DOI：10.7652/xjtuxb201412009

# 健康问题和医生匹配机制的研究

朱利，岳爱珍
（西安交通大学软件学院，710049，西安）

**摘要**：针对医疗社区问答系统中的健康问题，提出了一种新的问题回答者推荐机制，以提高问题解决的效率。该机制引入了医生回答问题的态度，将问题-医生的专业匹配程度和医生回答问题的态度关联起来一同考虑；使用概率超图和查询似然语言模型对问题-医生的专业匹配进行建模，利用历史数据对医生的态度进行建模；使用回归模型对问题-医生的专业匹配和态度进行自适应权衡。进行了大量基于真实数据集的实验对所提出的机制进行了验证。实验结果表明：与常用的方法相比，本文所提出的机制准确度能提高 30%；很大程度上提高了问题解决的效率。

**关键词**：医疗社区问答系统；问题推荐；概率超图；态度建模
**中图分类号**：TP181 **文献标志码**：A **文章编号**：0253-987X(2014)12-0057-06

## Routing Health-Oriented Questions to Appropriate Doctors

ZHU Li, YUE Aizhen
(School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

**Abstract**: A novel mechanism connecting health seekers to appropriate doctors is proposed to improve the efficiency of question resolving in community-based health service systems. Attitudes of doctors answering questions are introduced in the mechanism, and both the professional matching degree between doctors and questions and the doctor's attitudes are associated and considered at the same time. The probabilistic hypergraph and the query likelihood language model are used to model the professional matching degree, and a doctor's attitude is modeled from his historical data. Meanwhile, a regression model is used to trade off between the professional matching degree and the doctor's attitude. Extensive experimental results on several real-world datasets show that the matching precision of the proposed mechanism increases by about 30%, and the efficiency of resolving problems is greatly improved.

**Keywords**: community-based health service system; question routing; probabilistic hypergraph; attitude modeling

随着信息技术和互联网的发展，医疗社区问答系统作为一种健康知识交流和分享的有效平台开始出现，例如 HealthTap、HaoDF 和 39 健康网等。医疗社区问答系统不仅允许咨询者免费咨询健康问题，而且鼓励医生提供可靠的答案，同时也为健康咨询者提供相似问题的检索。

医疗社区问答系统中，问题的增长速度通常比医生的增长速度快得多。一方面，健康咨询者不得不长时间等待医生的回答，等待数小时甚至数天；另一方面，医生往往会淹没在海量的问题中。因此，健康咨询者和医生之间存在一个日渐变宽的鸿沟，急需一个高效的“问题-医生”映射机制来弥补这个鸿沟。

收稿日期：2014-5-20。 作者简介：朱利(1968—)，男，副教授。 基金项目：国家重点基础研究发展规划资助项目(2012CB327902HZ)。
网络出版时间：2014-10-31 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20141031.1643.017.html

已有的问题回答者推荐研究可以分为两类:全局专家发现方法和主题级专家发现方法。前者是在特定论坛使用来自帖子和回复的链接来衡量专家的权威度[1]。例如 Bouguessa 等提出使用入度(最优答案数目)来找专家[2];基于链接的 HITS 算法比入度算法性能要高[3]。除了 HITS 算法外,另外一个基于链接的 PageRank 算法在社交媒体中也取得了很大的成功[4]。全局专家发现方法只能找到特定论坛的专家,没有考虑问题内容和专家的专业是否匹配;主题级的专家发现方法则在更细粒度上考虑了问题敏感的专业匹配问题。Zhou 等在文献[5]中提出了基于专业的在线论坛问题推荐。Li 等在文献[6]中深入研究了分类敏感的语言模型。除了专业估算外,在文献[7]中提出了考虑专家是否在线问题的基于语言模型的框架。主题级专家发现方法虽然可以根据问题的内容找到相应领域的专家,但不能获取专业能力与问题的匹配程度。更重要的是,目前尚未发现考虑专家态度问题的研究。为此,我们提出了一种针对医疗社区问答系统中医生推荐的机制 QDM(question-doctor mapping),同时考虑了问题和医生的专业匹配以及医生的态度,能够准确地解决"健康问题医生"之间的映射问题,提高了问题解答的效率。机制主要由 3 部分组成:①问题和医生的专业匹配建模;②医生的态度建模;③问题-医生的专业匹配和医生态度的自适应权衡。其中,①用来表示医生对问题的专业程度,②用来表示医生回答问题的态度。对于不同的问题,对医生的专业和态度有不同的倾向。因此,③用来自适应地权衡问题-医生专业匹配和医生态度的重要性。

