摘 要

Internet 上数据量急剧膨胀使其成为企业竞争情报获取的重要来源，然而如何从这个信息海洋中找到企业所需要的情报成为困扰企业竞争情报获取的难题。商业信息抽取作为解决这一难题的重要手段，其抽取结果的好坏对最终竞争情报的形成有着重要的影响。

本文对 Web 环境上的商业信息抽取技术进行了研究，主要关注两个方面：商业信息中的关系抽取和实体抽取。针对抽取对象的不同特征，研究不同的技术方法，以提高抽取的召回率和准确率。其中关系信息抽取以职位关系抽取为例，分析了职位关系实例在网页中的呈现特征，设计了基于结构特征的职位关系抽取算法；实体抽取以机构名识别为例，基于语言学中语法对语义的依赖关系和共生性词场两个观点，提出了语义隐马尔可夫模型的机构名识别算法。两个算法有效改善了商业信息抽取效果，同时也为其它商业信息抽取提供了参考。

本文的主要贡献主要有：

（1）提出了基于 Web 的职位关系抽取算法。职位关系反映了一个人在一个组织所占据的职位，是一种重要的竞争情报。本文分析了网页中职位关系实例的特征，并利用结构化系数和结构化文件片断对这些特征进行描述，最后利用模式匹配的方法从结构化文件片断中抽取出职位关系。实验结果表明算法达到了准确率超过 96%、召回率超过 87%的较好结果。

（2）提出了基于语义隐马尔可夫模型的中文机构名识别算法。语义隐马尔可夫模型的构建以语言学中的语法对语义的依赖关系和共生性词场两个重要观点为理论依据。一个句子可以看作是一个词的序列，这个序列背后隐含着一个语义序列，且语义序列决定了句子的组成。我们首先对机构名及其上下文中的词进行语义标注，然后构建语义隐马尔可夫模型。在机构名上下文选择时利用共生性词场现象来决定上下文的边界。事实上，算法试图利用机构名与其上下文之间的语义关联性来提高机构名识别的效果。实验表明算法改善了机构名识别效果，而且普适性更好。

关键词：商业信息 竞争情报 信息抽取 关系抽取 命名实体识别

Abstract

As the rapid increasing of the data volume in Internet, Web has become an important source for competitive intelligence acquisition. However, it is still a difficult task for enterprises to obtain competitive intelligence from this information ocean. To solve this problem, researchers introduced the technology of business information extraction into competitive intelligence acquisition, in which the result of information extraction plays an important role in the quality of competitive intelligence.

In this paper, we study the issues of business information extraction from the Web and focus on two aspects in this area: relation extraction and entity recognition. For different extracting objects, we analyze their distinctive features and develop appropriate methods to extract these objects in order to improve the effect of business information extraction. Position relation extraction is set as an example for business relation extraction. We investigate the appearance features of position relation instances on the Web and adopt structure-based algorithm to extract position relations from the Web. For entity recognition, we research the organization name entity recognition and present an organization name entity recognition algorithm based on Semantic Hidden Markov Model. Two algorithms effectively improve the effect of two kinds of information extraction respectively and provide reference information for other business information extraction.

The main contribution of this paper can be summarized as follows:

(1) We present an algorithm to extract position relations from the Web. People's position in a corporation, which the term *position relation* refers to, is a kind of significant competitive intelligence for enterprises. Our algorithm is based on the structural feature of position relation in Web contents. We first introduce structural coefficient and structural file segment to describe these features and then employ a pattern-matching method to extract position relations from the structural file segments. Finally, we conduct experiments on a real data set and evaluate the precision and recall of our approach. The experimental results show that our algorithm has a high precision over 96% as well as a recall over 87%.

(2) We bring forward a SHMM-based Chinese organization name recognition algorithm. Semantic Hidden Markov Model is based on two important linguistics viewpoints: the dependence of syntax on semantics and symbiotic word field. A

sentence is considered as a sequence of words, this sequence implies a semantic sequence which decides the construction of the sentence. We first conduct semantic tagging on the words from organization name interior and its context, and then construct semantic Hidden Markov Model for organization name recognition. During the selection of organization name context, we employ symbiotic word field phenomenon to decide the boundary of the context. In fact, the algorithm attempt to make use of the relevancy between organization name and its context to improve the effect of organization name recognition. The experimental results show that our algorithm gains better outcome compared to other approaches and has a stronger ability to process different type of contents.

key words: business information, competitive intelligence, information extraction, relation extraction, named entity recognition

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文，是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：刘嘉宏　　签字日期：2010.5.20

中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

☑公开　□保密（____年）

作者签名：刘嘉宏　　导师签名：

签字日期：2010.5.20　　签字日期：2010.5.23

第一章 绪论

互联网的迅猛发展和快速普及使其成为人们获取各类信息的重要源泉。在改变人们生活和工作方式的同时，互联网也使企业的商业活动方式发生重大变化。越来越多的企业通过互联网来推广它们的产品、宣传它们的服务；各类网络媒体也争抢在第一时间将企业的相关信息发布到网上。在这样的环境下，互联网成为企业竞争的新战场，谁能够以最快的速度获取到真实、有效的商业信息，特别是竞争对手的信息，谁就能在这个新战场上获胜。然而互联网上的信息纷繁复杂，想要在海量的网页中找到对企业有价值的商业信息来还需要克服许多技术上的难题，这也是本文研究的重点。

1.1 研究背景与意义

市场经济的繁荣和发展加剧了企业间的竞争，一个企业要想在激烈的市场竞争中生存并获胜，不但要清楚自己的优势和劣势，还需要了解行业政策、市场需求变化、竞争对手等商业信息。以往的商业信息一般通过人际关系、纸制传媒等方式获取。但近几年，随着互联网的飞速发展，网络成为获取商业信息的一个重要途径，基于互联网的商业信息抽取成为企业和学术界研究的重点。

（1）互联网成为世界上最大的信息库

毫无疑问，互联网是世界上容量最大、内容最丰富的信息库，而且平均每天以千万级网页的数量增长。根据瑞典互联网流量监测机构 Pingdom 近期公布的数据[1]，2009 年全球网站数量已达到 2.34 亿家，其中 2009 年新增 4700 万家。中国互联网络信息中心（CNNIC）2010 年 1 月发布的第 25 次《中国互联网络发展状况统计报告》显示，截止 2009 年 12 月中国的网站总数达到 323 万个，网页总数达到 336 亿个[2]。从 2003 年开始，中国的网页规模基本保持翻番增长，年增长率超过 100%。这些数据充分表明互联网是世界上最大的电子图书馆，且无地理位置限制，使用成本低，已经成为人们获取信息的重要源泉。

（2）互联网上存在大量有价值的信息

第 25 次《中国互联网络发展状况统计报告》中关于主要网络应用使用行为的调查显示，基于网络新闻和搜索引擎的信息获取成为主要的网络行为，使用率分别达到 80.1%和 73.3%，使用率排名分别占到第二位和第三位，仅次于网络音乐的使用率，而且两项应用的用户增长率分别达到 31.5%和 38.6%。CNNIC 分析师认为网络应用的日趋丰富和网络信息量的与日俱增是网络信息获取行为增长的主要原因。网络使用行为能够反映人们的需求态势，以上数据表明，互联网不

仅信息量巨大，而且是一部百科全书。对于一个企业来说，最想获取的信息莫过于对自身成长和发展有利的商业信息。这些信息涉及行业政策、市场环境、竞争对手等，其中关于竞争对手的信息最为重要。商业信息经过汇总、分析后可以成为有价值的竞争情报，给企业的决策提供有力支持。由于互联网上存在大量的商业信息，基于互联网的竞争情报获取成为当前的研究热点。据美国海军高级情报分析员埃利斯·扎卡利亚斯讲，95%的竞争情报来自于公开资料，4%来自于半公开资料，仅1%或更少来自机密资料。而互联网无疑是获取竞争情报最重要的公开信息源。

（3）信息获取难度大

互联网上的海量信息在丰富信息来源的同时，也给信息的获取造成了困扰。互联网上的网页数以百亿计，靠人工一个一个地浏览网页来收集信息无疑是大海捞针。搜索引擎和门户类网站的产生给人们获取信息的方式带来了革命性的变化，使信息获取变得容易了很多。为了进一步提高信息获取的准确性，还发展出了垂直搜索技术。目前，利用网络获取商业信息的途径有利用搜索引擎技术、利用行业站点和黄页、利用竞争对手的网站、竞争情报获取软件等，然而这些方法仍不能很好地满足的企业对商业信息的需求。以搜索引擎为例，针对一个查询返回的网页往往数以百万计，人工从这些网页中找到所需的信息仍需大量的工作。另外，有时为了获取某种商业信息，需要使用不同的查询关键字进行多次查询，并将找到的信息进行汇总，这大大加大了商业信息获取的工作量。在这种背景下，信息抽取技术应运而生。信息抽取就是从文本中获取感兴趣的知识点，并以结构化的形式保存在数据库中，以便以后的查询和使用。信息抽取技术能够进一步把人们从人工查找信息的繁重劳动中解放出来，提高信息获取的效率。比尔·盖茨在其著作《未来时速》一书中讲到："将您的公司和您的竞争对手区别开来的最有意义的方法，使您的公司领先于众多公司的最好方法，就是利用信息来干最好的工作。您怎样收集、管理和使用信息将决定您的输赢。"。可见，及时、全面、准确地获取商业信息是决定一个企业成败的关键。互联网为商业信息的获取提供了资源，而如何从这个信息海洋中找到企业所需要的那根"针"是重点要解决的问题。

1.2 商业信息与竞争情报

我国著名情报专家包昌火指出：竞争情报是关于竞争环境、竞争对手与竞争策略的信息和研究[3]。竞争情报获取分为规划与定向、信息采集、信息加工、情报分析及情报传播五个阶段。商业信息获取涵盖了信息采集与信息加工两个阶

段。信息采集阶段主要是原始信息的收集，如网页、纸制资料等。在信息加工阶段，对采集到的信息进行初步处理，主要采用一些自动化的技术，如自动分类、自动摘要、文档去重、信息抽取等。经过加工后的信息必须经过情报分析后才能成为真正意义的竞争情报，主要的分析方法有 SWOT 分析法、德尔斐法和定标比超法及数据挖掘的方法。

表 1.1 竞争情报具体内容

编号	调研分类	调研内容
1	基本概况	企业简介、组织框架、股本结构、行业背景、产品概况、行业地位
2	人力资源	员工数量、素质、学历结构、主要管理者背景、经验、培训制度、聘用程序、奖惩制度、薪资水平/结构、福利体系人力资源
3	管理团队	高层架构、重要决策权利分布、高层领导背景
4	生产能力	生产线情况（设备明细、投资渠道、技术水平、生产能力、使用率、净值率）、技术人员分布、主要技术人员介绍（特长、背景、学历、工程项目）、生产员工操作熟练程度、生产环境、是否OEM
5	研发能力	研发队伍资历（人数、学历、结构）、专利资源（数量、技术含量）、开发费用现状（金额、来源）、与其他企业/学校/政府机关合作情况（合作数量、技术含量、未来发展趋势）
6	质量检查	认证情况、质量控制流程、方案、质量控制专员（数量、素质、经验、背景）、主要产品质量指标（产品平均寿命、合格率等）
7	原材料采购	原材料（采购量、来源、平均价格）、供应商情况（数量、供货评价、供应商关系）、采购支付情况（结款方式、期限、信用额度等）、供应商的挑选机制采供销
8	产品结构	产品明细、种类、详细介绍（规格、型号、性能、用途、优势）、产品服务、产品技术含量（技术水平、技术参数、技术性能）、原材料采购成本、产品价格体系、市场定位、供货能力
9	产品销售	销售部门设置、直接销售渠道（分产品、区域、行业）、分销渠道（代理商数目、结算方式、分销商评价、关系、未来的发展战略）、历年销售情况（数量、金额、趋势）、市场占有率及趋势、销售策略（利润为主或开拓市场为主）、销售价格体系（分出厂价、批发价、零售价）、价格回扣与折扣（回扣条件、折扣率）
10	广告推销	品牌模式、广告营销情况（现有媒体数量、广告投入金额、占销售比重、未来发展趋势）、主要营销分布（分产品、形象）、主要使用广告媒体（报纸、杂志、电视、广播、网络、交通广告、户外）
11	客户情况	主要客户（数量、购买力、购买的产品情况）、客户分布（区域、行业、金额）、客户满意程度、客户投诉及退货、客户维护模式
12	财务情况	连续N年的资产负债表、损益表、现金流量表、财务指标（销售增长率、资产负债率、存货周转、流动资产周转、总资产周转）

商业信息抽取属于信息加工阶段的最后一个环节，衔接着情报分析阶段，随着信息量的急增，这部分工作变的越来越重要。信息加工阶段的其它工作都是基于文档一级的处理，处理的结果仍是文档，还需用户人工在这些文档中查找信息，而随着采集到的信息成倍增长，使得这部分工作单纯依靠人工操作变得不太现实。在这种情景下，信息抽取工作变得尤为重要，抽取结果的质量直接关系到后面情报分析的准确性，进而影响企业决策的正确性。

广义的商业信息包括行业环境、市场态势、竞争对手等信息。狭义的商业信息主要是关于竞争对手的信息，也就是以一个企业为中心的相关信息。表 1.1 是北京东方策略科技有限公司能够提供的关于某个企业的竞争情报具体内容[4]。它将企业的竞争情报分为十二个大类，每个大类又分若干小类。可以说，这些竞争情报都是由零散的商业信息整合而成，如管理团队是由多个职位关系信息组成的。

本文主要研究两类商业信息的抽取：实体信息和关系信息。实体信息抽取是从文本中抽取商业信息涉及的命名实体，如机构名、人名、地名、产品名等。关系信息抽取主要是抽取实体间关系的描述信息。文献[5]通过本体来描述竞争情报中的实体信息和关系信息，表 1.1 中的商业信息可以按同样的方法进行描述，然后再抽取具体的实体信息和关系信息。

1.3 商业信息抽取的国内外研究现状

目前，商业信息抽取的研究主要集中在竞争情报研究领域，下面主要介绍国内外的竞争情报研究中关于商业信息抽取的研究现状。

1.3.1 国外研究现状

国外竞争情报研究中关于商业信息抽取方面的工作已经比较多，开始从只是简单地将现有的信息抽取工具集成到竞争情报系统中向专门研究特定应用的信息抽取技术转变。另外，一些商用的竞争情报软件也开始加入信息抽取的功能，使其更加实用化。

Byron Marshall 等提出一个商业信息集成工具 EBiZPort [6]，采用元搜索(meta-search)技术来收集信息，以提高信息收集的召回率和质量。另外，该工具还对收集到信息作了进一步的处理，如摘要提取、自动分类、可视化设计等，也涉及了信息整合问题，但只是文档级的，仍需要用户人工到返回的文档中寻找所需的信息。

Francois Paradis 等设计的 MBOI（Matching Business Opportunities On the

Internet）系统[7]试图从互联网上寻找与企业相关的招标（Call for Tenders）信息。系统使用 NStein NFinder 工具进行命名实体抽取，目的是为了查询时更准确地定位所需要的信息和改善后续的信息分类效果。NStein NFinder 工具采用词法规则与词典相结合的方法进行命名实体识别，主要是针对一些结构规范的网页，如表格和列表。

Robert Baumgartner 等提出一个基于 Web 的商业情报抽取系统 Lixto[8]。该系统采用包装器（wrapper）信息抽取技术从半结构化的电子商务网站中抽取商品信息，如商品的名称、制造商及价格等，并将抽取结果以 XML 文档的形式保存。包装器是出现最早的基于 Web 的信息抽取技术，专门针对结构化较强的网站，如招聘网站、购物网站。

h-TechSight 是 Diana Maynard 等设计的一个知识管理系统[9]，其功主要功能是对网上的敏感信息进行实时监控。系统对网页中敏感的概念信息进行抽取，并对这些概念的变化进行实时监控。概念抽取使用了 GATE 工具中的信息抽取（IE）组件，该组件采用基于规则的方法进行概念的识别，但抽取规则需要人工总结，不能自动生成。

