·业务研究·

一种融合标签和患者咨询文本的医生推荐算法

周　鑫,熊回香,肖　兵

(华中师范大学 信息管理学院,湖北 武汉 430079)

摘　要:【目的/意义】针对在线医疗信息结构松散,医疗平台医生推荐精度不足的现状,设计了一种基于标签和患者咨询文本的医生推荐算法,提升医生推荐效果。【方法/过程】利用Word2vec模型训练患者咨询文本得到特征向量,改进余弦相似度算法计算医生推荐集A;利用LDA模型训练医生标签得到医生在主题上投影的概率分布,改进KL距离算法计算医生推荐集B;基于社会网络分析理论设计相关算法重构医生网络链接,选择中心性指标得到最终医生推荐集C。【结果/结论】以"丁香医生"数据进行实证,面向UGC数据丰富了算法的可用程度,弥补了单一推荐方法的不足,提高了推荐的精度。本文所提方法有效提升了医生推荐精度。【创新/局限】通过融合标签和患者咨询文本,采用社会网络分析实现了医生混合推荐。虽然通过中心性指标进行重要医生挖掘,但挖掘效果有提升空间。

关键词:医生推荐;标签;Word2vec;LDA;社会网络分析

中图分类号:G254　　**DOI:**10.13833/j.issn.1007-7634.2023.03.017

1　引　言

近年来在政策的扶持下,为了满足健康医疗信息用户交流、分享、合作、协调一致行动的本能需求,互联网在线医疗社区快速兴起,代表社区如"丁香园""微博医疗""好医生"等。医生和患者在平台中浏览诊疗信息、参与病情讨论、寻求潜在的合作和诊疗机会[1],生成了海量却分散的健康医疗信息。这些指数式增长的信息呈现出大规模、异质多元、组织结构松散等特点,这给患者有效的信息获取和利用带来了诸多困难,因此需要对患者进行精准的医生推荐。当前大部分在线医疗社区(典型为丁香园、好医生)医生推荐的主要依据是医生"星级""在线情况""问诊价格"等网站指标,功能较为粗泛,推荐效果欠佳;进一步地,目前学界在线医生推荐研究方法上多基于传统的内容推荐[2]、协同过滤推荐[3],内容上多基于传统的医疗文本如病例。因此,利用混合推荐方法弥补单一推荐手段的不足[4],对用户生成的更具分析价值的标签、患者咨询文本等信息进行挖掘,无疑可以提高推荐的质量。

本文提出了一种融合标签和患者咨询文本的医生推荐算法,选取患者咨询文本、医生擅长领域标签等数据进行推荐算法研究。首先,构建患者病症集合与医生集合,通过Word2vec模型对预处理后的患者咨询文本进行词向量训练以表征病症特征和医生特征,进而改进相似度算法计算病症与医生之间的相似度得到医生推荐集A;其次以医生推荐集A作为输入数据,利用LDA主题模型对医生标签进行医疗主题训练,将各医生的标签关系映射到潜在的医疗主题上,根据改进的KL算法计算概率之间的分布距离以表示医生之间的相似度进而得到医生推荐集B;最后结合医生推荐集A和B的推荐结果构建医生社会网络,利用社会网络分析方法设计网络链接值并计算中心性指标,得到最终的医生推荐集C。

2　研究现状

国内外一些学者对医生推荐方法进行了相关研究,主要集中在近五年的医学和计算机科学领域,多为探讨推荐算法与模型的设计与实现。Huang Y等人提出了一种基于医生绩效模型和患者偏好模型的医生推荐算法,旨在解决医生信息过载和预约不平衡的问题[5];Makowski等人利用层次分析法提出了一种基于医患偏好模型的医生决策算法,辅助临床并提供治疗建议[6];杨晓夫等人为解决传统推荐方法在医疗领域推荐质量不佳的问题,提出了一种基于矩阵乘法构建的医生推荐模型[7];陈亚明对病历信息和处方信息进行挖掘,提出了一种基于DBSCAN聚类和KNN算法的疾病预测方法,进行医生处方推荐[8]。而在情报学领域中,医生推荐是个较新的课题,一些学者将传统的个性化推荐方法移植到了医生推荐课题中,取得了一定的进展。Waqar等人结合内容

收稿日期:2021-09-16

基金项目:国家社会科学基金重点项目"数智驱动的在线健康资源挖掘与智慧服务研究"(22ATQ004);2022年度华中师范大学基本科研业务费(人文社科类)交叉科学研究项目"基于量化自我技术的个体健康管理研究"(CCNU22JC033)。

作者简介:周鑫(1995-),男,江西吉安人,硕士研究生,主要从事网络信息组织和数据挖掘研究;熊回香(1966-),女,湖北鄂州人,教授,博士,主要从事网络信息组织和数据挖掘研究;肖兵(1993-),男,四川资阳人,博士研究生,主要从事网络信息组织和用户行为研究,通讯作者:xiaobing@mails.ccnu.edu.cn。

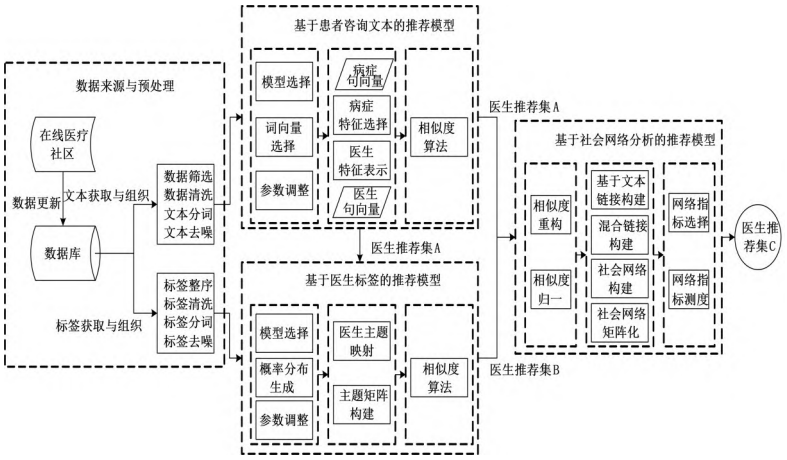

图1 医生推荐模型架构

Figure 1 The framework of physician recommendation model

库、协同过滤和人口学过滤方法，提出了一种医生推荐系统[9]；李勇等人基于推荐过程中语义相关度的考量缺陷，提出了一种协同过滤推荐算法[10]。

目前推荐方法多基于病历处方等传统医疗文本进行单一方法的推荐，混合推荐的研究较少，对海量的医疗UGC利用不足，对较新算法和模型的运用也有所欠缺，这使得在信息日益增长环境下丰富资源的利用程度、检索结果的可用程度、医生推荐的精确程度差强人意。因此，本文基于患者咨询文本和医生标签，融合机器学习领域、社会学领域相关方法和模型，设计了一种混合推荐算法。相关研究成果能够从独特视角丰富个性化推荐的理论与方法体系，促进情报学领域融合交叉学科发展；也可以帮助患者治疗和康复，节省医疗成本，节约社会资源，具有理论与现实意义。

