



北京航空航天大學

B E I H A N G U N I V E R S I T Y

第二十八届“冯如杯”学生学术科技作品 竞赛项目论文

基于自然语言处理的“智能问诊”系统

摘要

随着互联网医疗的发展，在线问诊网站也越来越多，但目前各种问诊网站鱼龙混杂，其中的医生资质也参差不齐，这些网站给患者医疗建议的质量堪忧。另一方面，由于我国医疗行业发展迅速，全科医生缺口逐渐加大，青年医师规范化培训资源短缺，我国医疗培训管理现状亟待改善。

面对现在问诊中为医生提供用药建议和培养医生、进行用药研究的需求，本文在传统的信息检索技术的基础上，结合近年兴起的深度学习技术和词向量模型，从文章主题提取的角度来处理这个问题，提出了基于 LDA 和 Word2vec 的用药推荐模型，构造一个新的“智能问诊”系统。该方法经过河北以岭医院等多家医院的测试，不同类型疾病系统推荐用药与《用药指南》相比前三项涵盖率达 81.6%，前五项涵盖率达 86.3%；病例匹配平均分诊准确度达 79%，平均疾病准确度达 58%。测试结果表明本文提出的方法对心内科、消化科、肿瘤科、呼吸科疾病的药物推荐和病例匹配效果良好，在辅助医生诊断、青年医师培养以及进行用药研究方面具有一定参考价值。

关键词：智能问诊；医疗问答；互联网医疗；Word2vec；LDA

Abstract

With the development of medical care on the Internet, there are more and more online medical consultation websites. However, various kinds of medical consultation websites are mixed and the doctors' qualifications are also uneven. The quality of these websites' medical advice to patients is worrying. On the other hand, due to the rapid development of the medical industry in China, the shortage of general practitioners has gradually increased, the training resources for standardized training of young doctors are short, and the current state of medical training management in China needs to be improved.

In the face of the current needs of doctors to provide medication advice and training for doctors and conduct medication research, this article combines the traditional deep information technology and the deep learning techniques and word vector models that have emerged in recent years. To deal with this problem, a drug recommendation model based on LDA and Word2vec was proposed to construct a new "intelligent inquiry" system. The method was tested in many hospitals such as Hebei Yiling Hospital, and the recommended drug use in different types of disease systems was 81.6% in the first three items compared with the "Guide to Medicine". The first five items covered 86.3%; The accuracy is 79%, and the average disease accuracy is 58%. The test results show that the method proposed in this paper has good results in drug recommendation and case matching for cardiology, gastroenterology, oncology, and respiratory diseases, and has certain reference value in assisting physician diagnosis, cultivating young physicians, and conducting drug research.

Keywords:

Smart Interrogation; Medical Q&A; Internet Health Care; Word2vec; LDA

目录

一. 绪论.....	1
二. 项目简介.....	1
(一) 主要创新点.....	1
(二) 项目背景.....	2
(三) 选题价值.....	3
三. 工具主要功能介绍.....	4
(一) 用药推荐.....	4
(二) 病例参考.....	5
(三) 诊断建议.....	6
四. 技术选型.....	6
(一) WORD2VEC 模型.....	6
(二) LDA 模型.....	7
(三) TF-INF	7
五. 技术难点.....	8
(一) 关键词提取不精确.....	8
(二) 药品与疾病匹配的库不健全	8
(三) 准确度问题.....	8
六. 工具使用范例	8
(一) 用药推荐.....	9
(二) 相似病例匹配	11
七. 改进方向.....	12
(一) 数据优化.....	12
(二) 模型优化.....	12
(三) 匹配算法优化	13

结论	13
参考文献.....	14

一. 绪论

目前网络医疗问答社区已是人们频繁访问的一类网站。根据有关统计调查结论表明，当身体不适时，有 90% 的人会通过网络搜索相关信息来了解病因，如拇指医生、丁香园、寻医问药等医疗问答社区正悄然改变着传统的医疗生态。在这类网站上，人们可以根据身体的状况，通过在线提出问题方式来获取有价值的信息。尽管这些医疗问答系统经过多年的发展和完善，使用起来已经很便捷，但它依然面临一系列挑战。例如对于大多数互联网用户，由于缺乏专业的医疗知识，在其身体不适时，对症状的描述非常模糊，甚至不够准确。这就导致在疾病知识问答中，对于同一个问题，在互联网上搜索到的答案也五花八门，且很难找到权威的解释，如果想知道大多数医生的共同回答，需要梳理归纳许多医生的回答才能总结出来。^[1]除此之外，医生在诊断过程中的用药也具有倾向性，使得不同医生的用药习惯可能有很大差异，这非常不利于对病人治疗，也给医院的用药研究制造了障碍。

