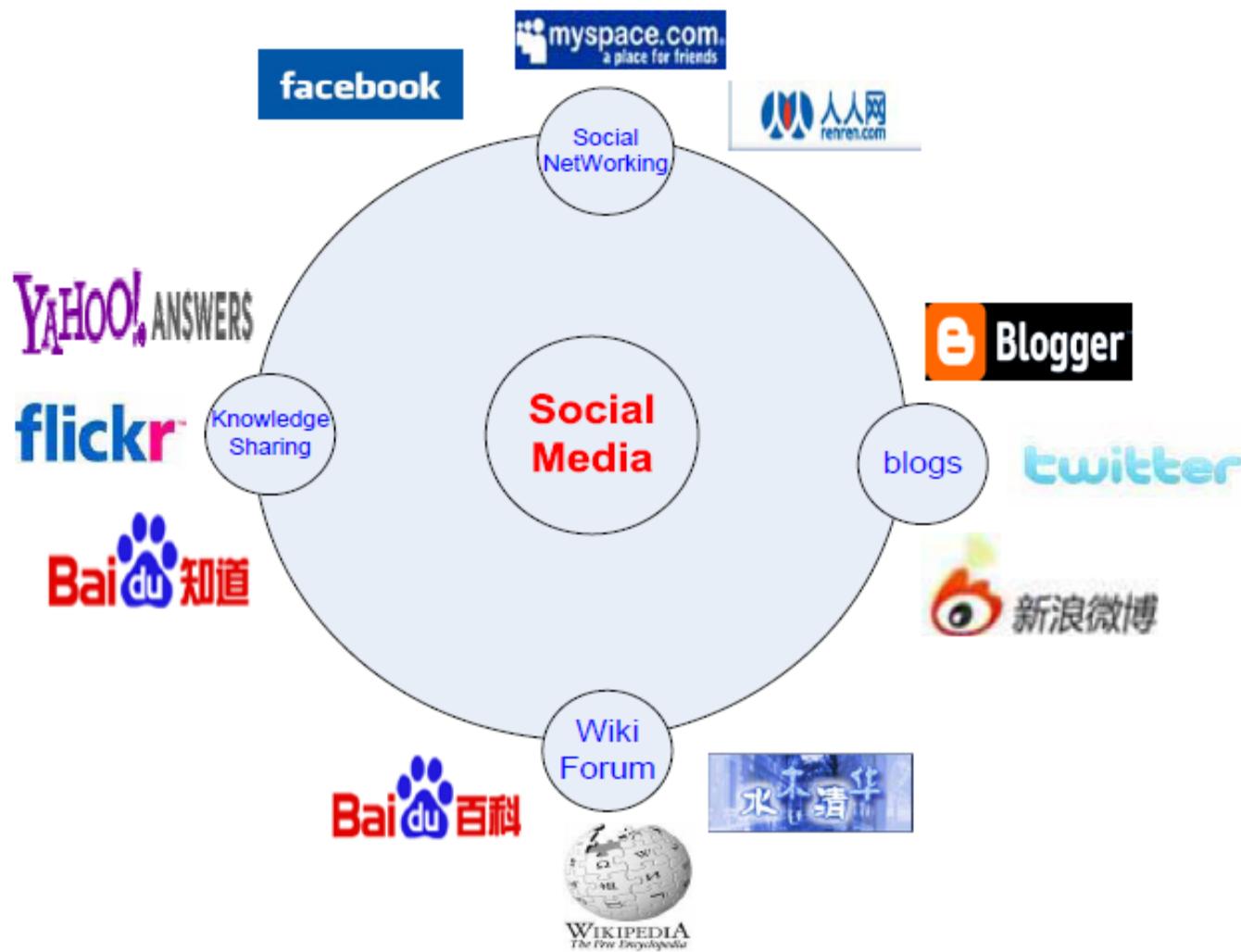
- ◆ 社区问答中的用户交互机制
- ◆ 社区问答中的专家用户挖掘
- ◆ 社区问答中的最佳回答者推荐

课题主要成果

- ◆ 发表论文5篇，其中高水平SCI期刊一篇，国际顶级会议3篇，NLPCC论文一篇（最佳论文）
- ◆ 申请专利一项
- ◆ 较好地完成了项目预期的任务

社区问答中的“专家用户”挖掘

Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)



Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. *Knowledge-based systems*, 66:136-145,2014. (SCI, IF=3.175)



生气时，为啥说话会很大声？

问 【知道日报】鸟撞：当“小鸟”遭遇“铁鸟”？

答 目前，鸟类撞击已经和风切变、机翼结冰并称为世界航空运输业的三大灾害。

回答者：fengfeixue0219

【知道日报】如何艺术地表现7种精神疾病？

【知道日报】少女妈妈的孩子学习差？

【真相问答机】曝：美国司法部利用飞机收集手机信息？

周光耀068
一级

财富值：20

活跃天数：0 天

去签到

0
我的提问0
我的回答 NEW

新的一天从答题助人开始

go 去答题，开启知道之旅吧！

【知道日报】鸟撞：当“小...

【知道日报】怎样判断自己...

【知道日报】为什么猫总是...

更多

问题分类

- 电脑/网络 >
- 硬件 常见软件 互联网
- 生活 >
- 服装/首饰 美容/塑身 购物
- 医疗健康 >
- 内科 妇产科 人体常识
- 体育/运动 >
- 足球 篮球 健身
- 电子数码 >
- 手机/通讯 照相机/摄像机
- 商业/理财 >
- 股票 财务税务 创业投资
- 教育/科学 >
- 理工学科 外语学习
- 社会民生 >
- 法律 求职就业 时事政治
- 文化/艺术 >
- 文学 历史话题 书画美术
- 游戏 >
- 网络游戏 单机游戏
- 娱乐/休闲 >

等待您来回答

更多提问 >

| 我关注的关键词 | 我关注的分类 | 为我推荐的问题 |
|--------------------------------|--------|---------|
| 30 谁能解释狗为什么对人那么忠诚 | | 1回答 |
| 大约有多少个叫袁盈盈的人 | | 0回答 |
| 建行挂号信 | | 0回答 |
| we can be a good friend, bu... | | 0回答 |
| 苹果4可不可以把垃圾相信屏蔽 | | 0回答 |
| 电动机mot2遥控飞机电源电压 | | 0回答 |
| 从什么时候开始教育孩子 | | 0回答 |
| 艾泽拉斯大战1.89如何用上帝秘籍 | | 0回答 |
| 两江新区公租房有没有宾馆 | | 0回答 |
| 黄俊涵以后的老婆是谁 | | 0回答 |

公告区

更多

- 百度知道APP5.0更新上线！ NEW
- 百度知道出书啦！火爆预售中
- 知道新品：宝宝知道App上线啦
- 帮助手册：如何使用知道
- 新品上线：作业帮APP发布啦



知道日报

品质生活的「靠谱读品」

和身体有关的知识

那些和身体有关的知识

Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

知道专家排行



晋波医生

专长: 无分类

上周回答数: 2926

好评率: ★★★★★

- | | |
|---------|-----------|
| 2 吴华娟医生 | 上周回答 2692 |
| 3 高进医生 | 上周回答 2341 |
| 4 李婕医生 | 上周回答 2192 |
| 5 李彦滨医生 | 上周回答 2159 |

名医义诊



汪波主任：冬季呼吸疾病的防治

回答数 40 帮助了 24847 人

北京大学医院急诊科

问：前些天突发心慌气短，医院检查是说是房颤...

答：房颤通常跟高血压、冠心病、风湿性心脏病以及甲亢有关系，所以要排除... 294

[进入专题](#)

推荐名医

[北京大学医院熊辉](#) | [北京军联骨科医院赵克明](#) | [北京东直门医院杨保林](#)

周采纳数上升排行

总积分排行



坏先生上帝

全部回答数: 23206

上升采纳数: 4661

[向TA求助](#)

2 一骑当后

3 最后天使毁灭

4 可以叫我表哥

5 坑爹的周太狼

达人风采



18910199239 : 知道之星

被赞同数: 18091 回答数: 49823

问：北京、母子之间房屋过户，请问现在怎么办...