3 推荐模型构建

本文构建的医生推荐模型是一种针对患者咨询病症，挖掘在线医疗社区中相关患者咨询文本和医生标签的语义关联，完成对症推荐的模型，完整架构如图1。该模型主要分为数据来源与预处理、基于患者咨询文本的推荐模型、基于医生标签的推荐模型及基于社会网络分析的推荐模型四个子模型。

由图1可知，该模型实时从在线医疗社区收集更新每一位医生的患者咨询文本数据和标签数据，对不同推荐子模型需要的数据提供预处理和组织。首先将患者咨询的病症作为参数输入基于患者咨询文本的推荐模型中，输出与该病症最匹配的医生集合A。然后将集合A输入基于医生标签的推荐模型，召回部分医生，扩展医生推荐集合B；最终将两个子模型推荐的医生集合(A+B)输入基于社会网络分析的推荐模型，得到最终的医生推荐集C。

3.1 基于患者咨询文本的医生推荐模型

患者咨询文本可以给予相似用户直观的建议，相关推荐能够起到改善推荐结果，扩大推荐覆盖面的作用。本节引入Word2vec词向量模型，将咨询文本内容映射成词语的向量形式进行相似度计算，使数据矩阵更为稠密，提高文本之间相关的可信度。

3.1.1 文本训练与评估

Word2vec是Google在2012年提出的一系列基于深度学习的用以产生词向量的模型，本质上是一种降维模型，利用上下词推测可能出现的词[11]。Word2vec分为词向量计算结构(Skip-gram)和连续的词袋模型(Continue bag-of-word CBOW)。其中，CBOW擅长处理300M以下的小型语料。因此，采用CBOW模型对文本进行训练。

为证明训练结果有效，可以作为下一步推荐过程的语料而不影响推荐精度，需要对训练结果进行可信度评估。为了不失一般性，选取代表性的词语作为训练测试集，计算词间的余弦相似度，找到与训练测试词语最相似的其他训练词语，进行训练效果评价。通过相似词的排序，判断得到的训练结果是否有效。有效的模型训练结果应当符合相似度预期，从而作为后续研究的基础。

3.1.2 医生与病症特征表示

(1)医生特征表示

将医生对应的患者咨询文本数量记为n。选取$k(k<n_{min})$个具有代表性的文本，这些文本是点赞数最高、查阅数最高、时间最近等最能表征该医生特征的文本。对k个预处理后的文本进行词向量表征，每个文本都被分解为m个特征词，每一个特征词都拥有一个词向量。整合该文本中所有保留下的训练词向量，并计算该向量集合的均值以表示该文本的文本句向量，以表征该文本对应医生的特征。那么，对于医生j的序号为k的文本，记其第i个训练词的向量为v_{ki}，其句向量则表示为$(v_{k1}+v_{k2}+v_{k3}+\cdots+v_{km})/m$。

(2)病症特征表示

患者输入的病症词包含在患者咨询文本中，经词向量训练后即拥有词向量。与医生特征表示同理，一个或多个病症可用词向量进行表征。当患者咨询的症状词数量为1时，该症状词的词向量即表征当前病症；当咨询的症状为组合症状，即病症词数量大于1时，由各病症向量取均值，构造病症

句向量,以表征病症特征。对于疾病训练测试集$W_1,W_2,W_3,\cdots,W_m$,m为训练集的词语数量,计算病症句向量为$(v_{W1}+v_{W2}+v_{W3}+\cdots+v_{Wm})/m$。

3.1.3 医生推荐

特征表示后,采用余弦相似度方法,通过计算表征后的医生和病症之间的相似度排序进行医生推荐。对于每位医生,选取表征文本k个,通过病症句向量和文本句向量计算得到的余弦相似度也有k个。那么,需要对这k个相似度进行整合,以表征医生和病症之间的关联,得到医生和病症之间的相似度,计算方法如公式(1)。

$$sim\left(D_n,U_m\right)=\frac{sim\left(O_{n1},O_m\right)+\cdots+sim\left(O_{nk},O_m\right)}{k} \tag{1}$$

式中D_n表示医生n,U_m表示相关病症集合m,$O_{ni}(i<k)$表示医生n的第i个文本句向量,O_m表示相关病症集合句向量。那么$sim(D_n,U_m)$为医生n和病症集合m之间的相似度,$sim(O_{ni},O_m)$为医生特征文本i的句向量和病症句向量的余弦相似度。

根据医生与病症相似度的排序,选择相似度最高的n个医生,作为最终的基于患者咨询文本的医生推荐结果,记为医生推荐集A。

3.2 基于医生标签的医生推荐模型

上节所述方法考虑到了UGC中的关联信息,但是单一的基于内容的推荐方法存在一定局限。除了医生与患者之间的关联值得挖掘,医生之间的关联同样具备提供推荐的可行性。因此,可通过召回医生推荐集A的相似医生达到协同过滤推荐的效果。医生标签为协同过滤推荐提供了直接可靠的数据源。

本节引入LDA主题模型量化相关相似度。隐含狄里克雷分配(Latenet Dirichlet Allocation LDA)模型本质上是基于概率分布,对文本中可能存在的不同主题出现的概率进行建模的方法[12-14]。某一文档资源可能存在不同概率分布下的文档主题,而这些主题依靠与之相关的词语来表达,不同的词语表达同一主题的概率也会随之不同。模型引入语词概率的度量方式,展现文档语义层面的关系。

