# 摘要

WWW 的飞速发展使其已成为全球信息传播与共享的重要平台，并成为人们获取信息的主要来源。但是随着信息量的激增，要想从 WWW 上获取一条有用信息的难度却越来越大。人们期望着一种理想情况的出现：像查询数据库一样地查询 WWW 上的信息。

Web 信息抽取技术正是随着这样的需求而出现并不断丰富的，而各种抽取技术的侧重点不同导致了抽取系统在精确度、可扩展性、适应性方面不能都令人满意。本课题较好地解决了基于自然语言理解的方式在处理半结构化文本时的不足，改进了现有的语言模型，并在此基础上实现了一个 Web 招聘信息抽取系统——JobHunter。

JobHunter 的实现如下：首先，构建 Spider，“爬行”WWW 上的若干招聘网站并抓取网页；然后，由基于自然语言理解的信息抽取模块将 Spider 抓取的网页抽取成结构化信息并存入数据库；最后，将用户所关注的招聘信息清楚地显示在界面上。

由于 JobHunter 基于自然语言理解方式进行信息抽取，可以从任何类型的网站抽取招聘信息，所以有着良好的可扩展性和适应性。经测试，本系统抽取准确率和召回率都达到 70%以上。

**关键词：** Web 信息抽取，自然语言理解，Spider，命名实体识别

# Abstract

With the quickly development of WWW, it has become the important platform of transmitting and sharing information all over the world. It's out of question that the Internet has become the primary source for people to get the information they needed. But the fact is that the difficulty of getting useful information is growing rapidly while the explosion of the data appears on the WWW. Ideally, people can query the information on the WWW just like a database.

For satisfy such the needs, web information extraction appeared and become abundant, but they cann't get high score at each aspect such as accuracy, extensibility, adaptability and so on. My research subject sloves the drawbacks on processing the half-structure text by Natural Language Understanding method and improves the existing language model.Based on this, the author design and development a web recruitment information extraction system called JobHunter.

The extraction processes are as follows. Firstly, construct a Spider to snatch the web pages from some employ sites. And then extract the employment information and saved to the database by information extraction model. Lastly, display the information extracted to the job hunters at the interface.

The system has a good extensibility and adaptability because it based on the natural language understanding method, and precision and recall can reach above 70%.

**Keywords:** Web Information Extraction, Natural Language Understanding, Spider, Name Entity Recognition

## 西北大学学位论文知识产权声明书

本人完全了解学校有关保护知识产权的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属于西北大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版。本人允许论文被查阅和借阅。学校可以将本学位论文的全部或部分内容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和汇编本学位论文。同时，本人保证，毕业后结合学位论文研究课题再撰写的文章一律注明作者单位为西北大学。

保密论文待解密后适用本声明。

学位论文作者签名：孟华蓉 指导教师签名：张蕾

07 年 6月18日

---

## 西北大学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。据我所知，除了文中特别加以标注和致谢的地方外，本论文不包含其他人已经发表或撰写过的研究成果，也不包含为获得西北大学或其它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：孟华蓉

07年 6月18日

# 第一章 引言

## 1.1 研究背景

在日益信息化和网络化的今天，如何找到所需要的信息并把有用的信息归类、过滤或提取出来，一直以来都是一个备受关注的实际问题。相应地，各种帮助人们查找、分类和存储信息的理论、技术、应用工具和系统始终在不断地发展和更新。近年来，一种叫做信息抽取(Information extraction,IE)的技术逐渐受到了人们的关注，它的提出和兴起有着特定的时代背景。20世纪80年代后期，美国政府提出了一个专门的文本处理研究计划——Tipster计划，其内容包括信息抽取、文档检索和文献摘要等，以期提高政府部门的信息处理速度和质量。该计划的一个重要的目标就是研究和实现文本信息的自动查找、收集汇总和存储，以便将人们从大量的、低效的文本阅读劳动中解放出来。

信息抽取的任务是把文本里包含的信息进行结构化处理，变成类似表格一样结构化的组织形式。输入信息抽取系统的是原始文本，输出的则是固定格式的信息点。

对于从大量文档中抽取所需要的特定事实来说，信息抽取技术非常有用。在WWW上，同一主题的信息通常分散存放在不同网站上，表现的形式也各不相同。若能将这些信息收集在一起，用结构化形式储存，那将是有益的，所以Web信息抽取技术就成为当前的一个研究热点[1]。

同时，WWW所具有的海量、异构、动态等特性也给Web信息抽取研究带来了挑战。首先，WWW是一个巨大的信息空间，Web页面数以几十亿计，而且仍在以几何级数增长[2]，如何自动高效地处理海量的Web信息就成为一个难点；其次，Web页面的异构性(即同一主题的信息分散在不同语种、组织形式各异的Web页面中)使得如何在这些异构的网页里准确识别所需要的信息变得更加困难；最后，WWW是一个动态的空间，网站的页面格式和内容瞬息万变，如何保持Web信息抽取的动态适应性也是一个有待解决的问题。

Web信息抽取系统可以看作是把Web信息从不同文档中转换成数据库记录的系统。因此，成功的Web信息抽取系统将把互联网变成巨大的数据库，它为海量Web信息的再利用提供了可能，有着明显的优势和广阔的应用前景，是当今自然语言处理领域的研究热点。

## 1.2 研究现状

目前，信息抽取的研究重点主要集中在英文领域，日文的研究也有一部分[3]。英文信息抽取在命名实体(TE)和实体关系(TR)识别方面，已经取得相当大的进步，但是在真正的事件抽取(ST)方面，还有许多问题需要探索，而这些问题大多涉及到了自然语言处理中的核心难题。比如，在MUC-7上，SRA公司的3项IE指标均取得了较高的成绩。

中文信息抽取的研究工作开展的较晚，仅有国立台湾大学和新加坡肯特岗数字实验室参加了MUC-7关于命名实体识别的评测。Intel中国研究中心开发了一个抽取命名实体和它们之间关系的信息抽取系统，该系统利用记忆获取规则抽取相关内容。北京大学孙斌采用有限状态自动机进行事件抽取，开发了InfoX信息抽取系统，对人民日报语料库中任职、离职、调职三个事件进行了抽取[4]。总体而言，中文信息抽取的研究主要集中在命名实体识别方面，设计并实现完整的中文信息抽取系统还处于起步探索阶段。

## 1.3 本文的研究内容

Web 信息抽取的一个直接应用就是帮助人们在 WWW 中快速准确地查找所需信息，加快人们获取信息的速度，从而提高工作效率。本着这样一个思想，本选题着眼于当前社会的“找工作”问题，将分散在不同 Web 页面的动态变化的招聘信息抽取出来，以简单明晰的结构显示给找工作者，帮助他们尽快找到称心满意的工作。

本课题采用基于自然语言理解的方式来进行 Web 信息抽取：首先，构建网络蜘蛛(Spider)，“爬行”WWW 上的若干招聘网站并抓取相关网页；然后，由基于自然语言理解的信息抽取模块将 Spider 抓取的网页抽取成结构化信息并存入

数据库；最后，将用户所关注的招聘信息清楚地显示在界面上。

本文的特色主要如下：

1) 解决了基于自然语言理解方式进行 Web 信息抽取时对处理半结构化文本的不足；

2) 改进了现有的语言模型并应用于命名实体识别，取得了较好的识别效果。

## 1.4 本文的结构安排

全文共分五章，各章的内容概括如下：

第一章，研究背景与现状，指出本文的研究内容。

第二章，信息抽取技术。本章从Web信息抽取开始论述，然后针对中文信息抽取中的关键步骤——中文命名实体的识别进行了分析，最后给出信息抽取过程中用到的语言模型。

第三章，Web招聘信息抽取系统设计。本章在对传统的基于自然语言理解的抽取方法以及语言模型改进的基础上进行系统的分析与设计。

第四章，系统实现与评测。本章实现Web信息抽取系统JobHunter，并进行分块和整体测试。

第五章，总结与展望。对全文的工作进行总结，并指出进一步的研究方向。

# 第二章 信息抽取技术

## 2.1 Web 信息抽取

### 2.1.1 Web 信息抽取背景及其分类

从自然语言文本中获取结构化信息的研究最早开始于20世纪60年代中期，它以两个长期的研究性的自然语言处理项目为代表[2]——美国纽约大学开展的Linguistic String项目和耶鲁大学的FRUMP项目。

20世纪80年代末，消息理解系列会议(Message Understanding Conference, MUC)的召开使得信息抽取的研究蓬勃开展起来。信息抽取技术发展成为自然语言处理领域一个重要分支，并一直推动这一领域的研究向前发展。

从1987年到1998年，MUC会议共举行了七届，它由美国国防高级研究计划委员会(the Defense Advanced Research Projects Agency, DARPA)资助[5]。MUC的显著特点并不是会议本身，而在于对信息抽取系统的评测。各届MUC都吸引了许多来自不同学术机构和业界实验室的研究者参加信息抽取系统竞赛。每个参加单位根据预定的知识领域，开发一个信息抽取系统，然后用该系统处理相同的文档库，最后用一个官方的评分系统对结果进行打分。

目前，除了强烈的应用需求外，推动信息抽取研究进一步发展的动力主要来自美国国家标准技术研究所(NIST)组织的自动内容抽取(Automatic Content Extraction , ACE)评测会议[6]。这项评测旨在开发自动内容抽取技术，以支持对三种不同来源(普通文本、由自动语音识别ASR得到的文本、由光学字符识别OCR得到的文本)的语言文本的自动处理；研究的主要内容是自动抽取新闻语料中出现的实体与关系等内容，即对新闻语料中实体与关系的识别和描述。

随着WWW的日益繁荣，信息抽取的研究重点已经逐渐转移到Web信息抽取上来，并涌现出许多算法和系统[7]。其中最知名的研究项目是卡耐基-梅隆大学自动学习和发现中心(Center for Automated Learning and Discovery)的“Web挖掘(Mining the World Wide Web)”项目。该项目的目标是通过从Web中自动提取事实，来创建大型的、结构化的和有用事实的数据库。它们的技术途径是研究机器学习算法，通过训练自动提取信息。

Web信息抽取技术有多种分类方式[1][8][12][17]，根据各种工具所采用的原理不同，可分为4类：基于自然语言理解的方式、基于包装器归纳的方式、基于Ontology的方式和基于HTML结构的方式。

- **基于自然语言理解方式的信息抽取**

自然语言理解技术通常用于自由文本的信息抽取，需要经过的处理步骤包括：句法分析、语义标注、专有对象的识别(如人物、公司)和抽取规则[12]。具体地说就是把文本分割成多个句子，对一个句子的句子成分进行标记，然后将分析好的句子语法结构和事先定制的语言模式(规则)匹配，获得句子的内容。也就是利用子句结构、短语和子句间的关系建立基于语法和语义的抽取规则实现信息抽取。规则可以由人工编制，也可从人工标注的语料库中自动学习获得。这类信息抽取主要适用于源文档中含有大量文本的情况，特别针对于合乎语法的文本。

基于自然语言理解的信息抽取技术是将Web文档视为文本进行处理的，其缺点是[1]：

1) 没有利用Web文档独特于普通文本的层次特性，抽取规则表达能力有限，缺乏健壮性，获得有效的抽取规则需要大量的样本学习，达到全自动的程序较难，而且速度较慢，对于操作网上海量数据来说这是一个大问题。

2) 只支持记录型的语义模式结构，而不支持复杂对象的抽取。

3) 由于Web页面中的文本通常不是结构完整的句子，所以适用范围较窄。

- **基于包装器归纳方式(Wrapper Induction)的信息抽取**

包装器由一系列的抽取规则以及应用这些规则的程序代码组成。通常，一个包装器只能处理一种特定的信息源。从几个不同信息源中抽取信息，需要一系列的包装器程序库。形式化地，每一类Web页面对应一个包装器[1]。

包装器归纳法可以自动分析出待抽取信息在网面中的结构特征并实现抽取，其主要思想是用归纳式学习方法生成抽取规则，该方法由Nicholas Kushmerick于1996年提出[17]。

与自然语言处理方式比较，包装器较少依赖于全面的句子语法分析和分词等复杂的自然语言处理技术，更注重于文本结构和表格格式的分析。使用包装器的

困难在于：

1) 包装器的针对性强，可扩展性 (scalability) 较差。由于一个包装器只能处理一种特定的信息源，所以若从几个不同的信息源中抽取信息，就需要一系列的包装器集，这样使得信息抽取的工作量巨大。

2) 可重用性(reusability)差。包装器对页面结构的依赖性强，当出现一类新的Web页面或旧的页面结构发生了变化后，原来的包装器就会失效，无法从数据源中获得数据或得到错误的数据。这使得一个新的问题出现，即包装器的维护问题。

(3) 缺乏对页面的主动理解。目前的包装器主要依赖于原网页或其后台数据库的模式，基本上是一种数据模式的还原，缺乏对数据语义的主动理解。

- **基于Ontology方式的信息抽取**

按照Stanford AI专家Tom Gruber的定义，Ontology是为了帮助程序和人共享知识的概念化规范，在知识表达和共享领域，Ontology描述了在代理之间的概念和关系(Concepts and Relations)。

基于Ontology的信息抽取主要利用了对数据本身的描述信息实现抽取，对网页结构的依赖较少。由Brigham Yong University开发的信息抽取工具就采用了这种方法。采用该方法，事先要由领域知识专家采用人工的方式书写某一应用领域的Ontology(包括对象的模式信息、常值、关键字的描述信息，其中常值和关键字提供了语义项的描述信息)。根据Ontology中常值和关键字的描述信息产生抽取规则，对每个无结构的文本块进行抽取获得各语义项的值。另外系统根据边界分隔符和启发信息将源文档分割为多个描述某一事物不同实例的无结构的文本块，还将抽取出的结果放入根据ontology的描述信息生成的数据库中。

基于ontology方式的最大的优点是对网页结构的依赖较少，只要事先创建的应用领域的Ontology足够强大，系统可以对某一应用领域中各种网页实现信息抽取。主要缺点是：

1) 需要由领域专家创建某一应用领域的详细清晰的Ontology，工作量大。

2) 由于是根据数据本身实现信息抽取，因此在减少了对网页结构依赖的同时，增加了对网页中所含的数据结构的要求，如要求内容中包含时间、日期、电

话号码等有一定格式的内容。

(3) 从大量异构的文档中提取公共模式工作量繁重，并且不支持对超链接的处理。

- **基于HTML结构的信息抽取**

该类信息抽取技术的特点是根据Web页面的结构定位信息。在信息抽取之前通过解析器将Web文档解析成语法树，通过自动或半自动的方式产生抽取规则，将信息抽取转化为对语法树的操作实现信息抽取[10][11]。

### 2.1.2 Web 信息抽取的任务

为了填充复杂的模板，研究人员发现系统必须能执行多种简单任务，这些任务包括实体抽取、属性抽取和关系抽取等[16][17]。

- **实体抽取(Entity Extraction)**

命名实体是文本中基本的信息元素，是正确理解文本的基础。常用的实体类型有：

1) 命名实体(Named individuals)：如组织、人、地点、书、电影、宾馆等。
2) 命名类型(Named kinds)：如蛋白质、化合物、药物、疾病、飞行器等。
3) 时间(Times)：时间表达式，日期、时刻等。
4) 量度(Measures)：金钱表达式、距离、大小、重量等。

对于每个参考文本必须识别它的范围和类型，比如“IBM和Microsoft今天宣布”，其中下划线被识别为组织或者公司名。但“戴尔”是公司名还是人名则需要根据具体情况来判断。

在信息抽取研究中，命名实体识别是目前最有实用价值的一项技术。根据MUC评测结果[6]，英文命名实体识别任务的F-指数(召回率与准确率的加权几何平均值，权重取1)能达到90%以上。

命名实体识别的难点在于：在不同领域、场景下，命名实体的外延有差异；数量巨大，不能枚举，难以全部收录在词典中；某些类型的实体名称变化频繁，并且没有严格的规律可以遵循；表达形式多样；首次出现后往往采用缩写形式[1]。

命名实体识别的方法主要分为：基于规则的方法和基于统计的方法。一般来说，基于规则的方法性能要优于基于统计的方法。但是这些规则往往依赖于具体语言、领域、文本格式，编制过程耗时且容易产生错误，并且需要富有经验的语言学家才能完成。相比而言，基于统计的方法利用人工标注的语料进行训练，标注语料时不需要广博的计算语言学知识，并且可以在较短时间内完成。因此，这类系统在移植到新的领域时可以不做或少做改动，只要利用新语料训练一遍即可。此外，基于统计的系统要移植到其他自然语言文本也相对容易一些。