答：过户：需交3%契税，差额20%个人所得税
(网签价减去原购价的20%)，超... 33

[向TA求助](#)

推荐达人

[yq_whut: 知道之星](#) | [一刻永远523: 知道之星](#) | [不随意123456: 知道之星](#)

Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

专家用户挖掘研究的意义

从提问用户的角度来看

- 当用户提交新问题后，必须被动地等待其他用户访问系统
- 提问用户需要等待很长时间获得问题的答案
- 30%的问题一直得不到解答而被系统关闭(Cong et al., 2011)

从答案质量的角度来看

- 用户的知识层次不齐，得到的答案质量良莠不齐
- “专家” 用户提供的答案质量往往较高

Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

专家用户挖掘研究的意义

专家用户挖掘的任务

- 从大量的社区用户中发掘能够提高质量答案的活跃用户

传统方法及存在的问题

- 传统方法主要采用链接分析技术
- 仅仅考虑了链接结构而忽视了用户在特定主题下的相似度
- 以及用户自身的行为（用户的知名度、用户的专业程度等）

Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

基于上下文主题信息的概率主题方法

基本思想

- 通过分析用户回答的问题来自动识别用户的兴趣所在，构建了主题敏感的提问-回答关系图。
- 根据提问-回答关系图中的链接结构和主题相似度来计算用户的排序得分。
- 提出了一个概率排序模型，对候选专家用户根据用户的知名度和专业程度进行排序。

Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

用户-主题模型

- 主题识别的目的是通过对用户回答的问题进行分析，自动挖掘用户感兴趣的主题。
- 每一个问句都非常短（平均长度为11.2个词），通常仅仅包含一个句子，直接应用LDA，效果并不理想。
- 提出了用户-主题模型，将用户的所有历史问题（用户资料）合并成一个单一的文档 d 。因此，每一个文档 d 对应一个用户。

Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

用户-主题模型

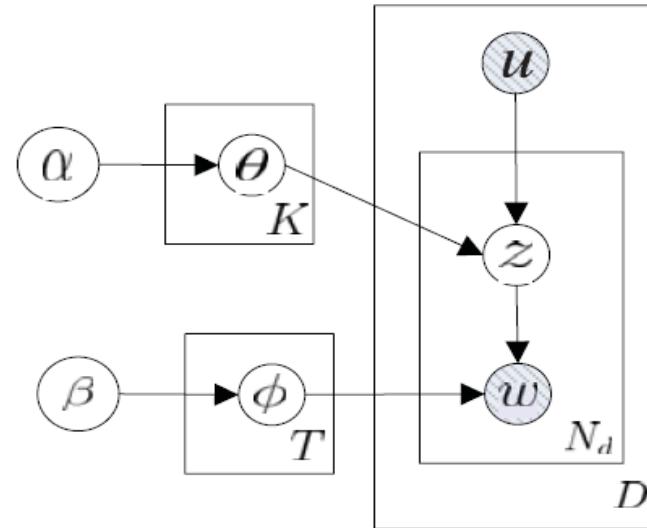


Figure: 用户-主题模型的图表示。

- 作者-主题模型认为所有的作者服从均匀分布
- 用户-主题模型中，每个用户都是显式给出的

Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

用户-主题模型

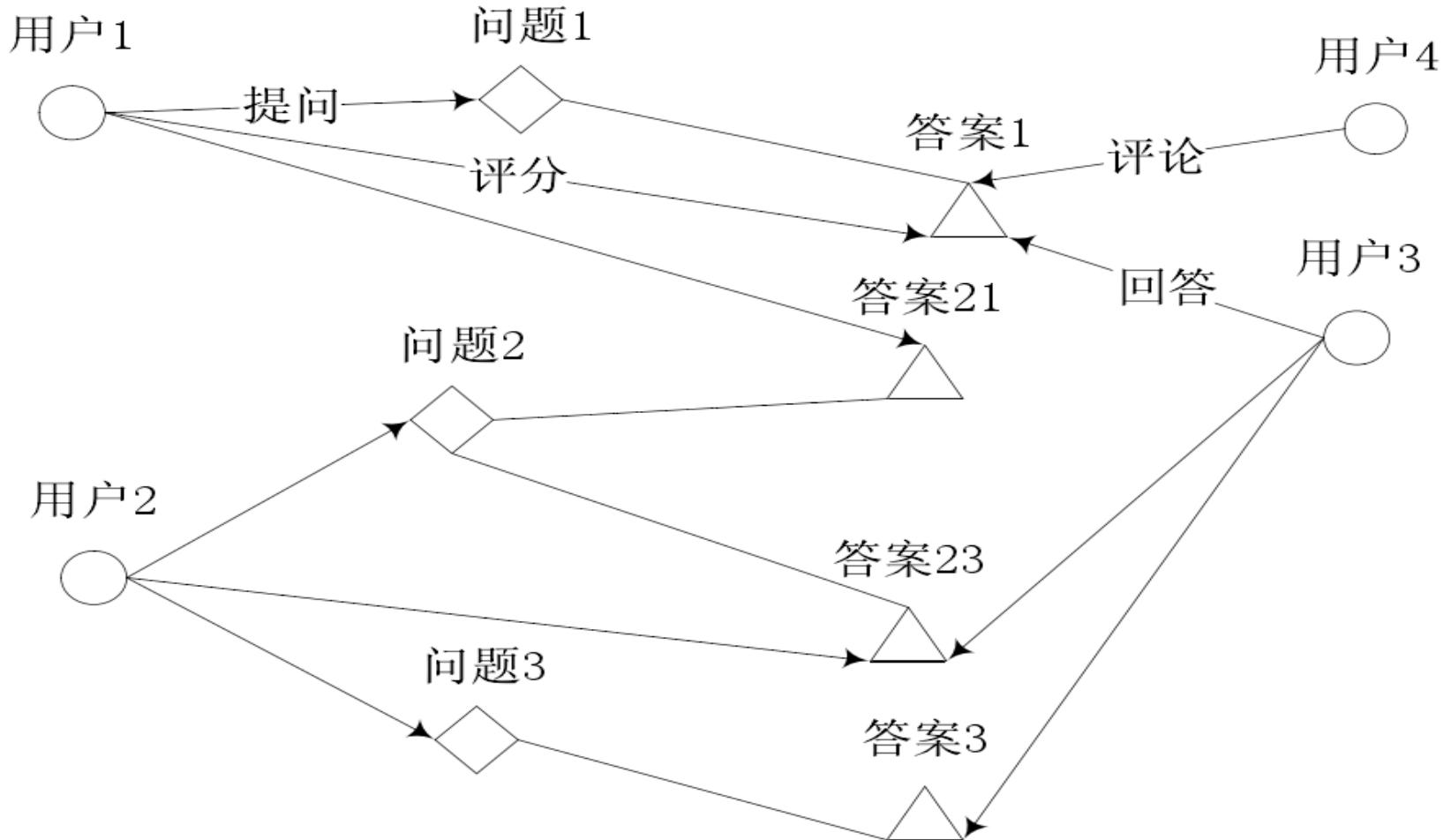
- 采用吉布斯采样 (gibbs sampling) 的方法来推导求解
- 得到两个矩阵: $|\mathcal{V}| \times T$ 表示词-主题, $K \times T$ 表示用户-主题

$$\phi_{wj} = \frac{C_{wj}^{WT} + \beta}{\sum_{w'} C_{w'j}^{WT} + \beta |\mathcal{V}|}$$

$$\theta_{kj} = \frac{C_{kj}^{UT} + \alpha}{\sum_{j'} C_{kj'}^{UT} + \alpha T}$$

其中, ϕ_{wj} 表示词 w 属于主题 j 的概率, θ_{kj} 表示用户 k 与主题 j 之间的概率。

基于上下文主题信息的概率主题方法



Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. *Knowledge-based systems*, 66:136-145,2014. (SCI, IF=3.175)

基于上下文主题信息的概率主题方法

提问用户 u_i 和回答用户 u_j 之间的权重用如下公式计算：

$$f(i \rightarrow j) = |Q(i) \cap A(j)|$$

- $Q(i)$ 表示用户 u_i 提出的问题集合， $A(j)$ 表示用户 u_j 回答的问题集合
- 同时定义 $f(i \rightarrow i) = 0$

