基于 jieba 中文分词的在线医疗网站医生画像研究*

李 岩 郭凤英 翟 兴 陈晓倩 佟金铎

（北京中医药大学管理学院 北京 100029）

〔摘要〕 以在线医疗网站患者对医生的评价数据为基础，基于 jieba 分词技术从医生基本信息、医生所在医院、患者评价信息 3 个维度构建用户画像，借助 PowerBI 实现用户画像可视化，分析患者对医生的关注侧重点，为智能医生推荐奠定基础。

〔关键词〕 在线医疗网站；用户画像；jieba 分词；情感分析

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2020.07.003

Study on Doctors′ Portraits of Online Medical Website Based on Jieba Chinese Word Segmentation *LI Yan, GUO Fengying, ZHAI Xing, CHEN Xiaoqian, TONG Jinduo, School of Management, Beijing University of Chinese Medicine, Beijing 100029, China*

〔**Abstract**〕 Based on data from patients´evaluations of doctors on online medical websites, the user portraits are built based on the jieba segmentation technology from three dimensions, namely, the basic information of the doctor, the hospital where the doctor is located, and the evaluation information of the patient. The visualization of user portraits is realized with the PowerBI, the patients′ focus on doctors is analyzed, and lay a foundation for the intelligent recommendation of doctors.

〔**Keywords**〕 online medical website; user portrait; jieba segmentation; sentiment analysis

1 引言

医疗水平与人民生活幸福指数密不可分，要想使患者获得更好的就医体验，就要提升医疗水平，以及为患者推荐更优秀的医生[1]。人们可通过在线医疗网站了解医生概况，就医后还可在网上给予医生评价，这部分数据的潜在价值在于可以从患者角度真实了解医生，收集整理这些数据使其成为有价值的标签，为医生构建画像，为患者提供智能推荐功能，使用户得到最优的治疗体验。本文选取好大夫网站中与高血压疾病相关的全国医生大数据（包括结构化和非结构化数据），如医生与医院等级、医生职称、医生好评度、以往患者在线评论等，采用 jieba 分词等用户画像技术刻画医生画像，为后续医生推荐奠定基础。

〔收稿日期〕 2020－06－08

〔作者简介〕 李岩，本科生；通讯作者：郭凤英，副教授。

〔基金项目〕 北京中医药大学基本科研业务费项目教育科学研究专项课题“基于教育大数据的学生画像模型研究”（项目编号：2019－JYB－JYZX－004）；北京中医药大学教育科学研究课题“慕课视角下‘SPOC＋翻转课堂＋PBL’教学模式研究与设计—以医学数据分析课程为例”（项目编号：XJYB1922）。

2 研究背景

2.1 用户画像

交互设计之父 Alan Cooper 于 1998 年首次提出用户画像这一概念，将用户画像理解为虚拟化用户数据[2]。可以认为是以用户为中心，在具体场景下还原用户典型特征，为用户描绘一幅数据的画像，以更好地研究用户行为习惯以及特点。国内对于用户画像的研究要晚于国外[3]，但近年来发展较为迅速。用户画像数据可分为两大类：动态和静态数据，动态数据即不断变化的数据，随用户行为的改变而改变，如用户在网上浏览的足迹、评论、参与的话题讨论[4]等。静态数据是用户的个人信息，如性别、年龄等。根据得到的动态和静态两类数据，进一步处理提取出用户特征，这个特征就是用户属性的共性，最后对用户特征进行精炼得到易于理解和使用的用户标签。因此可以说用户画像就是将用户标签化[5]，用很多标签来简化对用户的刻画，用户画像构建过程是收集用户相关信息，最终整理为标签形式。

2.2 智能推荐

用户画像应用可分为 3 类：宏观（战略制定）、中观（运营优化）、微观（价值发现）[3]。用户画像应用场景十分广泛，涉及金融、医疗等多种行业，其中大多用于精准营销和个性化推荐[6]，以及搜索引擎[7]等。根据对大量用户数据的分析，构建用户画像，进而实现用户需求分析，根据用户偏好进行相应智能推荐，例如单晓红等利用北京地区酒店在线评论信息构建酒店和用户评价的用户画像，实现酒店精准营销[8]。本研究从患者角度出发，将医生作为分析对象，根据医生个人信息和对医生的在线评论信息整理出医生标签，构建医生用户画像，以实现医生推荐，为患者提供最合适和满意的医生。