- **属性抽取(Attribute Extraction)**

实体常常是由感兴趣的属性联系起来的，如：

*西北大学肇始于1902年的陕西大学堂，1912年始称西北大学，1923年8月改称国立西北大学。1937年抗战爆发后，国立北平大学、国立北平师范大学、国立北洋工学院等内迁来陕，组成国立西安临时大学，1938年更名为国立西北联合大学，1939年8月复称国立西北大学。建国初期，西北大学为中央教育部直属的14所综合大学之一，1958年归属陕西省主管，1978年被确定为全国重点大学。现为国家"211工程"重点建设院校和西部大开发重点支持建设院校。*

对于上面这段文字，西北大学的属性信息可以用如下所示：

```
<OrganizationInfo>
    <ORGName>西北大学</ORGName>
    <Time>1902年</Time>
    <Address>陕西省</Address>
    <ORGType>学校</ORGType>
</OrganizationInfo>
```

属性值的发现常依赖于共指分析，即知道哪些属性是属于同一个实体。

- **关系抽取(Relation Extraction)**

在抽取实体和它们的属性之后，下一步就是抽取除实体之间的各种关系。如Employee_of是Person和Orgaization之间的关系；Product of是Artifact和Organization之间的关系等。再比如 Employee- of(张三，IBM)表示：张三是IBM

的employee(员工)；Product of (PC，IBM)表示：PC是IBM的product(产品)。

## 2.2 中文命名实体识别

命名实体识别是分词和标注过程中的一个重要环节，并在信息检索、信息抽取以及自动问答系统等领域中有直接的应用。

### 2.2.1 命名实体识别的任务

命名实体识别是一类特殊的模式识别问题，近年来有关这一问题的研究非常活跃，许多组织、学术机构每年都举办有关命名实体识别的研讨和评测。MUC中提到的命名实体包括人名(Person)、地名(Location)、机构名(organization)、日期(data)、时间(time)、百分数(percentage)和货币(monetary value)七类命名实体[15]，如图2-1所示。

<DAT>昨日下午</DAT>，<ORG>世界银行贸易局</ORG>经济顾问<PER>纽法默</PER>在<LOC>北京</LOC>指出，全球化会促进未来25年平均收入加快增长，但并非人人都能分享全球化的收益，随之而来的收入不平等在国与国间和国家内部都会加剧。<PER>纽法默</PER>在<ORG>北京大学</ORG>用简明的幻灯片演示，整体而言，发展中国家占全球产出份额从约为全球经济的<NUM>五分之一</NUM>增加到近<NUM>三分之一</NUM>。在回答提问时，<PER>纽法默</PER>表示，按照<ORG>世行</ORG>的计量标准，<LOC>中国</LOC>将在<DAT>2020年</DAT>就进入富裕国家队列。

图 2-1 中文命名实体的例子

“世界银行贸易局” 和“北京大学”都是机构名，而“北京”和“中国”都是地名。命名实体识别的关键有两个：一个是确定命名实体的左右边界，第二个就是识别改命名实体对应的类别。

命名实体识别中人名、地名、机构名是最难识别的三类，下文将有针对性地1讨论机构名和地名的识别。

### 2.2.2 中文命名实体识别的困难

相对于英文来说，中文命名实体识别的困难在于以下方面[8][9][15]：

1) 中文命名实体识别和中文分词是互相缠绕在一起的；

2) 在中文中，词的定义不清晰；

3) 中文不像英文那样在命名实体中有大小写的形态变化。

在方法上，目前识别命名实体所采用的方法往往把分词和命名实体识别分割为两个独立的步骤。如国立台湾大学的NTU系统先利用规则(3条规则和18条构词律)对文本进行分词，得到确定的分词结果后再识别人名、地名、机构名。在识别人名、地名以前进行最大匹配切分。文献[13]是在分词以前作姓名识别的。无论是先确定性切分还是先确定性命名实体识别都会存在一些问题：确定性切分的错误很可能会导致命名实体识别的错误，先确定性命名实体识别更多的是利用姓名构成的内部信息，没有充分考虑语境因素的影响。

先确定性切分可能导致的命名实体错误：

*<PER>王辉</PER>国家里有点急事。*

先姓名识别可能导致的分词错误：

*请问政府有<PER>何安</PER>全措施?*

另外，目前常用的命名实体识别策略往往是计算一个候选字符串作为人名或者地名或者机构名的概率大小，如果此概率大于某个特定的阈值，就认为是相应的命名实体。其实，这种方法更多的是利用某个字符串的内部构成规律，而没有充分利用语境信息。当然，在计算概率时可能会利用一些周围的语境信息对此概率大小进行一定的奖惩处理。但是，这样利用语境信息的方法实际上还是把其放在一个次要的位置上，只是作为一个补充手段而已。总之，这些方法存在的问题是没有把内部信息和语境信息有机的结合为一个整体，没有系统科学的方法能够准确的确定每一种情况的阈值大小。

在命名实体识别过程中，有些系统利用人名词典、地名词典、机构名词典来进行直接匹配。如果匹配成功，就作为一个命名实体。这种看似很有效的方法其实存在一些明显的问题，例如在人名词典中收集有“成方圆”、地名词典中收集有“山东”，如果采用直接匹配的方法，很容易导致下面的错误：

*没有规矩，怎<PER>成方圆</PER>*

*我家在泰<LOC>山东</LOC>边*

人名、机构名一般来说具有任意性并且是开放的集合，所以无论词典如何庞大，都不可能用穷举的方法将它们囊括进去。地名是一个相对封闭的集合，但是

在真实文本中，地名的变化形式很多，同一地名也有很多表达方式，实际上几乎也是不可能穷举的。从另一个方面来讲，词典中收集了这么多的名字，会对分词的精度有较大影响，大大增加了分词的交集歧义。所以，仅仅利用专有名词词典直接匹配的方法是不能从根本上解决问题的。

另外，除了以上所说的各类中文命名实体识别的共同困难，各类命名实体还有由其自身特点所决定的特殊困难。

- **地名识别的困难**

地名是一个相对封闭的集合，但是在真实文本中，地名的变化形式很多，同一地名也有很多表达方式，实际上几乎也是不可能穷举的，仅仅利用专有名词词典直接匹配的方法是不能从根本上解决问题的；此外，地名经常和其它词组合成机构名，地名和机构名的边界确定使得地名识别注定不可能简单地采用词典收集的办法解决地名识别问题。最后，音译名的归类也成为地名识别的另一个难题。

- **机构名识别的困难**

对于机构名识别来说，主要的瓶颈在于存在大量的未登录机构名。未登录词在人名、地名和机构名中都占有很大一部分的比例，未登录机构名的识别比未登录人名和地名的识别要难得多，这主要是由机构名的自身特点所造成的[15]：

第一，中文机构名组成方式非常复杂。机构名识别中的机构种类繁多，各类机构都有其自己独特的命名方式。例如，公私企业命名大多以地名作为开头，中间加以企业字号，如“金山”、“亿阳”等等，结尾一般都是“公司”、“集团”类的普通名词。而机关团体类的机构名则相对比较正规，一般以上级部门开头，结尾为“所”、“部”、“院”、“委”等单字。序数词在一般的机构名中很少出现，但是在军队、医院类的机构名中，序数词却占有相当大的比例。而且机构名中还嵌套的情况，机构名中包含有另一个机构名，如“北京电影学院青年电影制片厂”。

第二，机构名中含有大量的其它命名实体。在这些命名实体中，地名所占的比例最大，其中未登录地名又占了相当一部分的比例。其它命名实体的识别大大制约了机构名的识别。

第三，中文机构名用词非常广泛。通过对1998年1月人民日报语料中的10817个机构名所含的19986个词进行统计，共计27种词，其中名词最多(9941个)，地

名其次(5023个)。所用词如此之广泛，是命名实体中绝无仅有的。最为严重的是，在这些词中有很大一部分词是未登录词，例如大部分的企业字号。

第四，机构名的长度极其不固定。不像中国人名，一般为两到三个字，最多不超过四个字，地名最多也只是由三到四个词组成。机构名的长度少到两个字(“北大”、“首钢”)，多到几十个字(“中国人民政治协商会议第八届全国委员会常务委员会”)，在人民日报的真实文本中，由十个以上的词构成的复合机构名占了相当一部分的比例。机构名称长度的不确定性，导致机构名称的边界很难确定，加大了机构名识别的难度。

第五，大多数机构名都有其简称。简称一般都是取其全称中的几个关键字或关键词，例如“联想”、“人大”。大量的机构名简称的出现，使得本来已经十分困难的问题变得更加困难。

### 2.2.3 前人的相关工作

命名实体识别按照方法的不同，大体可以分为三类：基于规则的方法；基于统计的方法；统计与规则相结合的方法。后两种方法目前占主导地位。

#### 2.2.3.1 基于规则的方法

规则方法主要是利用两种信息：命名实体用字分类和限制性成分。即：分析命名实体用字，驱动对命名实体的识别过程，并采集命名实体前后相关的成分，对命名实体的前后位置进行限制。小规模测试的结果表明，其准确率可达97%。

如前面所提到的NTU系统，在地名和机构名识别上就采用了规则匹配的方法，其中地名规则的例子如下：

LocationName → PersonName LocationNameKeyWord

LocationName → LocationName LocationNameKeyWord

机构名规则的例子如下：

OrganizationName → OrganizationName OrganizationNameKeyword

OrganizationName → CountryName(D|DD) OrganizationNameKeyword

其中D表示一个内容词。

不过，我们可以发现，对于这类采用人工组织规则的系统，主要存在以下缺

点：

1) 人工组织规则的代价非常昂贵，并且主要依赖于有经验的计算语言学家；

2) 当把此系统移植到不同领域时，需要大量的人工修改工作；

3) 当把此系统移植到新的语种时，这些规则需要重新书写和组织；

4) 语言学家书写规则的经验和所花费的人力劳动的大小对性能的影响很大。

例如，文献[16]就是从10万条人名库、2亿字的真实语料库中将姓名用字分为9类，并总结了21条识别规则。无论是收集如此巨大的人名库、真实语料库，还是提炼规则，都是一个浩大的工程。这无疑是非常费时、费力的。一旦增加新特征的人名，就必须增加新的规则，并对以前的规则重新修订，因此规则方法很难扩展。规则可以保证很高的准确率，但是任何规则体系的覆盖面都是有限的，对于规则覆盖集合之外的人名就完全无能为力。

文献[13]虽然在封闭测试中能达到百分之九十多的准确率和召回率，但是在开放测试中仅能达到百分之六十多一点，远远不能满足人们的实际需求。在特定领域内尚且如此，如果把基于规则的方法推广到全领域内，其效果是可以想象的。但是，在缺乏特大规模熟语料库的时候，规则方法是唯一可行的方法。

#### 2.2.3.2 基于统计的方法

统计方法主要是针对命名实体语料库来训练某个字作为命名实体组成部分的概率值，并用它们来计算某个候选字段作为命名实体的概率，其中概率值大于一定阈值的字段为识别出的命名实体。

基于统计的命名实体识别主要包括以下方法：基于决策树模型(Decision Tree)、基于隐马尔科夫模型(HMM)、基于最大熵模型(Maximum Entropy)等。

实际上，现在实用的系统使用纯统计方法的很少，或多或少都应用了一些规则。

#### 2.2.3.3 统计与规则相结合的方法

规则与统计相结合的办法，可以通过概率计算减少规则方法的复杂性与盲目性，而且可以降低统计方法对语料库规模的要求。目前的研究基本上都是采取规

则与统计相结合的方法，不同之处仅仅在于规则与统计的侧重不同而已。

在MUC-7评测中，爱丁堡大学的A. Mikheev等在命名实体识别过程中采用了规则和最大熵模型相结合的方法。其明显的特点是把识别过程分五个步骤完成，每一步完成特定的任务。这五个步骤分别是：确定性触发规则、局部匹配1、约束较弱的规则、局部匹配2、标题的特殊处理。在识别过程中，采取的是多遍扫描的方法，每一遍扫描实施的操作不同。第一步实施的是确定性的触发规则。图2-2所示是确定性的触发规则的例子。

Capitalized_word + is a? JJ* PROF 其中
+ 表示一个或多个单词
* 表示零个或多个单词
? 表示零个或一个单词
JJ 表示一个形容词
PROF 表示称谓词 (director manager etc)
符合这条规则的一个例子：Bill Gates is a great software engineer

图 2-2 确定性触发规则

利用这些人工编写的正则表达式规则进行命名实体识别，能够获得很高的准确率，但是召回率较低。在局部匹配1阶段，首先收集该文本中已经识别出来的命名实体，对这些命名实体的各种局部(当然顺序保持不变。例如：已经识别出ABC作为机构名，那么AB, BC作为候选的机构名)都作为候选的对应类别的命名实体。第三步采用约束较弱的规则，这些规则具有较弱的语境约束，并且能够充分利用已经存在的信息和词典。在局部匹配2的处理中，系统充分利用已经识别出来的人名、地名、机构名进行局部匹配，然后经过最大熵模型进行进一步的确认和过滤。最后对文本标题中的候选命名实体通过另一最大熵模型(此最大熵模型是基于文本标题训练得到)进行确认。

## 2.3 语言模型

描述物理世界和自然语言的模型可以分为确定性模型和统计模型。确定性模型运用明确的规则来表述物理世界或自然语言的已知的特定属性，典型的例子是牛顿力学。然而并不是所有的物理世界和自然语言的现象都可以由确定的规则来刻画。在这种情况下，统计模型被用以描述物理世界和自然语言的统计属性。其

基本假设是，物理世界和自然语言可以用随机过程来刻画，而随机过程中的参数可以精确地估计。物理世界中统计模型的例子有统计力学，在自然语言中有概率语法。本节的主要内容就是介绍几种概率语法，如N元模型、隐马尔可夫模型等。

### 2.3.1 N 元模型

统计语言模型的实质就是刻画所有序列W=$w_1...w_n$的概率分布P(W)，此概率分布反映了字符序列W作为句子的概率大小。我们首先介绍被广泛运用于不同应用领域中的N元模型。

在N元模型中，P(W)可以分解如下：

$$P(W)=P(w_1w_2...w_n)$$
$$=P(w_1)P(w_2|w_1)P(w_3|w_1w_2)...P(w_n|w_1w_2...w_{n-1}) \quad \text{公式 (2-1)}$$
$$=\prod_{i=1}^{n}P(w_i|w_1,w_2...w_{i-1})$$

其中，$P(w_i|w_1,w_2...w_{i-1})$表示在给定序列$w_1,w_2...w_{i-1}$的条件下，后面紧跟$w_i$的概率。如果词典规模为|V|，$w_i$有$|V|^{i-1}$个不同的历史，所以为了刻画概率$P(w_i|w_1,w_2...w_{i-1})$，有$v^i$个参数需要估计。随着历史长度的增加，不同的历史数按照指数级增长。事实上，绝大多数的历史在训练数据中根本没有出现，要估计这么多的参数是不可能的。所以，可以假定$P(w_i|w_1,w_2...w_{i-1})$只依赖于等价类，而等价类的数目远远小于不同历史的数目。

一种简单的等价类可以近似的假定$P(w_i|w_1,w_2...w_{i-1})$只依赖于前面的N-1个词$w_{i-N+1}w_{i-N+2}...w_{i-1}$，这样得到的模型就是N元模型。特别地，当N=2时，$P(w_i|w_1,w_2...w_{i-1})\equiv P(w_i|w_{i-1})$，这就是二元模型(bigram)，也被称为一阶马尔可夫链；当N=3时，$P(w_i|w_1,w_2...w_{i-1})\equiv P(w_i|w_{i-2},w_{i-1})$，这就是三元模型(trigram)，也被称为二阶马尔可夫链[15]。

W = John read a book，为了使$P(w_1|w_0)$有意义，我们添加句首标志<BOS>，并相应增加句尾标志<EOS>。那么在二元模型下P(John read a book) = P(John|BOS)P(read|John)P(a|read)P(book|a)P(<EOS>|book)。

$P(w_i \mid w_{i-1})$可以通过最大似然估计的方法来得到，即：

$$P_{MLE}(w_i \mid w_{i-1}) = C(w_{i-1}, w_i) / C(w_{i-1}) \qquad \text{公式 (2-2)}$$

其中，C(*)表示训练语料库中观测到的序列次数。

假设训练集S为("John read mob dick","Mary read a different book"," she read a book by cheer")，那么

$$P(John \mid < BOS >) = C(< BOS >, John) / C(< BOS >) = \frac{1}{3}$$

$$P(read \mid John) = C(John, read) / C(John) = \frac{1}{1}$$

$$P(a \mid read) = C(read, a) / C(read) = \frac{2}{3}$$

$$P(book \mid a) = C(a, book) / C(a) = \frac{1}{2}$$

$$P(< EOS > \mid book) = C(book, < EOS >) / C(book) = \frac{1}{2}$$

从而

P(John read a book)

= P(John|BOS)P(read|John)P(a|read)P(book|a)P(<EOS>|book)