2007 年，Diana Maynard 等又提出了一个基于领域本体（Domain Ontology）的商业情报系统[10]。系统采用本体来描述领域概念、概念间的关系及属性，根据本体定义来抽取领域信息。该系统主要是针对跨国公司的商业情报需求开发的，主要抽取有关公司概况（公司名、所在国家、电话、邮编、分支机构、主要业务、进出口业务、营业额、雇员数量、股东及其它相关人员）、国家/区域概况（国家名字、人口数量、土地面积、官方语言、货币、汇率、外债、失业率、GDP、海外投资）等信息，主要为跨国公司的海外投资提供决策支持。系统采用 ANNIE 工具进行概念抽取。ANNIE 是一个基于规则的通用的概念抽取工具，主要抽取人名、地名和机构名。为了适应金融领域的概念，Diana Maynard 等对 ANNIE 工具进行了修改以适应新的需求。系统对不同来源的文本设计了不同的抽取规则，这些文本既有结构化的（主要是表格），也有非结构化的，并且对来源不同但涉及同一概念的信息进行了整合。

随着互联网上信息量急剧增长和信息处理技术的进步，国外开始出现了一些可以实用的商用竞争情报软件。起初这些软件只具备以文档为单位的信息收集功能以及一些简单预处理功能。近几年，信息抽取和数据挖掘技术逐步被集成到竞争情报软件中，提高了其实用价值。表 1.2 列出了目前国外主要竞争情报软件的功能。可以看出，数据挖掘技术（自动分类、自动摘要）在竞争情报软件中应用已经比较广泛，而信息抽取技术刚刚开始集成到竞争情报软件中，还有很大的发展空间。

表 1.2 国外竞争情报软件功能

竞争情报软件	支持多种文件格式	自动分类	自动摘要	关系抽取	关系可视化	自动排序	关键字搜索	监视和预警	情报发布	自然语言搜索
WebQL	√									
Knowledge Works	√		√			√	√	√	√	
Text Analytics				√	√					
Textanalyst		√	√	√			√			√
PolyAnalyst	√	√	√	√	√		√	√	√	√
TrackEngine								√		
Trendicate		√	√			√		√		√
wincite	√	√	√				√		√	

1.3.2 国内研究现状

与国外相比，国内的竞争情报研究和信息抽取研究都处于探索阶段，技术都不够成熟。表 1.3 是 2007 年 1 月到 2008 年 3 月国内竞争情报核心期刊上发表的有关竞争情报的 83 篇文献的统计信息。可以看出，信息加工已成为国内竞争情报研究的热点，主要采用数据挖掘方法，如文本分类、文本挖掘、Web 挖掘等对收集的文本或网页进行初步处理，还没有涉及句子级的处理，如实体和关系抽取等。

表 1.3 竞争情报研究方向统计表

研究方向	篇数	研究方向	篇数
信息加工	**17**	研究现状	**5**
情报理论	**9**	情报评估	**4**
人际网络	**7**	情报教育	**3**
网络组织	**7**	情报分析	**3**
知识管理	**5**	其它	**18**
产业发展	**5**		

2005年，刘非凡等提出了基于层级隐马可夫模型（HHMM）的产品名识别算法[11]，从自由文本中抽取产品名。这是第一次专门针对中文商业信息抽取的研究，突破了信息抽取研究只专注传统信息抽取的禁锢，推动信息抽取技术向实用化方向迈进了一步。

2006年，Wei Li等提出并实现了一个中文竞争情报系统C-CIS[12]，实现了四项功能：情报定制、信息采集、信息加工、和情报发布。信息加工模块包含了自动分类、文档去重和信息抽取三项功能。其中信息抽取子模块采用基于规则和词典相结合的技术，可以抽取命名实体、事件等信息。

2008年，Yan Chen等提出了基于本体的竞争情报抽取系统[5]，利用本体来描述企业对竞争情报的需求，然后采用基于规则和模板匹配的方法将抽取的信息填充到实例本体中。

目前，中文竞争情报软件主要提供网页收集功能，如天下互联的网络情报中心。给定关键字后，这些软件到事先定制好的网站上实时爬取网页，将包含关键字的网页返回给客户。一些功能较强的竞争情报软件还提供了自动分类、自动摘要等功能，如TRS竞争情报软件。然而这些软件没有提供对网页内部的关键信息进行抽取和集成，还需要人工从返回的网页中寻找所需的信息。

国内竞争情报领域关于商业信息抽取研究较少的一个原因是中文信息抽取技术目前还没有突破性的进展，这使得中文信息抽取技术无法集成到竞争情报系统中，成为影响竞争情报获取的重要技术瓶颈，所以研究中文商业信息抽取技术，使其达到实用化水平，对企业竞争情报获取有着重要意义。

综上所述，目前，国外商业信息抽取研究比较多，一般是将现有的信息抽取技术应用到商业信息抽取中，并开始探索专门针对特定领域的商业信息抽取技术，一些商用竞争情报软件已具备商用价值。而中文商业信息抽取研究比较少，中文竞争情报软件的商用价值不高，探索和发展中文商业信息抽取技术迫在眉睫。

传统信息抽取的处理对象主要是自由文本，且只对的常规信息（如人名、地名、机构名等）进行抽取。随着互联网上信息量的急剧增长，基于网页的信息抽取研究随之成为研究的热点。与自由文本相比，网页具有两个特征：

（1）网页是一种半结构化的文本。除了常规文本外，网页中包含了大量的HTML标签，这些标签使得网页文本具有了半结构化的特征。一方面 ，可以利用这些标签提高信息抽取的效果；另一方面，网页标签复杂多样，往往与常规文本混杂在一起，也成为信息抽取的障碍。因此，如何利用网页文本半结构化的优点，避免其缺点，成为网页信息抽取成败的关键因素。

（2）网页文本规范性较差。自由文本一般来自报纸或书籍，在词法、句法

和表述方式上比较规范。而网页文本缺乏严格要求，行文较自由，不像纸制文本那样规范。因此，基于纸制文本的信息抽取方法应用于网页信息抽取时，效果不一定理想。对于网页信息抽取，需要设计容错能力更强的信息抽取系统。

Internet 是商业信息的重要来源，研究基于网页的商业信息抽取技术既要利用传统信息抽取技术已有的成果，又要充分考虑网页自身的特性，利用网页提供的便利信息来提高信息抽取的效果，同时避免一些不利信息的负面影响。

1.4 本文的主要工作

本文以 Internet 中的网页为信息源，研究基于 Web 的商业信息抽取技术。并以职位关系和机构名两种具体的商业信息为对象，设计具体的信息抽取算法。主要研究内容有：

（1）商业信息中实体抽取技术研究

实体是商业信息中的基本信息单元，实体识别是实现其它商业信息抽取的基础。商业信息中的实体主要有机构名、人名、地名、产品名、商标名等。目前，中文命名实体识别的研究主要集中在机构名、人名和地名，其中机构名由于构成比较复杂，识别效果不佳，而机构名又是商业信息中最最要的一个实体。研究国内外现有命名实体识别技术，特别是机构名识别技术，设计网页环境下中文机构名识别技术，提高机构名识别的召回率和准确率是本文的主要研究工作之一。

（2）商业信息中关系抽取技术研究

商业信息中的实体关系描述了现实世界中两个实体由于企业的生产、销售等活动而发生的相互联系。实体关系是一种更为重要的商业信息，与实体相比，实体关系已经是一种浅层的知识，透过这种关系可以对某些事实有所了解。当将相关的实体关系组织起来，形成一个信息网时，商业信息就会成为一种重要的竞争情报，为企业的决策提供支持。中文关系抽取的研究较少，基于 Web 的关系抽取研究则更少。采用的方法主要借鉴国外现有的技术，如模式匹配方法和机器学习方法等。研究基于 Web 的中文商业关系抽取技术，并以职位关系抽取为例，设计具体的关系抽取算法是本文的另一项研究工作。

1.5 本文的组织结构

本文的组织结构如下：

第一章 绪论。对基于 Internet 的商业信息抽取研究作了整体介绍，包括研究的背景、意义，商业信息抽取在企业竞争情报获取中的地位，以及商业信息抽取

的国内外研究现状。

第二章 信息抽取技术。简要介绍了信息抽取技术研究的发展历程、研究内容、国内外主要的信息抽取系统以及信息抽取技术的评测指标。

第三章 详细论述了基于 Web 的职位关系抽取方法，提出了结构化系数、结构化文件片断、标准模式等概念，并利用这些概念来描述网页中职位关系的结构特征。最后，通过在真实语料集上的实验来检验算法的有效性。

第四章 基于中文句子中语法与语义之间的关联性，提出语义隐马尔可夫模型的机构名识别算法，并进行了具体的讨论。另外，通过实验比较对所提方法效果进行了验证。

第五章 总结与展望。对本文做出了总结，并指出其中的不足之处，展望下一步的研究方向。

第二章 信息抽取技术

2.1 前言

在日益信息化和网络化的当代社会，如何找到感兴趣的信息并对些信息进行归类、过滤和提取，一直是一个比较紧迫的实际问题。信息量的急增使得早期单纯依靠人工收集信息的方法不再可行，自动化的信息采集工具成为迫切需求。在这样的背景下，信息抽取技术应用而生。信息抽取的目标是对文本中感兴趣的信息进行提取，并以结构化的形式集中存储起来，以便于以后的查询和更高一级的应用。信息抽取系统的输入是一系列原始文本，输出的是具有一定格式的结构化信息集。

信息抽取的文本可以分为三类：自由文本、半结构化文本和结构化文本，后两者主要是指网页文本。对于不同类型的文本，所采用的信息抽取技术可能会有所不同。目前，由于网页数量的急剧膨胀，基于网页的信息抽取技术成为研究的重点。

2.2 信息抽取研究简史

根据信息抽取研究的发展轨迹，一般将信息抽取的研究划分为三个阶段：前期、中期和近期。

信息抽取研究的前期开始于 20 世纪 60 年代中期，结束于 80 年代中期，这是信息抽取研究的初始阶段，主要以两个长期的自然语言处理项目为代表。第一个项目是美国纽约大学的 Linguistic String 项目[13]。该项目研究如何构建一个大规模的英语计算语法，其中与信息抽取相关的应用是从医疗领域的 X 光报告和医院的出院记录中生成信息格式。事实上这种信息格式就是后来消息理解会议定义的模板。第二个项目是耶鲁大学 Roger Schank 及同事开展的有关故事理解的研究。他的学生 Gerald De Jong 设计实现了一个叫作 FRUMP 的信息抽取系统[14]。该系统从新闻报道中抽取信息，内容涉及地震、工人罢工等很多领域或场景。系统采用了期望驱动与数据驱动相结合的处理方法。后来，这种方法被许多信息抽取系统所吸纳。

20 世纪 80 年代末期到 90 年代是信息抽取研究的中期发展阶段。这一时期出现了一个对于信息抽取发展具有里程碑意义的研讨会 —— 消息理解会议 (Message Understanding Conference, MUC)。MUC 会议由美国国防高级研究计划

委员会(DARPA, The Defense Advanced Research Projects Agency)资助，从1987年到1998年共举办了七届。MUC系列会议对于信息抽取这一研究方向的确立及发展起了极大的推动作用。每届的MUC会议吸引若干学术机构来参加信息抽取竞赛，从第一届的只有6个系统到最后一届的18个系统，MUC的影响越来越大。总的看来MUC会议的主要贡献有两项：

（1）信息抽取任务的确立。从第二届MUC会议开始，信息抽取的任务被明确为模板填充，主要是对某一事件或场景中的关键信息进行填充。以后各届模板变的越来越复杂。1995年的第六届MUC会议在原有场景模板的基础上又加入了命名实体识别、共指关系确定和模板元素填充三项新的任务。在最后一届又增加了模板关系任务。这五项任务的确立指明了信息抽取研究的具体对象，使信息抽取的研究逐步走向规范。

（2）信息抽取评价体系的确立。这是MUC会议的另一重大贡献，参加信息抽取竞赛的每个单位按给定的知识领域提交一个信息系统，然后使用相同的测试数据集对这些系统的性能进行测试比较。第三届MUC会议引入了信息检索领域中的评价指标：召回率、准确率和F-Measure，并利用这些指标对信息抽取系统的性能进行打分。测试方法及评价体系的确立使得对信息抽取效果的评价更加客观和公正，成为信息抽取研究事实上的标准。

进入21世纪后，信息抽取研究又达到了新的高度，进入了信息抽取研究的第三个阶段。这个时期信息抽取研究的重点发生了转移，开始关注新的研究方法和研究内容。如基于机器学习的信息抽取技术、深层理解技术、篇章分析技术、多语言文本处理能力、基于Web的信息抽取以及对时间信息的处理等等[15]。这一时期的一个重要会议是美国国家标准技术研究所（NIST）组织的自动内容抽取（ACE, Automatic Content Extraction）评测会议。与MUC相比，ACE在任务和评测两个方面进行了改变，ACE将MUC定义的五种任务进行了合并，将命名实体和共指合并为“实体检测和识别（Entity Detection and Recognition, EDR）”，将模板元素和模板关系合并为“实体关系检测和识别（Relation Detection and Recognition, RDR）”，场景模板任务改名为“事件检测和识别（Event Detection and Recognition, VDR）”。另外增加了时间短语表达和数量值的识别任务。在评测方面，ACE采用基于漏报（标准答案中有而系统输出中没有）和误报（标准答案中没有而系统输出中有）的评价体系，还对系统跨文档处理（Cross-document processing）能力进行评测。

2.3 信息抽取的研究内容

根据 MUC 和 ACE 定义，可以将信息抽取任务分为四类：命名实体识别、实体关系抽取、共指分析和事件识别。下面我们分别从这四个方面详细介绍所涉及的信息抽取技术。

2.3.1 命名实体识别

命名实体识别（Name Entity Recognition, NER）的任务是识别出文本中有意义的专有名词或数量词，是其它信息提取任务、问答系统、句法分析、机器翻译和面向语义网的元数据标注等应用领域的基础性工作。ACE 定义的命名实体抽取任务分为七大类：人名（Person）、机构名（Organization）、地名（Location）、设备名（Facility）、武器名（Weapon）、交通工具名（Vehicle）和地理政治实体（Geo-Political Entity）。每一大类又分为若干小类，如机构名可分为政符类（Government）、商业类（Commercial）、教育类（Educational）等。其它的实体抽取还有日期、时间、货币、百分比等。目前，中文命名实体抽取主要集中在人名、地名和机构名上。

命名识体识别一般分为两个步骤：实体边界的识别和实体类型的确定。英文命名实体的边界比较容易识别，难点是实体类型的确定。相对于英文，中文命名实体没有明显的边界标志，识别难度较大。而且中文命名实体识别还涉及分词问题，分词工具的性能直接影响实体识别的效果。常用的识别方法有基于统计学习的方法、基于规则的方法和两者混合的方法。

2.3.2 实体关系抽取

实体关系抽取（Entity Relation Extraction）任务是对两个实体间存在的某种关系进行确定。关系抽取结果已经是一种简单的知识，在很多领域具有极高的应用价值，如自动问答系统、检索系统、本体学习、语义网标注任务等。ACE 定义的的关系抽取任务分为六大类：物理关系（Physical）、部分-整体关系（Part-Whole）、人与社会关系（Personal-Social）、组织机构从属关系（ORG-Affiliation）、代理关系（Agent-Artifact）、地理位置隶属关系（Gen-Affiliation）。每个大类同样又分为若干小类，如物理关系为包括了位于关系（Located）和邻近关系（Near）。

