# 《网络舆情分析技术说明文档》

## 第一章： 网络舆情技术概述（孔凯传）

### 1.1 网络舆情概述

随着我国网络媒体的快速发展和网民数量的高速增长，论坛、微博、微信群、QQ群等社交平台越来越成为舆论传播的主场，它的全球性、快捷性、开放性和互动性给网络舆情的传播带来了便捷，同时它的隐蔽性和虚拟性又使得舆情传播走向放大化和无序化。网络媒体成为民众表达民义、交流情感的主要窗口，同时也是网络舆情传播的载体。

如何准确把握舆情发展规律并对其进行有效管控，使用网络舆情分析系统，通过自然语言处理、数据分析与挖掘、机器学习等技术对海量舆情数据进行多种分析，从中发现重点、热点和事件的各种特征、知识及发展趋势，对互联网内容进行治理，有利于维护网络信息交往秩序，促进互联网健康发展，从而营造风清气正的网络空间，保障网络社会的安全。

#### 1.1.1 舆情与舆论

‘舆’的含义是指民众或公众，‘情’的含义是情绪或意愿，‘舆情’的含义是公众的情感或情绪。

舆情问题涉及社会学、心理学、新闻传播学、政治学等领域，目前还没有一个相对权威、统一的定义：

|  |  |
| --- | --- |
| **社会学** | “在一定的社会空间内，围绕中介性社会事项的发生、发展和变化，作为主体的民众对作为客体的国家管理者产生和持有的社会政治态度” |
| **社会心理学** | “由个人及各种社会群体构成的公众，在一定历史阶段和社会空间内，对自己关系或者自身利益紧密相关的各种公众事务所持有的各种情绪、意愿、态度和意见交错的总和。” |
| **舆论引导工作** | “在一定的社会空间内，围绕特定的舆情因变事项的发生、发展和变化，在民众中产生和存在的对执政者及其所持有的政治价值取向的社会政治态度。” |

舆论与舆情两个词一字之差，很容易混淆。实际上，两者的概念既有联系又有区别。

广义舆论指人们的认知、态度、情感和行为倾向的人群需要达到一定的数量，否则不能不能称为舆论。而舆情是指人们的认知、态度、情感和行为倾向的原始表露，可以是零散的、非体系化的，也不需要得到多数人的认同，是多种不同意见的简单集合。舆情聚集时有可能转化为舆论，通过舆论引导可以使舆情转化为良性舆论。

舆情和舆论都表现为公众的意见、情绪、和态度，舆情虽然是多种意见、情绪和态度的交织的总和，但是其中包含着小范围多数人的意见，就是舆论。当这类意见被社会大多数人认同时，就会转换为声势浩大的社会舆论。当正确和公正的意见被广大群众推崇和追随时，代表着历史发展的必然趋势，从而形成了主流民意。针对某一公共事务的分散和错综复杂的舆情，向一致有序的舆论与民意的转化是一种必然趋势。

#### 1.1.2 网络舆情

**网络舆情是指通过互联网表达和传播的舆情，反应了人们对某一公共事件所表达的认知、态度、情感和倾向性，具有虚拟化、快捷化、多元化、开放性、匿名性及互动性等特点。**

网络舆情通常有突发社会公共事件触发，突发社会公关事件包括自然灾害、重大事故、公共卫生、社会安全、媒体事件、地方经济、社会治理、官吏腐败等多个方面，推动网络舆情形成和传播的网络媒体主要是**网络新闻、微博、论坛帖子、博客博文**等，其中微博网站和网络论坛是网络舆情形成和传播的主要信息传播平台。

网络舆情作为舆情的网络表现形式，具有如下特点：

**（1）自发性**：在互联网上，人人都可以自由、自发的发表意见和表达态度，每个人是信息的发布者，又是信息的评论者，同时还是信息的传播者。任何一个公共事件发生后，网民都会自发地通过微博、论坛帖子、博客博文等网络媒体自由发表意见，表达自己的观点，情绪和态度。

**（2）多元性：**在虚拟的网络空间中，人们不再像现实生活中总是掩饰自己真实的想法和感受，而是更愿意表达自己真实的意见、情绪和态度，能够真实地反映人们不同的思想形态、文化观念、价值取向、生活准则以及道德规范等。

**（3）时效性：**由于网络媒体打破了时间和空间上的界限，网民可以随时随地在网上发表意见，网络舆情的形成非常迅速。当一个公共事件发生以后，网民可以立即在网上发表意见，网民意见由点到面，由散到聚，迅速汇聚成公共意见，形成强大的意见声势。另一方面，随着其他社会热点事件的发生，一个舆情事件很快被新涌现的公共事件所掩盖，持续事件通常为一周左右，呈现较强的时效性。

**（4）偏差性：**社情民意是基于网络舆情中最普遍、最尖锐的一部分，包含了广大网民不同的利用诉求，在一定程度上真实反映了他们的意见和呼声，但网络舆情不能完全等同于所有社会群体的立场。虽然网民在网络媒体上可以自由地发表自己的观点和情绪，但是网民的言论及其影响所承担的社会责任是脱节的，网民随意发表某些不负责任的言论甚至谣言，导致网络清晰表达失真，与真实的社会舆论存在偏差，极易引起网民情绪走向极端，导致“网络暴力”频频出现。

**（5）从众性：**在网络舆情形成过程中，一些网民并不直接发布信息，而是通过关注和转发他人的信息来表达自己的态度和倾向性，特别是意见领袖发布的信息。意见领袖是指在信息传播网络中经常发表意见，并具有相当影响力的“活跃分子”，在意见领袖的引导和影响下，网民通过转发微博、论坛跟帖等方式，推动网络舆情的形成和发展，局部意见可能演化成网络舆情。

#### 1.1.3 网络舆情演化

当一个突发社会公共事件发生后，随着事件的发展以及时间的推移，网络舆情经历一个个形成期、高涨期、波动期和消退期的演化过程，而事件的处置措施和应对能力对网络舆情的发展和走势起到至关重要的作用。

**（1）舆情形成期：**舆情形成期也称为发酵期，当一个突发社会公共事件发生后，经过传统媒体或网络媒体的报道，引起网民的关注和热议，通过微博、论坛帖子等方式表达自己的意见、情绪和态度，并在互联网上不断地传播，经过不断地发酵而成为被广泛关注的热点事件，形成了网络舆情。在发酵期，如果事件处置和应对得当，则有可能及时化解尚未形成的舆论热点。

**（2）舆情高涨期：**网络舆情形成后，随着网民的情绪等不断高涨，热点事件的关注度越来越高，不断吸引更多网民的关注，其关注度和影响力达到高峰。根据事件性质和发展，高涨期持续时间长短不一。在高涨期，如果事件处置和应对得当，则有利于消解高涨的网络舆论，释放公众的情绪。

**（3）舆情波动期：**网络舆情达到高峰后，随着事件的发展和处置，网络舆情进入波动期，呈现出波浪式发展态势，在事件处置不当或岀现新情况时，网络舆情可能出现又一波高峰。在波动期，凉糕的处置措施和应对能力能够使网络舆论得到有效的化解，进入舆情消退期。

**（4）舆情消退期：**舆情持续一段时间后，由于事件的处理、新事件的出现等原因，使人们的关注点发生了转移，转向新的目标。这时，原来关注热点逐渐变冷，网络舆情进入消退期，最终沉寂。在消退期，有效的应对措施能够起到形象修复和再造的作用。

### 1.2 舆情分析技术

随着网络舆论影响力的增强，网络舆情已经成为各级政府了解社情民意、改进工作作风、提高执政能力的重要窗口。近年来，国家大力推进网络舆情监控体系建设，各级宣传主管部门以及主流新闻媒体等大多设立了网络舆情监测机构，加强对网络舆情的检测和引导。

在网络舆情监测中，面对海量的网络信息，必须借助信息技术来实现网络舆情监测的自动化和智能化。于是，各种网络舆情监测系统便应运而生了。

#### 1.2.1 网络舆情监测系统

网络舆情监测系统的主要功能是实现网络信息的自动采集和网络舆情的在线监测与分析。尽管不同的网络舆情监测系统产品存在一定的差异，但在系统架构和核心技术上大同小异。通常，一个网络舆情监测系统可以按照层次化结构来构建，主要分为数据采集处理、数据预处理、舆情分析引擎和舆情分析服务等四个层次，如图1-1所示。



图1-1网络舆情监测系统

**（1）数据采集处理层：**主要提供网络数据采集和预处理功能，网络数据监测和采集的对象主要是主流的网络新闻、微博、网络论坛、网络博客等网站的文本信息，对于采集到的网络数据，首先需要进行初步的数据过滤、去重等预处理，经过数据格式转换及元数据标引后，存入数据库待进一步处理。

**（2）数据预处理层：**舆情数据预处理的目的是将违规识别系统和互联网采集系统收集的复杂不规则文本数据进行结构化清洗，结构化后的文本数据以词作为最小的语义单位，将海量文本中的舆情要素进行提取，为后续的舆情判别提供依据，并帮助分析引擎更高效准确地分析和检索舆情。预处理的手段主要有中文分词、命名实体识别和词性标注等。

**（3）数据分析引擎：**经过舆情预处理后的舆情数据，被按照语义切分成为词及词属性标注输入舆情分析引擎，多维度社评分析引擎主要是从采集内容中获取相关数据来完成社评事件主体分析、主题聚类、多维文本索引、低值元素剔除、大数据行为挖掘等分析；另外一方面主要是通过对采集的非结构化数据进行结构化获取相关结构数据来完成非结构化数据内容提取、摘要、取证、语义理解等工作。舆情分析引擎主要内容是进行主题、情感、文本分类等多层面的分析。

**（4）数据分析服务：**主要提供用户需求层的信息，具体包括热点搜索、关键词搜索、自动汇总统计、自动推荐关键词、采编辅助功能以及生成各类用户需要的统计报表等。系统对采集分析的数据可以通过负面舆情、分类舆情、最新舆情、专题跟踪、舆情简报、分类评、图表统计和短信通知等形式推送给用户。

#### 1.2.2 网络信息采集技术

网络数据采集是整个舆情预警系统的数据入口，通过对目标网络中舆情信息的集中采集和清洗处理，完成对实时和离线数据的外部数据收集工作，为下一步的舆情分析研判服务。网络舆情包含传统互联网网站媒体和移动新媒体，传统互联网网站媒体根据地域可分为港澳新闻网站和欧美新闻网站，网站内容涵盖纯新闻媒体和富媒体两类，对整体网站的所有文字、音频、图片、视频舆情内容进行统一的采集。

传统互联网网站的采集主要依赖于网络爬虫技术，根据入口位置对整体网站的内容进行规则性遍历，以发现有价值的内容信息并抓取下载，经过清洗后进行统一存储。移动新媒体由于上文提到的去域名化问题，需要通过模拟访问和流量截取技术实现数据的获取。

网络爬虫主要是通过Web页面之间的链接关系，利用网页自动抓取策略进行配置，保证其能高效运行，从Web上自动地获取页面信息，并且随着链接不断扩展，甚至向整个Web范围发展的过程。这一过程主要是由Web网页抓取器（WebCrawler/WebSpider，也称网络爬虫或网络蜘蛛）来完成的。它从一个初始的URL集合出发，将集合中的URL全部放入到一个有序的待抓取队列里。而抓取器从该队列中按顺序取出URL，通过相应的协议，获取URL所指向的页面，并保存到本地。然后从这些己获取的页面中提取出新的URL，并将它们继续放入到待抓取队列里。上述过程不断重复，直到抓取器根据自己的策略终止抓取。

利用专门针对境外互联网数据开发的网络爬虫程序，自动沿着互联网上任意网页中的URL链接爬到其它网页，重复此过程，并把爬过的所有网页收集到服务器中，同时提取网页所在URL、文本关键词、文件特征信息等，实现对境外网站及舆情信息的自动搜索和发现。

#### 1.2.3 网络信息预处理技术

**（1）中文分词：**

在中文自然语言中，词是最小的能够独立的包含和表现语义内容的结构成分。与英语等西方语言不同的是，汉语的语义组成结构使得词语之间无法通过空格等明显的标记区分，因此需要将文本、篇章、段落、句子等长结构中的字符串合理地切分成词语序列，在此基础上进行进一步的分析。因此，中文分词就是对中文句子的切分技术，其具体的过程是通过特定的语言使用规范，将多个邻近汉字之间的依赖关系映射到分词模型中，以求解出最符合使用规范的词语结构。

分词最简单的方法就是字典匹配，通过字典和文本遍历对应查找语义切分，但是无法对二义性的中文分词进行处理。而统计语言模型通过计算句子中符合语法规则分词的最大概率，能够基于语言的语法历史使用情况统计给出最符合常理的分词结构，能够有效地消除分词的二义性。因此，分词中最核心的内容是统计语言模型，模型的构建和训练决定了分词结果与语言使用实际的匹配程度。统计语言模型中，词的组成体现为词表，词表是分词系统中自带的通用词表，其中的词汇与领域无关。如果文本中的词语在此表中出现，则按照词表对词计算各自的出现概率，对于一个任意的一元字序列，按照出现总概率最高的句组合进行分词；如果未出现，则属于未登录词，需要依赖未登录词的分词机制，例如HMM的隐含模型

**（2）命名实体识别：**

命名实体作为一则信息中的基本信息元素，往往包含了其主要内容，是正确理解文本的基础。实体通常参照现实世界的一些对象来确定，是文本中最有价值的信息部分。命名实体是指现实世界中存在的具体或抽象的实体，例如人物、机构、地区、公司名等，通常能够用唯一专有的标识符标识。在不对文章全文进行遍历理解的情况下，命名实体识别是对文章内容摘要最重要的方式，能够帮助我们在海量舆情数据中提取出感兴趣的完整摘要。相较于传统摘要的关键词匹配法，命名实体识别能够提取出更为完整的对象词。

命名实体识别的过程通常与分词同时进行，在分词的词生成同时对该词的实体属性进行识别，基于分词的上下文相关文法结构和词实体边界信息对命名实体进行提取。命名实体识别的方法通常与分词方法相类似，包括基于词序列限长概率的N-gram语言模型方法、基于隐含状态转移结构的隐马尔科夫模型（HMM）方法以及通过概率熵凸优化的最大熵模型等。

**（3）词性标注：**

词性标注是语言结构理解中的又一大识别问题，其面向的对象是基于分词和命名实体识别的名词结果之上的所有字和词属性标注，目的是将句子中的名词、动词、助词、定冠词等词属性标注出来，同时在语义层面，将句子中的主语、谓语、宾语等组成结构进行标注，帮助识别引擎理解句子中的层级属性关系，使得舆情分析能够对句子和文章的主题和摘要进行提取。

词性标注的实现基础是文法语序之间的依赖关联，例如汉语中主谓宾及倒装的文法规则。中文中的动词、名词搭配通常也具有规律性，例如常见的动词+名词组合形式，这需要通过对所有组合路径上的组合输出概率进行统一计算，对于海量文本中庞大的HMM语义元连接，需要求解巨大网络中的全局最优化路径，即句子的最大概率值，就需要依赖高效的Vertibi算法进行最优路径搜索。Vertibi算法是求解HMM中路径最优化的一种动态时间规划算法，其核心是基于路径层级中的多阶段递进假设，通过选择此时刻包含最大局部概率的状态及其相应的局部最佳路径来确定全局最佳路径，即确定词性标注中所有词的最大词性组合概率。

#### 1.2.4 网络信息分析技术

**（1）主题分析：**

舆情分析是在海量的数据中寻找用户关注内容的一个过程，这个过程建立在一个使用前提上，即用户十分清楚自己想要获取的内容（即主题）。而由于舆情分析的对象存在模糊性和二义性，使得这一前提往往不成立，例如传统舆情分析中许多搜索引擎或文本检索工具的使用者，只是对自己希望得到的检索结果内容有一个粗浅的描述，并且大部分的使用者会通过不断的搜索-阅读-再搜索这一流程逐步优化自己的关键词选择。显然，这样迭代式的关键词选择并不是最高效的。另一方面，搜索引擎或文本检索工具返回的结果只能是针对关