$= (\frac{1}{3})*(1)*(\frac{2}{3})*(\frac{1}{2})*(\frac{1}{2})$

$\approx$ 0.06

但是，我们要注意到由于语言模型的训练语料不可能无限大，许多词之间的接续在训练语料中没有出现过，最大似然估计条件概率$P_{MLE}(w_i \mid w_{i-1})$的过程中必然会出现零概率的情况。例如对于P(Cher read a book)来说，

$$P(read \mid Cher) = C(Cher, read) / C(Cher) = \frac{0}{1}$$

很显然，句子的出现概率被低估了，Cher read a book有一定的出现概率。

这种训练不足引起的零概率问题称为数据稀疏问题。为了克服数据稀疏问题，数据平滑处理(smoothing)是非常必要的。

常用的数据平滑处理策略有模型整合(model combine)和折扣(discounting)两种。模型整合就是将不同模型结合起来，以取得综合的效果。折扣就是从观测到的事件最大似然概率中扣除一部分，把这部分余额分配给没有观测到的事件。我们在处理词典中未出现过的词时采用的平滑策略为Laplace折扣方法。

Laplace折扣方法表示为：

$$P(wi \mid wi-1) \cong \frac{C(w_{i-1}, w_i)+1}{C(w_{i-1})+|V|} \qquad \text{公式 (2-3)}$$

其中，|V|表示词典的规模，对于上例中的训练集S, |V| =11。

经过平滑处理后：

P(John read a book)
= P(John|BOS)P(read|John)P(a|read)P(book|a)P(<EOS>|book)
$= (\frac{1}{14})*(\frac{1}{12})*(\frac{3}{14})*(\frac{2}{13})*(\frac{2}{13})$
$\approx$ 0.00003

这些概率与原来相比更加合理。

### 2.3.2 基于类的语言模型

基于类的语言模型利用词的等价类的N元组来刻画它们之间的依存关系。由于一些单词可以归为一个类，基于类的语言模型的参数个数较少，所以可以在一定程度上克服数据稀疏问题。词分类是按照一定的语法或语义属性，将词汇集划分成若干等价类，使语法或语义属性相近的词属于同一个等价类。词分类是在语言学习中从特殊到一般的通用化的常用方法。

在建立基于类的语言模型之前，首先要确定从单词到等价类的映射函数。一般来说，有两种类型的映射函数：确定性的(每一个单词只能属于一个类)和概率性的(一个单词可能属于多个类，并且属于各个类的概率不同)。确定映射函数的方法也有两类：根据语言知识(语法信息、语义信息等)人为确定[14]和通过聚类方法来确定[17]。

如果是确定性映射函数且假设从单词到等价类的映射函数为：C：w->C(w)，那么$P(w_i \mid w_{i-1})$可以用基于类的N元语言模型来表示：

$$P(w_i \mid w_{i-1}) = P(w_i \mid C(w_{i-1})) \qquad \text{公式 (2-4)}$$

如果是概率性映射函数，那么，$P(w_i \mid w_{i-1})$可以表示如下：

$$P(w_i \mid w_{i-1}) = \sum_{C(w_i)} P(w_i \mid C(w_i)) \times P(C(w_i) \mid C(w_{i-1})) \qquad \text{公式 (2-5)}$$

例如：Season = {spring, summer, autumn, winter}。

在应用过程中，经常是把基于类的N元模型和词的N元模型结合在一起，基于类的N元模型在一定程度上克服数据稀疏问题。对于不同的任务，对于类的定义也不一样。

### 2.3.3 马尔可夫模型

马尔可夫(Markov，又译为马尔柯夫)模型描述了一类重要的随机过程。如果一个系统有N个状态S1，S2，...，Sn，随着时间的推移，该系统从某一状态转移到另一状态，我们将在时间t的状态记为qt。对该系统的描述通常需要给出系统的当前状态(时间为t的状态)及其之前的所有状态：系统在时间t处于状态S的概率取决于其在时间1，2，...，t-1的状态，该概率为：

$$P(q_t = S_j \mid q_{t-1} = Si, q_{t-2} = S_k, ...) \qquad \text{公式 (2-6)}$$

如果在特定情况下，系统在时间t的状态只与其在时间t-1的状态相关，则该系统构成一个离散的一阶马尔可夫链：

$$P(q_t = S_j \mid q_{t-1} = Si, q_{t-2} = S_k, ...) = P(q_t = S_j \mid q_{t-1} = S_i) \qquad \text{公式 (2-7)}$$

进一步，我们只考虑独立于时间t的随机过程：

$$P(q_t = S_j \mid q_{t-1} = Si) = aij, 1 \le i, j \le N \qquad \text{公式 (2-8)}$$

该随机过程为马尔可夫模型。其中状态转移概率aij必须满足：

$$aij \ge 0$$

$$\sum_{j=1}^{N} a_{ij} = 1$$

马尔可夫模型又可视为随机有限状态自动机。该有限状态自动机的每一个状态转换都有一个相应的概率，该概率表示自动机采用这一状态转换的可能性。例如，假定在一段时间内的气象可由一个三状态马尔可夫模型M描述：

状态S1：雨或雪

状态S2：多云

状态间的转移概率由下列矩阵给出：

$$A = [a_{ij}] = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

如果第一天为晴天，根据这一模型，在今后七天中天气为“晴晴雨雨晴云晴”的概率为：

$$\begin{aligned}P(O|M) &= P(S_3,S_3,S_3,S_1,S_1,S_3,S_2,S_3|M)\\ &= P(S_3)\times P(S_3|S_3)\times P(S_3|S_3)\times P(S_1|S_3)\times P(S_1|S_1)\times P(S_3|S_1)\times P(S_2|S_3)\times P(S_3|S_2)\\ &= 1\times a_{33}\times a_{33}\times a_{31}\times a_{11}\times a_{13}\times a_{32}\times a_{23}\\ &= 0.8\times 0.8\times 0.1\times 0.4\times 0.3\times 0.1\times 0.2\\ &= 1.536\times 10^{-4}\end{aligned}$$

### 2.3.4 隐马尔可夫模型

马尔可夫模型中，每一个状态代表一个可观察的事件，这限制了模型的适用性。在隐马尔可夫模型中，观察到的事件是状态的随机函数。因此该模型是一双重随机过程，其中模型的状态转移过程是不可观察(隐蔽)的。而可观察的事件的随机过程是隐蔽的状态转换过程的随机函数。也可以这样理解：马尔可夫模型的概念是一个离散时域有限状态自动机，隐马尔可夫模型HMM是指这一马尔可夫模型的内部状态外界不可见，外界只能看到各个时刻的输出值。

隐马尔可夫模型(Hidden Markov Model; HMM)[15]是经典的描述随机过程的统计方法，在自然语言处理中得到了广泛的应用。HMM模型可以看作一种特定的BayesNet，等价于概率正规语法或概率有限状态自动机，可以用一种特定的神经网络模型来模拟。HMM模型的优点主要有：研究已经非常透彻，算法成熟，效率高，效果好，易于训练。

隐马尔可夫模型的三大假设：对于一个随机事件，有一个观察值序列：$O_1,O_2...O_T$，该事件隐含着一个状态序列：$X_1,X_2...X_T$

假设1：马尔可夫假设(状态构成一阶马尔可夫链)

$$P(X_i|X_{i-1}...X_1)=P(X_i|X_{i-1})$$

假设2：不动性假设(状态与具体时间无关)

$$P(X_{i+1}|X_i)=P(X_{j+1}|X_j),\text{对任意i,j成立}$$

假设3：输出独立性假设(输出仅与当前状态有关)

$$P(O_1,...,O_T|X_1,...X_T)=\prod P(O_t|X_t)$$

一个隐马尔可夫模型(HMM)可以形式化为一个五元组：$(\Omega x,\Omega o,A,B,\pi)$，

其中：

$\Omega_X = \{q_1,...q_N\}$：状态的有限集合

$\Omega o = \{v_1,...v_M\}$：观察值的有限集合

$A = \{aij\}, aij = P(Xt+1 = qi \mid Xt = qi)$：转移概率

$B = \{b_{ik}\}, b_{ik} = P(O_t = v_k \mid X_t = q_i)$：输出概率

$\pi = \{\pi_i\}, \pi_i = P(X_t = q_i)$：初始状态分布

对于词性标注任务来说，已知的单词序列$w_1w_2...w_n$为观察值序列，词性序列$c_1c_2..c_n$便为隐含着的状态序列。训练过程实际就是统计词性转移矩阵$[a_{ij}]$和词性到单词的输出矩阵$[b_{ik}]$，而求解的过程实际上就是一个用Viterbi算法求可能性最大的状态序列。

# 第三章 Web 招聘信息抽取系统设计

本章主要介绍 Web 招聘信息抽取系统的设计。3.1 节在指明系统目标的基础上对系统框架进行分析；3.2 节对 Web 信息抽取系统的基础 Spider 模块进行分析和算法实现；3.3 节对文本预处理：从 HTML 标签过滤和词性标注两个方面进行论述；3.4 节是本章的核心，对待识别的各个命名实体的识别方法进行阐述。

## 3.1 系统目标和分析

### 3.1.1 系统目标

由于本系统是个Web招聘信息抽取系统，而各个招聘网站的结构各异，形态多样，传统的Web信息抽取方式，如基于HTML结构的方式以及包装器归纳方式就显得力不从心，所以本课题研究采用基于自然语言理解的方式来进行Web信息抽取，同时还要解决基于自然语言理解的方式对半结构化文本处理的不足。系统的重点和难点分析如下：

1) 要解决自然语言理解在处理半结构化文本时的不足，需要结合别的Web信息抽取方式来改进；

2) 需要改进现有的语言模型以提高系统的识别准确率和召回率；

3) Spider的构建。本信息抽取系统需要的是专用的Spider，因此必须自己构建，并且要求抓取效率足够的高；

4) 由于本文采用基于自然语言理解的方式来进行信息抽取，所以需要准备较多的识别资源来提高识别效果。

### 3.1.2 系统分析

从整体上看，本系统属于 B/S(浏览器/服务器)架构，用户浏览器负责显示抽取的结果，而服务器端是系统的核心，负责所有的 Web 信息抽取工作。根据前面的系统目标，进行以下分析：

首先需要构建 Spider，爬行网上各大招聘网站，Spider 爬行起始点除了若干默认网站外可以由 Browser 端用户指定；

对于 Spider 抓取的网页需要经过过滤 HTML/XML 标签，预处理为普通文本。为了结合“基于 HTML 结构的 Web 信息抽取”的优势，在该过程需要利用 HTML/XML 标签的提示作用，详见后文叙述；

中文分词，使用中科院计算所的开源系统 ICTCLAS[15]，该系统基于层叠隐马尔科夫模型实现，分词效果良好，在 2003 年 5 月 SIGHAN 举办的第一届汉语分词大赛中名列第一；

信息抽取模块是本系统的核心，该模块需要识别的命名实体有：招聘单位名、地址、职位名和联系方式等。采用统计加规则的方式进行识别，需要构建若干的知识库以及相应的规则库；

用户接口，Bowser 端用户应能在浏览器中浏览或者搜索招聘信息，这里可使用脚本语言如 ASP 来实现。

根据分析，得到相应的系统模块结构，见图3-1。

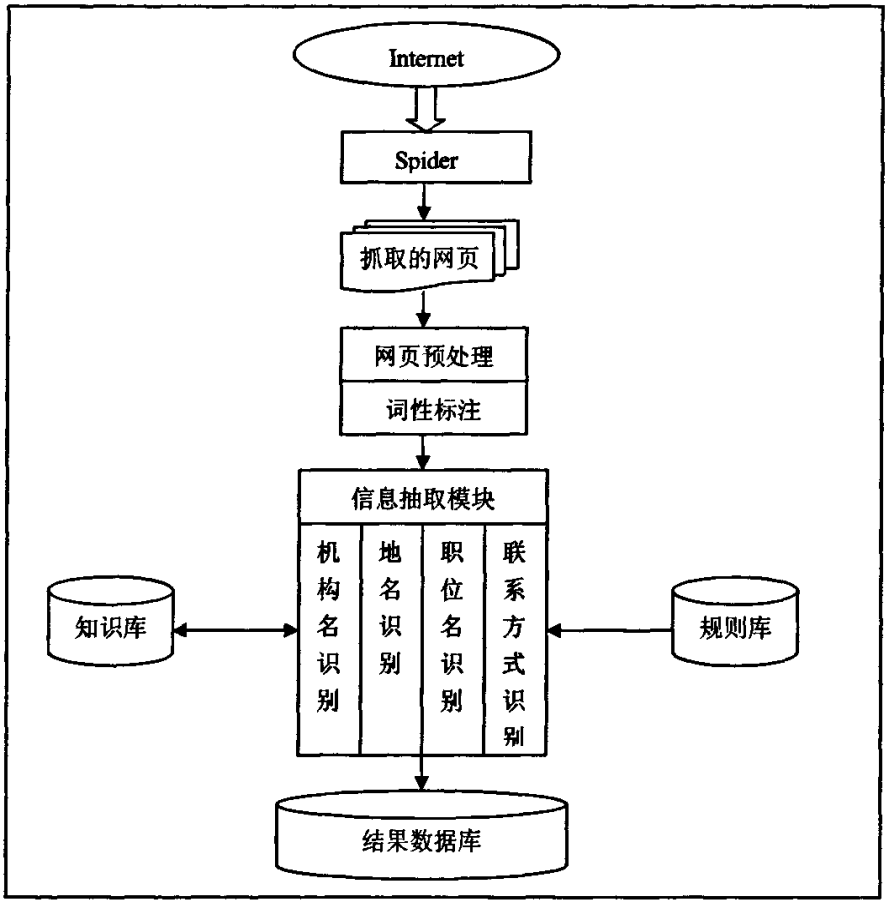


图 3-1 Web 信息抽取系统整体结构图

## 3.2 Spider 的构建

在WWW上海量存在的是网页，而网页与普通文本的最大不同在于网页内具有超链接(Hyperlink)，这些超链接又指向别的网页，于是数以几十亿的网页就牵连在了一起而构成WWW。理论上说，从任意一个网页就能够遍历整个WWW，这个遍历过程很像在蛛网上爬来爬去的蜘蛛，因此，搜索引擎的Robot程序被称为Spider程序[18]。

### 3.2.1 Spider 的工作原理

发现和搜集信息是Spider程序的根本目标[19]。一个典型的网络蜘蛛工作的方式为：查看一个页面，并从中找到相关信息，然后它再从该页面的所有链接中出发，继续寻找相关的信息，以此类推。网络蜘蛛在搜索引擎整体结构中的位置如图3-2所示。

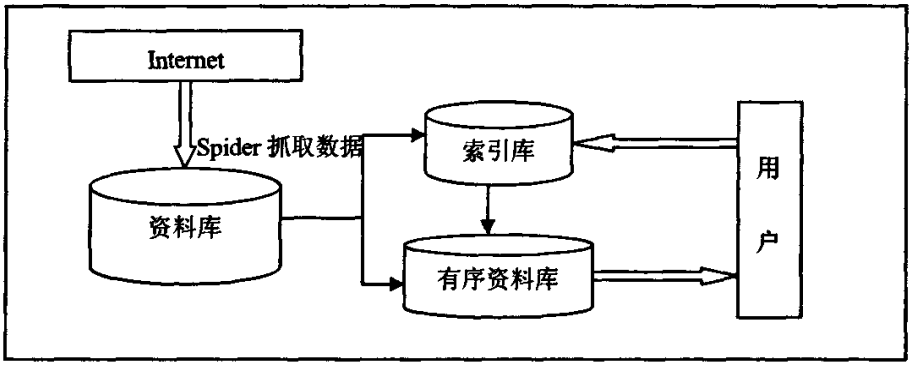


图 3-2 Spider 在搜索引擎中的结构示意

初始化时， 网络蜘蛛一般指向一个URL ( Uniform Resource Locator)池。在遍历WWW的过程中，按照深度优先、广度优先或其他启发式算法从URL池中取出若干URL 进行处理，同时将未访问过的URL放入URL池中，这样处理直到URL池空为止。对Web 文档的索引则根据文档的标题、首段落甚至整个页面内容进行，这取决于搜索服务的数据收集策略。网络蜘蛛在爬行过程中，根据页面的标题、头、链接等生成摘要放在索引数据库中。如果是全文搜索，还需要将整个页面的内容保存到本地数据库。

网络蜘蛛为实现其快速地浏览整个互联网，通常采用抢先式多线程技术实现

网上的信息搜索。通过抢先式多线程的使用，能索引一个基于URL 链接的Web 页面，启动一个新的线程跟随每个新的URL 链接[19]。当然在服务器上所开的线程也不能无限膨胀，需要在服务器的正常运转和快速收集网页之间找一个平衡点[20][21]。