一般来讲，关系抽取是指二元关系抽取，即两个实体间关系的确定。多元关系可通过二元关系来表示。关系抽取的一般也分为两个步骤：实体的识别和关系类型的确定。因此实体识别的结果直接影响关系抽取的结果。如果实体不能够被识别或类型识别错误，都会造成关系抽取的错误结果。实体关系抽取的重点在关

系类型的确定，常采用的方法有基于模式匹配的方法和机器学习的方法。相对于英语，中文关系抽取的研究较少，还处于探索阶段。

2.3.3 指代消解

指代消解（Anaphora Resolution）任务是确定篇章中的一个语言单位（通常是词或短语）与之前出现的语言单位存在的特殊语义关联性。指代，也称为提及，是自然语言表达中一种常见的语言现象，常见的提及有命名性提及、名词性提及和代词性提及。在 ACE 定义的抽取任务中将指代消解作为命名实体识别的一部分。

目前，指代消解方面的研究主要集中在对实体的代词（人称代词、指示代词等）提及和名词性提及的指代消解，主要采用是基于句法语义结构分析的规则方法进行指代消解。

2.3.4 事件识别

事件识别（Event Recognition）任务是识别一个事件中的关键信息，如事件发生的时间、地点，事件中所涉及的人物及其它信息。事件识别建立在其它三项识别任务的基础上，识别技术最复杂，难度也最大。目前，ACE 的事件识别任务分为五类：交互类（Interaction）、行动类（Movement）、迁移类（Transfer）、创建类（Creation）和破坏类（Destruction）。事件抽取面临着很多难题，如修饰短语的最大化识别、关键词抽取、事件区分、事件共指分析、浅层句法分析等。这些难题都在不同程度上影响着事件识别的效果。中文事件识别方面的研究比较少，基本上仍采用模板填充的方法进行事件抽取，领域和文本类型均比较受限。

2.4 信息抽取系统

Hobbs 在 1993 提出一个信息抽取系统的通用体系结构[16]，认为典型的信息抽取系统应当由文本分块、预处理、过滤、预分析、分析、片段组合、语义解释、词汇消歧、共指消解、模板生成等十个模块组成。针对不同应用的信息抽取系统可能在模块构成上与该体系结构有所不同，但在功能上就该相同。美国纽约大学在 MUC-5 上参与竞赛的 PROTUES 系统[17]被认为是具有代表意义的信息抽取系统体系结构，如图 1.1 的所示。

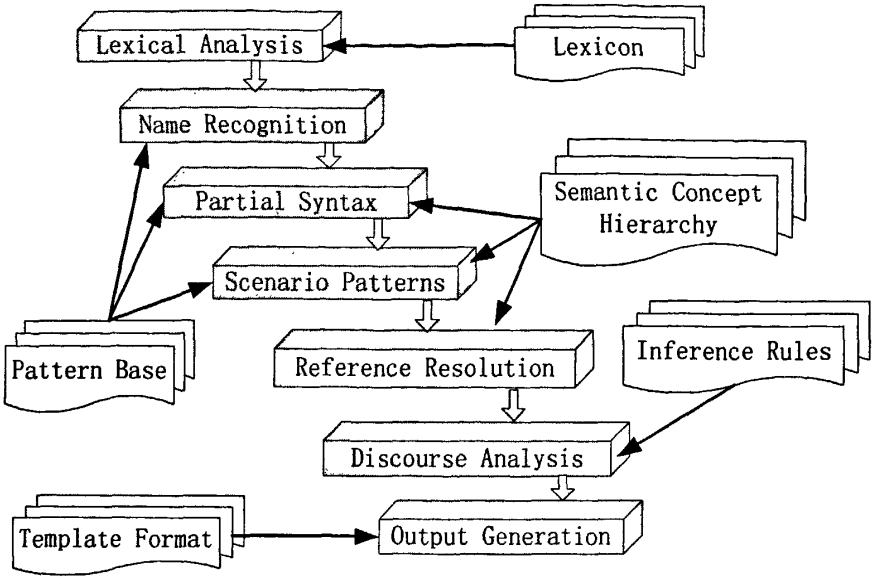


图 1.1 PROTUES 信息抽取系统的体系结构

目前，国外的信息抽取系统的研究比较深入，很多系统已经能在商业应用中发挥作用。相比而言，国内的研究还处于起步阶段，距实用还有一段距离。

2.4.1 国外信息抽取系统

本节介绍四个国外主要的信息抽取系统。

（1）ShopBot [18]

ShopBot 是一个价格比较代理系统，用于从卖家网站上抽取产品信息，并将产品信息按价格进行排序。ShopBot 主要对以表单形式提供查询的网页进行处理，搜索得到的产品信息保存为表格形式的页面，然后综合运用启发式搜索、模式匹配和归纳式学习等技术从这些页面中抽取信息。

ShopBot 的信息抽取过程分两个阶段：离线学习阶段和在线比价阶段。在学习阶段，系统分析每个购物网站，对查询结果页面的格式进行学习，如头部、正文和尾部，并获得其符号化描述。在比价阶段，利用获得的符号化描述，从相应的购物网站上抽取产品信息，并对获得的产品信息按价格进行排序。

（2）STALKER[19]

STALKER 系统用来抽取网页中的餐厅信息，如餐厅名称，菜肴种类、价格、烹调方法，餐厅地址、电话及评价信息。算法采用有指导的学习算法归纳抽取规则。训练样本由用户提供，并对页面中有用的数据进行标注。系统利用标注好的

页面生成一个符号序列，用来表示页面的内容。另外，还生成代表信息点开始的符号索引。符号系列（字或 HTML 标签）和通配符被作为定位标志，用于找到页面上的有用数据。

在 STALKER 系统中，网页文档采用所谓的“内嵌目录（Embedded Catalog）”的形式表示。内嵌目录是一个树形结构，根节点表示整个文档，内部节点表示同构的信息点列表或异构信息点元组，叶子结点表示用户需要抽取的信息，其中任一节点的内容代表其父节点内容的一个接续。

（3）SRV[20]

SRV（Sequence Rules with Validation）是一种自上而下的关系型信息抽取算法。其输入是一个网页集，且每个网页中标记了待抽取区域的实例以及一系列基于字串的特征，输出是一组抽取规则。 SRV 用于讲座信息的抽取任务，如演讲者、地点、时间等信息。

SRV 把信息抽取问题转化为一个分类问题。将文本中所有可能的最长短语看作实例，并针对实例提取特征。SRV 采用的特征分为两类：简单特征和关系特征。简单特征主要有字词的长度、类型、拼写及词性等；关系特征是字词的邻近度。最终，系统为每个短语分配一个测量值，用来反映该短语作为目标槽填充子的可信度。

（4）RAPIER [21]

RAPIER（Robust Automated Production of Information Extraction Rules）系统用于招聘广告中的信息抽取，如职位名称、工资、地点等。它以半结构化文本为处理对象，从中学习抽取规则。系统以待抽取信息的“文档 — 填充模板”对作为训练集，从中获取模式匹配规则。抽取时按获取的抽取规则抽取出“填充子”，用以填充模板中的空槽。

RAPIER 学习算法综合了多个归纳逻辑编程系统所采用的技巧，能够学习无界限模式。这些模式考虑了对词的限制条件和填充子邻近词的词性。学习算法是一个从具体到一般（即自下而上）的搜索过程，从训练中与目标槽匹配的最具体的规则开始。然后从规则库中随机抽取一对规则，并进行横向搜索，以图找到对这两条规则的最佳概括。搜索过程中，采用最少概括的概括方法，增加限制条件，并不断重复直到不再有进展为止。

3.4.2 国内信息抽取系统

中文信息抽取方面的研究起步较晚，目前主要的研究工作集中在命名实体识别方面，且主要是研究人名、地名、机构名三类命名实体的识别技术，出现了几个原型系统。关系抽取方面的研究近几年也有所增加，但还属于起步阶段。总得

来说中文信息抽取研究还处于探索阶段，目前还没有出现功能相对完整的中文信息抽取系统。

Intel 中国研究中心的张益民等人在 ACL2000 上演示了他们开发的一个中文命名实体和实体关系抽取系统[22]，该系统采用基于记忆的学习（MBL，Memory-Based Learning）算法获取规则用以抽取命名实体及它们之间的关系。

国立台湾大学的 NTU 系统[23]参加了 MUC-7 中文命名实体识别任务的评测。该系统主要采用基于规则的方法抽取人名、地名、机构名、日期、时间和百分比表达式。

上海交通大学提出了基于混合模型的信息抽取系统[24]，该系统的两个核心部分是文本分类系统和规则库。其中，规则库包含了三个子库：命名实体抽取规则库、歧义消除规则库和切词错误修复规则库。命名实体抽取规则库是个“群山”规则库，它是由来自不同领域的抽取规则组合而成的混合规则库。在进行识别时，首先对给定的文档进行类型识别，然后利用与该类型匹配的规则进行命名实体识别。

以上三个命名实体识别系统都采用基于规则的方法进行命名实体识别。2003 年，中科院计算所张华平等人提出了基于层叠隐马尔可夫模型（HMM）的命名实体识别算法[25]，并形成中文词法分析系统 ICTCLAS（Institute of Computing Technology, Chinese Lexical Analysis System）。该系统将中文分词、词性标注和命名实体识别集成在一起。其中分词和词性标注功能的效果较好，已达到商用水平，但命名实体识别的效果还不是很好，距实用还有一定的差距。此后，运用统计学习的方法进行命名实体识别成为研究热点，国立台湾大学的 Tzong-Han Tsai 等提出了基于最大熵（ME）的命名实体识别系统[26]，中国软件开发环境国家重点实验室的 Hongping Hu 等提出基于双层条件随机场（CRF）的中文命名实体识别算法[27]，中科院模式识别国家实验室的 YouZheng Wu 等提出了统计模型与人类知识相结合的中文命名实体识别方法[28]。

2.5 信息抽取的评测

MUC-3 评测中引入了信息检索领域的评价指标：准确率（Precision）、召回率（Recall）和 F 指标（F-Measure）。此后的研究者一直将这些指标作为评价信息抽取效果的标准，延用至今。

召回率是正确召回数占应召回数的比例，用于反映系统的查全率，见公式 2.1。

$$Recall = \frac{正确召回数}{应召回数} \tag{2.1}$$

准确率是正确召回数占全部召回数的比例，用于反映系统的准确率，见公式2.2。

$$Precision = \frac{正确召回数}{全部召回数} \tag{2.2}$$

在实际应用中，召回率、准确率往往相互矛盾。召回率提高的同时，准确率会有所下降；准确率提高时，召回率又会受到损失。为了兼顾召回率和准确率，研究者又提出了F指标来评价信息抽取系统的整体性能。F指标的计算公式如下：

$$F\text{-}Measure = \frac{(1+\beta^2)\times Precision\times Recall}{\beta^2\times Precision + Recall} \tag{2.3}$$

其中β是一个权重系数。当$\beta=1$时，表示准确率和召回率同样重要，当β小于1时，表示准确率更重要些。评测时可以根据实际需要调节β的大小。

2.6 本章小结

本章对信息抽取的研究工作进行了简要介绍，从信息抽取技术的研究历史、研究内容、目前国内外主要的信息抽取系统以及所采用的技术三个方面进行了讲解，最后引入了信息抽取技术的评价指标。

信息抽取是协助人们从海量信息里获取感兴趣知识点的关键技术。近几年国内外在信息抽取方面的研究工作取得了很大进展，一些信息抽取技术已经相当成熟，如命名实体识别。信息抽取的研究开始面对更复杂的抽取任务（如事件抽取），抽取的对象也更加广泛，不再只是关注一些通常的抽取内容（一般是国际研究组织定义的抽取任务），注意力开始向实用转移。

中文信息抽取技术比较落后，原因是多方面的。首先是起步比较晚，仍处在学习和借鉴国外研究经验的阶段。其次，中文本身固有的特征，使中文的信息抽取与英文相比更加困难。另外，中文信息技术整体水平不高也影响了信息抽取技术向更深的层次发展。值得庆幸的是，近几年，国内的信息抽取研究不断增多，研究水平提高很快，出现了几个重要的信息抽取研究机构：如哈工大的信息检索中心、中科院的计算所和人工智能实验室、上海交大等。相信依靠这么多研究力量的不懈努力，中文信息抽取技术一定会在不久的将来取得突破性进展。

第三章 基于 Web 的职位关系抽取

3.1 引言

职位关系描述了一个人在一个特定的组织机构中所占据的职位。职位关系典型地包含三个元素：机构名、职位名和人名，可被形式化为一个三元组<*O*，*P*，*R*>，其中 *O*、*P*、*R* 分别代表组织机构名、职位名和人名。职位关系抽取的目的就是从自然语言文本或网页中获取职位关系三元组。

在现代市场经济中，职位关系是一种重要的商业信息。企业管理团队和研发团队的组成和变化趋势在一定程度上反映了企业战略动态、市场倾向等信息。如果一个企业能够获取其竞争对手的职位关系信息，对了解竞争对手，达到知己知彼以增强其核心竞争力无疑大有卑益。

目前，专门针对基于网页的职位关系抽取的研究不多。传统的关系抽取方法主要基于命名实体识别来实现。这种方法首先识别出一个句子中的两个命名实体，然后采用基于模式匹配的方法[29-32]或机器学习的方法[33-38]判断两个命名实体是否具有某种关系。这种方法的有效性对命名实体识别的准确性有很强依赖性。由于目前中文组织机构名识别效果不是很理想，传统的关系抽取算法不适合职位关系抽取。

本章提出一种新的基于 Web 的职位关系抽取方法，该方法利用了网页中职位关系的结构特征，从而避免了中文组织机构名识别效果差带来的影响。经过大量观察，我们发现很多职位关系实例以特定的结构出现的网页的一个片断中，即职位关系实例在视觉上以类似于表格或列表的结构出现在网页中。这些结构具有以下特征：（1）每个职位关系实例在网页中占据一行，且仅占据一行。（2）每个职位关系实例包含至少一个特定的分隔符，这些分隔符将职位关系三元素分隔开来。典型的分隔符有空格和破折号。（3）一个片断中的所有关系实例具有相同的结构，也就是说职位关系三元素的相互位置关系相同。（4）除职位关系三元素外，关系实例中包含较少的噪声数据。这使得关系抽取工作变得相对容易，因为可将分隔符看作是职位关系三元素天然的分界线。

本章的主要贡献有以下几点：

（1）提出了基于结构的职位关系抽取方法。相对于传统的基于命名实体的抽取方法，我们的方法不依赖于机构名识别，只需少数机构特征词就能正确抽取出职位关系。

（2）引入了结构化系数和结构化文件片断两个新的概念，并用来描述网页

的结构特征。实验表明两个概念能够有效地捕捉网页的结构特征。

（3）基于一个真实的数据集进行实验，数据集包含了 6028 个通过百度搜索引擎获取的网页。实验结果表明我们算法达到了准确率 96.1%和召回率 87.16%的较好结果。

3.2 关系抽取研究

1998 年，Sergey Brin 提出了 DIPRE（Dual Iterative Pattern Relation Extraction）关系抽取算法[29]，奠定了模式关系抽取算法的基础。图 3.1 显示了 DIPER 关系抽取算法的流程图。首先由种子关系生成关系模式，然后基于关系模式抽取新的关系，得到新关系后，从中选择可信度高的关系作为新种子，再寻找新的模式和新的关系，如此不断迭代，直到没有新的关系或新的模式产生。此后基于模式的关系抽取算法基本上都采用了这个框架，不同之处主要在模式的形式和模式相似度的计算上。如 Eugene Agichtein 提出的 Snowball 算法[30]将 DIPRE 模式中的实体名替换为实体类型，模式的相似度的计算由简单的关键词比较改为关键词权重向量的余弦函数。而李维刚等人[32]针对中文的特殊性，在构建模式时将词性和词距加入到模式中去。