3.2.1 主题训练

计算医生—主题联合概率分布的目的是利用潜在的主题量化医生和医生之间、医生和标签之间的联系。传统的标签关系挖掘单纯从共现角度进行相似度计算,在Web2.0环境下会造成系统误差。医生在定义标签时没有特定规则,对于擅长同一疾病的医生集合,不同医生会用不同的标签进行表征,如“感冒”“咳嗽”“呼吸道感染”等,这些标签各自不同但含义接近。因此,LDA主题概率分布可以从语义角度弥补标签存在的不限定规则问题,提高标签表征相似医生的精度。

3.2.2 主题矩阵构建

本节通过挖掘不同医生基于标签的语义关系进行相似推荐。而构建主题矩阵,可以将医生之间的相似程度以主题概率分布的形式进行量化,进而计算医生之间的相似度。

(1)医生—标签矩阵构建

LDA三层结构中,主题层是隐性的,医生集和标签是显性的,因此可以根据显性的标签链接关系,构建医生—标签矩阵。那么有医生集合$U=\{u_1,u_2,\cdots,u_n\}$(n为医生总数),标签集合$T=\{t_1,t_2,\cdots,t_m\}$(m为标签总数),医生—标签矩阵表示为公式(2),其中,$t_{ij}(i<n,j<m)$为第i位医生的第j个标签。为阐述方便,该矩阵是一个m*n维矩阵;但LDA模型实际允许输入不规则的矩阵进行主题概率分布训练。

$$X=\begin{bmatrix}u_1\\u_2\\\cdots\\u_n\end{bmatrix}=\begin{bmatrix}t_{11}&t_{12}&\cdots&t_{1m}\\t_{21}&t_{22}&\cdots&t_{2m}\\\cdots&\cdots&\cdots&\cdots\\t_{n1}&t_{n2}&\cdots&t_{nm}\end{bmatrix} \tag{2}$$

(2)医生—主题矩阵构建

将显性的矩阵X作为模型训练的输入数据,设定潜在主题维度k,运用吉布斯迭代采样方法进行主题概率先验分布的拟合,构建医生—主题概率分布矩阵Y,如公式(3)所示。在医生集$U=\{u_1,u_2,\cdots,u_n\}$(n为医生总数)中对应标签集合$T=\{t_1,t_2,\cdots,t_m\}$(m为标签总数),那么有医生标签$t_{ij}(i<n,j<m)$,基于标签数据生成潜在主题$Z_p(p<k)$的概率分布。p_{ij}解释为第i个用户在第j个潜在主题上的分布概率。

$$Y=\begin{bmatrix}u_1\\u_2\\\cdots\\u_n\end{bmatrix}=\begin{bmatrix}topic_1&topic_2&\cdots&topic_k\\p_{11}&p_{12}&\cdots&p_{1k}\\p_{21}&p_{22}&\cdots&p_{2k}\\\cdots&\cdots&\cdots&\cdots\\p_{n1}&p_{n2}&\cdots&p_{nk}\end{bmatrix} \tag{3}$$

3.2.3 医生推荐

在计算各医生在主题层面的概率分布后,度量医生间的相似度则转化为比较潜在医疗主题的相似性。矩阵Y的概率分布距离通常采用KL(Kullback-Leibler Divergence)距离算法,如公式(4)。

$$D_{kl}\left(p,q\right)=\sum_{i=1}p_i\ln\frac{p_i}{q_i} \tag{4}$$

式中$D_{kl}(p,q)$为概率分布p和q的距离,当$D_{kl}(p,q)$趋近0时,概率分布相似度极高。但由于引入对数,使得$D_{kl}(p,q)\neq D_{kl}(q,p)$,是非对称的距离计算函数。为便于矩阵计算,改进KL距离公式,通过均值处理进行对称转换,构建新的概率分布距离计算函数,如公式(5)。

$$D_{kl}\left(p,q\right)=\frac{1}{2}\left[D_{kl}\left(p,\frac{p+q}{2}\right)+D_{kl}\left(q,\frac{p+q}{2}\right)\right] \tag{5}$$

该公式构建的是医生在主题分布中的距离,表示医生之间的距离差异。$D_{kl}(q,p)$取值区间为[0,1],越趋近0,表明医生之间的相似度越高。为便于后续计算,对其进行相似度转换,如公式(6)。

$$sim\left(a,b\right)=\frac{1}{1+D\left(a,b\right)} \tag{6}$$

通过相似度计算,获得基于医生标签的医生相似度集合。将医生推荐集A作为本节医生推荐的输入数据集,计算与A中每一位医生最相似的k名医生,构成新的基于医生标签的医生推荐集B。

3.3 基于社会网络分析的医生推荐模型

上文两种推荐模型分别存在患者选择医生的动机局限、标签定义和关联的随机性等问题。通过组合两个推荐集进行混合推荐,可以弥补各模型的弱点。组合的医生推荐集本身具有一定联系,擅长相关病症的医生集合可以构建一个社会网络,网络中最重要的结点可视作与病症最匹配的结点。因此,可基于社会网络分析进行医生推荐,得到网络中最重要的医生。

本节引入的社会网络分析用于描述和测量行动者之间的关系,是一种较成熟的定量分析方法[15-16],在情报学领域常被应用于合著网络、引文网络、竞争情报等研究。由于社会网络分析的点度中心性指标测度网络中节点与其他节点之间联系的重要程度[17-18],故本节采用该指标进行个体医生的推荐。

3.3.1 医生网络链接构建

网络是由网络结点和结点之间的链接组成的,如何测度两个结点的链接值,需要进行值的量化,且保证值的归一。在基于患者咨询文本的推荐模型中,相关病症与集合A中每一位医生的相似度是归一的;在基于医生标签的协同过滤推荐模型中,(A+B)中医生两两之间的相似度对于整个网络而言不是归一的。所以在构建医生网络链接的过程中,需要确定基于整个网络的归一的链接值,有以下三个步骤。

(1)病症与医生的相似度重构

在推荐集B中,医生与病症基于患者咨询文本的相似度由于词向量训练的系统误差,在一定程度上被低估。例如,对于推荐集A中排序第一的医生i、最后的医生j,与医生i最相似的医生k,当医生k在推荐集B中,没有充分理由认为医生k与病症的相关性会低于医生j,可以认为医生k与病症的相似度被低估。因此,需要重构推荐集B中的医生与病症的相似度,以提升网络链接的精度。