本文在传统的信息检索技术的基础上，结合近年兴起的深度学习技术和词向量模型，构造了一个新的“智能问诊”系统。首先在传统的词袋模型只考虑词法信息，不考虑句法信息的不足的基础上，提出利用基于贝叶斯模型的 LDA 模型，推测文章主题，从而提高对用户描述的整体理解，进而有效改善匹配的精度。此外，本文还突破了经典信息检索模型只匹配相同单词的限制，使用 Word2vec 模型引进词向量，从而降低了同义词的匹配差异。该方法经过河北以岭医院等多家医院的测试，测试结果表明本文提出的方法在药物推荐方面效果良好，在辅助医生进行诊断以及进行用药研究方面具有一定参考价值。

二. 项目简介

(一) 主要创新点

1. 本文在传统的词袋模型的基础上，利用 LDA 模型，推测文章主题，提高对用户描述的整体理解，进而有效改善匹配的精度。
2. 使用 Word2vec 模型引进词向量，从而降低了同义词的匹配差异。
3. 根据用户对症状的描述进行用药推荐，经不同类型疾病测试，与《用药指南》

相比前三项涵盖率达 81.6%，前五项涵盖率达 86.3%。

4.根据用户导入的自身病例匹配相似病例供医生参考。

(二) 项目背景

当前社会经济的发展以及医疗科技的进步促进了现代人寿命的增长和健康状况的改善，同时随着生活水平和民众健康意识的不断提高，人均期望寿命与健康寿命趋异，使得民众对健康服务消费的需求不断增长，并呈现多层次、多样化的结构特点，这就促使医疗机构由“以疾病治疗为中心”向“以健康促进为中心”的医疗服务模式转变中，大力促进健康管理服务发展。



图 1 互联网医疗产品细分领域分布

《2010-2014YTD 中国互联网医疗投融资报告》显示，在互联网医疗概念方兴未艾的今天，在线问诊产品占到了互联网医疗产品的近三成，较其他细分领域具有压倒性优势，这说明在线问诊不仅仅是互联网医疗的重要组成部分，还是未来发展的方向。目前国际上关于管理健康需求的主要实践包括：需方管理策略、完善服务供给、医保政策导向等，然而目前我国医疗病例管理流程欠规范、缺乏系统的、动态的病例管理服务，在健康管理服务中尚未发挥应有的作用。

