文章编号：2095-2163(2021)02-0035-05 中图分类号：TP391 文献标志码：A

基于 CTM 模型的在线轻问诊医生推荐研究

张锦红，张云华
（浙江理工大学 信息学院，杭州 310018）

摘 要：本文采用 CTM 主题模型对现有的在线医生专家推荐模型进行优化，首先利用患者提出的健康问题，得到问题-主题概率分布，然后根据医生历史回答的所有问题得到医生-主题概率分布，接着对得到的两项分布用杰卡德相似系数计算方法计算相似度，进而将主题相似度高的医生列表推荐给患者。实验阶段先对好大夫在线轻问诊模块的过敏反应科的数据进行采集和处理，再进行建模与测试，结果证实本文提出的医生推荐方法比该科室现存推荐方法更高效。

关键词：CTM；专家推荐；在线轻问诊

Research on doctor recommendation for online light consultation based on CTM model

ZHANG Jinhong, ZHANG Yunhua
(School of Informatics Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

【Abstract】 This paper uses the CTM topic model to optimize the existing online doctor expert recommendation model. Firstly, the paper uses the health questions raised by the patient to obtain the question-topic probability distribution, secondly obtains the doctor-topic probability distribution based on all the questions answered by the doctor's history. Then the paper uses the Jackard similarity coefficient calculation method to calculate the similarity of the obtained two distributions, finally recommends a list of doctors with high topic similarity to the patient. In the experimental stage, the data of the Allergic Reactions Department of the Doctor Online Inquiry Module is collected and processed, and modeling and testing are performed. The results confirm that the doctor recommendation method proposed in this article is more efficient than the existing recommendation method in the department.

【Key words】 CTM; expert recommendation; online light consultation

0 引 言

随着互联网技术的快速发展及广泛应用，医疗也不再局限于线下看医生，很多轻微疾病用户会选择在互联网上咨询疾病问题。此时，患者会在就医网站上诉说自己的身体状况，医生根据患者的病情描述回答患者的问题并同步给出健康问题解决方案[1]，可以达到资源合理配置的效果。虽然就目前来讲在线医疗轻问诊医生推荐研究取得了很大的突破，但有些方面仍然亟待优化，主要包括以下 3 点：

（1）当患者根据自身的健康状况在网络上寻求帮助时，往往因为信息量过大、且在描述上有失精准而显得无所适从。再者，部分患者几乎不了解相关医学知识，就可能在选择合适医生进行轻问诊上存在困难，而选定医生也因为患者问询诊治领域与自身专业方向并不匹配，如此就失去了在线医疗解决身体小疾患的意义。

（2）当前已推出不少提供患者和医生在线沟通的互联网平台，但医生却要在大量的咨询中耗费精力筛选自己可以解答的问题，医生资源得不到充分利用，大大降低了在线轻问诊的效率。

（3）目前在线医疗轻问诊平台中，用户不能及时得到解答服务，从寻求帮助到得到方案需要的时间具有不确定性[2]。因此，通过科学合理的专家推荐方法来充分利用医生资源以及提升用户满意度就显得尤为必要[3]。

综合前面问题所述，本文拟研究面向在线患者轻问诊的医生推荐主题模型，通过利用患者提出的待匹配健康问题与医生专家的历史回答健康问题的主题提取以及主题相似度的匹配，当患者提问时将合适的医生推荐给患者，并将患者的病情推送给专业的医生做病理解析，在一定程度上能够确保患者快捷、高效地获得健康问题解决方案，同时提高在线医疗轻问诊服务的效率和准确性及有效性[4]。

作者简介：张锦红(1996-)，女，硕士研究生，主要研究方向：软件工程、智能信息处理；张云华(1965-)，男，博士，研究员，主要研究方向：软件工程、系统仿真、智能信息处理。