### 3.2.2 Spider 的实现

考虑到抓取效率，本 Spider 采用多线程的广度优先策略，线程个数和抓取深度都可以指定，并采用 C 语言结合底层 socket 封装成一个动态链接库，便于主程序调用。另外，本 Spider 程序是个专用 Spider，它只针对特定网站甚至是网站的特定部分进行抓取。具体流程图见图 3-3，测试部分详见第四章。

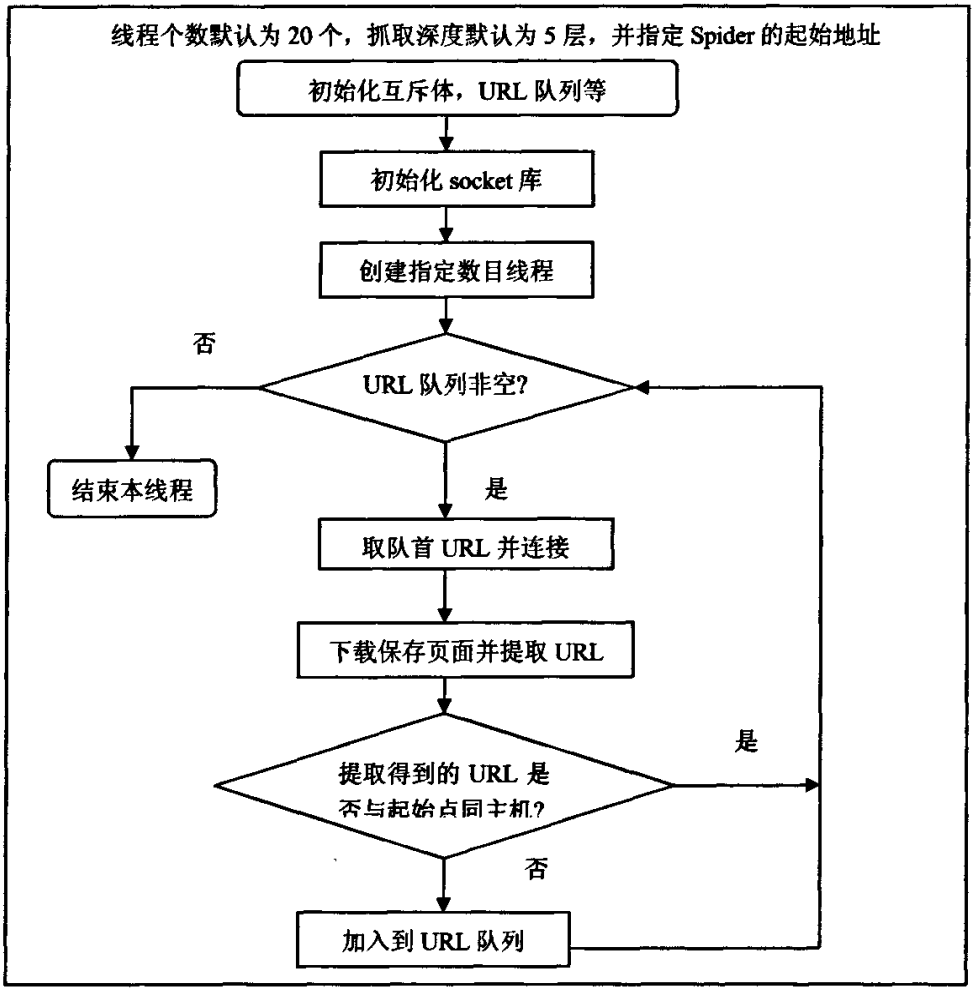


图 3-3 Spider 流程图

## 3.3 预处理

Spider抓取的网页需要两个步骤的预处理：1) HTML/XML标签过滤；2) 分词和词性标注。下面分别进行阐述。

### 3.3.1 HTML/XML 标签过滤

Web信息抽取的处理对象是半结构化文本，所以存在“基于HTML结构的方式”，它根据HTML/XML结构生成抽取规则进行抽取，并能取得很好的效果，而“基于自然语言理解的方式”实质上处理的是自由文本，也就是没有考虑HTML/XML结构。本文的特色之一在于利用HTML/XML结构进行基于自然语言理解的信息抽取。如果对于如图3-4所示招聘网页，如果纯粹使用“基于自然语言理解的方式”，那么将要处理的文本如图3-5所示。

当前位置：首页 >>> 企业介绍及招聘职位


广州市日立电梯有限公司

通讯地址：广东省广州市番禺区大石镇石北工业区
邮政编码：
传　　真：--
公司网站：Http://

广州日立电梯有限公司（简称 HELG）成立于 1996 年 1 月 15 日，是华南地区最大的电梯生产企业，是日立电梯在中国的唯一制造商。公司由株式会社日立制作所和广州广日集团有限公司共同投资建立，总投资为 9000 万美元，注册资本 6488 万美元。在全国各主要城市设立了 33 家分支机构。司主要生产"日立"牌和"广州日立"牌电梯和扶梯，生产规模至 2004 年底为年产电梯 13000 台。经营各类型电梯、扶梯、自动人行道、大楼设备管理网络技术系统等，集产品研发、制造、销售、进出口贸易、安装、维修、保养工程服务为一体，是广东省 50 户工业龙头企业之一。经过不断的努力，公司经济实力已跃居全国同行第二位，市场占有率达 13.6%，并连续三年获得"全国用户满意企业"称号。 1998 年，通过了英国劳氏质量认证公司 ISO9001 质量管理体系认证并同时取得美、日、英等国的授信证书；2000 年，通过英国劳氏质量认证公司 ISO14001 环境管理体系认证；2003 年 1 月，通过了英国劳氏质量认证公司 OHSAS18001 职业安全健康管理体系认证。 日立电梯依靠先进的技术，通过加强售后服务，延长产品的使用寿命使客户得到增值服务，让客户真正享受到"产品使用过程的服务"。

| 招聘职位：工艺工程师 | | | |
|---|---|---|---|
| 职位编号： | 115400 | | |
| 招聘日期： | 2006-12-27 ~ 2007-03-22 | 招聘部门： | 不限 |
| 联 系 人： | 王小姐/张科长 | 工作地点： | 广东省广州市 |
| 联系电话： | 合则约见，谢绝来电 | 传　　真： | - |
| 电子邮件： | （请通过系统发送求职意向） | | |
| 招聘人数： | 不限 | 学　　历： | 本科 |
| 工作年限： | 2 年以上 | 性别要求： | 不限 |
| 年龄要求： | 18 岁到 60 岁 | 所在地区： | 全国(不限) |

图 3-4 招聘网页示例

可以看到，该文本已经失去了半结构化文本的特征，这使得后面的命名实体识别阶段对半结构化文本的处理优势不复存在，所以为了保留半结构化文本信息，需要在过滤 HTML/XML 标记时进行特殊处理：考察 HTML/XML 标记后发现，在过滤结束标签时加上段落分隔符(自定义为双空格)就可以保留半结构化文本特征，处理流程见图 3-6：

当前位置：首页 >>> 企业介绍及招聘职位广州市日立电梯有限公司通讯地址：广东省广州市番禺区大石镇石北工业区邮政编码：传　　真：--公司网站：Http://广州日立电梯有限公司（简称 HELG）成立于 1996 年 1 月 15 日，是华南地区最大的电梯生产企业，是日立电梯在中国的唯一制造商。公司由株式会社日立制作所和广州广日集团有限公司共同投资建立，总投资为 9000 万美元，注册资本 6488 万美元。在全国各主要城市设立了 33 家分支机构。　公司主要生产“日立”牌和“广州日立”牌电梯和扶梯，生产规模至 2004 年底为年产电梯 13000 台。经营各类型电梯、扶梯、自动人行道、大楼设备管理网络技术系统等，集产品研发、制造、销售、进出口贸易、安装、维修、保养工程服务为一体，是广东省 50 户工业龙头企业之一。经过不断的努力，公司经济实力已跃居全国同行第二位，市场占有率达 13.6%，并连续三年获得“全国用户满意企业”称号。　1998 年，通过了英国劳氏质量认证公司 ISO9001 质量管理体系认证并同时取得美、日、英等国的授信证书；2000 年，通过英国劳氏质量认证公司 ISO14001 环境管理体系认证；2003 年 1 月，通过了英国劳氏质量认证公司 OHSAS18001 职业安全健康管理体系认证。　日立电梯依靠先进的技术，通过加强售后服务，延长产品的使用寿命使客户得到增值服务，让客户真正享受到“产品使用过程的服务”。招聘职位：工艺工程师职位编号：115400 招聘日期：2006-12-27 ~ 2007-03-22 招聘部门：不限联　系　人：王小姐/张科长工作地点：　广东省广州市联系电话：　合则约见，谢绝来电　传真：-电子邮件：（请通过系统发送求职意向）招聘人数：不限 学　　历：本科工作年限：2 年以上性别要求：不限年龄要求：18 岁到 60 岁所在地区：全国(不限)

图 3-5 过滤掉 HTML 标记后得到的文本

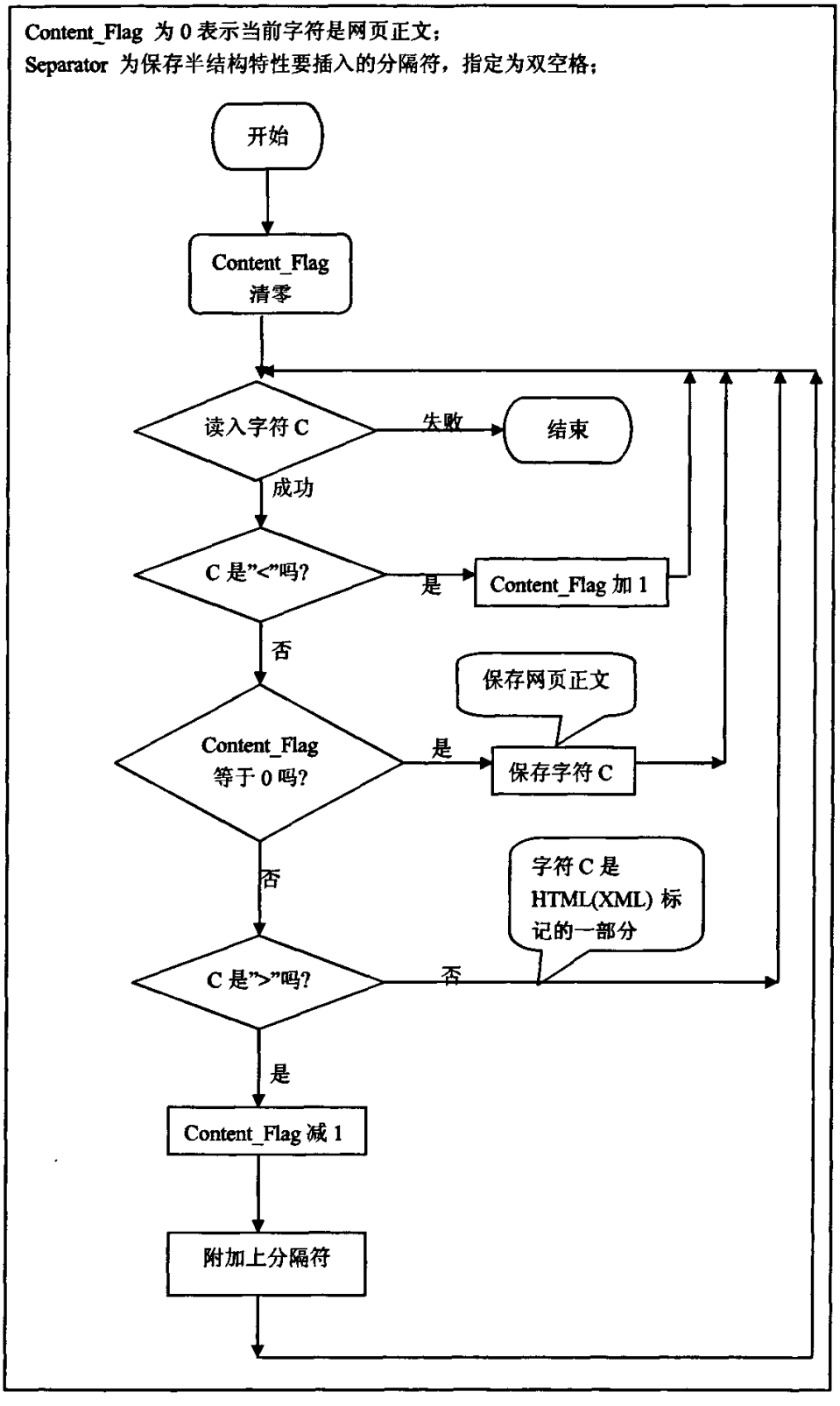


图 3-6 过滤 HTML(XML)标记流程

经过测试，该过滤方式效果良好，抽取数据的纯度比开源工具 HTMLParser 要高，并可以同时处理 HTML 和 XML 两种文档。本过滤方式被封装成动态链接库 HTMLParse.dll 以便后面调用。图 4-4 所示的网页经特殊方式过滤后得到文本见图 3-7。

当前位置：首页 >>> 企业介绍及招聘职位　　广州市日立电梯有限公司
通讯地址：广东省广州市广州市番禺区大石镇石北工业区　　邮政编码：　　传　　真：一
公司网站：Http：//
广州日立电梯有限公司（简称 HELG）成立于 1996 年 1 月 15 日，是华南地区最大的电梯生产企业，是日立电梯在中国的唯一制造商。公司由株式会社日立制作所和广州广日集团有限公司共同投资建立，总投资为 9000 万美元，注册资本 6488 万美元。…… 2003 年 1 月，通过了英国劳氏质量认证公司 OHSAS18001 职业安全健康管理体系认证。 日立电梯依靠先进的技术，通过加强售后服务，延长产品的使用寿命使客户得到增值服务，让客户真正享受到“产品使用过程的服务”。
招聘职位：工艺工程师　　职位编号：115400　　招聘日期：2006-12-27 ~ 2007-03-22　　招聘部门：不限　　联系人：王小姐/张科长　　工作地点：广东省广州市　　联系电话：合则约见，谢绝来电　　传　　真：-　　电子邮件：（请通过系统发送求职意向）　　招聘人数：不限　　学　　历：本科工作年限：2 年以上　　性别要求：不限　　年龄要求：18 岁到 60 岁　　所在地区：全国(不限)

图 3-7 经特殊方式过滤掉 HTML 标记后得到的文本

### 3.3.2 词性标注

词是最小的能够独立活动的有意义的语言成分，但汉语是以字为基本的书写单位，词语之间没有明显的区分标记，因此，中文词法分析是中文信息处理的基础与关键[22][23]。由于本课题着重研究Web信息抽取系统的命名实体识别部分，所以对于中文词法分析本身并不准备做深入的探讨和研究。Web信息抽取系统采用第三方的词法分析系统：ICTCLAS（Institute of Computing Technology Chinese Lexical Analysis System[24]），以作为Web信息抽取系统进行分词处理和词性标注的基础性构件。选择ICTCLAS系统，主要基于以下几个理由：

1) 功能全面，既包括传统的分词处理功能，也支持词性的一级、二级标注以及未登录词的识别。

2) 分词准确率高[25]。按照973专家组评测结果，ICTCLAS分词识别率97.58%，基于角色标注的未登录词识别能取得高于90%的召回率，其中中国人名的识别召回率接近98%。

3) 应用范围广。ICTCALS已经为超过2000人次的研究人员下载和使用，并收到良好的效果。在文献[26]中，作者实现的基于互联网的中文问答系统也使用了ICTCLAS作为词法分析组件，并收到良好的效果。

实践表明，选用ICTCLAS作为词法分析组件，获得了时间效率和词法处理准确率的很好的平衡，为Web信息抽取的后续处理提供了良好的基础。

在 ICTCLAS 基础上，作者把它简单的封装成一个库，命名为 WordTag.dll，接口尽量简单，只保留段落的词法分析功能，并将分词结果输出为 XML 格式，便于后面信息抽取模块处理。