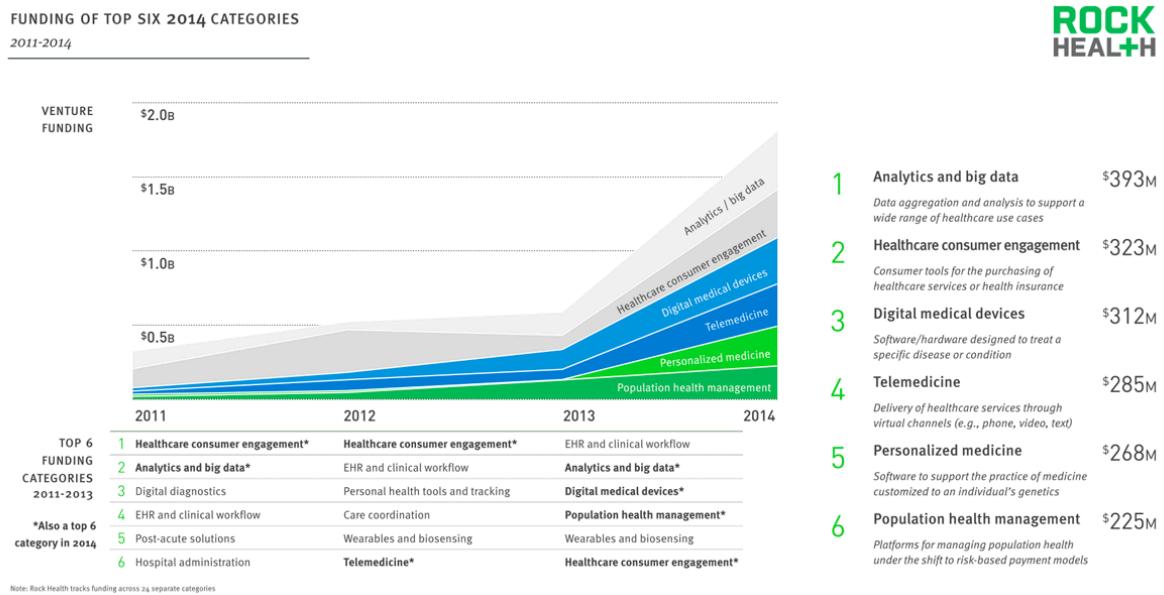


图 2 Funding of top six 2014 categories

而在技术领域，互联网发展如火如荼，美国著名互联网医疗孵化器 Rock Health《2014 互联网医疗投资年度回顾》显示，2014 年 258 家互联网企业获得超过 200 万美元的投资，单季投资额超过 100 亿美元。随着移动互联网条件的不断成熟，未来智能医疗的应用将更加广阔，更好地应对紧急突发状况。^[2]从 2014 年互联网医疗投资获得最多的六大类来看，我们提出的创意分别与大数据分析、消费者参与、个性化医疗以及健康管理类别相关，是今后重点投资的主方向。

（三）选题价值

1. 辅助医生进行诊断

随着“互联网医疗”这一概念的兴起，如百度“拇指医生”一类的在线问诊网站纷纷出现，但由于我国医疗行业尚无统一的病例、诊断标准，导致各种问诊网站鱼龙混杂，医生资质也参差不齐，而“智能问诊”以机器学习为基础，数据更为可靠，流程更加便捷，能够提高医生工作效率；同时也使更多的民众更快地获得更优质的医疗服务成为可能；与各类在线问诊网站相比有着很大优势。

2. 助力青年医师培养及用药研究

我国全科医生规范化培训的临床教学基地为总床位数大于等于 500、科室设置齐全的三级甲等综合类医院。部分三甲医院虽然拥有丰富的教学经验、先进的教学设备与齐

全的病种，但整体仍处于专科理论系统状态，缺乏正规的全科医学科，接受过全科教育、具有全科带教资格的老师也异常缺乏。^[3]

“智能问诊”系统可以为全科医生、处于规范化培训的住院医师提供外脑，降低青年医生培养成本，提高培养效率，缩短青年医生成长时间，为青年医生获取经验提供有效途径；

以上国人身体及医疗管理现状、民众健康意识的提升、科学技术的支持以及国家政策的带动等因素共同促使了“智能问诊”的诞生。

三. 工具主要功能介绍

(一) 用药推荐

用户可以选择输入自身症状，经过以 TF-INF 为基础的关键词提取解析其含义，再通过 Word2vec 模型自动进行用药推荐。

“智能问诊”系统经河北以岭医院心内科、呼吸科、消化科的测试，收到了较高的评价。

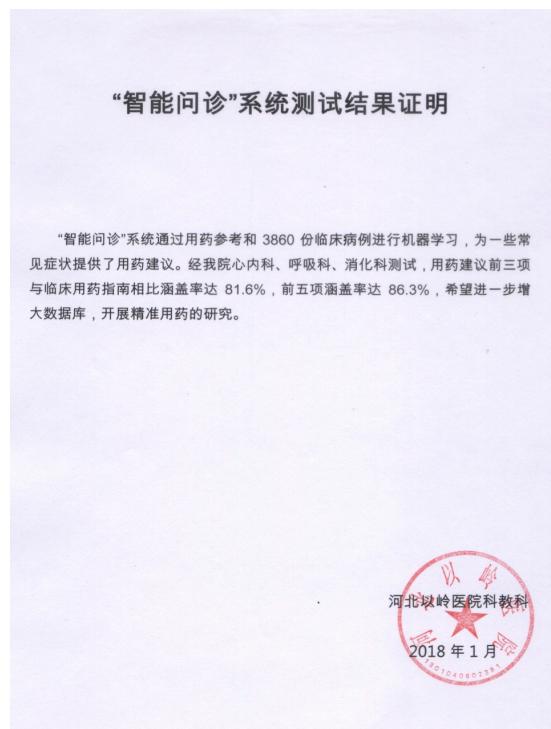


图 3 “智能问诊”系统测试结果

测试过程中，我们将输入症状得到的推荐用药与医院使用的临床用药指南进行对比，分析临床用药指南某一疾病所包含的用药与我们的推荐用药的包含关系，最终对于心内科、呼吸科、消化科的一些常见疾病的用药建议前三项和前五项在用药指南中的涵盖率分别达 81.6% 和 86.3%。具有较高的临床参考价值。