从提问用户 u_i 到回答用户 u_j 之间的转移概率定义如下：

$$p(i \rightarrow j) = \begin{cases} \frac{f(i \rightarrow j)}{\sum_{k=1}^{|V|} f(i \rightarrow k)} & \text{if } \sum f \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

基于上下文主题信息的概率主题方法

对于给定的主题 z , 从提问用户 u_i 到回答用户 u_j 之间的转移概率定义如下:

$$p_z(i \rightarrow j) = \begin{cases} \frac{f(i \rightarrow j) \times sim_z(i \rightarrow j)}{\sum_{k=1}^{|V|} f(i \rightarrow k) \times sim_z(i \rightarrow k)} & \text{if } \sum f \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$sim_z(i \rightarrow j)$ 表示在该主题下用户 u_i 到 u_j 的相似度。

$$sim_z(i \rightarrow j) = \frac{1}{2} \left\{ p_{KL}(u_i || u_j) + p_{KL}(u_j || u_i) \right\}$$

其中, $p_{KL}(u_i || u_j) = p(z|u_i) \log \frac{p(z|u_i)}{p(z|u_j)}$, $p(z|u_i) = \theta'_{iz}$ 。

基于上下文主题信息的概率主题方法

新的行归一化的矩阵定义如下：

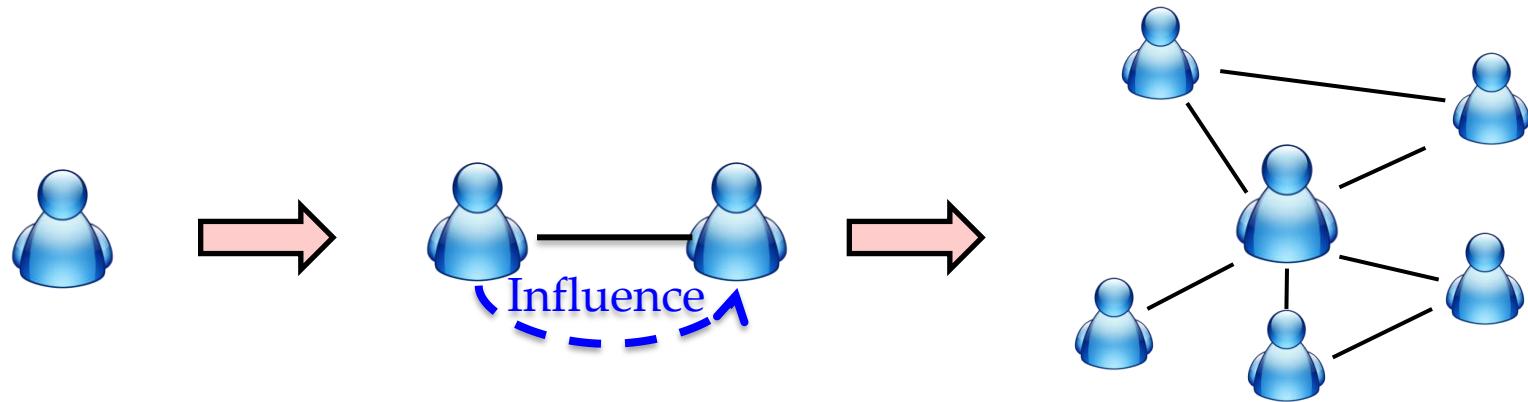
$$\widetilde{M}_{ij}^* = p_z(i \rightarrow j)$$

对于给定的主题 z ，利用TSPR得到的迭代公式如下：

$$R_z^*(u_i) = \lambda \sum_{j: u_j \rightarrow u_i} R_z^*(u_j) \cdot \widetilde{M}_{ji}^* + (1 - \lambda)p_z(u_i)$$

当利用TSPR对用户排序后，选择排序最靠前的 N 个用户作为候选专家用户。

基于上下文主题信息的概率主题方法



皮之不存，毛之焉附？
水无常形，随物附形

表层关联特征与真实
环境存在“语义鸿沟”
扭曲：夸大、缩小

基于上下文主题信息的概率主题方法

- 传统的方法仅仅利用用户之间的问答关系获得用户的打分，忽略了专家用户的先验信息。
- 提出了一个概率模型，对于给定的一个主题，认为一个候选用户是一个真正的专家，需要满足下面两个假设：