3 用户画像模型构建流程

3.1 概述

基于在线医疗网站患者评价的医生用户画像构建主要包括数据准备、数据处理、用户画像可视化 3 个步骤，见图 1。一是数据准备，从好大夫网站上抓取医生基本信息、医生所在医院、患者评价数据，从 3 个维度构建用户画像。二是数据处理，包括数据预处理、标签构建和情感分析。三是用户画像可视化，使用 PowerBI 软件实现可视化并分析。

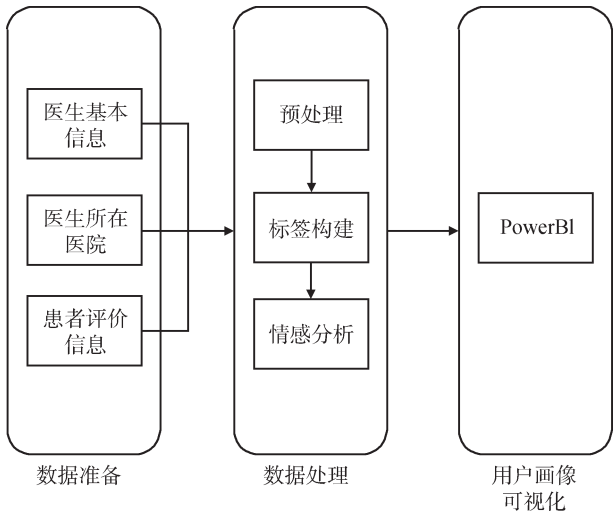

图 1 用户画像构建流程

3.2 数据准备

3.2.1 概述 早期用户画像一般采用访谈调研方法，随着互联网数据量的增加，目前用户画像主流方法为数据挖掘[9]。一般采用网络爬虫抓取 Web 页面数据，按照用户自定义规则实现自动抓取网页信息的程序。本研究使用八爪鱼爬虫软件实现数据抓取。该软件可直观模仿鼠标手动点击页面，根据用户需求选择需要抓取的数据，可选择各种可以避免反爬虫机制的方式，例如使用代理 IP、定时切换浏览器版本、定时清除 Cookie，在进行点击或抓取操作时可选择延迟时间，从而更加接近手动点击状态。

3.2.2 整体流程 首先在好大夫网站中检索治疗高血压疾病的医生，范围选择全国，将此页面的网址链接输入软件并打开，然后鼠标单击需要抓取的数据，包括医生姓名、职称、所在医院、擅长、评分（推荐热度），软件可自动解析 html，提取所选位置的文本数据，之后点击“查看详情”继续采集患者评价信息，选择在每个医生所在列表中循环抓取操作。进入“查看详情”页面后，发现患者评论

信息并未全部显示出来，需要继续点击查看全部投票，进入详细页面。之后可循环抓取患者评价信息，软件可循环点击“下一页”，抓取规则设置好后，在设置中选择定时清除 Cookie，最后运行程序即可。在运行过程中查看全部投票页面存在反爬虫机制，爬取过程中会出现网页错误，无法正常打开，笔者对点击选项进行延迟操作，随后再次运行程序便可正常抓取数据。最后将爬取的数据导出为 csv 格式，共抓取到数据 1 006 条。

3.3 数据处理

3.3.1 存在的问题 对获取的数据进行预处理，提高其质量[10]。通过分析数据，发现存在以下问题：一是个别列中会出现极少重复行，原因可能是数据抓取过程出现未知错误，需要批量进行查询和删除。二是评价原始列中的数据中存在空值，占比并不多，直接删除即可，对结果影响不大。三是医生评价数据中每条评价之间存在空行，需要进行批量处理。

3.3.2 处理过程 针对以上问题首先对数据进行删除重复行操作，使用 python 工具进行数据处理时需要使用第 3 方 python 库[11]，pandas 库是 python 中功能强大、用于数据分析的库，包括 DataFrame 和 Series 等很多实用工具，使数据处理过程更加简单、快速，其中 .duplicated 可以得到重复值判断的布尔值，而 drop_ duplicated（）方法可以直接去除重复项，返回唯一值，代码如下：

```
#导入 pandas 包并重命名为 pd
import pandas as pd
# 读取 Excel 中 Sheet1 中的数据
data = pd.DataFrame（pd.read_ csv（'all.csv', 'Sheet1'））
# 查看是否有重复行
re_ row = data.duplicated（）
#删除重复行
no_ re_ row = re_ row.drop_ duplicated（）
# 将去除重复行的数据输出到 csv 表中
no_ re_ row.to_ csv（" all2.csv"）
```

