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

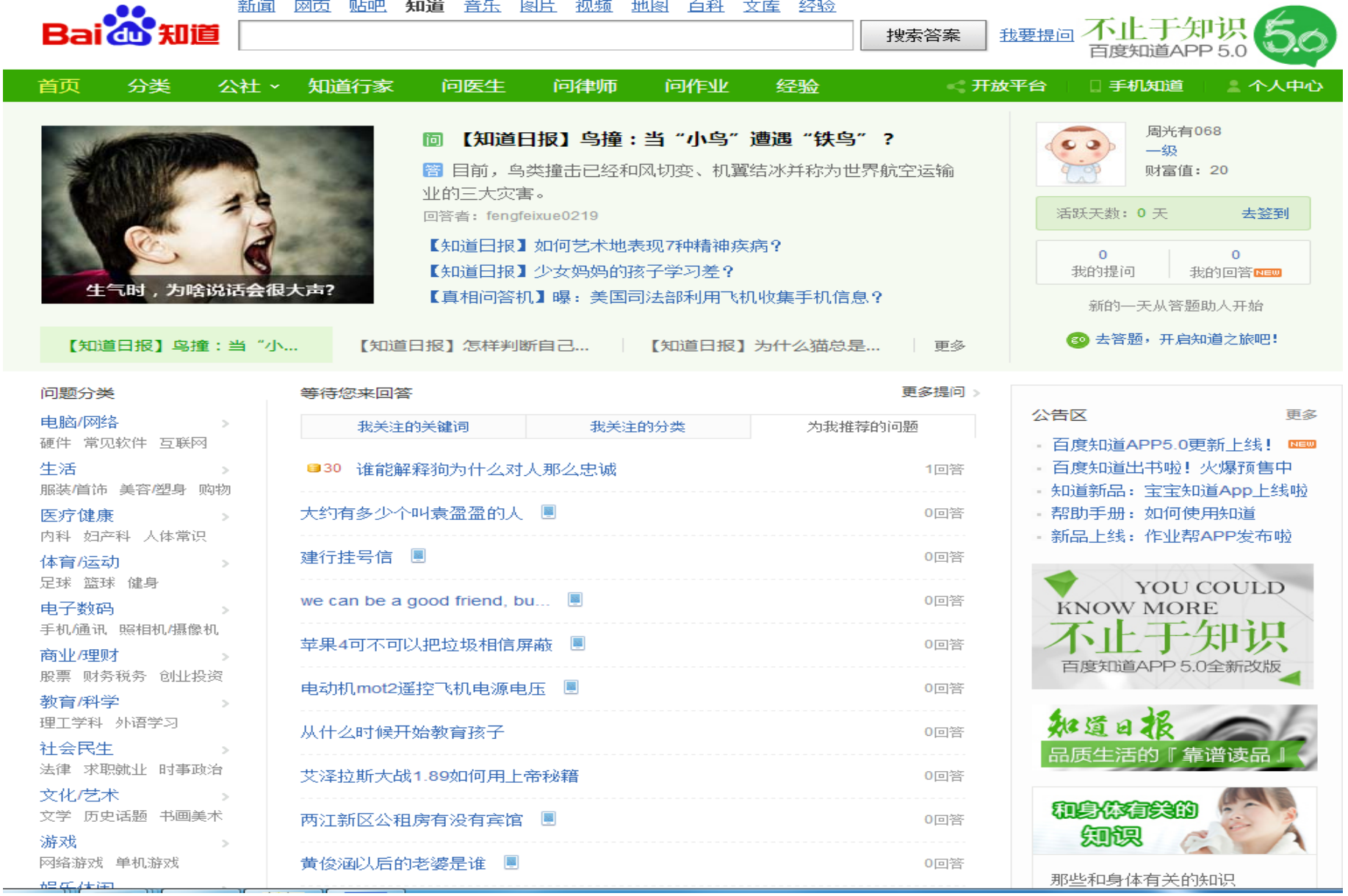
课题主要成果

- 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)



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)

