基于 HSM_LDA 模型的在线医院特色挖掘研究*

黄锦泉[1] 张 楚[2] 刘灵涛[1] 潘 玮[1] 翟菊叶[1] 刘玉文[1]

([1] 蚌埠医学院卫生管理学院 蚌埠 233030 [2] 蚌埠医学院护理学院 蚌埠 233030)

〔摘要〕 **目的/意义** 挖掘在线医院的医疗特色对在线医疗推荐具有重要作用。当前，虽然部分在线医院具备特色标注功能，但只能实现医院内部特色提示，无法从全局角度衡量不同医院之间的特色差异。**方法/过程** 提出一种基于在线医院问诊文本的医院特色识别模型（hospital special medical based LDA，HSM_LDA）。该模型以医院 ID 为文本划分依据，将语料库中的“文本－词汇”矩阵转换成“医院－词汇”矩阵，联合建模医院、主题、词汇 3 个变量，生成“医院－主题”（E）和“主题－词汇”（F）两个分布。最终结合 E 和 F 两个分布识别出每个医院的医疗特色。**结果/结论** 以“好大夫在线”平台中的医院问诊文本作为实验数据集，运用 HSM_LDA 模型进行特色挖掘分析，识别精度为 87%，效果良好。

〔关键词〕 在线医院；医院特色；HSM_LDA 模型；主题识别；医疗推荐

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2023.09.005

Study on Online Hospital Feature Mining Based on HSM_LDA Model

HUANG Jinquan[1], *ZHANG Chu*[2], *LIU Lingtao*[1], *PAN Wei*[1], *ZHAI Juye*[1], *LIU Yuwen*[1]

[1]*School of Health Management, Bengbu Medical College, Bengbu 233030, China*; [2]*School of Nursing, Bengbu Medical College, Bengbu 233030, China*

〔**Abstract**〕 **Purpose/Significance** Mining the medical characteristics of online hospitals plays a very important role in online medical accurate recommendation. At present, although some online hospitals have the feature label function, they can only realize the internal feature prompt, and can not measure the feature differences between different hospitals from a global perspective. **Method/Process** The paper proposes a hospital special medical based LDA (HSM_LDA) model based on online hospital inquiry text. The method takes hospital ID as the modeling entry, converts the “text－vocabulary” matrix in the inquiry corpus into the “hospital－vocabulary” matrix, jointly models the three variables of hospital, topic and vocabulary, and generates two distributions, “hospital－topic” (E) and “topic－vocabulary” (F). Finally, the medical characteristics of each hospital are identified by combining E and F distribution. **Result/Conclusion** The hospital consultation text provided by the “haodf.com” platform is used as the experimental data set, and the HSM_LDA model is used for mining and analysis. The experimental results show that the hospital characteristic recognition accuracy of the proposed method is 87%, which has achieved a good effect.

〔**Keywords**〕 online hospital; hospital characteristics; HSM_LDA model; topic identification; medical recommendation

〔修回日期〕 2023－03－28

〔作者简介〕 黄锦泉，硕士研究生，发表论文 3 篇；通信作者：刘玉文，教授，硕士生导师。

〔基金项目〕 安徽省哲学社会科学规划项目（项目编号：AHSKQ2019D070）；安徽省人文社科高校重点项目（项目编号：SK2021A0444）；蚌埠医学院研究生科研创新项目（项目编号：Byycxz22024）。

1 引言

随着我国“互联网 + 医疗健康”事业的迅速发展，以“好大夫在线”“春雨医生”等为代表的在线健康社区（online health communities，OHCs）逐步涌现[1]，为在线医院的兴起提供了平台基础。截至目前，众多国内医院已在健康社区内注册账号[2]成为在线医院。与传统线下就医模式相比，在线医院打破时空局限，实现了患者与医生的跨时空交互，对提高医疗资源利用率[3]、促进医疗均衡发展具有推动作用。但 OHCs 尚缺乏全局性的在线医院特色导航服务，用户在线问诊时无法根据自身病情选择合适的医院[4]，这在一定程度上限制了在线医院服务质量的提升。所以，从全局角度挖掘在线医院的医疗特色，实现医疗特色精准导航，对提升在线医院服务质量、改善用户问诊体验具有重要意义。