对于图 3-7 所示的文本通过 WordTag 标注后见图 3-8(词性标注集见附录一)。

当前<nt>位置<n>：<w>首<m>页<q> >>><ws> 企业<n>介绍<v>及<c>招聘<vn>职位<n>
广州市<ns>日立<nz>电梯<n>有限公司<n> 通讯<n>地址<n>：<w>广东省
<ns>广州市<ns>番禺<ns>区<n>大石<ns>镇<n>石<j>北<nd>工业区<n>
邮政编码<n>：<w> 传<v> 真<a>：<w> --<m> 公司<n>网站<n>：<w>
广州<ns>日立<nz>电梯<n>有限公司<n>（<w>（<w>简称<v>HELG<ws>）<w>）<w>
成立<v>于<p>1996 年<nt>1 月<nt>15 日<nt>，<w>，<w>是<v>华南<ns>地区<n>最<d>
大<a>的<u>电梯<n>生产<vn>企业<n>，<w>，<w>是<v>日立<nz>电梯<n>在<p>中国
<ns>的<u>唯一<f>制造商<n>。<w>公司<n>由<p>株式会社<n>日立<nz>制作<v>所<u>
和<c>广州<ns>广<j>日<j>集团<n>有限公司<n>共同<d>投资<v>建立<v>，<w>，<w>总
<f>投资<n>为<p>9000 万<m>美元<q>，<w>，<w>注册<v>资本<n>6488 万<m>美元<q>。
<w>…<w>…<w> 2003 年<nt>1 月<nt>，<w>，<w>通过<v>了<u>英国<ns>劳<gn>氏<gn>
质量<n>认证<vn>公司<n>OHSAS18001<ws>职业<n>安全<an>健康<an>管理<vn>体系
<n>认证<vn>。<w> 日立<nz>电梯<n>依靠<v>先进<a>的<u>技术<n>，<w>，<w>通过
<p>加强<v>售后服务<l>，<w>，<w>延长<nz>产品<n>的<u>使用<vn>寿命<n>使<v>客
户<n>得到<v>增值<vn>服务<vn>，<w>，<w>让<v>客户<n>真正<d>享受<v>到
<v>"<w>"<w>产品<n>使用<v>过程<n>的<u>服务<vn>"<w>"<w>。<w>
招聘<v>职位<n>：<w>工艺<n>工程师<n> 职位<n>编号<vn>：<w> 115400<ws>
招聘<v>日期<n>：<w> 2006-12-27<ws>~<ws>2007-03-22<m> <ws>招聘
<vn>部门<n>：<w> <ws>不<d>限<v> 联<gv> 系<v> 人<n>：<w> 王<nh>
小姐<n>/<ws>张<q>科长<n> <ws> 工作<vn>地点<n>：<w> <ws>广东省<ns>
广州市<ns> 联系<v>电话<n>：<w> <ws>合<v>则<d>约<d>见<v>，<w>，<w>谢绝
<v>来电<v> <ws> 传<v> 真<a>：<w> -<ws> 电子<n>邮件<n>：
<w><ws> （<w>（<w>请<v>通过<p>系统<n>发送<v>求职<vn>意向<n>）<w>）
<w> 招聘<v>人数<n>：<w><ws>不<d>限<v> <ws>学<v> 历<gv>：<w>
<ws>本科<n>工作<vn>年限<n>：<w> 2<ws>年<q>以上<nd> <ws> 性
别<n>要求<n>：<w> <ws>不<d>限<v> 年龄<n>要求<n>：<w> 18<ws>岁<q>
到<v>60<m>岁<q> 所在<n>地区<n>：<w> <ws>全国<n>(<w>(<w>不<d>限

图 3-8 使用 WordTag 进行词性标注后的文本

## 3.4 机构名识别

### 3.4.1 机构名识别现状

近年来，中文人名、地名的识别研究已经取得了较大的进展，而对中文机构名识别还未能获得较好的效果[27]。 2004年度国家863中文信息处理与智能人机接口技术评测的命名实体识别评测结果显示：中文机构名识别的召回率仅为57.41%，准确率仅为64.64%。这表明对中文机构名的识别研究目前仍处在探索阶段[28]。

相对于中文人名、地名的识别来说，中文机构名的识别存在较大的困难[29]：一方面，由于中文机构名称数目庞大，层出不穷，给自然语言处理带来困难；另一方面，机构名称变化较大，很不稳定。目前国内外对中文组织机构名称的研究仅限于学校、企业这些特定的类型，其识别技术大多数基于人工书写规则，这种方法虽然取得了较好的研究成果，但与实际需要还有相当大的差距。例如，文献[30]和文献[31]提出了基于启发式规则的机构名识别方法，虽然取得了一定的效果，但论文所报告的测试结果只是基于一个很小规模的测试数据集。由于机构名种类繁多，对各类机构名要总结出统一的识别规则，这基本上是不可行的。也有基于统计的，例如文献[32]提出了基于隐马尔可夫模型的机构名识别方法，但由于语料库规模的有限性并且对于小概率稀疏事件处理方面的不足导致召回率不太理想。

### 3.4.2 机构名识别理论基础

机构名称一般泛指机关、团体、企事业单位和协会等，包括：学校、研究所、公司、医院、银行和公检法等。这些词都称之为机构名核心词，本课题的机构名识别部分研究的是含有机构名核心词的中文机构名识别，不考虑诸如“中国电信”、“网通”等不含特征词或者机构名缩略语这样的机构名。形式上，中文机构名称的构造是“W+X+G”。其中“W”代表描述性词，X+表示X 的元素出现一次或多次，G是机构名核心词。即：机构名称是由一个或一个以上的描述性词加上机构核心词组成。从句法角度看，机构名称属“定语+名词性中心语”型的名词短语。

简称定名型短语。这类短语的定语部分一般只能含名词、代词、形容词、动词、数量词或短语，语法结构简单[33]。

通过对机构名的语法和构成方式两方面的分析，我们发现机构名的形成过程是个随机的过程，这个随机过程具有两个层面的含义：词性选择的随机性和词语选择的随机性。例如，“中国光大银行”这个机构名在形成过程是这样的：首先选择“中国”这个处所词来进行范围上的限定，这个步骤选择处所词作为机构名的描述性词和选择“中国”作为处所词都是具有随机性的；然后分别选择“光”和“大”两个形容词再次修饰机构名核心词“银行”，同样这里选择两个形容词与分别选择“光”和“大”作为修饰词都是随机的。于是，机构名形成了一个隐马尔可夫链，并且该链具有两个层面的随机性特征。可以用词性的转移概率和词语的转移概率共同描述机构名的生成概率。

形式化描述如下：

词组 $W=w_1w_2...w_n$，$S=s_1s_2...s_n$，$s_i$ 是词组 $W$ 经过切分后 $w_i$ 对应的词性，$C=c_1c_2...c_n$，其中 $c_i$ 表示机构名用词。要计算该词组是一个机构名的概率：$P_1(C|W)$ 和 $P_2(C|S)$，由Bayes法则得，$P_1(C|W)=P(C,W)/P(W)$，$P_2(C|S)=P(C,S)/P(S)$，对于给定词组 $P(W)$ 和 $P(S)$ 是固定的，所以只要求 $P(C,W)$ 和 $P(C,S)$ 即可，而

$$\begin{cases} P_1(C,W)=\prod_{i=1}^{n}P_1(c_i,w_i)=P_1(c_1,w_1)*P_1(c_2,w_2)*...P_1(c_n,w_n) \\ P_2(C,S)=\prod_{i=1}^{n}P_2(c_i,s_i)=P_2(c_1,s_1)*P_2(c_2,s_2)*...P_2(c_n,s_n) \end{cases}$$

其中，

$$P_1(c_i,w_i)=N_c(w_i)/N_t(w_i)=\begin{cases}(1-\partial)*N_c(w_i)/N_t(w_i) & if \quad N_c(w_i)>0 \\ \partial/N_t(w_i) & otherwise\end{cases}$$

$$P_2(c_i,s_i)=N_c(s_i)/N_t(s_i)=\begin{cases}(1-\partial)*N_c(s_i)/N_t(s_i) & if \quad N_c(s_i)>0 \\ \partial/N_t(s_i) & otherwise\end{cases} \qquad 公式\ (3\text{-}1)$$

$N_c(w_i)$ 是词 $w_i$ 在真实文本中作为机构名出现的次数，$N_t(w_i)$ 是词 $w_i$ 在真实文本中出现的总次数，$N_c(s_i)$ 是词性 $s_i$ 在真实文本中作为机构名出现的次数，$N_t(s_i)$

是词性$s_i$在真实文本中出现的总次数，$\partial$是采用线性折扣采用的平滑参数[34]；

计算出$P_1(C,W)$和$P_2(C,S)$后和阈值$\sigma_1$和$\sigma 2$比较，只有当$P_1(C,W)>\sigma_1$并且$P_2(C,S)>\sigma 2$，我们认为词组$W$是一个机构名。

具有了以上基于统计的形式化描述之后，我们通过对大量招聘信息的统计还发现，招聘单位往往会在网页的醒目位置给予提示，例如，作为网页的标题或者单位名字位于网页中单独一行，这就会给我们的识别过程中机构名的定界问题有很好的提示作用，这就是前文的HTML(XML)标签过滤部分加上分隔符的部分原因，并且这也正是半结构化文本的处理优势所在。

### 3.4.3 知识库的获取

依据上节给出的机构名识别方式，需要两个知识库的支持：机构名核心词库和机构名录，下面分别说明：

- 机构名核心词库

本库的存在目的在于产生机构名的候选，在待处理文本中顺序扫描到收录在机构名核心词库中的词语时，我们就默认产生了一个机构名的候选，然后再确定该机构名的候选是一个机构名的概率。

考虑到运行时间和机构名的出现频率，本库共收录约15个常见机构名核心词，如“公司”、“有限公司”、“研究所”、“设计院”、“集团”和“厂”等，详见附录三。

- 机构名录

机构名录是计算机构名生成概率的主要工具，本名录来源于中文自然语言处理平台(http://www.nlp.org.cn)，绝大多数是公司名，有22731条记录，部分内容参见附录二。

- 语料库

采用已切分标注过的1998年1月《人民日报》语料库，在该语料库中,机构名全部被显式标注了。按照标注的形式,其中的机构名可以被分为两类：简单机构名（如“致公党/nt”）和复合机构名（如“美国/ns 加利福尼亚/ns 理工学院/n”）。

### 3.4.4 机构名识别

具有了机构名识别理论基础和获取了相关的知识库后，就可以进行机构名的识别工作了。算法描述如下：($\partial$、$\sigma_1$、$\sigma 2$都是统计值，经过大量测试后取值分别为0.05、0.08、0.00075时识别效果最好)。

1) 读入分词后文本；

2) 从机构名核心词库中取一条机构名核心词，如果已达词库末尾则结束；

3) 在读入文本中扫描该核心词，如果找不到则转步骤2，否则记下位置p1，然后向前扫描到分隔符处并记下位置p2，于是在位置p1和p2间的字符产生一个机构名候选；

4) 按照公式(3-1)计算$P_1(C,W)$和$P_2(C,S)$；

5) 如果$P_1(C,W)>\sigma_1$并且$P_2(C,S)>\sigma 2$，则保存该识别出的机构名，否则转步骤2。

以图4-8词性标注后的文本为例，并且当前所取的机构名核心词为“有限公司”，在算法描述的步骤3中将会产生：“广州市<ns>日立<nz>电梯<n>有限公司<n>”的机构名候选。通过机构名录和语料库分别得到如下数据：

Nc(广州市)=21；Nc(<ns>)=3734；Nt(广州市)=10+21；Nt(<ns>)=31633；

Nc(日立)=7；Nc(<nz>)=1678；Nt(日立)=8；Nt(<nz>)=5379；

Nc(电梯)=6；Nc(<n>)=71689；Nt(电梯)=23；Nt(<n>)=383952；

Nc(有限公司)=22335；Nc(<n>)=71689；Nt(有限公司)=188+22335；

Nt(<n>)=312263+71689

计算得

$P_1$ =(0.95*21/31)*(0.95*7/8)*(0.95*6/23)*(0.95*22335/22523)= 0.1315

$P_2$ = (0.95*3734/31633)*(0.95*1678/5379)*

(0.95*71689/383952)*(0.95*71689/383952)

= 0.00105

通过和既定阈值$\sigma_1$和$\sigma 2$比较而确定“广州市日立电梯有限公司”是一个机构名，本次识别完成。具体测试部分见第五章。

## 3.5 地名识别

### 3.5.1 中文地名识别的现状

与其他中文命名实体相比，中文地名识别相对比较少。文献[35] 构建了一个基于最大熵原理的汉语人名地名自动识别混合模型，并达到了比较满意的识别效果；文献[36] 以词语级的中文地名为识别对象,根据地名内部用字的统计信息和地名构成特点产生潜在地名，然后对句子进行切分，并在确定句子最佳切分时识别句子中的中文地名，也取得了较好的识别效果。

### 3.5.2 中文地名的特点

中文地名主要有以下特点[35][36]：

1) 中文地名数量大，没有明确规范的地名定义，并不断有新的地名出现。

2) 地名结尾经常有地名特征词出现，如“省、市、区”。但地名特征词出现的情况比较复杂：既可以作为普通用词出现，又可以出现在地名的其它位置。

3) 地名长度没有严格限制，短的如“京”，长的如“陕西省西安市郭杜镇小居安村”。

4) 单字词在地名中经常出现，如“南| 稍| 门”。

5) 地名有时与一些介词、动词、方位词之类的指示词一起出现，但有些指示词也可以作为地名组成部分。

6) 经常多个地名一起出现，如：“河南省| 渑池县| 果元乡| 峪峒村”。

正是由于以上特点增加了中文地名识别的难度。

### 3.5.3 中文地名识别模型

#### 3.5.3.1 基本定义

类似于前文的机构名识别，结合前人所做的工作，我们给出了地名识别模型的相关定义：

**定义1** 设SpNameSpecialWord为地名特征词表，SpNameChar为地名前部词表。

则中文地名定义为：$SP = F_0F^+S$

其中 $F^+ = F_1...F_n, Fi \in \text{SpNameChar}(i=1,...,n)$，$S \in \text{SpNameSpecialWord}$。$F_0$定义为地名首字，$F^+$为地名中部，$F_0F^+$统称为地名前部词，$S$为地名特征词(如：省、市等)，即地名是由地名前部词和地名特征词组成的。

根据是否可以作为地名的前部词，地名特征词分为：

1) 只能作为地名特征词而不能作为地名前部词(“省”、“开发区”、“三角洲”)

2) 既能作为地名特征词，又能作为地名前部词(“江”、“湖”、“岗”)根据组成地名的长度，地名特征词又可以分为：

I 组成的地名可以少于三个单字长度(“县”、“山”、“盟”)

II 组成的地名至少三个单字长度(“路”、“观”、“坡”)

根据在地名中出现的位置，地名前部词可以为：

1) 不能作为地名首字的词(“满族”、“现”、“敢”)

2) 不能作为地名中部的词，这样的一般也多为多字词(“黄粱梦”)根据与特征词的关系，地名前部词又可以为：

I 不能单独和特征词连用作为地名(“可”、“并”、“个”)

II 只能和特征词连在一起用，这样的一般为多字词(“平等”、“中央”、“胜利”)

**定义2** 地名前词是指在真实文本中地名的前一个词。如“在| ～大连市～|”，“在”是地名“大连市”的地名前词。其分类如下：

肯定前词：地名识别过程中，遇到该词不需要继续向前搜索，如“针对”、“关于”等。

可能前词：地名识别过程中，遇到该词还需其他的信息来判断究竟是不是地名前词，这是地名识别的难点。

**定义3** 地名后词是指在真实文本中地名的后一个词。

**定义4** 常规切分是指不含地名识别的分词；按地名切分是指含地名识别的分词。

**定义5** 地名构词可信度是根据地名用字用词情况，判断它作为地名的可信度；

地名接续可信度是根据地名的构成及其上下文的接续关系来判断它作为地名的可信度。

**定义6** 地名构词评价系数是指根据地名用词之间搭配情况，对其构词可信度或增大或减小的评价系数；地名接续评价系数是指根据地名的构成以及其上下文对其接续可信度或增大、或减小的评价系数。

#### 3.5.3.2 识别资源

本文根据《中国地名录》[37](含地名约4万条)，建立了地名特征词表和地名前部词表。其中，地名前部词表记录为3655条，地名特征词表记录为127 条。另外，根据中科院计算所ICTCLAS分词系统的切分词典(来自于1998年1月份人民日报语料)建立了单词频度词典和双词接续词典，本项工作可以通过程序自动化处理。

#### 3.5.3.3 算法模型

根据上述定义，给出本模型的计算公式：

1 地名特征词的可信度 $P_l(S)$

$$P_l(S)=\frac{P_{l0}(S)}{\sum_{y\in SpNameSpecialWord} P_{l0}(y)}$$

$$P_{l0}(S)=\log_2(D(S)+2)$$

公式 (3-2)

$D(S)$ 是字串S作为地名特征词在中国地名库中出现的次数。

2 地名首字可信度 $P_h(F_0)$

$$P_h(F_0)=\frac{P_{h0}(F_0)}{\sum_{y\in SpNameChar} P_{h0}(y)}$$

$$P_{h0}(F_0)=\log_2(C(F_0)+2)$$

公式 (3-3)