(二) 病例参考

用户可以选择导入自身病例，系统将会将用户导入病例与病例库数据进行比对，得到相似病例供医生参考。

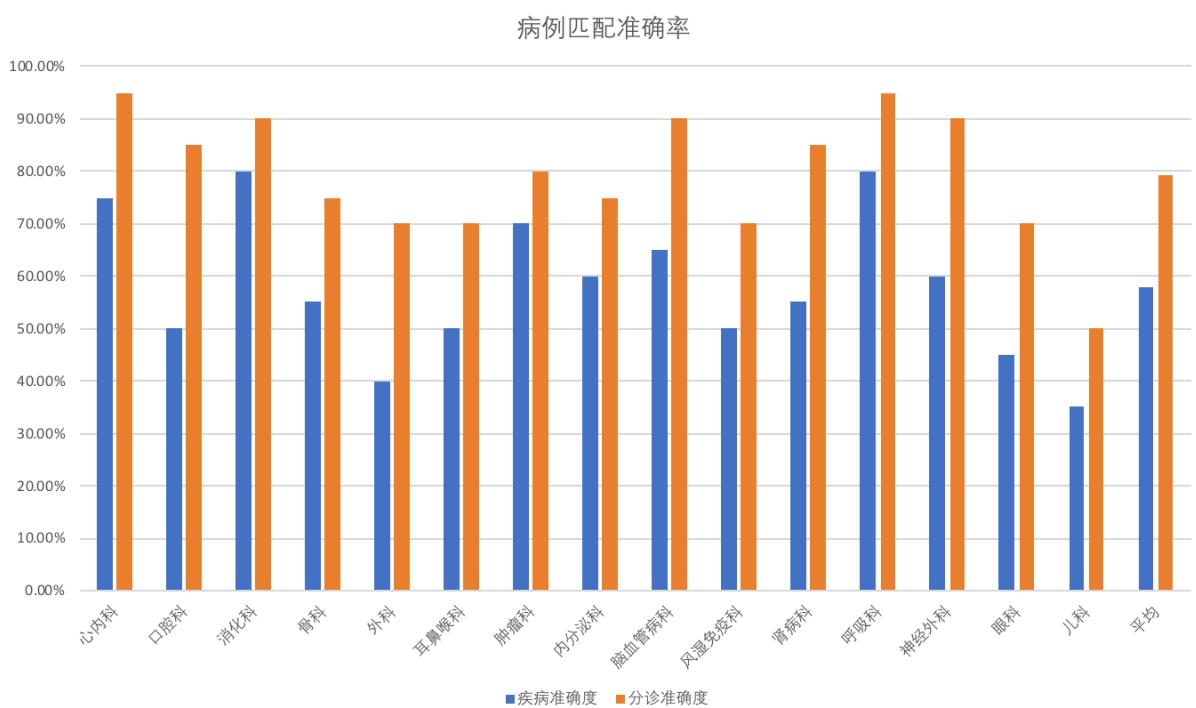


图 4 病例匹配准确度

由于本身医药方面并不是一对一的答案，所以评价通常机器学习模型的召回率和 F 值这种评价标准不太适用，所以我们建立了一个依托于医生的经验参数来测试训练效果，医生将真实诊断病例导入本系统，并对匹配到的相似病例进行评估，如果匹配到的病例所属科室与该科室相同，分诊准确度的分子分母加一，否则分母加一，如果匹配到的病例诊断与导入病例相同，疾病准确度的分子分母加一，否则分母加一。最终在心内科、口腔科、消化科等 15 个科室测试所得结果

如图 4 所示，从测试结果可知，本系统的病例匹配在心内科、消化科、肿瘤科、呼吸科准确度较高，而对于外科、眼科、儿科等科室准确度欠佳，平均分诊准确度达 79%，平均疾病准确度达 58%。