当前，在线医疗特色识别相关研究主要围绕医生和医院两方面展开。其中，医生特色识别相关研究较多，主要是利用机器学习、自然语言处理等方式探索 OHCs 中医生的专业领域，为患者提供高效便利的医生推荐服务。例如，孟秋晴等[5]利用文本相似度和隐含狄利克雷分布（latent Dirichlet allocation，LDA）主题模型对患者问诊文本和医生回答文本进行挖掘，试图分析在线医生的诊疗特色。梁建树等[6]利用 Word2Vec 和 LDA 等技术对 OHCs 中的医生特征进行挖掘，并结合三支决策思想提出多维度的三支医生推荐方法。该方法深入挖掘医生特色，大幅度提高医生推荐精准度。Li Y Y 等[7]提出一种组合条件的目标医生挖掘模型，该模型分为相似患者、相似领域和医生绩效 3 部分，最后采用线性加权整合 3 部分结果，挖掘符合患者需求的目标医生。武家伟等[8]以 OHCs 中用户评论文本作为数据源，融合知识图谱和深度学习技术挖掘医生服务特色。叶佳鑫等[9]利用 Word2Vec 模型对 OHCs 中医生相关文本进行挖掘，从而找寻与目标医生相似的医生人群，进而对目标医生进行标注，丰富医生特征。在医院特色识别方面，诸多学者开始挖掘目标医院的特色科室，帮助患者解决挂错号等问题。例如，宁建飞等[10]使用词向量和句子相似度方法分析患者在线问诊文本的语言特征，并进一步以词向量代替词频比对问诊文本和问答知识库的相似度，从而挖掘目标医院特色科室。郑姝雅[11]提出一种基于线性支持向量机的医院科室匹配方法，利用科室内的接诊记录推算符合目标患者需求的特色科室。何慧茹[12]利用统计学原理对医疗资源进行收集与分析，通过径向基函数（radical basis function，RBF）神经网络模型和模糊算法模型推导医院中不同科室具备的特色。以上研究使用不同方式对在线医院特色进行挖掘，虽然有助于改善 OHCs 的患者体验，挖掘用户需求，但无法从全局角度挖掘不同医院之间的特色差异，且患者与医院匹配不精准问题仍未得到较好解决。

因此，本研究将医院 ID 融入传统 LDA 模型中，构建医院特色识别模型（hospital special medical based LDA，HSM_LDA）。该模型将原始的“文本 - 词汇”矩阵转化为“医院 - 词汇”矩阵，联合医院、主题、词汇 3 个变量进行建模，生成“医院 - 主题”（E）和“主题 - 词汇”（F）两个分布矩阵，从而识别出医院特色。

2 相关技术介绍

2.1 词频 - 逆文本频率指数算法

词频 - 逆文本频率指数[13]（term frequency - inverse document frequency，TF - IDF）是文本数据挖掘的重要方法，主要用于度量文本中词语的重要程度。一般情况下，词语的重要程度不仅与该词在文本中出现的次数有关，还与包含该词语的文本数量有关。如果某个词语在文本中出现的次数越高，且包含它的其他文档数量越少，则该词的重要程度就越高。

$$TF-IDF(w_i) = TF(w_i) \times IDF(w_i) \quad (1)$$

其中，$TF(w_i)$ 表示词语 w_i 在文档 d_i 中出现的频率，$IDF(w_i)$ 表示词语 w_i 的逆向文档频率。

2.2 LDA 模型

LDA 模型[14]是一种无监督学习的文档生成模

型，于2003 年被提出，可以计算文档集中每篇文档的主题概率分布和每个词语的概率分布，主要用于文档主题的聚类和分类。LDA 建模过程可以分为 4 步：一是选择一篇文档，以 α 为超参数进行 Dirichlet 分布采样生成“文档 - 主题”概率 θ；二是由 θ 分布生成所有文档中词语的主题 Z；三是以 β 为超参数进行 Dirichlet 分布采样生成“主题 - 词汇”概率 φ；四是由 φ 分布生成词语 W。