通讯作者：张锦红 Email：1913415651@qq.com

收稿日期：2020-11-18

1 研究综述

与传统的关键字检索相比,社区问答系统能更好地满足用户对快速、准确获取信息的需求。因此,对问题的精准处理可以有效帮助社区问答系统抽取出更好的答案[5]。

主题识别主要通过共词分析和概率模型来实现,并抽取词汇来对主题进行表征[6]。迄今为止,主题模型已经发展了 20 余年,作为篇章级别文本语义理解的重要工具,pLSA (probabilistic Latent Semantic Analysis)就成为早期概率主题模型的典型代表。随后,Blei 等人在 2003 年提出的 LDA 模型则标志着对主题模型的研究进入热潮。

图 1 主题模型发展史

Fig. 1 The history of the theme model

隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)是常见的主题模型。由于 LDA 是非监督学习模型,本身不可直接用于分类,需将其嵌入到适合的分类算法中。许多学者基于 LDA 模型建立主题模型,包括 Blei 和 Lafferty 提出的相关主题模型(CTM)[7]、Li 和 Andrew McCallum 用无向图表示文档隐含主题结构的 PAM 模型[8]以及 RosenZvi 等人提出的作者主题模型(ATM)[9]等等。

其中,CTM 主题模型可以很好地展现主题间的相关性,并且文本主题数目对 CTM 模型的性能相当重要。LDA 主题模型采用狄利克雷分布(Dirichlet distribution)模拟文档生成过程, CTM 用对数正态分布替换 LDA 的狄利克雷分布对文档集隐含的主题进行提取,并引入协方差矩阵来描述主题间的相关性,解决了 LDA 主题之间不相关的问题[10]。

CTM 主题模型的框架如图 2 所示。此模型假定某个词汇拥有丰富的语义信息,某个主题的语句会含有和此主题相关的词汇。便可以通过探索语料库中频繁组合出现的词汇组来挖掘深层次的主题信息。利用这一方法,把待分析的文档建模成为拥有潜在主题信息的随机混合模型,模型中的语句含有的每个主题特征取决于语句中单词的特定分布,即为主题-词汇分布。

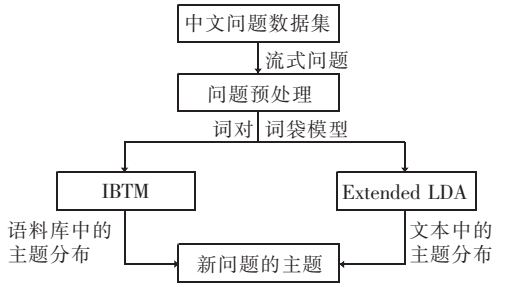

图 2 CTM 主题模型框架图

Fig. 2 CTM theme model framework diagram

2 基于 CTM 构建在线轻问诊医生推荐模型

常规的推荐算法大体上是根据问题和医生的二元关系来建立推荐模型,与传统的推荐算法相比,本文拟要建立的是问题-专长-医生的三元关系模型。三元模型能最大化地提高医生回答效率以及改善用户体验。当对文本进行提取时,对于健康问题的主题之间则会存在相关性,语句中包含的每个主题并非是完全独立的,本文选用的 CTM 模型就能很好地解决这个问题。本次研究分 3 个步骤完成在线轻问诊医生的推荐,整体的步骤流程框架如图 3 所示[11]。