将医生i与相关病症基于咨询文本的相似度记为$sim_{(i_text)}$,同理医生k相似度记为$sim_{(k_text)}$,医生i与医生k基于医生标签的相似度最高记为$sim_{(ik_tag)}$,重构后的医生k基于咨询文本的相似度记为$sim_{(k_text_new)}$。那么,基于患者咨询文本的病症与医生相似度重构算法如公式(7)。

$$sim_{(k\_text\_new)}=\frac{sim_{(i\_text)}\times sim_{(ik\_tag)}+sim_{(k\_text)}}{2} \tag{7}$$

医生k是由医生i推荐而来,在医生i的视角里,医生k与病症的相似度可以定义为$sim_{(i_text)}$与$sim_{(ik_tag)}$的乘积。而实际上由词向量训练模型得到的医生k与病症的相似度为$sim_{(k_text)}$,因此取二值的均值代表重构后的医生k与病症的相似度。

(2)基于病症与医生相似度的网络链接值计算

上一步获得重构后归一的相似度集合,因此进一步计算医生之间基于病症与医生相似度的网络链接值。事件A发生的概率记为P(A),事件B发生的概率为P(B),两者互相独立,那么A、B至少发生一件的概率如公式(8)。

$$P(AorB)=1-P(\overline{A})\times P(\overline{B}) \tag{8}$$

当医生i与病症的相似度为$sim_{(i_text_new)}$时,可以认为医生i能够治愈该病症的概率P(i)=$sim_{(i\_text\_new)}$,对医生k有P(k)=$sim_{(k\_text\_new)}$。当两位医生同时被推荐,该病症能被治愈的概率P(i or k)记为$sim_{I(ik_text_new)}$。该值代表两位医生组合的病症相似度,将该值作为基于病症与医生相似度的网络链接值,记为$v_{(ik_text)}$,其计算方法如公式(9)。

$$v_{(ik\_text)}=1-\left(1-sim_{(i\_text\_new)}\right)\times\left(1-sim_{(k\_text\_new)}\right) \tag{9}$$

(3)融合医生与医生相似度的网络链接值计算

上一步计算的链接值仅是基于患者咨询文本推荐得到的链接值。为实现协同过滤的推荐效果,需要引入基于标签得到的医生相似度对网络链接值进行改进。在基于医生标签的协同过滤推荐维度上,医生i、k的网络链接值是归一的,直接使用医生间相似度$sim_{(ik_tag)}$来表征。将任意两位医生基于医生标签的网络链接值记为$v_{(tag)}$,融合医生与医生相似度的网络链接值记为$v_{(ik)}$,直接由公式(10)计算,将式(9)代入式(10),得到最终的医生网络链接值计算公式(11)。

$$v_{(ik)}=v_{(ik\_text)}\times v_{(ik\_tag)} \tag{10}$$

$$v_{(ik)}=\left(1-\left(1-sim_{(i\_text\_new)}\right)\times\left(1-sim_{(k\_text\_new)}\right)\right)\times sim_{(ik\_tag)} \tag{11}$$

经上述步骤计算的医生网络具有结点、结点链接和链接测度值。而此类关系数据常见形式是列表,但该形式不便于社会网络分析,需要将其转换成矩阵。社会网络矩阵是一个大小等于网络结点量的方阵。由此,将上述计算结果转化为医生社会网络矩阵,以进行社会网络分析。

3.3.2 医生推荐

在社会网络分析中,挖掘潜在的网络信息有多种角度,如中心性分析、凝聚子群分析和核心—边缘结构分析等。本文最终目的是挖掘出网络中最重要的医生结点,并将其作为与病症相匹配的最核心结点,进行医生混合推荐。因此,选择对网络进行中心性分析。运用中心度计算方法,对获取到的医生网络矩阵进行社会网络分析,挖掘出中心度最高的医生排序,获得网络中最重要的医生集合作为最终推荐结果,记为医生推荐集C。

4 实证分析

4.1 数据获取及预处理

“丁香医生”是医学文献检索知识传播的专业医疗媒体平台,核心用户达550万,医生用户超200万。系统的就医决策引擎基于医生同行的评议数据,着重考虑医生层面,忽略了患者间的联系。使用“丁香医生”主页提供的医生标签和患者咨询文本作为实证的数据来源,不仅可以验证算法的可行性,也可优化“丁香医生”的推荐精度。

4.1.1 数据收集

“丁香医生”中包含丰富的UGC文本,选择适中大小的

表1 患者咨询文本数据(部分)

Table 1 Patient consultation text data (part)

医生	文本序号	文本内容
张宝	1	我29岁,男,昨天出现低烧体温37.4,无咳嗽,无咽喉疼…
	2	我从大年初二开始一直喉咙痛、中间有段时间长溃疡…
	3	喉咙疼大概有十天了,之前喝了蒲地蓝口服液效果不太好…
	…	…
	100	前天低烧37.6,吃了退烧药,昨天体温36.7,早上又37.5…
艾永梅	1	我女朋友今天体温37.5了,有点担心是不是新冠…
	2	今年40岁,近期有咳嗽症状,咽喉肿痛伴随呼吸有点困难…
	3	早上起床后,嘴里很苦,不知但怎么了!
	…	…
	100	酒喝多了,第二天胃痛吃什么吐什么
陆金帅	1	武汉人,咳嗽有好几天了,然后昨天开始胸口正中间有点疼痛…
	2	陆医生,您好,我这两天有点轻微咳嗽,目前每天都有量体温…
	3	陆医生你好 我老公现在出现嗓子痛 嗓子痒 不怎么咳嗦…
	…	…
	35	头痛有两天,喉咙也痛,上肢痛,不咳嗽,不发烧…
谢琪		……
	1	我的母亲63岁了,患有骨质疏松,在治疗中。我觉得她平时…
	2	老师你好,我外公76岁,膀胱癌,做了膀胱全切除…
	3	您好,我朋友41岁,这周一做的脑部肿瘤手术,手术过程…
	…	…
	10	医生我到了冬天穿很多,背心也特别冷,刺骨的冷…

表2 医生标签数据(部分)

Table 2 Physician tag data (part)