两个假设

- **专业程度：**领域知识应该与给定的主题密切相关。
- **知名度：**贡献大量高质量答案，具有领袖的能力和很高的知名度。

Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

基于上下文主题信息的概率主题方法

- 定义 E 是一个二元变量，表示候选用户的专业程度
- 定义 R 是一个二元变量，表示候选用户的知名度
- 根据概率关联模型(Lafferty, 2003)，对于给定的主题 z ，用 $p(E = 1, R = 1|u, z)$ 对候选用户 u 进行排序：

$$\begin{aligned} & p(E = 1, R = 1|u, z) \\ = & \quad p(E = 1|u, z)p(R = 1|u, z, E = 1) \end{aligned}$$

- 对于已经给定的主题 z ，认为知名度 R 与专业程度 E 是相互独立的：

$$\begin{aligned} & p(E = 1, R = 1|u, z) \\ = & \quad p(E = 1|u, z)p(R = 1|u) \end{aligned}$$

基于上下文主题信息的概率主题方法

进一步定义 $\delta = \frac{p(E=0)}{p(E=1)}$, 假设 $p(E=0) \gg p(E=1)$, 即 $\delta \gg 1$:

$$\begin{aligned} & \log p(E=1, R=1|u, z) \\ = & \log \frac{1}{1 + \frac{p(E=0)}{p(E=1)} \times \frac{p(u|z, E=0)}{p(u|z, E=1)}} + \log p(R=1|u) \\ \approx & \log \frac{p(u|z, E=1)}{p(u|z, E=0)} \times \frac{1}{\delta} + \log p(R=1|u) \\ = & \log \frac{p(u|z, E=1)}{p(u|z, E=0)} + \log p(R=1|u) - \log \delta \end{aligned}$$

最终的排序分数 $\log p(E=1, R=1|u, z)$ 被分解成两部分：专业程度分数 $\log \frac{p(u|z, E=1)}{p(u|z, E=0)}$ 和知名度分数 $\log p(R=1|u)$ 。

基于上下文主题信息的概率主题方法

定义 $D(u) = (w_1, w_2, \dots, w_n)$, 表示候选用户 u 的历史问题, 在给定专业程度 E 和主题 z 的情况下, 我们有:

$$\begin{aligned} & \log \frac{p(u|z, E = 1)}{p(u|z, E = 0)} \\ = & \log \frac{p(D(u)|z, E = 1)}{p(D(u)|z, E = 0)} \\ = & \log \frac{p(w_1 w_2 \cdots w_n | z, E = 1)}{p(w_1 w_2 \cdots w_n | z, E = 0)} \\ = & \sum_{i=1}^n \log \frac{p(w_i | z, E = 1)}{p(w_i | z, E = 0)} \end{aligned}$$

基于上下文主题信息的概率主题方法

最后，得到最终的专业程度分数，计算公式如下：

$$\begin{aligned} & \log \frac{p(u|z, E = 1)}{p(u|z, E = 0)} \\ &= \sum_{w \in D(u)} \left\{ \log \frac{C_{wz}^{WT} + \beta}{\#(w, \mathcal{C}) + \mu} + \log \frac{\#(\cdot, \mathcal{C}) + \mu|\mathcal{V}|}{\sum_{w'} C_{w'z}^{WT} + \beta|\mathcal{V}|} \right\} \\ &= \left\{ \sum_{w \in D(u)} \log \frac{C_{wz}^{WT} + \beta}{\#(w, \mathcal{C}) + \mu} \right\} + |D(u)|\eta \end{aligned}$$

其中， $\eta = \log \frac{\#(\cdot, \mathcal{C}) + \mu|\mathcal{V}|}{\sum_{w'} C_{w'z}^{WT} + \beta|\mathcal{V}|}$ ， $|D(u)|$ 表示 $D(u)$ 词的数目。

Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

基于上下文主题信息的概率主题方法

- 候选用户 u 的知名度可以通过计算最佳答案的采纳率。
- 最佳答案的数目也是作为衡量用户知名度的一个重要指标。
- 专家的相对知名度分数可以用如下公式表示：

$$\begin{aligned} & \log p(R = 1|u) \\ &= \log \left\{ \frac{AR(u)}{\max AR(u')} \times \left[\gamma + (1 - \gamma) \frac{\#(BAs, u)}{\max \#(BAs, u')} \right] \right\} \end{aligned}$$

基于上下文主题信息的概率主题方法

采纳率 $AR(u)$ 可以定义如下：

$$AR(u) = \frac{\#(BAs, u)}{\#(As, u)}$$

其中， $\#(As, u)$ 表示用户 u 回答的答案 As 数目。

最后，得到如下的排序模型：

$$\begin{aligned} & \log p(E = 1, R = 1 | u, z) \\ &= \left\{ \sum_{w \in D(u)} \log \frac{C_{wz}^{WT} + \beta}{\#(w, \mathcal{C}) + \mu} \right\} + \log \left\{ \frac{AR(u)}{\max AR(u')} \right\} \\ &+ \log \left[\gamma + (1 - \gamma) \frac{\#(BAs, u)}{\max \#(BAs, u')} \right] + |D(u)|\eta \end{aligned}$$

数据集

数据集包含237,083个已解决的问题，593,107个答案，286,053个用户

| | |
|-------------------------------------|---------|
| Number of questions | 237,083 |
| Number of answers | 593,107 |
| Number of best answers | 162,733 |
| Number of total users | 286,053 |
| Number of askers | 180,166 |
| Number of answerers | 135,441 |
| Number of both askers and answerers | 29,554 |

实验结果

| # | Method | USER-15 | | | USER-10 | | |
|---|----------------------|---------|--------------------|--------|--------------------|--------------------|--------------------|
| | | nEQM@1 | nEQM@10 | MAP | nEQM@1 | nEQM@10 | MAP |
| 1 | HITS | 0.289 | 0.312 | 0.397 | 0.277 | 0.302 | 0.374 |
| 2 | InDegree | 0.280 | 0.296 | 0.372 | 0.262 | 0.293 | 0.348 |
| 3 | ExpertiseRank | 0.345 | 0.415 | 0.513 | 0.345 | 0.361 | 0.474 |
| 4 | PersonPR | 0.330 | 0.392 | 0.494 | 0.322 | 0.355 | 0.461 |
| 5 | CB | 0.361 | 0.445 | 0.543 | 0.351 | 0.378 | 0.482 |
| 6 | TSPR | 0.407* | 0.463 [†] | 0.576 | 0.366 [†] | 0.406 [†] | 0.512 [†] |
| 7 | TSPR + <i>ExpRep</i> | 0.430* | 0.492* | 0.641* | 0.399* | 0.443* | 0.556* |