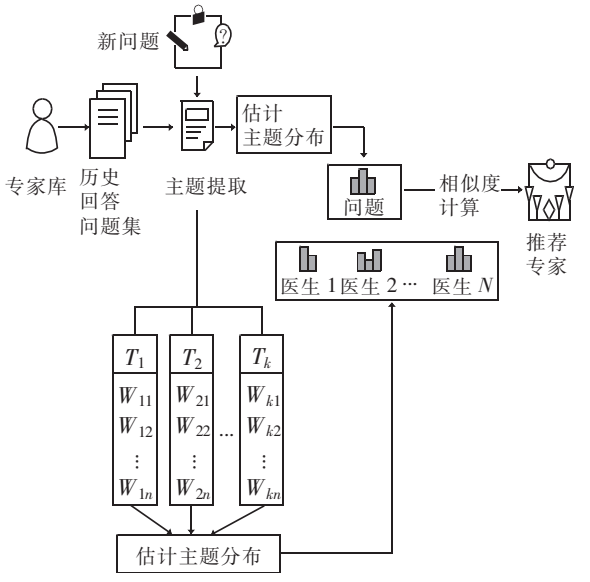

图 3 在线轻问诊医生推荐框架

Fig. 3 Online light consultation doctor recommendation framework

2.1　医生专长信息提取

医生专长信息提取主要思想是采集某科室中某医生历史回答问题集合进行建模，在此基础上进行监督学习，从而得到该医生回答问题的主题信息，对医生来说，该主题即是其在某科室的专长。为找到医生专长，本文用到的是 CTM 主题模型，其模型如图 4 所示。

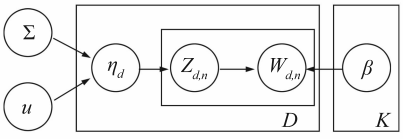

图 4　CTM 主题模型图表示

Fig. 4　CTM topic model diagram representation

图 4 中，K 表示某科室内医生以往回答健康问题的集合，D 表示某个问题的长度，矩形框表示进行迭代的次数，$W_{d,n}$ 表示第 d 个问题中的第 n 个词，问题库中所有词构成集合 V，$\boldsymbol{W}_d$ 表示问题 d 中所有 N_d 个词构成的 N_d 维向量，主题 β 是 V 上的分布。每个医生的过往回答健康问题集合都对应一个主题混合比例向量 θ_d，θ_d 是主题上的分布，既反映了问题库 d 中单词取主题集中每个主题的概率，也考虑了使用多项式分布 $\eta = \log(\theta_i/\theta_k)$ 进行自然参数化处理[12]。

2.2　待匹配健康问题主题提取

由于患者的医学涉猎较为有限，一个健康问题的医学专用术语并不明确，很难清晰地得到含蕴其间的医学主题。基于此，通过访问待匹配健康问题科室的问题集合文本，从中提取该科室涉及到的医学主题，可以得到访问科室健康问题的主题分布，即可推断待问答健康问题所含有的医学主题。因为健康问题是流动的，即使一句简单的问题也可能涉及到多个医学主题，为了获得健康问题主题分布，本文采用增量吉布斯采样（Incremental Gibbs Sample）对访问科室内健康问题集合进行参数估计，获取健康问题-主题的概率分布 θ 以及主题 - 词项的概率分布 β。

2.3　医生推荐

在线轻问诊医生推荐的目的是为患者提出的健康问题高效地匹配到专业的医生，当提取到科室医生的专长信息以及轻问诊健康问题的主题时，只需要计算相关的主题相似度，就能够为提出问题的患者找到最适宜的医生专家。本文采用的是杰卡德相似系数（Jaccard Similarity）计算方法，系数越大，表明医生专长与待回答轻问诊健康问题的内容就越相似。主要步骤为：

Step 1　从科室医生名单中获取某位医生的专长关键词记为 U。

Step 2　选取一个访问该科室的健康问题，记问题关键词集合为 V。采用杰卡德相似系数方法计算医生回答问题库与待回答轻问诊问题的相似度，即集合 U 和集合 V 的交集元素/并集元素。

Step 3　选取下一个访问该科室的健康问题，重复 Step 1 和 Step 2，直到所有访问该科室的健康问题遍历完毕。

Step 4　选取下一个医生，重复 Step 1 和 Step 2，直到所有医生遍历完毕。至此，得到了该科室医生与健康问题的主题相似度集合，根据集合中最大的前 n 个数给轻问诊问题匹配合适的 n 个医生。

3　实验结果与结果分析

3.1　数据收集与处理

考虑到数据的真实有效以及规模性，本文的数据来源为知名互联网医疗网站好大夫。皮肤科中的过敏反应科是比较常见并且涉及到的健康问题比较轻微的科室，寻求在线轻问诊解决健康问题的患者比较普遍。因此本文采用网络爬虫技术收集该网站截止到 2020 年 11 月 15 日的所有过敏反应科医生在线轻问诊的过往回答问题为研究样例，其中过敏反应科医生为 235 位，健康问题为最新产生的 30 万条轻问诊问题，问题中的 28 736 条被患者接受。