3 基于 HSM_LDA 模型的医院特色识别方法

基于 HSM_LDA 模型的医院特色挖掘过程主要包括 3 个步骤：下载在线医院问诊数据，并对问诊文本进行分词、去停用词等，生成问诊文本语料库；将预处理后的文本进行 TF - IDF 运算，计算文本中词汇的重要程度；建立 HSM_LDA 模型并对问诊语料库进行建模，生成“医院 - 主题”（E）和“主题 - 词汇”（F）两个分布矩阵；根据分布 F 人工标注特色主题含义，再根据分布 E 获取特色主题在医院的分布，见图 1。

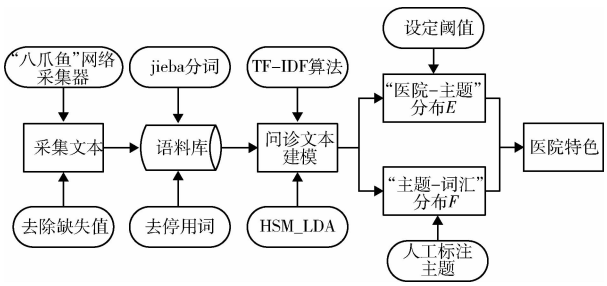

图 1 研究总体框架

3.1 HSM_LDA 模型建立

设在线医院的问诊文本语料库为 $D = [d_1, d_2, \cdots, d_m]^T$，其中，$d_i = <H\_ID, Text>$ 是个二元组，H_ID 表示医院编号，$Text$ 表示问诊文本。K 表示 D 中的主题数，W 表示 D 中所有词汇组成的集合。根据 *HSN_LDA* 模型的生成关系作如下定义。

3.1.1 定义 1：“医院 - 主题”分布 E 对任意医院 H_i 的问诊文本，生成主题的概率分布为 $E_{Hi} = <p_1, p_2, \cdots, p_k>$，$p_z = n_z/n$，其中，$n_z$ 表示医院 H_i 问诊文本中分配给主题 z 的词汇数量，n 表示医院 H_i 问诊文本的词汇总数，则主题 z 在医院 H_i 中生成概率如下：

$$P(z) = \sum_{n=1}^{n} P(z \mid H_i = n) P(H_i = n) \tag{2}$$

3.1.2 定义 2：“主题 - 词汇”分布 F 对任意主题 z_i，生成词汇的概率分布可表示为 $F_{zi} = <p_1, p_2, \cdots, p_n>$，$p_i = n_i/n$，其中，$n_i$ 表示词语 w_i 在主题 z_i 中的频数，n 表示属于主题 z_i 的词汇总数，k 表示主题的数量。词汇 w 在主题 z_i 中的生成概率如下：

$$P(w) = \sum_{k=1}^{k} P(w \mid z_i = k) P(z_i = k) \tag{3}$$

与 LDA 模型相比，HSM_LDA 通过医院 ID 参数在迭代采样时，将属于同一医院 ID 的文本进行连接视为一条文本，从而将传统 LDA 模型生成的“文本 - 主题”分布转化为“医院 - 主题”分布，见图 2。

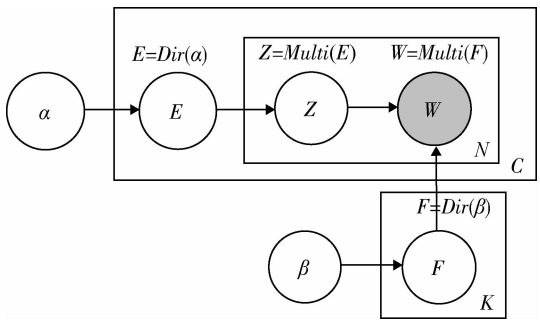

图 2 HSM_LDA 模型结构

HSM_LDA 与 LDA 的不同之处表现在以下两方面：框架的外层表示医院层，C 表示医院数量；E 表示“医院 - 主题”分布，F 表示“主题 - 词汇”分布。

3.2 模型公式推导

HSM_LDA 模型运用超参数 α 生成一个“医院 - 主题”概率分布，再运用 β 生成 N 个“主题 - 词汇”概率分布，最后再生成问诊文本的 N 个词的联合概率公式：

$$P(E, Z, W \mid \alpha, \beta) = P(E \mid \alpha) \prod_{n=1}^{N} P(Z_n \mid E) P(W_n \mid Z_n, \beta) \tag{4}$$