总的来说，匹配得到的相似病例与原病例有较高的相关度，可供医生作为参考，以更好确定患者的疾病类型并进行并发症分析。

(三) 诊断建议

综合推荐用药和相似病例，医生可以更准确地进行诊断。而对于医院而言，可以借此分析不同疾病之间可能存在的联系，推测并发症，或进行用药研究。

四. 技术选型

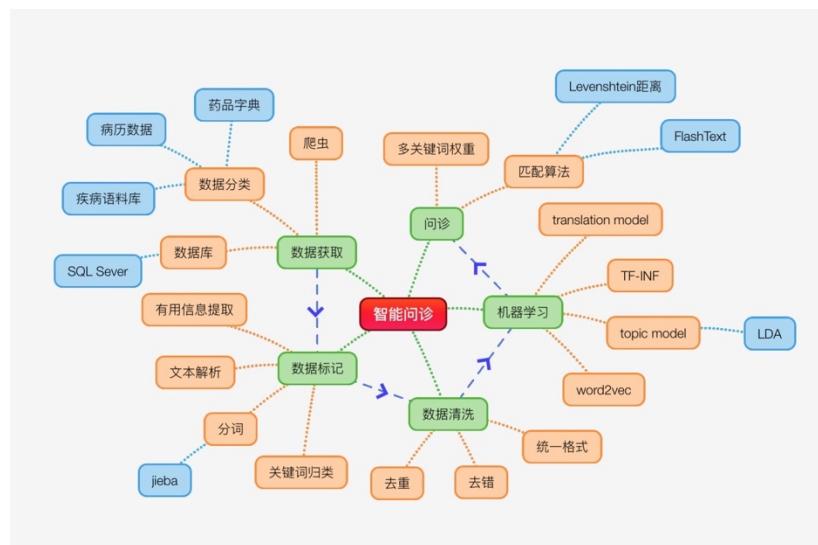


图 5 “智能问诊”技术框架

“智能问诊”基于北京朝阳医院提供的 3860 份完整病例以及临床诊断中使用的《用药参考》中的药品信息进行模型训练，经过数据标记和清洗得到训练集，最终训练得到可供问诊使用的模型。

(一) Word2vec 模型

word2vec 的训练模型是在于其具有一个隐含层的神经元网络。它输入词汇表向量，当看到一个训练样本时，对于样本中的任意一个词，我们把相应的在词汇表中出现的位

置的值设置成 1，否则设置成 0。^[4]它输出是词汇表向量，对于训练样本的标签中的每一个词，就把对应的在词汇表中出现所在的值设置成 1，反之设置成 0。对所有的样本，训练这个神经元网络。收敛后，我们可以把从输入一层到隐含层的那些权值，把它用作每一个词汇表中的词的向量。第一个向量就是 $(v_{1,1} \ v_{1,2} \ v_{1,3} \dots \ v_{1,p})$ ， p 是表示向量的维度数。所有虚框中的权值就是所有词的向量的值。有了每个词有限维度数的向量，就可以用到其它的应用中，因为它们就像图，有了有限维度数的统一意义的输入。训练 word2vec 的思想，是用一个词在文本中的上下文来表示这个词，这样就可以省去了人工去标注。

word2vec 模型在本文中用于近义词匹配，以减小误差。

（二）LDA 模型

LDA 是一种典型的词袋模型，即它认为一篇文档是由一组词构成的一个集，词与词之间没有先后顺序的关系。一篇文档可以包含多个 topic，文档中每一个词都由其中的一个 topic 生成。

看到一篇文章后，我们喜欢推测这篇文章是怎样生成的，我们可能会认为编写者先确定这篇文章的几个 topic，之后围绕这几个 topic 造句，表达成文。LDA 就是要根据给定的一篇文档，推断其 topic 分布。

LDA 是一种典型的 topic 模型，它可以将文档集中每篇文档的 topic 按照概率分布的形式给出；同时是一种无监督学习算法，在训练时不需要手工标注的训练集，需要的仅仅是文档集提取出来的关键词及指定 topic 的数量 k 即可；

LDA 可以被认为聚类算法：

1. topic 对应聚类中心，文档对应数据集中的例子。

2. topic 和文档在特征空间中都存在，且特征向量是词频向量。

3. LDA 不是用传统的距离来衡量一个类簇，它使用的是基于文本文档生成的统计模型的函数。^[5]

LDA 模型在本文中用于病例归类，以进行相关病例匹配。

（三）TF-INF

TF-IDF (term frequency-inverse document frequency) 是一种用于检索与探勘的常用

的一种加权技术。它是一种统计学方法，以评估字词对于一些文件或语料库中的其中一份文件的紧要程度。字词的重要程度随着它在文件中出现的次数正比增加，但会随着它在语料库中出现的频率次数反比下降。TF-IDF 加权形式常被搜寻引擎，作为文件与用户查询之间相关程度的评级。除 TF-IDF 以外，网上的搜寻引擎还会使用基于连结分析的评级方法，以确定文件在结果中出的顺序。^[6]

