分类号： TN919 单位代码： 10335

密 级： 学 号： 21131035

硕士学位论文



**中文论文题目：****文本挖掘在智慧医疗病例循证系统中的应用研究**

**英文论文题目：****Research and Application on ECG Automatic Analysis and Regional Medical Health Services**

申请人姓名： 朱寒阳

指导教师： 金心宇

合作导师：

专业名称： 电路与系统

研究方向： 智慧医疗

所在院系： 信息与电子工程学系

**论文提交日期 二〇一四年一月**

**心电图智能分析及区域医疗健康服务研究与应用**



**论文作者签名：**

**指导教师签名：**

论文评阅人1：

评阅人2：

评阅人3：

评阅人4：

评阅人5：

答辩委员会主席：

委员1：

委员2：

委员3：

委员4：

委员5：

答辩日期：

**Research and Application on ECG Automatic Analysis and Regional Medical Health Services**



**Author’s signature:**

**Supervisor’s signature:**

External Reviewers:

Examining Committee Chairperson:

Examining Committee Members:

Date of oral defence：

浙江大学研究生学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **浙江大学** 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名： 签字日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解 **浙江大学** 有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 **浙江大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名： 导师签名：

签字日期： 年 月 日 签字日期： 年 月 日

## 致 谢

时光荏苒，岁月如梭。转眼之间，将近三年的研究生生涯即将结束。三年来，书山有路，学海无涯。我以求是为径，创新为舟，勤奋刻苦，脚踏实地。三年来，我在实验室浓厚的学习氛围中，夯实了自己的理论基础，提高了自己的专业技能，也理解了许多深刻的人生哲理。而这些，都离不开老师和同学们的支持和鼓励。

值此毕业论文定稿之际，首先要向我的导师金心宇教授表达诚挚的谢意。金老师学识渊博，治学严谨，因材施教，春风化雨，润物无声。在这三年里，不管是在学习抑或是生活上，金老师都给予我一丝不苟的指导和无微不至的关怀，使得身在杭城异乡的我，倍感温暖。不仅如此，金老师还针对性的指出了我的缺点和不足，让我受益匪浅。我会谨记金老师的谆谆教诲，在今后的人生道路上谱写更加精彩的篇章。

其次，我要感谢实验室的师兄师妹和师弟们。因为你们，实验室的生活充满了乐趣和精彩。我们每天一起学习，一起研究，一起编程，一起吃饭，一起玩耍，一起说笑，俨然一个快乐的大家庭。在这里，我要特别感谢蒋路茸、吴端坡、欧阳博、余长春、杨胜林、赵长勇、钱铮铮、武海涛、杨雄、马文涛等，在我烦恼忧愁的时候，是你们给了我巨大的信心和鼓励，让我变得坚强而自信。你们的陪伴，使得我的生活变得更加美好。

最后，我要感谢我的家人和亲朋好友，是你们的默默付出和支持，才有了我的今天，才使得我可以全身心的专注于浙大研究生的学业、杭城象牙塔中的生活。你们对我的殷切期望和默默支持，给予了我前进道路上的不竭动力，谢谢你们！

朱寒阳

2014年1月于求是园

## 摘 要

随着医疗信息化的发展，区域医疗健康服务的发展越来越重要。区域医疗健康服务不仅是当前解决群众“看病难，看病贵”的手段之一，而且可以在一定程度上改变医疗系统之间“信息孤岛”的现象，是国家“十二五”规划中医疗领域的重要内容之一。

心电图是人体心脏健康状况最直接的反映。对于心电图的监测和自动分析可以为大量的健康、亚健康和病患人群提供心电的预分析和异常预警，对于心电图数据的保存可以为病人提供心脏健康状况的趋势分析。当前对于心电图的检查只停留在临床方面，而对于病人心电图的保存和长期跟踪服务还很缺乏，因此本文对心电图智能分析和医疗健康服务进行了深入研究，并在此基础上构建了心电图智能分析的区域医疗健康服务应用。

在心电图的波形检测中，本文对比了差分阈值法和小波变换法，仿真结果显示小波变换法在QRS波群检测方面有着更好的表现。在心电图PVC自动识别中，本文分析了神经网络BP算法的不足，并且为了适应心电图千变万化的复杂性，本文提出了基于Gentle AdaBoost算法的心电图PVC自动识别，仿真结果验证了BP算法泛化能力的不足，同时Gentle AdaBoost对PVC自动识别比BP算法有着更好的识别精度。

本文在实现算法的基础上，设计了系统的整体框架，并使用MVC框架编码实现了电子健康档案功能模块，在心电图智能分析中，将算法核心模块用软件进行实现，并完成了心电图分析的基本功能。本文构建的电子健康档案子系统部分目前已经投入线上使用当中，有着一定的实用价值。

**关键词：** 心电图，智能分析，医疗健康服务，分类，机器学习

## Abstract

With the development of medical information, the development of regional medical healthcare is more and more important. Regional medical healthcare is one of the means in solving the “difficult and expensive medical treatment” currently, and to some extent, it can change the “information isolated island” phenomenon in medical system. Meanwhile, it’s one of the important contents in medical field of the Twelfth Five-Year Plan of our country.

ECG is the most direct response to human heart health. Monitoring and Automatic analysis of ECG can provide early analysis and warning of ECG abnormalities for a lot of healthy, sub-healthy and sick group. The ECG data can be stored to analyze patients’ heart health. Currently, check for ECG only remains in clinical aspects, and there are still lack of storage and long-term tracking service for patient ECG. Therefore, this paper makes in-depth in ECG analysis and medical healthcare, and build up applications on regional medical healthcare with automatic ECG analysis.

This paper compares the differential threshold method and wavelet transform method in the ECG waveform detection. And the simulation results show that the wavelet transform method has a better performance to detect the QRS complex. In the ECG PVC recognition, this paper analyzes the shortcomings of BP algorithm，as well as proposes a new method to recognize PVC based on Gentle AdaBoost algorithm, which overcomes the drawbacks in generation ability of BP algorithm. The simulation results demonstrate that the Gentle AdaBoost has better accuracy then BP in PVC recognition.

With the realization of algorithms, this paper designs the whole system framework and implements EHR system coding with MVC framework. In the intelligent analysis of ECG, this paper makes accomplishments in conversion from algorithm module to software and initial function of ECG analysis. The EHR system designed by this paper is has been put online with a good operation. And it has certain practical value for establishing the regional medical health service system.

**Key words:** ECG, Automatic Analysis, Medical Healthcare, Classification, Machine Learning

## 目 次

[致 谢 IV](#_Toc408324544)

[摘 要 V](#_Toc408324545)

[Abstract VI](#_Toc408324546)

[目 次 VIII](#_Toc408324547)

[第1章 绪论 1](#_Toc408324548)

[1.1 课题研究背景和意义 1](#_Toc408324549)

[1.2 国内外研究现状 3](#_Toc408324550)

[1.3 本文研究内容和意义 6](#_Toc408324551)

[1.4 论文的组织结构 8](#_Toc408324552)

[第2章 基本理论知识 9](#_Toc408324553)

[2.1 信息论基础 9](#_Toc408324554)

[2.2 文本信息抽取技术 14](#_Toc408324555)

[2.3 文本相似度算法 16](#_Toc408324556)

[2.4 本章小结 25](#_Toc408324557)

[第3章 基于openEHR的电子病历管理系统构建 26](#_Toc408324558)

[3.1 系统需求分析与设计目标 26](#_Toc408324559)

[3.2 系统总体设计 27](#_Toc408324560)

[3.3 系统功能设计及工作流程 29](#_Toc408324561)

[3.4 数据库设计 31](#_Toc408324562)

[3.5 本章小结 33](#_Toc408324563)

[第4章 文本处理和挖掘技术在病例循证子系统中的应用 35](#_Toc408324564)

[第5章 系统测试和分析 35](#_Toc408324565)

[5.1 心电图波形检测测试分析 35](#_Toc408324566)

[5.2 系统功能测试 38](#_Toc408324567)

[5.3 本章小结 44](#_Toc408324568)

[第6章 总结与展望 45](#_Toc408324569)

[6.1 总结 45](#_Toc408324570)

[6.2 展望 46](#_Toc408324571)

[参考文献 47](#_Toc408324572)

**作者简历及在学习期间取得的科研成果………………………………………..76**

## 第1章 绪论

### 1.1 课题研究背景和意义

随着我国经济和社会信息化的迅速发展，医疗领域也紧随其他行业与基础信息网络融合，进入了信息化时代。《2013年中国医卫行业信息化建设与IT应用趋势研究报告》的研究结果显示，2012年中国医卫行业IT投入达185.6亿元，较2011年同比增长22.6%；2013年医卫行业信息化建设投入将继续保持理性状态，呈现平稳增长趋势。2013年中国医卫行业的IT投资规模约为225.5亿元人民币，较2012年同比增长21.5%。信息集成平台将会整合包括HIS、PACS、CIS等系统的数据，实现信息共享、流程协作。电子病历系统作为医疗信息化系统的核心得到重视并将逐步健全和稳步发展[[1]](#endnote-1)。2009年IBM提出的“智慧地球”概念已经得到商界和政府的一致认可，并上升为美国郭嘉战略，在其包含的六大推广领域中，“智慧医疗”将成为未来医疗卫生信息化的主要发展趋势，同时伴随着智慧医疗的发展，医疗领域网络互联和智能终端的普及，信息量成爆炸性增长。据有关市场调研报告，医疗行业在数字宇宙中占比显著，数据量每年以48%的速度增长，是增速最快的行业之一，而信息技术将成为未来成功的关键推动者[[2]](#endnote-2)。如今炙手可热的互联网技术、云计算技术和人工智能技术如果以这些海量数据为基础将会有巨大的施展空间和应用前景[[3]](#endnote-3)。

医院数字化建设逐步展开，音频、视频、图像及各种格式的文档涌向医疗机构的数据中心，然后目前有关医学领域的数据标准并不完善，大部分数据是不兼容的，同时存在着大量的非格式化数据，比如在医疗行业，逐渐普及的电子病历的建设中，既存在结构化的电子病历数据，也存在非结构化的电子病历数据，而非结构化的电子病历数据的重要性并不比结构化数据低。因为描述病人病情的自然语言要比患者基本信息等结构化数据更丰富形象[[4]](#endnote-4)。从整个发展趋势来看，未来趋势应该是非结构化占主流，数据管理未来发展方向一定不可能完全基于表单式、结构式的，一定是基于一种半智能化的数据，数据挖掘也不可能完全基于标准化、结构化数据来的。非结构化数据的管理、分析利用成为新挑战[[5]](#endnote-5)。

另一方面，随着生物医学的快速发展，大量的生物医学文献发表出来。美国国立医学图书馆(The National Library of Medicine, 简称NLM)建立了当前国际上最权威的生物医学文献数据库：美国医学文献在线分析和检索系统(Medieal Literature Analysis and Retrieval System Online)，简称Medline。Medline收录1966年以来世界70多个国家和地区出版的3400余种生物医学期刊的文献，近960万条记录，目前每年递增30-35万条记录。这些重要的医学科研信息，80%是以非结构化的文本形式存在的。

文本挖掘技术在这种非结构化数据的大趋势下逐渐兴起，它处理的是以自然语言处理形式存在的各种文本，自然语言文本包含着丰富的信息和知识,可以从不同的层次如词汇、语法、语义和语用等多个层次处理和理解文本数据。目前文本挖掘已经成为数据挖掘中一个日益流行而重要的研究领域[[6]](#endnote-6)。在生物医学领域，文本挖掘技术可以辅助研究者们的工作，信息检索技术可以帮助用户在大量的医疗信息中快速定位有效信息；信息抽取技术可以在各种医疗文本数据中抽取中人们感兴趣的描述；文本分类技术帮助用户对信息分门别类，可以提供粗粒度的主题性质的筛选；假设生成技术可以从文档中挖掘出可能存在的实验假设事实等。目前文本挖掘研究在新闻、互联网等领域已经取得了很多成果，而医疗领域，由于各种资源例如医学语义网络、统一医学语言系统正在建立和完善，文本挖掘技术在医疗领域的研究工作仍然处于起步阶段。在我国，相对于英语语系国家来说，加上中文语言的复杂性，各种可利用的医学资源更加缺乏。因此，研究文本挖掘技术如果更好处理目前存在于医疗信息系统中的非结构化、半结构化文本数据，对提高医疗健康服务水平，具有重要的现实意义。