由于只有 W 是唯一可观测值，如果要计算 W

的生成概率就需要对 E 和 Z 进行边缘概率求解，从而消除 E 和 Z。最终的词汇生成概率计算方式如下：

$$P(W|\alpha,\beta) = \int P(E|\alpha)(\prod_{n=1}^{N}\sum_{Z_n} P(Z_n|E)P(W_n|Z_n,\beta))dE \quad (5)$$

得到词汇生成概率后，可以通过采样算法对模型中的 E 和 F 参数进行估计。常用估计方法是吉布斯采样，通过期望最大化（expectation - maximum，EM）算法对 E 和 F 进行反复迭代，使其逐步收敛。基于 HSM_LDA 模型的医院特色识别算法描述如下：

Input：α，β
Output：E，F
（1）Get $\{D, V\}$ //读入文本语料库
（2）For$k$ = 1 to K
（3） //计算问诊文本主题
（4）Compute α，β
（5） //计算模型超参数
（6）Run Gibbs（α，β）
（7）//进行 Gibbs 采样
（8） For each wd
（9） Choose a w from E w ~ Multi（α）
（10） Choose a w from F w ~ Multi（β）
（11）End For
（12）Get E，F
（13） End For

3.3 最优主题数计算

在运用 HSM_LDA 模型进行主题挖掘时，主题数 K 是影响主题挖掘效果的关键因素，本研究采用主题困惑度曲线估计 K 值。困惑度是主题不确定性的一种表达方式，困惑度越低表明主题聚类效果越好，其计算方式如下：

$$P = e^{\frac{-\sum\log(P(w))}{N}} \quad (6)$$

其中，N 表示语料库中的词语总数，p（w）表示词语 w 出现的概率。利用困惑度 P 和主题数 K 建立主题困惑度曲线，当 P 值最小时 K 最优。但困惑度只是判定最优主题数的一个粗略指标。所以，本研究在实验过程中以困惑度最低点作为参考值，在最低点两边取值进行多次实验，选择效果最好的 K 值作为 HSM_LDA 模型的主题数。

4 实验分析

4.1 数据来源及预处理

“好大夫在线”是我国常用的网络医疗资源平台，集合了 1 万多家在线医院，注册医生近 60 万人[15]。众多访问用户在平台内积累了大量问诊数据。以“好大夫在线”为数据源，运用“八爪鱼”网络数据采集器获取 2022 年 4—5 月患者问诊数据 148 376 条，每条数据包括医院 ID、医院名称、患者性别、患者年龄、问诊时间、问诊文本和科室等信息。然后，按照问诊数量由高到低对医院排序，并选择前 100 家医院的问诊记录作为实验数据集。运用 jieba 分词工具[16]对问诊文本进行分词，并去除停用词、介词以及无用词，建立医院问诊文本矩阵 D。

4.2 实验结果及分析

4.2.1 医院医疗特色识别 HSM_LDA 模型需要设置 4 个参数：超参数 α、β，主题数 K 以及迭代采样次数。通常情况下：α 设置为 $0.5/K$，β 设置为 0.1，迭代次数设置为 1 000。当 K 设置为 13 时，主题困惑度最低。因此以 13 为主题数参考点，在 13 ±5 范围内进行多次实验，最终结果显示当主题数设置为 15 时主题识别效果最佳。运行 HSM_LDA 模型得到“主题 - 词汇”（F）和“医院 - 主题”（E）两个分布矩阵。在 F 分布中，主题生成词汇概率越大，词汇的主题属性越强。按照生成概率的大小选择生成概率前 15 位的词语作为主题关键词，然后，根据词汇表达出的语义，对特色主题含义进行人工标注，见表 1。