## 1 问题和医生的专业匹配建模

使用 $D=\{d_1,d_2,d_3,\cdots,d_n\}$ 表示 $n$ 个医生集合,$d_i$ 表示一个医生。每个医生的信息包括两部分:医生的信息简介和医生回答过的问题答案对。简介部分包括医生的教育背景、出版物、奖励以及来自其他医生和健康咨询者的投票等信息。问题答案对包括问题、答案、和答案相关的标签以及其他医生对答案的赞同情况。我们将医生回答过的问题答案对看作其积累的经验。

使用 $E(d_i,q)$ 表示医生 $d_i$ 对问题 $q$ 的专业程度。受文档主题生成模型(latent dirichlet allocation, LDA)的启发,我们认为每个医生的专业知识是综合的,呈现加权分布特征。每种专业被解释为可以使用相同专业知识回答的问题集合。因此,可以得到医生和所给问题的相关程度为

$$E(d_i,q)=\sum_j p(\varepsilon_j\mid d_i)p(q\mid\varepsilon_j) \tag{1}$$

式中:$p(\varepsilon_j\mid d_i)$ 表示医生 $d_i$ 在专业 $\varepsilon_j$ 上的专业程度,即医生-专业分布,而 $p(q\mid\varepsilon_j)$ 表示问题 $q$ 需要专业 $\varepsilon_j$ 解决的概率,即专业-问题分布。

### 1.1 医生-专业分布

通过对 HealthTap 上的 3 000 多个医生的简介进行分析,发现每个医生平均有 3.4 个专业技能。因此,$p(\varepsilon_j\mid d_i)$ 的计算就可以看作是在医生集合 $D$ 上的模糊专业聚类。不像传统的硬聚类,每个医生只属于一个集合。在这里,医生可根据相应概率 $p(\varepsilon_j\mid d_i)$ 属于多个集合。

目前存在多种计算 $p(\varepsilon_j\mid d_i)$ 的聚类技术,例如 $K$ 均值、LDA 和基于简单图[8]的聚类,可是这些方法都存在一些约束。首先,这些方法通常假设所感兴趣的物体之间的关系成对存在。在我们研究的工作中,医生之间的关系比成对关系更复杂,如果简单地将这种复杂关系转换成成对关系将会导致我们想要了解的信息丢失。其次,它们不能处理异构信息。在医疗社区问答系统中,医生通常同时拥有几种关系,例如医生间的社交关联、相似的简介和经验等。为了解决这些问题,我们构建概率超图,然后在超图上进行模糊划分。超图允许超边连接两个及以上的顶点。同时,不同类型的超边可以表示不同的异构关系。

超图 $(V,E,W)$ 由顶点集合 $V$、超边集合 $E$ 和超边的权重集合 $W$ 组成。每条超边 $e$ 连接两个或两个以上的顶点,且 $e$ 都分配一个权值 $w(e)$。在所研究的问题中,医生集合 $D$ 中的 $n$ 个医生被看作 $n$ 个顶点。根据医生的信息,可以构建 3 种类型的超边。对于第一种类型的超边,每个医生作为一个顶点,该医生和与他简介相似度最高的 $k$ 个医生组成一条超边。这种构建超边的方式在文献[9]中第 1 次被采用。第 1 种类型的超边集合定义为 $E_1$。第 2 种类型的超边是基于医生积累的经验。对于每个医生,将他回答过的所有问题答案对合并成一个文档,使用这个文档来表示该医生积累的经验。将具有相同积累经验的医生构建成一条超边,这种超边集合定义为 $E_2$。第 3 种类型的超边利用了医生之间的社交关系。对于超图中的每个医生,将该医生和与他回答过相同问题的医生组成一条超边,这种超边集合定义为 $E_3$。

概率超图 $G$ 可以被表示为 $|V|\times|E|$ 的关联矩

阵 $\boldsymbol{H}$，$\boldsymbol{H}$ 中的元素为

$$h(d_i,e_j)=\begin{cases}p(d_i,e_j), & d_i\in e_j\\ 0, & d_i\notin e_j\end{cases} \tag{2}$$