### 1.2 国内外研究现状

#### 1.2.1文本挖掘技术的发展现状

文本挖掘涉及多个学科领域，如信息论、机器学习、自然语言处理、计算语言学、统计学、图论等。在通用领域，通常能取得大量的可以供学习的样本以及一些人工标注过的数据，例如专门为1961年美国书面英语设计的Brown语料库，美国宾州大学整理的带有语法或者语义结构树标记的Treebank语料库，由北京大学计算语言学研究所完成的《人民日报》标注语料库，由国家语言文字工作委员会主持建立、面向全社会应用需求而建立的现代汉语通用语料库等。通用领域种应用文本挖掘技术已经取得了一些不错的成果，例如国内外大部分邮箱都已经设有垃圾邮件过滤系统，基于文本内容分析的过滤器性能要比基于简单模式或者规则匹配的过滤器好很多，通常能够过滤掉99%的垃圾邮件[[7]](#endnote-7)[[8]](#endnote-8)；Shota Ishikawa等人利用语义分析、文本聚类的方法从Twitter和Wikipedia等社交网络数据中挖掘出区域性的热点内容，对于市场预测，探测紧急事件等很有帮助[[9]](#endnote-9)；Lei Zheng等人使用多主题分布模型从Twitter中发现热点话题[[10]](#endnote-10)；任彬等人使用基于依存句法分析的文本挖掘方法，通过规则匹配的方式从社会媒体文本如微博文本中提取信息，实现了性别、地区、时间等维度的饮食习惯特色分析[[11]](#endnote-11)。

在医疗领域，随着信息化建设加速发展，传统基于规范格式化数据的处理方法面对医疗行业成爆炸性增长的非结构化、半结构化数据已经显得力不从心，而基于文本挖掘的技术则展现出其长远发展潜力和应用前景。目前国内外为了便于交流和信息获取，都建立了大量的生物医学数据库和生物医学信息检索引擎，例如美国国家医学图书馆提供的在线生物医学文献数据库Medline，它是现代生物医学研究发展和高价值文献数据库的代表资源；由中国医学科学院医学信息研究所/图书馆开发研制，整合了中国生物医学文献数据库（CBM）、西文生物医学文献数据库（WBM）、协和医大博硕学位论文数据库等多种资源的生物医学文献服务系统等。海量的医学文献给文本挖掘带来了巨大的施展空间，同时也带来了强有力的挑战。很多研究工作者利用文本挖掘技术对大量医学文献做了相关的研究。Aida Bchir等使用自然语言处理技术对Medline文献做预处理，然后使用机器学习分类器从文献中提取出疾病用药之间关系[[12]](#endnote-12)。Zhihao Yang开发的BioPPISVMExtractor系统利用SVM方法从生物医学文献中提取蛋白质互作用关系，达到了41.84%的召回率和55.41%的准确率[[13]](#endnote-13)。李毅等人通过建立命名实体规则、分类词表和领域本体对中文电子病历进行信息抽取和统计分析[[14]](#endnote-14)。

国内关于医学领域的文本挖掘的研究较少，发表的相关研究文章也不多，大部分研究工作都是基于国际上典型的生物医学文献数据库，这些数据库以英语语系为主。生物医学文本有其特有的复杂性，例如专业术语及其庞大，而且几乎不可能全覆盖；有多种术语系统，对同一个术语常常存在不同形式的名称。美国国立医学图书馆于1986年开始建立的一体化医学语言系统（Unified Medical Language System，简称UMLS）。UMLS提供的是医学科学领域内许多受控词表之间的映射结构，对医学知识进行统一的控制和组织，使概念的术语、属性、关系得到集中的展示[[15]](#endnote-15)[[16]](#endnote-16)[[17]](#endnote-17)。国内于2000年开始了中文一体化医学语言系统（Chinese Unified Medical Language System，简称CUMLS）的研究与建设，相继开展了一系列基础研究、试验与构建工作[[18]](#endnote-18)[[19]](#endnote-19)。 UMLS和CUMLS秉承语义网络与本体的建设理念，构建面向计算机应用的，集自然语言、主题语言、分类语言于一体的知识组织体系，这对于医学信息资源的组织和利用都起到了积极的推进作用[[20]](#endnote-20)。但CUMLS要晚于UMLS建设，目前也处于建设和完善阶段，对外并不开放使用，而UMLS每季度更新一次，且可以免费使用。再加之中文语言本身的复杂性，使得国内研究医学领域内中文文本处理、数据挖掘相比国外、其他语系要少。

#### 1.2.2 医学循证及其发展现状

循证医学(Evidence-based Medicine，EBM)是近十余年来在临床医学领域内迅速发展起来的一门新兴学科，它是一门遵循科学证据的医学，即遵循证据的医学，是国际临床领域近年来迅速发展起来的一种新的医学模式。循证医学是指认真、明确和明智地应用现有的最好证据，同时结合医生的个人专业技能和临床经验，考虑患者的愿望，对患者作出医疗决策。循证医学重视证据，证据及其质量是循证医学的关键，收集研究证据是循证医学实践一个不可缺少的重要组成部分。目前大量可供医学研究证据查询的来源，包括数据库、互联网、杂志、指南等[[21]](#endnote-21)[[22]](#endnote-22)[[23]](#endnote-23)。

循证医学研究人为搜集所有可靠的、高质量的同类科学研究进行系统评估后的证据最为可靠。具体实践就是结合临床经验与最好的证据对患者进行处理，包括提出问题，检索证据，评价证据，结合临床经验应用最佳证据。图1.1是循证医学在实践中的实施过程。

图1-1 循证医学实施过程

医学问题是根据当前具体的临床病例，总结和归纳所面临的临床问题，这些问题往往围绕着这个病例,包括病人的一些基本情况以及对病人的治疗措施；循证，即根据前面提出的一系列问题，收集大量在这问题上前人有所研究的方法和措施， 任何医疗决策的制定，都依赖于临床科研实践中所获得的最佳证据。证据的获得有很多，最常用的是检索相关的研究文献，结合他人的研究成果，往往能够在最大范围内更加客观和真实地匹配医学问题，这对评价证据的可靠性有很大帮助。卢延鑫等人应用自然语言处理和文本挖掘技术，对Medline数据库中的流行病学研究文献进行信息提取，解决循证医学中研究者人工搜寻证据时面临的数据量庞大的问题[[24]](#endnote-24)。Aaron M. Cohen等人提出一种基于流水线式的多重文本挖掘方法来帮助循证医学系统自动化摘要和更新最新的医学文献，这大大简化了医学领域专家手工归纳和循证的工作[[25]](#endnote-25)。评价证据。它是指运用方法学的评价标准将所检索的证据进行科学的评价，主要对真实性、可靠性和实用性方面进行评估，以确定最佳的证据。结合临床经验应用最佳证据是指结合医师对该临床问题的治疗经验和患者个体的情况进行治疗决策的定制。

在循证和结合临床经验的时候，除了研究文献以外，很多医疗机构还会有大量的历史病例数据，这部分数据中成功治疗的临床案例也可以做为证据，一个医师在行医过程中也常常会参考历史经验临床案例做出最终医疗决策。

### 1.3 本文研究内容和意义

目前国内的大部分医院都有医院信息化系统HIS（Hospital Information Sytem），但是各个系统基本上是独立的，相互之间也是不兼容，无法互通的。病人的病症、诊断信息、健康档案信息分散存在于各家医院系统中，由于医学数据的复杂性，现阶段依然缺少一个完善的标准来统一医学数据，各个HIS基本上是自行设计，数据标准也是各家一份，没有一个统一的标准使得分散的数据无法集中在一起运用人工智能、数据挖掘等技术手段分析和发现潜在的价值信息。2009年5月国家卫生部制定了《健康档案基本架构与数据标准(试行)》，明确提出了健康档案的基本概念，并强调了医疗健康档案的功能作用和实现标准[[26]](#endnote-26)。2011年，“3521工程”作为国家“十二五”期间卫生信息化建设的总体规划被提出，为我国医疗卫生信息化的发展指明了方向[[27]](#endnote-27)。在国外，医疗标准化的进程起步早，发展快。美国九十年代就开始制定和使用HL7标准。目前各厂家的医疗系统产品基本都符合HL7标准。同时openEHR规范作为另一个优秀的开放的电子健康档案（EHR）体系架构被提出，其目标是实现EHR系统内部以及EHR系统之间的健康信息共享，由openEHR机构专门负责制定，有关数据标准化的工作一直在努力着。

本文以浙江省某医疗合作公司提出的需求和所提供的数据为基础，以openEHR为电子健康档案标准构建了一个病人信息管理应用系统，旨在将各家不同医院信息系统中的数据统一起来，这使得有价值的医疗数据集中在一起，并在其上运用文本挖掘算法，提取出病例文本数据中的病症信息，以此为特征构建病例循证子系统，帮助医师快速检索相似的历史病例数据。

集合本课题的所有研究任务,将本文研究目的分解为如下几类:

1. 病人电子病历数据统一管理平台： 构建了一个基于openEHR的病人信息管理应用系统，能够将各家不同医院信息系统中的数据转换为统一的数据形式组织在一起。
2. 医学资源收集和预处理：中文医学文本处理目前还处于起步阶段，不像英文医学语言研究已经拥有丰富的知识资源库，例如UMLS中有医学词表（[Metathesaurus](http://zh.wikipedia.org/w/index.php?title=UMLS_Metathesaurus&action=edit&redlink=1)），语义网络（[Semantic Network](http://zh.wikipedia.org/w/index.php?title=UMLS_Semantic_Network&action=edit&redlink=1)），专家词典（[SPECIALIST Lexicon](http://zh.wikipedia.org/w/index.php?title=UMLS_SPECIALIST_Lexicon&action=edit&redlink=1)）三部分，对英文医学语言处理中词性标注，句法分析，语义分析都很有帮助。但国内医学领域内的专业资源例如中文医学主题词表（CMeSH）、中文一体化医学语言系统（CUMLS）等都还不完善并且没有对外开放使用。本文力图在现有条件下，搜集网络上已经存在的许多分类词表，包括通用领域词表（如国家语委现代汉语语料库[[28]](#endnote-28)）和医学领域词表（如国际疾病分类ICD-10、搜狗医学词库[[29]](#endnote-29)、学术名词资讯网[[30]](#endnote-30)等）。这样搜集的预料往往很庞大，而且很杂，文本利用语料特征适合度来筛选出合适的语料，最后以此为基础语料，通过分析和利用医学词库内在的特征和信息对词表进行扩展，作为最后的词表语料。
3. 病例数据标注和病症信息抽取：本文中构建的病人电子病历数据统一管理平台中提供了一个供医师对病例数据进行病症信息标注的模块，可以通过这个功能获得许多简单标注过的病例文本用于训练，结合之前词表语料，使用条件随机场（CRFs）模型进行病症信息抽取。
4. 病例数据相关度排序：使用病症信息作为病例数据的特征信息，构建一个病例循证子系统，根据医师输入的病症描述，利用文本相似度算法对历史病例数据进行相关度打分和排序，检索出最有可能相似病例。

本文构建的病例数据管理平台是针对当前医疗信息行业发展中对于数据统一化需求而进行的，这对于数据集中化，标准化都有一定的使用价值。在目前缺乏有效的中文医学语言资源的前提下，利用网络上现有的资源，通过文本处理和挖掘技术，提取病例数据中的病症信息并以此构建历史病例循证子系统对于帮助医师做出更加正确和客观的医疗决策有很重要的意义。

### 1.4 论文的组织结构

本文的各章节组织如下：

第一章，绪论。

第二章，基本理论知识。

第三章，基于openEHR的电子病历管理系统构建。

第四章，文本处理和挖掘技术在病例循证子系统中的应用。

第五章，系统测试和分析。

第六章，总结与展望。

## 第2章 基本理论知识

文本挖掘同时也成为文本知识发现（Knowledge Discovery in Text），是人工智能、机器学习、自然语言处理、数据挖掘、信息论、计算机科学等理论和技术相结合的产物。它的主要处理过程是对大量的文本内容进行预处理、特征信息抽取、结构分析、分类聚类、关联分析、语义分析等。

### 2.1 信息论基础

信息论是由香农（Claude Shannon）在20世纪40年代建立的理论体系，只不过当时他研究的实际目的是如何破解第二次世界大战时的地方密码，如何在非理想化的通信信道中传输尽可能多的信息。事实上，信息论的意义和应用范围早已经超出了通信的领域，它对许多学科如物理学、生物学、控制论、统计学、语言学、经济学等都产生了巨大的影响和作用。信息论在文本挖掘领域也也发挥着极大的作用。李晓光等人针对词、潜在概念、文本和主题之间的模糊关系，利用了信息论中熵压缩编码理论，提出了一种基于信息论的潜在概念获取与文本聚类方法，结果要优于基于词空间的文本聚类方法以及双向硬聚类方法[[31]](#endnote-31)。在中文文本分类的任务中，常常会用信息论中的期望交叉熵、信息增益、互信息、文档频率等作为分类特征[[32]](#endnote-32)。Adwait Ratnaparkhi用最大熵模型对通用预料进行词性标注，达到了96.6%的精度[[33]](#endnote-33)。在机器翻译中也常会使用互信息的方法来处理歧义词的问题。