表 1 医院特色主题识别结果（主题前 10 位）

主题	主题词及概率（前 15 位）	人工标注
主题 1	早搏 0.003 7/心衰 0.003 7/心脏病 0.003 7/胆固醇 0.003 7/脂蛋白 0.003 7/房颤 0.003 7/心梗 0.003 7/颈动脉 0.001 1/高血脂 0.003 7/血压高 0.003 7/心肌 0.001 1/心动过速 0.003 7/肺动脉 0.003 7/心血管 0.003 7/瓣膜 0.003 7	心血管系统疾病
主题 2	眼压 0.001 6/眼皮 0.005 6/视网膜 0.001 6/眼角 0.001 6/右耳 0.005 6/口腔 0.005 6/眼科 0.001 6/嘴唇 0.001 6/脸颊 0.005 6/肿痛 0.001 6/患处 0.001 6/眉毛 0.001 6/鼻腔 0.001 6/牙齿 0.005 6/全麻 0.005 6	眼科/口腔科疾病
主题 3	鼻涕 0.002 3/鼻塞 0.002 3/鼻窦炎 0.002 3/耳聋 0.002 3/鼻中隔 0.002 3/喉镜 0.002 3/鼻咽 0.002 3/中耳炎 0.002 3/食管炎 0.002 3/通气 0.002 3/扁桃体 0.002 3/腺样体 0.002 3/流鼻涕 0.002 3/交流 0.002 3/障碍 0.002 3	耳鼻喉科疾病
主题 4	蛋白尿 0.002 2/肾炎 0.002 2/潜血 0.002 2/支原体 0.002 2/抗核 0.002 2/霉素 0.002 2/右肾 0.002 2/肾内科 0.002 2/肾脏 0.002 2/空腹 0.002 2/血症 0.002 2/结核 0.002 2/高尿酸 0.002 2/尿痛 0.002 2/紫癜 0.002 2	肾功能系统疾病
主题 5	淋巴瘤 0.005 3/胃窦 0.005 3/胃胀 0.005 3/胃痛 0.005 3/结核 0.005 3/抗病毒 0.005 3/血症 0.005 3/脓肿 0.005 3/贲门 0.005 3/胰腺 0.005 3/肌瘤 0.005 3/右叶 0.005 3/胸片 0.005 3/个人 0.005 3/右乳 0.005 3	消化系统疾病
主题 6	精子 0.004 4/精索 0.004 4/睾丸 0.004 4/肌瘤 0.004 4/精液 0.004 4/射精 0.004 4/阴茎 0.004 4/性生活 0.004 4/红斑狼疮 0.004 4/男性 0.004 4/性激 0.004 4/阴囊 0.004 4/睾酮 0.004 4/包皮 0.004 4/风湿 0.004 4	男性生殖系统疾病
主题 7	韧带 0.005 0/髋关节 0.005 0/半月板 0.005 0/股骨头 0.005 0/腰椎间盘 0.005 0/肌腱 0.005 0/脊椎 0.005 0/踝关节 0.005 0/骨科 0.005 0/软骨 0.005 0/股骨 0.005 0/脚趾 0.005 0/肩膀 0.005 0/石膏 0.005 0/颈椎病 0.000 1	骨科疾病
主题 8	气喘 0.007 4 /雾化 0.007 4/支气管炎 0.007 4/牙龈 0.007 4/肺结核 0.007 4/哮喘 0.007 4/支原体 0.007 4/胸片 0.007 4/口腔 0.007 4/紫癜 0.007 4/点状 0.007 4/气管 0.007 4/呼吸 0.007 4/根部 0.007 4 /化痰 0.007 4	呼吸系统疾病
主题 9	免疫治疗 0.003 5/右肾 0.003 5/肌瘤 0.003 5/肉瘤 0.003 5/骨质 0.003 5/肝癌 0.003 5/甲胎蛋白 0.003 5/放疗 0.003 5/化疗 0.003 5/直径 0.003 5/右叶 0.003 5/根治术 0.003 5/直肠癌 0.003 5 /肾癌 0.003 5/公分 0.003 5	肿瘤科疾病
主题 10	骨髓 0.003 7/白血病 0.003 7/淋巴细胞 0.003 7/淋巴瘤 0.003 7 /肾盂 0.003 7/重症 0.003 7/角度 0.003 7/生长 0.003 7/实质 0.003 7/贫血 0.003 7/慢性 0.003 7/感染 0.003 7/免疫 0.003 7/功能 0.003 7/细胞 0.003 7	血液科疾病

从 F 分布中只能识别出特色主题的含义，不能确定医院特色主题。因此，还需要进一步结合 E 分布来确定医院的特色主题。问诊量排名前 10 位的医院主题识别结果，见表 2。