在好大夫网站采集到的原始数据存在着噪声，需在做处理后才能将其用于分析和主题挖掘。在使用 CTM 模型对健康问题集合进行建模前，通过利用中文分词、医学专业词识别、停用词过滤等方法对每个健康问题集合进行预处理，这样就降低了问题集的空间维度，从而提高了建模效率[13]。对于中文分词，因为健康问题集合数据庞大，本文采用的是统计分词的算法，基于统计学的机器学习模型对数据进行训练[14]。对于医学专业词识别，考虑到健康问题中涉及到例如药名、疾病名称等医学健康词汇，因此就要在用户词典中添加从互联网收集到的医学词库，旨在能够高效识别涉及到的医学方面用语[15]。对于停用词过滤，是因为分词后得到的问题集还是会存在大量的冗余，比如“在”、“的”等词汇，这些词汇对于文本语义分析以及主题的提取并无用处，而且还会降低建模效率。针对这个问题，本文使用哈工大停用词表来筛选语料中的高频通用词和低频词，以获得噪声较小的数据集，藉此来提高建模的效

率[16]。

3.2 模型构建

截止到 2020 年 11 月 15 日,好大夫在线过敏反应科的 235 名医生全都参与过最新的 30 万个问题。选取 25 万个健康问题作为训练集,其余的 5 万个健康问题作为测试集。好大夫在线从用户的健康提问和医生对问题的解答中自动识别出关键词来作为主题,这就完善了用户因为不了解医学专有名词而导致的健康状况不明确等问题。对过敏反应科以往回答过的过敏反应问题集的主题标签进行统计,合计获取了 13 026 个主题标签。使用停用词过滤后,选取出现频率最多的前 600 个主题作为模型训练的主题标签。

把这 600 个主题分布在 235 名过敏反应科医生的健康问题集合上,通过 CTM 模型训练,获取到每一位过敏反应科医生在各个主题上的概率分布,即获取医生专长,部分实验结果如图 5 所示。

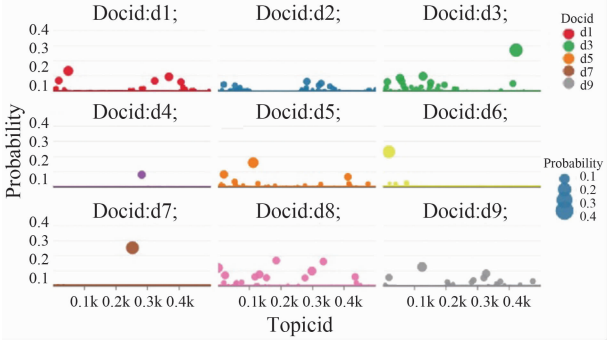

图 5 过敏反应科医生主题分布样例

Fig. 5 Sample distribution of allergy doctor topics

图 5 中的每个子图就是一个过敏反应科医生的主题分布,其中主题标识为横坐标,医生与主题的分布概率为纵坐标,每个点的大小反映了分布概率的大小。通过观测实验结果,可以发现不同的医生存在不同的专长分布,并且有些医生可以解答多个主题的健康问题,有些医生却仅会解答某个主题的健康问题,还存在一些医生对多个主题虽都有涉及,但却未能提取出特别擅长的主题。

3.3 模型测试

使用训练后的模型对 600 个主题测试集进行主题分布预测,其中主题标签为横坐标,测试问题集里面的健康问题为纵坐标。经过 CTM 主题模型训练得到每个健康问题在主题标签库上的概率分布情况,部分实验结果如图 6 所示。