#### 2.1.1 熵和互信息

熵(Entropy)的概念最早起源于物理学，用于度量一个热力学系统的无序程度。在[信息论](http://zh.wikipedia.org/wiki/%E4%BF%A1%E6%81%AF%E8%AE%BA)里面，熵是对不确定性的测量。在信息世界，熵越高，则能传输越多的信息，说明了不确定性越大，而信息量越多；熵越低，则意味着传输的信息越少，不确定性越小，信息量越少。对于一个随机变量，取值空间为，它的概率密度函数为，那么随机变量的熵定义为：

 (2.1)

将定义推广到两个随机变量的情况，则可以将视为单个向量值随机变量，服从联合分布，那么可以定义两个变量的联合熵为：

 (2.2)

定义一个随机变量在给定另一个随机变量的情况下的熵，它条件熵为：

 (2.3)

互信息(Mutual Information)也是一种测度，用来度量一个随机变量包含另一个随机变量的信息量。表示在给定另一个随机变量知识的条件下，原随机变量不确定度的缩减量，也就是信息量的增加量。信息量定义为：

 (2.4)

在文本处理中，互信息的方法非常有用，例如互信息可以表示一个词和其他词语有多大的联系，如果在一个语料中，词出现了42次，词出现了20次，两个词共同出现了20次，那么互信息，意味着词的出现使的下一个词是的信息量增加了18.38比特。

#### 2.1.1 上下文熵

熵用于表示信息的不确定性，在文本处理中，各种不同的熵经常用在无监督学习任务中。例如中文分词，新词发现等。Jin Hu Huang等人使用上下文熵对中文文本进行了分词，在边界判断上获得了93.2%的准确率和93.1%的召回率，在成词判断上获得了81.2%的准确率和81.1%的召回率[[34]](#endnote-34)。Moises Garcia等人使用上下文熵来对文本进行分类，方法简单而且训练时间很短[[35]](#endnote-35)。

上下文熵(Contextual Entropy，简称CE，又称边界熵，Branching Entropy，BE)，可用于刻画一个词的边界信息，定义如下：

 (2.5)

其中是上下文，是词表的词，对于文本来说，表示为字符串，一个词的词中位置的上下文熵会比词边界的上下文熵要低。定义左上下文熵(left contextual entropy)和右上下文熵(right contextual entropy)为：

 (2.6)

 (2.7)

如果一个字符串的左右上下文熵越大，其成词的可能性越大。

### 2.2 文本信息抽取

信息抽取是自然语言处理新兴的研究方向和技术，文本挖掘研究需要借助信息抽取的方法从非结构化、半结构化文本中抽取结构化内容，然后在此基础上进行各种维度的数据挖掘工作。从自然语言处理的角度来看有三个层面，文本理解、信息抽取和信息检索。最低层次的是信息检索(Information Retrieval)，信息检索系统操作的是词语，使用统计学计算文档中共现词语的相对频率，它对文本的理解仅仅是文档所包含的词语，很少的语法和语义分析会用到信息检索系统中，这也使得信息检索能够应用到大规模文本中，例如万维网搜索引擎。文本理解(Text Understanding)是语言处理的最高层次，它要求计算机不仅仅要理解文本的显示意思，还要理解文本的内在含义。信息抽取是介于信息检索和文本理解之间，需要对文本进行一定程度的理解，但与真正的文本理解还是不同的。在信息抽取中,一般只关心有限的感兴趣的信息，而不关心文本意义的细微差别以及作者写作意图等深层理解问题，因此也可以被看作是受限的文本理解( a restricted form of full natural language understanding)[[36]](#endnote-36)。信息抽取的主要任务是从文本中提取部分文本，然后赋予他们特殊的属性，典型的处理方法包括分词(tokenization)、句子切分(sentence segmentation)、词性标注(part-of-speech assignment)、命名实体识别(identification of named entities)。在更高的层级上，短语、句子得到解析，例如浅层句法分析(shallow parsing)，再以此为基础进一步做语义分析(semantic analysis)、情感分析(sentimental analysis)等。最后将需要的信息片段抽离出来存储到数据库库。文本研究的其中一个任务即是从病人病例文本中抽取出循证系统所需要的感兴趣信息，即病症信息。

#### 2.2.2 信息抽取算法简介

总体来说，文本信息抽取采用的模型主要有三类：基于词典的抽取模型，基于规则的抽取模型，基于机器学习和统计模型的信息抽取方法。基于词典的文本信息产后去需要构造模式词典，建立相关的模式知识需要很大的人力投入，最佳词典的产生也不容易，领域相关性强，单独使用很少，常和其他模型结合在一起使用[[37]](#endnote-37)。

基于规则的模型也叫做包装器模型(Wrapper模型)。它是一种软件构件，负责将数据和查询请求由一种模式转换为另一种模式。



图2-1 基于规则的模型信息抽取过程示意图

一般包装器包括三部分：规则库、规则执行模块和信息转换模型。应用包装器模型需要先构造抽取规则集合，然后包装器从文本信息，例如万维网的web页面，根据规则抽取特定的信息出来，转换成特定的格式描述信息，提供给其他信息系统作进一步处理。整个抽取过程如图2-1所示。

基于规则的抽取模型不是很好的适应领域的变化，对于每个领域都要有相应的包装器，维护成本高。目前有半自动化生成包装器的技术，通过图形界面，让用户标示出需要抽取的领域。通过归纳式的学习方法生产抽取规则；全自动包装器的生成利用机器学习开发学习算法来学习规则，但是即使是全自动的方法也需要人工专家的参与。

信息抽取中很多任务，例如实体识别，都可以被认为是标注问题(tagging problem)：实体中开始的词被标注为B，实体的延续部分被标注为I，实体之外的词被标注为O。在医学问题实体识别中也可以采用BIO标注的方法，例如对于句子“患者经心超检查后显心影增大”，经过BIO标注法正确标注后应该为“患者/O

经/O 心超/O 检查/O 后/O 显/O 心影/B 增大/I”。给词标注从另一个角度可以看成是分类问题，词周围的词可以作为特征向量。文献[[38]](#endnote-38)使用支持向量机来识别中文文本中的组织机构名。支持向量机(Support Vector Machine，简称SVM)是一个非常优秀的分类器，建立在统计学习理论的VC维理论和结构风险最小原理，在训练数据集中寻找一个高维空间中的超平面，使得不同类别之间的数据间隔最大[[39]](#endnote-39)。但是使用分类器的一个问题是无法利用周围其他词的标注。使用统计模型能够解决这个问题。

隐马尔科夫模型(Hidden Markov Model，简称HMM)是一个基于时序的有向图概率模型。其已经成功应用于连续语音识别和在线手写体识别当中，在生物医学中也得到了广泛的应用。它包括具有状态转移概率的马尔科夫链和输出观测值的随机过程。状态是不可见的，只有通过观测序列的随机过程才能表现出来。



图2-2 HMM模型的示意图

从结构上看HMM包含两层：一个可观察层和一个隐藏层，如图2-2。其中所处的层是隐藏状态序列，所处的层是可观察层。HMM是对隐藏序列和可观察序列的联合分布进行建模：

 (2.8)

其中隐藏状态序列，观察状态序列。对于标注问题来说，对标注之间的依赖做了建模，解决了分类器无法考虑周围词标注的问题。但是HMM缺陷是其遵循两个假设：

1. 马尔科夫假设。隐马尔可夫链在任意时刻的状态只依赖于前一时刻的状态，与其他时刻的状态无关，即只依赖于。
2. 输出独立性假设。任意时刻的观测只依赖于该时刻的隐藏状态，即只依赖于。

马尔科夫假设限制了模型的“视野”，HMM没有考虑上下文信息和远距离依赖问题，即当前信息片段不应该仅仅依赖于前一个或者几个信息片段的标注。输出独立性假设又使得观测状态之间是条件独立的，仅仅依赖于隐藏状态。

最大熵模型(Maximum Entropy Model，简称MEM)是一个条件概率模型，它基于最大熵准则得出。在满足给定约束条件下，未知情况的概率分布要满足无偏见原则，也就是熵最大，简单的描述是从满足约束条件的模型集合中选取熵最大的模型，既满足训练数据的分布，又使得熵最大，如公式(2.9)所示。

 (2.9)

最大熵模型将各种依赖关系简化成特征函数(Feature Function)，用来表示。是一个二值函数，用来表示输出和输出之间的某一个事实，这个事实一般来自于训练数据集。其定义见公式(2.10)。

 (2.10)

根据最大熵的原理，利用特征函数关于经验分布的期望值与特征函数关于模型与经验分布的期望值相等这个约束，多次运用数学方法求极值后，得到的最大熵模型的一般式 (2.11)，其中是归一化因子。它将所有信息都由特征函数和其参数来表示。

 (2.11)

最大熵模型中简化了各种依赖关系的表示，这也就克服了HMM中缺乏远距离和上下文信息的缺陷。但是最大熵模型与HMM不同，它不是一个时序模型，而是一个用于分类的判别模型。在文本标注任务中，不具备时序性使得模型缺乏状态到状态的转移概率，这样观测值能够影响，但是不影响，如图2-3所示。



图2-2 最大熵模型的示意图

最大熵模型中无状态转移概率的缺陷会导致一些不应该出现的状态序列出现，例如词性标注任务中出现了BB（连续两个词首）这样错误的标注。最大熵马尔科夫模型(Maximum Entropy Markov Model，简称MEMM)结合了在最大熵模型中引入了马尔科夫链的概念，在特征函数中加入了状态之间的转移，使得依赖，改善了MEM存在的问题[[40]](#endnote-40)。虽然MEMM效果要好于MEM和HMM模型，但是由于MEMM单独训练每个以状态的概率分布，没有考虑全局的状况，会存在标注偏置问题[[41]](#endnote-41)。

针对MEMM的标注偏置问题，John Lafferty等人提出了一个基于无向图的概率模型，条件随机场(Conditional Random Fields, 简称CRFs)。CRFs既可以较好的表达长距离依赖问题和交叠性特征的能力，能够很好的解决MEMM的标注偏置问题[[42]](#endnote-42)。目前CRFs在文本处理的各种任务中具有很好的表现[[43]](#endnote-43)[[44]](#endnote-44)[[45]](#endnote-45)[[46]](#endnote-46)。文本利用条件随机场抽取病例文本中的病症信息，因此下面具体介绍条件随机场以及其训练算法。

#### 2.2.3 条件随机场模型

##### 2.2.3.1 模型介绍

条件随机场(CRFs)是给定随机变量X条件下，随机变量Y的马尔科夫随机场。它是对条件概率建模的概率无向图模型，其中输出，输入(也成为观测值)。根据概率无向图理论，用概率无向图表示的联合概率分布可以写成图中所有最大团上的势函数的乘积形式，如(2.12)所示。

 (2.12)

其中，是归一化因子，且。最大团(maximal clique)是指无向图中任何两个节点均有边连接的最大结点子集。如图2-3所示。{*Y1,Y2,Y3*}和{*Y1,Y2,Y3*}为最大团，{*Y1,Y2*}和{*Y1,Y2,Y3,Y4*}不是。



图2-3 无向图的最大团

一个条件随机场可以从式(2.12)推导得出，可以写成式(2.13)的形式。

 (2.13)

最终可以得到条件随机场的一般式(2.14)，其中全局归一化因子，势函数结合了最大团内输入和输出随机变量的所有特征：

 (2.14)

对于自然语言文本处理来说，大部分任务中的随机变量序列都是线性链表示的，例如文本观测序列，标注序列。这可以用CRF中一个特殊形式来建模，线性链条件随机场（Linear-chain CRFs）。图2-4显示了线性链条件随机场的无向图表示，其中各个势函数是相应最大团内观测序列和标注序列特征结合。



图2-4 线性链条件随机场

根据图2-3，可以得到线性链条件随机场的一般式(2.15)、(2.16)：

 (2.15)

 (2.16)

式(2.15)中相比最大熵模型(2.11)多了势函数累加项，因为CRFs对序列做了建模，其结合了无向图各个最大团内的所有特征。HMM和MEMM是有向图模型，结点之间依赖是单向的，即依赖于。CRFs是无向图模型，无向图更适合表示复杂的相互作用关系，特别是标注任务中，要确定的标注结果时，依赖于、和。