式中：$p(d_i,e_j)$ 是超边 $e_j$ 连接顶点 $d_i$ 的概率。$p(d_i,e_j)$ 的定义为

$$p(d_i,e_j)=\begin{cases}S(d_i,d_{e_j}), & e_j\in E_1 \text{ 或 } e_j\in E_2\\ 1, & e_j\in E_3\end{cases} \tag{3}$$

式中：$d_{e_j}$ 是超边 $e_j$ 连接的医生；$S(d_i,d_{e_j})$ 是医生 $d_i$ 和医生 $d_{e_j}$ 简介（第 1 类超边）或积累的经验（第 2 类超边）的相似度。

超边的权重大小表示超边中的顶点属于同一组的可能程度。对于一条超边，它的权重定义为

$$W(e_j)=\sum_{d_i\in e_j}h(d_i,e_j) \tag{4}$$

式中：$d_i\in e_j$ 表示超边 $e_j$ 连接顶点 $d_i$。

对于每条超边，它的度定义和权重相同

$$\delta(e_j)=\sum_{d_i\in e_j}h(d_i,e_j) \tag{5}$$

根据 $\boldsymbol{H}$ 的定义，顶点 $d_i$ 的度为

$$d(d_i)=\sum_{e_j\in E}W(e_j)h(d_i,e_j) \tag{6}$$

在本文中，利用了一种高效且实现简单的算法[10]来划分超图，但是与算法中的超图不同，我们构建的超图是一个概率模型。在构建的超图上定义了正则项

$$\Omega(\boldsymbol{f})=\frac{1}{2}\sum_{e\in E}\sum_{u,v\in V}\frac{w(e)h(u,e)h(v,e)}{\delta(e)}\cdot\left(\frac{f(u)}{\sqrt{d(u)}}-\frac{f(v)}{\sqrt{d(v)}}\right)^2 \tag{7}$$

式中：矩阵 $\boldsymbol{f}\in\mathbf{R}^D$ 包含了每个医生和想要学习的、潜在的专业类别相关概率。通过定义，可得

$$\Omega(\boldsymbol{f})=\sum_{e\in E}\sum_{u,v\in V}\frac{w(e)h(u,e)h(v,e)}{\delta(e)}\cdot\left(\frac{f^2(u)}{d(u)}-\frac{f^2(v)}{d(v)}\right)=\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{\Theta})\boldsymbol{f} \tag{8}$$

式中：$\boldsymbol{I}$ 是单位矩阵。定义 $\boldsymbol{\Delta}=\boldsymbol{I}-\boldsymbol{\Theta}$，$\boldsymbol{\Delta}$ 是一个半正定矩阵，即超图的拉普拉斯算子[10]。$\Omega(\boldsymbol{f})$ 可被重新写为

$$\arg\min_{\boldsymbol{f}\in\mathbf{R}^D}\Omega(\boldsymbol{f})=\boldsymbol{f}^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{f} \tag{9}$$

通过最小化 $\Omega(\boldsymbol{f})$，可以确保相关概率函数在语义空间内是连续且光滑的。这意味着具有相似专业医生的相关概率是相似的。以上最优化问题的理论解决方法是求出 $\boldsymbol{\Delta}$ 最小非零特征值对应的特征矩阵。由于需要在超图中实现多路聚类，我们求出 $\boldsymbol{\Delta}$ 的 $T$ 个最小非零特征值对应的特征向量作为医生在 $T$ 个专业类别上的分布，$f_i^{(j)}$（$1\leqslant j\leqslant T$）即为 $p(\varepsilon_j|d_i)$。