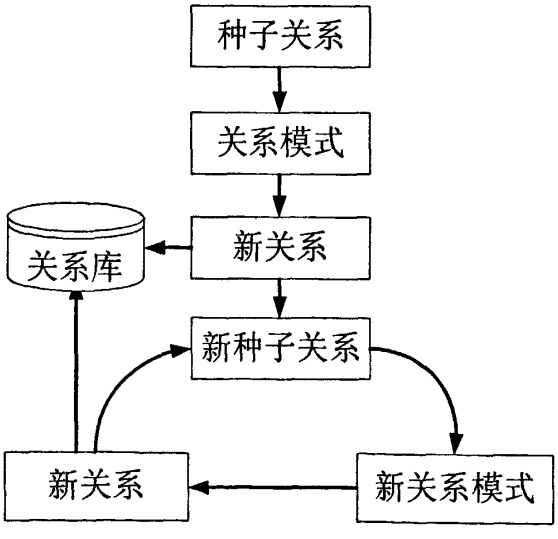


图 3.1 DIPER 关系抽取模型

基于模式的关系抽取算法对结构性较强的网页（如购物网站）有较好的效果。但对普通文本识别效果不太理想。基于机器学习的关系抽取算法由于考虑了更多特征，更适应普通文本的关系抽取。其中基于核函数的关系抽取方法研究较多，

Dmitry 等提出了基于解析树核函数的关系抽取算法[33]，Aron Culotta 等采用了依赖树核函数的关系抽取方法[34]，刘克彬等提出了基于序列核函数的中文实体关系抽取算法[35]，而文献[36, 37]将多种核函数进行组合，提出了多特征融合的核函数关系抽取算法。其它的基于机器学习的关系抽取算法还有基于聚类的方法[38]、半监督学习的方法等[39]。

无论是基于模式的关系抽取算法，还是基于机器学习的关系抽取算法，都涉及相似性计算的问题。在计算相似性时，候选实例中的命名实体必须被标注。当命名实体的类型与所要求的类型不匹配时，相似度被认为是零。由于中文组织机构名识别效果不佳，中文职位关系抽取不能采取基于命名实体的抽取方法。一个完整的中文组织机构名通常以一些特殊的词（如公司，集团等）结尾，我们称这类词为机构特征词。中文组织机构名识别算法对这些词的依赖性很强，然而网页中有许多机构名以简写的形式出现，这些机构名中往往不包含机构特征词，因此被正确识别的概率很小。如“百度网络技术有限公司”通常在网页中以“百度”的形式出现。

目前，国内外专门针对职位关系抽取的研究不多，ACE 定义的雇佣关系与我们的工作比较接近，但它的抽取任务只要求判定一个人与一个组织是否存在雇佣关系，对职位信息没有具体要求。也就是说，雇佣关系抽取的结果是两元组<机构名，人名 >，而不是一个三元组。文献[32]将职位信息看作是雇佣关系的补充说明，但没有进行深入的讨论，而且也没能去除中文组织机构名识别效果差的负面影响。

本章所提的基于 Web 的职位关系抽取方法是一种基于模式的方法，与传统基于模式的关系抽取方法相比，我们方法有三个显著特点：

（1）不需要选择种子关系。

（2）不依赖于中文组织机构名识别。

（3）抽取结果是一个三元组<*O*，*P*，*R*>，而非传统的二元组<*O*，*R*>。

3.3 职位关系抽取框架

职位关系抽取主要由结构化文件片断获取和关系抽取两大模块组成（见图 3.2）。

（1）结构化文件片断获取

模块以一个网页作为输入，输出为一个结构化文件片断或为空。一个结构化文件片断由结构相似的职位关系候选实例组成，通过结构化系数来衡量其结构化程度。当一个文件片断的结构化系数达到某一阈值时，就将其视为一个结构化文

件片断。

（2）关系抽取

模块对获取的结构化文件片断进行处理，输出最终的职位关系三元组。这个过程分为两步聚，首先为每个结构化文件片断生成一个标准模式，然后利用模式匹配的方法抽取出职位关系。

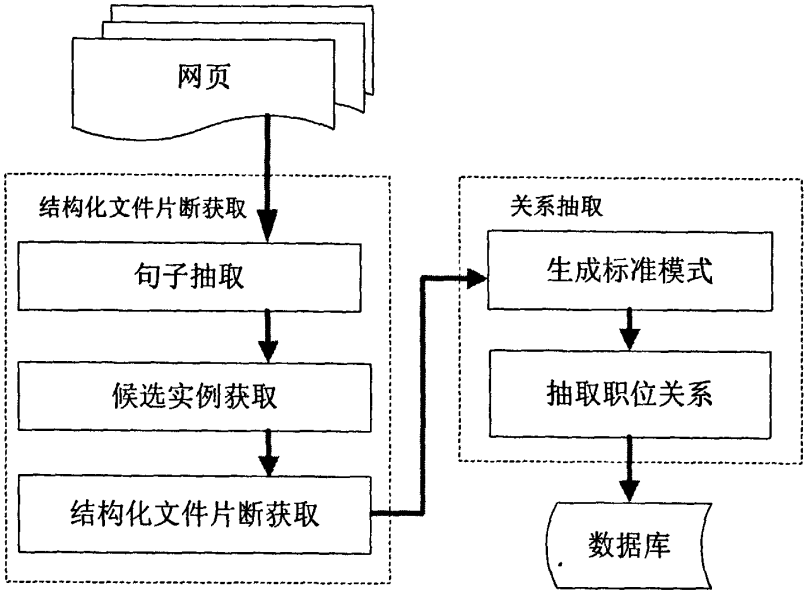


图 3.2 职位关系抽取框架

3.4 获取结构化文件片断

3.4.1 结构化系数和结构化文件片断

职位关系抽取算法涉及一些新的概念，为了便于理解，在介绍具体算法之前，首先给出这些概念的定义。

定义 1 职位关系候选实例

如果一个句子同时包含人名，职位名和分隔符，称这个句子为职位关系候选实例，简称为候选实例。

一个候选实例形如“f_1 —f_2 —... —f_i —... —f_n”，其中“ —”表示分隔符，表 3.1 给出了网页中常见的分隔符。f_i 表示第 i 个字串，这里我们称为 f_i 为段。由于候选实例中段的数目会对抽取结果产生影响，因此设置一个阈值对候选实例中允许的最大段数进行限制，我们称这个阈值为最大段数（Max Field Count， MFC）。当一个句子包含的段数超过 MFC 时，就不能成为候选实例。MFC 对抽取结果的

影响将在实验中讨论。

表 3.1 网页中常见的分隔符

分隔符	名称	例子
	空格	杨宁　空中网总裁
:	冒号	千橡集团总裁：陈一舟
—	破折号	大贺集团董事长 — 贺超兵
<td> ...</td>	表格标签	<td>中华广告网董事长</td><td>姜杉</td>
(...)	括号	雷军（金山总裁）

定义 2 文件片断

文件片断是一个网页中所有候选实例的集合。

定义 3 结构化系数

结构化系数用来衡量一个文件片断的结构化程度， 计算公式如下：

$$sc = \frac{{p_m}^2}{{p_m}^2 + c\sum_{\substack{i=2,\\ i\neq m}}^{mfc} t(p_i)\sum_{\substack{i=2,\\ i\neq m}}^{mfc} p_i} \tag{3.1}$$

其中：

p_i 是段数为 i 的候选实例的数目；

p_m 是所有 p_i 中的最大值，即 $p_m = \max(p_i)$；

mfc 为一个常数，表示候选实例允许包含的最大段数，即 MFC。

当 $p_i = 0$ 时，$t(p_i) = 0$，$p_i > 0$ 时，$t(p_i) = 1$；

c 是一个负影响因子，反映候选实例段数差异对文件片断结构化的负面影响。

结构化系数反映了一个文件片断整齐划一的程度，结构化系数越大，文件片断的结构化程度越高。根据定义，段数为 m 的候选实例决定了文件片断的结构特征。此外，候选实例段数差异也将影响文件片断的结构化程度，公式中通过引入$\Sigma t(P_i)$来表示这一影响。实际上$\Sigma t(P_i)$通过改变影响因子 c 来发挥作用，$\Sigma t(P_i)$越大，负影响因子也越大，文件片断的结构化程度越低。

定义 4 结构化文件片断

如果一个文件片断的结构化系数大于某一给定的阈值α，则这个文件片断为

结构化文件片断。

3.4.2 句子抽取

与传统的关系抽取一样，我们试图从一个单一的句子中抽取职位关系。所以结构化文件片断获取的第一步是句子抽取。由于我们试图从网页中具有特定结构的部分（如表格、列表）抽取职位关系，因而句子抽取方法与一般的基于标点符号进行切割的方法有所不同。网页中这些具有特定结构的部分通常由一些特殊的 HTML 标签分隔成句子，并使这些句子在浏览器中独自占据一行。基于这种观察，我们将网页文本看作一个长字符串，然后根据特定的 HTML 标签对这个长字符串进行切割，从而得到一个句子集。

3.4.3 候选实例获取

由于中文组织机构名识别效果不佳，所以候选实例的定义没有显式要求组织机构名的出现。在 3.4.4 节，将介绍如何判断候选实例中是否包含机构名。此外，一个候选实例必须包含至少一个分隔符，因为它将用来确定职位关系三元素的边界。不同的分隔符虽然形式上不同，但作用相同，因此不会影响抽取结果。故在以后的算法中，将表 3.1 中的分隔符全部转换为破折号（“—”）。图 3.3 显示了由一个句子得到一个候选实例的过程。

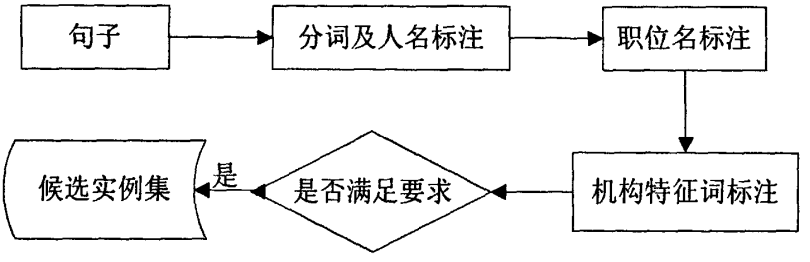


图 3.3 生成候选实例

（1）人名标注

这里我们采用中科院计算所的分词工具 ICTCLAS[41]进行分词和人名标注。然而 ICTCLAS 对网页中某些特殊形式出现的人名不能正确识别，如“杨 皓”，在姓与名之间有一个空格。这里我们设计了两条规则来处理这种以特殊形式出现的人名，称这一过程为人名补全。

规则 1：分词及词性标注后，如果出现形式为“A/*nr1* # B/*x*”或为“A/*nr1*B/*x* — ”的字串，“A”与“B”将合并到一起并标注为“*r*”，即结果为“AB/*r*”。

规则 2：分词及词性标注后，如果出现形式为“A/*nr1* BC/*x*—”或“A/*nr1* B/*x*

C/*x* －”，“A”、“B”与“C”将合并在一起并标注为“*r*”，即结果为“ABC/*r*”。

规则中的“A”、“B”和“C”表示一个中文字符，标签“*nr1*”表示“A”是一个中文姓氏。标签“*x*”表示任一词性标注。标号“#”表示空格，“－”表示分隔符。标签“*r*”表示一个完整的中文人名。

例如子串“杨/*nr1* # 皓/*ng*”根据规则 1 将被转换为“杨皓/*r*”。而子串“蒋/*nr1* 天龙/*nz*”根据规则 2 被转换为“蒋天龙/*r*”。

（2）职位名标注

职位名标注通过人工收集的职位名词典和职位名修饰词典来识别。职位名词典只包含职位名核心词。职位名核心词是一个完整职位名中起表意作用的词，如“副总裁”中的核心词是“总裁”，而“副”是一个修饰词。我们这将些修饰词收集起来建立职位名修饰词典。职位名识别通过前向匹配的方法来实现。这里将职位名标注为“*/p*”。

（3）机构特征词标注

如果一个句子同里包含一个人名、一个职位名和一个机构特征词，那么在机构特征词处是一个机构名的概率就会很大。我们首先假设在这样的机构特征词处存在着一个机构名，然后在后续步骤中判断这个机构名的合法性。这里我们只对最常见的两个机构特征词“公司”和“集团”进行标注，标签为“*/o*”。

表 3.2 给出了算法中使用到的标注，图 3.4 给出了生成的候选实例

表 3.2 候选实例中使用的标注

标注	意义
/r	人名
/p	职位名
/o	机构特征词

句子: 天极公司总裁－李志高
候选实例: 天极 公司*/o* 总裁*/p* － 李志高*/r*

图 3.4 职位关系候选实例示例

3.4.4 结构化文件片断

一个文件片断能否成为一个结构化文件片断是由它的结构化系数来确定的，当一个文件片断的结构化系数超过某一阈值α时，它就是一个结构化文件片断。

同时一个结构化文件片断又要满足另外两个条件：

（1）至少包含两个候选实例。如果一个文件片断只包含一个候选实例，它几乎不可能来自于网页中的结构化部分，因而包含一个职位关系的可能性就很小。

（2）至少有一个候选实例包含机构特征词，即存在标注“*o*”。如果所有候选实例都不包含机构特征词，我们就无法推断出机构名在候选实例中的位置。

这里我们没有考虑候选实例所包含的职位关系三元素在位置上的一致性，这可能会在结构化文件片断中引入错误的候选实例，在 3.5.2 节将介绍如何从结构化文件片断中过滤这些错误的候选实例。

3.5 抽取职位关系

得到结构化文件片断后，分两步抽取职位关系。首先为每一个结构化文件片断生成标准模式，标准模式表示了候选实例中职位关系三元素相对位置关系。然后依据标准模式，从候选实例中抽取出职位关系三元素。

3.5.1 标准模式生成

一个结构化文件片断的标准模式是基于候选实例的职位关系模式生成的。

定义 5 职位关系模式

职位关系模式表示了一个候选实例的基本结构，形式如图 3.5 所示：

$$S(I)\underline{\Delta}\{w \mid w=<s_I^1,...,s_I^i,..,s_I^n> \wedge s_I^i \in \{O,P,R,N,OP,OR,PR,OPR\}\}$$

图 3.5 职位关系模式

其中 I 是一个候选实例，w 是一个有序列表，其每一个元素是一个段类型标志，表示候选实例中对应段的类型。段的类型有 8 种 O、P、R、N、OP、OR、PR、OPR。

O，P，R 和 N 是四个基本段类型，分别代表机构名、职位名、人名和其它信息。四个基本段类型的组合构成其它四种复合段类型 OP，OR，PR 和 OPR。需要注意的是任一段类型与 N 的组合，仍是原类型。段类型按如下规则判断：

（1）如果一个段仅包含标注“*/o*”，段类型为 O。

（2）如果一个段仅包含标注“*/p*”，段类型为 P。

（3）如果一个段仅包含标注“*/r*”，段类型为 R。

（4）如果一个段不包含“/o”、“/p”和“/r”中任一标注，段类型为 *N*。

（5）如果一个段仅包含标注“/o”和“/p”，段类型为 *OP*。

（6）如果一个段仅包含标注“/o”和“/r”，段类型为 *OR*。

（7）如果一个段仅包含标注“/p”和“/r”，段类型为 *PR*。

（8）如果一个段同时包含标注“/o”、“/p”和“/r”，段类型为 *OPR*。

表 3.3 显示一个结构化文件片断中的三个候选实例及它们的职位关系模式。每一个候选实例有四个段，职位关系模式显式了它们的段类型信息。

表 3.3 职位关系模式示例

序号	候选实例	职位关系模式
1	姚忠良/*r* — 男 — 白象 食品集团/*o* — 董事长/*p*	<*R, N, O, P*>
2	杨宁/*r* — 男 — 空中 网 — 总裁/*p*	<*R, N, N, P*>
3	肖志国/*r* — 男 — 路明/*r* 集团/*o* — 董事长/*p*	<*R, N, OR, P*>