$C(F_0)$ 是字串$F0$ 作为地名首字在中国地名库中出现的次数。

3 地名中部可信度 $P_f(F^+)$

$$P_f(F^+)=\sum_{i=1}^{n}1/2^{n-i-1}\frac{P_{f0}(F_i)}{\sum_{y\in SpNameChar}P_{f0}(y)}$$

公式 (3-4)

$$P_{f0}(F_i)=\log_2(C(Fi)+2)$$

$C(Fi)$ 是字串 $Fi$ 作为地名中部在中国地名库中出现的次数。考虑到每个字离地名特征词距离的远近影响整个字串是地名的中部的可能性,故公式前面加了系数 $1/2^{n-i-1}$ 。

4 地名构词可信度 $P'_w(SP)$

$$P'_w(SP)=P_l(S)\times(P_h(F_0)+P_f(F^+))\times(1+P'_{ap}(SP))$$ 公式 (3-5)

其中： $P_l(S)$ 是公式3-2计算出来的地名特征词S作为地名的可信度。$P_h(F_0)$ 是公式(3-3)计算出来的F0作为地名首字的可信度。$P_f(F^+)$ 是公式(3-4)计算出来的 $F^+$ 作为地名中部的可信度。$P'_{ap}(SP)$ 为地名构词评价系数。用来作为地名内部词与词之间搭配的衡量尺度。$P'_{ap}(SP)$ 是以规则的形式来描述的，分为奖励规则和惩罚规则在之间取值。如：

奖励规则：

1) 潜在地名中，如果词与词之间的接续在双词接续词典中没有出现。

2) 潜在地名中，如果词是未定义词(没有词性的单汉字)或者是名词。

3) 有一些比较少用的词如果出现在潜在地名中，作为地名用字的可能性较大。

惩罚规则：

4) 在地名中，词与词之间的接续在双词接续词典中出现，则要进行相应的惩罚。

5) 如果潜在地名中出现的词性接续是<量词+ 数词+ 名词>，或是<动词+ 副词>等短语结构且后面不是特征词或名词。

6) 如果潜在地名中出现高频词，且词的长度超过1个汉字长。

7) 在地名中出现多个不能和特征词结合形成地名的单字词(见地名前部词分

类)。

对于地名链中的每一潜在地名，根据满足的不同规则，对其 $P'_{ap}(SP)$ 进行鼓励或惩罚，例如，“临| 潼| 区”，满足规则1)的条件，对 $P'_{ap}(SP)$ 增加0.1，同时，还满足规则2)，$P'_{ap}(SP)$ 增加0.1 。又如“| 合| 乡| 并| 镇|”符合规则7),进行惩罚后就从潜在地名链中过滤掉。

识别算法描述如下：

1) 初始化。对输入文本按常规切分得到单词序列$W_1$；

2) 从右向左扫描该单词序列，根据SpNameSpecialWord词表、SpNameChar词表、Unigram与Bigram 词典建立潜在地名链。(注意：一个句子可能不只包含一个中文地名，且随着中文地名边界如前词或特征词的取法不同，可能存在多个相互交叉的潜在地名)；

3) 根据公式(3-5)计算每一个潜在地名的构词可信度 $P'_w(SP)$ 。

4) 扫描潜在地名链，当潜在地名的构词可信度 $P'_w(SP) < \sigma$ 时，则从将该潜在地名从地名链中删除(这里的 $\sigma$ 为潜在地名的阈值，经过测试后给取值为0.03)。

5) 对连在一起的多个地名进行合并。重新建立含地名识别的单词切分序列$W_2$。

由于地址识别与机构名识别采用的语言模型基本一致，这里不再给出示例，具体评测部分见第四章。

## 3.6 联系方式识别

职位名的识别相对于前文的机构名和地址来说难度要小得多，因为职位名的数目有限，是可以列举的，另外，职位名相对稳定，很少变化。我们可以构造一个尽可能全面的职位名列表来进行职位的匹配，但这并不总是个有效的办法，例如对于“程序员”这个职位，我们会发现在相当一部分 IT 行业的招聘信息中往往会以如下形式出现：“Web 程序员”、“Java 程序员”、“C++程序员”等，而在其他行业中也会有类似的形式如“语文教师”、“数学教师”、“物理教师”等。我们可以在职位列表里包含诸如“程序员”和“教师”等职位名核心词，但我们不能包

含所有的细分职位名如“*程序员”和“*教师”等。所以为了更为精确的识别一个职位名，我们还需要更为有效的方法。

我们要重点考虑的根据职位名核心词进行职位名识别的方法基本上可以通过三个方面入手，半结构化文本的特点以及职位名词法信息和某些关键性词的提示，下面分别论述：

- 半结构化文本对识别过程的帮助

前文已经多次提及我们的处理对象是含有招聘信息的网页，它是半结构化文本，在结构上能够对我们的处理提供便利。通过分析招聘信息我们还发现，招聘单位同样会把要招聘的职位以清晰的格式列出(这里清晰的格式包括加粗，斜体，单独一行等形式)，这种结构标志在 HTML(XML)标签过滤阶段已经通过分隔符的形式包含在处理过的文本中，我们可以利用这些结构性提示信息。例如图 3-4 的招聘信息，要招聘的职位“工艺工程师”位于表格中并且单独成为一行，所以当通过职位名核心词驱动而找到“工程师”的时候，通过向前扫描到前面的分割符处，也就准确的识别出了“工艺工程师”这个职位名。

- 词法信息对识别过程的帮助

职位名和机构名类似，也是由一个或一个以上的描述性词加上职位名核心词组成。前者是定语，后者是中心语。从句法角度看，职位名称属“定语+名词性中心语”型的名词短语。简称定名型短语。这类短语的定语部分更特殊，一般只能含名词、代词、形容词、动词等，语法结构很简单。

另外，职位名不会含有任何形式的符号如：”，。：《》”等。因为我们要处理的是中文职位名，所以即便职位名中含有英文单词也是个别情况，可以特殊对待。

基于职位名这样一个词法特征，我们可以借助它进行职位名的合理排除。例如如果识别出“招聘职位：工艺工程师”，我们就可以根据规则二“职位名不会含有任何形式的符号”而给予排除，从而提高召回率。

- 关键词提示对识别过程的帮助

招聘信息中有不少关键词对于职位名的识别具有提示作用，例如：“招聘职位”，“职位如下：”等短语片段指明了附近就是招聘的职位名。我们可以通过引入一些类似的规则来进行职位名的识别。

根据以上的分析可以把识别算法描述如下：

1) 根据职位名列表来匹配待处理文本中出现的职位名，如果匹配成功，则保存识别出来的职位名(识别资源职位名列表从网上搜集得到，共有 438 条职位名，部分内容见附录四)；

2) 以职位名核心词为驱动再次在文中遍历，根据文本中的分隔符进行职位名的识别(职位名核心词表收录了若干条，如“程序员”、“工程师”等)；

3) 对识别出的职位名通过词法信息和关键词的提示进行排除，然后与步骤(1)识别出的职位名相比较，如果不重复则保存该职位名并重复步骤(2)和步骤(3)直至职位名核心词表为空。

职位名识别的测试部分见第四章。

## 3.7 联系方式识别

一条完整的招聘信息除了单位名，职位名以及地址外，联系方式也是不可缺少的。所以，对于招聘信息的抽取还需要实现联系方式的识别。联系方式包含电话号码和 E-mail 两部分内容，都宜采用基于规则的方法来识别，详述如下：

### 3.7.1 电话号码的识别

通常的电话号码书写形式是这样的：029-88301722 或者(029)88301722，其中 029 是区号，剩余部分就是受话号码。很容易发现电话号码的构词特征：

1 由两部分构成：区号(由 3 个或 4 个数字构成)和受话号码(由 7 个或者 8 个数字构成)；

2 只能由“0,1,2…9”这十个数字组成；

3 区号对电话号码的识别具有驱动作用并且对地址的识别具有反馈作用。

经过上面的构词分析，可以得出电话号码的识别采用统计加规则的方法来进行，采用统计的方法由区号开始驱动识别，然后附以规则判断进行识别。特别注意的是上面的构词特征 3 的分析，区号对前面部分地址的识别具有的反馈作用来提高识别准确率。识别资源主要是各地的区号列表(见附录五)，共收录全国 83 个市区的区号，格式如：“北京市 010，上海市 021”等，前面的市区名主要用于对地址的反馈和纠正。

识别算法如图 3-9 所示：

### 3.7.2 E-mail 的识别

E-mail 的识别和电话号码的识别类似，适合采用基于规则的方式并附以一定的统计信息，下面分析 E-mail 的格式：

1 一个完整的 E-mail 由两个部分组成，格式如下：loginname@full host name.domain name 即：登录名@主机名.域名。其中登录名只能由英文字母、数字、下划线组成；

2 E-mail 的域名部分具有可统计性，对 E-mail 的识别有提示作用；

3 由于 E-mail 里面必须包含字符“@”，所以字符“@”对 E-mail 的识别能起到驱动作用。

识别算法略。

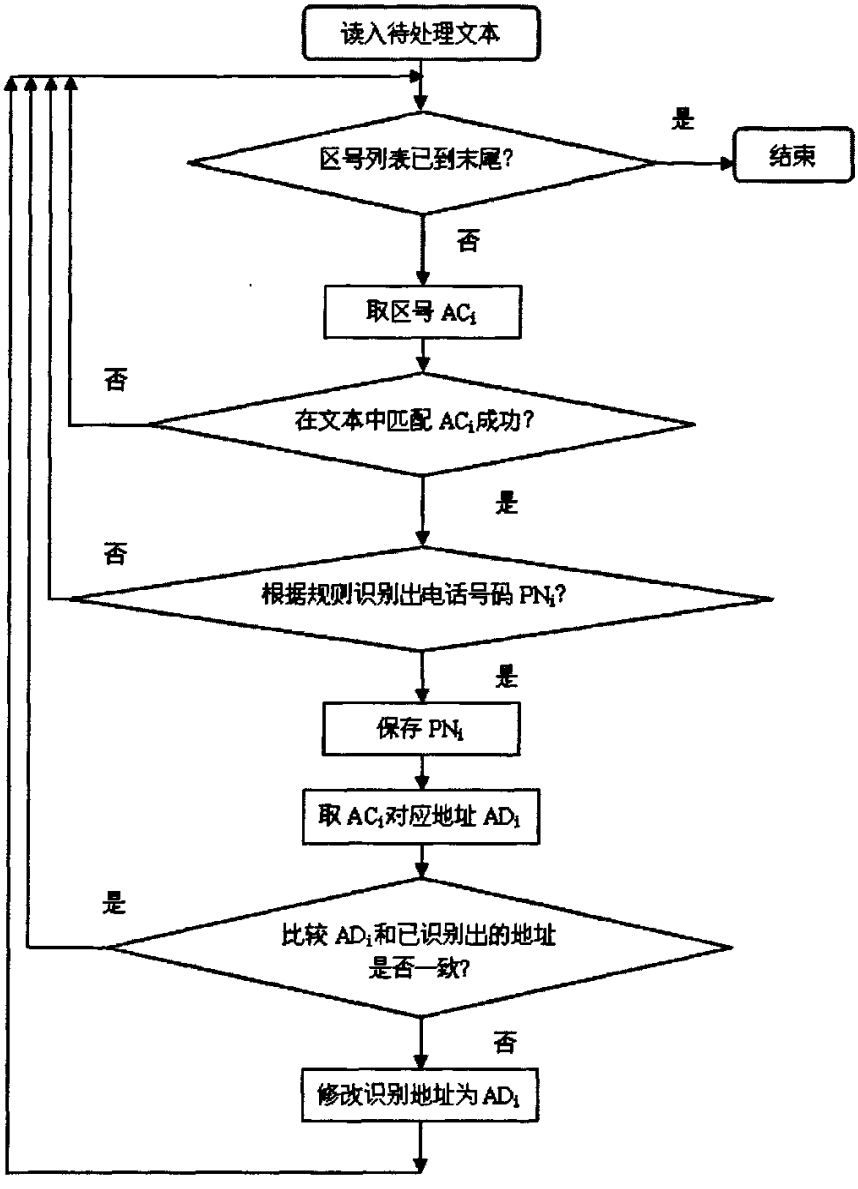


图 3-9 电话号码的识别

## 3.8 关系抽取

命名实体识别完成后，进行信息抽取的第二个阶段：关系抽取[38]。近年来，关系抽取的研究热点是基于支撑向量机(Support Vector Machine, SVM)[39][40]的分类方法，它把实体关系的抽取形式化为二元分类问题，在一些英文处理领域中是一种方法。而中文信息抽取的研究目前主要集中在命名实体识别阶段，关系抽取

的正确率和召回率一般都不高[41]，所以本文不对关系抽取展开研究，只是结合实际给出一个合理化的假设来实现关系抽取。

假设：对于招聘网站的任意网页，如果在命名实体识别阶段识别出本网页含有一个(仅一个)机构名，一个以上(含一个)的职位名，那么该网页就是个招聘页面，识别出的这个机构名和若干职位名之间就是招聘关系(Employ-of)。

该假设能以极少的处理时间换取同等的抽取效果。很容易证明上述假设的正确率至少在 50%以上，而目前对于中文信息抽取来说无论采用什么样的关系抽取方式，很少有抽取的正确率超过 50%[42]，并且还要进行大量的数据处理，所以，基于这样一个假设来进行关系抽取是合理的。

# 第四章 系统的实现与测试

## 4.1 系统的实现

本系统(取名为 JobHunter)属于 B/S 架构，其中 Server 端在 Visual C++ 6.0 环境下实现，在 Windows XP 和 Windows 2003 上都正常通过测试，其组成框架见图 4-1。可以看到 Server 端由 Spider 部分和 Name Entity Recognition 两部分组成，其运行状态如下：首先 Spider 模块根据用户指定的 URL(要求该 URL 最好是个招聘网站的网址)开始抓取网页并保存在默认文件夹中，直至该网站所有网页被抓取完毕。在 Spider 开始抓取网页的同时，Name Entity Recognition 模块也开始启动，依次通过 HTMLParser 部分过滤掉网页中的标记，WordTag 部分对文本内容进行分词和词性标注。接着进行命名实体识别，并把识别出的命名实体保存在数据库中。JobHunter 的 Server 端的运行界面如图 4-2 所示。

JobHunter 的 Browser 端由 ASP 实现，Browser 端的用户可以在浏览器中浏览并查询招聘信息，浏览器中显示的抽取后的招聘信息如图 4-3 所示。

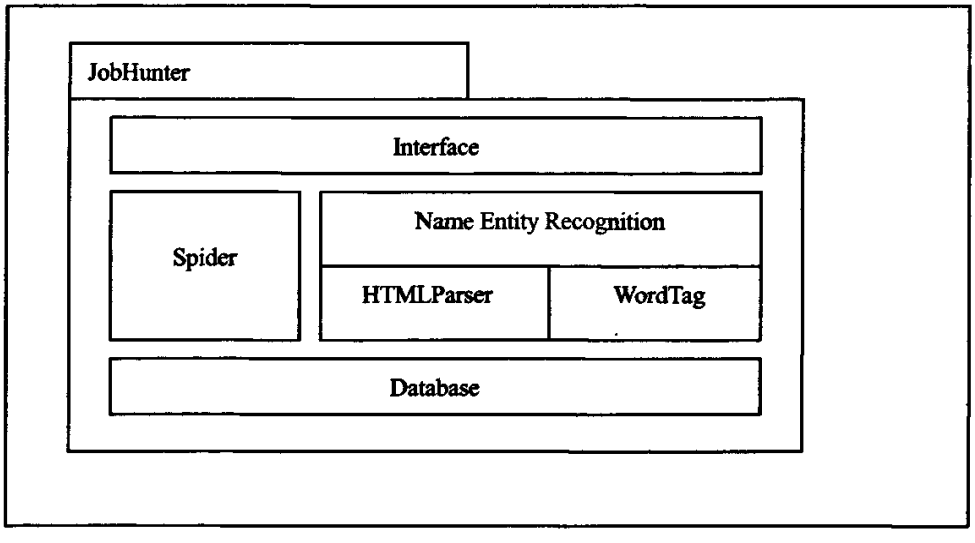


图 4-1 Web 信息抽取系统的模块组成

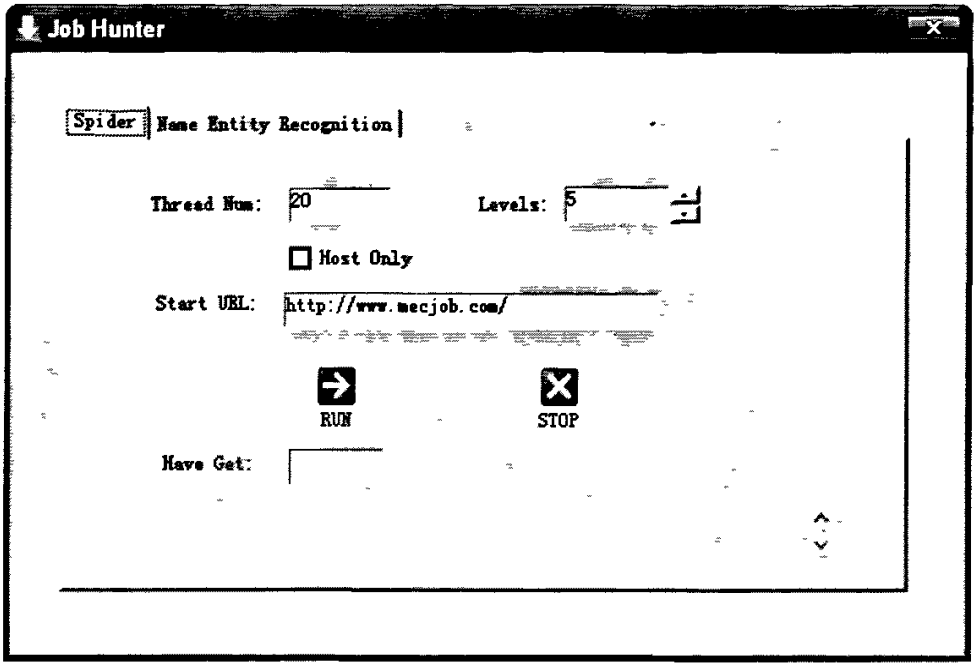


图 4-2 JobHunter Server 端的运行界面


图 4-3 JobHunter Browser 端的显示界面

## 4.2 系统的测试

### 4.2.1 Spider 部分

Spider 部分的运行界面如图 4-4 所示。界面元素如下：Thread Num 指定 Spider 线程个数，默认为 20；Levels 指定抓取深度，默认为 5；Host Only 复选框为了扩展选择是否仅限于指定网站；Start URL 是 Spider 搜索的初始地址。特别指出，在 Spider 实现部分，对于当前抓取的网页保存其 URL，以便 Browser 端用户通过它获得更完整的招聘信息。