图 6 中的每个子图反映的是测试集中的一个健康问题在主题上的概率分布情况。从分布情况来看,有些患者提出的健康问题主题特点明确,只涉及少数的主题,有些患者提出的健康问题涉及到多个主题并且概率都偏高,表明这些医学主题之间都将存在相关性,而本文采用的 CTM 模型能有效解决该问题。

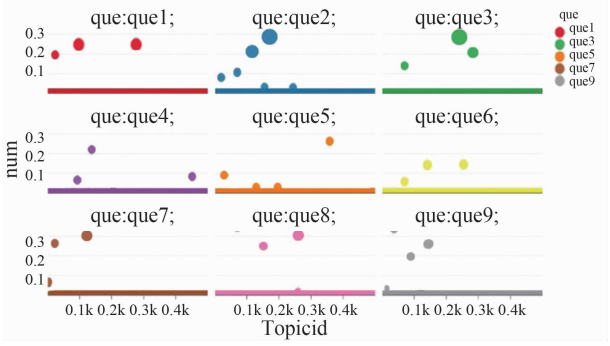

图 6 测试集健康问题主题分布

Fig. 6 Distribution of health issues in the test set

3.4 结果分析

对于本文提出的在线医生推荐模型的效果测评,先将测试集中的 5 000 个健康问题随机分成 5 组,即每 1 000 个为一组,使用本文方法产生医生推荐列表,其中限制的在线医生数量为 8,对 5 组问题集分别计算准确率、召回率和 *MRR*[17]。结果见表 1。

表 1 医生推荐测评结果

Tab. 1 Doctor recommendation test results

组别	准确率/%	召回率/%	*MRR*
组 1	45	42	0.315
组 2	40	39	0.321
组 3	42	43	0.289
组 4	36	35	0.258
组 5	48	47	0.314
组平均值	42	41	0.299
测试集总情况	41	45	0.302

由表 1 中数据可以看到,5 个分组的推荐情况都相对稳定,准确率和召回率都在 40% 左右,变化浮动小,并且两者相差较小。存在一些组的 *MRR* 值变化较大,容易被极端值所影响,经分析是由于 235 名过敏反应科医生参与网站回答医疗的时间跨度很大,有些医生注册网站时间久、回答的问题规模比较大,所以主题分布更高效清晰,还有一些医生新近加入网站,在线回答问题量偏少,仍无法完全提取得到其专长。由于新医生主题分布不明显,容易排在推荐医生列表的后面,如果某个患者采纳的是新加入医生的解答,那么就会对 *MRR* 值产生影响。

为了验证本文提出的在线医生推荐的有效性，对过敏反应科使用该方法与好大夫在线已存在的指标展开对比，对比结果见表 2。

表 2　医生推荐方法对比

Tab. 2　Comparison of doctor recommended methods

对比指标	好大夫现有指标	专家推荐方法
过敏反应科问题总数	300 000	5 000
过敏反应科问题采纳次数	96 000	2 050
过敏反应科医生回答总次数	806 722	7 795
所有医生回答总次数	2 304 920	17 322
准确率/%	32	41
召回率/%	35	45
回答采纳比/%	11.9	26.3

由表 2 中数据分析可知，准确率为过敏反应科问题采纳次数/过敏反应科问题总数，召回率为过敏反应科医生回答总次数/所有医生回答总次数，回答采纳比为过敏反应科问题采纳次数/过敏反应科医生回答总次数。结合好大夫网站现有指标对比发现，本文提出的专家推荐系统从准确率、召回率以及回答采纳比都优于好大夫在线过敏反应科的现有指标，充分证实了该系统对在线医生推荐的高效性。

4　结束语

目前的在线医生推荐研究中，现有的一些方法忽略了医生专长之间有关联以及描述的健康问题主题之间的关联性，导致获取的主题分布繁杂且无侧重。对于在线医生推荐，不仅要关注模型的主题词提取效果和分类准确性，同时还需要考虑模型能否兼顾主题之间的联系。在这种情景下，本文采用的 CTM 模型可以很好地解决这个问题：先用模型训练患者提出的健康问题，得到问题-主题概率分布，其次利用科室内的每个医生历史回答问题集合得到医生-主题概率分布，接着对得到的 2 项分布用杰卡德相似系数计算方法计算相似度，稍后将杰卡德相似系数大的、即主题相似度高的医生列表推荐给患者。最后，通过对好大夫在线过敏反应科的数据进行建模与测试，实验结果充分证明了本文提出的医生推荐方法比网站该科室现存推荐方法更高效。