表 2 “医院－主题”识别结果（H_ID 前 10 位）

ID	主题 1	主题 2	主题 3	主题 4	主题 5	主题 6	主题 7	主题 8	主题 9	主题 10	主题 11	主题 12	主题 13	主题 14	主题 15
H_ID_1	**0.120**	0.034	**0.109**	0.098	0.011	**0.138**	0.060	0.011	0.062	0.077	**0.143**	0.051	0.029	0.050	0.002
H_ID_2	0.024	0.001	**0.130**	**0.101**	0.006	**0.165**	0.020	0.033	0.001	0.044	**0.121**	0.054	0.063	**0.238**	0.001
H_ID_3	0.001	0.041	0.087	**0.143**	0.003	**0.197**	0.024	0.001	0.080	0.061	0.031	0.015	0.081	**0.227**	0.007
H_ID_4	**0.167**	0.022	**0.145**	**0.123**	0.044	0.035	**0.117**	0.035	0.042	0.026	**0.146**	0.010	0.039	0.017	0.031
H_ID_5	**0.182**	0.049	0.086	**0.113**	0.034	0.028	0.075	0.031	0.096	0.072	0.053	0.075	0.064	0.016	0.022
H_ID_6	0.050	0.006	**0.105**	**0.167**	0.023	0.029	0.077	0.008	**0.131**	**0.157**	0.024	0.035	0.065	0.066	0.061
H_ID_7	0.060	0.009	0.090	**0.115**	0.032	0.067	0.059	0.018	**0.145**	**0.150**	0.032	0.032	0.068	0.037	0.084
H_ID_8	0.045	0.001	**0.132**	**0.126**	0.019	0.019	0.012	**0.123**	0.036	**0.304**	0.019	0.011	**0.112**	0.026	0.016
H_ID_9	0.029	0.009	0.098	**0.100**	0.030	0.010	**0.278**	0.017	0.019	0.022	0.029	**0.112**	**0.244**	0.001	0.001
H_ID_{10}	0.016	0.014	**0.163**	**0.135**	0.084	0.049	0.010	0.001	0.057	0.029	0.044	0.008	0.040	0.046	**0.306**

根据医院生成主题的概率结果，在保证医院特色主题有较高鲜明度的情况下，特色主题不至于太多。设置医院特色主题概率阈值为 0.1，高于阈值的主题定义为医院特色主题。例如，编号为 H_ID_1 的医院特色主题包括主题 1、主题 3、主题 6 和主题 11。

结合表 1 和表 2 可以得出医院诊疗特色。例如，编号为 H_ID_1 的医院诊疗特色有：心血管系统疾病、耳鼻喉科疾病、男性生殖系统疾病与肛肠科疾病。依此类推，获得每家医院的医疗特色。

4.2.2 医院医疗特色对比 由于多家医院中会存在相同医疗特色，不利于患者选择就诊医院，所以对相同医疗特色下的医院进行排名。以医院主题概率值表示医院特色强度，对比同一特色下的多家医院，见图 3。该排名有助于患者在多家医院特色相同的情况下优先选择特色强度最高的医院进行就诊。