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

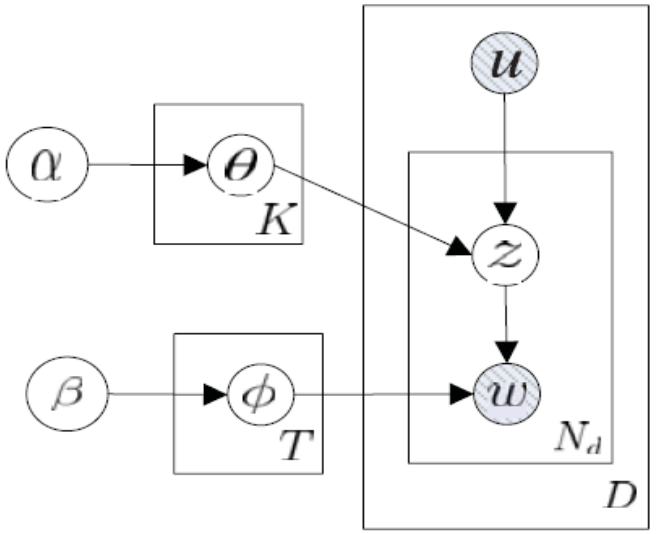


Figure: 用户-主题模型的图表示。

- 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

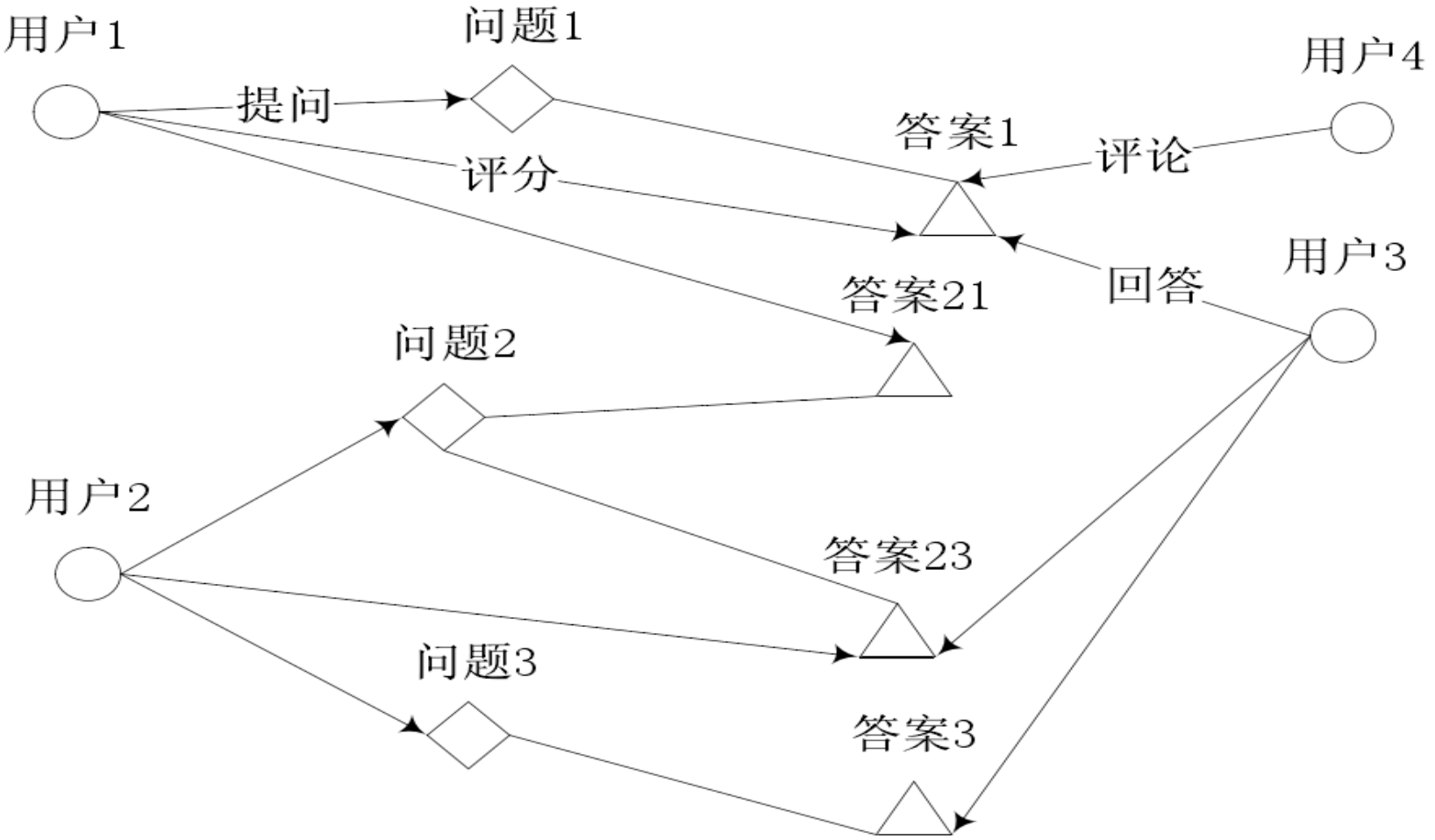
$$\phi_{wj} = \frac{C_{wj}^{WT} + \beta}{\sum_{w'} C_{w'j}^{WT} + \beta|\mathcal{V}|}$$