键词的实时内容，而无法实现对于该关键词发展的有效跟踪及关联的有效挖掘的展现，这一部分的工作依旧需要使用人工完成。那么，是否有相应的方法能够既展现某个搜索内容（主题）的相关结果，同时又能给出这一主题根据时间的发展情况以及主题内容中的其他词汇和相关主题。机器学习研究者给出的答案是概率主题模型，即一系列为了自动标注大规模语料主题的算法。

**（2）情感分析：**

情感分析是一类高层次的语义分析技术，自动聚合网络新闻并自动进行褒贬倾向性的分析。对于每个话题，对每个发布人发表文章的观点、倾向性进行分析与统计。采用文本聚类和褒贬分析技术，对网站、微博、论坛帖子的网民评论进行聚类分析和褒贬分析，更准确了解社情民意。

情感分析的本质是对词、句、文本的一个三分类问题，分类的目标包括正面、负面和中性三种，分类是基于文本中VSM的特征空间对文本特征向量进行情感模型分类，文本特征空间维度通常有数百万维，维度较高，通常将局部敏感性哈希的方法引入情感分析中，对文本进行压缩表示；同时建立情感词库，包括褒义词表、贬义词表、否定词表等作为情感分析情感模型训练的基础，对文章进行基于词库规则的情感分析；同时，在细粒度情感分析中，发现隐式表达词，并集合显式特征词进行细粒度情感分析，进一步直接发现隐含在文章中的作者情感倾向。

**（3）文本分类**

面向舆情分析的语义理解和摘要提取需求，需要依赖于舆情文本分类模型对舆情数据进行文本内容分析和语义理解，并依赖于社评知识库中的已入库主体进行智能匹配，根据机器学习中相关概率预测模型的输出确定舆情的具体主体分类。

在得到了文本分类模型之后，系统所采集的舆情数据在抽取过相同的特征后，通过得到的文本分类模型能够得到所输出的部门主体标签，若基于概率模型输出，则其结果可以通过阈值限定条件产生最终的文本分类结果。根据项目定义的准确率约束，可以在整个流程的某些环节中添加人工干预的过程，进一步优化分类效果。同时这一文本分类的准确率会随着系统使用时间的增长和标注数据的增加而日臻完善，因此也意味着系统的文本分类功能需要一定的预热期才能稳定的使用。

## 第二章： 网络信息采集技术（陈林翠）

随着互联网快速的发展，互联网日益成为舆情产生、传播的重要媒介。越来越多的人，通过互联网表达政治诉求，从互联网这一自由开放的媒体中了解民意，开展互联网舆情信息的监控与分析，已经成为目前各级政府部门面临的重要课题。网络信息采集就是应对这一发展势头，为互联网舆情分析提供基础性的数据，帮助分析互联网舆情指数，捕捉敏感事件，分析民意走向。网络信息数据的采集应具备几个特点：自主的数据采集技术，分布式部署，统一管理，采集节点覆盖面广，异步抓取保持稳定。

### 2.1 数据选择

网络上的舆情数据，从数据类型来看可以分为文本、图片、音频和视频等，由于音频、视频需要的处理信息计算量较大、处理技术比较复杂且相比文本和图片承载的数据量较少，因此，将网络舆情数据的主要来源选择为文本与图片两类最为合适。考虑到数据的客观性和网络用户的覆盖度，舆情数据采集应选择用户量大的社交媒体作为舆情数据俩元，会使得所得的结果更具有通用性。

为保证舆情数据具有代表性,采集舆情数据的目标网站需要具有日常活跃用户多、涵盖话题领域广泛等特点。一般而言，微博、微信、知乎、贴吧、论坛等平台数据覆盖范围广、信息更新快、数据产量大是研究网络舆情的优质信息源。

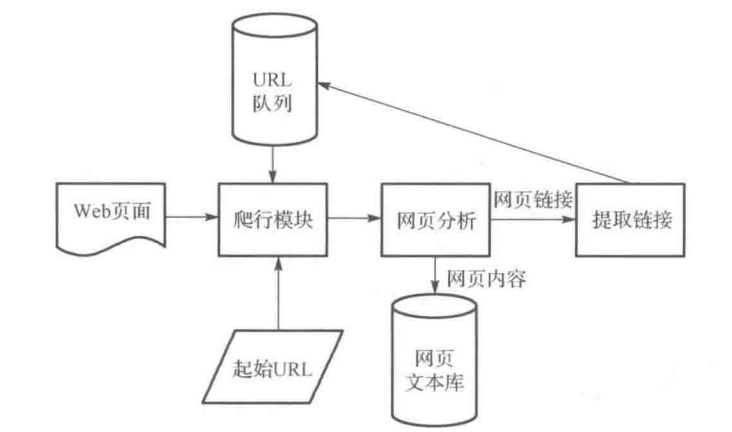
### 2.2 采集技术

大数据技术的发展，使人们处在一个信息爆炸的时代，如果通过人工处理数据信息，必将跟不上信息的更迭速度，只有借助软件、程序、代码等自动搜索才可以大幅提高对网络数据信息的获取与处理效率，网络爬虫技术由此诞生。它可以对网络信息进行纵、横双向的数据与信息抓取,处理之后保存在数据库中。

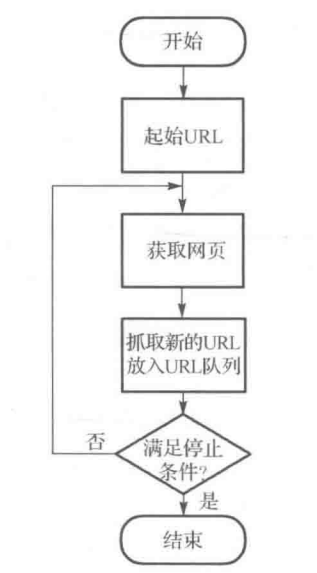
网络爬虫最早诞生于1993年，是舆情信息采集的重要方法，也是数据获取和分析最常用的方法。它的工作流程大致如下：选取初始页面的统一资源定位符(Uniform Resource Locator, URL)，将之放入待爬取队列，通过解析获得的IP地址就可以实现自动爬取网页的信息并进行保存，将爬取过的URL放入已爬取队列，从已爬取URL中获得新的链接，对新链接再进行解析处理，循环以上步骤直到爬取数据的数量达到规定的条件则停止。

#### 2.2.1 通用网络爬虫

通用网络爬虫是一种试图满足各类用户的需求，所搜集的网页内容广而泛。通用网络爬虫系统结构如图所示：



2-1 网络爬虫系统结构图

1. 爬行模块。
2. 网页分析模块。
3. 提取链接模块。该模块的主要作用是去除重复连接和循环链接。
4. 网页文本库。用来存储经过网页分析处理的网页，供后期处理使用。
5. URL队列。用来存放从网页中提取出的URL，当URL队列为空时，网络爬虫程序停止爬行。
6. 起始URL。提供种子URL，用来启动网络爬虫程序。
7. 通用网络爬虫工作原理

通用网络爬虫的目标就是尽可能多地采集网页，在这一过程中并不关注网

页采集的顺序和被采集网页的相关主题。这就需要消耗较多的系统资源和

网络带宽，并且这些资源的消耗并没有换来对所采集网页的较高利用率。

通用网络爬虫从网站某一个网页(通常是首页)开始，读取网页内容，并抽

取出网页中的其他超链接地址，然后通过这些链接地址寻找下一个网页，

这样一直循环下去，直到满足系统的停止条件，其工作流程如右图所示。

1. 通用网络爬虫搜索策略

通常，通用网络爬虫在抓取网页时采用两种搜索策略：深度优先搜索策略

和广度优先搜索策略。

深度优先搜索策略是指网络爬虫从起始网页开始，一个链接一个链接

地跟踪下去，处理完这条路径之后再转入下一个起始网页，继续跟踪链接，

直到遍历所有的网页及链接，搜索过程结束。这种搜索策略的优点是网络爬虫

能够遍历一个Web站点或深层嵌套的文档集合。缺点是当Web站点的网

页文件结构比较深时，有可能发生陷入进去而出不来的情况，即网络爬虫的

陷入问题。

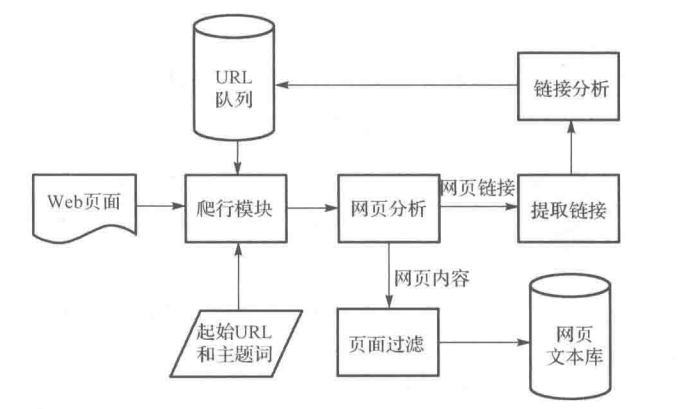
广度优先搜索策略是指网络爬虫从起始网页开始，首先搜索完-一个网页中所有的链接，然后再继续搜索下一层， 直到底层为止。这种搜索策略的优点是能够保证对浅层链接的优先处理，即使遇到一个深层分支时，也不会导致发生网络爬虫陷入深层文件中出不来的情况，并且能够在两个网页文件之间找到最短路径。广度优先搜索策略通常是网络爬虫的最佳策略，不仅容易实现，并且还能够实现并行处理，提高其抓取速度。然而，对于深层嵌套的网页文件集，广度优先搜索策略需要花费较长的时间才能搜索深层的网页文件。

在主流的网络爬虫系统中，一般采用广度优先搜索策略为主、深度优先搜索策略为

的搜索策略。对于某些不被引用或很少被引用的网页，广度优先搜索策略可能会遗漏这些孤立的网页，而深度优先搜索策略可以搜索到这些网页。

#### 2.2.2 主题网络爬虫

主题网络爬虫是一种只需满足某一类用户的需求，只搜集与主题内容相关的网页。主题网络爬虫系统结构如图所示：



主题网络爬虫主要由三个关键模块组成:

1. 网页分析模块。该模块采用了文本分类模型和算法对网页相关性进行分析。在爬行开始前，首先需要对样本网页信息进行学习，训练成一个主题相似度模型。当爬行到一个被认为与主题相关的网页后，将该网页提交给页面过滤模块进行主题相似度值计算，如果该网页相似度值大于或等于给定的阈值，则说明该网页与主题相关，则存入网页文本库；否则说明该网页与主题相关性小，丢弃该网页。
2. 链接分析模块。该模块是主题网络爬虫的核心模块，主要用于评估从主题相关网页中解析出来的URL链接与主题的相似度，并根据相似度值提供相应的搜索策略来指导网络爬虫的爬行过程。URL链接的主题相似度值越大，爬行的优先级就越高，在URL队列中排列越靠前。反之，如果某一URL链接的主题相似度值小于给定的阈值，则将该URL链接及其所隐含的子链接一并去除。
3. 爬行模块。该模块是任何网络爬虫程序都不可缺少的通用模块。爬行模块首先从待爬行URL队列中取出排在首位的URL,将该URL对应的网页抓取到本地，然后将该页面交给页面分析模块处理。在整个爬行过程中，爬行的次序和搜索策略都是由链接分析模块提供的。
4. 主题网络爬虫工作原理

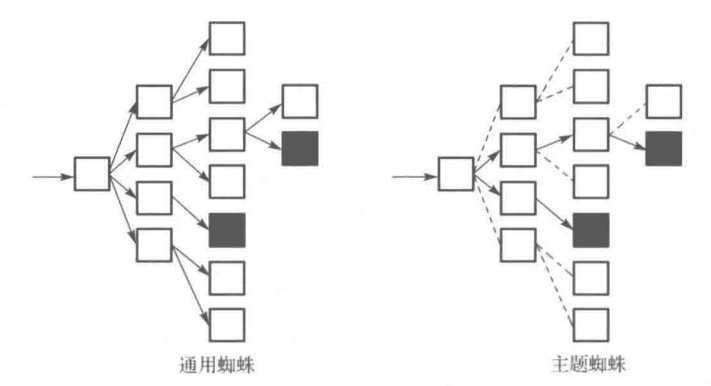
与通用网络爬虫不同的是，主题网络爬虫在爬行开始前需要就某个主题对样本网页信息进行学习和训练，建立相应主题相似度模型。在启动爬行后，首先从起始URL开始爬行，按照设定的搜索策略来搜索网页，对于所获取的网页，首先进行网页相关性分析，去除与主题不相关的网页;然后对所提取的URL进行链接相关性分析，设置该URL的优先级并存入URL队列。,当满足停止条件时，爬行过程结束。可见，存入网页文本库中的网页都是与主题相关的网页。

1. 主题网络爬虫搜索策略

由于主题信息一般只占整个Web空间很小的一部分，并且具有分散性，因此传统的深度优先搜索策略和广度优先搜索策略在Web信息搜集的效率上难以达到期望要求。由于主题网络爬虫的特点是采集的信息内容只限于特定的主题或专门的领域，因此在搜索信息过程中没有必要对整个Web空间进行遍历搜索，只需要选择与主题相关的网页进行访问即可。对主题搜索引擎而言，提高网络爬虫搜索效率的关键在于如何将不相关的网页快速地过滤掉，因为网页过滤的速度和准确性将会直接影响网络爬虫的性能。

主题网络爬虫在爬行过程中，如果对所发现的URL链接都不加选择的话，则大量的无关URL链接就会极大地浪费网络爬虫的处理时间。为了避免或减少这种现象的发生，主题网络爬虫必须对所发现的URL链接进行预测，只访问那些预测值符合给定阈值的URL链接。对于预测值符合要求的网页，系统还要对所抓取的网页进行主题相关性分析，以保证最终得到的网页是与主题相关的。

相比于通用网络爬虫，主题网络爬虫能够有效地发现主题相关的网页，并且能通过网页内容和链接结构来指导其资源发现过程。下图反映了通用网络爬虫和主题网络爬虫的爬行过程中的差异。



其中，浅色框代表主题无关的网页，黑色框代表主题相关的网页，虚线代表链接，箭头代表访问次序。

1. 通用网络爬虫以广度优先搜索策略，沿着每个链接进行爬行。假设从起始网页到目标网页需要爬行i步，那么在爬行到目标网页前必须先将i-l步内的网页爬行完。
2. 主题网络爬虫首先确定最有可能与主题相关的链接，忽略主题无关的网页。假设从起始网页到目标网页需要爬行i步，那么在爬行到目标网页前仅爬行i-l步内的网页的一个子集，在理想情况下，只要爬行i个链接就可以达到目标网页，这大大节省了爬行时间，提高了爬行效率。

如果同一个主题相关的网页在Web网站上是平均分布的，则主题网络爬虫和通用网络爬虫在效率上没有明显的差别。经过大量的研究表明，Web网站上的网页分布具有主题聚合性。利用这一\_特性，主题网络爬虫就可以优先对主题相关网页中的链接进行爬行。

与通用网络爬虫相比，主题网络爬虫更加专业化和可定制化。通用网络爬虫的目标是尽可能多地采集网页，在这一过程中它并不关注网页采集的顺序和被采集网页的主题，而主题网络爬虫能够定向性地采集与主题相关的网页，忽略无关的网页，并且还可以根据主题相似度值进行优先采集。因此，主题网络爬虫具有如下特点:

1. 抓取目标描述和定义。在主题网络爬虫采集网页之前，首先要确定一个抓取的主题。
2. URL搜索策略。它决定着待爬行URL的访问次序。通用网络爬虫以搜集海量数据为目标，在爬行时只需按照广度优先搜索策略或者深度优先搜索策略来抓取网页，尽量提高其爬行的范围。而主题网络爬虫因其自身的特点，在爬行过程中需要考虑优先访问与主题相关的网页，即考虑如何管理待爬行的URL队列，从而使得每次爬行都从相似度值最大的URL开始。
3. 网页相关性分析与过滤。对于已经下载的网页，需要通过适当的主题相似度模型和算法来判别该网页是否与主题相关。因此，主题相似度模型和算法的优劣将对网页和链接的主题相似度计算性能产生很大的影响。