### 1.2　专业-问题分布

根据超图划分，使用 $f_i^{(j)}$ 表示医生 $d_i$ 属于专业类别 $\varepsilon_j$ 的概率。因此，对于每个专业类别，得到按照 $p(\varepsilon_j|d_i)$ 降序排列的医生列表。合并排在前 100 个医生回答过的问题答案对，记为 $q_{\varepsilon_j}$，使用 $q_{\varepsilon_j}$ 表示专业类别 $\varepsilon_j$。我们使用 QLLM[11] 的方法来估计给定问题和每个专业类别的关联程度。由 QLLM 得

$$p(q\mid\varepsilon_j)=p(q\mid q_{\varepsilon_j}) \tag{10}$$

$$p(q\mid q_{\varepsilon_j})=\prod_{w\in q}P(w\mid q_{t_j}) \tag{11}$$

根据 Jelinek-Mercer 平滑法得

$$P(w\mid q_{\varepsilon_j})=(1-\alpha)P(w\mid q_{\varepsilon_j})+\alpha P(w\mid C) \tag{12}$$

$$P(w\mid q_{\varepsilon_j})=\frac{f(w,q_{\varepsilon_j})}{\sum_{j_{w'}}f(w')} \tag{13}$$

$$P(w\mid C)=\frac{f(w,C)}{\sum_{j_{w'}}f(w')} \tag{14}$$

式中：$C$ 是所有的问题集合；$\alpha$ 是一个调整平滑权重的加权系数；$f(w,q_{\varepsilon_j})$ 表示项 $w$ 在专业 $\varepsilon_j$ 的问题集合 $q_{\varepsilon_j}$ 中出现的频率；$f(w,C)$ 表示项 $w$ 在所有问题集合 $C$ 中出现的频率。根据经验值，设置 $\alpha=0.8$。至此，可以得到每个医生和给定问题的专业相关程度 $E(d_i,q)$。

## 2　医生态度建模

除了问题和医生的专业匹配外，根据文献[12]中的研究，我们假设问题答案的质量也取决于医生的态度。根据医疗问题回答系统中的可用信息，本文从积极性、责任感和声誉 3 个不同的角度对医生的态度进行建模。这些都需要使用历史数据进行估算，它们的乘积表示医生的态度。

积极性可用来测量问题出现时医生的积极程度。积极性的定义为

$$A(d_i)=\frac{N_{\mathrm{f}}(d_i)}{N_{\mathrm{a}}(d_i)} \tag{15}$$

式中：$N_{\mathrm{f}}(d_i)$ 表示医生 $d_i$ 是第 1 个问题回答者的问题数目；$N_{\mathrm{a}}(d_i)$ 表示医生 $d_i$ 回答的问题数目；$A(d_i)$ 表示医生 $d_i$ 回答问题的积极性，$A(d_i)$ 越大，医生 $d_i$ 回答问题的时间就越短。

责任感用来测量医生对给定问题回答的满意程度，它直接反映在答案质量上。我们认为如果一个

医生 $d_i$ 回答了问题 $q$，则 $d_i$ 有相应的专业能力解决 $q$。同时，如果提供的答案被其他医生选为最优答案，我们认为医生 $d_i$ 对该答案有责任。$d_i$ 的责任感的估算为

$$R(d_i)=\frac{N_b(d_i)}{N_a(d_i)} \tag{16}$$

式中：$N_b(d_i)$表示医生 $d_i$ 提供的答案被选为最优答案的数目。

医生的声誉是指其他医生和健康咨询者对这个医生的看法或想法。本文使用取值范围在 0～1 之间的 Sigmoid 函数来估算名声。

$$S(d_i)=\frac{1}{1+e^{(N_p(d_i)+N_s(d_i))}} \tag{17}$$

式中：$N_p(d_i)$和 $N_s(d_i)$分别表示支持医生 $d_i$ 的医生和健康咨询者的数目。

## 3 专业匹配和态度自适应权衡

对于一个新的用自然语言描述的问题 $q$，我们的目标是从 $D$ 中选择出一些匹配的医生，并且将 $q$ 推荐给这些医生。匹配分数为

$$S(d_i,q)=(1-\lambda)E(d_i,q)+\lambda A(d_i) \tag{18}$$

式中：$E(d_i,q)$是问题和医生的专业匹配模型，表示从专业的角度考虑医生 $d_i$ 可以回答问题 $q$ 的可能性。$A(d_i)$是态度模型，可以从历史行为中推测出来。另外，$\lambda$ 是一个自适应的参数，用来平衡专业和态度的影响。