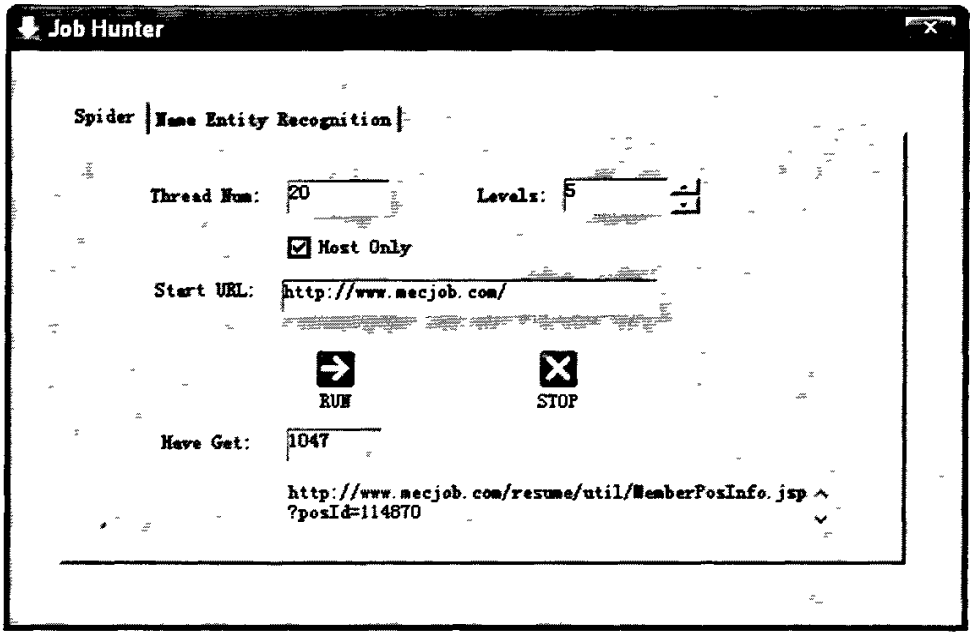


图 4-4 Spider 部分运行界面图

Spider 的运行效率见表 4-1。

表 4-1 Spider 的运行效率对比图

| 起始地址 | 线程个数 | 抓取深度 | 运行时间(秒) | 抓取页面数 | 抓取速度(页面/秒) |
|---|---|---|---|---|---|
| http：//www.51job.com | 10 | 5 | 270 | 1487 | 5.5 |
| http：//www.mecjob.com | 20 | 5 | 311 | 3030 | 9.7 |
| http：//www.mecjob.com | 30 | 8 | 230 | 2491 | 10.8 |
| http：//www.51job.com | 40 | 10 | 190 | 1198 | 6.3 |

### 4.2.2 Name Entity Recognition 部分

命名实体识别部分是整个系统的核心，运行界面如图 4-5 所示。界面元素介绍如下：

URLs Directory 给出 Spider 保存网页的文件夹位置。

在 Study mode 下，系统每识别出一个命名实体都会给 Server 端用户一个反馈，由他来决定该命名实体是否识别正确，从而决定是否将识别出的命名实体放入知识库来扩充训练样本，进而提高识别的准确度。而 Recog mode 是系统的识别模式，通过前文给出的命名实体识别方式把识别出的命名实体存入后台数据库。

其余的界面元素是数据的显示部分。包括待处理的 URL 总数(Total File)、已处理的 URL(Have Processed)、显示处理进度的进度条(Recog Status)以及处理速度(Seconds/Page)。

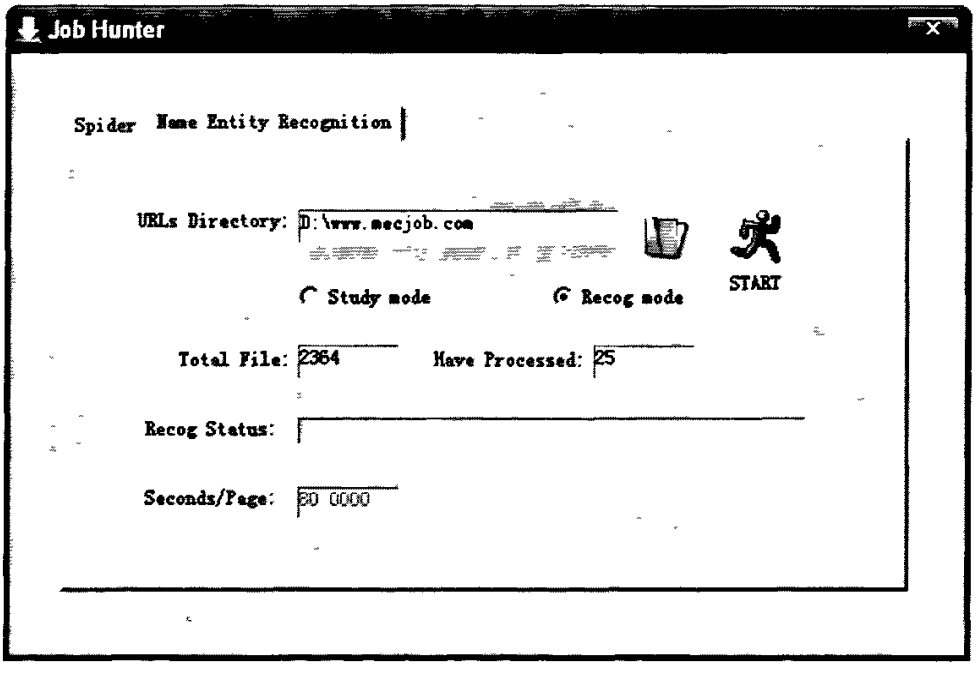


图 4-5 Name Entity Recognition 运行界面

测试数据的选择来自多个招聘网站(http：//www.mecjob.com, http：//www.51job.com, http：//yingjiesheng.cn 等)的 300 个网页，其中招聘信息有 198 条(没有完全选取招聘信息是为了客观衡量识别的正确率)，机构名有 231 个，地址有 132 个，职位名 207 个(含重复)，电话号码 52 个，E-MAIL 有 13 个。评测

标准按照惯例采用自然语言处理中使用最广泛的两个评测指标：正确率(precision)和召回率(recall)，它们分别定义如下：

正确率　　P＝A/(A+B)*100%

召回率　　R＝A/(A+C)*100%

其中 A 代表抽取出的信息个数，B 代表错误抽取的信息个数，C 代表未抽取出的信息个数。则 A+B 代表抽取出的信息总数，A+C 代表信息总数。另外，测试部分分为命名实体识别测试和招聘信息的抽取测试。测试结果如下：

- 机构名识别测试(见表 4-2)：

表 4-2　JobHunter 对机构名识别的测试结果

| 阈值$\sigma_1$ | 阈值$\sigma_2$ | 正确识别的机构名个数 | 错误识别的机构名个数 | 正确率 | 召回率 |
|---|---|---|---|---|---|
| 0.07 | 0.0007 | 185 | 43 | 81.12% | 80.09% |
| 0.07 | 0.00075 | 187 | 42 | 81.54% | 80.95% |
| 0.08 | 0.00075 | 191 | 39 | 83.21% | 82.68% |
| 0.08 | 0.0008 | 187 | 39 | 82.75% | 80.95% |

经过多次测试发现，识别错误的部分机构名如下(绝大多数识别错误出现在机构名的定界上)：

(1) 重机有限公司(北京市三一重机有限公司)

(2) 纬电源科技有限公司(惠州亿纬电源科技有限公司)

(3) 转向器有限公司(恩斯克(NSK)转向器有限公司)

为了便于对比，我们把前面的 300 个网页过滤掉 HTML 标签，然后依次经过哈工大信息检索教研室的语言技术平台(http：//ltp.ir-lab.org，这是国内首家提供公开评测的自然语言处理工具)进行命名实体识别并统计结果，见表 4-3：

表 4-3 哈工大语言技术平台对机构名的识别结果

| 正确识别的机构名个数 | 错误识别的个数 | 正确率 | 召回率 |
|---|---|---|---|
| 165 | 45 | 78.57% | 88.71% |

通过对比可以看出，哈工大信息检索教研室的语言技术平台识别的召回率高于我们的 JobHunter，而 JobHunter 的准确率要高些。这个原因通过分析主要在于如下两方面：

1) 系统的训练集目前还比较少，并且训练集的构成没有经过科学挑选，这与本人的时间与经历有关，希望下一步进行这方面的完善；

2) 没有经过充分的训练，进行系统测试时，JobHunter 还没有经过学习过程，知识库还不够丰富，这也是造成召回率低的重要原因。

- 地名识别(见表 4-4)：

表 4-4 JobHunter 对地名的识别测试结果

| 阈值$\sigma$ | 正确识别的地址个数 | 错误识别的地址个数 | 正确率 | 召回率 |
|---|---|---|---|---|
| 0.04 | 86 | 34 | 71.54% | 65.15% |
| 0.06 | 86 | 33 | 72.09% | 65.15% |
| 0.08 | 88 | 30 | 74.78% | 66.67% |
| 0.1 | 87 | 32 | 73.27% | 65.91% |

地名识别的难点与出错点大多在于地名片断的识别及合并，有些是地址名片断识别错误，有些是合并错误。由于地名识别部分错误较多，这里不给出识别错误的例子。

同样，我们也在哈工大信息检索教研室的语言技术平台上测试了对地名的识别情况，见表 4-5。

表 4-5 哈工大语言技术平台对地名的识别结果

| 正确识别的机构名个数 | 错误识别的个数 | 正确率 | 召回率 |
|---|---|---|---|
| 57 | 28 | 67.06% | 54.81% |

- 其他命名实体的识别

职位名、电话号码以及 E-mail 的识别主要采用基于规则的方法，正确率和召回率都在 90%以上，识别效果较好。

- 招聘信息的识别

根据前文给出的关于关系抽取的假设：如果一个页面中含有唯一的一个机构名、若干个职位名，则认为该页面是个招聘页面而不再进行关系抽取。于是，招聘信息的识别依赖于机构名和职位名二者的识别效果，而职位名识别的正确率和召回率都在 90%以上，所以，招聘信息的识别效果主要取决于机构名的识别。测试分析见表 4-6：

表 4-6 JobHunter 对招聘信息的识别测试

| 阈值 $\sigma_1$ | 阈值 $\sigma_2$ | 正确识别的机构名 | 正确识别的职位名 | 正确识别的招聘信息 | 错误识别的招聘信息 | 正确率 | 召回率 |
|---|---|---|---|---|---|---|---|
| 0.07 | 0.0007 | 185 | 193 | 141 | 41 | 77.56% | 71.21% |
| 0.07 | 0.00075 | 187 | 193 | 142 | 39 | 78.23% | 71.72% |
| 0.08 | 0.00075 | 191 | 193 | 143 | 37 | 79.16% | 72.22% |
| 0.08 | 0.0008 | 187 | 193 | 142 | 37 | 79.08% | 71.72% |

通过上面的一系列测试可以看出，JobHunter 在没有通过 Study mode 训练的的情况下，作为一个应用系统，它的信息抽取准确率和召回率均在 70%以上。在界面呈现部分还提供了必要的提示信息，如在显示给用户信息源的 URL，便于用户进行重点关注。总体来说，它可以满足人们的在 Web 招聘信息获取方面的基本需要。

# 第五章 总结与展望

## 5.1 论文工作总结

本文采用基于自然语言理解的方式进行 Web 中文信息抽取。通过构建信息采集模块(Spider)、HTML/XML 标记过滤模块、信息抽取模块以及多个知识库，采用基于自然语言理解的方式来实现信息抽取系统，并通过大量的实验检验了该方式的抽取效果。

本文的主要工作如下：

1) 解决了基于自然语言理解方式进行 Web 信息抽取时对处理半结构化文本的不足；

2) 改进了现有的语言模型并应用于命名实体识别，取得了较好的识别效果；

3) 构建有效的 Spider 模块，给后面的信息抽取工作打好了基础；

4) 实现了 HTML/XML 标记过滤模块。不仅很好地过滤掉 HTML/XML 标记，而且保留了半结构化文本的信息，为信息抽取模块的识别工作带来很大方便；

5) 完成了信息抽取系统的核心部分——信息抽取模块。该模块在大量的知识库支撑下，通过基于自然语言理解的方式能够较好地识别机构名、地址、职位名以及联系方式等命名实体。

## 5.2 展望

在信息抽取系统的进一步研究中，有两个主要的发展方向：

1) 提高系统向其他新领域的可移植性；

2) 提高系统的性能。

传统的手工编制规则方式是领域相关的，虽然能保证一定的效率，但可移植性很差。本文的研究正是为了从很大程度上解决这个问题而采用自然语言理解的方法，从理论上讲，只要有合适的知识库支撑，那么对于新领域的信息抽取工作就可以很容易做到，所以，在可移植性方面可以考虑更新为别的信息抽取系统，

例如通过对系统进行产品名录的训练可以扩展其在电子商务中的应用。可以预见，随着自然语言处理能力的逐步加强，基于自然语言理解的信息抽取方法是未来的大势所趋。

基于自然语言理解的方式虽然具有良好的可移植性，同时，我们也注意到，它的预处理工作包括标记的过滤以及分词等以及识别过程本身的时间复杂度都是 $O(n^2)$，但是信息抽取本身对及时性的要求远不如信息检索，所以目前来说运行时间的问题不是主要问题。

由于目前的自然语言理解还仅仅停留在词法、语法阶段，还远没有达到语义方面的真正的自然语言理解，所以，基于自然语言理解的信息抽取方法受限于这个大环境，在抽取的正确率和召回率方面还没有达到我们的期望值。虽然从长远看来，自然语言理解的方法能给信息抽取技术带来质的突破，但是目前要想提高信息抽取系统的性能，只能结合别的方法，采用进一步优化的方式来逐步提高系统的性能。

# 参考文献

[1] Line Eikvil 著，陈鸿标译．网上信息抽取技术纵览, 2003

[2] Alberto H. F. Laender, Berthier A. RibeiroNeto, Altigran S. da Silva, Juliana Teixeira. A Brief Survey of Web Data Extraction Tools. SIGMOD Record, 2002,31(2): 84-93

[3] Kiyoshi sodo. computer Information science department Univercity Extracted Japanese Extraction with New York Automaticly Patterns, 2004

[4] 孙斌．文本信息提取技术.ppt．北京大学计算机系计算语言所，2005

[5] N. Chinchor, E. Marsh. MUC-7 Named Entity Task Definition. In Proceedings of theSeventh Message Understanding Conference (MUC-7), Available at http: //www muc.saic.com/

[6] http://www itl.nist.gov/lad/894.01ltests/ace/

[7] Bing Liu, Yiming Ma, Philip S. Yu. Discovering Unexpected Information from Your Competitors' Web Sites. SIGMOD, 2001