之前描述的是一阶线性链条件随机场(first-order linear-chain CRFs)，高阶的条件随机场参考论文[[47]](#endnote-47)，还有其他结构更加复杂的条件随机场，根据不同的用于可以使用不同结构条件随机场[[48]](#endnote-48)[[49]](#endnote-49)。

##### 2.2.3.2 训练算法

对于所有的条件随机场，都可以像最大熵模型一样，使用最大似然估计的来对参数进行估计。设训练集，训练过程即使最大化训练集上log似然函数*L*：

** (2.16)

其中第二项高斯平滑因子是为了防止过拟合的惩罚项[[50]](#endnote-50)。对可观察到的特征频数和特征权重的平方之间做了折中平滑。由于式(2.16)是凸函数，将(2.15)式带入到(2.16)式然后对求偏导数最后可以得到式(2.17)，其中是特征经验分布的期望值，是模型分布下的期望值：

 (2.17)

令对数似然函数的梯度(2.17)为0，即可求解使得似然函数最大化的参数，但是这样并不总是可行的。参数选择过程中需要使用到一些数值迭代技术。目前的一些关于CRFs的参数估计训练算法主要有改进的迭代缩放算法(Improved Iterative Scaling，IIS)[[51]](#endnote-51)，归纳的迭代缩放算法(Generalised Iterative Scaling，GIS)[[52]](#endnote-52)，序列条件归纳迭代缩放算法([Sequential conditional generalized iterative scaling](http://dl.acm.org/citation.cfm?id=1073086)

，SCGIS)[[53]](#endnote-53)，LBFGS算法[[54]](#endnote-54)等，算法的具体流程参考相关论文，这里不再赘述。

##### 2.2.3.2 推理与解码算法

在文本处理任务中，例如本文要研究的信息抽取，实际上是一个标注任务[[55]](#endnote-55)，条件随机场的一个重要用途就是推理预测，即对于给定观测序列找出最可能的标注序列。与HMM模型一样，在已知模型参数的情况下，条件随机场的对于最佳标注序列的预测也是使用维特比算法解码(Viterbi Decoding)得到[[56]](#endnote-56)。维特比算法是著名的动态规划算法(dynamic programming)，用于寻找最优路径问题（概率最大路径）。在标注任务中，即寻找使得条件概率最大的标注序列：

 (2.18)

带入条件随机场一般式(2.15)，得到：

 (2.19)

根据动态规划原理，最优路径具有这样的特性：如果最优路径在*t*时刻经过某一结点，那么对于其从开始到在*t*时刻的部分路径来说，必须是最优的。因为如果不是这样，那么可以选择从开始到*t*时刻的所有路径中的那条最优路径然后将*t+1*到*T*的部分路径连接起来就会形成另一条更好的最有路径，这与之前所假设的最有路径是矛盾的。根据这一原理，我们可以从开始时刻，递推地计算在时刻*t*的各条部分路径的最大概率，直至得到时刻*t=T*的最大概率，最优路径的终结点标注也同时得到。之后，为了找出最优路径的各个节点标注，从终结点标注开始，由后向前逐步求得，最终得到最优路径，即为最优标注序列。

具体过程是，先设有部分概率表示为第*j*个位置取得标注为*s*的部分最优路径概率：

 (2.20)

其中，*S*是标注集。然后递推得到：

 (2.21)

开设数组用于记录*j*位置上标注为*s*的部分最优路径的前一个标注，即*j-1*位置上的标注。算法具体工作如下：

1. 初始化：

设第一个位置的前一个标注为观察序列的开头，标注为特殊符号，那么初始化所有以开始的变量：

 (2.22)

1. 递归：

根据式递推式(2.21)不断计算下一个位置的最优路径概率，并记录：

 (2.23)

1. 终止：

 (2.24)

1. 路径回溯：

根据从后往前找出最优标注序列的所有标注：

 (2.25)

#### 2.2.4 信息抽取评价方法

文本信息抽取研究开始于20世纪60年代，由于一系列有关会议，例如消息理解会议(Message Understanding Conference，MUC)、文本检索会议(Text Retrieval Conference，TREC)、自动内容抽取(Atomatic Content Extraction ，ACE)等评测会议的展开而获得很大的发展，并且已经取得了广泛的应用。在这些会议中，衡量信息抽取系统性能的主要依据是两个评价指标：召回率(recall)和准确率(precision)。

召回率是系统正确抽取的信息占所有可能的正确抽取的信息的比例；准确率是系统抽取的所有信息中正确抽取的信息所占的比例。

表2-1 召回率和准确率

|  |  |  |
| --- | --- | --- |
|  | 正确的 | 不正确的 |
| 抽取到的 | *tp* | *fp* |
| 未抽取到的 | *fn* | *tn* |

如表2-1所示，*tp*代表被系统抽取到的正确的信息，*fp*表示被系统抽取到的不正确信息，*fp*代表系统未抽取到的正确信息，*tn*代表未抽取的不正确信息。定义recall和precision如下：

 (2.26)

为了综合评价系统的性能，通常还是用召回率和准确率的加权几何平均值，即F指数来评价：

 (2.27)

其中，是召回率和准确率的相对权重。它可以调整召回率和准确率之间的比重。当等于1时，两者同样重要；小于1时，准确率更加重要一些； 大于1时，召回率更加重要一些。

在文本所研究的工作中，也使用召回率和准确率来评价算法系统的性能。

### 2.3 文本信息检索

文本信息检索是信息检索(Information Retrieval，IR)的一个子领域，通常指将文本信息资源按照某种特定的方式组织起来，并按照用户的需求找出相关信息的技术和过程。文本信息检索技术在许多领域有了非常广泛的应用。例如互联网的迅速发展与信息检索技术的普及紧密相关，为了在海量的网页中搜寻相关信息，搜索引擎技术应运而生，久而久之，有些用户都开始将信息检索等同于搜索引擎了。对于特定领域的也有很多计算机信息检索系统，例如企业信息检索，期刊文献信息检索。

信息检索主要有三种：全文检索、数据检索和语义检索。全文检索是把用户提交的查询请求以关键词的语法形式来与全文内容进行匹配，搜索引擎大部分都是这种方式。这种检索方式简单但是会返回很多与用户无关的信息。数据检索是针对结构化的信息进行检索，我国大部分论文数据库采用这种方式。这种方式需要人工标注资源，检索结果的优劣基本上决定于人工标注方法的优劣。语义检索是利用自然语言处理、人工智能等技术，结合语义网络和基于本体的方法分析出查询语句的语义，通过计算语义相似度的方法检索信息。语义检索弥补了前两种方法的局限性[[57]](#endnote-57)。语义检索受到语义信息处理发展的限制，而语义资源例如语义网络对于语义处理来说至关重要[[58]](#endnote-58)。在普通领域存在着一些著名的语义词典，例如由[普林斯顿大学](http://zh.wikipedia.org/wiki/%E6%99%AE%E6%9E%97%E6%96%AF%E9%A1%BF%E5%A4%A7%E5%AD%A6)大学整理的WordNet[[59]](#endnote-59)，国内则有知网整理的汉语英语双语知识库HowNet[[60]](#endnote-60)。但是在医学领域，中文医学语义词典并不多，典型的作为CULMS的组成部分之一的医学语义网络正在完善当中，而且并不免费开放。

文本使用信息检索领域目前使用最广泛的空间向量模型来计算病例文本数据之间的相似度。

#### 2.3.1 空间向量模型

向量空间模型(Vector Space Model, VSM)由G. Salton等人于上世纪60年代末提出，并成功应用于著名的SMART系统，是目前最为成熟和应用最为广泛的文本表示模型之一。在文本分类、自动索引、信息检索等许多领域取得了较好的应用成果[[61]](#endnote-61)。

向量空间模型假设一个文本所表达的内容与某些特定的特征项在该文本中出现的频数有关，而与这些特征项在文本中出现的位置和顺序无关。也就是说，一个文本所蕴含的内容在向量空间模型中被认为是可以用其包含的特征项在文档中出现频数来表示的。以下给出向量空间模型中的几个相关概念：

1. 文挡（Document）

文档是人类用自然语言组织的文本，可以是由文字构成文章或者片段。在文本中是病例文本。

1. 特征项（Term）

文本所标内容由它包含的基本语言单位组成，则用于构成特征向量空间可以使用这些文本语言单位。在文本中使用的是词。

1. 特征项权重（Term weight）

对于文本中某个特征项赋予的权重，表示该特征对于文本内容的重要程度，权重越高的特征越能反应文本的内容特征。

向量空间模型将文本建模为一个n维特征项向量：

 (2.27)

向量的每一维表示一个独立的特征项，以特征项的权重作为分量值：

 (2.28)

其中是文档document的特征向量。图2-5是向量空间模型对文本建模的示意图。



图2-5 空间向量模型对文本建模示意图

在空间向量模型中，文本之间的距离是以向量的距离来体现出来的，向量的距离计算依赖于特征权重的确定。

#### 2.3.2 特征权重计算

特征权重的计算影响文本相似度的计算。向量空间模型中最常用的权重计算方法就是基于统计的TF-IDF方法，这是由Salton等人提出的[[62]](#endnote-62)。它由两部分组成：特征项频率（Term Frequency，TF）和逆文档频率（Inverse Documentation Frequency，IDF）。对于特征词，其对应某一文档*D*的权重计算公式如下：

 (2.29)

其中表示特征词在文档*D*中出现的频率，*N*是文档集合中的文本总数，是在文本集合中出现过特征词的文本数。TF权值体现了出现次数多的特征词的重要性，而IDF则是对那些在多数文档中出现的词进行了权重的降低，因为这些词区分能力弱。TF-IDF权重计算公式只是一种经验公式,并没有严格的理论基础作为支捧。但是该算法简单易行,在实际应用中往往有较好的效果。

通常各个文本之间长度不同，缺乏可比性，需要对特征向量进行归一化处理：

 (2.30)

#### 2.3.2 文本相似度衡量

在得到文本的特征向量后，就可以通过计算向量之间的距离来衡量两个文本的相似度。目前主流的计算向量距离方法是计算向量的内积或者内积的某种相关系数，这几种方法简单介绍如下：

1. 内积：

 (2.31)

1. Jaccard系数：

 (2.32)

1. Dice系数：

 (2.33)

1. 余弦系数：

 (2.34)

其中最常用的是余弦系数，用两个向量的夹角来衡量文本的相似度。*S*越大，夹角越小，相似度越高。

### 2.4 本章小结

本章对本文涉及到基本理论知识进行了梳理。第一节对文本处理中用到的基本信息论知识做了介绍，包括熵和互信息等，然后给出了文本中发现词边界的常用无监督学习评价依据上下文熵。第二节对文本信息抽取算法做了介绍，并对比了隐马尔可夫模型、最大熵模型、条件随机场等几个常用的模型。第三节介绍了文本信息检索，给出了文本相似度衡量的方法。这几节为后续的章节做了理论上的铺垫。

## 第3章 基于openEHR的电子病历管理系统构建

### 3.1 系统需求分析

#### 3.1.1 目前电子健康档案系统的不足

#### 3.1.2 openEHR标准概念

### 3.2 系统总体设计

### 3.3 系统功能设计及工作流程

### 3.4 数据库设计

### 3.5 本章小结

## 第4章 文本处理技术在病例循证子系统中的应用研究

第三章整体介绍了电子病历管理系统的总体系统结构，在软件、数据存储、数据标准化层面给出了解决方案。本章注点其中智能分析系统的一个子模块——病例循证系统的设计、构建与实现以及文本处理技术在处理文本病例数据时的具体应用和实现过程。

### 4.1 文本处理与病例循证处理流程介绍

病例循证的主要任务是帮助医疗机构有效利用历史病例数据作为之后医师诊断过程中的参考证据，起到辅助作用。医师能够根据待确诊病人的临床症状描述，从系统中检索出最有可能的已经做出确诊的历史病例，这些病例信息是曾经临床案例的总结，往往详细记录了病人的描述症状，确诊信息以及详细治疗措施和治疗过程，结合这些数据，医师可以做出更加正确和客观的决策。