#### 2.2.3 增量式网络爬虫

增量式网络爬虫是指：只抓取新产生或已发生变化的网页，对于无更新的页面不采取任何操作，增量式爬虫可以确保采集的页面均是新页面。其优点是减小了数据的下载量，有效避免了时间和空间上的浪费。其缺点是算法的复杂度和实现难度大大增加了。在实际应用中，通常都是将这几种爬虫方式结合起来使用。

#### 2.2.4 反爬虫机制及应对措施

爬虫作为一种计算机技术就决定了它的中立性，爬虫本身在法律上并不被禁止，但是利用爬虫技术获取数据这一行为是具有违法甚至是犯罪的风险的。具体问题具体分析，互联网上的数据本身就是公开信息，浏览器解析并显示了页面内容，爬虫同理，只不过爬虫会批量下载，故具备合法性。不合法的情况则是配合爬虫，利用黑客技术攻击网站后台，窃取后台数据（比如用户数据等）。

随着搜索引擎的发展，网络爬虫已经成了很普及网络技术，但大部分爬虫不严格遵守robots协议，因此网站为了保护自己内容不让别人抓取而诞生的反爬虫需求。常见反爬虫机制主要包括以下几个方面：

1. 身份验证

爬虫抓取部分网站时，经常会遇到需要登陆验证（输入账号、密码）之后才能获取数据的情况。

实现爬虫的自动登录主要分为两种方式：数据包分析与模拟点击方式。

数据包分析登录利用浏览器的网络监听功能获取登录后的Headers与Cookies等信息，然后传递用户名、密码等参数完成登录。该方法执行速度快，但抓包分析比较费时，有时不确定Header中的必须参数。

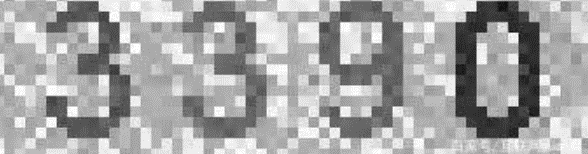
模拟点击方式登录，调用一个无头浏览器来访问登录网页。Python中可以使用Selenium库来调用浏览器，并代码方式操作页面。该种登录方式能几乎可以对付任何网页，但会导致代码效率低。

1. 验证码识别

验证码（CAPTCHA）是“Completely Automated Public Turing test to tell Computers and Humans Apart”（全自动区分计算机和人类的图灵测试）的缩写，是一种区分用户是计算机还是人的公共全自动程序。

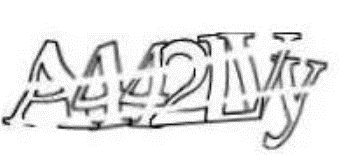
验证码常出现在登录页面或者相同IP重复或者频繁访问网页的场景中。常见验证码类型有：图片验证码，短信验证码、计算题图片验证码、滑动验证码、图案验证码、标记倒立文字验证码。

1. 图片验证码
2. 对于比较简单的图片验证码，字符规整，清晰，如下图所示。



应对措施：直接调用用OCR识别技术，如利用python第三方库：tesserocr来识别，对于图像中字符较为模糊的情况，在进行字符识别前可加入图像预处理工作，去除干扰，强化字符特征。

1. 对于容易迷惑人的图片验证码，噪声干扰大，字符存在重叠，如下图所示。



应对措施：收集数据，利用图像识别算法+深度学习训练识别模型。

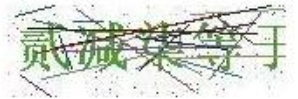
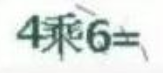
1. 短信验证

该类验证码需要输入相应的手机号，然后向手机发送验证码。

应对措施：尝试用Webbrowser技术，模拟用户打开短信的行为，最终获取短信验证码。

1. 计算题图片验证码

该类验证码通常由数字或者汉子组成的数学表达式，如下图所示：



应对措施：优先建议方案为：收集数据，利用图像识别算法+深度学习训练识别模型。对于建立图像库，然后模板匹配方法，效率低下，不建议使用。

1. 滑动验证码

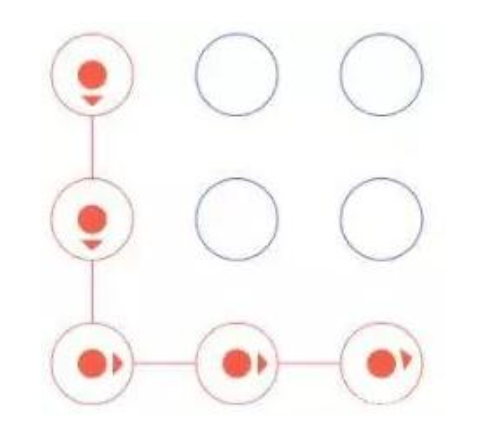
需要将滑块拖动到图像缺失的位置，如下图所示：



应对措施：该类型验证码的关键难点在于找出找出背景图中缺口的准确位置，先依据图像处理技术将缺块的位置，然后模拟用户拖动滑块先快后慢的运动轨迹特征，构造运动轨迹函数，最后使用无头浏览器将滑块按照运动轨迹函数拖动到检测出来的位置。

1. 图案验证码

此类验证码需要按照给出的顺序依次拖动鼠标轨迹，完成验证，如图所示。



应对措施：利用机器学习所有的拖动顺序或者穷举所有可能的图案，然后识别，最后使用无头浏览器完成类似人的操作，完成验证。

1. 标记倒立文字验证码

在验证码中随机生成一定量的中文汉字，部分汉字为倒立显示，对于汉字而言，有中华五千年庞大的文字库，加上文字的不同字体、文字的扭曲和噪点，识别难度更大了。如下图所示：



应当措施：利用图像识别技术识别出图中倒立文字的坐标序号，按照指定格式发去请求，完成验证。但识别倒立文字难度巨大。

综上，验证码的种类繁多，构建自己的识别模型所耗费的精力巨大，最便捷的应当措施：打码平台，神箭手库等。

1. 数据动态加载

部分网站是经过JavaScript渲染的，或者是通过Ajax请求得到,此类型网站在源代码中无法获取在浏览器网页中所看到的信息。

应对措施：

1. 重构Ajax请求

如果能够找到Ajax请求，也能分析出具体的参数和响应的具体含义，则能采用常规方法，即直接利用requests或者urllib2模拟Ajax请求，对响应的json数据格式进行分析得到需要的数据。能够直接模拟Ajax请求获取数据是优先选择的，但有些网站把Ajax请求的所有参数全部加密了。我们根本没办法构造自己所需要的数据的请求。遇到这样的网站，则需另找方法。

1. 借用无头浏览器动态渲染

调用浏览器内核，并利用无头浏览器执行JavaScript来模拟人为操作以及触发页面中的JavaScript脚本。从填写表单到点击按钮再到滚动页面，全部都可以模拟，不考虑具体的请求和响应过程，只是完完整整的把人浏览页面获取数据的过程模拟一遍。使用无头浏览器几乎能绕过大多数的反爬虫。

1. User Agent限制

部分爬虫框架的User Agent是特殊的，与正常浏览器的不一样，可通过识别User Agent，进而识别爬虫，最终直接封掉爬虫请求。

应对措施：构建User Agent池，包含不同系统、不同浏览器，不同版本的请求头列表。当爬虫每次发送的时候随机从User Agent池中随机选择一个请求头作为本次请求的请求头，防止暴露爬虫身份。特别地，对于需要制定浏览器的网站，则需要制定的请求头方可成功请求页面。

1. IP访问频率限制

网站通过检测用户行为，当同一IP短时间内多次访问，或者同一账户短时间内多次进行相同操作，网站则在一段时间内禁止该IP的所有请求。

应对措施：类似User Agent，构建IP池，这个需要大量的IP资源，可以通过抓取网上免费公开的IP建成自有的IP代理池或者向IP代理机构购买IP代理服务。然后再每次请求时使用不同的代理IP。

1. 请求时间窗口过滤统计

爬虫爬取网页的频率都是比较固定的，而用户去访问网页，中间的间隔时间比较无规则。给每个IP地址建立一个时间窗口，记录IP地址最近N次访问时间，每记录一次就滑动一次窗口，比较最近访问时间和当前时间，如果间隔时间很长判断不是爬虫，清除时间窗口，如果间隔不长，就回溯计算指定时间段的访问频率，如果访问频率超过阀值，则判断为爬虫。

应对措施：对于每个请求设置下载延迟，延迟时间越长越安全，但数据获取速度也就越慢。

1. 并发量限制

通常爬虫的并发是很高的，统计并发最高的IP，当统一IP的并发量高于设定阈值，则判断为爬虫。

应对措施：合理控制爬虫的并发量。

1. 蜜罐资源

爬虫解析离不开正则匹配，适当在页面添加一些正常浏览器浏览访问不到的资源，一旦有IP访问，则可检测为爬虫。

应对措施：合理筛选链接，控制请求广度。

#### 2.2.5 常用框架

|  |  |  |
| --- | --- | --- |
| 名称 | 开发语言 | 特征 |
| Apache Nutch | Java | 高度可扩展和可伸缩性的分布式爬虫框架；  支持Hadoop； |
| Arachnid | Java | 微型爬虫框架，含有一个小型HTML解析器 |
| crawlzilla | Java | 安装简易，拥有中文分词功能；  crawlzilla 除了爬取基本的 html 外，还能分析网页上的文件，如（ doc、pdf、ppt、rss ）等多种文件格式 |
| WebMagic | Java | 框架精简、模块化；  无须配置、便于二次开发的爬虫框架，它提供简单灵活的API； |
| Scrapy | Python | 基于Twisted的异步处理框架，文档齐全；  纯python实现的爬虫框架，模块化，开发便捷 |
| PyRailgun | Python | 简单易用，国内开发；  支持抓取javascript渲染的页面；  简洁、轻量、高效的网页抓取框架 |
| Hispider | C++ | 目前只是能提取URL， URL排重， 异步DNS解析， 队列化任务， 支持N机分布式下载， 支持网站定向下载 |
| Larbin | C/C++ | 高性能的爬虫软件，只负责抓取不负责解析 |
| Methabot | C/C++ | 过速度优化、可抓取WEB、FTP及本地文件系统 |
| NWebCrawler | C# | 可配置：线程数，等待时间，连接超时，允许MIME类型优先级，下载文件夹  统计信息：URL数量，总下载文件，总下载字节数，CPU利用率和可用内存 |

## 第三章： 网络信息数据预处理技术（付强）

### 3.1 网络信息数据预处理概述

由于词时最小的能够独立运用的语言单位，而中文文本不像英文文本，词与词之间没有任何空格之类的显式标志指示词的边界；不同词的时态、语态、词性没有明显的区分；不同的词是否是命名实体也没有明显的区分。因此，自动分词、词性标注和命名实体识别问题就成了计算机处理中文文本时面临的首要基础性工作，对于舆情分析系统而言更是一个不可或缺的重要环节。多年来，国内外的众多学者在这一领域做了大量的研究工作，发表了大量的学术专著和论文，并取得了一定的成果，为我们的舆情分析系统奠定了基础。

本章首先介绍中文自动分词所面临的一些主要问题和近年来取得的最新进展，然后介绍词性标注和命名实体识别的相关问题和研究进展。

### 3.2 汉语分词方法

#### 3.2.1 汉语分词方法概述

自汉语自动分词问题被提出以来，人们提出了很多分词方法。如正向最大匹配法(forward maximum matching method , FMM)、逆向最大匹配法(backward maximum matching method , BMM)、双向扫描法、逐词遍历法等，这些方法大多是在上世纪80年代或更早的时候提出来的。由于这些分词方法大多都是基于词表进行的，因此一般被统称为基于词表的分词方法。随着统计方法的发展，人们又提出了许多基于统计模型的中文分词方法，如基于隐马尔可夫模型的分词方法和n元语法等。除此之外，还有人提出了规则与统计方法相结合的中文分词技术。以上研究工作使得中文分词的问题得到了更加深入的研究。

#### 3.2.2 N-最短路径方法

考虑到中文自动分词中存在切分歧义消除和未登录词识别两个主要问题，有人将分词过程分为两个阶段。首先，采用切分算法对句子词语进行初步切分，得到一个相对好的粗分结果。然后，将粗分结果进行歧义排除和未登录词识别。粗分的结果准确性直接影响到了后续的歧义排除和未登录词识别的效果，并最终影响到整个分词系统的正确率。为此，有人提出了一种旨在提高粗分结果的词语粗分模型——N-最短路径方法的中文词语粗分模型。该方法的基本思想是：首先根据词典，找出字符串中所有可能的词，构造词语切分有向无环图。每个词对应图中的一条有向边，并赋给相应的权值。然后针对该切分图，在起点到终点的所有路径中，求出长度值按严格升序排列依次为第一、第二、…第N的路径集合作为相应的粗分结果集。如果两条或两条以上路径长度相等，那么，他们的长度并列第i，都要列入粗分结果集，而且不影响其他路径的排列序号，最后的粗分结果集合大小大于或等于N。

假设待分字符串S=c1c2…cn ，其中ci为单个汉字，n为字符串的长度。建立一个结点数为n+1的切分你有向无环图G，各结点编号依次为V0，V1，…Vn。

通过以下两步建立G所有可能的词边：

1. 相邻结点Vk-1，Vk之间建立有向边，边的长度值为<Vk-1，Vk>，边的长度值为Lk，边对应的词默认为ck。
2. 如果是词表中的词，则结点Vi-1，Vj之间建立有向边<Vi-1，Vj>，边的长度值为，边对应的词为。

这样，待分字符串S中包含的所有词与切分有向无环图G中的边一一对应，如图3.1所示。

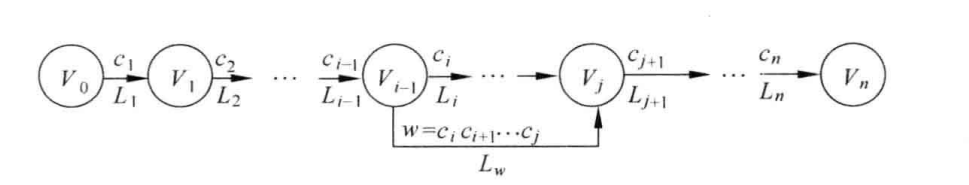


图3.1 切分有向无环图

考虑到切分有向无环图G中每条边权重的影响，该方法可分为统计粗分模型和非统计粗分模型两种。所谓非统计粗分模型即假定切分有向无环图G中所有词的权重都是对等的，即每个词对应的边长均设为1.