医生	医生标签
张宝	心力衰竭 心脏病 心血管系统疾病 高血压病 冠心病 重症肌无力 肺炎
艾永梅	高脂血症 糖尿病 痛风 慢性肾衰竭 高血压 肥胖症 冠心病 妊娠期糖尿病
陆金帅	急性酒精中毒 心脑血管疾病 急性上呼吸道感染 慢性阻塞性肺疾病 高血压急症
…	…
谢琪	蛋白质能量 营养不良 肿瘤

实验数据集,获取范围限定在“普通内科”医生,医生主页信息包含姓名、基本信息、擅长方向标签、患者咨询文本等。通过Python对2020年2月26日0时在线的296位“普通内科”医生的患者咨询文本和医生擅长方向标签分别进行采集。用患者咨询文本表征医生特征,文本数量不宜过小,不然会导致医生特征表示不准确,增大推荐误差。设定患者咨询文本数少于5的医生不能被患者咨询文本所表征,故筛选出110位医生,8376条患者咨询文本,其中,医生对应的患者咨询文本数量最大值为100(平台展示的最大数量),最小为5。

4.1.2 数据预处理

(1)数据筛选和清洗

对采集的数据进行清洗,去除数据中的无效信息,如姓名、时间、标签等,将每一条文本处理到一行,并去除行间空格,整合成方便代码处理的格式。

(2)文本分词与去噪

在进行Word2vec词向量训练前,需要把数据处理成分词后的格式,并且去除无意义的词语,以降低训练结果噪声。本文采用目前Python中文分词包中效果最好的Jieba进行分词,并选择常用且效果好的“哈工大停用词表”过滤无意义的词。患者咨询文本数据如表1所示,医生标签数据见表2。

4.2 基于患者咨询文本的医生推荐

4.2.1 文本训练与评估

将患者咨询文本输入到CBOW模型中,参数选择100维,利用Python+Gensim对该文本进行词向量训练。Gensim是一款开源的第三方Python工具包,支持包括LDA和Word2vec等多种主题模型算法,在小语料处理中具有方便快捷且准确度高的特征。保留训练词最低词频默认为5,部分词向量训练结果如表3所示。

词向量训练后,需要评估训练结果。模拟普通患者实际使用中的情形,通常只会笼统简短地描述病症,因此选取患者症状“咳嗽”+“胸闷”进行测试,测试结果见表4。

表3 患者咨询文本词向量训练结果(部分)

Table 3 Word vector training result of patient consultation text (part)

	1	2	3	4	5	…	100
咳嗽	-0.01790669	-0.33933419	0.34496513	0.92208272	0.93792659	…	0.10166514
胸闷	-0.43444124	-0.23471412	-0.00300824	0.89045662	0.80176687	…	0.62328011
发烧	0.003195192	-0.16810364	0.20673151	0.6885705	0.39605367	…	-0.45278853
体温	-0.2303288	-0.29572448	-0.70795316	0.99125737	0.025519924	…	-0.30792186
医生	0.98656034	-0.2499271	0.31704214	-0.79299921	0.55289787	…	-0.38054213
…	…	…	…	…	…	…	…
蛋白粉	0.05604941	0.017167469	-0.04463150	-0.01846731	0.086673833	…	0.029833119

表4 测试词相似度(部分)

Table 4 The similarity of test words (part)

	词1 (相似度)	词2 (相似度)	词3 (相似度)	…	词10 (相似度)
咳嗽	打喷嚏(0.9442)	干咳(0.9362)	鼻涕(0.9235)	…	咽痛(0.8903)
胸闷	气短(0.9585)	酸痛(0.9579)	心慌(0.9498)	…	无力(0.9401)

与“咳嗽”最相似的病症词语为“打喷嚏”“干咳”“鼻涕”;与“胸闷”最相似的病症词语为“气短”“酸痛”“心慌”。可以看出,训练结果符合预期,可在此基础上进行后续研究。

4.2.2 医生推荐

按照3.1.2所述方法对医生和病症进行特征表示后,通过计算医生和病症之间的余弦相似度进行医生推荐。设定公式(1)中的k值与上文筛选医生时同步取5,即选择患者咨询文本中的前五个文本(丁香医生中点击量越高的文本排名越靠前)。沿用上文病症测试集“咳嗽+胸闷”,医生推荐的相似度结果见表5。

表5 文本句向量与病症句向量相似度

Table 5 The similarity of text sentence vector and disease sentence vector

	1	2	3	4	5	平均相似度
陈曼	0.879991	0.900939	0.707817	0.960965	0.685274	0.8269972
刘旭阳	0.874163	0.635074	0.810391	0.842782	0.816315	0.795745
张磊	0.795795	0.7807	0.835395	0.760892	0.805154	0.7955872
门士虎	0.807409	0.843177	0.789391	0.725856	0.745279	0.7822224
张宝	0.868617	0.772871	0.611848	0.836801	0.815071	0.7810416
…	…	…	…	…	…	…
黄茂梁	0.0478244	0.137337	0.603907	0.802272	0.047173	0.30883342

根据表5所示,对病症“咳嗽+胸闷”,最终基于患者咨询文本的医生推荐选择相似度排名前五的集合。推荐医生为“陈曼”“刘旭阳”“张磊”“门士虎”“张宝”,记为医生推荐集A。

4.3 基于医生标签的医生推荐

4.3.1 主题训练

通过Python+Gensim工具包输入LDA主题模型进行联合概率分布训练。综合考虑普通内科可能存在的病症主题数量与训练集,将训练主题维度设为8,主题词数为5。训练后各主题词概率分布如表6所示。

4.3.2 主题矩阵构建

根据3.2.2所述方法,将训练后的医生—主题概率分布转化为公式(3)形式的概率分布矩阵,如表7所示,各医生均依据标签关系被映射到了潜在的医疗主题上。基于语义层面的计算依据,相较于传统的标签共现计算方法,具备更高的推荐精度。

4.3.3 医生推荐

将上节基于患者咨询文本推荐的医生集合A作为输入数据集,对相似医生进行召回,设定每一位医生召回两位相似医生,构建医生推荐集B。推荐集A和推荐集B中医生的相似度见表8。