TF-IDF 算法在本文中用于关键词提取。

五. 技术难点

(一) 关键词提取不精确

虽然 jieba 分词与 TF-INF 能很好的从一句话里面提取出关键词，但是并不能保证非常好的提取效果，需要再校检。在进行数据清洗的过程中，需要导入疾病相关词库来提高分词准确率。

(二) 药品与疾病匹配的库不健全

目前各种疾病用药信息繁杂，没有一个系统的用药与疾病的库，很难找到一个全面可靠的库。这就要求我们根据已有数据自行建立一个疾病语料库供匹配推荐使用。

(三) 准确度问题

词对齐与结构对齐错误累积，可能导致精度不高。一种疾病可能翻译出来的多种药品的作用不一，很难推荐出最适合的。使用 Word2vec 模型分析词语之间的关系，解决近义词匹配结果相差较大的问题。

六. 工具使用范例

下面以界面图示介绍主要使用流程。展示网页使用基于 django 的 html+CSS+Javascript，并使用 ajax 进行数据传输。



图 6 网站主页

(一) 用药推荐

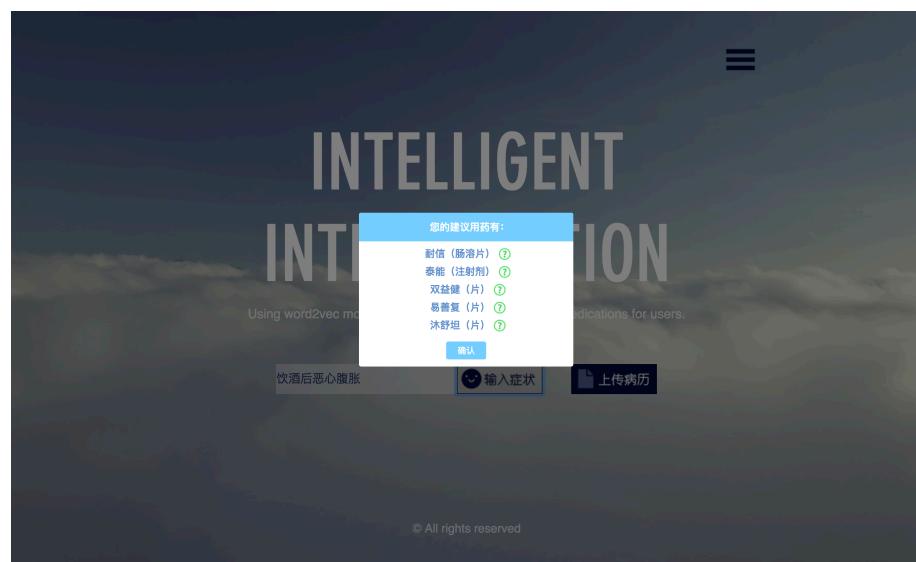


图 7 用户输入症状得到推荐用药

用户输入自身的症状后，点击“输入症状”，系统将会从用户输入的自然语言中提取关键词并进行用药匹配，最终得到五个推荐用药。



图 8 系统根据用户的反馈对模型进行修正

用户可以对推荐用药进行反馈，系统将会获得一个标注，随着用户的增多，反馈增加，模型将会不断进行迭代修正，其准确度也将逐渐提升。

The screenshot shows a pharmaceutical information website. At the top, there is a navigation bar with links for "医脉通首页", "软件中心", "频道导航", "登录", and "注册". The main header features a logo with a stylized "Rx" and the text "用药参考" (Drug Reference). Below the header is a search bar with the placeholder "请输入拼音首字母或药名" (Please enter the first letter of Pinyin or drug name) and a "检索" (Search) button. To the right of the search bar is a handwritten-style slogan "助您合理用药!" (Help you use drugs reasonably!). The main content area is titled "埃索美拉唑镁肠溶片" (Esomeprazole Magnesium Enteric-coated Tablets). It includes a product image, the drug name, and a chemical structure diagram. The chemical structure shows a complex organic molecule with various functional groups and substituents. On the right side of the page, there is a sidebar titled "相关药品:" (Related Drugs) which lists "耐信 [注射用埃索美拉唑钠]" (Nexium [Intravenous Esomeprazole Sodium]). Below this is a section titled "最专业的 医学资讯APP 医脉通" (The most professional medical information APP - Medical Pulse). The bottom right corner features a green download button with the text "六千多份指南 免费下载" (Over 6,000 guidelines available for free download).