在表 3.3 中，第一个职位关系模式正确反映了候选实例中职位关系三元素的位置关系。第二个职位关系模式中缺少机构名信息，而在第三个职位关系模式的第三个段中出现了噪声标志“*R*”，这给确定人名在候选实例中的正确位置造成了干扰。我们期望后两个职位关系模式与第一个相同，为了实现这个目的，我们为每一个结构化文件片断生成一个标准模式。

结构化文件片断的标准模式是文件片断中所有候选实例共享的正确模式，反映了这些候选实例的共同结构。标准模式的存在，可以实现以下几个目的：

（1）推断未知机构名在候选实例中的位置；

（2）消除个别职位关系模式中噪声数据的干扰；

（3）移除结构化文件片断中的噪声候选实例。

经过大量的观察，我们发现机构名、职位名和人名在候选实例中的相对位置关系不是任意的，而是存在一定的规律。根据这些规律，将候选实例分为两种类型，在此基础上给出一个有效的算法来生成标准模式。

（1）双段型候选实例：在这种类型的候选实例中，职位关系三元素隐含在候选实例的两个段中，其中一个段是 *OP* 型、一个段是 *R* 型。候选实例中的其它段是 *N* 型段。例如候选实例“大贺集团/*o* 董事长/*p* — 贺超兵/*r*”是一个双段型的候选实例，它的关系模式是<*OP*，*R*>。所谓“双段型”是指这种候选实例只

有两个有效段，即两个包含职位关系三元素的段。而且段类型分别为 *OP* 型和 *R* 型。其余段都没有包含职位关系信息，因此段类型为 *N*。

（2）三段型候选实例：这种类型的候选实例包含三个有效段，分别包含机构名、职位名和人名。因而这类候选实例的标准模式由 *O*、*P*、*R* 和 *N* 构成。如表 3.3 中第一个候选实例，它的关系模式为<*R*，*N*，*O*，*P*>。

Algorithm: *Generating Standard Schema*

Input: *a structural file segment S*

Output: *F: the Standard Schema of S*

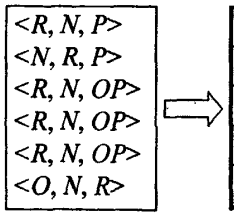
Precondition: *mfc is the number of field types in Standard Schema*

```
1    E ← all the position relation schemas of S;
2    F ← <'N', ..., 'N'>; //total mfc elements with field type 'N'
3    For i ← 1 to mfc  Do //Constructing matrix M[4 × mfc]
4        M[1, i] ← count of "O" in the i-th position in E;
5        M[2, i] ← count of "P" in the i-th position in E;
6        M[3, i] ← count of "R" in the i-th position in E;
7        M[4, i] ← count of "OP" in the i-th position in E;
8    End For
9    v1 ← max(M[1, i], i = 1, 2, ..., mfc); suppose v1 = M[1, p1]
10   v2 ← max(M[2, i], i = 1, 2, ..., mfc); suppose v4= M[2, p2]
11   v3 ← max(M[3, i], i = 1, 2, ..., mfc); suppose v4= M[3, p3]
12   v4 ← max(M[4, i], i = 1, 2, ..., mfc); suppose v4= M[4, p4]
13   If v4 > v1 Then  //double-typed
14       F[p4] ← 'OP';
15       F[p3] ← 'R';
16   Else //triple-typed
17       F[p1] ← 'O'; F[p2] ← 'P'; F[p3] ← 'R';
18   End If
19   Return F;
```

图 3.6 标准模式生成算法

标准模式生成算法如图 3.6 所示。首先为结构化文件片断中的每个候选实例生成关系模式，然后统计每类有效段出现的次数，并将其保存于矩阵 $M[4 \times mfc]$ 中。最后确定候选实例的类型及标准模式各元素的位置。

图 3. 7 显示了算法的工作过程，图 3. 7（a）是由某个结构化文件片断中获得的一组职位关系模式，图 3. 7（b）的矩阵记录了关系模式中不同位置中各有效段出现的次数。因为 *OP* 型出现的最大次数大于 *O* 型段出现的最大次数，因此这些候选实例是双段型的，生成的标准模式为<*R*，*N*，*OP*>。



the matrix *M*[4×3]

	1	2	3
O	1	0	0
P	0	0	2
R	4	1	1
OP	0	0	3

(a) (b)

图 3.7 标准模式生成算法示例

3.5.2 抽取职位关系三元素

从一个候选实例抽取职位关系三元素，首先检查候选实例的关系模式是否与标准模式匹配，如果匹配，则依据标准模式从候选实例的相应位置抽取出职位关系三元素，否则将该候选实例过滤掉。关系模式与标准模式的匹配性主要是看它们的对应元素对是否匹配，因而如何选择对应元素对是匹配成功的关键因素。设 *F* 是标准模式，*S* 为一个职位关系模式，有四种匹配方案可供选择：

（1）左部优先：将 *F* 的元素与 *S* 的元素按前后顺序，顺次组成元素对。这种方法不允许 *S* 中有任何噪声元素，容错能力差。

（2）机构名优先：将 *F* 中包含“*O*”的元素与 *S* 中包含“*O*”的元素组成一个元素对，然后以这两个元素为基准，将剩余的元素组成元素对。当 *S* 中没有包含“*O*”的元素时，这种方法就会失去效用。由于很多候选实例中的机构名没有包含机构特征词，由此生成的关系模式没有包含“*O*”的元素，故这种方法会使最终的召回率降低。

（3）人名优先：将 *F* 中包含“*R*”的元素与 *S* 中包含“*R*”的元素组成一个元素对，然后以这两个元素为基准，将剩余的元素组成元素对。这种方法的缺点是，当 *S* 中包含“*R*”的元素数大于 1 时，无法确定选择其中的那个元素。不少中文机构名中包含人名，这会使关系模式中有多个元素包含“*R*”。

（4）职位名优先：将 *F* 中包含“*P*”的元素与 S 中包含“*P*”的元素组成一个元素对，然后以这两个元素为基准，将剩余的元素组成元素对。相对于以上方法，这种方法的适应性最强，能够克服以上三种方法的缺点。

基于以上分析，我们选择职位名优先的匹配方案来检验关系模式与标准模式

的匹配性。算法流程如图 3. 8 所示，我们只对包含“*P*”和“*R*”的元素对进行匹配性检查，忽略包含“*O*”的元素。这是因为一部分关系模式中没有包含“*O*”的元素。在算法中，首先将标准模式中的“*O*”去掉，即将元素“*O*”或“*OP*”转换为“*N*”或“*P*”。图 3. 9 给出了匹配算法的一个示例。首先将标准模式中的“*OP*”转换为“*P*”，然后以包含“*P*”的元素为基准，组成元素对，检查各元素对是否匹配。根据算法，图中虚线椭圆包围的两个元素组成元素对。

Algorithm *Matching*

Input: *a standard schema F, containing the elements F*[1], ..., *F*[*mfc*]
a position relation schema E: E[1], ..., *E*[*k*], *k* ≤ *mfc*
Output: *Return true if F matched E, otherwise false*
Precondition: *size*(*F*) = *mfc, i.e. F contains mfc elements.*
size(*E*) = *k*, *k* ≤ *mfc.*

```
1    For t ← 1 to mfc Do //replace 'O' in normal form
2        If F[t] = 'O' Then F[t] ← 'N';
3        If F[t]= 'OP' Then F[t] ← 'P';
4    End For
5    i ←  the index of 'P' in F; //F[i] = 'P'
6    j ←  the index of the element containing 'P' in E;
7    f ← i; e ← j;
8    While (f  ≥ 1 and e ≥ 1) Do  //check the left side
9        If F[f] is not contained in E[e]
10           Return false;
11       End If;
12       f—; e—;
13   End While;
14   f ← i; e ← j;
15   While (f  ≤ size(F) and e ≤ size(E)) Do //check the right side
16       If F[f] is not contained in E[e]
17           Return false;
18       End If;
19       f++; e ++;
20   End While;
21   Return true;
```

图 3. 8 关系模式与标准模式匹配算法

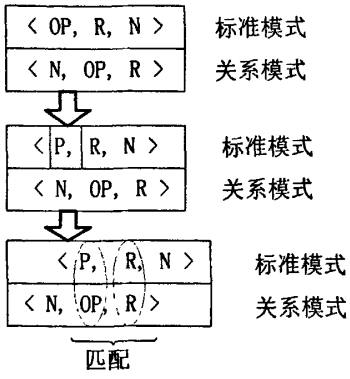


图 3.9 匹配算法示例

当一个候选实例的关系模式与标准模式匹配时，就可以从候选实例中抽取出职位关系三元素。由于关系模式的元素与候选实例的段是一一对应的，只要将关系模式中与标准模式相匹配元素替换成候选实例相对应的段就得到了职位关系三元素。表 3.4 给出了两个例子，在第一个例子中，得到了两个字串："陈一舟 */r*" 和 "千橡 集团/*o* 总裁/*p*"。根据机构特征词对第二个字串再进行切割，可以得到一个职位关系三元组<千橡集团，总裁，陈一舟>。第二个例子的第二个字串中没有包含机构词，我们采用基于词典的前向匹配法得到职位名的前界，使之与机构名分割开来。

表 3.4 职位关系三元素抽取示例

标准模式	***<OP，R>***	
示例 1	候选实例	千橡 集团/*o* 总裁/*p* — 陈一舟/*r*
	关系模式	*<OP，R>*
	职位关系	<千橡集团，总裁，陈一舟>
示例 2	候选实例	盛大 总裁/*p* — 谭群钊/*r*
	关系模式	*<P，R>*
	职位关系	<盛大，总裁，谭群钊>

3.6 实验与分析

3.6.1 实验设置

实验使用的网页集通过百度搜索引擎获取，搜索关键字形式为“职位名 + 中文姓氏”，如“总裁 + 张 | 王 | 李 | 赵 | 刘”。这种方法使得网页中含有职位关系候选实例的概率比较大。共收集网页 6028 个，并选择“总裁”、“经理”、“工程师”、“首席执行官”、“董事长”五种主要的职位关系对我们的算法进行验证。另外，职位名词典采用人工收集的方法，共收录了 66 个职位名核心词。

实验结果由信息检索中常用的三项指标进行评估，分别为准确率、召回率和 F 值。其中 F 值综合考虑了标准率和召回率，公式如下：

$$F-measure=\frac{(1+\beta^2)P*R}{\beta^2*P+R} \tag{3.2}$$

其中 P 是准确率，R 是召回率，β 是一个调节系数。当 β为 1 时，表示准确率和召回率同样重要，当β小于 1 时，表示准确率更重要些。根据实际需要，实验中β 取值 0.5。

为了提高说服力，实验中使用的应召回职位关系数设为结构化系数为 0 时，所有结构化文件片断中包含的职位关系。

我们进行了四项实验来评估所提方法的效果：

（1）人名补全实验

（2）职位关系抽取实验

（3）结构化系数的影响

（4）最大段数的影响

实验中使用处理器为 Pentium (R) 4，主频为 3.0 GHz，内存为 512M 的计算机。操作系统为 Windows XP，算法的实现程序采用 Visual Studio 2005 中 C#语言编写。

3.6.2 实验结果

3.6.2.1 人名补全

在 3.4.3 节，我们使用 ICTCLAS 工具进行人名标注。在一些句子中，ICTCLAS 能够正确识别出姓氏，而不能识别出名。这些句子在后续的候选实例生成过程中会因不包含人名而被过滤掉。但这些句子中有一部分确实包含有职位关系。为了弥补 ICTCLAS 的这种不足，在 3.4.3 节中采用基于规则的人名补全方法来解决这种特殊情形下的人名标注问题。下面的实验验证了这种方法的效果。

实验在 18501 个句子上进行，这些句子由前面提到的 6028 个网页抽取得来。且这些句子经 ICTCLAS 标注后，只包含有中文姓氏的标注“*/nr1*”，而没有名的标注“*/nr*”。采用 3.4.3 节的人名补全规则对这些句子进行处理，并随机选择 1000

个句子作为检验样本。

在 1000 个句子中，经 ICTCLAS 标注后有 713 个句子的姓氏标注是正确的。再经人名补全后，有 438 句子中的人名被标注，其中正确标注的 427 个，准确率达到 97%。相对而言，召回率比较低，将近 60%，但仍然对最终的职位关系召回率有一定的贡献。剩余 287 个姓氏标注不正确的句子，被人名补全的句子有 68 个，这些句子经过候选实例生成规则过滤后，未能成为候选实例，因此并没有对最终结果造成影响。另外，人名补全对职位关系抽取最终结果的影响也比较明显。在最终抽取得到的 28512 个正确的职位关系中，由人名补全得到的有 3194 个，占到 11%，这表明，人名补全对最终的职位关系抽取结果有一定的贡献。

3.6.2.2 职位关系抽取

表 3.5 给出了结构化系数为 0.9，最大段数（MFC）为 7 时的五种职位关系抽取结果。平均准确率超过 96%，召回率超过 87%。可以看出抽取结果的准确率比较高，主要原因是我们的方法利用了网页中职位关系的结构特征。只要能够准确描述这些特征，就能够得到正确的结果。标准模式有效地反应了这些特征，并克服了中文机构名识别难的问题。另外，基于职位名对齐的匹配方法不仅提高了抽取结果的召回率，也将非法候选实例拒之门外。两种方法的结合提高了抽取结果的准确率。召回率相对较低的主要原因是有些网页中包含了多个具有特定结构的片断，这些片断分布在网页中的不同位置，且结构不同，而本文方法只能获取到职位关系实例较多的那个片断。另外，不同职位关系的 F 值存在着一些差别，这主要与不同的职位关系在网页中出现的复杂程度有关。职位关系出现形式越复杂，F 值越低。

表 3.5 五种职位关系抽取结果

	实际召回数	*R*	*P*	*F*
总裁	4991	88.2%	97.0%	95.1%
经理	12490	84.3%	95.1%	92.7%
工程师	2547	88.1%	90.1%	89.7%
董事长	9027	90.1%	98.7%	96.9%
首席执行官	613	84.8%	95.4%	93.1%
总和/平均值	29668	87.2%	96.1%	94.2%

另外，我们采用基于层叠隐马尔可夫模型的方法对数据集中的机构名进行识别，得到召回率为 48%、准确率为 61.8%。机构名识别效果较差的原因是机构名出现在格式化较强的句子中，而不是一般的自然语言句子中。由于传统的基于模

型匹配或机器学习的关系抽取方法依赖于命名实体识别的结果，如果采用传统的关系抽取方法，其抽取结果的召回率和准确率不会超过机构名识别的召回率和准确率，因而职位关系抽取不适宜采用传统的关系抽取方法。

3.6.2.3 结构化系数的影响

文件片断的结构化系数反应了文件片断的结构化水平，较高的结构化系数意味着文件片断中含有较少的噪声。因此结构化系数的阈值越高，抽取结果的准确率会越高，但召回率越低。反之，结构化系数的阈值较低时，准确率会降低，但会提高召回率。为了在准确率和召回率之间达到平衡，必须选择一个合适的结构化系数。

表 3.6 结构化系数对职位关系抽取的影响

α	*Recall*	*Precision*	*F-measure*	*Number of newly Recalled Pos. Relations*	*Precision of newly recalled pos. relations*
1	0.253	0.973	0.620	8498	0.973
0.99	0.847	0.965	0.938	20220	0.961
0.98	0.859	0.963	0.940	463	0.88
0.97	0.864	0.962	0.941	206	0.83
0.96	0.866	0.962	0.941	43	0.907
0.95	0.868	0.962	0.941	87	0.736
0.925	0.870	0.961	0.941	104	0.846
0.9	0.872	0.961	0.942	47	0.894
0.8	0.872	0.961	0.942	40	0.7
0.7	0.873	0.961	0.942	11	0.82
0.6	0.873	0.961	0.942	3	1
0.3	0.873	0.960	0.942	20	0.75
0	0.873	0.960	0.942	0	—