其次，针对评价列中存在的空值，即缺失值，使用 csv 的筛选功能，找到医生姓名，返回网页中查看发现没有评价数据，可能是患者只为其打了分或默认评价，导致为空，对于这部分数据采取删除整行操作[12]，不将其纳入到本次分析范围。最后对于每条评价之间存在空行，原因是单元格中存在换行符，此处新插入一列辅助列，使用 SUBSTITUTE 函数，该函数可实现将单元格内的换行符进行替换，在辅助行输入公式：=SUBSTITUTE（E2, CHAR（10）,""），其中 CHAR（10）代表换行符，向下填充即可将单元格中的换行符删除，再将辅助列的文本数据粘贴回原所在列。清洗后共得到 911 条有效数据，损失 95 条数据。

3.3.3 用户画像标签构建 对于医生基本信息，如姓名、所在医院、职称这类静态数据，直接使用作为医生标签即可。对于评价信息需要使用特征提取，找到最能代表医生的一个词语。特征即为关键词，是代表一篇文章或一段话中重要内容的一组词，可以使人们更加快捷地浏览和获取信息。本文采用 jieba 中文分词技术实现对文本特征提取。首先使用词频分析，选择 jieba 默认模式，即精准模式，定义高频词统计函数，

```
def get_ TF（words, topK = 10）:
tf_ dic = {}
for w in words:
tf_ dic {w} = tf_ dic.get（w, 0） + 1
return sorted（tf_ dic.items（）, key = lambda x:
x [1], reverse = True）[: topK]
然后定义过滤停用词函数
def stop_ words（path）:
with open（path）as f:
return [l.strip（）for l in f]
```

列出第 1 位医生评价数据中数量最多的前 10 个词语，发现其中最能体现医生特征的是形容词，反映医生性格特点。使用 jieba. analyse 中的 textrank 提取评论中的形容词，作为患者对医生评价的标签。打开经过预处理的 csv 文件：

```
data = pd.read_ csv（" all2.csv", encoding = " utf_ 8 _ sig"）;
```

遍历表格中评价所在列的所有单元格，对每个单元格进行特征词提取操作，将特征词保存在评价列右侧新列：

```
for i in range（len（data））：
split_ words = jieba. analyse. textrank（data. iloc［i，4］，topK = 1，withWeight = True，allowPOS = （'a'））
for item in split_ words：
corpus. iloc［i，5］ = item［0］；
```

3.3.4　情感分析　基于粒度不同可分为 3 类：文档、句子和词语。文档级情感分析通过训练分类器实现情感分类，句子级直接分析句子的情感类型从而判断文档情感，而词语级通过判断情感词的极性来实现情感分析[13]。本研究借鉴钟泉[14]使用的情感得分算法，使用知网 Hownet 情感词典，根据医生评价特殊性，在情感词典中加入一些表达对医生评价情绪的词语，如“医术了得”、“医术高明”、“医术高超”、“仁心仁术”、“名不虚传”、“医品”、“医德”等，这些词语表现患者对医生的积极评价；“匆匆的”表现患者对医生的消极评价。第 1 步，对每位医生的评论按“：”进行分句，检索每条评论情感词，与情感词典对照，记录情感词是积极或消极，设置变量存储相应的情感值。第 2 步，在当前情感词位置向前遍历，与程度词表进行匹配，为得到的程度词设置权值，然后与情感值相乘。第 3 步，利用与第 2 步同样的方法检索否定词，对否定词进行计数，若为奇数则将结果乘以 -1，反之乘以 1。第 4 步，对于每位医生，计算完一条评论的情感值，使用数组格式记录，从而记录该医生所有评论的情感值。第 5 步，将每位医生所有评论的积极和消极情感值分别计算均值并储存。第 6 步，利用郭丽环等[15]提出的情感判断算法，积极情感均值 > 消极情感均值则判断为积极，记为 1，反之记为 0，最后写入表格文件中。

3.4　可视化

将 PowerBI 作为可视化工具，利用 word cloud 插件，首先选择患者评价信息这一维度，见图 2，可以看出患者在评价医生时，是否耐心是一个重要指标。然后选择医生所在医院这一标签进行词云分析，见图 3，可以看到安贞医院、北京 301 医院、阜外医院心内科在治疗高血压疾病上在全国范围内居于前列。最后将医生姓名、职称、擅长、情绪以表格形式展现出来，当单击词云中的标签时，表格中的信息会相应跳转至所选标签对应的医生。将所属医院、医生姓名、评价和情绪分别以列表形式展现出来，得到 1 个级联列表，3 个列表实现联动，单击医院显示其治疗高血压的医生以及他们的评价和评价的情绪，直观展现所得结果，方便查找所需医院医生评价结果。