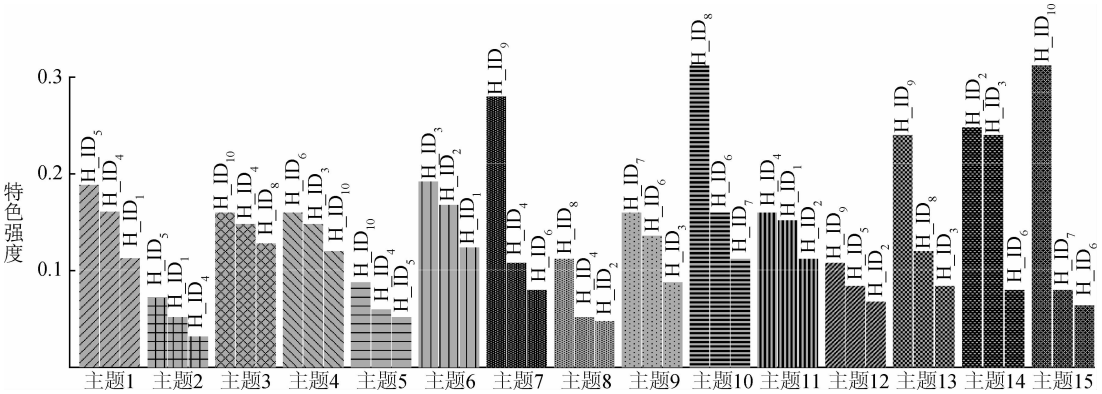

图 3 多家医院特色强度对比

以安徽医科大学第一附属医院为例，统计并分析其各科室问诊量条数，见图 4。

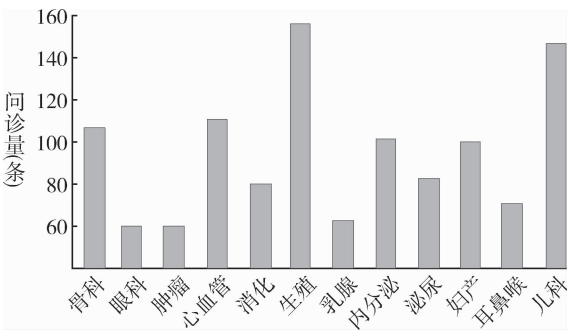

图 4 不同科室问诊量大小

再利用本文提出的 HSM_LDA 模型识别该医院特色包含男性生殖系统疾病、心血管系统疾病和儿科疾病。与前文图 4 中问诊量前 3 位的科室相符，说明该模型识别出的医院特色具有一定准确性。

4.2.3 模型评价 为验证 HSM_LDA 模型的有效性，以医院特色官方介绍作为评价标准，用准确率作为评价指标。准确率表示模型识别结果中符合官方特色数量（*DS*）除以模型识别出的特色总量（*HS*）：

$$Accuracy = \frac{DS}{HS} \tag{7}$$

结果发现，本文提出的 HSM_LDA 模型识别准确率达 87%（100 家医院识别准确率均值），见表 3。

表 3 官方特色与模型识别结果对比（H_ID 前 5 位）

医院 ID	官方特色	模型识别结果	准确率（%）
H_ID_1	肿瘤科、肾内科、内分泌科、泌尿外科、耳鼻喉科、心内科、消化内科、皮肤科、妇科、老年医学科	心内科、耳鼻喉科、泌尿外科、肛肠科	75
H_ID_2	耳鼻咽喉科、老年医学科、肾内科、呼吸内科、生殖科、神经外科、骨科、烧伤外科、普外科	耳鼻喉科、肾内科、泌尿外科、肛肠科、神经外科	80
H_ID_3	心内科、消化内科、肾内科、神经外科、肿瘤科、感染科、皮肤科、泌尿外科、儿科、骨科	心内科、肾内科、泌尿外科、神经外科	100
H_ID_4	儿科、泌尿外科、心血管内科、骨科、血液内科、眼科、耳鼻喉科、神经内科、消化内科	心内科、耳鼻喉科、肾内科、骨科、肛肠科	60
H_ID_5	心内科、消化内科、肾内科、神经外科、肿瘤科、感染科、皮肤科、泌尿外科、儿科、骨科	心内科、肾内科、肿瘤科	100

5 结语

本文在梳理 OHCs 相关研究时发现其无法从全局角度衡量不同医院之间的特色差异。为弥补这一缺陷，提出一种基于在线医院问诊文本的医院特色挖掘模型（HSM_LDA）。该模型在传统 LDA 模型 3 层结构的基础上，用医院层代替文本层，建立医院、主题、词汇之间的依赖关系，通过吉布斯多次采样生成“医院－主题”和“主题－词汇”两个分布矩阵，利用人工标注对主题词汇进行识别，从而挖掘医院特色。实验证明，HSM_LDA 模型在医院特色识别中能达到较好效果。

本文提出的 HSM_LDA 模型易于挖掘 OHCs 中的医院特色，有助于满足患者选择最佳就诊医院的需求，对推动 OHCs 发展具有一定积极意义。在后续研究中，可加入医院官网公布的问诊记录，以增强医院特色的鲜明程度；进一步细化特色主题含义，提高特色判定的准确性。目前模型的评价指标较少，后续研究会加入多种定量指标，以更好地展

示模型性能以及更全面、细致的医院医疗特色。

参考文献

1 DONG Q X, ZHOU X, MAO F H, et al. An investigation on the users' continuance intention in online health community —based on perceived value theory [J]. Journal of the China society for scientific and technical information, 2019, 39 (3): 3 - 14, 156.

2 曹卓琳, 蔡静, 刘人境. 基于 SWOT 分析的我国在线健康社区发展研究 [J]. 卫生软科学, 2022, 36 (1): 32 - 35.