图 9 点击查看药品详细信息

用户点击推荐用药可以查看“用药参考”中关于该药物的详细信息。

(二) 相似病例匹配

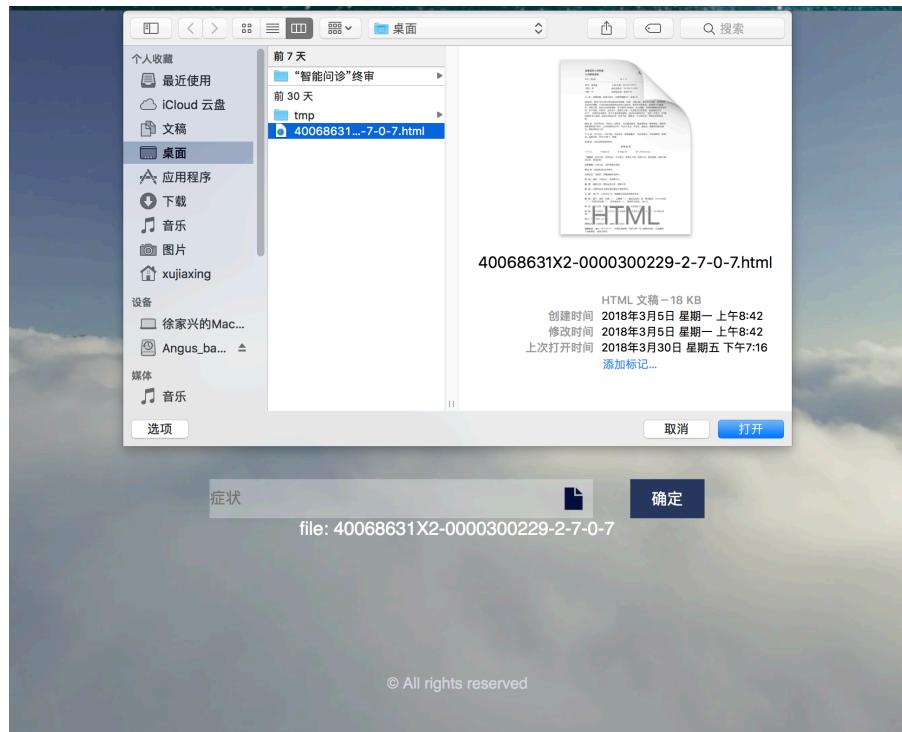


图 10 选择导入文件

用户选择导入本地病例进行匹配。

主诉: 间断咳嗽、咳痰10余年，加重伴喘憋4天，发热1天。

现病史: 患者10余年前无明显诱因出现咳嗽、咳痰，为粘白痰，秋冬季节加重，发作时伴有活动后喘憋。4天前无明显诱因再次出现上述症状，夜间可平卧休息，家属给予头孢地尼、切诺口服，症状未见明显缓解，曾于我科门诊就诊，查白细胞、中性粒细胞未见明显异常，给予切诺、沐舒坦、金荞麦片、祛痰灵口服。1天前患者出现发热，最高温度可至38.7°C，来我院急诊就诊，给予头孢美唑抗感染、沐舒坦祛痰等治疗，为进一步治疗，以“肺部感染”收入我科。患者自发病以来，饮食不佳，睡眠差，大小便正常。体重未有明显变化。

既往史: 否认肝炎史、疟疾史、结核史，否认糖尿病史、脑血管病史、精神病史。慢性阻塞性肺疾病10余年，心房纤颤病史20年。否认手术史、外伤史、输血史。磺胺类药物过敏史。预防接种史不详。

个人史: 生于北京，久居本地，否认疫水，疫源接触史。否认性病史。否认嗜酒史、吸烟史。适龄结婚，育有4个孩子，体健。

家族史: 否认家族性遗传病史

入院初步诊断:

慢性阻塞性肺病伴有急性加重

心房颤动(心房纤颤)

