文章编号：2095-2163(2023)06-0181-04 中图分类号：TP301.6 文献标志码：A

基于标签预测和奇异值分解的医生推荐系统的实现

徐志坚，易丽莹
（广州市第一人民医院，广州 510000）

摘 要：随着信息技术的发展，医疗卫生领域中医疗管理系统、医生推荐系统等得到广泛的应用。但是目前的医生推荐系统存在着信息混乱等问题，使得医生推荐的准确度不高。本文基于标签预测和奇异值分解的医生推荐系统的实现，可以为患者推荐合适的医生，减少了患者选择医生所用的时间，选择医生更有针对性，给患者带来便利。

关键词：标签预测；奇异值分解；医生推荐；推荐系统

Implementation of doctor recommendation system based on label prediction and singular value decomposition

XU Zhijian, YI Liying
(Guangzhou First People's Hospital, Guangzhou 510000, China)

【Abstract】 With the development of information technology, traditional Chinese medicine management systems and doctor recommendation systems have been widely used in the field of medical and health care. However, the current doctor recommendation system has problems such as information confusion, which leads to low recommendation accuracy. This article focuses on the research of doctor recommendation system based on label prediction and singular value decomposition. This system can recommend suitable doctors for patients, reduce the time for patients to choose doctors, and bring convenience to patients.

【Key words】 label prediction; singular value decomposition; doctor recommendation; Recommendation system

0 引 言

随着信息技术的不断发展，医疗信息系统也得到了快速发展，很多医院建设了医生推荐系统，提供推荐医生的服务工作。但是目前的大多数系统存在着计算复杂、精度不高等问题，亟需对医生推荐系统的算法进行优化，本文基于标签预测和奇异值分解算法，优化设计医生推荐系统，并进行了性能分析。

1 标签预测和奇异值分解算法概述

1.1 标签预测

使用多标签分类等算法对医生所擅长治疗疾病进行标签预测，根据医生的标签为患者推荐符合需求的医生。在对医生的擅长治疗标注时，需要建立特征向量和标签向量，如医生有 d 个标签，则标签向量 $\boldsymbol{Y}$ 的维度数量为 d，其中各维度分别表示该医生所擅长治疗的一种疾病[1]。如果医生擅长治疗一种疾病，则该维度的值为 1；如果不擅长，该维度的值为 0。

特征向量 $\boldsymbol{X}$ 包含以下 3 种信息：

(1)分类信息。包括医生所在医院、科室、职称以及其他医生对其评价等信息，需要对每项信息特征进行编码。以科室举例说明，设科室取值数量为 $\boldsymbol{O}_i$，表示为 P 维度的向量，每个维度表示一个科室，一个医生按照其所在科室将向量对应的纬度值定为 1。对于这类特征信息，各向量中只有一项为 1，其余各项均为 0。如果分类信息特征数量为 P，则特征维度为 $\sum_{i=1}^{P} \boldsymbol{O}_i$。

(2)数值信息。包含图文咨询平台数、电话咨询平台数、站内关注数等。对于这些数值信息，按照数值特征的数量表示为 q 维向量的方式进行处理[2]。

(3)文本信息。主要是对每位医生的介绍，对

作者简介：徐志坚(1995-)，男，学士，中级工程师，主要研究方向：信息系统管理。
通讯作者：徐志坚 Email：xuzhijian0107@163.com

收稿日期：2023-03-13

文本信息进行分词处理,分为 r 个不同的单词,则文本信息为 r 维变量。如果该医生的介绍中有向量各维度中的单词,则该项维度值为单词出现的次数;如果单词没有出现,则维度值为 0。

通过以上 3 项处理,每位医生的向量维度数 m 为: $\sum_{i=1}^{P} O_i + q + m$,$n$ 位医生的特征向量可以用 $n \times m$ 矩阵 $\boldsymbol{X}$ 表示,$\boldsymbol{X} = [x_1; x_2; \cdots; x_n]^{\mathrm{T}}$;标签向量用 $n \times d$ 矩阵 $\boldsymbol{Y}$ 表示,$\boldsymbol{Y} = [y_1; y_2; \cdots; y_n]^{\mathrm{T}}$。

1.2 奇异值分解算法

对于矩阵 $\boldsymbol{M}$,使用奇异值分解算法 $\boldsymbol{M} = \boldsymbol{U\Sigma V}^{\mathrm{T}}$ 将其分解为 3 个矩阵,其中 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 均为正交矩阵,$\boldsymbol{\Sigma}$ 为对角矩阵。如果将 $\boldsymbol{M}$ 作为物品的评分矩阵,则 $\boldsymbol{\Sigma V}^{\mathrm{T}}$ 表示将该物品映射到媒介空间[3]。