病例循证的过程类似于检索，但是一般的检索系统统往往是过滤少量无用信息之后对全文做索引，例如过滤掉大部分停止词。但是对于病例循证，这种信息过滤并没有做到最大化区分有用和无用的信息。例如对于病例中出现的如下描述“病人于2010年11月12日来我院检查，心超显示心房颤动”，对于病例循证，最重要的是“心房颤动”这个强特征信息，而对于“病人”、“2010年11月12日”、“检查”等这些词都是弱特征信息，甚至是无用的信息。在信息检索的向量空间模型中，由于特征词都是独立平等的，这些弱特征、无用信息会给检索系统带来噪声，影响检索结果的正确性。为了去除这些无用的信息，文本在构建病例循证系统时，在病例信息检索模块前加入了一个病症信息抽取模块，利用文本处理和自然语言处理技术，从病例文本中提取出最有对比性的信息片段作为后续信息检索模块的输入，相当于去掉了文本中对于病例相似度排序无用的信息。整个病例循证过程如图4-1所示，



图4-1 病例循证处理流程

本文要处理的数据是浙江省某医疗合作公司提供的1131份病人病例数据，来自于本课题的基础平台，即第三章所述的电子病历管理系统。数据是文本描述的半结构化信息。取其中一份病例数据的部分诊断信息如图4-2所示。

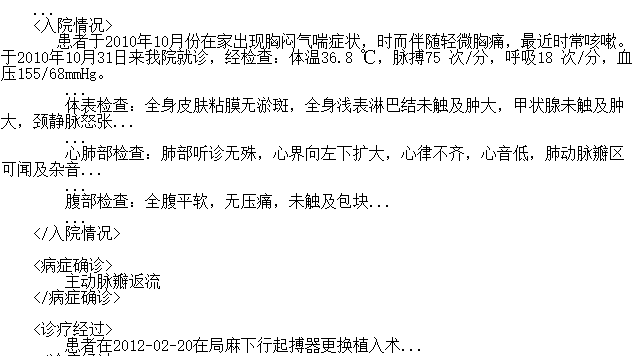


图4-2 病例半结构化数据片段示意图

文本利用条件随机场（CRFs）来抽取病例文本中的病症信息，信息抽取过程便是对观察序列进行标注，得到一个标注序列，其中是病例文本中的句子，如=(心 界 向 左 下 扩 大 ，心 律 不 齐)，利用第三章提到的BIO标注法，则正确标注的=(O-B O-I O O O S-B S-I O O-B O-I S-B S-I)，其中O-B表示病症对象的开始，O-I标志病症对象的延续，S-B表示症状描述的开始，S-I表示症状描述的延续，O表示非病症信息。CRFs属于有监督学习模型，即模型根据给定标注过的训练数据集学习文本中的特征信息，例如文本的字，字组成的词，语言符号，词性甚至是句法结构，语义等各个层面上的特征。本课题中为了方便医师对病例数据进行标注，在病例管理系统中专门开发了一个基于Web的标注系统，如图4-2所示。



图4-3 基于Web的病症信息标注

医师对病例文本仅仅做了做了简单病症信息标注，即病症对象（object）和症状描述（status），并没有像自然语言语料库一样对文本做全面系统的标注，例如分词、词性、句法结构等都是没有的。缺乏足够的上下文特征会造成抽取模型性能很差。本文利用收集到的医学语料，通过筛选和扩充词典，分析医学词语和病例数据的特点，在原标注数据的基础上构建出包含更多上下文特征信息的数据作为最终的训练数据训练CRFs模型，使得CRFs模型能够学习到足够的信息用于病症信息抽取。最后还对比各种特征集的效果，并做了分析和总结。

在病症信息抽取后，病例检索模块对病症信息进行倒排索引建立，用于构建后续用于文档相似度计算的空间向量模型。后续利用病症中语义信息对空间向量模型做了改进，在两个不同的层面上分别利用空间向量模型进行相似度计算，然后通过融合计算得到最终相似度，最后进行实验对比和总结。

### 4.2 医学病例文本数据特点分析与预处理

病例数据中入院情况描述片段是我们用来做病症信息提取的主要文本对象。可以看到文本描述具有自然语言的语法结构特征，但存在很多医学领域用语，如“皮肤巩膜”，甚至还有“慢性细菌性痢疾”这样的长医学词语，这些词容易成为词典或者测试语料中都没有的词语，即未登录词（Out-of-Vocabulary，OOV）。如果测试语料中有比较高的未登录词比率将会大大影响分词的精度，黄昌宁等人在Bakeoff-2003的四个语料库上，对未登录词和歧义对分词精度造成的影响做了比较，未登录词要比歧义切分造成的精度失落至少大10倍左右[[63]](#endnote-63)。普通领域OOV一般比例较低，专业领域OOV比例很高，达50%以上也是很正常的。分词精度的下滑会影响后续病症信息的抽取，最终造成病例检索的准确率下滑。例如“胃肠吻合口机能障碍”被切分成“胃肠/吻合/口/机能障碍”，那么病症信息抽取可能会得到病症部位是“口”，症状是“机能障碍”。目前国内尚缺乏完善的中文医学语料资源，利用无监督学习的方式能够从病例数据中发现一些频度比较高的词汇，从网络上也能够收集到一些医学分类词表，例如ICD-10疾病词汇、搜狗医学词汇、学术名词资讯网中的医学词库。收集整理这些医学词能够一定程度上降低未登录词带来的影响，但是在另一方面，由于词库良莠不齐，文献[[64]](#endnote-64)指出无针对性地添加语料（词表）将会带来词语切分歧义的问题，随着语料增大，模型学习复杂度变大，同时词语切分歧义引起的问题较OOV问题会显得越来越严重。

针对这几个问题本文提出了以下几种方法来减少医学领域未登录词所带来的影响，同时使用语料特征适合度来作为筛选语料的衡量标准，避免为了减少登录词的影响而带来过大词语切分歧义的影响：

1. 从网络上收集各种医学词库，同时利用第二章介绍的边界熵的方法从原始病历数据集中通过无监督学习方法找出使用率较高的医学词汇。计算各个语料之间的特征适合度，筛选出最终的语料作为本文的分类词表。
2. 分析和利用词表语料中医学词的内在构词信息，提取构词子串并对词库扩充，使用基于子串的标注单元让CRFs模型学习到医学构词信息，使模型能够抽取包含未登录词的病症信息，提升信息抽取的能力。

文本中还存在着数字，特殊字符（如摄氏度符号℃），字母（如单位）等非汉字字符。为了减少词形数量，本文先对病例数据做预处理，将非汉字字符映射到特定类别上，如数字0-9映射成NUMBER类别。如表4-1所示。

表4-1 类别映射

|  |  |  |
| --- | --- | --- |
| 类别 | 原字符 | 举例 |
| NUMBER | 一个或者多个数字 | 2010 165 |
| CHAR\_SEQUENCE | 一个字母或者字母序列 | mmHg |
| NON\_CHAR | 特殊字符 | ℃,β |
| PUNCTUATION | 标点符号 | 。，、 |

### 4.3 病症信息抽取

#### 4.3.1 医学语料收集和筛选

病例文本数据本身具有语言的特点。Harris等人于1955年研究并给出了一个语言学假设：文字串后面的字符不确定性可以用来指示词的边界[[65]](#endnote-65)。Tanaka-Ishii等人使用边界熵(Branching Entropy，BE)具体描述了这个语言学假设：语言的字符序列的边界熵随着字符长度的增加会不断减少，即变化边界熵 (Variation of the Branching Entropy，VBE)应该是负的，如果出现BE增长的情况，即VBE>0，则这个位置很可能是词边界[[66]](#endnote-66)。使用BE和VBE可以无监督学习出文本中的词，但是这种方法需要人工指定判断的阈值且不同长度的词具有不同的阈值，分割词的时候又容易出现局部最优。在VBE基础上，Pierre Magistry提出了归一化变化边界熵(Normalized Variation of the Branching Entropy，nVBE)，使得不同长度的词的nVBE可以比较[[67]](#endnote-67)。

使用nVBE可以从原始病例数据中学习出一些医学领域词汇，这可以在一定程度上解决医学预料不足的问题。

在第2章给出了边界熵的定义，对于一个字符串，其左边界熵为，右边界熵为。定义左变化边界熵和右变化边界熵：

 (4.1)

归一化变化边界熵则定义为，其中是长度为*k*的字符串的*VBE*平均值，这里的长度为*k*：

 (4.2)

结合左右归一化变化边界熵作为最后字符串的得分：

 (4.3)

如果直接使用式(4.3)遍历病例文本数据，对其中所有的n元文法打(*n=1,2,3..,k*，*k*为指定最大长度)，然后按照分数从高到低筛选出可能的词会由于局部最优而带来大量的歧义切分问题。因为式(4.3)仅仅考虑了各个n元文法独立成词的最大可能性，并没有综合考虑句子中其他词汇成词可能性。例如“主动脉瓣反流”、“ 三尖瓣反流”，如果不考虑句子全局分词情况，最后得到的n元文法中会有很多类似于“反流”这种独立成词的词，甚至会出现“瓣反流”这种错误的词。文本先从全局最优的角度对病例文本无监督分词，然后从最终的分词结果中将各个n元文法按得分排序后再筛选出来作为最后的词。句子最优分词序列通过计算式(4.4)得到，其中是所有可能的分词序列集合，*W*是中某个分词序列包含的所有词的集合：

 (4.4)

对预处理后的病例文本数据计算(4.3)和(4.4)，最后按照词的得分排序如表4-2所示。

表4-2 病例文本数据中各个文法的得分和筛选

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | N元文法 |  | 是否筛选 |
| 1 | 的 | 30.66 | 否 |
| 2 | 在 | 30.21 | 否 |
| … | … | … | … |
| 54 | NUMBER年NUMBER月 | 20.78 | 否 |
| … | … | … | … |
| 83 | 胸 |  | 是 |
| 84 | 左心室 |  | 是 |
| 85 | 淋巴结 |  | 是 |
| 86 | 房室传导阻塞 |  | 否 |
| 87 | 三尖瓣 |  | 是 |
| … | … | … | … |
| 95 | 性心肌炎 |  | 否 |
| … | … | … | … |

表中前部分有很多助词，介词，连词等，这些词得分很高是因为普遍使用，并且左右边界熵很大，被认为成词可能性很大；之后由于预处理时候做了类别映射，像“2010年12月”这种词串变成了“NUMBER年NUBMER月”，由于词形减少，词次增多，也被认为成词可能性很大；之后就是根据文本内容而学习得到的医学词，由于本文中使用的数据大部分关于心脏内科的诊断病例，因此其中有关心脏疾病的医学词汇使用频率高，这部分词容易被发现；表中还有一些误认为是词的词串，例如“性心肌炎”是由于受到多个构词方式类似的医学疾病名词的影响，例如“病毒性心肌炎”、“感染性心肌炎”、“细菌性心肌炎”等“AA性BBB”形的医学词造成“性BBB”被误认为是一个词。

本文通过规则匹配和人工筛选相结合的方法来筛选出最后的词：

1. 通过对比通用领域词表（如国家语委现代汉语语料库）中的词和词性来排除，例如“的”，“在”，“是”这些词将会被排除。
2. 指定一些筛选规则模板，通过正则表达式匹配来排除一些以医学词语组成词开头的词，如“性XXX”、“样XXX”、“状XXX”等。

除了从原病例文本中抽取医学词语外，本文还从网络上收集一些医学词表库，本文主要收集了ICD-10疾病词汇、搜狗医学词汇、学术名词资讯网中的分类医学词库。包括普通医学、解剖学、内分泌学、精神病、内科学、临床医学、病理学、药学医学词汇。4.2节已经分析了不同语料之间有差异，并不是语料越大越好，需要针对性地添加语料，否则会带来词语切分歧义的问题。语料不仅存在量的问题还存在着质的问题，文献64在研究专业领域分词任务时提出了一种衡量语料特征适合度的方法用于在缺乏专业语料的情况下选择合适的训练集，其实验显示训练集和测试集的特征适合度越大，机器学习的效果越好。本文利用特征适合度来衡量收集到的语料的质量，筛选出适合度相对高的语料作为文本后续处理的词库。

元素特征适合度的思想类似于向量空间模型。设有两个语料和，语料中所有元素集合的某一个元素记为，元素的输出特征集合中某一个特征记为，那么两个语料关于元素的特征适合度：

 (4.5)

其中是元素在语料中以特征出现的概率。每个元素在文本中出现的频率不同，将基本元素在各个语料中出现概率的平均值作为每个基本元素适合度的权重。则语料和的关于元素集合的适合度为：

 (4.6)