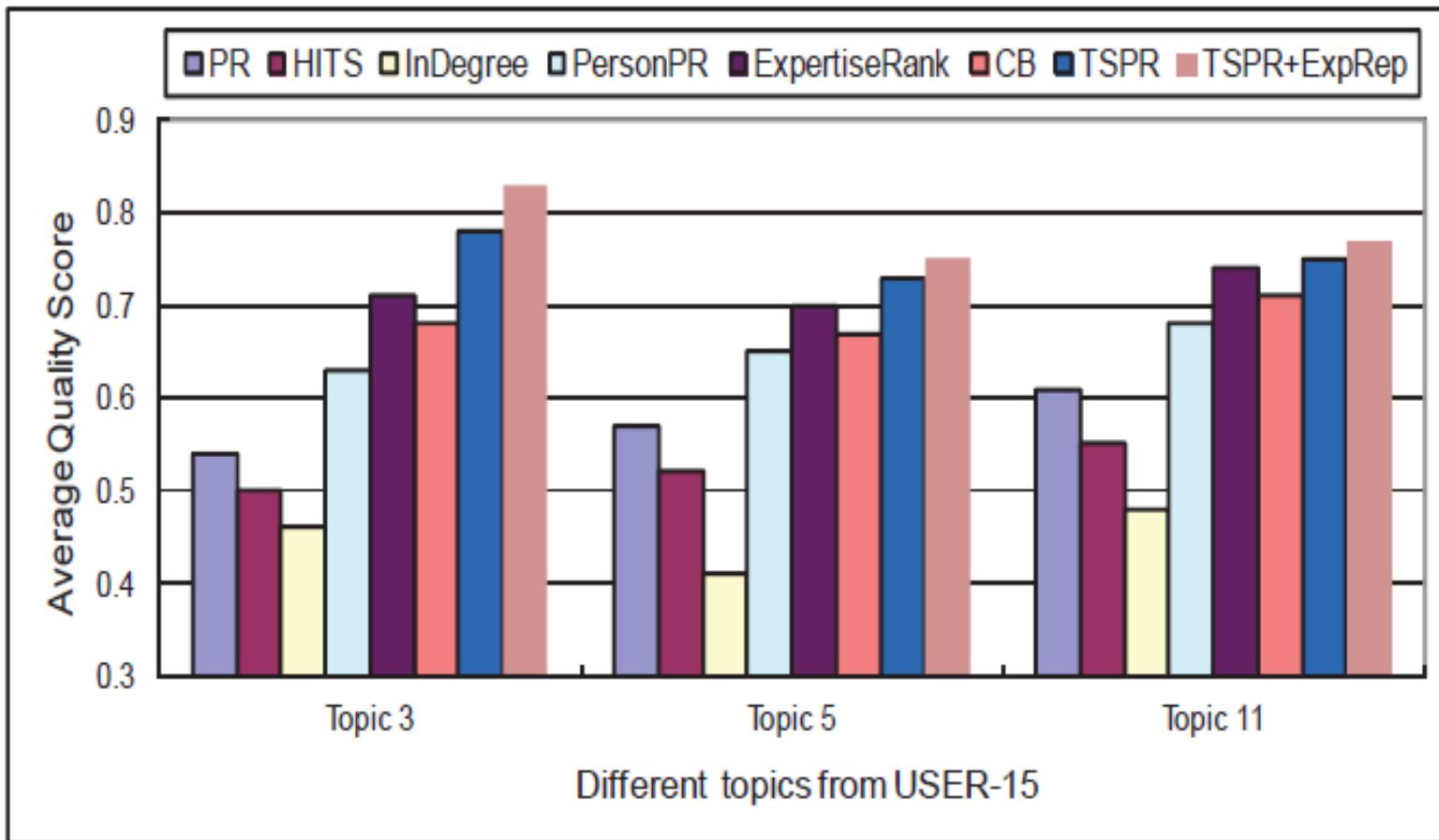
Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

实验结果：显示的数据做了处理

| 经济领域 | | | 法律领域 | | | 教育领域 | | |
|--------------|------|-----|----------------|------|-----|---------------|------|------|
| 用户名 | 出勤天数 | 回答量 | 用户名 | 出勤天数 | 回答量 | 用户名 | 出勤天数 | 回答量 |
| ① 靖立群 | 15 | 842 | ① 戚家人 | 15 | 917 | ① howshineyou | 15 | 1690 |
| ② yangss1230 | 15 | 791 | ② 曹正坤 | 15 | 612 | ② 马兴武 | 15 | 1144 |
| ③ 梁晓燕 | 15 | 538 | ③ 罗笙铭 | 15 | 517 | ③ 宇文仙 | 15 | 1123 |
| 4 多多HC信仰爱 | 15 | 427 | 4 周斌 | 15 | 492 | 4 兔二玉六 | 15 | 947 |
| 5 李迎 | 15 | 400 | 5 刘辉律师 | 15 | 451 | 5 黄祝荣 | 15 | 836 |
| 6 厦门平安保险人 | 15 | 396 | 6 sunlianxiang | 15 | 370 | 6 初中化学老师 | 15 | 815 |
| 7 监理师 | 15 | 379 | 7 北京中圣晖1 | 15 | 336 | 7 翁祖清 | 15 | 727 |
| 8 谢承 | 15 | 339 | 8 amidaya | 15 | 301 | 8 尹强 | 15 | 645 |

Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

实验结果



Guangyou Zhou, Tingting He, Jun Zhao and Wensheng Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. Knowledge-based systems, 66:136-145,2014. (SCI, IF=3.175)

致谢

CCF中文信息技术开放基金的支持

谢谢大家！