为了了解结构化系数对职位关系抽取综合性能的影响，我们让结构化系数的阈值（α）从 1 到 0 不断递减，并观察准确率、召回率和 F 值的变化。另外，对每次因结构化系数降低而增加职位关系也进行统计。实验结果如表 3.6 所示。当结构化系数为 1 时，召回率只有 0.235，意味着完全不含有噪声的结构化文件片断很少。但当结构化系数稍微降低时，召回率增长很快，准确率只降低一小点。这表明大多数职位关系存在于含有较少噪声的结构化文件片断中。当结构化系数阈值降低到 0.99 后，它对召回率、准确率和 F 值的影响比较小，特别是在 0.9

到 0 之间变化时。原因是新召回的职位关系相对于召回的职位关系总数来说很小。这也表明结构化系数较小的文件片断中只含有少量的职位关系。此外，我们也注意到新召回的职位关系的准确率随着结构化系数的降低而降低。这表明结构化系数较小的文件片断含有较多的噪声，从而带来更多的错误。基于以上分析，比较合适的结构化系数的阈值取在 0.9 到 0.99 之间。

3.6.2.4 最大段数（MFC）的影响

最大段数是为限制一个候选实例允许包含的最多段数。职位关系有三个元素，最多分布在 3 个段中。因此一般情况下，MFC 取 3 就可以了。但有时网页中的职位关系信息与其它不相关信息混合在一起，如性别、年龄等。这样的候选实例往往包含 3 个以上的段，如果 MFC 为 3 时，这些候选实例将被过滤掉，因此 MFC 应大于等于 3。但当这些不相关的信息过多时，会增加关系抽取的复杂性。

图 3.10 显示了结构化系数为 0.9 时，MFC 对召回率、准确率和 F 值的影响。根据图示 MFC 的值应大于 5。

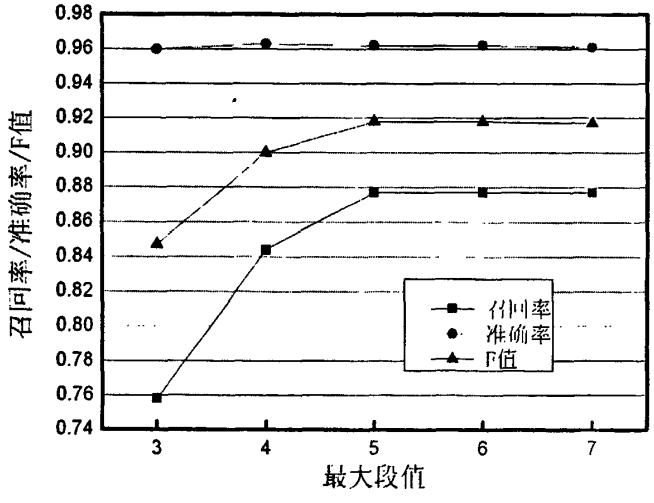


图 3.10 最大段值的影响

3.7 本章小结

职位关系反映了一个人在一个组织所占据的职位，是一种重要的竞争情报。本章基于职位关系在网页中的结构特征，提出一种新的职位关系抽取算法。算法基于职位关系在网页中的结构特征，引入结构化系数和标准模式来描述网页的这种特征，最后利用模式匹配的方法抽取职位关系。实验结果表明算法达到了准确

率超过 96%、召回率超过 87%的较好结果。

虽然本章主要是针对中文职位关系抽取，但所提出的利用网页结构特征抽取商业信息的思想可以用来抽取其它信息。因此，未来的工作将集中于利用网页中的结构特征抽取其它商业信息。

第四章 基于语义隐马尔可夫模型的中文机构名识别

4.1 引言

商业信息中最核心的部分是关于竞争对手的信息。商业中的竞争对手不是个人，而是企业或组织。要从网页中抽取一个企业的相关信息，必须首先识别网页中的组织机构名。组织机构名识别属于信息抽取中命名实体抽取范畴。目前，中文的人名、地名、时间等命名实体抽取研究已取得了一定的效果，但组织机构名的抽取效果不是很理想。原因是中文组织机构名的构成比较复杂，几乎没有规则可寻。

英文机构名的每一个词都以大写字母开始，具有明显的形态特征，一般采用基于规则的方法进行识别。中文机构名没有明显的形态特征，很难判断其前后界，而且中文机构名识别受中文分词结果的影响，因此识别难度较大。目前，中文机构名识别主要采用基于规则的方法和基于统计学习的方法，每种方法各有优缺点，相对而言基于统计学习的方法效果更好一些，但识别效果还是不能令人满意，达不到商用要求。基于语言学中的两个观点，我们提出一种基于语义隐马尔可夫模型（Semantic HMM, SHMM）的机构名识别算法，利用机构名与其上下文之间的语义关联性来提高机构名识别的效果。

4.2 机构名识别研究

英文机构名的每个词都以大写字母开头，且词之间都以空格分隔。这些特征使英文机构名识别比较容易，一般采用基于规则的识别方法。与英文相比，中文机构名没有以上这些特征，英文机构名识别方法不适用于中文机构名的识别，中文机构名识别必须探索适合中文特征的识别方法。因此下面主要对中文机构名的研究现状进行简单介绍。

在中文信息抽取领域，研究最多的就是命名实体识别。对于组织机构名识别，既有独立进行研究的，也有将机构名与其它命名实体看作一类，共同研究的，采用的方法主要有两种：基于规则的方法[22, 23, 24, 42]和基于统计学习的方法[25, 26, 27, 43-50]。基于规则的方法在机构名识别研究的早期使用的比较多，主要是借鉴了国外命名实体识别的经验。这种方法一般将机构名分为前部和后部两个部分，后部称为机构特征词，如“公司”、“集团”等，前部是修饰词。基于规则的方法试图对机构名的前部分进行分析，总结出若干规则，形成规则库。识别时首

先根据机构特征词确定出机构名的后部，然后根据规则库识别机构名的前部。然而机构名前部构成复杂，人工寻找规则的难度较大，而且容易引入错误。随着机器学习技术的发展，一些研究者开始探索基于统计学习的机构名识别技术。目前，基于各种统计学习模型的机构名识别研究已经比较多，主要有基于支持向量机（SVM）的方法[43]、最大熵模型（MEM）的方法[26,44,45]、隐马尔可夫模型（HMM）的方法[25,46,47]和条件随机场（CRF）的方法[27,48,49,50]等。这些方法首先对学习语料进行人工标注，然后对标注后的语料进行学习以得到相应的模型。统计学习的方法实际上是用机器去自动寻找规则，避免了人工寻找规则的缺点，可以找到普适性较强的模型。但这种方法有两个问题需要解决：(1)标注集的选取。标注集选取合适与否直接影响机构名识别的召回率和准确率。目前标注集选取方法主要对组成机构名的词进行分类，每一类分配一个标注。对机构名所处的上下文语境关注不足。(2) 普适性问题。虽然基于统计学习的方法比基于规则的方法普适性强，但它只局限于与学习语料类型相同的文本。对于不同类型的文本需要建立不同的模型。

本文提出一种新的基于语义隐马尔可夫模型(SHMM)的机构名识别算法。这种算法是对文献[47]所提算法的扩展。文献[47]提出一种层叠隐马尔可夫模型的机构名识别算法，首先在底层抽取人名、地名，然后在高层抽取机构名。由于算法对机构名的内部组成进行了细分，提高了机构名识别的效果，但其关注点主要是机构名本身的组成，对于机构名上下文的信息关注不多，只关心机构名前一个词及后一个词。我们的算法基于语言学中语法对语义的依赖关系和共生性词场现象，分析了机构名与其上下文之间的语义关联性，对机构名及其上下文同时进行语义分类标注，从而构建语义隐马尔可夫模型。新标注类型反映了机构名与其上下文之间的语义关联性，借助这种语义关联性来提高机构名识别的召回率和准确率。此外，机构名上下文信息的加入提高了算法的普适性，使得模型对不同类型的文本同样具有较好的识别效果。

4.3 语言学中的两个观点

信息抽取技术与语言学有着密切的关系，语言学研究的一些成果为信息抽取技术提供理论依据。这里简要介绍语言学中的两个观点，是本章所提的语义隐马尔可夫模型的语言学理论基础。

（1）语义决定语法，语法表现语义。

认知功能语言学派的学者认为，语法和语义是密不可分的，语法是词语内容的结构化，语义在很大程度上决定语法 [51] 。

武汉大学教授赵世举的观点则更为明确，他认为意义是语言的出发点和落脚点，在表达时，先有“意”，然后具“意”择“形”，而“形”则随“意”转；在理解时，则是据“形”索“意”[52]。也就是说语义决定语法形式，语法形式表现语义。赵教授还提出语言的这种特性在中文中表现的尤为明显。

根据上述观点，一个句子是一个词的有序序列，表现的是一个语法结构，其背后则隐藏着一个词的语义序列，这个序列决定了词的语法序列。词的语法序列和语义序列可以分别对应到隐马尔可夫模型的观察值序列和状态序列。

（2）共生性词场[52]

词汇是构成句子的基本单位，词汇语义角色决定词的组配方式或组配框架，一个词能与哪些词组配，不能与哪些词组配，都取决于词汇的语义。如“中国通信设备商华为”中的“设备商”具有商业组织语义角色，由它修饰的“华为”应该是一个企业名。

共生性词场是一个词汇系统，在这个系统中，词的组配方式由核心词的语义角色决定。具体讲共生性词场就是以核心词所表事物的属性和功用为基础而自然形成的绕核共生的词团，场中的词之间是客观存在的，与生俱来的血肉关系。共生性词场中的各项要素具有相对稳定的数目和层次顺序。这表现为以核心词为中心，伴生着性质各异、亲疏有别、分层的词团。核心词为名词的词场如图 4. 1 所示。

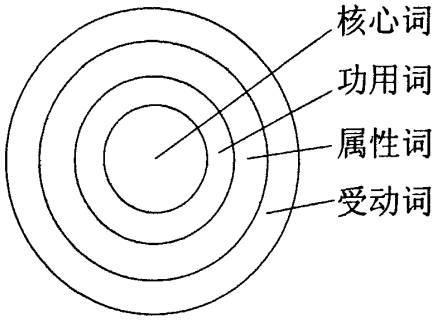


图 4. 1 核心词为名词的词场一般状态

共生性词场现象说明，距核心词越近的词，与核心词的关系越密切。在句子“中国通信设备商华为”中，“华为”是核心词，“设备商”是功用词，与“华为”的关系最密切，而“中国”和“通信”则是属性词。

4.4 机构名语义环境及分类

4.4.1 机构名及其上下文语境分析

机构名的组成比较复杂，可以由人名、地名、常见机构名、机构特征词和其它词组成。另外，机构名中未登录词一般比较多，各类词之间的组配也没有统一的规律可寻，这导致仅以机构名本身的组成为基础来识别机构名的效果比较差。

与机构名相比，机构名的上下文语境有着另外的特征。基于共生性词场现象，经过大量的观察分析，我们发现机构名的上下文存在两个特点：

(1) 机构名提示信息多。机构名的上下文往往隐含着大量的提示信息，这些信息显示它们所包围的字串是一个机构名。这符合词汇的语义角色决定词的组配方式的语言学观点。如字串“目前搜狐畅游所运营《天龙八部》”中的“运营”，“通信设备商华为”中的“设备商”等。很多时候，人的认知过程就是利用这些提示信息来识别机构名。这里我们同样试图利用这些提示信息来协助识别机构名。

(2) 内容和形式稳定。尽管机构名的组成变化不定，但其上下文在内容和形式上都比较稳定，存在一定的规律。如机构名在作主语时，常与谓语动词“成立”、“发布”、“收购”等搭配，而在作状语时多与“在”，“对于”等词一起使用。这也符合共生性词场现象，即词场中的各要素按性质功能、亲疏远近形成数目一定的层状词团。

机构名与其上下文的这种语义关联性为利用机构名上下文语境协助识别机构名提供了可能。本章正是基于这种观点，对语料中的机构名及其上文进行分类，并进行语义标注，对标注后的语料进行学习，得到机构名识别的语义隐马尔可夫模型。

4.4.2 机构名及其上下文分类

我们将机构名的组成成份划分为五类：地名、人名、常见机构名、机构特征词和其它词。其中人名、地名由中科院 ICTCLAS 工具进行标注。常见机构名和机构特征词采用基于词典的方法进行识别。文献[47]也采用了同样的方法对机构名的组成进行分类和标注，这里不再详述。

机构名所在的句子可能很长，如果对其全部的上下文进行分类标注，需要为上下文设置较多类型，容易引入错误，所以需要为机构名的上下文确定一个界限。常用的设定界限方法是数值窗口，按窗口的大小，取机构名前后若干词作为上下文。但这种方法只考虑词的数目，割裂了机构名与其上下文在语义上的连贯性。经过观察，我们发现机构名在句子中一般作主语、宾语、状语和定语。根据共生性词场理论，机构名在作主语、宾语和状语时一般扮演核心词角色，而在作定语

时，一般作为属性词，修饰核心词。以机构名所在的共生性词场作为机构名上下文的界限比较合理，因为词场中的词与机构名的关系最为密切，能为识别机构名提供最丰富的提示信息。

共生性词场是根据词的语义关系确定的，不方便问题的描述。根据 4.3 节讲述的语法和语义的关系，我们利用语法块（Syntactic Piece, SP）来描述共生性词场。语法块是句子中语法相对独立的一个子串，如句子中的主语、谓语、宾语等。机构名所在的共生性词场由机构名所在的语法块或相邻语法块构成。构成规则为：如果机构名在句子中作主语或宾语的中心词，且与之相关的谓语动词所在的语法块与机构名所在的语法块相邻，则机构名所在的共生性词场由机构名所在的语法块和谓语动词所在的语法块构成；在其它情况下，机构名所在的共生性词场由机构名所在的语法块构成。图 4.2 中的句子被划分为四个语法块，机构名“华为”所在的共生性词场由第一个语法块（SP1）组成，机构名的上下文由“中国”、“通信”、“设备”、“商”四个词组成。以后我们就用语法块代替共生性词场来描述机构名的上下文。

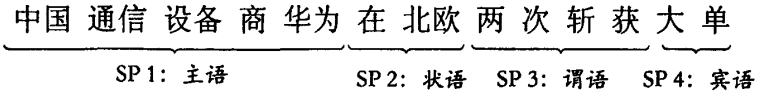


图 4.2 语法块划分图

根据上述分析，将机构名的上下文按其与机构名之间的语义关联性分为以下几类：

（1）修饰词：用来修饰机构名的词，这类词在主语、宾语、状语和定语中都可能会出现，图 4.2 中的修饰词“中国”、“通信”、“设备”、“商”出现在主语中。修饰词只能来自于机构名所在的语法块。修饰词包含了共生性词场中所涉及的功能词和属性词，这里没有进行细分的原因与中文分词有关。在图 4.2 给出的例子中，“设备商”是功能词，但经分词后，“设备商”被划分为“设备”和“商”两个词，每个词都不能称为功能词。这种现象在中文分词中大量存在，所以这里将功能词和属性词不作区分，统称为修饰词。

（2）中心词：机构名所修饰的词，这种词一般出现在定语中。如“华为 市场 份额”中的“市场”和“份额”。中心词也只能取机构名所在的语法块。

（3）谓语动词：机构名在作主语或宾语中心语时，如果其所在语法块与对应的谓语所在语法块相邻，则谓语动词成为机构名上下文。如“诺基亚 终于 发布 了 其 第一 款 TD 产品”中的“发布”。谓语动词就是图 4.1 中的受动词，作为表达句意的关键词，谓语动词与动作的发出者主语和动作的承受者宾语有着