纯粹的奇异值分解算法中,$\boldsymbol{U}$、$\boldsymbol{\Sigma}$ 和 $\boldsymbol{V}$ 分别为 $m \times \hat{k}$、$\hat{k} \times \hat{k}$、$\hat{k} \times n$ 矩阵,$\boldsymbol{M}$ 为 $m \times n$ 矩阵,秩的数值为 $\hat{k}$。将 $\boldsymbol{\Sigma}$ 中各值取最大的 k 个数值保留,得到矩阵 $\sum_k$,此时 M 的近似值为 $U_k \sum_k V_k^{\mathrm{T}}$。

对于 $\boldsymbol{\Sigma}$ 取 k 个最大值,实现了两个目的,一是使向量的维度减少,而且使模型存储所需空间减少;二是将数值较小的奇异值去除,消除了影响小的因素,只保留了具有最强影响的因素,使推荐效果更加理想[4]。

奇异值分解法不仅能够预测,而且能够形成客户或物品的邻近点。使用奇异值分解对客户 - 物品的评分矩阵 $\boldsymbol{R}$ 分解,分别得到矩阵 $\boldsymbol{U}$ 和矩阵 $\boldsymbol{V}$,其中矩阵 $\boldsymbol{U}$ 每行代表客户的特征向量,能用来计算客户间的相似度;矩阵 $\boldsymbol{V}$ 每行代表物品的特征向量,能用来计算物品间的相似度[5]。

2 基于标签预测和奇异值分解的医生推荐系统设计思路

医生推荐系统由信息采集、信息管理和信息推荐三部分组成,系统结构如图 1 所示。首先,从权威网站上获取医疗信息数据,其次,对其进行去噪、抽取,将有用信息保存在数据库,根据患者需要为其进行医生推荐服务。

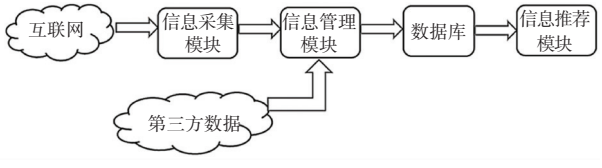

图 1 医生推荐系统结构图

Fig. 1 Structure diagram of doctor recommendation system

2.1 信息采集模块

医生推荐系统结构中,信息采集模块的作用是从网络上抓取与医疗相关的数据信息,核心是网络爬虫系统。网络上的医疗信息资源多种多样,各政府公开网站、医院官网、专业医疗信息网站等众多的医疗数据信息。为了提高数据信息的抓取速度,需要建立分布式网络爬虫,使多台计算机能共同合作。

2.2 信息管理模块

信息管理模块是将网络爬虫系统从网络上抓取到的信息和第三方数据进行处理,存入数据库。信息管理模块主要承担的任务:一是对抓取到的原始数据信息进行去噪处理;二是去除待存数据信息中的重复数据,并进行匹配;三是从需要存储的数据信息中发掘有价值的信息数据,如从与该医生进行论文合作的其他医生的信息中发掘与该医生专业相关的学术信息。

2.3 信息推荐模块

信息推荐模块为患者推荐合适的医生。该模块主要承担的工作任务:一是对标签预测模型进行训练,二是根据训练模型对数据库中医生擅长治疗的疾病和对医生的评价进行标签预测,三是根据患者的需要为其推荐合适的医生。

3 基于标签预测和奇异值分解的医生推荐系统的实现

3.1 信息采集系统实现

(1)信息搜索策略。网络爬虫在进行内容抓取前,需要前端数据机构提供一个 URL 集合,该集合可以基于网络重要性或其他结构特征选取,也可以人工选择;网络爬虫开始抓取该 URL 集合指向的网页,并通过新的网页提取新的 URL,直至没有新的网页需要抓取。在网络抓取时,采用广度优先策略,按照网页发现先后顺序进行抓取。

(2)信息采集实现。网络爬虫由三部分组成:一是爬虫客户端,用来接收来自爬虫服务器的下载要求,并对需要下载的网页页面发送 HTTP 请求,在响应完成后,将结果反馈给爬虫服务器;二是爬虫控制器,用来从服务器接受抓取数据,并在本地进行保存,同时从网页中提取 URL 地址,并将没有抓取的 URL 封装为 HTTP 请求,并将其返回到服务器。三是爬虫服务器,从爬虫控制器接收下载请求,并对相应客户端进行下载分配,将下载的结果反馈给爬虫控制器。爬虫服务器具有两种队列,一是请求队列,对待下载的 HTTP 请求信息进行记录;二是响应队