根据观察，不同的问题对专业和态度有不同的倾向。对于简单问题，入门级医生就可以回答。这种问题答案的质量主要由医生的态度决定而不是医生的专业知识。对于高难度问题，需要根据病人症状找出发病的原因以及告诉病人应该如何做，此时医生的专业知识对于给出高质量回答起到了主要作用。因此可以得出结论，参数 $\lambda$ 是一个关于给定问题的自适应函数，它平衡专业和态度对答案质量的重要性。当 $\lambda=1$ 时，给定的问题将被推荐给态度最好的医生，而他们的专业能力将被忽视。相反，如果 $\lambda=0$，将不考虑医生的态度问题。

将 $\lambda$ 的自适应估计任务看作监督回归问题，其目标是根据类似训练问题的有效权重，为每个新的问题预测一个合适的权重。对于训练集中的每个问题，首先使用固定的 $\lambda$ 来估算问题推荐的性能。$\lambda$ 的最优值通过在[0,1]之间使用固定步长得到。利用这些最优值作为真实值来训练回归模型。在实验中，使用了线性回归、保序回归和 pace 回归等不同的回归模型。

## 4 实 验

### 4.1 实验设置

在实验中使用了从 HealthTap 上收集的 3 123 个医生的简历。每个医生的简历包括医生的简介和该医生以前回答过的问题答案对。表 1 展示了我们统计的实验数据。

表 1 收集到的数据的统计

| 医生数目 | 问题数目 | 答案数目 | 答案标签数目 |
|---|---|---|---|
| 3 123 | 142 025 | 235 162 | 1 659 668 |
| 答案的平均标签数目 | 答案的平均医生支持数目 | 来自医生投票数的平均值 | 来自健康咨询者投票数的平均值 |
| 7.06 | 1.40 | 82.37 | 77.65 |

本文使用了基于 LDA 的主题级别的数据表示。对于一个数据集，LDA 按照语义将内部相关联的健康概念分配到一个潜在的组，它可以按照主题描述健康数据的底层语义结构。每个潜在组被视为一维特征。特征空间维度通过困惑度(perplexity)[10]得到。困惑度是一种计算统计模型，设为

$$p_e=\exp\left\{\frac{\sum_{i=1}^{m}\log p(d_i)}{\sum_{i=1}^{m}l_i}\right\} \tag{19}$$

式中：$l_i$ 表示 $d_i$ 的词数。困惑度的值越小表示所用的 LDA 模型越好。我们将医生的简历分为两部分，80%的数据用来训练 LDA 模型，20%用来评估性能。LDA 建模和困惑度矩阵通过 Stanford 建模工具集来实现。当潜在组的组数变化时，困惑度取值如图 1 所示。由图 1 可知，当潜在组数为 110 时，困惑度最低。因此，对于给定的一个医生或者问题，它可以表示成 110 个语义主题级别特征的混合。

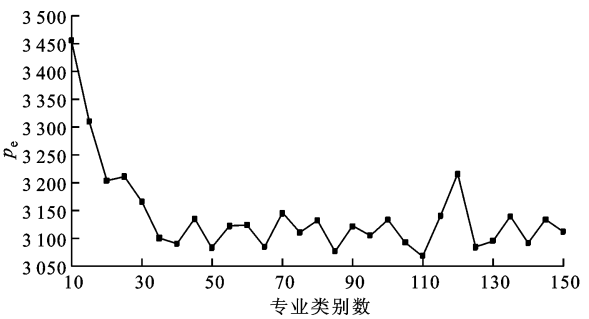


图 1 困惑度随专业类别数变化曲线

对于随后的主观评价，我们邀请了 3 个来自不同背景的自愿者并进行了训练。在 3 个志愿者中采用

多数表决方案来解决有分歧的问题。对于那些有两类选票数相同的情况，通过讨论来获得最后的决定。

### 4.2　问题推荐性能比较

对于专家推荐问题，准确度是最重要的性能指标。因此，采用了客观评估和主观评估两种指标从不同的方面获取准确度。从数据集中随机选择了 1 000个问题作为测试样本。对于客观评估，采用了平均的 $H@K$[13]。如果真正回答该问题的医生排在前 $K$ 位，则认为 $H@K=1$，否则 $H@K=0$。这种评估方法的优点是使用了真实数据且不需要构建其他的真实值。可是这种评估方法可能受这种场景的影响，医生 $d_i$ 回复了问题 $q$，尽管医生 $d_j$ 有能力回答问题 $q$，但由于未知的原因没有回答问题。尽管本文提出的机制很可能将 $d_j$ 排在一个高的位置，但是 $H@K$ 却忽略了。因此，$H@K$ 不能全面公平地评估本文提出的推荐机制。