假设NSP为结点V0到Vn的前N个最短路径的集合，RS为最终的N-最短路径粗分结果集。那么，N-最短路径方法将词语粗分问题转化为如何让求解有向无环图G的集合NSP。

求解有向无环图G的集合NSP可以采用贪心算法，在N-最短路径模型中作者采用的是基于求解单源最短路径问题的Dijkstra贪心算法的一种简单扩展。改进之处在于：每个结点处记录N个最短路径值，并记录相应路径是上当前结点的前驱。如果同一长度对应多条路径，则必须记录这些路径上当前结点的前驱，最后通过回溯即可求出NSP。

Dijkstra算法的时间复杂度为O(n2)，它求的是图中所有点到单源点的最短路径，而应用于切分有向图时，有两个本质区别：首先有向边的源点编号均小于终点编号，即所有边的方向一致；其次，算法最终求解的是有向图首位结点之间的N-最短路径。因此，在该算法中，运行时间与n(字符串长度)、N(最短路径数)以及某个字作为词末端字的平均次数k(等于总词数除以所有词末端字的总数，对应的是切分图中结点入度的平均值)成正比。所以，整个算法的时间复杂度是O(n×N×k)。

考虑到在非统计模型构建粗切分有向无环图的过程中，给每个词对应边的长度赋值为1.随着字符串长度n和最短路径数N的增大，长度相同的路径数急剧增加，同时粗切分结果数量必然上升。考虑到大量的切分结果对后期处理以及整个分词系统性能的提高非常不利。因此，该方法的作者又提出一种基于统计信息的粗分模型。

假定一个词串W经过信道传送，由于噪声干扰而丢失了词界的切分标志，到输出端便成了汉字串C。N-最短路径方法词语粗分模型可以相应地改进为：求N个候选切分W，使概率P(W|C)为前N个最大值：

(3.1)

其中，P(C)是汉字串的概率，它是一个常数，不必考虑。从词串恢复到汉字串的概率P(C|W)=1。

因此，粗分的目标就是确定P(W)最大的N种切分结果。为了简化计算，假设是字串的一种切分结果。是一个词，表示词出现的概率，在大规模训练语料的基础上通过最大似然估计方法求得。切分W的概率为：

(3.2)

为了方便处理，令，可以看作是词在切分有向无环图中对应的边长。于是，求式(3.2)的最大值问题转化为求的最小值问题。

针对修改了边长后的切分有向无环图G\*，直接使用非统计粗分模型的求解算法，就可以获得问题的最终解。

#### 3.2.3 基于词的n元语法模型的分词方法

基于词的n元语法模型是一个典型的生成式模型，早期很多统计分词方法均以它为基本模型，然后配合其他未登录词识别模块进行扩展。其基本思想是：首先根据词典对句子进行简单匹配，找出所有可能的词典词，然后将它们和所有单个字作为结点，构造的n元切分词图，图中的结点表示可能的词候选，边表示路径，边上的n元概率表示代价，最后利用相关搜索算法从图中找到代价最小的路径作为最后的分词结果。

由于未登录词的识别是中文分词过程中的关键问题之一，因此很多人认为未登录词的识别与歧义切分应该是一体化处理的过程，而不是相互分离的。基于这种思想，有人提出了一种基于加权的有限状态转换机(weighted finite-state transducer)与未登录词识别的一体化切分的实现方法，以及基于改进的信源信道模型的分词方法。

该分词方法将中文定义成以下4类：

1. 待切分文本中能与分词词表中任意一个词相匹配的字段为一个词。
2. 文本中任意一个经词法派生出来的词或短语为一个词，入重叠形式、前缀派生、后缀派生、中缀派生、动词加时态助词、动词加趋向动词、动词的分离形式等等。
3. 文本中被明确定义的任意一个实体名词是一个词。
4. 文本中任意一个专有名词是一个词。

假设随机变量S为一个汉字序列，W是S上所有可能切分出来的词序列，分词过程应该是求解使条件概率P(W|S)最大的切分出来的词W\*，即：

(3.3)

根据贝叶斯公式，式(3.3)可以改写为：

(3.4)

由于分母为归一化因子，因此：

(3.5)

为了把4类词纳入同一个语言模型框架，可以将专有名词的人名(PN)、地名(LN)、机构名(ON)各作为一类，实体名词中的日期(dat)、时间(tim)、百分数(per)、货币(mon)等作为一类处理，简称实体名，对词法派生词(MW)和词表词(LW)则每个词单独作为一类。这样，按表3.1可以把一个可能的词序列W转换成一个可能的词类序列。则式3.5可以改写成：

(3.6)

其中，P(C)称为语言模型，P(S|C)称为生成模型。

表3.1 生成模型P(S|C)

|  |  |  |
| --- | --- | --- |
| 词类 | 生成模型P(S|C) | 语言知识 |
| 词表词(LW) | 若S是词表词，P(S|LW)=1；否则为0 | 分词词表 |
| 词法派生词(MW) | 若S是派生词，P(S|MW)=1；否则为0 | 派生词词表 |
| 人名(PN) | 基于字的二元模型 | 姓氏表，中文人名模板 |
| 地名(LN) | 基于字的二元模型 | 地名表、地名关键词表、地名简称表 |
| 机构名(ON) | 基于词类的二元模型 | 机关名关键词表，机构名简称表 |
| 实体名(FT) | 若S可用实体名词规则集G识别，P(S|G)=1；否则为0 | 实体名词规则集 |

如果P(C)采用三元语法，可以表示为：

(3.7)

三元模型的参数可以通过最大似然估计在一个带有词类标记的训练语料上计算，并采用回退平滑算法解决数据稀疏问题。生成模型在满足独立性假设的条件下，可以近似为：

(3.8)

#### 3.2.4 基于感知机算法的汉语分词方法

假设是输入句子，是切分结果，其中X是训练语料集合，Y是X中句子标注结果集合。用GEN(x)表示输入句子x的切分候选集，用表示训练实例(x,y)对应的特征向量，α表示参数向量，其中是模型的特征空间。则给定一个输入句子x，其最优切分结果满足如下条件：

(3.9)

平均感知机用来训练参数向量α。首先将α中所有参数初始化为0，然后在训练解码过程中不断更新。对每一个训练样本，用当前的模型参数进行解码得到切分结果，如果切分结果与标注结果不一致，则更新模型参数。每一次都保留参数的加和，直到进行多轮迭代以后，取参数的平均值以避免模型过拟合。

平均感知机训练算法如下：

input：训练样本

初始化：α=0，=0

算法过程：

for t=1…T do # T轮迭代

for i =1…N do # N个训练样本

计算 # 用当前参数解码

if != then # 更新权重

# 权重加和

# 平均化模型参数

output：

基于感知机算法的汉语自动分词方法的基本思路是，对于任意给定的一个输入句子，解码器每次读一个字，生成所有的候选词。生成候选词的方式有两种：(1)作为上一个候选词的末尾，与上一个候选词组合成一个新的候选词；(2)作为下一个候选词的开始。这种方式可以保证在解码过程中穷尽所有的分词候选。在解码的过程中，解码器维持两个列表：源列表和目标列表。开始时，两个列表都为空。解码器每读入一个字，就与源列表中的每个候选泽合生成两个新的候选，并将新的候选词放入目标列表。当源列表中的候选都处理完成之后，将目标列表中的所有候选复制到源列表中，并清空目标列表。然后，读入下一个字，如此循环往复直到句子结束。最后，从源列表中可以获取最终的切分结果。

解码算法描述如下：

Input：任意一个汉语句子；

初始化：置源列表src和目标列表tgt均为空，即src=[[]]，tgt=[]；

说明：变量item用于记录已生成的候选词序列，item1用于记录当前字作为新词开始时新生成的候选序列(separated)，item2用于记录当前文字与前面的词合并时新生成的词序列(appended)。

算法过程：

Step-1 从句子中任取一个字，执行如下循环，直到句子结束：

1. 对于src中的每个候选词序列item：
2. 将当前字作为新词的开始，形成新的候选词item1；
3. 将当前字附加到item中最后的候选词上，形成新的候选词item2；
4. 将item1和item2添加到tgt列表；
5. 将目标列表tgt中中候选词复制到源列表src中，并清空目标列表tgt。

Step-2 输出源列表src中的最佳候选词序列。

该算法类似于全切分方法，理论上会生成所有的个切分结果(l为句长)。为了提升切分速度，可以对目标列表tgt中候选词的数目进行限制，每次只保留B个得分最高的候选。

#### 3.2.5 由字构词的汉语分词方法

由字构词的汉语分词方法的主要思想是，将分词过程看作字的分类问题。在以往的分词方法中，无论是基于规则的方法还是基于统计的方法，一般都依赖于一个事先编制的词表，自动分词过程就是通过查词表做出词语切分的决策。与此相反，由字构词的分词方法认为每个字在构造一个特定的词语时都占据着一个确定的构词位置。

这里所说的“字”不仅限于汉字，也可以指标点符号、外文字母、注音符号和阿拉伯数字等任何可能出现在汉语文本中的文字符号，所有这些字符都是由字构词的基本单元。

分词结果表示成字标注形式之后，分词问题就变成了序列标注问题。对于一个含有n个字的汉语句子，可以用下面的公式来描述分词原理：

(3.10)

其中，表示第k个字的词为，即。

通常情况下，使用基于字的判别模型时需要在当前字的上下文中开一个个字的窗口，在这个窗口里抽取分词相关的特征。常用的特征模板有：

前面三类特征模板是窗口内的字及其组合特征，是指的字符类别。有了这些特征以后，我们就可以利用常用的判别式模型，如最大熵、条件随机场、支持向量机和感知机等进行参数训练，然后利用解码算法找到最优的切分结果。

由字构词的分词技术的重要优势在于，它能够平衡地看待词表词和未登录词的识别问题，文本中的词表词和未登录词都是用统一的字标注过程来实现的，分词过程成为字重组的简单过程。在学习架构上，既可以不必专门强调词表词信息，也不用专门设计特定的未登录词识别模块，因此大大简化了分词系统的设计。

#### 3.2.6 分词方法的比较

自开展汉语自动分词方法研究以来，人们提出的各类方法不下几十种甚至上百种，不同方法的性能各不相同，尤其在不同领域、不同主题和不同类型的汉语文本上，性能表现出明显的差异。为了对比现有的常见分词方法的性能，有人利用语料对不同分词方法性能做了对比试验。其中，语料的统计情况如表3.2所示。

表3.2 三种分词语料

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 语料名称 | 编码 | 训练语料规模  (字数/字节) | 测试语料规模  (字数/字节) | OOV比例 |
| AS | Big5 | 5.45M/141K | 122K/19K | 0.046 |
| CITYU | Big5 | 1.46M/69K | 41K/9K | 0.074 |
| MSR | GB | 2.37M/88K | 107K/13K | 0.026 |

模型训练和测式过程中，只使用了提供的数据集进行了测式和训练。评价指标包括准确率(P)、召回率(R)、F-测度、未登录词的召回率(ROOV)和词典词的召回率(RIV)。各指标的计算公式如下：

(3.11)

(3.12)

(3.13)

在实际应用中，F值计算时一般取β=1，所以F值又被称为F1值。当β=1时：

(3.14)

最后，选取了七个常见分词系统，对比如表3.3所示。

表3.3 七个分词系统的性能对比

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Corpus** | **Participants** | **R** | **P** | **F** | **ROOV** | **RIV** |
| **AS** | Asahara05 | 0.952 | **0.951** | 0.952 | 0.696 | 0.963 |
| Tseng05 | 0.950 | 0.943 | 0.947 | **0.718** | 0.960 |
| Zhang06 | 0.956 | 0.947 | 0.951 | 0.649 | 0.969 |
| Z＆C7 | N/A | N/A | 0.946 | N/A | N/A |
| Jiang08 | 0.958 | 0.949 | 0.953 | 0.692 | 0.970 |
| Sun10 | N/A | N/A | 0.952 | N/A | N/A |
| Wang10 | **0.963** | 0.949 | **0.956** | 0.652 | **0.977** |
| **CITYU** | Asahara05 | 0.937 | 0.946 | 0.941 | 0.736 | 0.953 |
| Tseng05 | 0.941 | 0.946 | 0.943 | 0.698 | 0.961 |
| Zhang06 | 0.952 | 0.949 | 0.951 | **0.741** | 0.969 |
| Z＆C7 | N/A | N/A | 0.951 | N/A | N/A |
| Jiang08 | 0.946 | 0.950 | 0.948 | 0.695 | 0.966 |
| Sun10 | N/A | N/A | **0.956** | N/A | N/A |
| Wang10 | **0.959** | **0.952** | **0.956** | 0.700 | **0.980** |
| **MSR** | Asahara05 | 0.952 | 0.964 | 0.958 | **0.718** | 0.958 |
| Tseng05 | 0.962 | 0.966 | 0.964 | 0.717 | 0.968 |
| Zhang06 | 0.972 | 0.969 | 0.971 | 0.712 | 0.976 |
| Z＆C7 | N/A | N/A | **0.972** | N/A | N/A |
| Jiang08 | 0.964 | 0.967 | 0.966 | 0.686 | 0.972 |
| Sun10 | N/A | N/A | 0.969 | N/A | N/A |
| Wang10 | **0.975** | **0.970** | **0.972** | 0.632 | **0.984** |

表中N/A表示没有相关数据，从表中可以看出这些系统性能都达到了相当高的水平。其中Wang10系统在多个测试中取得了相比其他系统更优的性能。

### 3.3 中文词性标注

#### 3.3.1 词性标注概述

在对语句进行高层分析之前准确把握语句的每个词汇的词性对分析结果有很大帮助，比如语句的谓语往往是一个动词或相当动词的短语。 因此词性标注对高层分析发挥着基础性作用，其主要人物是结合语句的上下文环境，对语句中的每一个词汇标注合适的词性，如：名词、动词、形容词、数词、量词等。 词性标注是与语言的语法密切相联的。 每一门语言都有一些语法规则，但几乎每条规则都有例外，比如英语的动词第三人称用法一般是加ｓ，但是有好多动词有各种例外的形式。 词性也有很多存在例外，并不是一个词汇对应一种固定的词性。 这也是词性标注的最大困难。词性标注的另一个困难就是新词的标注，所谓新词也有人称未登录词汇，是指训练模型的时候没有遇到的词汇。 通常的方法就是改进以前的基于统计的方法，从语料库训练得到全面知识，改进模型，但这种方法有其先天的局限性，只能解决近程依赖的词汇的词性标注，对于长程依赖的词汇还是存在困难。

常见的词性标注方法可以分为以下两类：

1. 基于规则的词性标注方法。

词性标注传统方法是采用基于规则的方法，规则集的完备程度会影响到标注方法的性能。 规则越完备效果越好，但是规则也就越来越复杂。 规则通常是利用现有语言的语法知识，人工总结出来的，准确程度依赖于语言专家的语言知识。 标注之前先建立规则库，再对语句进行词汇切分，对切分的词汇序列进行标注。 从某一词汇所有可能的词性中，依据上下文信息筛选出唯一合适的词性作为最终的标注结果。

规则的方法确实有很多优点，比如标注效率高，适用范围也比较广，但是基于规则的方法最大的缺点就是规则集的构造很难自动从特定语言中提取。 人工构造规则集是一项费时费力的任务，而且要依赖特定语言专家的领域知识。 如果规则描述的粒度过细，自适应性自然大大降低。 而且仅仅依据规则判断词性又可能往往会抹杀上下文的语境而出现歧义。

1. 基于统计的词性标注方法。

统计词性标注方法主要是将词性标注问题抽象为一个数学统计模型，通过某个已经标注词性的语料库来训练这个模型从而对未知的词汇序列进行标注。 最典型的就是基于隐马尔可夫模型的标记方法，与基于规则的方法相比，基于统计的方法更加客观、适应性更强。

目前较为流行的方法是词性标注方法，训练词性标注模型的语料库规模越大，覆盖面越广，词性标注的性能就越高。 因此在利用词性标注之前先要准备人工标注的大规模语料库。 但当训练语料达到一定规模后，就不能再通过扩大语料规模来提高词性标注性能。 仅依靠扩大语料库的规模是不够的，最大限度地提高词性标注的性能需要设计更为严格精细的统计模型，参考更多的特征信息来进行词性标注。

#### 3.3.2 基于规则的词性标注方法

基于规则的词性标注方法时人们提出较早的一种词性标注方法，其基本思想时按兼类词搭配关系和上下文语境建造词类消歧规则。早期的词类标注规则一般由人工构造。

随着标注语料库规模逐步增大，可利用资源越来越多，以人工提取规则的方式显然时不现实的。于是，人们提出基于机器学习的规则自动提取方法。其中，最具代表性的是E.Brill提出的基于转换错误驱动的学习方法，该方法可以用图3.2描述。