3 肖洁, 高晶磊, 赵锐, 等. 我国城市医疗联合体实施现状及综合评价 [J]. 中国医院管理, 2021, 41 (2): 9 - 13.

4 ZHOU H, LIU J, ZHANG P Y, et al. A study on the usefulness of comments in online health community based on complex network perspective [J]. Information science, 2022, 40 (9): 88 - 97.

5 孟秋晴, 熊回香. 基于在线问诊文本信息的医生推荐研究 [J]. 情报科学, 2021, 39 (6): 152 - 160.

6 梁建树, 叶晓庆, 刘盾. 面向在线问诊平台的三支推荐方法 [J]. 西北大学学报 (自然科学版), 2022, 52 (5): 784 - 796.

7 LI Y Y, XIONG H X, LI X M. Recommending doctors online based on combined conditions [J]. Data analysis and knowledge discovery, 2020, 4 (8): 130 - 141.

8 武家伟, 孙艳春. 融合知识图谱和深度学习方法的问诊推荐系统 [J]. 计算机科学与探索, 2021, 15 (8): 1432 - 1440.

9 叶佳鑫, 熊回香, 童兆莉, 等. 在线医疗社区中面向医生的协同标注研究 [J]. 数据分析与知识发现, 2020, 4 (6): 118 - 128.

10 宁建飞, 黄发良. 基于词向量句子相似度量的医疗科室推荐 [J]. 福建师范大学学报 (自然科学版), 2018, 34 (4): 10 - 15.

11 郑姝雅. 面向在线问诊平台的精准导医模型构建研究 [D]. 南京: 南京大学, 2020.

12 何慧茹. 基于推理算法的导医系统设计与实现 [D]. 合肥: 安徽大学, 2016.

13 叶雪梅, 毛雪岷, 夏锦春, 等. 文本分类 TF - IDF 算法的改进研究 [J]. 计算机工程与应用, 2019, 55 (2): 104 - 109, 161.

14 BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of machine learning research, 2003, 3 (4 - 5): 993 - 1022.

15 XU X T, YANG M Q, SONG X K. Exploring the impact of physicians' word of mouth on patients' selection in online health community—taking the website of www. Haodf. com as an example [J]. Journal of modern information, 2019, 39 (8): 20 - 28, 36.

16 王杨, 许闪闪, 李昌, 等. 基于支持向量机的中文极短文本分类模型 [J]. 计算机应用研究, 2020, 37 (2): 347 - 350.

(上接第 36 页)

25 EBERT D D, CUIJPERS P, MUÑOZ R F, et al. Prevention of mental health disorders using internet - and mobile - based interventions: a narrative review and recommendations for future research [J]. Frontiers in psychiatry, 2017 (8): 116.

26 DOMHARDT M, GEßLEIN H, VON REZORI R E, et al. Internet - and mobile - based interventions for anxiety disorders: a meta - analytic review of intervention components [J]. Depression and anxiety, 2019, 36 (3): 213 - 224.

27 ORJI R, MOFFATT K. Persuasive technology for health and wellness: state - of - the - art and emerging trends [J]. Health informatics journal, 2018, 24 (1): 66 - 91.

28 FERNANDEZ - LLATAS C, GARCÍA - GÓMEZ J, VICENTE J, et al. Behaviour patterns detection for persuasive design in nursing homes to help dementia patients [C]. Boston: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2011.

29 INTILLE S S. A new research challenge: persuasive technology to motivate healthy aging [J]. IEEE transactions on information technology in biomedicine, 2004, 8 (3): 235 - 237.

30 URDA J L, LYNN J S, GORMAN A, et al. Effects of a minimal workplace intervention to reduce sedentary behaviors and improve perceived wellness in middle - aged women office workers [J]. Journal of physical activity and health, 2016, 13 (8): 838 - 844.

31 WUNSCH M, STIBE A, MILLONIG A, et al. What makes you bike? Exploring persuasive strategies to encourage low - energy mobility [C]. Chicago: International Conference on Persuasive Technology, 2015.

32 BAUMEISTER R F. The self [M]. Oxford: Oxford University Press, 2010.

33 MALONE T W, LEPPER M R. Aptitude, learning, and instruction [M]. London: Routledge, 2021.