图 11 用户导入病例部分内容

主诉：喘憋10日，咳嗽4日，发热1日。

现病史：患者于10日前受凉后出现喘憋，安静状态下即出现，活动后明显。4日前无明显诱因出现咳嗽，咳白色粘痰，痰易咳出，自服阿奇霉素抗感染3日效果不佳。1日前患者出现发热，最高体温38.1°C，伴畏寒，无寒战，伴咳嗽，咳黄粘痰，痰中无血丝，痰不易咳出，伴尿频、尿急，无尿痛，夜尿2-3次，于外院就诊，予以克林霉素抗感染后仍发热，就诊于我院，血常规示WBC 7.84×10⁹/L, NE 74.6%, HGB 182g/L, PLT 107×10⁹/L, 生化示LDH 571U/L, HBDH 479U/L, TBIL 35.8umol/L, DBIL 9.12umol/L, IBIL 26.68umol/L, 乙型流感病毒抗原(+)，予以阿奇霉素抗感染、达菲抗病毒治疗，现为进一步诊疗收入我科。病程中无胸痛、无恶心、呕吐，无腹痛、腹胀，精神、食欲、睡眠欠佳，一般，睡眠一般，大小便如上述，体重无明显减轻。

既往史：风湿性心脏病及房颤病史5年，5年前行二尖瓣、主动脉瓣置换术，三尖瓣成形术，目前规律华法林抗凝治疗。否认冠心病、高血压、糖尿病等慢性病史。否认肝炎、结核等传染病史。否认重大外伤及输血史。对青霉素、头孢类、磺胺类过敏。

个人史：生于青海并久居，否认疫水、疫源接触史。吸烟20余年，平均20支/日，已戒7年。无饮酒嗜好。适龄婚育，育有3女，家人体健。

家族史：否认家族性遗传病史

入院初步诊断：

- 流行性感冒
- 风湿性心脏病
- 二尖瓣机械瓣置换状态
- 主动脉瓣机械瓣置换状态
- 三尖瓣成形术后
- 心房颤动(心房纤颤)

图 12 匹配病例部分内容

可以看出，用户导入的病例与系统匹配得到的相似病例的主诉和初步诊断十分相近，表明本系统具有较高的准确度。

七. 改进方向

(一) 数据优化

现阶段LDA模型主要是基于3860份完整病例进行的，在一定程度上体现出了机器学习在问诊领域的优势，但对于实际应用来说，这个量级的训练集还是远远不够的，所以我们将会通过增加高质量语料提高现有模型准确度。

(二) 模型优化

尝试translation、Seq2seq等模型，与现有模型效果进行对比。调整LDA模型中迭代次数、关键词个数等参数，获得效果更好的模型。

(三) 匹配算法优化

目前使用的 levenshtein 算法在搜索效率上差强人意，我们希望能在此基础上尝试 Flashtext 等更加精确高效的搜索算法。

此外，目前的关键词提取是基于 jieba 分词与 TF-INF 算法的，虽然基本可以正确提取出所需关键词，但对于专业性较强的医学词汇在识别上还存在困难，接下来我们将扩展用户词典，扩大关键词提取的适用范围。

结论

随着深度学习技术的成熟，未来各行各业都可以使用深度学习，以实现自动化的智能社会构建。智能问诊与人工智能技术结合是未来的研究趋势。

面对现在问诊中为医生提供用药建议和进行用药研究的需求，本文从文章主题提取的角度来处理这个问题，提出了基于 LDA 和 Word2vec 的用药推荐模型。本文提出的算法相对于传统的模型的算法有两大优势：

- 在传统的词袋模型的基础上，利用 LDA 模型，推测文章主题，提高对用户描述的整体理解，进而有效改善匹配的精度。
- 使用 Word2vec 模型引进词向量，从而降低了同义词的匹配差异。

从实验可见，上述的两个特点使得新模型的问答匹配准确率相对于传统方法有了显著的提升，实现了本研究最初设定的目标。

参考文献

- [1] 李超. 智能疾病导诊及医疗问答方法研究与应用[D]. 大连理工大学, 2016.
- [2] Malay Gandhi. Digital Health Funding: 2014 Year in Review [EB/OL]. [2018-4-4].
<https://rockhealth.com/reports/digital-health-funding/>
- [3] 刘小艳, 黄海. 全科医生规范化培训存在的问题及对策[J]. 医学与社会, 2016, 29(11): 88-91.
- [4] 待字闺中. 如果看了此文还不懂 Word2Vec , 那是我太笨 [EB/OL]. [2018-4-4].
http://www.sohu.com/a/128794834_211120
- [5] 箬 笠 蓑 衣 . 泡 沫 [EB/OL]. [2018-4-4].
<http://www.cnblogs.com/ruo-li-suo-yi/p/7813031.htm>
- [6] allenshi_szl.TF-IDF(term frequency-inverse document frequency)[EB/OL]. [2018-4-4].
https://blog.csdn.net/allenshi_szl/article/details/6283375