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

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。

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



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：

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

- $Q(i)$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$f(i \to i) = 0$

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：

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

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

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：

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

$sim_z(i \to j)$表示在该主题下用户u_i到u_j的相似度。

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

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

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

新的行归一化的矩阵定义如下：

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

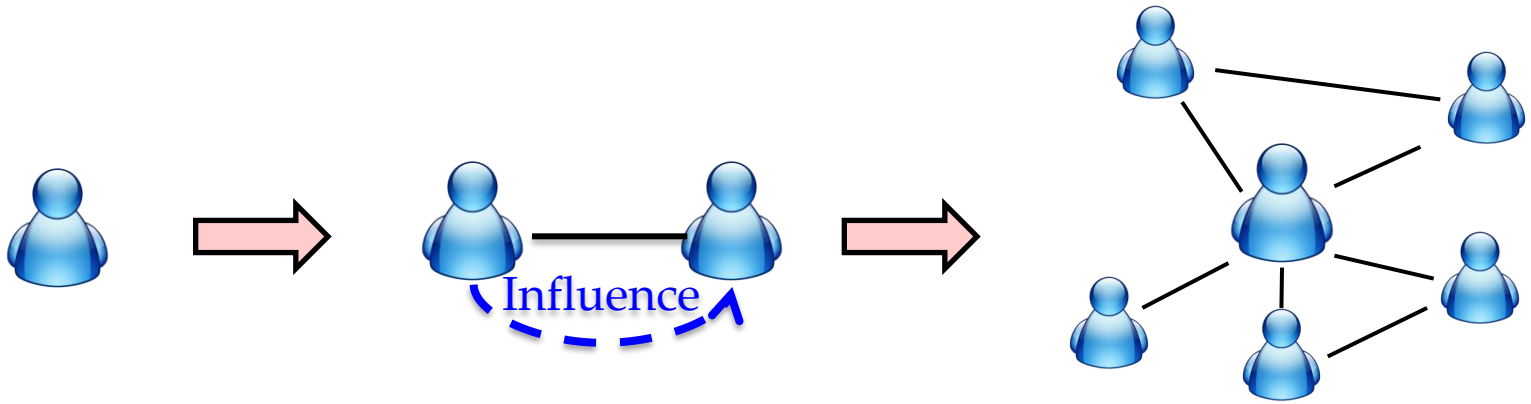
对于给定的主题z，利用TSPR得到的迭代公式如下：

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

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。

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



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

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

- 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：

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

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

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$\log p(R=1|u)$。

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$D(u)=(w_1,w_2,\cdots,w_n)$，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：

$$
\begin{aligned}
& \log \frac{p(u|z,E=1)}{p(u|z,E=0)} \\
= & \log \frac{p(D(u)|z,E=1)}{p(D(u)|z,E=0)} \\
= & \log \frac{p(w_1w_2\cdots w_n|z,E=1)}{p(w_1w_2\cdots w_n|z,E=0)} \\
= & \sum_{i=1}^{n} \log \frac{p(w_i|z,E=1)}{p(w_i|z,E=0)}
\end{aligned}
$$

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

最后，得到最终的专业程度分数，计算公式如下：

$$\log \frac{p(u|z,E=1)}{p(u|z,E=0)}$$

$$= \sum_{w \in D(u)} \left\{ \log \frac{C_{wz}^{WT} + \beta}{\#(w,\mathcal{C}) + \mu} + \log \frac{\#(\cdot,\mathcal{C}) + \mu|\mathcal{V}|}{\sum_{w'} C_{w'z}^{WT} + \beta|\mathcal{V}|} \right\}$$

$$= \left\{ \sum_{w \in D(u)} \log \frac{C_{wz}^{WT} + \beta}{\#(w,\mathcal{C}) + \mu} \right\} + |D(u)|\eta$$

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：

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

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

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：

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

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)

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

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

#	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 + *ExpRep***	**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

用户名	出勤天数	回答量
1 靖立群	15	842
2 yangss1230	15	791
3 梁晓燕	15	538
4 多多HC信仰爱	15	427
5 李迎	15	400
6 厦门平安保险人	15	396
7 监理师	15	379
8 谢承	15	339