表6 医生标签主题联合概率分布

Table 6 The joint probability distributions of physician tag topics

主题	联合概率分布
1	0.049*"急性" + 0.038*"感染" + 0.038*"发热" + 0.026*"病毒性" + 0.026*"肝炎"
2	0.077*"急性" + 0.064*"感染" + 0.044*"肺炎" + 0.026*"冠心病" + 0.023*"螺杆菌"
3	0.079*"急性" + 0.063*"感染" + 0.053*"疾病" + 0.041*"糖尿病" + 0.033*"冠心病"
4	0.109*"急性" + 0.058*"感染" + 0.052*"肺炎" + 0.041*"出血" + 0.029*"胃肠炎"
5	0.066*"急性" + 0.045*"腹泻" + 0.034*"咳痰" + 0.034*"腹痛" + 0.034*"咳嗽"
6	0.047*"糖尿病" + 0.035*"胃炎" + 0.035*"高血压" + 0.027*"高血压病" + 0.027*"溃疡"
7	0.053*"疾病" + 0.039*"急性" + 0.038*"感染" + 0.030*"肺" + 0.030*"阻塞性"
8	0.074*"冠心病" + 0.059*"糖尿病" + 0.053*"高血压病" + 0.045*"高血压" + 0.036*"疾病"

表7 医生—主题概率分布矩阵(部分)

Table 7 Probability distribution metrix of physician-topic (part)

	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8
张宝	0.01391	0.01391	0.0139	0.0139	0.01394	0.01391	0.90261	0.01391
艾永梅	0.01389	0.01391	0.0139	0.01389	0.55515	0.01391	0.0139	0.36144
陆金帅	0.01138	0.01138	0.34512	0.58658	0.01138	0.01138	0.01139	0.01138
李柏林	0.01251	0.91241	0.01251	0.01251	0.01252	0.01251	0.01251	0.01251
魏伟	0.01044	0.01043	0.01043	0.01043	0.48619	0.01043	0.28569	0.17595
…	…	…	…	…	…	…	…	…
谢琪	0.03133	0.03125	0.03127	0.03125	0.03125	0.78104	0.03127	0.03132

表8 医生推荐集A与B中医生相似度

Table 8 The similarity of physicians in recommendation set A and B

医生推荐集A	医生推荐集B	相似度
陈曼	王泽华	0.999999993
	韩旭	0.996482494
刘旭阳	巩雷	0.999999998
	李津	0.999941010
张磊	龙胜规	0.976713849
	范永周	0.868754894
门士虎	张海容	0.999999993
	蒋芳	0.999939286
张宝	于翠萍	0.999999989
	许蕊	0.999857501

由上表可知,对病症“咳嗽+胸闷”,基于患者咨询文本和医生标签推荐的医生集合(A+B)={“陈曼”“刘旭阳”“张磊”“门士虎”“张宝”“王泽华”“韩旭”“巩雷”“李津”“龙胜规”“范永周”“张海容”“蒋芳”“于翠萍”“许蕊”}。

4.4 基于社会网络分析的医生推荐

4.4.1 医生网络链接构建

(1)病症与医生的相似度重构

通过上节得到了医生推荐集合B。根据上文3.3.1小节中的公式(7)对集合B中医生与病症的相似度进行重构,重构结果见表9。

表9 医生集B与病症相似度重构结果

Table 9 The similarity reconstruction result of recommendation set B and disease

医生	医生与病症相似度	重构后的相似度
陈曼	0.8269972	0.8269972
刘旭阳	0.795745	0.795745
张磊	0.7955872	0.7955872
门士虎	0.7822224	0.7822224
张宝	0.7810416	0.7810416
王泽华	0.7100298	0.768513497
韩旭	0.6247114	0.724399816
巩雷	0.6343226	0.715033799
李津	0.5878008	0.69174943
龙胜规	0.7186998	0.747880418
范永周	0.3275622	0.509366237
张海容	0.7337292	0.757975797
蒋芳	0.6176042	0.699889554
许蕊	0.7203172	0.750623751
于翠萍	0.7049834	0.743012496

(2)医生网络链接矩阵构建

获得重构后归一的相似度集合后,按照3.3.1所述方法进一步构建医生之间基于病症与医生相似度的网络链接矩阵,由公式(9)计算得到,其结果如表10所示。再根据公式(11),计算融合医生与医生相似度的网络链接值,构建的网络链接矩阵见表11。链接值取小数点后三位。

表10 基于病症与医生相似度的网络链接矩阵

Table 10 The network link metrix based on the similarity of disease and physician

医生	陈曼	刘旭阳	张磊	门士虎	张宝	…	于翠萍
陈曼	1	0.642	0.655	0.640	0.650	…	0.645
刘旭阳	0.642	1	0.603	0.600	0.605	…	0.600
张磊	0.655	0.603	1	0.803	0.819	…	0.812
门士虎	0.640	0.600	0.803	1	0.603	…	0.598
张宝	0.650	0.605	0.819	0.603	1	…	0.943
…	…	…	…	…	…	…	…
于翠萍	0.645	0.600	0.812	0.598	0.943	…	1

表11 融合医生与医生相似度的网络链接矩阵

Table 11 The network link metrix fused the similarity of physician and physician

医生	陈曼	刘旭阳	张磊	门士虎	张宝	…	于翠萍
陈曼	1	0.964	0.965	0.962	0.962	…	0.956
刘旭阳	0.965	1	0.958	0.956	0.955	…	0.948
张磊	0.965	0.958	1	0.955	0.955	…	0.947
门士虎	0.962	0.956	0.955	1	0.952	…	0.944
张宝	0.962	0.955	0.955	0.952	1	…	0.944
许蕊	…	…	…	…	…	…	…
于翠萍	0.955	0.947	0.947	0.944	0.943	…	1

4.4.2 医生推荐

社会网络分析工具多达数十种,在参考王陆对典型社会网络分析软件工具的详细比较[19]后,本文选定UCINET,该工具是一种综合型的社会网络分析软件,是处理小型网络的

