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：**

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

**关键词：**

## Abstract

**Key words:**

## 目 次

[致 谢 IV](#_Toc408833100)

[摘 要 V](#_Toc408833101)

[Abstract VI](#_Toc408833102)

[目 次 VII](#_Toc408833103)

[第1章 绪论 1](#_Toc408833104)

[1.1 课题研究背景和意义 1](#_Toc408833105)

[1.2 国内外研究现状 3](#_Toc408833106)

[1.3 本文研究内容和意义 6](#_Toc408833107)

[1.4 论文的组织结构 8](#_Toc408833108)

[第2章 基本理论知识 9](#_Toc408833109)

[2.1 信息论基础 9](#_Toc408833110)

[2.2 文本信息抽取 11](#_Toc408833111)

[2.3 文本信息检索 22](#_Toc408833112)

[2.4 本章小结 22](#_Toc408833113)

[第3章 基于openEHR的电子病历管理系统构建 23](#_Toc408833114)

[3.1 系统需求分析与设计目标 23](#_Toc408833115)

[3.2 系统总体设计 23](#_Toc408833116)

[3.3 系统功能设计及工作流程 23](#_Toc408833117)

[3.4 数据库设计 23](#_Toc408833118)

[3.5 本章小结 23](#_Toc408833119)

[第4章 文本处理和挖掘技术在病例循证子系统中的应用 23](#_Toc408833120)

[第5章 系统测试和分析 23](#_Toc408833121)

[5.1 信息抽取和病例检索测试分析 23](#_Toc408833122)

[5.2 系统功能测试 23](#_Toc408833123)

[5.3 本章小结 23](#_Toc408833124)

[第6章 总结与展望 24](#_Toc408833125)

[6.1 总结 24](#_Toc408833126)

[6.2 展望 24](#_Toc408833127)

[参考文献 25](#_Toc408833128)

**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（如搜狗医学词库[[29]](#endnote-29)、学术名词资讯网等[[30]](#endnote-30)）。这样搜集的预料往往很庞大，而且很杂，文本利用语料特征合适度[[31]](#endnote-31)来筛选出合适的预料，最后以此为基础预料，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，作为最后的词表预料。
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[[32]](#endnote-32)。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[[33]](#endnote-33)。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[[34]](#endnote-34)。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[[35]](#endnote-35)。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[[36]](#endnote-36)。

上下文熵(Contextual Entropy，简称CE，又称边界熵，Branching Entropy)，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)[[37]](#endnote-37)。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[[38]](#endnote-38)。

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。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。



图2-1 基于规则的模型信息抽取过程示意图

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[[39]](#endnote-39)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[[40]](#endnote-40)。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[[41]](#endnote-41)。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[[42]](#endnote-42)。

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[[43]](#endnote-43)。目前CRFs在文本处理的各种任务中具有很好的表现[[44]](#endnote-44)[[45]](#endnote-45)[[46]](#endnote-46)[[47]](#endnote-47)。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2-3 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[[48]](#endnote-48)，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[[49]](#endnote-49)[[50]](#endnote-50)。

##### 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[[51]](#endnote-51)。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)[[52]](#endnote-52)，归纳的迭代缩放算法(Generalised Iterative Scaling，GIS)[[53]](#endnote-53)，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)[[54]](#endnote-54)，LBFGS算法[[55]](#endnote-55)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，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；大于1时，准确率更加重要一些； 小于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

基于领域本体的语义检索系统的研究与实现\_武俊丽-page18

简述

http://www.52nlp.cn/tag/%E6%B5%85%E5%B1%82%E8%AF%AD%E4%B9%89%E5%88%86%E6%9E%90

#### 2.3.1 相似度算法

#### 2.3.3 语义

文本病例信息抽取方法研究 ---语义 page15

### 2.4 本章小结

## 第3章 基于openEHR的电子病历管理系统构建

### 3.1 系统需求分析与设计目标

### 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 文本处理和挖掘技术在病例循证子系统中的应用

## 第5章 系统测试和分析

### 5.1 信息抽取和病例检索测试分析

### 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

## 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. 适应于不同领域的中文分词方法研究与实现\_修驰 [↑](#endnote-ref-31)
32. 基于信息论的潜在概念获取与文本聚类 [↑](#endnote-ref-32)
33. 中文文本分类中的特征选择算法研究 [↑](#endnote-ref-33)
34. A maximum entropy model for part-of-speech tagging [↑](#endnote-ref-34)
35. Chinese Word Segmentation Based on Contextual Entropy [↑](#endnote-ref-35)
36. Contextual Entropy and Text Categorization [↑](#endnote-ref-36)
37. Introduction to information extraction [↑](#endnote-ref-37)
38. 文本信息抽取模型及算法研究\_周顺先 [↑](#endnote-ref-38)
39. 基于支持向量机的中文组织机构名识别 [↑](#endnote-ref-39)
40. 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-40)
41. [Maximum Entropy Markov Models for Information Extraction and Segmentation.](http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/gidofalvi.pdf) [↑](#endnote-ref-41)
42. 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-42)
43. [Conditional random fields: Probabilistic models for segmenting and labeling sequence data](http://repository.upenn.edu/cis_papers/159/) [↑](#endnote-ref-43)
44. 基于条件随机场的自动分词技术的研究 [↑](#endnote-ref-44)
45. [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-45)
46. [Shallow parsing with conditional random fields](http://dl.acm.org/citation.cfm?id=1073473) [↑](#endnote-ref-46)
47. [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-47)
48. 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-48)
49. 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-49)
50. 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-50)
51. 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-51)
52. The Improved Iterative Scaling Algorithm:A Gentle Introduction [↑](#endnote-ref-52)
53. Generalized Iterative Scaling for Log-Linear Models [↑](#endnote-ref-53)
54. [Sequential conditional generalized iterative scaling](http://dl.acm.org/citation.cfm?id=1073086)s [↑](#endnote-ref-54)
55. [A limited memory algorithm for bound constrained optimization](http://epubs.siam.org/doi/abs/10.1137/0916069) [↑](#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)