其中为元素在语料中出现的概率。

基本元素和输出特征集的选择依赖于不同的任务，例如词性标注任务中，元素即是“词”，输出特征即是“词性”。在本文研究的病症信息抽取任务中，信息抽取的性能依赖于两点：词语识别的性能和病症信息识别的性能。正确识别词语是信息抽取的前提，本文收集的语料主要是医学领域词汇，语料之间的适合度体现在词语的构词特征上。因此文本以“字”为基本元素，“字位置”为输出特征，通过计算语料关于“字”-“位置”的特征适合度来衡量医学词表之间的适合度。“字位置”使用中文分词标注集中常使用的6分类标注（6-Tag标注集）来表示：B为词首，B2为词中的第二个字，B3为词中的第三个字，M为词中居于B3后且不是词尾的字，E为词尾，S为单字词。例如“三尖瓣闭锁不全”标注为“三/B尖/B2瓣/B3闭/M锁/M不/M全/E”。本文之前通过无监督学习的方式从原病历数据中抽取了部分词汇，称为原病例词表。利用式(4.6)计算原病例词表和网络上收集到的各医学领域词表之间的特征适合度，显示在表4-3中。

表4-3 病例文本数据中各个文法的得分和筛选

|  |  |
| --- | --- |
| 医学词表 | 特征适合度 (与原病例词表) |
| 普通医学 | 0.72 |
| 解剖学 | 0.48 |
| 内分泌学 | 0.32 |
| 精神病 | 0.29 |
| 内科学 | 0.76 |
| 临床医学 | 0.61 |
| 病理学 | 0.47 |
| 药学 | 0.12 |

由于病例文本描述中大量使用的是临床医学和普通医学的领域词语，从表4-3中可以看到特征适合度基本上能够反应语料的适用程度。文本将选用普通医学、解剖学、临床医学、内科学词库和之前整理的原病例词表作为最终的语料。

#### 4.3.2 医学词特征分析

文本之前通过收集语料扩充词典，可以在一定程度上解决语料不足的问题，降低OOV率，但是不可能像通用领域一样达到高词典覆盖率，尤其在医学专业领域，专业词汇数量巨大且构词丰富，高OOV率已成必然，仅仅基于词典来处理自然语言问题很难得到好的效果。通常在处理自然语言的问题时都会结合各种方法，综合各种模型的优点来提高处理效果。岳中原使用基于机械匹配与基于统计相结合的中文分词算法，解决了部分条件下的未登录词和歧义引起的切分问题，然后又使用统计和词典相结合的方法对之前结果进行改进[[68]](#endnote-68)。王琦在生物命名实体识别中使用一种基于词典和机器学习相结合的方法，在JNLPBA2004语料上实验取得了较好的结果[[69]](#endnote-69)。

与通用领域词表相比，专业领域词表包含着一定的构词信息。医学领域的词表构词信息也很明显，如“三尖瓣闭锁不全”、“三尖瓣返流”都以部位“三尖瓣”开头；“病毒性心肌炎”、“感染性心肌炎”都以心肌炎作为结尾，以“性”作为两种病症的结合词。这些丰富的构词信息在普通领域很少，但是在医学领域很常见，这也是导致医学领域词汇高OOV率的主要因素。文本利用统计的方法挖掘出医学词表中的构词信息，并以子串的形式重构医学词典，让之后的CRFs信息抽取模型能够从中学习到医学构词信息，提高对OOV的识别能力，进一步提升病症信息抽取的性能。

#### 4.3.4 基于构词子串的标注单元

第二章已经提到信息抽取任务可以看成标注任务，标注可以基于不同的单元。GC Ling等人使用基于字的标注和分块方法来识别中文未登录词[[70]](#endnote-70)，Hwee Tou Ng等人对基于词和基于字的中文词性标注方法做了对比[[71]](#endnote-71)。李莹在对文本病例抽取医学问题的时候为了避免分词错误影响问题抽取，采用了字作为标注单元[[72]](#endnote-72)。这种方法只不过将分词和信息抽取合并为一个过程，并没有减少分词错误带来的影响。基于字的标注在后续应用CRFs信息抽取模型时出现很多词表中已经有的词，但在CRFs信息抽取时被切分开了。对于医学领域词表中的词，不需要CRFs重新生成。同时为了利用医学词表中丰富的构词信息，本文使用词的构词子串作为最小标注单元，构词子串可以是词缀，词的组成词等，而子串之间的组合蕴含构词信息。相当于在CRFs信息抽取之前，强化了标注单元和构词信息，使得CRFs尽量避免错误的切词，同时又能通过学习构词信息识别未登录词，最终使得信息抽取性能提高。本文利用无监督学习方法提取医学词表中的构词子串，并以此对词库做了扩展，最终词典中是所有词和词串的合并，本文将其统称为“构词子串”或者“子串”，所有的子串都赋予类型信息，具体类别如下：

1. 独立词：

独立成词，几乎不用于构成其他医学词，这类词归类为独立词（Independent Word，IW）。

1. 组成词：

医学领域词表中的词经常利用其他词来构成。例如“急性淋巴细胞性白血病”可以由“急性”+“淋巴细胞”+“性”+“白血病”构成。其中“白血病”在词典中存在，且常常用在构成其他医学词的词尾，则将这一类词设为“后缀词”（Suffix Word，SW），相应的还有“前缀词”（ Prefix Word，PW）。

本文使用如下公式来找出词典中所有的词缀，其中是词出现的频数，*a*和*b*是判定阈值，一般根据词库特点使用经验值，在本文的医学词库中使用0.75：

 (4.6)

如果词出现在医学词典中且能用于构成其他医学词，但是不是SW或者PW，则归类为普通组成词（Component Word，CW）。

1. 组成串：

除了组成词，还有很多几乎不用做单独词，但是常常用作其他词的前后缀，例如“病毒性”，“先天性”。这类词串设为前缀字符串（Prefix String，PS）或者后缀字符串（Suffix String，SS）。这类子串的边界熵要比其他字符串边界熵大，使用式(4.7)就可以找出所有的前缀和后缀字符串。

 (4.7)

其中是长度为k的前(后)缀字符串的平均边界熵。*a*和*b*为经验系数，在本文中为1.2。

1. 字：

在医学词汇中经常有类似“样”、“状”、“性”等结合字，如“淋球菌性心肌炎”、“冠状动脉狭窄”。这些字具有强烈的构词特征，很多医学未登录词都使用这些字连接两个医学词来构成新词。因此本文将这些词归类出来成一类，称为结合字（Joint Character，JC），文本使用左右边界熵的和来给词典中的各个字打分，然后从达到小排序，筛选出分数最大的那些字。其他的不成词但又不是结合字的字归为其他字（Other Character，OC）。

1. 通用领域词：

这些词是医学领域词表和构词信息中都不存在的，但是在通用领域词表中出现的词，将其归类为通用词（Gerneral Word，GW）。

以这些构词信息为词典对医学词应用正向最大匹配分词方法(Forward Maximum Matching, FMM)[[73]](#endnote-73)，将医学词拆分成构词子串的组合，最后合并医学词和构词信息成为最终的词典，如表4-4所示。

表4-4 加入构词信息后的词典示例

|  |  |  |
| --- | --- | --- |
| 医学词 | 子串切分 | 子串类型组合 |
| 状 | 无 | JC |
| 风湿性主动脉瓣狭窄 | 风湿性/主动脉/瓣/狭窄 | PS CW OC SW |
| 风湿性 | 无 | PS |
| 三尖瓣 | 无 | PW |
| 三尖瓣返流 | 三尖瓣/返流 | PW CW |
| 带状疱疹性脑膜炎 | 带/状/疱疹/性/脑膜炎 | OC JC CW JC SW |

#### 4.3.5 训练数据生成

CRFs信息抽取模型最后是根据已标注数据对未标注数据进行处理。文本所处理的病例数据仅仅只有医师标注的病症对象和症状描述信息。为了构建出更多的上下文特征信息，基于4.3.4节整理得到的带构词信息的词典，利用FMM分词法对标注数据进行分词得到子串序列，同时也得到了子串类型序列。例如对于句子 “肺部听诊无殊，心界向左下扩大，心律不齐，心音低，肺动脉瓣区可闻及杂音”的处理结果为“肺部/PW/O-B听诊/IW/O无殊/GW/S-B，心界/CW/O-B向/GW/O左/GW/O下/GW/O扩大/SW/S-B，心律/CW/O-B不齐/SW/S-B，心音/CW/O-B低/GW/S-B，肺动脉/CW/O-B瓣/OC/O-I区/OC/O-I可/GW/O闻及/GW/O杂音/GW/S-B”，其中前一项标注是子串类型，后一项标注是病症对象和症状描述。

FMM属于基于词典的机械分词，不能利用上下文信息，不能处理OOV问题，这里使用FMM处理语料是为了给后续CRFs模型生成学习语料。语料的好坏决定了后续CRFs模型抽取模型的性能。这里需要从语料中筛选出那些正确性比较大的句子作为最终的训练语料。由于词典中的医学词汇以及构词信息大部分是正确的，因此本文中训练语料的筛选规则是句子对医学词典的覆盖率，如果句子中存在多个与医学词典匹配的词，则该句子放入训练集中。

#### 4.3.6 CRFs特征选择

由第二章可知，条件随机场（CRFs）使用特征函数简化了各种依赖关系的表示，这使得我们非常容易地将观察序列中的任意特征加入到模型中，例如传统HMM模型中的状态转移特征和状态特征，还可以是语言本身的特征，如词性、构词规则、语义、语法结构等。特征一般在上下文窗口内选取，所谓上下文窗口，是指包括当前标注单元及前后若干个标注单元组成的标注串。窗口越大，能利用的上下文越多，但是模型的训练效率就会降低；窗口过小，可选的特征就不够多，影响模型的效果。特征可以是单一特征，也可以是多种特征的组合。窗口的大小和特征的组合选择都需要根据数据集来确定，选择冗余少、有代表性的特征能够很好反映出目标信息的上下文特征，有助于识别目标信息，同时让模型收敛得更快；差的特征可能在训练集当中也没出现过几次，同时特征组合又很多，特征空间很大，数据稀疏问题严重，模型收敛慢，识别精度也不高。

在训练数据集中，对目标信息的长度进行统计，结果显示90%的目标信息是8个字以下，而60%的构词子串是4个字以下，90%的构词子串是6个字以下，因此取当前标注单元前后各1-2个标注单元作为上下文特征是足够的。

本文中标注单元是构词子串，子串本身以及构词类型这两个特征能够反映文本中医学领域用语的特征和成词特征。子串性质是构词子串的医学范畴所属或者普通词的词性，在本文中使用的是普通医学、解剖学、临床医学、内科学词库和原病例词表五个词表，从不同的词表中提取出来的构词子串或者词具有不同的所属性质，例如解剖学词典主要是人体器官组织等词汇，那么其中的“扁桃体”的性质就是A（A代表解剖学词汇），子串性质蕴含了一定程度上的医学语义信息，能够帮助识别是病症对象还是症状描述。在病例文本中还有一些通用领域的词汇，例如 “的”、“了”等助词，“在”、“对”等介词，“患者”、“检查”等通用领域词汇，这些词的用法在医学领域和通用领域几乎一样，因此本文为这些词赋予通用领域最常用的词性作为相应的子串性质。表4-4显示了子串性质的归类。最后本文还加入了指示词这种特征，对病例文本进行统计发现，病症对象和症状描述等目标信息容易出现在一些特定词之后，例如“出现”、“伴随”、“致”、“诊断为”等。本文更加病例文本收集和整理了28个词作为指示词。

表4-4 子串性质归类

|  |  |  |  |
| --- | --- | --- | --- |
|  | 性质 | 标记 | 举例 |
| 医学领域 | 普通医学(general) | G | 手术 |
| 解剖学(anatomy) | A | 主动脉窦 |
| 内科学(internal) | I | 室性 |
| 临床医学(clinical) | C | 急诊 |
| 原病例词表(original) | O | 心律 |
| 通用领域 | 助词 | U | 的，了 |
| 介词 | P | 对，于，在 |
| 名词 | N | 患者 |
| … | … | … |

综上所述，文本经过分析最后选用的特征是子串、子串构词类型、子串性质和指示词。

#### 4.3.7 基于CRFs的信息抽取与实验测试

CRF++是著名的条件随机场开源工具，也是目前综合性能最佳的CRF工具[[74]](#endnote-74)。本节使用CRF++工具来训练模型，CRFs模型在特征集上进行训练，从训练集中抽取特征使用的是特征模版，特征模版定义了选取特征样本的模式。如图4-4所示。



表4-4 特征模板从训练集中提取特征集

将子串、子串构词类型、子串性质和指示词4种特征归纳成特征模版，根据病例数据特点选择合适的特征类型。表4-5显示了本文实验中使用的几个模版。例如T2模板使用了构词子串、构词类型、子串性质三种特征以及其不同上下文形式的组合。其中T1模板仅仅使用了字和词性的特征，这是为了后续实验做对比用。