密切的联系。不同的谓语动词要与相匹配的主语和宾语搭配，反之亦然。所以与机构名搭配的谓语动词对识别机构名有很大的启发作用。

（4）主谓之间的词：机构名作主语时，如果谓语动词可以成为机构名的上下文，则主谓之间的词也作为机构名上下文中的一种类型。如下例中的“终于”

（5）谓宾之间的词：机构名作宾语时，如果谓语动词可以成为机构名的上下文，则主谓之间的词也作为机构名上下文中的一种类型。如“中国联通 联合 了 中国 电信”中的“了”。

（6）介词：机构名在作状语时，有时由一些介词引导，这些介词也构成机构名上下文的一种类型。如“在 央 视 广告 招标 中”的“在”。

表 4. 1 显示了本文用到所有类型标注，包括了机构名及其上下文的类型标注。另外，我们发现机构名与职位名共同出现的频率比较高，是识别机构名的一个重要信息。因此将职位名作为一类特定的信息，以提高对机构名识别的贡献。

表 4.1 机构名及上下文类型标注

标注	类型	示例
F	机构特征词	北京搜狐畅游时代网络技术*有限公司*
R	机构名中的人名	法国*马蒂尼埃*集团
NR	其它人名	*林增伟*提出了“蓄水池”的概念
S	机构名中的地名	*北京市*文化局相关领导表示
NS	其它地名	在前不久的*中国*游戏行业年会上
O	常见机构名	*中国人民银行*
E	机构名中的其它词	侵犯*腾讯*公司相关游戏著作权一案
L	机构名之间的连接词	中国移动*和*中国联通慢慢掌控了很多版权
M	修饰词	*国内知名厂商*长虹
C	中心词	华为*市场份额*
W	谓语动词	诺基亚终于*发布*了其第一 款 TD 产品
N	主谓之间的词	诺基亚*终于*发布了其第一 款 TD 产品
K	谓宾之间的词	中国联通联合*了*中国电信
J	介词	*在*央视广告招标中
P	职位名	友达*董事长*李焜耀
B	机构名上文的前一个词	瑞典*正是*爱立信的总部所在地。
A	机构名下文的后一个词	北京市文化局相关领导*表示*
Z	其它词	

4.5 基于语义隐马尔可夫模型的机构名识别算法

4.5.1 隐马尔可夫模型

隐马尔可夫模型（HMM）是计算语言学中经常用到的一个统计学习模型，一般常用的是一阶隐马尔可夫模型（见图 4. 3）。隐马尔可夫模型包含了两个随机过程，一个是马尔可夫过程，描述状态之间的转移，用转移概率表示，对于一阶隐马尔可夫模型，它表示当前状态只与前一个状态有关，与其它因素无关。另一个是一般的随机过程，描述状态与观察值之间的关系，用观察值概率表示。HMM 的状态是不确定的或不可见的，只有通过观察值序列的随机过程才能表现出来。可以用一个五元组表示隐马尔可夫模型：HMM = { Ω_Q, Ω_O, Π, A, B}。其中 Ω_Q 是状态集合、Ω_O 是观测值集合，Π是状态的初始分布矩阵，其元素 π_i 表示初始时状态为 Q_i 的概率，A 是状态转移概率矩阵，其元素 a_{ij} 表示从状态 Q_i 转移到状态 Q_j 的概率，B 是观察值的分布矩阵，其元素 b_{ik} 表示状态为 Q_i 时，观察值为 O_k 的概率。

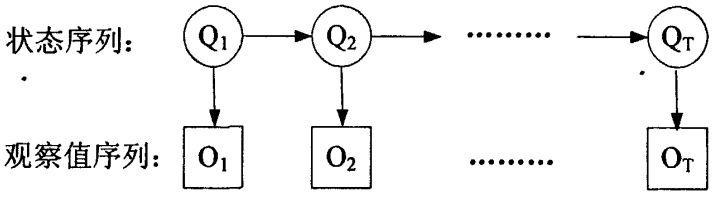


图 4. 3 一阶隐马尔可夫模型

隐马尔可夫模型常用来解决三类问题：

（1）评估问题：给定模型，求某个观察值序列的概率。

（2）解码问题：给定模型和观察值序列，求可能性最大的状态序列。

（3）学习问题：给定一个观察值序列，调整模型参数，使观察值序列出现的概率最大。

计算语言学中常用 HMM 模型来解决解码问题，如词性标注、命名实体识别等。基于语义隐马尔可夫模型的机构名识别也属于这类问题。给定 HMM 模型和一个观察值序列<O_1, O_2, ..., O_T>，出现概率最大的状态序列<Q_{i1}, Q_{i2}, ... , Q_{iT}>就是我们要找的答案。设这个概率为 P，则 P 可表示为：

$$P = (\pi_{i1} b_{i1,1})(a_{i1,i2} b_{i2,2}) \ldots (a_{i(T-1),iT}, b_{iT,T}) \tag{4.1}$$

从公式 4.1 可以看出，概率 P 与一系列观察值概率和状态转移概率有关，这

实质上反映了观察值序列各元素之间的相互关系。当一组观察值按照某些规则组成一个序列时，它隐含了一个状态序列，这个状态序列反映了观察值序列形成的规则。隐马尔可夫模型就是利用了观察值序列和状态序列的这种关系来解决实际问题的。

4.5.2 语义隐马尔可夫模型

语义隐马尔可夫模型中的“语义”是指模型中的状态序列是一个语义序列，即表 4.1 列出的标注组成的序列。表 4.1 所示的标注集是基于共生性词场原理，在机构名与其上下文之间语义关联性分析的基础上得到的类型标注，是一种语义标注集。而现有的基于统计学习的机构名识别算法一般采用 BIEO 标注法对机构名本身进行标注，即 B 标注机构名的首词，E 标注机构名的尾词，I 标注组成机构名内部的词，O 对其它词进行标注。这种标注法只考虑了机构名内部的组成，割裂了机构名作为组织机构这样一种语义角色与其上下文之间的语义关联性。根据 4.4.1 节的分析，机构名上下文包含了丰富的对识别机构名具有启发作用的信息，这些信息与机构名之间具有一种天然的语义关联性，可以利用这些语义关联性来改善机构名识别效果。我们所提的分类标注就是为了反映机构名与其上下文之间的语义关联性。

图 4.4 显示了一个语义标注示例，由词序列和对应的标注序列构成。根据共生性词场原理，我们利用机构名所在的共生性词场来确定机构名的上下文。在图 4.4 中，机构名“华为”所在的共生性词场由“中国”、“通信”、“设备”、“商”和“华为”五个词组成，其中前四个词构成机构名的上文，用于修饰机构名，故标注为“M”。这个例子中机构名没有下文。词序列和标注序列的关系反映了 4.3 节所讲的语言学中语法与语义间的相互关系，即语义决定语法，语法表现语义。具体到我们的应用就是标注序列决定词序列，而词序列表现标注序列。语义隐马尔可夫模型就是基于词序列和标注序列间的这种语法和语义关系而构建的，其中词序列作为隐马尔可夫模型的观察值序列，标注序列作为状态序列。给定语义隐马尔可夫模型和和一个词序列，可以计算出一个可能性最大的标注序列，根据标注序列识别机构名。

词序列：	中国	通信	设备	商	华为	在	北欧	两	次	斩	获	大	单
标注序列：	M	M	M	M	E	A	Z	Z	Z	Z	Z	Z	Z

图 4.4 句子标注示例

4.5.3 机构名识别算法

基于语义隐可尔可夫模型的机构名识别算法的实质就是利用机构名与其上下文之间的语义关联性来识别机构名，识别过程如图 4.5 所示。识别过程分为学习和识别两个部分，这与一般的隐马尔可夫模型建立与识别过程相同，不再详述。这里讲一下主要的不同之处：预处理和人工标注。

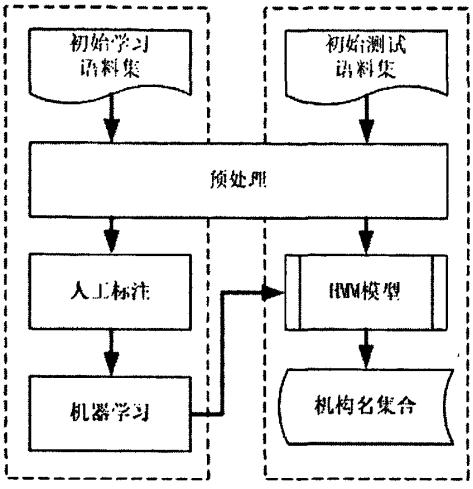


图 4.5 基于语义隐马尔可夫模型的机构名识别流程图

预处理主要包括分词，人名、地名、常用机构名和职位名的识别。其中分词，人名、地名的识别由中科院计算所的分词工具 ICTCLAS 处理。常用机构名由一个词典库识别，职位名采用 3.4.3 节介绍的方法识别。识别后将人名、地名、常用机构名和职位名进行转换，转换规则见图 4.6。表 4.2 是命名实体转换示例。

人名 ---> 〈PER〉， 地名 ---> 〈LOC〉

常用机构名---> 〈ORG〉， 职位名---> 〈POS〉

图 4.6 命名实体转换规则

表 4.2 命名实体转换示例

转换前	转换后
北京 文化局 相关 领导 表示	**<LOC>** 文化局 相关 领导 表示
艾瑞 咨询 研究 认为	**<NR>** 咨询 研究 认为
中国移动 广东 公司	**<ORG>** 广东 公司
大唐 电信集团 总工程师	大唐 电信集团 **<POS>**

学习语料经过预处理后，就要进行人工标注。按 4.4.2 节描述的上下文类型和表 4.1 的类型标注对预处理后的句子进行人工标注。表 4.3 列出了四个人工标注后的示例，黑体部分显示了机构名及其上下文。

表 4.3 人工标注语料示例

示例 1	消息/Z 称/B 央/E 视/E 新/C 媒体/C 业务/C 3/A 年/Z 要/Z 占/Z 总收入/Z 25%/Z 以上/Z
示例 2	**<LOC>/S 搜狐/E 畅游/E 时代/E 网络/E 技术/E 有限公司/F <POS>/C <PER>/C** 告诉/A 记者/Z
示例 3	**<LOC>/M 通信/M 设备/M 商/M 华为/E 将/N 持续/N 推动/W** LTE/A 产业/Z 发展/Z
示例 4	华为/E 再/N 获/W <LOC>/A 大/Z 单/Z ，/B 在/J 爱立信/E 老巢/C 击败/A 对手/Z

学习语料经人工标注后，就可以用于训练隐马尔可夫模型了。得到模型后，采用韦特比算法进行机构名识别。

4.6 实验与分析

由于我们的目的是为了从网页中抽取商业信息，所以实验数据集来自 Internet 上的真实网页。表 4.4 显示了实验中使用的数据集，所有数据都来自于腾讯网(http://www.qq.com)。其中 Data-1 来自 318 个网页，共 5649 个句子；Data-2 由来自 368 个网页的 6568 个句子随机选择 1000 个句子构成；Data-3 由来自 2209 个网页的 47364 个句子随机选择 1000 个构成。Data-1 作为训练集，而 Data-2 和 Data-3 用作测试集，其中 Data-3 是为了检验学习器对类型不同的语料的识别效果。

表 4.4 实验数据集

名称	句子数	机构名数	来源
Data-1	5649	1793	腾讯网 2009 年 12 月 21 日科技版
Data-2	1000	703	腾讯网 2009 年 12 月 22 日科技版
Data-3	1000	517	腾讯网 2009 年 12 月 21 日新闻版

实验结果的评价方法采用信息抽取中常用的召回率(R)、准确率(P)和 F 指标(F)。为了验证算法的效果，将我们的方法与以下三种方法进行了比较：

（1）Y-HMM：俞鸿魁等提出的基于层叠隐马尔可夫模型的机构名识别算法[47]。

（2）Y-MEM：杨华提出的基于最大熵模型的机构名识别算法[45]。

（3）C-SVM：陈霄等提出的基于支持向量机的机构名识别算法[43]。

另外，为了验证所提方法的普适性，我们利用由科技语料（Data-1）学习得到的隐马模型对新闻语料(Data-3)进行了测试。实验结果如表 4.5 所示。

表 4.5 中文机构名识别效果比较

		Data-1	Data-2	Data-3
SHMM	R	85.6	73.28	68.37
	P	91.62	82.84	64.96
	F	88.51	77.77	66.62
Y-HMM	R	79.54	68.28	58.33
	P	89.95	79.47	55.7
	F	84.38	73.45	56.98
Y-MEM	R	86.52	62.47	51.79
	P	93.35	84.86	69.53
	F	89.81	71.96	59.36
C-SVM	R	67.42	52.47	43.6
	P	72.38	62.4	53.93
	F	69.81	57.01	48.22

实验结果表明，与其它方法相比，我们的方法取得更好的效果，尤其是在新闻语料（Data-3）的实验中，无论是召回率还是准确率都有明显的提高。这表明我们的方法对类型不同的语料有更好的适应能力。在数据集 Data-1 的实验中，基于 Y-MEM 方法的 F 值略高于我们的方法，这是因为 Y-MEM 方法在进行特征选取时，选择了较多的组合特征，使其对已知的机构名（学习语料中的机构中）识别效果较好，但对未知的机构名识别效果较差，这导致 Y-MEM 在数据集 Data-2 和 Datat-3 上的召回率急剧下降。在数据集 Data-2 的实验中，我们方法的优越性主要表现在召回率上，表明利用机构名与其上下文之间的语义关联信息可以改善机构名识别效果，证明了我们的设想。在数据集 Data-3 的实验中，我们的方法取得了较好的效果。这主要是在类型不同的语料中，组成机构名的未登录词增多，其它方法只专注于构成机构名的内部词，所以对不同类型的语料识别效果不佳。我们的方法既考虑了机构名的组成，也顾及了机构名的上下文信息。由于机构名

上下文信息的变化相对较小，所以当未登录词增多时，这些上下文信息就能协助识别机构名。

另外，我们注意到 Y-MEM 方法的准确率一直比较高，这是因为 Y-MEM 方法的召回率比较低，所识别的机构名中大部分是已知的机构名，故准确率相对较高。基于 C-SVM 方法的识别效果较差，原因是 C-SVM 方法将标注问题转化为分类问题，类别就是 4.5.2 节提到的 B、I、E 和 O 四类。而训练样本中 O 类的样本比重很大，导致分类器学习效果比较差。而且在实验中，我们发现由于 SVM 将 n 分类问题转化为多个 n*（n-1）个二值分类问题，使得识别的时间复杂性较高。Y-HMM 方法在数据集 Data-1 和 Data-2 的实验中与我们的方法相差不多，而在数据集 Data-3 上的 F 值与我们的方法相差近 10 个百分点。这表明我们在 Y-HMM 方法的基础上加入机构名上下文信息，可以提高机构名的识别效果，更重要的是我们的方法具有更强的普适性。

从表 4.5 可以看出，我们方法对召回率的改善比较明显，但对准确率的提高不是很大，原因是机构名的上下文信息使得召回了一些伪机构名，即召回的机构名不是一个真正的机构名。如句子“汪洋书记给我们题词说”中的“汪洋”被错误识别为机构名。这也说明不合适的机构名上下文信息及不完善的类型标注集也会对机构名识别带来负面影响。总的来说，机构名上下文信息对识别机构名有很大的辅助作用，但要慎重选择。另外，从召回率、准确率的绝对数值上来看，几种方法的效果都不是很好，这可能是由于我们的语料来自于网页的原因。网页中的语言表述没有书面表述规范，增加了模型学习和识别的复杂性。

4.7 本章小结

本章提出了一种新的基于语义隐马尔可夫模型的机构名识别算法。算法考虑了机构名与其上下文之间的语义关联，对机构名的上下文进行了分类标注。借助机构名上下文语境来提高机构名识别的召回率和准确率。实验表明，与传统只关注机构名本身组成的方法相比，我们的方法取得了更好的效果。