法律领域

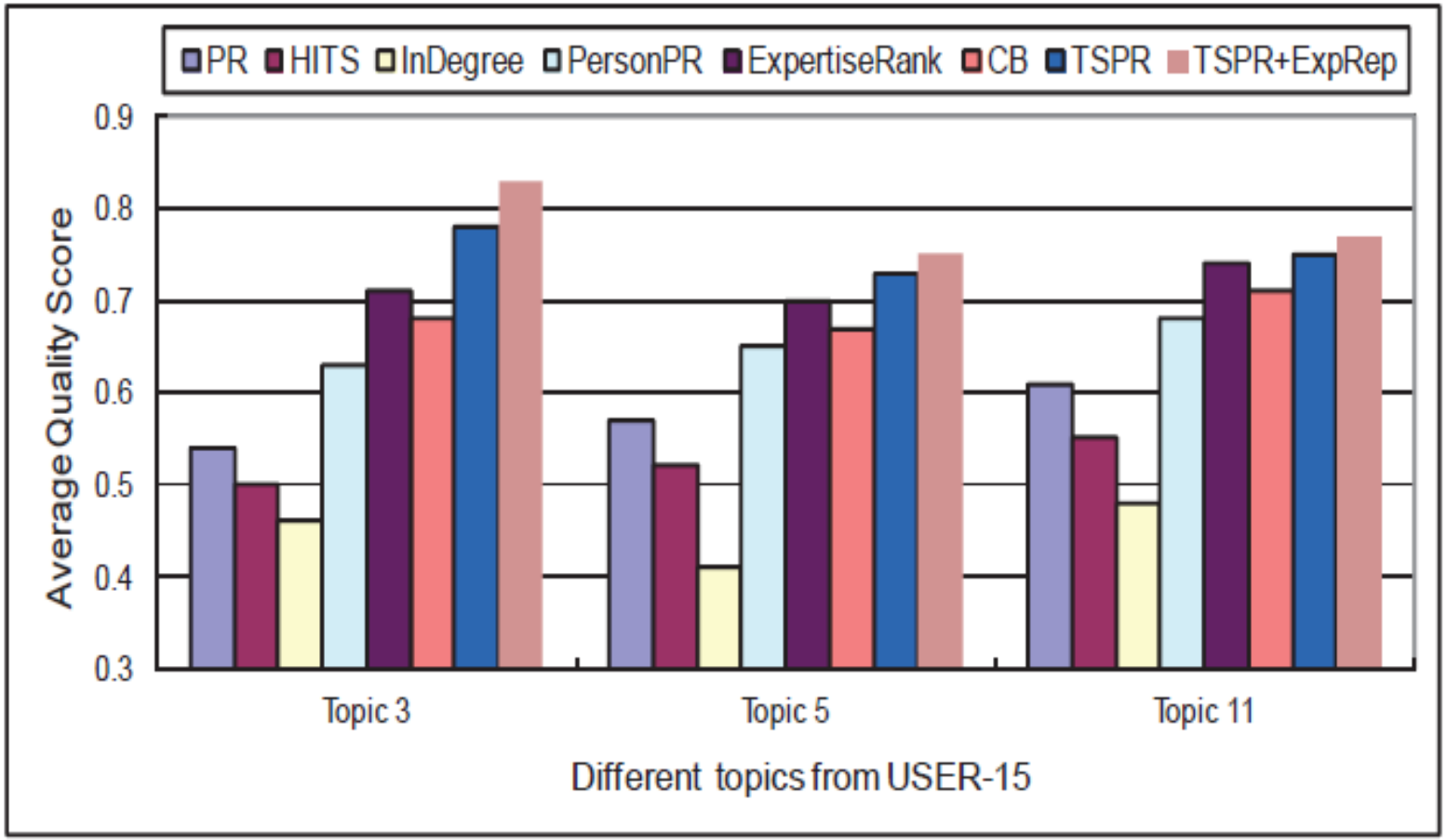
用户名	出勤天数	回答量
1 戚家人	15	917
2 曹正坤	15	612
3 罗笙铭	15	517
4 周斌	15	492
5 刘辉律师	15	451
6 sunlianxiang	15	370
7 北京中圣晖1	15	336
8 amidaya	15	301

教育领域

用户名	出勤天数	回答量
1 howshineyou	15	1690
2 马兴武	15	1144
3 宇文仙	15	1123
4 兔二玉六	15	947
5 黄祝荣	15	836
6 初中化学老师	15	815
7 翁祖清	15	727
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！