表4-5 CRFs使用的特征模板T1、T2、T3、T4

|  |  |  |
| --- | --- | --- |
| 所在模板 | 特征类型 | 特征 |
| T1 | 字 | Cn(n=-1,0,1) |
| T1 | 双字邻接 | CnCn+1(n=-1,0) |
| T1 | 字跳跃 | C-1C1 |
| T2,T3,T4 | 子串 | Sn(n=-1,0,1) |
| T2 | 子串邻接 | SnSn+1(n=-1,0) |
| T2,T3, T4 | 构词类型邻接 | TnTn+1(n=-1,0) |
| T2,T3, T4 | 构词类型跳跃 | T-1T1 |
| T1,T2,T3, T4 | 子串性质(或词性) | Pn(n=-1,0,1) |
| T1,T2,T3, T4 | 子串性质(或词性)邻接 | PnPn+1(n=-1,0) |
| T4 | 指示词特征 | In(n=-2,-1) |

抽取模型的性能评价标准使用第2章介绍的召回率（recall）、准确率（precision）和F指数。本论文的实验中，一共有1131份病人病例数据，将数据随机打乱，采用K次交叉验证[[75]](#endnote-75)（文本中K=5）的方法来估计模型在数据集上整体性能。表4-6显示了实验中使用不同特征模板的CRFs信息抽取模型的性能，其中F指数中的，即召回率和准确率同等重要。

表4-6 CRFs使用的特征模板T1、T2、T3、T4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 病症对象(Object) | | | | |
| 特征模板 | Precision | Recall | F | Left-F | Right-F |
| T1 | 0.8426 | 0.7865 | 0.8136 | 0.8578 | 0.8459 |
| T2 | 0.9013 | 0.8378 | 0.8684 | 0.9054 | 0.8702 |
| T3 | 0.8893 | 0.8265 | 0.8568 | 0.9241 | 0.8769 |
| T4 | 0.9141 | 0.8467 | 0.8791 | 0.9365 | 0.8891 |
|  | 症状描述(Status) | | | | |
| 特征模板 | Precision | Recall | F | Left-F | Right-F |
| T1 | 0.7486 | 0.7243 | 0.7362 | 0.7402 | 0.7954 |
| T2 | 0.8567 | 0.8136 | 0.8346 | 0.8394 | 0.8835 |
| T3 | 0.8465 | 0.8287 | 0.8375 | 0.8440 | 0.8912 |
| T4 | 0.8613 | 0.8390 | 0.8500 | 0.8546 | 0.8906 |

其中Left-F和Right-F是指模型标注信息左右边界的性能，F则是正确抽取抽取信息的性能。T1模板使用的是以字为标注单元，T2、T3、T4模板都使用以构词子串为标注单元。从实验数据中可以看到，本文研究的基于构词子串的信息抽取模型要比基于字的效果好。T3模板仅比T2模板少利用了一个子串转移（SnSn+1）的特征，信息抽取的性能并没有下降很多，但是训练时间大大减少，这是因为词形很多，子串转移的特征空间很大，而且由于数据稀疏问题严重，特征的统计估计变得不准确，该特征提供的性能提升也就受到限制。T4在T3的基础上加入了指示词的特征，从实验结果上看，引入这个特征确实可以提升信息抽取的性能。病症对象的Left-F普遍比Right-F要好，而症状描述的Right-F普遍比Left-F要好，这是由于大部分病症对象出现在症状描述的左边，而两者的分界处容易产生歧义，例如“皮肤巩膜黄软”可以标注为“皮肤/Obejct 巩膜黄软/Status”，也可以标注为“皮肤巩膜/Obejct 黄软/Status”。当病症对象和症状描述信息连接在一起，医学词典以及标注训练集的歧义容易造成中间切分点的错误。

### 4.4 病例信息检索

#### 4.4.1 信息检索模型简介

信息检索系统的核心是信息检索模型。孙坦等人研究了近几年国内外信息检索模型的进展。最早的布尔模型由于以布尔表达式为基础具有二值评价文档的局限性已经很少研究。Salton等人于60年代末提出的向量空间模型通过计算文档向量相似度来排序具有广泛的适应性和生命力。Roberston等人于1976年以概率论为基础，提出概率模型用于信息检索，主要用概率相关性来表达相似度，以解决空间向量模型中特征独立性的缺陷，概率模型主要缺点是条件概率值很难估计。

近年来由于受到自然语言处理发展的影响，语言模型也逐步引入到信息检索中，同时本体方法广泛用于处理自然语言语义的问题中，基于本体的信息检索通过对知识和语义上的匹配，具有更好的查全率和查准率，也成为一个新的发展方向。无论是经典的布尔模型、向量空间模型、概率模型，还是随着环境的变化和技术的发展涌现出来的语言模型、基于本体的信息检索模型，都从不同层面丰富了信息检索模型的研究内容[[76]](#endnote-76)。不同的检索模型受到不同的约束，例如今后基于本体的信息检索能够更加智能地从概念语义上搜集信息，但是受到知识获取的约束，知识库的建立仍然是构建强大的人工智能系统的关键。目前本体库建立的并不多，国内外主要有4个通用领域本体库WordNet、DBPedia、Cyc、HowNet，两个专业领域本体库生物医学本体库TMO、UMLS以及一些企业领域本体库，同时还存在着本体库之间标准化程度不够高的问题[[77]](#endnote-77)。在国内生物医学领域，相关生物医学知识库相比国外要更加缺乏，例如CUMLS都还处于建立和完善，并没有对外开放。目前信息检索模型中居于主导地位的依然是传统的向量空间模型[[78]](#endnote-78)，因此文本使用空间向量模型来实现病例信息检索。

#### 4.4.2 信息检索系统的组成

信息检索系统一般包括信息源、信息处理、信息特征组织管理、检索模型、用户接口等几部分组成。系统先对信息源进行处理，将信息的特征按照检索模型的需求组织并管理起来，然后检索模型根据用户的查询与信息源的特征通过模型计算相似度，最终按照相似排序通过用户接口输出。文献[[79]](#endnote-79)给出了基于空间向量模型的系统结构，如图4-5。



图4-5 信息抽取系统结构

其系统中的文档是半结构化的，例如文献，具有关键词和摘要等内容，这些内容精炼地反映整片文档的主要内容。因而其设计的系统中具有类库，即根据关键词和摘要对文档分门别类，用户的查询先经过类库特征比较，找出最相关的类别，然后在最相关类别的文档中应用传统的向量空间模型进行检索。

本文中文档是病例数据经过4.3节信息抽取之后的病症信息。这部分信息精炼反映文档的医学信息，但所用的数据并没有分类信息，因此本文直接对病症信息进行特征项库的建立。特征项使用的是词，特征项的组织是使用倒排索引（Inverted Index）[[80]](#endnote-80)的数据结构来管理。倒排索引也称为反向索引，被用来存储在全文检索下某个词在一个文档或者一组文档中的存储位置的映射。它是文档检索系统中最常用的数据结构，图4-6显示了倒排索引的示意图，其中第一列结点是特征项，在特征项之后的结点表示该特征项出现在相应文档中，该节点一般存储特征项的归一化TF-IDF权重以及在文档中的位置等信息。



图4-6 倒排索引的示意图

#### 4.4.3 病症信息的相似度衡量

利用向量空间模型计算病例相似度需要将所有的文档转换成特征向量。利用4.4.2节中以倒排索引组织的特征项库可以很容易得到该文档的特征权值向量，使用式(2.34)可以计算任意两个文档的相似度。

向量空间模型具有一些改进算法[[81]](#endnote-81)[[82]](#endnote-82)。陈治平等人提出的基于N层向量空间模型的信息检索算法，将一篇文档从划分为N个相对独立的段，然后按照文本段内容建立特征向量和本段的权值向量，最后对所有段相似度求平均值作为最后的相似度，此方法具有较高的查找速度和查准率[[83]](#endnote-83)。但这个方法有一定的局限性，首先文本段具有比较明显的段落分割，例如标题、摘要、正文，其次是每一个段落相似度的重要性并不一样，使用普通平均值没有考虑这种情况。本文实验中的病例文本数据不存在明显的分段特征，但是4.3节提取出病症对象和病症描述具有很大的区分度，病症对象主要描述的是部位、器官、医学对象，例如“淋巴结”、“心律”，而病症描述主要是状态、现象等描述词，如“肿大”，“不齐”。本文借鉴N层向量空间模型的思想，但分层则不是对文本分层，而是对信息类型分层，即病症对象和病症描述分别用向量空间模型计算相似度，然后使用调和平均数融合两个相似度，如式(4.8)所示：

 (4.8)

其中是病症对象的相似度，是病症描述的相似度。调整可以调整两个量之间的重要性。当等于1时，两者同样重要；小于1时，病症对象更加重要一些；大于1时，病症描述更加重要一些。使用调和平均值能尽量压制极端值，例如当病症对象和病症描述同样重要的情况下，，时，，而当，时，，这说明两种信息都有相关性的情况要比一种信息很相关而另外一种信息几乎不相关要有更高的相似度。这与具体实际情况是相符合的，例如两种完全不相关的病例，“心肌劳损”和“心肌梗死”，病症对象完全一样，但病症描述完全不同，使用(4.8)式得到，使用算数平均。

#### 4.4.4 实验测试与总结

本节对病例信息检索做了实验的测试，主要包括以下两个：

1. 基于病症信息和基于全文的病例检索性能比对

实验中本文随机抽取1份标注数据，用其病症对象和症状描述作为检索系统的输入，统计系统返回的前50个病例数据，由医师判断是否为同一类型的病例，如果是则记录正确数，否则记录错误数，同时统计出1131份病例中认为与这份病例为同一类型的病例数，计算准确度和召回率，最后以F值作为评价指标。取10次实验结果的平均值，如表4-7所示。

表4-7 基于病症信息和基于全文的病例检索实验数据

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实验序号 |  | 检索系统 | | |
| 数据源 | Precision | Recall | F |
| 1 | 病症信息 | 0.7840 | 0.7059 | 0.7429 |
| 2 | 病例全文 | 0.7563 | 0.6347 | 0.6902 |

实验1是以病症信息进行检索的，实验3是以病例全文进行检索的，从实验数据可以得出，通过病症信息抽取对病例数据进行医学信息提取确实可以提高病例检索系统的准确率和召回率。本论文研究的方法具有一定的实用价值。

1. 基于分层向量空间模型的检索性能测试

文本借鉴分层向量空间模型的思想，在病症对象和症状描述两个层面计算相似度，最后通过式(4.8)式融合。本文将该方法与普通向量空间模型进行实验对比，实验数据如表4-8。

表4-8 分层和传统向量空间模型的病例检索实验数据

|  |  |  |  |
| --- | --- | --- | --- |
| 数据源 | Precision | Recall | F |
| 分层模型 | 0.8006 | 0.7335 | 0.7656 |
| 传统模型 | 0.7840 | 0.7059 | 0.7429 |

由实验结果可以得到，本文所使用的病症信息分层方法可以提高信息检索的性能。

1. 病症对象和症状描述信息对检索性能的影响比对

式(4.8)融合病症对象和症状描述信息相似度，用来调节相对比重。实验中，本文让从0到2变化，即病症对象相对于症状描述相似度的比重从大到小变化，然后测试检索系统的性能，测试方法和(1)相同。测试数据如表4-9：

表4-9 不同下病例检索的F值

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 实验序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|  | 0 | 0.3 | 0.6 | 0.7 | 0.8 | 1 | 1.3 | 1.5 | 2 |
| F | 0.475 | 0.485 | 0.621 | 0.694 | 0.753 | 0.701 | 0.635 | 0.594 | 0.426 |

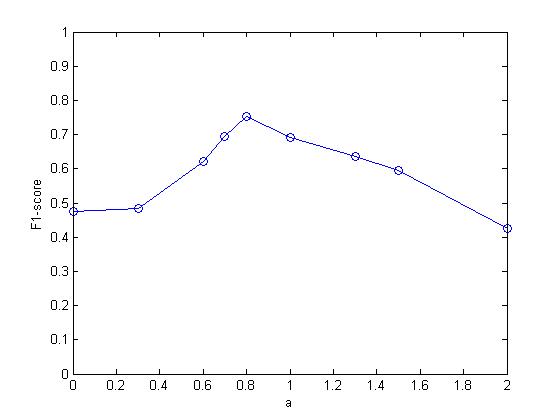


图4-7 F值随着变化的曲线

从实验数据可以得到，在0.7到0.9左右，系统的病例检索效果最好，这说明了病症对象的相似度要比病症信息的相似度值对最终病例检索的性能影响更大，具体分析实验数据发现，如果两份病例属于同一类，那么病症对象由于大部分是医学领域术语，变化不会太大，相同居多，但是病症信息有很多是通用领域词汇，表述不同会导致症状描述相似度降低。因此在融合公式中偏重病症对象相似度能够提高检索效果，这与实验数据是相符合的。