列，对获得响应的信息进行记录。网络爬虫系统结构如图 2 所示。

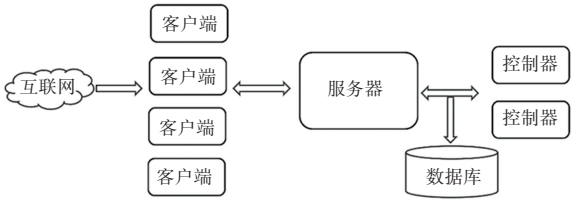

图 2　网络爬虫系统结构

Fig. 2　Network crawler system structure

通过对信息的分析，系统获取更加深入的信息，如医生的学术圈子、综合能力和医院的专长等信息。

在存储过程中，实现应用程序与系统之间的联系。使用存储过程能够使编程的复杂程度降低，因为存储表数量众多且关系复杂，对一个存储表进行修改会对其他存储表带来影响，使用存储过程使程序和数据库之间的联系简化；提升数据的安全性，通过存储过程中程序和数据库的交互，使数据库中的信息不必直接暴露在程序中，减少了恶意程序对信息的破坏；缩短程序和数据库开发周期，使用存储过程能够将程序和数据库存储分离，能够同时对两者进行开发，缩短开发周期。

(2)数据管理。系统读取数据后，对数据进行切分、过滤、校验和补全等操作后存储。在数据录入时会出现多种名称代表同一医院问题，需要人工进行干预，进行数据匹配和清洗；在存储过程中，需要将每项数据写入操作日志，对数据的操作进行分析，并将有错误的数据及出现错误的原因写入错误日志中，管理人员可通过错误信息进行特殊处理并将该类信息重新存储。

(3)自定义模板和插件。模板为 XML 文件，其中包含读取字段、读入接口和执行插件等内容，在管理人员录入新信息时，首先会读取模板信息，选择与信息的格式相应的模板即可完成信息录入工作。在将信息读入内存之后，为使信息符合数据库存储要求，需使用插件对信息进行处理。管理人员通过系统的插件接口，编写插件程序，完成读入内存后信息的处理工作。

(4)信息后台处理。后台管理可以根据医院、医生名字或其他关键字搜索，找到符合条件的医生的基本资料、学术论文发表以及学术圈子信息。医生的基本资料包括年龄、工作医院及科室、擅长治疗疾病等信息；学术论文发表能够展示学术论文发表的趋势，并对各学术论文详细介绍。信息管理系统根据医生学术论文找出与其合作的其他医生信息，并对医生的学术圈子进行总结，对医生的业务能力进行综合判断。

3.2　信息管理系统实现

信息管理系统的工作：将从各数据源获取的数据信息通过数据处理模块进行切分、过滤、校验和补全等处理，最终将处理后的信息存储到相互关联的数据表中，并在后台对信息进行处理。

(1)数据库存储。数据存储在数量众多、相互关联的信息表中，各信息表之间关系复杂。数据库中存储的医生信息主要内容由 4 部分组成，见表 1。

表 1　数据库存储医生信息主要内容

Tab. 1　The main content of doctor information stored in the database

序号	主要内容	包含内容
1	医生信息	医生年龄、擅长治疗疾病、所在医院科室等基本信息
2	医院信息	医院的等级、简要介绍和地址等信息
3	论文信息	发表学术论文的作者、刊号、关键词等信息
4	患者对医生评价信息	对医生的诊断、治疗和服务评价等信息

3.3　信息推荐系统

信息推荐系统使用标签预测和奇异值分解算法，对医生擅长治疗的疾病进行预测，并进行综合评分和排序。信息推荐系统主要是由模型训练、存储和推荐三部分组成，其主要步骤包括模型训练和存储、信息补全和预测、推荐和排序。信息推荐系统首先使用奇异值分解算法进行标签预测模型训练；其次，使用标签预测模型对医生信息进行预测和补全；最后，信息推荐系统根据患者的需要进行医生推荐，并按照推荐值顺序显示。信息推荐系统结构如图 3 所示。

离线进行标签预测模型训练和存储的原因主要是奇异值分解算法所需时间较长，在大量信息需要更新时才进行模型训练工作；其次，在训练模型时，需要人为的参与调整。在模型训练后以二进制形式存储，在需要时从文件中进行读取即可。

（下转第 188 页）