[8] 张玲．WEB信息提取技术研究与应用：[硕士学位论文]．北京：中科院计算技术研究所，2003

[9] 史忠植．知识发现．北京：清华大学出版社，2002

[10] Sahuguet A，Azavant F. Building intelligent web applications using lightweight wrappers. Data and Knowledge Engineering, 2001,Vol. 36(3): 283-316

[11] Liu L, Pu C, Han W. XWRAP: An XML-enable Wrapper Construction System for WebInformation Sources. Proceedings of the 16th IEEE International Conference on Data Engineering, California, 2000: 611-621

[12] Crescenzi V, Mecca G, Merialdo P. RoadRunner: Towards Automatic Data Extraction from Large Web Sites. Proceedings of the 2e International Conference on Very Large Database Systems. Rome, 2001: 109-118

[13] 郑家恒，刘开瑛．自动分词系统中姓氏人名的处理策略探讨，1993

[14] 陈少飞，郝亚南，李天柱等．Web信息抽取技术研究进展．河北：河北大学学报(自然科学版)，2003，Vol. 23(l)：106-112

[15] 愈鸿魁，张华平．基于层叠隐马尔科夫模型的中文命名实体识别．全国网络与信息安

全技术研讨会，2005

[16] S. Debnath, P. Mitra, and C. Giles. Automatic Extraction of Informative Blocks from Webpages. In Proc. of the Special Track on Web Technologies and Applications in the ACM Symposium of Applied Computing, 2005:1722-1726

[17] Yanhong Zhai, Bing Liu, Extracting Web Data Using Instance-Based Learning. in Proceedings of 6th International Conference on Web Information Systems Engineering, 2005: 318-331

[18] （美）Heaton J. 高克宁，柴桥子．支持Web分类的高性能蜘蛛程序．小型微型计算机系统，2006，7

[19] 徐宝文．搜索引擎与信息获取技术．北京：清华大学出版社, 2003

[20] 骆斌，费翔林．多线程技术的研究与应用．计算机研究与发展，2000，Vol.37(4)：407-412

[21] Cameron Hughes Tracey Hughes著，肖和平，张杰良译．C++并行与分布式编程．北京：中国电力出版社，2004

[22] 李盛． 面向真实文本的汉语词义排歧系统：[硕士学位论文]．太原：山西大学，2004

[23] 刘开瑛．中文文本自动分词和标注．北京：商务印书馆，2000

[24] 张华平，刘群基于N-最短路径方法的中文词语粗分模型．中文信息学报，2002,5

[25] 李向阳，张亚非．一种网上图书信息抽取方法．情报学报，2004，Vol. 23(6)： 655-660

[26] 崔桓，蔡东风，苗雪雷．基于网络的中文问答系统及信息抽取算法研究．中文信息学报，2004，Vol. 18(3) ： 24-32

[27] Hong I Ng and Kim Teng Lua. A Word Finding Automation for Chinese Sentence Tokenization, submitted to ACM Transaction of Asian Languages Processing, 2002

[28] 罗智勇，宋柔，现代汉语自动分词中专名的一体化、快速识别方法．新加坡：国际中文电脑学术会议，2001：323-328

[29] 陈小荷，现代汉语自动分析．北京：北京语言文化大学出版社, 2000：104-114

[30] 张小衡，王玲玲．中文机构名称的识别与分析．中文信息学报，1997，11 (4) ： 21－32

[31] Wang Houfeng, ShiWuguang. A simple rule-based approach to organization name recognition in Chinese text [A] . Proc of 5th Cycling[C ]. LNCS 3406, Heidelberg, German: Springer2 Verlag, 2005: 769－772

[32] 冯丽萍．基于统计的中文组织机构名识别．福建电脑，2006，5

[33] 王宁，中文金融新闻中公司名的识别．中文信息学报，2005，Vol.16(2)

[34] ChristinaY.C, MichaelC, Neel S. Reverse engineering for web data: From visual to semantic structures. In proceedings of the 18'Th International Conference on Data Engineering San Jose, Califomia, 2002.

[35] 钱晶，张杰，张涛．基于最大熵的汉语人名地名识别方法研究．小型微型计算机系统，2006，9

[36] 高红，黄德根，杨元生．汉语自动分词中中文地名识别．大连理工大学学报，2006，4

[37] 国家测绘局地名研究所．中国地名录．北京：中国地图出版社，1997

[38] 朱永盛，武港山．基于Web的新闻信息抽取．计算机工程, 2006

[39] D.M. Campbell, Y. Ding, D.W. Embley, et al. Demonstration: A Robust Web Data-Extraction Technique with High Recall and Precision. Brigham Young University DEG Technical Report.

[40] 牛成. IE-Hit.ppt．微软亚洲研究院2005年信息抽取技术暑期研讨班，2005

[41] 叶娜，吴雪军．基于相似计算的信息抽取模板自动获取办法．第二届全国学生计算机语言学研讨会论文集，2004

[42] 张清军，朱才连．基于主动学习的Web页面信息抽取．情报学报，2004，Vol. 23(6)：667-671

# 硕士期间参加的课题与发表的文章

孟伟涛，张蕾，张晓孪等. 一种基于位置概率模型的中文人名识别方法. 计算机应用与软件.

# 致谢

在这篇论文完稿之际，回顾在西北大学计算机系三年的研究生生活，我由衷地感谢我的导师张蕾教授的精心指导、严格要求和亲切关怀。她严谨务实的治学态度、独特敏锐的学术直觉和朴素无华的人格力量总在潜移默化中激励着我。在学习上，张老师对我悉心指导，鼓励我大胆创新，抓住机会，扩展研究领域；在思想上，她对我循循善诱，让我时时感到前进的动力；在生活上，她对我关心照顾，给了我很多帮助。再次对张老师的无私教诲和热情帮助表示感谢。

衷心地感谢院、系和教研室的各位领导、教师在我学习期间给予我的培养、教育和帮助。

近三年的学习和课题研究期间，我得到了教研室同学的亲切关怀、大力支持和帮助。感谢我的同学对我的关心和支持，我们将成为永远的朋友。

感谢中科院计算所搭建的自然语言理解平台，它给广大自然语言处理研究者提供了互相交流的环境；感谢张华平等的开源分词程序ICTCLAS为我们带来的便利；感谢SourceForge开源平台的广大同仁。

特别要感谢我的父母，多年来他们给予我极大的理解、鼓励和支持，并为此做出了很多的奉献与牺牲，我的人生道路上每跨出一步，都倾注了他们对我的殷切期盼与无限的爱。

感谢对本论文进行评审并提出宝贵意见的各位专家。

最后，感谢所有关心、支持和帮助过我的老师和朋友们。

# 附录

## 附录一(汉语文本词性标注标记集)

| 代码 | 名称 | 帮助记忆的诠释 |
|---|---|---|
| Ag | 形语素 | 形容词性语素。形容词代码为 a，语素代码 g 前面置以 A。 |
| a | 形容词 | 取英语形容词 adjective 的第 1 个字母。 |
| ad | 副形词 | 直接作状语的形容词。形容词代码 a 和副词代码 d 并在一起。 |
| an | 名形词 | 具有名词功能的形容词。形容词代码 a 和名词代码 n 并在一起。 |
| b | 区别词 | 取汉字“别”的声母。 |
| c | 连词 | 取英语连词 conjunction 的第 1 个字母。 |
| Dg | 副语素 | 副词性语素。副词代码为 d，语素代码 g 前面置以 D。 |
| d | 副词 | 取 adverb 的第 2 个字母，因其第 1 个字母已用于形容词。 |
| e | 叹词 | 取英语叹词 exclamation 的第 1 个字母。 |
| f | 方位词 | 取汉字“方”的声母。 |
| g | 语素 | 绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母。 |
| h | 前接成分 | 取英语 head 的第 1 个字母。 |
| i | 成语 | 取英语成语 idiom 的第 1 个字母。 |
| j | 简称略语 | 取汉字“简”的声母。 |
| k | 后接成分 | |
| l | 习用语 | 习用语尚未成为成语，有点“临时性”，取“临”的声母。 |
| m | 数词 | 取英语 numeral 的第 3 个字母，n，u 已有他用。 |
| Ng | 名语素 | 名词性语素。名词代码为 n，语素代码 g 前面置以 N。 |
| n | 名词 | 取英语名词 noun 的第 1 个字母。 |
| nr | 人名 | 名词代码 n 和“人(ren)”的声母并在一起。 |

| | | |
|---|---|---|
| ns | 地名 | 名词代码 n 和处所词代码 s 并在一起。 |
| nt | 机构团体 | “团”的声母为 t，名词代码 n 和 t 并在一起。 |
| nz | 其他专名 | “专”的声母的第 1 个字母为 z，名词代码 n 和 z 并在一起。 |
| o | 拟声词 | 取英语拟声词 onomatopoeia 的第 1 个字母。 |
| p | 介词 | 取英语介词 prepositional 的第 1 个字母。 |
| q | 量词 | 取英语 quantity 的第 1 个字母。 |
| r | 代词 | 取英语代词 pronoun 的第 2 个字母，因 p 已用于介词。 |
| s | 处所词 | 取英语 space 的第 1 个字母。 |
| Tg | 时语素 | 时间词性语素。时间词代码为 t，在语素的代码 g 前面置以 T。 |
| t | 时间词 | 取英语 time 的第 1 个字母。 |
| u | 助词 | 取英语助词 auxiliary。 |
| Vg | 动语素 | 动词性语素。动词代码为 v。在语素的代码 g 前面置以 V。 |
| v | 动词 | 取英语动词 verb 的第一个字母。 |
| vd | 副动词 | 直接作状语的动词。动词和副词的代码并在一起。 |
| vn | 名动词 | 指具有名词功能的动词。动词和名词的代码并在一起。 |
| w | 标点符号 | |
| x | 非语素字 | 非语素字只是一个符号，字母 x 通常用于代表未知数、符号。 |
| y | 语气词 | 取汉字“语”的声母。 |
| z | 状态词 | 取汉字“状”的声母的前一个字母。 |

# 附录二(部分公司名录)

(香港)飞豪国际投资有限公司； 二十一世纪生活事业股份有限公司； J-M 塑胶有限公司； OK 便利 富群超商股份有限公司； 一友企业股份有限公司； 一心工业有限公司； 一心机械工业股份有限公司；一卡国际 一卡国际科技股份有限公司； 一民塑胶股份有限公司； 一丞冷冻工业股份有限公司； 一全工业股份有限公司； 一全行股份有限公司； 一全兴业股份有限公司； 一名工业股份有限公司； 一宏钢铁股份有限公司； 一廷企业有限公司； 一协荣股份有限公司；一忠建设 一忠建设股份有限公司；一升股份有限公司；一品光学 一品光学工业股份有限公司； 一品光学工业股份有限公司； 一品建材股份有限公司； 一品国际科技股份有限公司； 一品钻石工业股份有限公司； 一恒电脑企业股份有限公司； 一飞印媒体资讯股份有限公司； 一通实业股份有限公司； 一等高 一等高科技股份有限公司；一等高科技股份有限公司； 一华半导 一华半导体股份有限公司； 一华半导股份有限公司；一华半导体股份有限公司； 一阳电子股份有限公司； 一新控股有限公司； 一诠精密 一诠精密工业股份有限公司；一诠精密工业股份有限公司；一路得股份有限公司；一零四资 一零四资讯科技股份有限公司； 一零四资讯股份有限公司； 一澈科技 一澈科技股份有限公司； 一银证券 一银证券股份有限公司； 一亿机器厂股份有限公司； 一德金属工业股份有限公司； 一辉制衣股份有限公司； 一兴开发股份有限公司； 一点零科 一点零科技股份有限公司； 乙元木业股份有限公司； 乙正五金有限公司； 乙光精机厂股份有限公司； 乙先公司； 丁台工业有限公司； 丁立实业有限公司； 丁守企业有限公司； 丁守企业股份有限公司； 丁年豆 丁年豆企业股份有限公司； 七味之友实业有限公司； 七和工具厂股份有限公司； 七和工业股份有限公司； 七和实业股份有限公司； 七星轮胎股份有限公司； 七阳实业股份有限公司； 七盟电子 七盟电子工业股份有限公司； 七盟电子工业股份有限公司；七福工业股份有限公司； 七联重工 七联重工股份有限公司； 七联重工股份有限公司； 乃上电器有限公司； 乃辉企业股份有限公司； 乃兴企业股份有限公司； 九大纺织股份有限公司； 九禾电子股份有限公司； 九如股份有限公司； 九如航空货运承揽股份有限公司；九州铝业股份有限公司； 九邦卫星 九邦卫星通信股份有限公司； 九和汽车股份有限公司；九昱光电 九昱光电科技股份有限公司； 九泰营造 九泰营造工程股份有限公司

# 附录三(部分机构名后缀)

公司，有限公司，研究所，研究院，设计院，集团，中心，大学，厂

# 附录四(部分职位名)

首席执行官,总经理,副总经理,总监,总裁,总经理助理,行政总监,人事总监,人事经理,人事主管,人事专员,人事助理,招聘经理,招聘主管,薪资福利经理,薪资福利主管,薪资福利专员,薪资福利助理,培训经理,培训主管,培训专员,培训助理,行政经理,行政主管,办公室主任,行政专员,行政助理,经理助理,秘书,经理秘书,前台接待,后勤,资料管理员,电脑操作员,打字员,销售总监,销售经理,区域销售经理,客户经理,销售主管,销售代表,销售工程师,销售助理,渠道分销经理,渠道主管,医药代表,保险代理,商务经理,商务专员,商务助理,销售行政经理,销售行政主管,售前售后技术服务经理,售前售后技术服务主管,售前售后技术服务工程师,售后客户服务((非技术))经理,售后客户服务((非技术))主管,售后客户服务((非技术))专员,财务总监,财务经理,财务主管,总帐主管,财务助理,会计助理,会计经理,会计主管,会计,出纳员,财务分析经理,财务分析主管,财务分析员,成本经理,成本主管,成本管理员,审计经理,审计主管,审计专员,审计助理,税务经理,税务主管,融资经理,融资主管,投资项目经理,基金项目经理,投资顾问,证券经纪人,清算人员,高级客户经理,客户经理,客户主管,客户专员,市场总监,广告总监,市场营销经理,市场主管,市场营销专员,市场助理,市场分析,调研人员,产品经理,品牌经理,促销经理,促销员,公关,公关经理,媒介经理,媒介人员,企业发展经理,业务发展经理,企业策划人员,广告策划,广告设计,文员,厂长,总工,总工程师,副总工程师,项目经理,项目主管,项目工程师,营运经理,营运主管,车间主任,生产主管,督导,领班,生产计划协调员,技术主管,工艺设计经理,技工,高级技工,工程师,电气工程师,电子工程师,机械工程师,维修工程师,质量经理,质量主管,质量工程师,质量检验员,测试员,质检员,认证工程师,安全主管,健康主管,工程绘图员,机械制图员,实验室负责人,工程师,化验员,电工,首席技术执行官,技术总监,技术经理,信息技术经理,信息技术主管,信息技术专员,项目经理,项目执行人员,项目协调人员,系统分析员,高级软件工程师,软件工程师,高级硬件工程师,硬件工程师,通信技术工程师,数据库工程师,ERP 技术顾问,软件测试工程师,硬件测试工程师,测试员,技术支持经理,技术支持工程师,系统管理员,网管,系统工程师,信息安全工程师,网站营运经理,网站营运主管,网络工程师,网页设计师,网页制作人员,技术助理,建筑工程师,结构工程师,土建工程师,电气工程师,给排水工程师,暖通工程师,工程造价师,建筑工程管理,工程监理,室内外装潢设计,施工员,房地产开发,房地产策划,房地产评估,房地产中介,物业管理,物流经理,物流主管,物流人员

# 附录五(部分区号列表)

北京市 010
上海市 021
天津市 022
太原市 0351
石家庄市 0311
沈阳市 024
长春市 0431
吉林市 0432
哈尔滨市 0451
南京市 025
合肥市 0551
济南市 0531
杭州市 0571
绍兴市 0575
宁波市 0574
南昌市 0791
九江市 0792
福州市 0591
长沙市 0731
武汉市 027
郑州市 0371
广州市 020
珠海市 0756
海口市 0898
成都市 028
昆明市 0871
西安市 029
咸阳市 0910
兰州市 0931
西宁市 0971
重庆市 023
湖州市 0572
嘉兴市 0573
海宁市 0573
余姚市 0574
舟山市 0580
临海市 0576
椒江市 0576
金华市 0579
兰溪市 0579
丽水市 0578
衢州市 0570
江山市 0570
温州市 0577
义乌市 0579
东阳市 0579
瑞安市 0577
乐清市 0577
台州市 0576
福州市 0591
莆田市 0594
南平市 0599
邵武市 05906
厦门市 0592