对于本文提出的推荐模型也存在不足，例如有一些医生注册该网站时间不长，回答患者问题的积累量偏少，其专长无法得到完全提取，会导致该新医生即便很适合回答某个健康问题，但因为自身的主题分布不明显，而排在该问题推荐医生列表的后面将无法反馈给患者。后续亟需对这个问题进行特殊的处理，即对新加入医生的专长进行优化提取，以此来提高系统整体效率和用户满意度。另外，本文提出的方法默认患者是知道自己的健康问题属于哪个科室，在该科室有医生能帮助自己，所以针对一些对自身疾病存在盲区的患者，需要配合健康问题和医院科室选择的系统结合使用。

参考文献

[1] 林悦. “互联网+智慧医疗”现状及发展展望[J]. 中国医疗器械信息，2019，25(18)：15-16.

[2] 刁必颂. 基于在线患者咨询数据的在线医生推荐系统研究[D]. 北京：北京理工大学，2016.

[3] 朱利，岳爱珍. 健康问题和医生匹配机制的研究[J]. 西安交通大学学报，2014，48(12)：57-62，139.

[4] 杨晓夫，秦函书. 基于电子病历利用矩阵乘法构建医生推荐模型[J]. 计算机与现代化，2019(06)：81-86，97.

[5] 朱龙霞. 面向中文问答系统问题分析与答案抽取方法研究[D]. 石家庄：河北科技大学，2018.

[6] 张金柱，于文倩. 基于短语表示学习的主题识别及其表征词抽取方法研究[J/OL]. 数据分析与知识发现：1-13[2020-10-22]. https://kns.cnki.net/kcms/detail/10.1478.g2.20201022.1158.002.html.

[7] JURCZYK P，AGICHTEIN E. Discovering authorities in question answer communities by using link analysis[C]// Sixteenth ACM Conference on Information and Knowledge Management，CIKM 2007. Lisbon，Portugal，November. DBLP，2007：919-922.

[8] BOUGUESSA M，DUMOULIN B，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. Las Vegas，Nevada，USA：ACM，2008：866-874.

[9] BLEI D M，LAFFERTY J D. Correlated topic models[C]// Advances in Neural Information Processing Systems. Vancouver，British Columbia，Canada：dblp，2005，18：147-154.

[10] 史盛楠. CTM 主题模型在学科主题识别与学科文献分类中的应用研究[D]. 曲阜：曲阜师范大学，2019.

[11] 潘有能，倪秀丽. 基于 Labeled-LDA 模型的在线医疗专家推荐研究[J]. 数据分析与知识发现，2020，4(4)：34-43.

[12] 杨正良. 优化特征选择的 CTM 模型在文本分类中的应用研究[D]. 武汉：华中师范大学，2016.

[13] 丁勇，程家桥，蒋翠清，等. 基于主题和关键词特征的比较文本分类方法[J/OL]. 计算机工程与应用：1-9[2020-11-02]. http:///KCMS/detail/11.2127.tp.20201026.0911.002.html.

[14] 李国垒，陈先来，夏冬，等. 中文病历文本分词方法研究[J]. 中国生物医学工程学报，2016，35(4)：477-481.

[15] 王月瑶. 面向医疗文本检索的查询重构技术研究与实现[D]. 上海：华东师范大学，2018.

[16] 王凡，夏晨曦. 中文医学摘要主题建模方法评估[J]. 医学信息学杂志，2018，39(2)：60-64.

[17] 单国栋，肖彦翠，王皓. 基于主题模型的中外期刊文献挖掘对比研究[J]. 长春大学学报(自然科学版)，2019，29(3)：23-29.