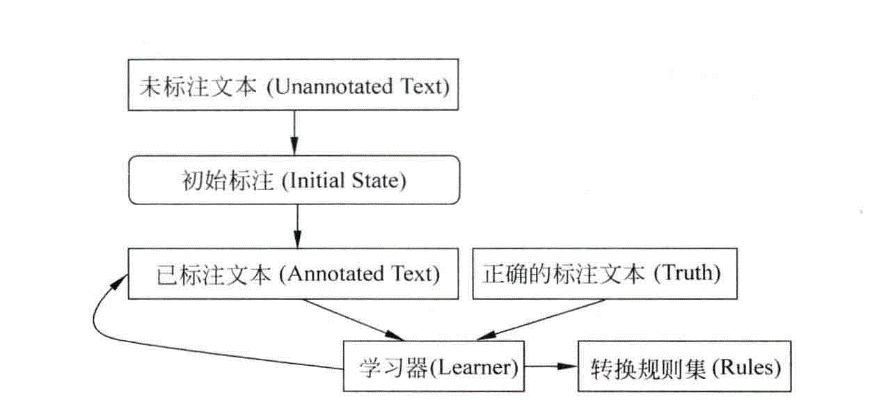


图3.2 基于转换规则的错误驱动的机器学习方法

从图3.2可以看出，基于转换规则的错误驱动的学习方法的基本思想是，首先运用初始状态标注器标识未标注的文本，由此产生已标注的文本。文本一旦被标注后，将其与正确的标注文本进行比较。由于初始标注器标注的文本一般会含有错误，学习器通过将这标注文本与正确的标注文本相比较，可以学习到一些转换规则，从而形成一个排序的转换规则集，使其能够修正已标注的文本，使标注结果更接近参考答案。这样，在所有学习到的可能的转换规则中，搜索那些使已标注文本中的错误数减少最多的规则加入到规则集，并将该规则用于调整已标注的文本，然后对已标注的语料重新打分。不断重复该过程，直到没有新的转换规则能够使已标注的语料错误数减少。最终的转换规则集就是学习到转换规则结果。

基于转换规则的错误驱动的学习方法用于词性标注，解决了传统的规则方法中由手工构造规则的不足，而且与统计方法相比，系统标注速度有很大的优势。但是，该方法存在一个很大的问题就是时间过长。为此，有人提出了相应的改进方法，在改进算法的每次迭代过程中，只调整收到影响的小部分转换规则，而不需要搜索所有的转换规则。通过研究发现，每当一条获取的规则对训练语料实施标注后，语料中只有少数的词性会被改变，而只有在词性发生改变的地方，才影响与该位置相关的规则的得分，可见在改进前的算法中，大量时间花在数量占绝大多数的分值不需要修改的变换上。如果在用新的规则标注的过程中，当上下文环境满足，规则成立，而导致某一位置的词性发生改变时，准确判断哪些规则受到影响，然后相应修改这些规则的得分，而不必理会那些未受影响的其他所有规则，这样大大地节省了处理时间，提高了学习算法的速度。

#### 3.3.3 基于统计模型的词性标注方法

基于统计模型的词性标注方法最早在上世纪80年代就有人提出，其基本思想是利用n元语法与一阶马尔可夫转移矩阵进行词性标注。该方法通过对n元语法概率的统计优化，实现了133个词类标记的合理标注。

实现基于HMM的词性标注方法时，模型的参数估计是其中的关键问题。假设输入符号表由单词构成，如果某个对应的“词汇-词性标记”对没有被包含在词典中，则令该词的生成概率为0。否则，该词的生成概率为其可能被标记的所有词性的倒数，即：

(3.15)

其中，为词l由词性标记j生成的概率，为词出现的次数，分母为在词典中所有词汇范围的求和，而：

(3.16)

式中，为词允许标记的词性个数。

一旦初始化完成以后，HMM参数就可以利用前向后向算法进行训练。

此外，在HMM模型实现中，还有另外一个问题需要注意，就是模型参数对训练语料的适应性。也就是说，由于不同领域语料的概率有所差异，HMM的参数也应随着语料的变化而变化。这个问题涉及两个方面，一个是对原有的训练语料增加新的语料以后，模型的参数需要重新调整；另一个是在经典HMM理论框架下，利用标注过的语料对模型初始化以后，已标注的语料就难以再发挥作用。而折两方面的问题Baum-Welch方法都不能解决。针对此问题，有人提出对原训练方法做出如下修改：

给定两个训练语料C1和C2，N为状态个数，即不同词性的个数。模型的参数估计如下：

(3.17)

(3.18)

(3.19)

给定训练语料C1，首先用Baum-Welch方法从该语料中训练得到的模型，但是在保存该模型时，并不直接保存的值，而是在中保存所有的C1的期望变量。由这些变量，可以很容易的利用下面的公式计算出的值：

(3.20)

(3.21)

(3.22)

在此基础上，假如有新的语料C2引入，我们希望建立一个既能反映C1又能反映C2的模型。则使用作为初始模型，利用前向后向变量，通过Baum-Welch方法可以得到C2的期望值变量。然后将中保存的C1的期望值变量与C2的期望值变量相加，得到了反映C1和C2的期望值变量的值，将这些值保存下来就得到了新的模型。既反映了C1的信息，又反映了C2的信息。而且，由于保存了期望值变量，不需要再为后续的训练而保存用过的训练语料了。因此，模型和训练模型的语料能够分离开来，具有良好的灵活性。

### 3.4 命名实体识别

#### 3.4.1 命名实体识别方法概述与研究现状

自从在 MUC-6（Message Understanding Conference）会议上作为一个子任务被提出之后，国内外学者在该领域做了很多有意义的工作。命名实体识别任务主要是从自然语言文本中抽取出特定的名词，如人名、地名、机构名、时间、货币等，为机器翻译、问答系统、知识图谱构建等高级自然语言处理任务提供信息。作为信息抽取的基础任务，命名实体识别的结果影响着舆情分析系统的下游任务，如文本分类、摘要提取、情感分析等。目前存在的方法主要包括基于规则的方法、基于机器学习的方法、基于深度学习的方法等。

1. 基于规则的命名实体识别

学者们最早利用基于规则的方法，通过相关领域专家和语言学家制定有效的规则模板，如专家系统等，进行命名实体识别。但是，这种方法严重依赖专家制定的规则方法，且规则方法之间不能冲突，维护代价高，运转周期时间长，只能适用于单一的领域，一旦跨领域识别，就需要重新制定规则方法，领域迁移性较差。

1. 基于机器学习的命名实体识别

随着机器学习方法的流行，学者们采用基于机器学习的方法对实体进行识别，并取得了较好的效果，基于规则的方法逐渐被淘汰。学者们将命名实体识别工作看成序列标注任务，给输入句子中的每一个词语标注相应的标签，根据标签确定实体的边界位置和类型。例如标签策略 SBIEO，其中S表示该字为单个实体，B表示该字在实体的开始位置，I表示该字在实体的中间位置，E表示该字在实体的结束位置，O表示该字不属于实体的一部分。学者们利用机器学习方法在命名实体识别领域做了很多有效的工作，提出了隐马尔可夫模型（Hidden Markov Model, HMM）、最大熵马尔可夫模型（Maximum Entropy Markov Model, MEMM）、条件随机场（Conditional Random Field, CRF）等模型，这些机器学习方法需要研究者手工提取有效的语法特征，设定模型的模板进行实体识别。

Zhou和Su等人提出使用四种不同的特征，如词语的内部特征、触发器内部的语义特征、地名实体内部特征、上下文特征等提高 HMM 在命名实体识别的性能。张等人提出将 MEMM 作为模型的基本框架，融合局部与全局多种特征，并引入启发式知识进行有效的命名实体识别。Lafferty 等人2001年首先提出将CRF 模型应用在序列标注任务上。McCallum和Li等人提出特征感应的方法和Viterbi方法，用于寻找最优命名实体识别序列。邱海泉等人利用知识库，设计合适的特征模板，将CRF应用于微博文本数据的命名实体识别。

另外，由于不同领域存在差异性，文本格式等存在区别，命名实体识别模型的领域迁移性较差，如机构名、地名等实体识别比较困难。陈钰枫等人提出一种双语实体交互式对齐模型，利用双语对齐信息帮助实体识别并对结果进行修正，很大的提高了中文命名实体识别的自适应性。在军事领域，冯蕴天等人提出基于CRF模型的半监督学习方法，针对军事领域文本的语言特点，建立高效的特征集合，然后采用CRF模型进行序列标注，结合字典比对和规则匹配的方法对识别结果进行校正，取得了很好的效果。在地理领域，何炎祥等人为了识别文本中海量的地理命名实体，根据地理命名实体的特点，设计适合的特征模板，基于CRF模型并加入规则的方法进行命名实体识别，有效地提高了实体识别的召回率。Chen 等人提出利用条件随机场和最大熵模型对中文命名实体进行识别，识别效果F1值达到了86.2%。但是基于机器学习的命名实体识别研究比较依赖于特征的选择，特征的选择很大程度上影响模型的性能好坏。

1. 基于深度学习的命名实体识别

近年来，随着深度学习方法的成熟，基于深度学习的命名实体识别研究在通用领域达到了很好的效果，并成为了命名实体识别的主流方法。神经网络通过搭建多层网络结构，抽取文本中局部和全局特征，然后进行序列标注工作。目前比较常见的神经网络主要包括循环神经网络、卷积神经网络等。同基于机器学习方法的命名实体识别研究一样，基于神经网络的方法将中文命名实体识别看作序列标注问题。

针对通用领域的命名实体识别，柏兵等人提出基于CRF和BiLSTM方法的命名实体识别方法，BiLSTM 神经网络能够很好地保留上下文信息的特性,该模型在 1998 年人民日报语料库上取得了很好的效果。针对于文本自身，在不使用外部资源、不进行多任务联合训练的情况下，李明扬等人提出在BiLSTM-CRF模型的基础上，加入自注意力机制进一步获取文本的特征，进而提高中文命名实体识别的性能，该模型在新浪微博命名实体识别语料库上的实验结果达到了 58.76%。在深度学习网络模型的基础上，秦娅等人提出一种基于特征模板的CNN-BiLSTM-CRF 的网络安全实体识别方法，利用人工特征模板提取局部上下文特征，进一步利用神经网络模型自动提取字符特征和文本全局特征。实验结果表明，在大规模网络安全数据集上，提出的网络安全实体识别方法，相关评价指标由于其他算法，F1值达到86%。zhang等人在2018年提出基于字级别序列标注模型为基础，融入大规模词表匹配信息，提出 Lattice LSTM 模型，在中文命名实体识别领域取得了较好的效果。在中文命名实体识别研究领域，由于神经网络模型能够自动提取文本的特征，如卷积神经网络可以有效地提取局部特征，LSTM 网络可以记忆长文本信息，因此，目前大部分的中文命名实体识别研究都采用基于神经网络的深度学习方法。

但是，关于中文嵌套命名实体识别领域的研究，目前比较匮乏，由于识别技术限制以及任务要求等原因，大部分集中于普通的命名实体识别。中文领域目前研究进展比较慢。尹迪等人将嵌套实体识别看成是一种联合切分和标注任务，基于联合模型，并采用柱搜索算法进行解码，实现嵌套命名实体的识别。Chen等人首先预测实体的左右边界，根据预测的边界生成候选实体，然后采用分类器判断实体真假以及实体类型等，该方法在 ACE 2005 中文数据集上取得了较好的效果。关于英文的嵌套实体识别，目前研究较多。Wang等人将嵌套实体识别任务看成二叉树的构建过程，基于状态转移模型，不断进行动作生成和状态转移。

#### 3.4.2 基于CRF的命名实体识别方法

基于CRF的命名实体识别与基于字的汉语分词方法的原理一样，就是把命名实体识别过程看作一个序列标注问题。其基本思路是：将给定的文本首先进行分词处理，然后对人名、简单地名和简单的组织机构名进行识别，最后识别复合地名和复合组织机构名。所谓的简单地名是指地名中不嵌套包含其他地名。同样，简单的组织机构名中也不嵌套包括其他组织机构名。

基于CRF的命名实体识别方法属于有监督的学习方法，需要利用大量已标注的语料对CRF模型参数进行训练。在训练阶段首先需要将分词语料的标记符号转化成用于命名实体识别序列标注的标记。然后确定特征模板。特征模板一般采用当前位置的前后位置上的字机器标记表示，即以当前位置的前后n个位置范围内的字串及其标记作为观察窗口。

由于不同命名实体一般出现在不同的上下文语境中，因此对于不同的命名实体识别一般采用不同的特征模板。同时，考虑到出现在人名左右两边的字串对于确定人名的边界有一定的帮助作用。因此，某些总结出来的“指界词”也可以作为特征。

特征函数确定以后，剩下的工作就是训练CRF模型参数λ。针对基于CRF的中文命名实体识别，大量的研究人员做了大量的研究工作。实验表明，在人名、地名、组织机构名三类实体中，组织机构名的识别性能最低。一般情况下，英语和汉语人名识别的F1值可以达到90%左右，而组织机构名识别的F1值一般在85%左右。

#### 3.4.3 基于多特征的命名实体识别方法

在命名实体识别中，无论采用哪一种方法，都是试图冲突分发现和利用实体所在的上下文特征和实体的内部特征，只不过特征的颗粒度大小不同。考虑到大颗粒度特征和小颗粒度特征有互相补充的作用，应该兼顾使用的问题。有学者提出了基于多特征相融合的汉语命名识别方法，该方法是在分词和词性标注的基础上进一步进行命名实体的识别，有词形上下文模型、词性上下文模型、词形实体模型和词性实体模型4个子模型组成的。其中，词形上下文模型估计在给定词形上下文语境中产生实体的概率；词性上下文模型估计在给定上下文语境中产生实体的概率；词形实体模型估计在给定实体类型的情况下词形串作为实体的概率；词性实体模型估计在给定实体类型的情况下词性串作为实体的概率。

1. 模型描述

在基于多特征模型的命名实体识别系统中，词形包括以下几种情况：字典中任何一个字或词单独构成一类；人名(Per)、人名简称(Aper)、地名(Loc)、地名简称(Aloc)、机构名(Org)、时间词(Tim)和数量词(Num)各定义为一类。也就是说，词形语言模型中共定义了|V|+7个词形，其中|V|表示词典的规模。有词形构成的序列称为词形序列WC。

根据北京大学计算语言学研究所开发的汉语文本词性标注集，加上人名简称词性和地名简称词性，共47个词性标记。由词性标记构成的序列称为词性序列TC。

命名实体识别可以看作是一个序列化数据的标注问题。输入时带有词性标记的词序列，如式(3.23)所示。

(3.23)

其中，n表示句子中被分词程序切分出来的词的个数，是标注的词的词性。

在分词和词性标注的基础哈桑进行命名实体识别的过程就是对部分词语进行拆分、组合和重新分类的过程，最后输出一个最优的“词形/词性”序列，用式(3.24)表示：

(3.24)

由式(3.24)计算最优“词形/词性”序列的方法有三种：

1. 词形模型

词形特征模型根据词形序列W产生候选命名实体，用Viterbi确定最优词形序列。目前大部分系统都是从这和层面来设计命名实体识别算法的。

1. 词性模型

词性特征模型根据词性序列T产生候选命名实体，用Viterbi确定最优词性序列。目前只有较少的系统在命名实体识别过程中引入词性的知识。

1. 混合模型

词形和词性混合模型式根据词形序列W和词性序列T产生候选命名实体，一体化确定最优序列。

1. 词形和词形上下文模型

上下文模型估计在给定的上下文语境中产生实体的词形和词性概率。词形上下文模型和词性上下文模型均可采用三元语法模型近似描述：

(3.25)

(3.26)

其中，m、和的含义与式(3.24)一致。当i取1时，取一元语法概率和；当i取2时，取二元语法概率和。

1. 实体模型

考虑到每一类命名实体都具有不同的内部特征，因此不能用一个统一的模型刻画人名、地名和机构名等实体模型，应分别建立相应的实体模型。

1. 人名实体模型

基于字的人名实体词形模型用下式描述：

(3.25)

其中，表示组成人名实体的单字。BNe，MN和ENe分别表示实体的首字、中间字和尾字，在具体计算人名时，分别将其替换成具体的人名。

由于人名的词性实体模型的训练语料很难得到，因此为了简化起见，使用词形实体模型替代词性实体模型，但乘以一个加权因子，如式(3.26)所示：

(3.26)