机构名识别是抽取其它商业信息的基础，因此机构名识别的效果直接影响到其它商业信息的抽取。面对复杂多样的网页文本，我们的方法虽然取得了比传统方法更好的效果，但从抽取结果本身来看还不是很理想。如何进一步提高网页中机构名识别的效果是下一步研究的重点。

第五章 总结与展望

5.1 本文工作总结

在商品经济高度发达的今天，市场竞争更加激烈和残酷，企业对竞争情报的需求如饥似渴。信息爆炸的时代背景下，商业信息抽取作为企业竞争情报获取过程中的重要一环，对及时、准确地形成竞争情报的影响越来越大。

Internet 上巨量而丰富的信息使其成为商业信息获取的重要源泉。本文研究了基于 Internet 的商业信息抽取，以 Internet 上的中文网页为信息源，从中抽取商业信息。主要工作如下：

（1）提出了基于 Web 的职位关系抽取算法。职位关系反映了一个人在一个组织所占据的职位，是一种重要的竞争情报。本文分析了职位关系在网页中特殊的表现形式，提出一种新的职位关系抽取算法。通过调查，我们发现大量的职位关系实例以列表或表格的形式集中出现在网页的某一个区域中，而且这些职位关系实例具有相同的结构。职位关系抽取算法基于职位关系在网页中的这些特征，利用结构化系数对这些特征进行描述，并据此获取网页中的职位关系实例块——结构化文件片段。最后利用模式匹配的方法从结构化文件片断中抽取出职位关系。我们利用百度搜索引擎获取了 6028 个中文网页作为实验的语料集，验证算法的召回率与准确率。实验结果表明算法达到了准确率超过 96%、召回率超过 87%的较好结果。

（2）提出了基于语义隐马尔可夫模型的中文机构名识别算法。算法以语言学中语法对语义的依赖关系和共生性词场为理论依据，将一个句子划分为若干语法块，并以语法块为单位，构建机构名的上下文语境，借助机构名与其上下文之间的语义关联性来识别机构名。中文机构名识别困难的一个重要原因是组成机构名的词复杂多样，而且未登录词较多。而机构名的上下文相对比较稳定，变化较少，而且与机构名之间存在语义上的关联性。当构成机构名的未登录词比较多时，机构名的上下文对识别机构名的作用就会变的特别突出。这也符合人类的认知规律，当面对一个生词时，我们需要借助上下文的信息来判断该生词的类型。本文正是基于这样的认识，同时考虑机构名和及其上下文，将机构名及其上下文划分为若干类，以此构建语义隐马尔可夫模型。实验表明我们的算法达到了较好的效果，特别是对类型不同的文本具有更好的适应能力。

5.2 下一步工作展望

为了获取有效的竞争情报，需要抽取很多种商业信息，每种商业信息都有自身的特点，可能需要采用不同的抽取方法。另外，中文信息抽取技术涉及很多其它技术，如中文分词、句法分析、统计学习模型等，这些技术的发展，可能带来信息抽取技术的变革。基于这些考虑，对未来工作作了如下安排：

（1） 探索其它商业信息的抽取方法

本文针对商业信息中的职位关系和机构名识别进行了研究，还有很多其它重要的商业信息有待抽取，如客户关系、供应关系、产品信息等。这涉及两个问题，网页收集方法和具体的信息抽取算法。互联网上的网页数以亿计，如何找到包含待抽取信息的网页是首先面临的难题。借助搜索引擎技术是一种方法，但必须考虑查全率。得到网页集后，就需要设计具体的信息抽取算法，需要对所抽取信息及其上下文环境进行分析，设计有效的实现算法。

（2） 商业信息抽取方法的集成

具体的商业信息很多，为每一种商业信息设计相应的抽取算法显然不太现实，需要设计能够适应多种商业信息抽取的技术方法。在对几种重要的商业信息抽取技术研究的基础上，可以对商业信息及所采用的抽取方法进行分析比较，找到技术上的突破口，设计出少数几个典型的商业信息抽取算法实现多数商业信息的抽取。

（3） 设计基于本体的商业信息抽取系统

在多种商业信息抽取技术实现后，接下来就需要将这些抽取技术进行整合，构建商业信息抽取系统。本体是描述实体、实体属性及实体关系的有效技术，已在应用到多种专用领域的信息抽取系统中。可以利用本体对待抽取的商业信息进行表示，并将相应的抽取技术集成的本体中，实现一个基于本体的商业信息抽取系统。

参考文献

[1] 中国信息产业网 (http://www.cnii.com.cn), 2010.

[2] 中国互联网络中心 (http://www.cnnic.net.cn), 2010.

[3] 包昌火, 谢新洲. 企业竞争情报系统[M]. 北京: 华夏出版社, 2002.

[4] 中国企业监测网（http://www.cichina.com.cn）,2010.

[5] Yan chen, Peiquan Jin, Lihua Yue. Ontology-driven Extraction of Enterprise Competitive Intelligence in Internet[C]. In Proceeding of the 2008 Second International Conference on Future Generation Communication and Networking Symposia, 2008, 35-38.

[6] Byron Marshall, Daniel McDonald, Hsinchun Chen et al. Collecting and Analyzing Business Intelligence Information[J]. Journal of the American Society for Information Science and Technology, 2004, 55(10):873-891.

[7] Francois Paradis, Jian-Yun Nie, Arman Tajarobi. Discovery of Business Opportunities on the Internet with Information Extraction[C]. In Workshop on Multi-Agent Information Retrieval and Recommender Systems (IJCAI), 2005, 47-54.

[8] Robert Baumgartner, Oliver Frlich, Georg Gottlob et al. Web data extraction for business Intelligence: the lixto approach[C]. In Proceeding of BTW 2005, 2005, 48-65.

[9] Diana Maynard, Milena Yankova, Alexandros Kourakis et al. Ontology-based Information Extraction for Market Monitoring and Technology Watch[C]. In Proceedings of the Workshop on User Aspects of the Semantic Web (UserSWeb) at European Semantic Web Conference (ESWC 2005), 2005, 1-10.

[10] Diana Maynard, Horacio Saggion, Milena Yankova et al. Natural Language Technology for Information Integration in Business Intelligence[C]. In W. Abramowicz, editor, 10th International Conference on Business Information Systems, Poland, 2007, 25-27.

[11] 刘非凡, 赵军, 吕碧波. 面向商务信息抽取的产品命名实体识别研究[J]. 中文信息学报, 2005, 20(1):7-13.

[12] Wei Li, Jianyi Liu, Yixin Zhong et al. C-CIS: A Chinese Competitive Intelligence System Based on the Internet[C]. 2006 IEEE International conference on Industrial Informatics, 2006, 715-719.

[13] Naomi Sager. Natural Language Information processing: A Computer Grammar of English and Its Applications [M]. Massachusetts USA: Addison-Wesley Longman Publishing Co., 1981.

[14] G Dejong. An Overview of the FRUMP System[C]. In W.G. Lehnert and M. H. Ringle, editors, Strategies for Natural Language Parsing, Lawrence Erlbaum, 1982, 149-176.

[15] 李保利, 陈玉忠, 俞士汶. 信息抽取研究综述[J]. 计算机工程与应用 2003,39(10):1-6.

[16] J Hobbs. The Genetic Information Extraction System[C]. In Proceedings of the Fifth Message

Understanding Conference(MUC-5), Morgan Kaufman, 1993, 87-91.

[17] Ralph Grishman, John Sterling. Description of the PROTEUS System as used for MUC-5[C]. In Proceeding of the Fifth Message Understanding Conference(MUC-5), 1993.

[18] R. B. Doorenbos, O. Etzioni, D. S. Weld. A Scalable Comparison-Shopping Agent for the World-Wide-Web[C]. In Proceedings of the first International Conference on Autonomous Agents, California, February 1997, 39-48.

[19] Muslea Ion, Minton Steve, Knoblock Craig. A hierarchical Approach to Wrapper Induction[C]. In Proceeding of third International Conference on Autonomous agents,1998, 190-197.

[20] Freitag Dayne. Information Extraction from HTML: Application of a General Learning approach[C]. In Proceeding of the 15th Conference on Artificial Intelligence, 1998, 517-523.

[21] CaliH Mary Elaine, Mooney Raymond. Relational Learning of Pattern-Match Rules for Information Extraction[C]. In Proceedings of the National Conference on Artificial Intelligence, 1999, 328-334.

[22] YImin Zhang, Zhou Joe F. A Trainable Method for Extracting Chinese Entity Names and Their Relations[C]. In proceeding of the Second Chinese Language Processing Workshop, Hong Kong, 2000, 66-72.

[23] Hsin-His Chen, Ding YungWei, Chung Shih. Description of the NTU System Used for MET2[C]. In Proceeding of the Seventh Message Understanding Conference, 1998.

[24] 王睿,张洁,张由仪等. 基于混合模型的中文命名实体抽取系统[J]. 清华大学学报(自然科学版), 2005, 45(S1): 1908-1914.

[25] Huaping Zhang, Qun Liu, Xueqi Cheng et al. Chinese Lexical analysis using hierarchical hidden Markov model[C]. Sapporo Japan, 2003:63-70.

[26] Tzong-Huan Tsai, Shih-Hung Wu, Cheng-Wei Lee et al. Muncius: A Chinese Named Entity Recognizier Uising the Maximum Entropy-based Hybird Model[J]. International Journal of Computational Linguistics and Chinese Language Processing, 2004, 9(1) : 65-82.

[27] Hongping Hu, Hiu Zhang. Chinese Named Entity Recognition with CRFs: Two Levels[C]. In Proceedings – 2008 International Conference on Computational Intelligence and Security, CIS 2008, 2008, 1-6.

[28] Youzheng Wu, Jun Zhao, Bo Xu. Chinese Named Entity Recognition Combing a statistical Model with Human Knowledge[C]. In proceeding of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 2005, 427-434.

[29] Sergey Brin. Extracting Patterns and Relations from the World-Wide Web[C]. In Proceedings of the 1998 International Workshop on the Web and Databases (WebDB'98), 1998, 172-183.

[30] Eugene Agichtein, Luis Gravano. Snowball: Extracting Relations from Large Plain-text Collections[C]. In Proceedings of the Fifth ACM International Conference on Digital Libraries, 2000, 85-94.

[31] Deepak Ravichandran, Eduard Hovy. Learning Surface Text Patterns for a Question Answering System[C]. In Proceedings of the ACL Conference, 2002, 41-47.

[32] 李维刚, 刘挺, 李生. 基于网络挖掘的实体关系元组自动获取. 电子学报, 2007, 35(11): 2111-2116.

[33] Dmitry Zelenko, Chinatsu Aone, Anthony Richardella. Kernel Methods for Relation Extraction[J]. Journal of Machine Learning Research, 2003, 3(6) : 1083-1166.

[34] Aron Culotta, Jeffrey Sorensen. Dependency Tree Kernels for Relation Extraction[C]. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 2004, 423-429.

[35] 刘克彬,李芳, 刘磊等. 基于核函数中文关系自动抽取系统的实现[C]. 计算机研究与发展, 2007, 44(8) : 1406-1411.

[36] Shubin Zhao, Ralph Grishman. Extracting Relations with Integrated Information Using Kernel Methods[C]. In Proceedings of the 43rd Annual Meeting of the ACL, 2005, 419-426.

[37] Claudio Giuliano, Alberto Lavelli, Daniele Pighin et al. FBK-IRST: Kernel Methods for Semantic Relation Extraction[C]. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), 2007, 141-144.

[38] Takaaki Hasegawa, Satoshi Sekine, Ralph Grishman. Discovering Relations among Named Entities from Large Corpora[C]. IN Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics(ACL-04), 2004, 415-422.

[39] Min Zhang, Jian Su, Danmei Wang et al. Discovering Relations between Named Entities from a Large Raw Corpus Using Tree Similarity-based Clustering[C]. Lecture Notes in Computer Science (LNCS 3651), 2005, 378-389.

[40] Seokhwan Kim. Minwoo Jeong, Gary Geunbae Lee et al. An Alignment-based Approach to Semi-supervised Relation Extraction Including Multiple Arguments[C]. AIRS, LNCS vol.4993, 2008, 526-536.

[41] ICTCLAS, Available at http://www.ictclas.org (accessed in May 18, 2009)

[42] 张小衡, 王玲玲. 中文机构名称的识别与分析[J]. 中文信息学报, 1999,11(4):21-31.

[43] 陈霄, 刘慧, 陈玉泉. 基于支持向量机方法的中文组织机构名的识别[J]. 计算机应用研究, 2008, 25(2):362-365.

[44] 冯冲, 陈肇雄, 黄河燕等. 采用主动学习策略的组织机构名识别[J]. 小型微型计算机系统, 2006, 27(4):710-714.

[45] 杨华. 基于最大熵模型的中文命名实体识别方法研究[D]. 哈尔滨工程大学, 2008.

[46] 郑家恒, 张辉. 基于HMM的中国组织机构名自动识别[J]. 计算机应用, 2002, 22(11):1-3.

[47] 俞鸿魁, 张华平, 刘群等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2): 87-94.

[48] 周俊生, 戴新宇, 尹存燕等. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006,34(5): 804-809.

[49] 沈嘉懿, 李芳, 徐飞玉. 中文组织机构名称与简称的识别[J]. 中文信息学报,2007,21(6):17-21.

[50] 冯元勇,孙乐,张大鲲等. 基于小规模尾字特征的中文命名实体识别研究[J]. 电子学报, 2008, 36(9):1833-1838.

[51] 白解红, 石毓智. 从认知心理学的角度看语义和语法的关系[J]. 湖南师范大学社会科学学报, 2006,25(5):99-104.

[52] 赵世举. 试论词汇语义对语法的决定作用[J]. 武汉大学学报, 2008, 61(2): 173-179.

致 谢

时光飞逝，转眼即将毕业，科大三年的学习生活仍历历在目。感谢科大为我的学习与成长提供了一个美好的环境！感谢科大的全体教职工和同学们，正是他们的热情和关爱，让我为成为科大的一员而倍感荣幸和骄傲。感谢计算机学院的所有老师和同学们，是他们的支持和帮助让我在攀登知识高峰的过程中感到无限的乐趣！

由衷感谢我的导师金培权副教授！他渊博的知识和对研究工作的执着与热爱使我受益匪浅，时刻督促我不断进步；而论文写作过程中一遍又一遍的耐心指导，让我倍感温暖的同时，也让我体验到他人格的伟大。他授于我的不仅是分析问题和解决问题的能力，还有为人处世的真理。再次对金老师表示深深的谢意，祝他及他的家人平安幸福。

感谢岳丽华教授！她对实验室同学无微不至的关怀让我们倍感温馨；对工作认真负责的态度让我由衷敬佩。岳老师对我的研究工作给予了很多的关心和指导，让我取得了很大的进步。感谢万寿红老师！每一次的实验室活动都是她精心策划的结果，使我们三年的学习生活充满了乐趣。

感谢知识与数据工程实验室的所有同学们，特别是电四楼420的张奇、田明辉、秦富童、赵旭剑、艾国红、汪启伟、李筱文、陈鸿、夏瑜、汪娜，还有已经毕业的任真、张旭、陈艳、孙逸雪等同学。他们在学习和生活上都给予了我很多帮助，在实验室如同在家一样，温暖无外不在。衷心祝愿实验室的所有兄弟姐妹们每天都幸福快乐！

感谢我的父母！是他们给予了我生命，教会我作人的道理，培育我成长，也是因为他们的支持和关爱，才有了我的今天。

最后，谨向百忙之中抽出宝贵时间评审论文的老师们表示深深的谢意！

在读期间发表的学术论文与取得的研究成果

已发表论文：

[1] Yanhong Liu, Peiquan Jin, Lihua Yue. Extracting Position Relations from the Web[C]. In Proceedings of International Conference on Information and Knowledge Management 2009. 2009 :59-62.