首选。将获得的网络链接矩阵导入UCINET,进行点度中心度计算,结果如表12,医生网络可视化见图2。

表12 医生网络点度中心度结果

Table 12 The degree centrality of physician network

医生	Degree	NrmDegree	Share
张磊	10.257	76.317	0.072
龙胜规	10.074	74.961	0.070
张宝	9.822	73.084	0.069
于翠萍	9.725	72.363	0.068
许蕊	9.716	72.292	0.068
陈曼	9.626	71.623	0.067
门士虎	9.600	71.429	0.067
张海容	9.541	70.990	0.067
王泽华	9.481	70.550	0.066
范永周	9.408	70.006	0.066
蒋芳	9.385	69.832	0.066
韩旭	9.239	68.749	0.065
刘旭阳	9.156	68.127	0.064
巩雷	8.970	66.743	0.063
李津	8.919	66.363	0.062

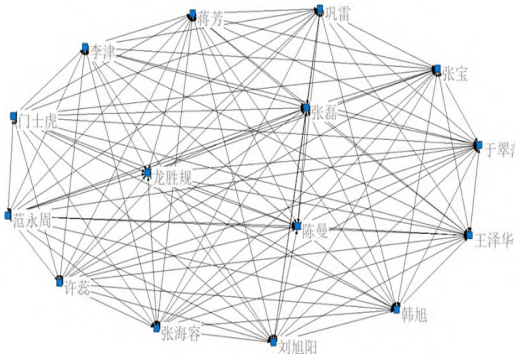

图2 医生网络可视化

Figure 2 The visualization of physician network

基于社会网络分析的推荐方法将推荐结果进行了重排,排序前五的医生为“张磊”“龙胜规”“张宝”“于翠萍”“许蕊”,构成最终推荐集C;当推荐医生数量为1时,最终医生推荐结果为“张磊”医生。在图2所示的可视化网络中,可以看到“张磊”“龙胜规”位于网络的中心。

5 推荐结果分析

在基于患者咨询文本的推荐过程中,排序最高的是“陈曼”医生;而经过本文算法推荐重排后,排序最高的是“张磊”医生。同时,最终推荐集召回了“龙胜规”“张宝”“于翠萍”“许蕊”医生。为衡量推荐结果的精度,对最终推荐集的五位医生擅长领域标签进行分析,标签如表13所示。

从表13可知,对于病症“咳嗽+胸闷”推荐的五位医生,擅长领域都包括呼吸系统疾病,如“呼吸道感染”“支气管炎”“肺炎”,这与患者咨询的病症吻合,推荐效果较好。本文提出的推荐算法的有效性得到了证实。接下来通过比较平台现在自有的检索推荐结果、基于患者咨询文本的推荐结果与本文算法的推荐结果,证明本文算法对推荐精度的提高。三者排序第一分别为“冯高科”医生、“陈曼”医生和“张磊”医生,通过表14对三位医生的标签进行比较。

由表14可知,“冯高科”医生擅长领域涉及胸闷,但主要擅长领域其实是心血管类疾病;“陈曼”医生的擅长病症标签较为笼统;而“张磊”医生的擅长病症较为详细且明确。所以系统从语义的角度判定推荐“张磊”医生的效果更好。接下来引入召回率指标进行咨询文本精度判断,计算方法如公式(12)。式中TP代表与病症相关的文本,FN代表与病症不相关的文本,Recall代表召回值,意为与推荐病症相关的文本占全部文本的比例。以三位医生的患者咨询文本是否含有“咳嗽”“胸闷”及其近义词,判定相关医生的召回值。部分与病症相关患者咨询文本如表15,召回情况如表16,可知“张磊”医生的Recall值也高于“冯高科”医生与“陈曼”医生。

平台的推荐方法受星级、价格等因素影响较大(冯高科医生为5星医生被推荐但病症相关度不足);而单独的推荐方法中,用于表征医生的文本选择会影响到推荐的效果,同

表13 推荐医生擅长领域标签

Table 13 The specialized field tag of recommendation physician

医生	医生标签
张磊	急性上呼吸道感染、急性呼吸窘迫综合征、慢性阻塞性肺疾病、肺炎、流感、上消化道出血、胰腺炎、脓毒血症、心力衰竭、冠心病
龙胜规	急性支气管炎、肺炎、急性上呼吸道感染、脑梗死、冠心病、高血压病、糖尿病、胃炎、胃溃疡、脑出血
张宝	肺炎、心力衰竭、心脏病、心血管系统疾病、高血压病、冠心病、重症肌无力
于翠萍	呼吸道感染、糖尿病、高血压病、痛风、心血管系统疾病、消化性溃疡
许蕊	呼吸道感染、消化性溃疡、肾盂肾炎、前庭神经炎、糖尿病、心脑血管疾病

表14 “冯高科”医生、“张磊”医生与“陈曼”医生标签比较

Table 14 The physician tags of "Feng Gaoke","Zhang Lei" and "Chen Man"

医生	医生标签
冯高科	冠心病、高血压、高脂血症、胸闷、普通心电图检查、心率和脉搏
陈曼	呼吸系统疾病、消化系统疾病
张磊	急性上呼吸道感染、急性呼吸窘迫综合征、慢性阻塞性肺疾病、肺炎、流感、上消化道出血、胰腺炎、脓毒血症、心力衰竭、冠心病

表15 病症相关患者咨询文本(部分)

Table 15 Patient consultation text related disease (part)

医生	文本序号	文本内容
冯高科	6	您好冯主任,本人男性,25岁。近三四日感到:胸闷、胸口偶尔疼下…
	13	病史描述:胸闷,偶有胸痛,查心电图部分导联J点抬高,st段抬高…
	22	冯医生您好!患者男52岁今年的2月25号胸闷、突发中风入院导致偏瘫…
	…	…
	49	偶尔感冒和压力大的时候心脏附近会感觉一阵疼痛,可能是神经痛…
陈曼	4	发热,咽喉发炎,流鼻涕,时长5小时,有点胸闷…
	9	嗓子痒的刺激性咳嗽,偶尔有痰,刚开始光鼻子不通气…
	10	最近几天嗓子有点不舒服,想咳,但咳得不厉害,晚上嗓子…
	…	…
	76	医生,你好。我今年51岁了,最近一个月咳嗽,咳白色黏痰…
张磊	10	干咳,嗓子略微疼痒,有一点点痰,鼻塞,少量鼻涕,之前…
	16	今年40岁,近期有咳嗽症状,咽喉肿痛伴随呼吸有点困难…
	23	鼻塞,胸闷,要用嘴巴呼吸,呼出来的气很热
	…	…
	81	你好,我想问一下喉咙痒,白天不怎么咳嗽,晚上有点咳…