其中，γ为小于1的加权因子，通常取经验值为0.5。

1. 地名和机构名实体模型

对于地名和机构名，其实体模型要复杂得多，这是因为地名除了普通词汇意外，还常嵌套人名和其他地名。基于词得嵌套地名和机构名词形实体模型可以用式(3.27)描述：

(3.27)

其中，和分别是实体被分词切分出的首词和尾词；和分别是的首词和尾词它们都是按照分词模块的词形定义切分出来的最基本的词形。是由原分词序列组合的可能的词，假设组合后含有k个词或子实体名，即长度为k，子实体可能是人名或地名。若子实体是人名时，将被符号PER替换，如果子实体是地名时，将用标记Loc替换。BNe为实体被正确切分时的首词，记作Boo；MN…MN为实体被正确切分时中间部分的k-2个词，记作Moo；ENe为实体被正确切分时的末尾词，记作Eoo。

1. 单字地名实体模型

单字地名词形实体模型和词性实体模型均可采用最大似然估计方法计算，分别运用以下公式估计：

(3.28)

(3.29)

其中，和分别是语料中作为单字地名和其词性出现得次数。为训练语料中单字地名出现的次数，即。

1. 简称机构名实体模型

简称机构名是对机构名全称的缩略叫法，机构名简称的出现形式大致可分为连续简写、不连续简写和混合简写三种方式。包括机构名关键词的机构名简称的识别同机构名全称的识别过程是一样的，但对于那些省略了机构名关键词的简称机构名的识别则是非常困难的问题。

简称机构名在文本中的出现基本上有以下三种形式：

1. 某些简称可以作为常用词收录进词典中。
2. 有些简称机构名无法被收录进词典，但该简称的全称形式在文本中出现过。
3. 文本中直接出现省略了机构名关键词的简称机构名。
4. 专家知识

在基于统计模型的命名实体识别中，最大的问题是数据稀疏严重，搜索空间太大，从而影响系统的性能和效率。因此，引入专家知识来限制候选实体的产生，从而达到提高系统性能和效率的目的。专家知识主要包括以下几类：

1. 人名识别的专家知识
2. 地名识别的专家知识
3. 机构名识别的专家知识
4. 模型训练

根据前面的介绍，基于多特征的中文命名实体识别模型由4个参数组成，这些参数使用最大似然估计从不同训练语料中学习。尽管使用了大规模的训练语料，数据稀疏问题还是非常严重。因此，可以采用Back-off数据平滑方法，并引入逃逸概率计算权值，如式(3.30)所示：

(3.30)

其中，，，n是待求的n元阶数，是各阶逃逸概率，可使用下面的经验公式计算：

(3.31)

其中，分子表示对于词序列后面跟随的不同的的个数，分母表示词序列出现的次数。

1. 测试结果

系统性能表现主要通过准确率P、召回率R和F-测度值3个指标来衡量，计算公式如下：

(3.32)

(3.33)

(3.34)

为了充分测试基于多特征模型的命名实体识别系统的性能，研究者对系统进行了性能测试，测试结果如表(3.4)所示。

表3.4 基于多特征的混合模型测试性能

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 性能 | 正确识别数 | 总计识别数 | P | R | F |
| 人名 | 19051 | 20220 | 94.06 | 95.21 | 94.63 |
| 地名 | 20861 | 22159 | 93.98 | 93.48 | 93.73 |
| 机构名 | 9390 | 11094 | 84.69 | 86.86 | 85.76 |

混合模型的人名、地名、机构名识别性能比单独使用词形模型时的性能分别提高了5.4%，1.4%，2.2%，比单独使用词性模型时分别提高了约0.4%，2.7%，11.1%。也就是说，混合模型的命名实体识别模型优于单一特征的命名识别模型。

## 第四章： 网络信息数据分析技术（孔凯传）

### 4.1 话题（主题）检测与跟踪技术

话题检测与跟踪（Topic Detection and Tracking， TDT）是自然语言处理和信息检索领域提出的研究课题，最初是由美国国防高级研究计划署（DAPPA）发起研究的，旨在没有人工干预的情况下自动检索、判断和识别新闻数据流中的话题。

TDT是从一片文章的主题作为切入点，通过对文章的主题的发现和跟踪，把各种分散的信息进行有效的汇集，并且组织成线索提供给用户进行查阅，理清一个主题事件的来龙去脉，把握整个事件的整体和细节。例如，在网络舆情监测中，通过TDT技术对各种信息源的监测和分析，从中识别出出各种突发事件以及事件的演化过程。TDT技术还可以应用于证券市场分析等领域，用途比较广泛。

（本节内容主要涉及TDT基本概念、相关技术、话题检测算法、话题跟踪算法、热点话题监测等内容。）

#### 4.1.1 基本概念

**1、相关术语：**

（1）话题（Topic）:在TDT 中，话题通常被定义为与显示世界中某个事件相关的新闻和故事的集合。在最初的研究中，话题和事件含义基本相同。一个话题是指某些原因、调节引起，发送在特定事件和地点，有一定的参与者或者涉及者，并可能伴随某一些必然结果的一个事件。例如： “2001年9月11日美国世贸大厦遭受到恐怖袭击”。后来的话题概念要相对宽泛一些，它包括一个核心事件或活动以及所有与之直接相关的事件和活动。如果一篇报道讨论了某个话题的核心事件直接相关的事件或者活动，那么也认为该报道与此话题相关。例如，搜救9。11事件的幸存者、安葬死难者等都被看作与“9。11 恐怖袭击”这个话题相关。

（2）事件（Event）：事件通常是指在特定事件和地点发生的事情。可以简单地认为话题就是若干对某个事件相关报道的集合。例如，“2001年9月11日美国世贸大厦遭受到恐怖袭击”是一个事件而不是话题，而“美国世贸大厦遭受到恐怖袭击”是话题而不是事件，事件是话题的实例，与一定的活动相关。

（3）故事( Story):故事是对某个事件的相关报道。在TDT中，通常是指一个与话题紧密相关的、包含两个或多个独立陈述某个事件的子句的新闻段落。

（4）主题( Subject):主题的含义更广一些，主题可以涵盖多个类似的具体事件或者根本不涉及任何具体事件，而话题则与某个具体事件相关。例如，“恐怖袭击”是一个主题，而“美国世贸大厦遭受到恐怖袭击”则是一个话题。又如，“自然灾害”是一个主题，而属于该主题类别的文本未必有与之直接相关的事件发生，例如讲述自然灾害预防的文章等。

**2、TDT任务**

TDT是一种交叉性技术，以自然语言处理技术作为主要技术支撑。因此TDT测评会议对TDT任务进行了细化，根据不同的应用需求，将TDT任务分成5个子任务。

(1)**新闻报道切分:**该任务的目标是将一个语言信息流分割为不同的独立新闻报道。由于文本信息流本身就是以单个报道形式出现的，不存在切分问题。因此该任务只适用于对来自广播、电视等媒体的音频数据处理。一段新闻节目通常包含了很多条报道，这些节目本身很少在不同的新闻报道间设置明显的分隔标记，例如商业广告很可能出现在某一报道的中间。要切分的语料或数据可以是音频记录本身，也可以是从音频记录得到的文字记录

(2)**新事件检测:**该任务的目标是从新闻报道信息流中检测出对一个新话题的首次报道，该任务也被看作对一个话题检测系统的透明测试。在新事件检测中常用的典型方法是采用向量或概率分布形式的特征项集合来代表每篇报道，每遇到新来的报道，就将其特征项集合与过去所有报道的特征项集合进行相似度比较，以此来判断该报道是否描述了一个新的话题

(3)**报道关系检测:**该任务的目标是判断两个随机选择的新闻报道是否讨论同一个话题。与其他任务相比，该任务没有直接的应用目标，因此对该任务的研究并没有受到重视在报道关系检测中，需要使用某种相似度度量方法来计算报道关系的相似度，经过测试表明，余弦系数方法是比较有效的。

(4)**话题检测:**该任务的目标是检测出未知的话题以及相关报道。话题检测关注的是将某个话题的所有报道归入一个话题类，因此它是一个无监督学习的聚类问题。通常的聚类可以看作基于全局信息的聚类，即在整个数据集合上进行聚类，而话题检测中的聚类是增量式的，即做出最终的决策前，不能或只能看到前面有限数量的报道。在话题检测研究中，大多数采用传统的自然语言处理方法，如中心向量法、K最近邻居( K-Nearest Neighbor，KNN)、K-均值(K- MEANS)、单边聚类算法等。

(5)**话题跟踪:**该任务的目标是给出某一话题的一组样本报道，通过训练得到话题模型，然后在后续报道中找出所有讨论目标话题的报道。该任务的实质是通过有监督的学习过程，利用非常少的正例数据和大量的反例数据来获得一个分类器，用于区分新报道与话题的相关性，因此可以吧话题跟踪看作一种特殊的二值分类问题。在话题跟踪的研究中，常见的分类放到有KNN、决策树、Rocchio算法、SVM、语言模型、概率模型、基于查询的方法等，其中比较有效的方法是KNN 算法以及多种算法的组合。

#### 4.1.2 相关技术流程

在话题检测与跟踪中，需要解决一下几个问题：

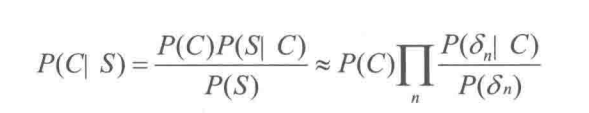
1. 报道和话题的表示模型；
2. 特征权重计算；
3. 话题和报道间相似度计算；
4. 文本分类与聚类的决策选择。

**1、表示模型：**

为了判断一个报道是否与某个话题相关，需要使用适当的模型来表示报道和话题，一边对两者的相关度进行计算和比较，常用的表示模型有向量空间模型和语言模型。

下面主要解释语言模型：

语言模型是一种概率模型，假设报道中出现的词各不相关，则某个报道S与话题C相关的概率为：

 （4-1）

式中，可以表示成一个两台的混合模型，如图4-1所示

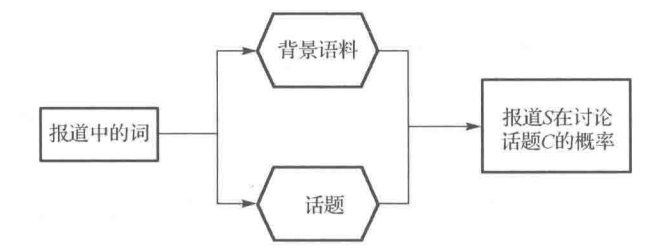


图4-1 混合模型示意图

在两态混合模型中，一个状态是词在该话题中所有报道的分布，另一个状态是词在整个语料中的分布，这样就构成了词的生成模型。

**2、相似度计算：**

文档是由相互独立特征词构成，令表示由m个特征次构成的n个文档的文档集合，其中是文档向量，表示特征词i出现在文档中的权重，设，则与的相似度用表示如下:

权重的选择可以使用多种方法，如果选择二值权重计算方法，特征词i出现在文档j中，；特征词i不出现在文档j中，。

|  |  |
| --- | --- |
| 相似度计算方法： | 公式： |
| 1．内积 |  |
| 2．Dice系数 |  |
| 3．Jaccrad系数 |  |
| 4．余弦系数（余弦相似度） |  |
| 5．欧几里得度量 |  |

**3、特征项选取：**

1．使用映射或者变换的方法特征把原始特征变换为较少的特征；

2．从原始特征中挑选出一些最具有代表性的的特征

3．根据专家的知识挑选最具有影响的特征

4．使用统计方法找出最具分类信息的特征，这种方法是一种比较精确的方法，人为因素的干扰较少，尤其适合于文本自动分类挖掘。

**4、文本聚类、分类方法：**

**话题检测属于文本聚类。**

* **文本聚类过程大致分为3个步骤：**

（1）文本表示：把文档表示成聚类算法能够处理的形式，最常用的文本表示方法是向量空间模型。

（2）聚类算法：使用无监督学习算法对文档集合进行划分，文本聚类算法有很多种，但是没有一个通用的算法能够解决所以的聚类问题。在文本挖掘中，最常用的相似度计算方法是余弦系数法。

（3）效果评估：通常使用准确率、召回率、漏报率和误报率等评价指标。

文本聚类算法有很多种，常用的有基于层次方法、划分方法、基于密度的方法、基于网格的方法、基于模型的方法等。

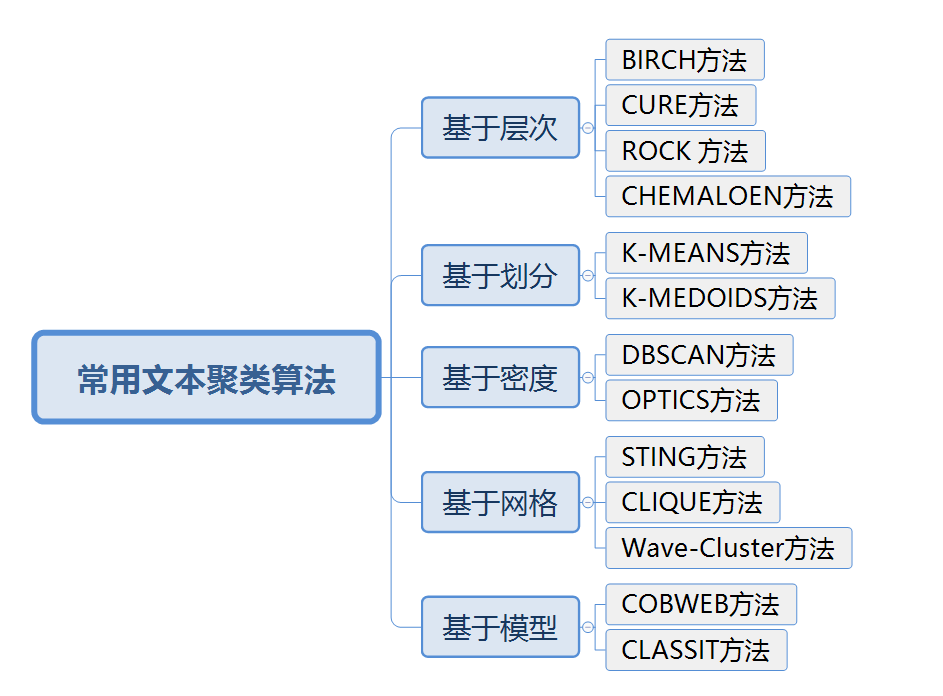


图4-x 文本聚类常用算法

**话题跟踪属于文本分类。**

* **文本分类过程大致分为3个步骤：**

（1）文本表示：把文档表示成分类器能够处理的形式，最常用的文本表示方法是向量空间模型。

（2）分类器构建：选择和设计分类器构建方法，应当根据所要解决问题的特点来选择一个分类器。在选定方法之后，在训练集上为每个类别构建分类器，然后把分类器应用于测试集上，得到分类结果。