作为对 $H@K$ 的补充，采用了主观的指标 $S@K$，来测量在推荐的 $K$ 个医生中能找到一个匹配的医生回答问题的概率，即如果推荐的前 $K$ 个医生中有能力回答该问题，则 $S@K=1$，否则 $S@K=0$。不同于主观评估，此时的真实值需要志愿者手动构建。他们需要查看医生的简介和历史数据，如果认为某个医生有能力并且有可能回答这个问题，将标记该医生有能力回答该问题，否则标记不能回答该问题。本文中使用 Kappa 分析[14]来评估志愿者间的一致性。Kappa 的值在 0～1 之间，值越大，一致性越高。Kappa 值大于 0.7 表示一致性很强。在本文的工作中，采用了在线的 Kappa 计算工具。表 2 展现了分析结果，结果证明它能标记志愿者之间的一致性。

表 2　使用 Kappa 方法的志愿者之间的一致性评估

| 案例数 | 类别数 | 自愿者人数 |
|---|---|---|
| 1 000 | 2 | 3 |
| 全部一致率 | 固定边界 Kappa 值 | 自由边界 Kappa 值 |
| 97.33% | 76.36% | 94.67% |

我们将本文提出的问题推荐机制与最新技术进行了比较。为了确保公平，它们将同时考虑或不考虑态度问题。

$K$ 均值为医生-专业相关程度使用 $K$ 均值估算的机制。LDA 为医生-专业相关程度使用 LDA 估算的机制。

由于 K 均值的结果是离散的，只能得到一个医生是否属于某个专业类别，而得不到该医生属于某个专业类别的概率。为了得到医生属于某个专业类别的概率，本文使用下式计算

$$p(\varepsilon_j \mid d_i) = \exp\left(-\frac{d_i - c_j}{\sigma^2}\right) \tag{20}$$

式中：半径参数 $\sigma$ 是所有医生对之间的欧氏距离的均值，$c_j$ 是专业类别 $\varepsilon_j$ 的中心。

图 2 和图 4 分别列出了使用 $H@K$ 和 $S@K$ 来评估推荐性能的对比结果。当引入医生的态度时，结果在图 3 和图 5 中给出。由图可见，本文方法的性能明显高于其他方法。综合分析这 4 个图可以看出，考虑医生态度时的性能要高于不考虑医生态度时的性能，这证明了医生的态度确实影响答案的质量。另一方面，实验结果也证明使用自适应的 λ 来平衡专业匹配和态度的影响比使用固定 λ 的要优越。

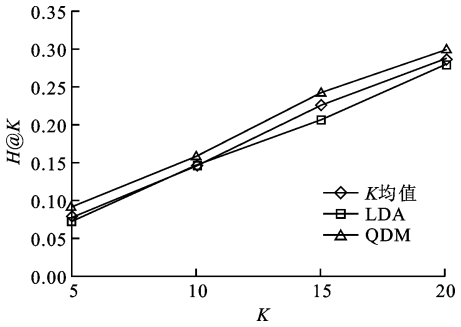


图 2　不考虑医生态度时使用 $H@K$ 评估的性能比较

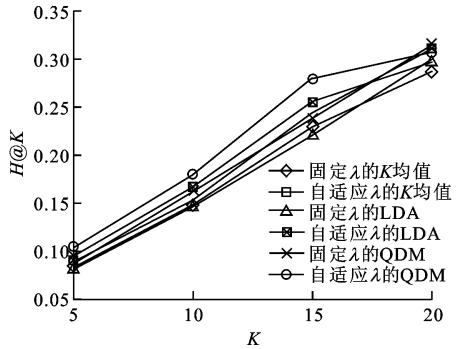


图 3　考虑医生态度时使用 $H@K$ 评估的性能比较

### 4.3　专业和态度权衡

从数据集中随机选择了 1 000 个问题。对于每个问题，使用固定步长 0.05 在[0，1]之间找到了最优性能。为了节省时间，性能通过 $H@K$ 测量。这些问题分为两部分，80%用来训练，20%用来测试。我们采用了 4 种回归模型，它们的性能比较见表 3。可以看出，pace 回归模型取得了最好的性能。对于一个新问题，它的自适应参数 λ 是可预测的，这个值影响到问题需要对医生的专业更关注还是对态度更