图 2　医生评价标签词云

图 3　所在医院词云

4　结语

本文以好大夫网站上患者对医生的评价为基础，从医生基础信息、所在医院、患者评价信息 3 个维度刻画“好大夫”网站上全国治疗高血压医生特征，借助 PowerBI 实现用户画像可视化，其中患者评价信息使用 jieba 分词和情感分析技术实现标签化，展现患者眼中医生真实面貌，反映出患者对医生的关注点，研究具有一定价值。实验结果表明“耐心”这一词被患者所提及的频率最高。医生与患者在医学专业领域知识方面有着巨大差距，如果

医生不能耐心为患者解答很容易引起医患矛盾。“耐心”只是在性格或态度层面评价医生的一个重要标准，医生专业技能更为重要，然而本次实验中患者评价信息并未提出有关医生技能方面的信息，可能由于评价信息多从患者主观出发，更多突出其体验、感受。本次实验尚有不足，数据来源还应扩大，获得有关医生专业技能方面的信息，有效利用医生职称等信息，使医生画像构建得更加完善。

参考文献

1 李贺，袁翠敏，李亚峰. 基于文献计量的大数据研究综述［J］. 情报科学，2014，32（6）：148－155.

2 刘漫. 基于用户画像的高校图书馆阅读推广模式构建［J］. 图书馆理论与实践，2019（7）：21－26.

3 宋美琦，陈烨，张瑞. 用户画像研究述评［J］. 情报科学，2019，37（4）：171－177.

4 曾鸿，吴苏倪. 基于微博的大数据用户画像与精准营销［J］. 现代经济信息，2016（16）：306－308.

5 王凌霄，沈卓，李艳. 社会化问答社区用户画像构建［J］. 情报理论与实践，2018，41（1）：129－134.

6 王智囊. 基于用户画像的医疗信息精准推荐的研究［D］. 成都：电子科技大学，2016.

7 王苗苗. 基于数据挖掘的互联网企业用户画像分析［J］. 现代经济信息，2018（10）：33－34，47.

8 单晓红，张晓月，刘晓燕. 基于在线评论的用户画像研究——以携程酒店为例［J］. 情报理论与实践，2018，41（4）：99－104，149.

9 朱钰涵. 在线视频社区中弹幕信息交互群体的用户画像研究［D］. 南京：南京大学，2019.

10 赵荣霞. 基于用户画像的 WordPress 博文推荐研究［D］. 北京：北京交通大学，2018.

11 高雅，苏艳，席方园. 基于 Python 的新浪微博用户数据采集与分析［J］. 电子设计工程，2019，27（20）：157－160，165.

12 殷丽凤，张浩然. 基于 Python 网上招聘信息的爬取和分析［J］. 电子设计工程，2019，27（20）：22－26.

13 尹裴，王洪伟. 面向产品特征的中文在线评论情感分类：以本体建模为方法［J］. 系统管理学报，2016，25（1）：103－114.

14 钟泉. 基于在线评论情感分析的手机推荐方法［J］. 安徽理工大学学报（自然科学版），2019，39（1）：80－86.

15 郭丽环，韩越，王伟. 在线评论对旅游者酒店选择的影响——基于细粒度文本情感分析［J］. 泉州师范学院学报，2019，37（6）：93－100.

2020 年《医学信息学杂志》征订启事

《医学信息学杂志》是国内医学信息领域创刊最早的医学信息学方面的国家级期刊。主管：国家卫生和计划生育委员会；主办：中国医学科学院；承办：中国医学科学院医学信息研究所。中国科技核心期刊（中国科技论文统计源期刊），RCCSE 中国核心学术期刊（武汉大学中国科学评价研究中心，Research Center for Chinese Science Evaluation），美国《化学文摘》、《乌利希期刊指南》及 WHO 西太区医学索引（WPRIM）收录，并收录于国内 3 大数据库。主要栏目：专论，医学信息技术，医学信息研究，医学信息组织与利用，医学信息教育，动态等。读者对象：医学信息领域专家学者、管理者、实践者，高等院校相关专业的师生及广大医教研人员。

2020 年《医学信息学杂志》国内外公开发行，每册定价：15 元（月刊），全年 180 元。邮发代号：2－664，全国各地邮局均可订阅。也可到编辑部订购：北京市朝阳区雅宝路 3 号（100020）医科院信息所《医学信息学杂志》编辑部；电话：010－52328673，52328672，52328686，52328687，52328670。

《医学信息学杂志》编辑部