（3）效果评估：当分类算法在测试集（而不是训练集）上完成分类过程后，需要对算法的分类效果进行评估，常用的有准确率、召回率、误报率和漏报率

文本分类中使用的学习算法有很多种，比如Rocchio算法、KNN、决策树、简单贝叶斯算法、神经网络、最大熵、SVM等

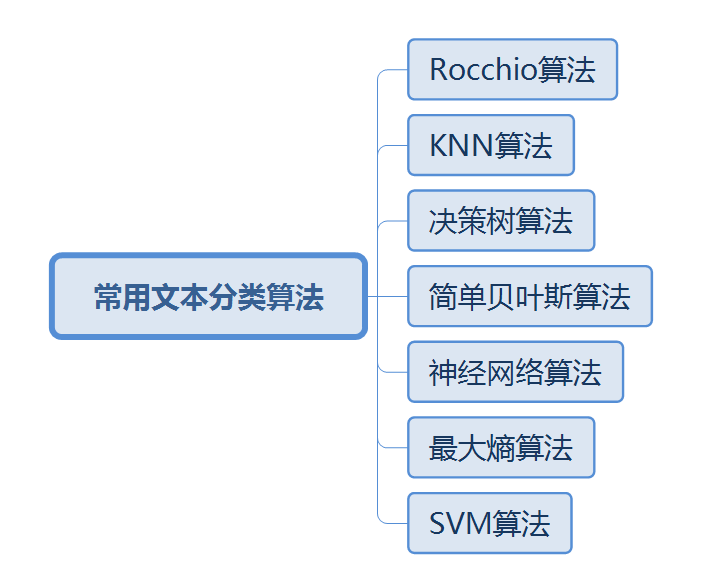


图4-x 文本分类常用算法

#### 4.1.3 话题检测算法

话题检测的目的是按照话题对文档进行聚类，从一组新闻报道中发现新话题，它是是按在实现没有分类体系和训练语料的情况下对报道进行聚类分析，因此它是一个无间断学习的聚类问题。主要介绍：K-means、FCM算法。

* **K-means算法：**

算法的基本思想，把n个对象划分成k个类，其中聚类数量k是输入参数，通过不断地迭代来聚类，当算法收敛于一个结束调节时，终止迭代过程，输出一个聚类结果。

k-means算法

|  |
| --- |
| **算法4.1 K-means算法** |
| 算法：k-means。算法基于簇中对象的平均值。  输入：簇的数量k和包含n个对象的数据库。  输出：平方误差总和最小条件下的n个簇。  方法：   1. 任意选择k个对象作为初始的簇中心； 2. 将所有对象划分到相应的簇中； 3. 计算每个簇中对象的平均值，将所有对象重新赋给最类似的簇； 4. 重复步骤（3），知道不在发生变化，收敛为止； 5. 算法结束 |

* **FCM算法（Fuzzy C-Means）：**

FCM算法有两个非常重要的参数：聚类数量C和加权指数m。通常，C要远小于聚类样本的总数，同时还要保证C>1。算法的输出是C个聚类中心点向量和一个的模糊划分矩阵，这个矩阵表示每个样本点属于每个类的隶属度，根据这个矩阵并按照模糊集合中的最大隶属度原则就能确定每个样本点归为哪个类，聚类中心点表示每个类的平均特征，可以认为是这个类的代表点。

加权指数m是控制算法柔性的参数，如果m取值不合适，不仅会影响FCM 算法的收敛性，而且会影响模糊聚类的分类性能。如果m取值过大，则聚类效果比较差；如果m取值过小，则算法将退化为HCM算法。研究表明，m的最佳取值区间为[1。5，2。5]，在没有特殊要求时，可以去中间值m=2。

#### 4.1.4 话题跟踪算法

话题跟踪算法是一个文本分类过程，与文本聚类不同，文本分类是一种有监督的学习过程，需要事先给定一个分类体系和标注好类别的文本集合，利用这些资源来构造一个分类器，将待分类文本归入不同的、预先定义的类别中。主要介绍：KNN算法。

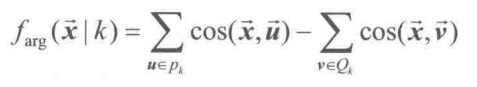
* **KNN算法：**

KNN是一种基于机器学习的分类算法，其实质就是记忆，即把新的只是存储起来，供需要时候使用，而不需要推理和计算。KNN是性能比较号的文本分类算法之一，其他的比较好的方法还有SVM、决策树、神经网络等。

KNN算法思想比较简单，对于给定的一个测试文档，在训练集中查找距离它最近的*k*个邻居，并根据这些邻居的类别，给该文档的候选类评分，把邻居文档和测试文档的相似度作为邻居文档所在类的权重。如果这*k*个邻居中的部分文档属于同一个类别，则将该类中的每个邻居的权重之和作为该类和测试文档的相似度，通过对候选类评分的排序，然后给出一个阈值，就可以判定测试文档的类别。

话题跟着实际上是二元分类问题，利用KNN算法进行话题跟踪的基本思路也比较简单，在给定一个行的文本之后，考虑在训练文本集中与该文本距离最近（相似度）的*k*个文本，根据这*k*个文本所属的正例、反例的相似度值大小来判定新文本是否属于该话题，具体的算法步骤如下：

1. 使用余弦系数法作为相似度度量方法，分别计算新文本与训练集文本的相似度值，选出新文本最相似的*k*个文本。
2. 在这*k*个文本中，抽取属于正例样本集合的所有文本，将这些文本与新闻本*x*的相似度值求和，作为新文本*x*与正例集合的相似度值。同样，在最相近*k*个文本抽取属于反例集合的所有文本，将这些文本与新闻本*x的相似度值*求和，作为新文本*x与反例集合的相似度值。*
3. 比较新文本与正例、反例的相似度值，计算公式如下：



1. 当，判定新文本x属于该话题，当，判定新闻本*x*不属于该话题。

#### 4.1.5 热点话题检测（舆情分析领域）

在网络论坛中，网民的观点是通过发帖或回帖进行传播的，再检测突发性热点话题时，需要考虑网络论坛结构、话题特征以及突发性热点话题所具备的高关注度和事件突发特性，采用噪声过滤、文本聚类等方法实现对突发新热点话题的检测与跟踪。

面对网络论坛的突发性热点话题检测方法包括候选话题集构建、噪声过滤、热点话题检测、热点话题跟着4个步骤。

1. 候选话题集构建：对于采集到的网络论坛数据，构建以主帖标题为索引的候选话题集；
2. 噪声过滤：对候选话题集进行去噪处理，过滤掉热度值较低的主帖和不具有事件突发性的主帖，这样可以过滤掉网络论坛中大部分不会演变成突发性热点话题的帖子；
3. 特点话题检测：通过分词工具提取主题帖标题中所包含的主题词，并采用聚类方法对主题词进行合并，进而抽取出突发性热点话题；
4. 热点话题跟踪：计算针对突发性热点话题的事件序列，通过绘制其对于的回帖加速度变化曲线进行跟踪。

**具体流程：**

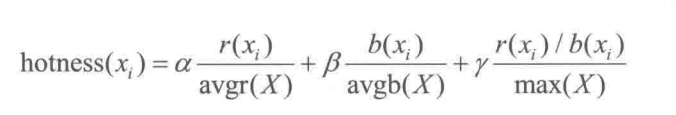
（1）候选话题集构建

在网络论坛中，主帖的标题通常代表了用户讨论的主题。根据2011年中国互联网舆情分析报告中列出的前10大网络舆情热点事件，以影响力较大的网络论坛-天涯社区为对象，检索出该社区中对应10大网络热点事件的回复数最大的主贴，统计发现，能够概括事件的主题的主帖标题占80%以上。因此，可以讲完了论坛中的主帖标题作为索引来构建候选话题集。

（2）噪声过滤

网络论坛的热点话题的形成分为内容驱动和形式驱动两种不同的方式。内容驱动时通过发表内容丰富的铁质来吸引大量的网民浏览和回复，从而形成热帖和热点话题；形式驱动时推手通过调动网络水军发表大量有关事件的帖子，将该事件强行推入公众视野，形成热点话题。

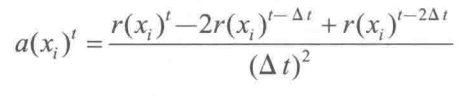
通过计算热点帖子热度，给帖子热度评分，过滤掉热度值较低的帖子。帖子热度值计算公式如下：

 （4-2）

式中，均为加权值，计算得到、、；为主帖的回复数；为主帖的点击数；为所有帖子*X*的平均回复数；为所有帖子*X*的平均点击数；为所有帖子*X*中最大回复数与点击数之比。

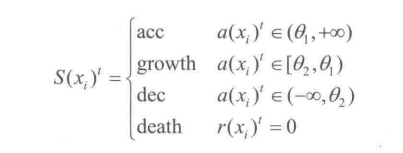
一般而言，用户必须点击某个帖子才能对它进行恢复，因此帖子点击数一般要大于回复数。在（4-2）中，以来描述用户围绕主贴讨论的活跃程度，比值越大、表明该帖子相对应的话题关注度越高。

根据网络论坛舆情演化理论，帖子的生命周期可以分为突发、成长、衰退、消解四个阶段。处于衰退阶段的帖子即使其热度值很高，已不可能演变成具有事件突发性的热点话题。因此，采用回复加速度的概念来识别和莲花帖子的热度随事件的变化特性：

 （4-3）

公式中，为主帖的回复加速度；为主帖在事件*t*的回复数；为介于之间的时间粒度。

同时，定义一个状态函数：，以标识帖子生周期中的不同阶段，该函数定义为：



公式中，acc表示帖子生命周庆的突发期，growth表示帖子生命周期的成长期，dec表示帖子生命周期的衰退期，death表示帖子生命周期的消解期，均为预先设定的阈值，为常量0。

在突发期和衰退期，帖子回复数或急剧增加，或强烈衰退；在成长期，帖子在单位时间内的回复数基本不变；而在消解期，帖子回复数为0。

通过对帖子的热度度量和热度随时间变化的特性，不仅可以过滤掉热度值较低或已出阈衰退、消解阶段的帖子，保留出阈加速或成长阶段的帖子，还可以对回复加速度急剧增加的帖子给予足够的关注，而这些帖子往往容易发展成为突然性热点话题。

对去噪后保留下来的帖子建立主题集合，其中为主题词。

（3）热点话题检测

在对主题进行聚类之前，需要对主题进行分词处理，抽取出能够反映突发性热点话题特征的主题词。在中文分词中，采用中文分词系统进行分词，按照词性仅仅抽取其中的名词与动词，删除重用词。

在实际的操作过程中，需要对主题集合ST中的主题词分别进行分词处理，并分别建立集合。同时，考虑到网络论坛中很多主题讨论的话题事显示中的新时间，如“房妹”、“房姐”等，在分词之前需要手动添加到中文分词系统的词库中。

经过分词处理之后，抽取突发性热点话题的主题词，基于主题词对热点话题进行检测。

热点话题检测步骤：

|  |
| --- |
| **算法4.2 热点话题检测算法** |
| 输入：和预设阈值  输出：热点话题集合 H  方法：   1. 初始化主题集合，分为每m个主题； 2. 对于，选取最大主题max()，且有 ，加入到热点话题集合H中，否则转入（3）； 3. ，，迭代计算转入（2），送ST中再次选取最大主题|，直到集合ST为空。 |

### 4.2 文本情感分析技术

网络文本内容通常是对一些新闻时事、社会热点、法规政策、公众人物、消费产品等话题的个人评价，反映了网民个体的观点。由于每个网民的立场、出发点、个人偏好的不同，对现实世界中各种事物和事件所表达出的立场、态度、意见和情绪的倾向性必然存在很大的差异。对于相同的事物或事件，不同的人从不同视角出发，往往持有不同的观点和态度。例如，对于某个产品，一些用户可能因喜欢该产品某些功能或特点而给出正面(即褒义)评价;而另一些用户也可能因不喜欢该产品某些功能或特点而给出负面(即贬义)评价。

面对互联网中的海量信息，不仅在数量上难以逐一浏览，而且由于受到用户主观认识影响，往往表现出复杂多样的特点。如何快捷而准确地了解人们对某一产品、事件或政策等所持的观点是褒义还是贬义，必须借助于自动化分析技术，这种技术就是文本情感分析技术

文本情感分析技术主要研究如何对文本所表达的观点、情感、立场、态度等主观性信息进行自动分析，从海量文本中识别出人们对某一产品、事件或政策等所持有的观点是褒义还是贬义，提高对文本情感分析的效率。文本情感分析技术涉及自然语言处理、计算语言学、人工智能、机器学习、信息检索、数据挖掘等多个研究领域，属于交叉性技术。

文本情感分析技术作为一种组织和管理数据的有效手段，在网络舆情监测、电子商务等领域中得到了应用。随着互联网发展和大数据时代的到来，将会不断地扩展文本情感分析技术的应用领域。

在网络舆情监测中，对于一个突发社会公共事件的网络舆情，网民所持有的态度倾向性往往是多元化的，包括正面或负面、赞扬或批评、支持或质疑、肯定或否定等。通过文本情感分析技术，能够自动识别出其态度倾向性，并给出分类统计结果，有助于及时采取应对措施。

在电子商务中，通过文本情感分析技术，对产品评论的褒贬倾向性进行分析，可以帮助生产者和商家及时了解消费者对产品的反馈意见，从而做出准确的商业决策。例如，在电子商务推荐系统中，通过分析消费者对产品及服务的反馈评论和意见，向其他用户推荐受到好评的产品和服务，同时利用用户的反馈信息，对消费市场进行深入分析，对产品和服务进行总结和改进。  
文本情感分析技术主要研究如下3个问题:

(1)主客观分析，即识别出文本内容是主观性表达还是客观性表达，文本情感分析主要关注主观性表达的文本内容

(2)情感倾向分析，对于主观性表达的文本内容，识别出情感倾向性:褒义、贬义还是中性

(3)情感强度分析，即判定文本情感倾向的强弱程度，例如，将情感强度分为强烈贬抑、一般贬抑、客观、一般褒扬、强烈褒扬等类别，判定当前文本的情感倾向所属于的类别。

情感强度分析比较复杂，并且情感强度分类也缺乏被广泛认可的标准。因此，文本情感分析技术主要研究前面两个问题。需要指出的是，文本情感分析技术主要将文本的情感倾向分为褒义和贬义两类，对于网络舆情分析中来说，还不够细致。在此基础上，还需要人工做进一步的统计分析。本节将从句子情感分析方法、段落情感分析方法以及文本情感分析模型3个层面来介绍文本情感分析技术。

#### 4.2.1 基本概念

文本情感分析层次划分：

1. **词语情感分析**

词语情感分析的对象是在特定的句子中出现的词和短语。表达情感的词大多是名词、动词、副词和形容词，其情感倾向可以分为褒义、贬义和中性等三类，词语情感分析包括对词的情感极性、情感强度以及上下文模式等进行分析。在词语情感分析时，需要借助于标注有倾向性的情感词典。不论是英语还是汉语，词汇都是非常丰富和庞大的，并且有相当数量的词语随着语境的变换，其倾向性也会发生相应的变化。因此，构建一个涵盖所有情感词倾向性的完整词典是非常困难的，一般都是面向领域应用来构建情感词典。在构建情感词典时，大多采用在已有的电子词典或词库上进行扩展的方式。例如，在知网(HowNet)的词库上进行扩展。

词语情感分析是文本情感分析的基础,主要涉及以下方面的工作:

1)情感词典的构建及扩展

2)语料库的情感倾向标注

3)新词语的情感倾向判断

4)文本中情感词重要性分析

1. **句子情感分析**

句子情感分析的对象是在特定的上下文中出现的句子,其目的是通过分析句子中的各种主观性信息,判断该句子是主观句还是客观句,对于主观句,进一步提取出句子中的主观关系,实现对句子的情感倾向的判断,同时还要分析与情感倾向性相关的各个要素,如评价对象、情感极性、情感强度等。

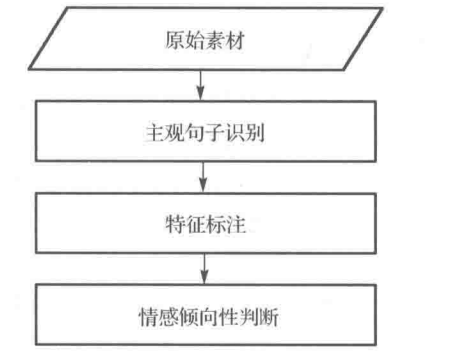


图4-2 句子情感分析流程

句子情感分析是文本情感分析的重点,主要涉及以下方面的工作

1)主题句和主观句识别和提取;

2)主观句中的主观关系识别和提取;

3)主观句的情感倾向性判断;