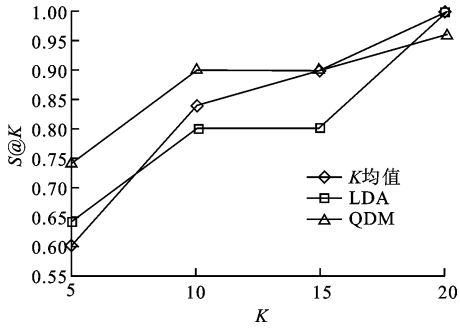


图 4 不考虑医生态度时使用 $S@K$ 评估的性能比较

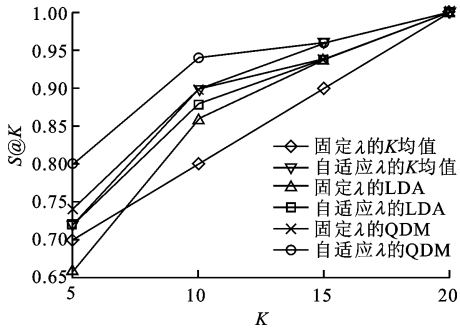


图 5 考虑医生态度时使用 $S@K$ 评估的性能比较

关注。

表 3 使用平均绝对误差回归模型的性能比较

| 平均绝对误差 | | | |
|---|---|---|---|
| Pace 回归 | 线性回归 | 保序回归 | 简单线性回归 |
| 0.027 4 | 0.027 9 | 0.027 7 | 0.029 7 |

## 5 结 语

本文研究的医疗社区中的问题推荐方法包含 3 个主要步骤:①对医生和问题的专业匹配度进行建模;②从积极性、责任感和名声 3 个角度对医生的态度进行建模;③根据问题的内容自适应地权衡问题和医生的专业匹配和医生态度的影响。实验证明,本文提出的"健康问题-医生"推荐机制具有很高的准确度和问题回答效率,可实际应用于网上医疗社区中,能够弥补目前这个应用方面的空白。

由于目前很难获得中文的问题集,本文的方法和实验都是针对英文问题集的。后面我们将通过技术方法收集中文数据集,来验证本文提出机制的通用性,并尝试将该机制应用到其他领域。

**参考文献:**

[1] JURCZYK P, AGICHTEIN E. Discovering authorities in question answer communities by using link analysis [C]// Proceedings of the 16th ACM Conference on Information and Knowledge Management. New York, USA: ACM, 2007: 919-922.

[2] MOHAMED B, BENOIT D, WANG Shengrui. Identifying authoritative actors in question-answering forums: the case of yahoo! answers [C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2008: 866-874.

[3] KLEINBERG J M. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM, 1999, 46(5): 604-632.

[4] ZHOU Guangyou, LAI Siwei, LIU Kang, et al. Topic-sensitive probabilistic model for expert finding in question answer communities [C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2012: 1662-1666.

[5] ZHOU Yanhong, CONG Gao, CUI Bin, et al. Routing questions to the right users in online communities [C]// Proceedings of the 25th IEEE International Conference on Data Engineering. Piscataway, NJ, USA: IEEE, 2009: 700-711.

[6] LI Baichuan, KING I, LYU M R. Question routing in community question answering: putting category in its place [C]// Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2011: 2041-2044.

[7] LI Baichuan, KING I. Routing questions to appropriate answerers in community question answering services [C]// Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2010: 2041-2044.

[8] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展 [J]. 软件学报, 2006, 17(9): 1848-1859. SU Jinshu, ZHANG Bofeng, XU Xin. Advances in machine learning based text categorization [J]. Journal of Software, 2006, 17(9): 1848-1859.

[9] HUANG Yuchi, LIU Qingshan, ZHANG Shaoting, et al. Image retrieval via probabilistic hypergraph ranking [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2010: 3376-3383.

[10] HUANG Yuchi, LIU Qingshan, LV Fengjun, et al. Unsupervised image categorization by hypergraph partition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(6): 1266-1273.

(下转第 139 页)