### 4.5 本章小结

本章对第三章系统中的病例循证子系统的做了设计与实现，同时研究了文本处理技术在病症信息抽取和信息检索中的具体应用。在病症信息抽取中利用医学术语构词信息提升了抽取效果，在病例检索中使用分层方法提升了检索效果。最后通过实验验证了本文所研究的方法，同时结合实验数据和结果，给出了总结。

## 第5章 系统测试和分析

### 5.1 测试方法介绍

### 5.2 系统功能测试

### 5.3 本章小结

## 第6章 总结与展望

### 6.1 总结

### 6.2 展望

基于领域本体的语义检索系统的研究与实现\_武俊丽

可以结合语义分析技术，语法处理，句子过滤，主干提取，句型识别，修饰词语义提取，语义信息生成。参考基于浅层语义分析技术的语义检索

## 参考文献

1. http://www.ccwresearch.com.cn/report\_detail.htm?id=3 [↑](#endnote-ref-1)
2. http://news.hc3i.cn/art/201412/32087.htm [↑](#endnote-ref-2)
3. 漫谈\_智慧医疗 [↑](#endnote-ref-3)
4. http://news.hc3i.cn/art/201107/14474.htm [↑](#endnote-ref-4)
5. http://news.hc3i.cn/art/201001/1549.htm [↑](#endnote-ref-5)
6. 文本挖掘技术研究进展\_袁军鹏 [↑](#endnote-ref-6)
7. 基于内容的垃圾邮件过滤研究\_潘文锋 [↑](#endnote-ref-7)
8. 基于贝叶斯分类的中文垃圾邮件过滤方法研究和改进\_胡睿 [↑](#endnote-ref-8)
9. Hot Topic Detection in Local Areas Using Twitter and Wikipedia [↑](#endnote-ref-9)
10. Multi Topic Distribution Model for Topic Discovery in Twitter [↑](#endnote-ref-10)
11. 基于依存句法分析的社会媒体文本挖掘方法 [↑](#endnote-ref-11)
12. Extraction of Drug-disease relations from MEDLINE Abstracts [↑](#endnote-ref-12)
13. Zhihao Yang, Hongfei Lin and al, 2010. BioPPISVMExtractor: A protein–protein interaction extractor for biomedical literature using SVM and rich feature sets. J Biomed Inform.88-96. [↑](#endnote-ref-13)
14. 中文电子病历的信息抽取研究\_李毅 [↑](#endnote-ref-14)
15. http://www.nlm.nih.gov/research/umls/ [↑](#endnote-ref-15)
16. http://en.wikipedia.org/wiki/Unified\_Medical\_Language\_System [↑](#endnote-ref-16)
17. [The unified medical language system (UMLS): integrating biomedical terminology](http://nar.oxfordjournals.org/content/32/suppl_1/D267.short) [↑](#endnote-ref-17)
18. 李丹亚, 胡铁军, 诸文雁等.中国生物医学文献数据库的建设与发展[ M] .数字图书馆的建设与发展, 天津:天津人民出版社, 2001:412 -419 [↑](#endnote-ref-18)
19. 张爱连, 刘春艳, 李丹亚.一体化医学语言系统研究进展[ J] .医学情报工作, 2005, 26(2):81 -84 [↑](#endnote-ref-19)
20. 中文一体化医学语言系统的构建与应用\_李丹亚 [↑](#endnote-ref-20)
21. 循证医学的定义、发展、基础及实践 [↑](#endnote-ref-21)
22. Dunn G, Everitt B. Clinical biostatistics: an introduction to evidence-based medicine. London New York: E. Arnold; Halsted Press; 1995. [↑](#endnote-ref-22)
23. Friedland DJ. [Evidence-based medicine: a framework for clinical practice](http://medical-textbook-deals.info/wp-content/uploads/pdfs/Evidence%20Based%20Medicine%20A%20Framework%20For%20Clinical%20Practice%20by%20Daniel%20J%20Friedland%20-%205%20Star%20Review.pdf). Stamford, Conn. : Appleton & Lange; 1998. [↑](#endnote-ref-23)
24. 基于自然语言处理技术的循证医学信息提取研究\_卢延鑫 [↑](#endnote-ref-24)
25. Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools [↑](#endnote-ref-25)
26. 《健康档案基本架构与数据标准(试行)》. [↑](#endnote-ref-26)
27. 卫生部将在“十二五”期间加快建立卫生信息标准.

    http://www.chinadaily.com.cn/dfpd/2012-03/20/content\_14870883.htm [↑](#endnote-ref-27)
28. http://www.cncorpus.org/Resources.aspx [↑](#endnote-ref-28)
29. http://pinyin.sogou.com/dict/cate/index/132 [↑](#endnote-ref-29)
30. http://terms.naer.edu.tw/forum/noun/ [↑](#endnote-ref-30)
31. 基于信息论的潜在概念获取与文本聚类 [↑](#endnote-ref-31)
32. 中文文本分类中的特征选择算法研究 [↑](#endnote-ref-32)
33. A maximum entropy model for part-of-speech tagging [↑](#endnote-ref-33)
34. Chinese Word Segmentation Based on Contextual Entropy [↑](#endnote-ref-34)
35. Contextual Entropy and Text Categorization [↑](#endnote-ref-35)
36. Introduction to information extraction [↑](#endnote-ref-36)
37. 文本信息抽取模型及算法研究\_周顺先 [↑](#endnote-ref-37)
38. 基于支持向量机的中文组织机构名识别 [↑](#endnote-ref-38)
39. P. H. Chen, C. J. Lin, and B. Schölkopf, A tutorial on ν-support vector machines, Appl. Stoch. Models. Bus. Ind. 2005,   21, 111-136.  [↑](#endnote-ref-39)
40. [Maximum Entropy Markov Models for Information Extraction and Segmentation.](http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/gidofalvi.pdf) [↑](#endnote-ref-40)
41. LafFerty J D, McCallum A,Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufinann Publishers Inc., 2001: 282-289 [↑](#endnote-ref-41)
42. [Conditional random fields: Probabilistic models for segmenting and labeling sequence data](http://repository.upenn.edu/cis_papers/159/) [↑](#endnote-ref-42)
43. 基于条件随机场的自动分词技术的研究 [↑](#endnote-ref-43)
44. [基于层叠条件随机场模型的中文机构名自动识别](http://epub.cnki.net/kns/detail/detail.aspx?QueryID=8&CurRec=3&recid=&FileName=DZXU200605007&DbName=CJFD2006&DbCode=CJFQ&pr=) [↑](#endnote-ref-44)
45. [Shallow parsing with conditional random fields](http://dl.acm.org/citation.cfm?id=1073473) [↑](#endnote-ref-45)
46. [Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons](http://dl.acm.org/citation.cfm?id=1119206) [↑](#endnote-ref-46)
47. McDonald, Ryan; Pereira, Fernando: Identifying gene and protein mentions in text using conditional random fields. In: BMC Bioinformatics 6 (Suppl 1) (2005), May, No. S6. [↑](#endnote-ref-47)
48. Sutton, Charles; McCallum, Andrew: An Introduction to Conditional Random Fields for Relational Learning. In: Getoor, Lise; Taskar, Benjamin (Editors.): Introduction to Statistical Relational Learning. MIT Press, November 2007, Chap. 4, pp. 93–127. [↑](#endnote-ref-48)
49. Finkel, Jenny R.; Grenager, Trond;Manning, Christopher: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL), 2005, pp. 363–370. [↑](#endnote-ref-49)
50. Chen, Stanley F.; Rosenfeld, Ronald: A Survey of Smoothing Techniques for ME Models. In: IEEE Transactions on Speech and Audio Processing 8 (2000), No. 1, pp. 37–50. [↑](#endnote-ref-50)
51. The Improved Iterative Scaling Algorithm:A Gentle Introduction [↑](#endnote-ref-51)
52. Generalized Iterative Scaling for Log-Linear Models [↑](#endnote-ref-52)
53. [Sequential conditional generalized iterative scaling](http://dl.acm.org/citation.cfm?id=1073086)s [↑](#endnote-ref-53)
54. [A limited memory algorithm for bound constrained optimization](http://epubs.siam.org/doi/abs/10.1137/0916069) [↑](#endnote-ref-54)
55. S. Bandyopadhyay and U. Maulik, Knowledge Discovery and Data Mining, in Advanced Method for Knowledge Discovery from Complex Data, S. Bandyopadhyay, U. Maulik, L.B. Holder and D.J. Cook, eds, USA: Springer, 2006, pp. 164. [↑](#endnote-ref-55)
56. Rabiner, Lawrence R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Proceedings of the IEEE 77 (1989), No. 2, pp. 257–286. [↑](#endnote-ref-56)
57. 基于领域本体的语义检索系统的研究与实现\_武俊丽 [↑](#endnote-ref-57)
58. 语义检索研究综述\_黄敏 [↑](#endnote-ref-58)
59. http://wordnet.princeton.edu/ [↑](#endnote-ref-59)
60. http://www.keenage.com/ [↑](#endnote-ref-60)
61. 基于向量空间模型的中文文本相似度算法研究\_陈飞宏 [↑](#endnote-ref-61)
62. Salton G, Buckley B. Term-weighting Approaches in Automatic Text Retrieval [J]. Information Processing and Management, 1998, 24(5); 513-523. [↑](#endnote-ref-62)
63. 中文分词十年回顾\_黄昌宁 [↑](#endnote-ref-63)
64. 适应于不同领域的中文分词方法研究与实现\_修驰 [↑](#endnote-ref-64)
65. Zellig S. Harris. 1955. From phoneme to morpheme.Language, 31(2):190–222. [↑](#endnote-ref-65)
66. Unsupervised segmentation of Chinese text by use of branching entropy [↑](#endnote-ref-66)
67. Unsupervized Word Segmentation the case for Mandarin Chinese [↑](#endnote-ref-67)
68. 词典与统计相结合的中文分词的研究\_岳中原 [↑](#endnote-ref-68)
69. 词典和机器学习相结合的生物命名实体识别\_王琦 [↑](#endnote-ref-69)
70. [Chinese unknown word identification using character-based tagging and chunking](http://dl.acm.org/citation.cfm?id=1075215) [↑](#endnote-ref-70)
71. [Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based?](http://www-new.comp.nus.edu.sg/~nght/pubs/emnlp04.pdf) [↑](#endnote-ref-71)
72. 文本病历信息抽取方法研究\_李莹 [↑](#endnote-ref-72)
73. 梁南元. 书面汉语自动分词系统-CDWS[J].中文信息学报, 1987, 1(2):44-52. [↑](#endnote-ref-73)
74. http://crfpp.googlecode.com/svn/trunk/doc/index.html [↑](#endnote-ref-74)
75. Kohavi, Ron. [A study of cross-validation and bootstrap for accuracy estimation and model selection](http://citeseer.ist.psu.edu/kohavi95study.html). Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. 1995, 2 (12): 1137–1143 [↑](#endnote-ref-75)
76. 近几年来国外信息检索模型研究进展\_孙坦 [↑](#endnote-ref-76)
77. 国内外主要本体库比较分析研究\_白如江 [↑](#endnote-ref-77)
78. 信息检索模型及相关性算法的研究\_吴丽华 [↑](#endnote-ref-78)
79. 基于向量空间模型的信息检索系统的研究与实现\_赵耀红 [↑](#endnote-ref-79)
80. Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. Cambridge: Cambridge university press, 2008:6-8. [↑](#endnote-ref-80)
81. [基于向量空间模型的中文文本相似度算法研究](http://epub.cnki.net/kns/detail/detail.aspx?QueryID=0&CurRec=8&recid=&FileName=1011193114.nh&DbName=CMFD2011&DbCode=CMFD&pr=) [↑](#endnote-ref-81)
82. [基于改进向量空间模型的话题识别与跟踪](http://epub.cnki.net/kns/detail/detail.aspx?QueryID=0&CurRec=19&recid=&FileName=WJFZ200609021&DbName=CJFD2006&DbCode=CJFQ&pr=) [↑](#endnote-ref-82)
83. 基于N层向量空间模型的信息检索算法\_陈治平 [↑](#endnote-ref-83)