4)情感倾向性相关要素分析。

1. **文档情感分析**

文档情感分析的对象是一篇完整的文章,从整体上分析某个文章的情感倾向性。由于文档情感分析属于文本分类问题,通常采用机器学习方法,如朴素贝叶斯、最大熵、支持向量机等方法来解决文本情感分析问题,首先构建语料库,人工标注语料库中每个文本的情感倾向,并将语料库分为训练集和测试集,然后对模型进行训练和算法测试,对模型和算法的

文档情感倾向识别能力进行评价。

文档情感分析主要涉及以下方面的工作:

1. 语料库的构建和情感倾向标注；
2. 分类模型选取和模型参数确定；
3. 分类模型训练和测试；
4. 文档全局情感倾向识别。

#### 4.2.2 相关技术流程

目前常用的情感倾向性分析方法是基于情感词典及基于机器学习方法两种。基于机器学习的情感倾向性分析方法，实质上是有监督的学习方法，首先通过标注一些训练数据作为训练样本，同时提取文本的特征，然后利用训练数据训练分类模型得到情感分类器，再通过分类器对测试文本进行预测，得到测试文本的情感倾向值。常用的分类器包括KNN、SVM、Naive Bayes、最大熵模型等。

基于情感词典的方法：

1. 构建情感词典
2. 语义规则处理
3. 计算情感倾向值
4. 情感极性判断

基于机器学习的方法：

#### 4.2.3 基于情感词典方法

**1、构建情感词典：**

根据情感词典分析情感倾向的方法其分析结果的精确度与所选用的情感词典的情感词规模和情感极性标注质量息息相关。为了获得能够满足分析需求的情感词典，首先需要选取一本基础词典，例如知网情感词典等。在这类基础词典中，包含了一些最为常用的情感词，例如“高兴”、“生气”、“开心”、“难过”等；其次需要在基础词典之上搭配同义词词词典，通过寻找基础情感词的近义词来拓展基础词典，可以将加入了基础情感近义词的词典称为拓展词典；最后就需要在进行文本情感分析时，对拓展词典进行扩充与完善。可通过使用PMI(点互信息)方法来进行实现，PMI方法的核心思想是假设如果有一个词，它与积极词一起出现的频率很高，那么这个词在很大程度上就有可能是积极词；反之亦然。因此，可分别计算某个词与积极词和消极词一起出现的频率，计算频率差并设置一个阈值，当频率差超过阈值时，就可判断这个词为情感词应该加入拓展词典，从而对情感词典进行了补充。

当然现在也有很多成熟的情感词典，例如一些常用的情感词典有 BosonNLP、WordNet、GI(General Inquirer)、NTUSD 和 HowNet 词典等。

**2、情感极性分析与情感强度计算方法**

情感极性分析是对一段带有主观意见的文本进行文本情感倾向判断，区分这段文本中所隐含的情感倾向是正面的还是负面的；情感强度则是对不同类型情绪的强烈程度进行了量化表示。情感极性的判断方法目前主要有两种：基于 PMI(pointwise mutual information)的判断方法和基于知网的判断方法

1）情感极性判断方法通过计算待分析的词与预先获取的情感基准词之间的共现频率来判断当前文本的情感倾向。基准词中包含积极词和消极词，且积极词和消极词作为语义对立的词汇通常都会成对出现，因此标准的情感基准词集中正向词汇和负面词汇的总数是非常接近的。

设基准词中积极词集合表示为，消极词集合表示为，则某一个词语的PMI极性SO\_PMI的计算公式如（xxx）所示：

PMI的计算公式如（xxx）所示：

其中，和分别表示和再文本集中出现的概率，表示了和同时再文本集中出现的概率。

2）基于知网的判断方法与基于 PMI 的判断方法类似，也是预先获取一些基准词，然后计算待分析情感词与基准词之间的相似度。

设基准词中积极词的集合表示为，消极词集合表示为，则待分析词的基于知网相似度的词语极性SO\_HOWNET的计算公式如（xxx）所示：

其中，表示的是词与中的个基准词之间的相似读，而则表示词与中第*j*个基准词之间的相似度。

**3 基于情感词典的短文本的情感分析方法：**

短文本是指由一个或几个短句所构成的文本信息；本文所研究的微博平台上的博文、评论就是典型的短文本信息。短文本由于文本量小，因此对上下文的依赖很低，需要通过单句或几个词就能清楚地表达一个意思。所以短文本中体现用户个人意志的相关词就会非常聚集，而基于情感词典的文本情感分析方法非常适合对这类关键词非常聚集的文本进行情感分析。基于情感词典的文本情感分析本质上就是使用情感词典对短文本中的情感词进行提取，通过判断情感词的情感极性来概括该条短文本的情感倾向。基于情感词典的文本情感分析方法的核心是将情感词典和人为制定的语言的使用规则相结合，用具有人工标注的情感词典去与待判定文本进行匹配，统计文本中所包含的正负向情感词的数量，按照公式(xxx)判断情感极性：

#### 4.2.4 基于机器学习方法

随着文本情感分析工作中涉及领域的多样化发展和文本数据量的增加，机器学习算法展现了其优势，结合特征提取使用机器学习算法来完成情感分析工作也引起了研宄人员的关注，取得了一定的成果。

机器学习理论方法可以分为三种：有监督学习、无监督学习以及半监督学习。其中，通过训练有标注的数据集进行的有监督的机器学习模型是应用比较广泛的主流方法。使用机器学习算法研究文本情感分析工作主要是把情感分析看作为对情感态度进行分类的过程。提取文本中的关键语言特征信息，输入到机器学习分类模型，通过样本数据训练的模型来预测其他文本的情感类别。对于不同类型的数据，在任务中可以选择使用不同的机器学习模型，下面将具体介绍常用于情感分析任务的分析模型。

基于机器学习的情感分析研究归纳为如下五个主要步骤：

（1）数据预处理。筛选与情感分类无关的数据，去噪，降低情感维度。

（2）文本分词及词性标注。将文本分成一个个分开的、具有词性标注的词序列。

（3）特征词提取。提取出具有分类能力并且能够表征文本信息的情感特征词，情感特征词的好坏将直接影响分类结果的准确性。

（4）文本结构化表示。本文采用最常用的向量空间模型 VSM 来表示计算机可以识别的结构化表达方式。

（5）权重计算及情感分类实现。权衡特征向量的权重并结合机器学习方法，训练分类器并对测试数据进行预测。

传统的基于机器学习情感分类算法主要有朴素贝叶斯、支持向量机、K近邻模型；

1. **朴素贝叶斯模型：**

朴素贝叶斯模型是一种基于贝叶斯定理以及特征条件独立假设的分类模型。首先在训练数据处理时，通过特征条件独立假设计算输入／输出的联合概率分布，对于给定的输入数据，结合贝叶斯定理得到后验概率最大的预测项，从而得到最终分类结果。

假设类别集合共有*ｎ*种类别，定文本样本*ｘ*可表示为，其中*d*为特征维数。根据贝叶斯公式，给定文本*ｘ*属于类别的概率由公式（xxx）可计算。

其中、分别表示为文本x和类别的先验概率，为条件概率，有可表示为。由于朴素贝叶斯模型基于特征条件独立假设，即文本中各词语之间相互独立且不存在依赖关系，则可以通过式子（xxxx）计算。

其中为样本*x*在第*ｊ*个特征上的取值，由于文档*ｘ*的先验概率对所有类别皆相同，可省略计算，由此得到朴素贝叶斯模型最终计算如公式2.5所示：

朴素贝叶斯模型的优点在于它所需参数较少，对缺失数据敏感度相对较低，算法实现简单。但朴素贝叶斯模型假设特征之间相互独立，这个假设在实际应用中通常情况下是不成立的，这对朴素贝叶斯模型的分类效果造成一定影响。

1. **支持向量机模型**

支持向量机，即通过支持向量运算的分类器。在求解的过程中，只需要根据部分数据就可以确定分类器。支持向量机的基本思想是将输入空间的输入转换到特征空间，在特征空间中找到一个最优分类超平面，将该超平面用于分类。最优分类超平面又被称为最大间隔超平面（maximum-margin hyperplane），分类器又被称为最大间隔分类器（maximum-margin classifier）。寻找最优分类超平面的准则是该平面能够将两类点集分开且最近点的距离最大，对于给定训练集，其中*x*为输入*，y*为标签，超平面记为。

支持向量机分类算法有以下几点优点：首先在小样本情况下该方法相比其他分类算法仍可得到较好的结果。其次该算法可解决非线性问题，提高分化性能，解决高维问题。但它也有一些不足之处，例如该算法对缺失数据敏感、在大规模数据集下将耗费大量的机器内存和运算时间。

1. **K最近邻模型**

Ｋ最近邻（K-Nearest Neighbor, KNN）算法是数据挖掘中基本的分类算法之一，它的主要思想是通过邻居来判断其所属类别。ＫＮＮ算法的工作原理是通过计算待分类样本与训练样本之间的差异性，将差异性从小到大按顺序排序，选出差异最小的前面Ｋ个分类样本，统计在Ｋ个分类样本分别属于的类别，其中出现次数最多的类别即为待分类样本最相似的类别，最终将待分类样本分到该类别中。Ｋ最近邻算法的具体流程如图xxx所示，其中可以发现该算法的三个基本要素，分别为Ｋ取值选择、距离度量和分类决策规则，以下详细描述各要素对Ｋ最近邻算法的影响。



Ｋ取值的不同会造成算法分类结果的不同。当Ｋ值较小时，容易产生过拟合现象。极端情况下取Ｋ＝１，待分类样本只与最接近的训练样本有关，训练误差很小，但若该训练样本为噪声，分类出现误差。当Ｋ值较大，极端情况下取Ｋ＝ｎ，所有待分类样本的结果都是训练数据集中出现次数最多的分类，这样会产生欠拟合。因此在应用中，通常采用交叉验证的方法来选取合适的Ｋ值。

距离度量指的是样本之间的距离，通过距离度量可以得到样本间相似度。通常有以下几种方法进行距离度量，包括闵可夫斯基距离（Minkowski Distance），欧氏距离（Euclidean Distance），曼哈顿距离（Manhattan Distance），切比雪夫距离（Chebyshev Distance）等。其中最常用的是闵可夫斯基距离，由公式（xxx）可见，当ｐ=2时，该式即为欧氏距离公式，p=1时为曼哈顿距离公式。

*Ｋ*最近邻的分类决策通常采用多数表决的方法。对给定待分类样本*x，Ｎ（ｘ）*为该样本最邻近的*Ｋ*个训练样本的集合，为该集合涵盖的类别，采用0-1损失函数（0-1 loss function），则分类误差率计算如公式(xxx)所示。由该公式可知要使得分类误差率小应采用多数表决方法。

Ｋ最近邻算法实现简单、有效，计算时间和空间与训练集的规模呈线性关系，对于离群点不敏感。但该算法不足之处在于当样本不平衡时，即当某些类别的样本容量很大，而其他类别样本容量很小时，待分类样本的Ｋ个邻居中大容量类的样本占多数而被错误分类至该类别。同时Ｋ最近邻算法计算量和内存空间都较大。

#### 4.2.5 基于深度学习方法

由于基于传统机器学习的文本情感分析方法需人工提取文本情感特征，方法的效果也依赖于特征的选取，近年来，基于深度学习的方法也被广泛应用在情感分析任务中。深度学习方法也是机器学习的分支之一，本节简单介绍常见的深度学习方法，包括神经网络、卷积神经网络、循环神经网络及长短时记忆模型。

基于深度学习的情感分析研究归纳为如下五个主要步骤：

（1）数据预处理。筛选与情感分类无关的数据，去噪，降低情感维度。

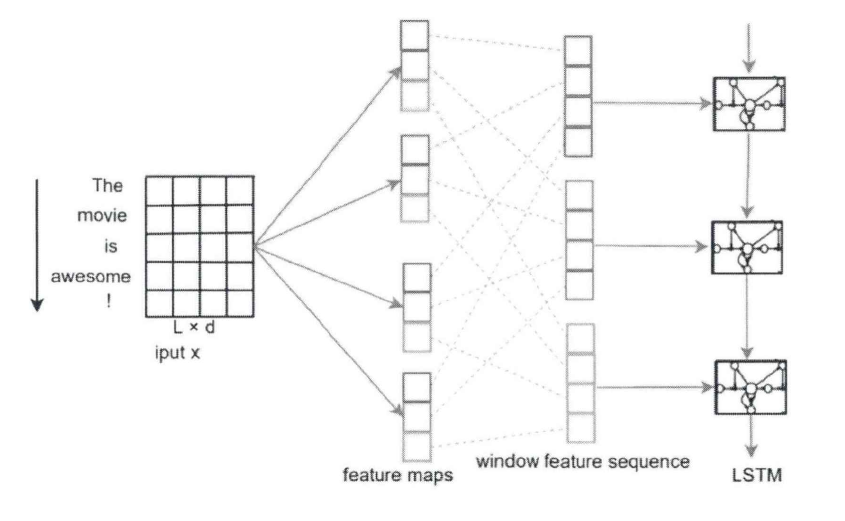
（2）文本分词及词性标注。将文本分成一个个分开的、具有词性标注的词序列。

（3）特征词提取。提取出具有分类能力并且能够表征文本信息的情感特征词，情感特征词的好坏将直接影响分类结果的准确性。

（4）文本结构化表示。本文采用最常用的向量空间模型 VSM 来表示计算机可以识别的结构化表达方式。

（5）权重计算及情感分类实现。权衡特征向量的权重并结合机器学习方法，训练分类器并对测试数据进行预测。

**1）基于CNN和LSTM的情感分析模型**

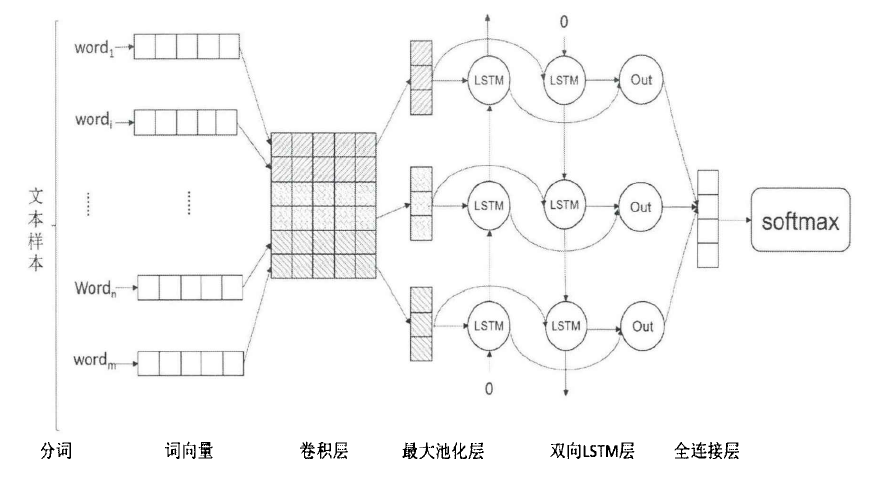


**图3.3 C-LSTM模型神经网络结构**

C-LSTM模型通过使用卷积神经网络对文本词向量进行卷积抽象操作，得到经过卷积的抽象文本序列，并使用长短时记忆神经网络学习文本序列，实验证明该模型优于传统的长短时记忆神经网络。

2)**基于CNN和双向LSTM的情感分析模型**

基于双向长短时记忆单元和卷积神经网络的文本情感分析模型（以下简称CNN-BLSTM模型），是在C-LSTM模型的基础上提出的改进方法。CNN-BiLSTM模型的神经网络结构如图3.4所示，模型的输入是文本词向量，通过文本预处理并通过大规模语料使用Word2vec工具进行训练而来。CNN-BiLSTM模型大致包括两个部分，分别为卷积神经网络和双向长短时记忆神经网络。由于普通的长短时记忆神经网络中每个输出都根据前一步的输出得到，这导致输出只与其前置上下文信息有关，而事实上在文本情感分析任务中，应与前置上下文信息和后置上下文信息都相关，而使用双向长短时记忆单元则能够避免这一点。

图3.４CNN-BiLSTM模型神经网络结构