时在文本内部病症词语的位置、词频,文本的长度、病症词在文中所占的比例等问题都会影响词向量训练的效果。本文提出的方法从患者咨询文本入手,通过医生标签的LDA主题模型训练进行了召回,利用社会网络分析方法进行混合推荐。至此,本文提出的融合标签和患者咨询文本的推荐模型在有效性和推荐精度提升方面都得到了证实。

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

表16 患者咨询文本召回情况

Table 16 The recall of patient consultation text

医生	患者咨询文本数	与病症相关文本数	Recall值
冯高科	55	8	14.5%
陈曼	100	17	17%
张磊	94	24	25.5%

6 结 语

目前的在线医疗社区功能粗泛,个性化的医生推荐效果难言理想。针对以上情况,本文设计了一种融合标签和患者咨询文本的推荐算法。算法分为三个子模型,首先利用Word2vec词向量模型对患者咨询文本建模,并通过特征表示算法和相似度算法得到推荐结果;其次通过医生标签进行推荐医生召回,利用LDA模型训练主题概率分布,改进KL算法得到推荐结果;最后基于医生社会网络构建网络链接,以中心性指标计算最重要的医生结点。文章利用"丁香医生"数据实证算法的有效性,提高了推荐精度。该算法价值在于,着眼于UGC等更具挖掘价值的内容而非传统的文本,采用较新的机器学习模型进行混合推荐以弥补单一推荐方法的缺陷;既考虑了患者与医生之间的匹配,也考虑了医生之间的关联,使算法更具普适性和实用性。本文的不足之处在于:限于研究条件,实证数据集较小,一定程度影响推荐的效果;仅通过中心性指标进行重要医生挖掘,可以进一步丰富挖掘效果。针对以上不足,可以进一步扩大数据样本,训练精度更高的模型,提高推荐精度;构建更大更丰富的医生社会网络,挖掘网络更全面的指标,实现医生推荐。

参考文献

1 马骋宇.在线医疗社区医患互动行为的实证研究——以好大夫在线为例[J].中国卫生政策研究,2016,9(11):65-69.

2 王嫣然,陈梅,王翰虎等.一种基于内容过滤的科技文献推荐算法[J].计算机技术与发展,2011,21(2):66-69.

3 刘继,邓贵仕.基于最近邻评价矩阵的混合协同过滤推荐算法[J].情报学报,2007,26(6):808-812.

4 Mehrbakhsh N,Othman B I,Norafida I.Hybrid recommendation approaches for multi-criteria collaborative filtering[J].Expert Systems with applications,2014,41(8):3879-3900.

5 Huang YF,Liu P,Pan Q,et al.A doctor recommendation algorithm based on doctor performances and patient preferences[C]//2012 International Conference on Wavelet Active Media Technology and Information Processing (ICWAMTIP),2012.

6 Makowski C,Marung H,Callies A,et al.Emergency doctors in Palliative patients Algorithm for Decision-making and Treatment recommendations[J].Anästhesiol Intensivmed Notfallmed Schmerzther,2013,48(2):90-96.

7 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(6):81-86,97.

8 陈亚明.基于医疗数据分析的疾病预测与处方推荐算法研究[D].洛阳:河南科技大学,2019.
9 Waqar M, Majeed N, Dawood H, et al. An adaptive doctor-recommender system[J]. Behaviour & Information Technology,2019,38(9):959-973.
10 李勇,黄俊.一种混合医生推荐算法的研究[J].信息通信,2018(2):67-70.
11 熊回香,李跃艳.基于Word2vec的学者推荐与跨语言论文推荐模型研究[J].情报科学,2019,37(12):19-26.
12 Blei D M,Ng A Y,Jordan M I.Latent dirichlet allocation[J]. The Journal of Machine Learning Research,2003,3(1):993-1022.
13 吴江,侯绍新,靳萌萌,等.基于LDA模型特征选择的在线健康社区文本分类及用户聚类研究[J].情报学报,2017,36(11):1183-1191.
14 张卫卫,胡亚琦,翟广宇等.基于LDA模型和Doc2vec的学术摘要聚类方法[J].计算机工程与应用,2020,56(6):180-185.
15 Perry B,Freeman P R,Oser C B,et al.Can Social Network analysis be used to Identify Doctor Shoppers[J]. Value in Health,2015,18(3):A128.
16 Rattanavayakorn P,Premchaiswadi W.Analysis of the social network miner (working together) of physicians[C]//2015 13th International Conference on ICT and Knowledge Engineering,2015:121-124.
17 Freeman L C.Centrality in social networks conceptual clarification[J].Social Networks,1978,1(3):215-239.
18 Kang G J,Ewing-Nelson S R,Mackey L,et al.Semantic network analysis of vaccine sentiment in online social media[J]. Vaccine,2017,35(29):3621-3638.
19 王陆.典型的社会网络分析软件工具及分析方法[J].中国电化教育,2009(4):95-100.

(责任编辑:赵红颖)

A Physician Recommendation Algorithm Based on the Fusion of Label and Patient Consultation Text

ZHOU Xin,XIONG Huixiang,XIAO Bing

(School of Information Management,Central China Normal University,Wuhan 430079*,China)*

Abstract:【Purpose/significance】Aiming at the current situation of loose structure of online medical information and insufficient precision of online physician recommendation,a physician recommendation algorithm based on label and patient consultation text is designed to improve the effect.【Method/process】Word2vec model is used to train patient consultation texts to obtain feature vectors and improve cosine similarity algorithm to calculate physician recommendation set A; LDA model is used to train labels to obtain probability distribution of projection on the subject and KL distance algorithm is improved to calculate set B; based on social network analysis theory,relevant algorithms are designed to reconstruct physician network links and select centrality index to obtain final set C.【Result/conclusion】Based on the relevant data of "doctor clove",the non-traditional UGC data enriches the availability of the algorithm,makes up for the deficiency of single recommendation method,and improves the accuracy.The proposed method effectively improves the accuracy of doctor recommendation.【Innovation/limitation】By fusing label and patient consultation text,social network analysis is used to realize physician hybrid recommendation.Though the central indicators of important doctors mining,the effect has room to improve.

Keywords:doctor recommendation; tag; Word2vec; LDA; social network analysis