第 3 章　网络信息内容预处理技术

3.1　网络信息内容预处理概述

计算机和 Internet 的普及，带来了现代社会的信息爆炸，每天都会有海量的信息需要处理，信息的存在方式和形式可以归纳为四个“多”：多媒体、多语言、多文种、多格式。多媒体是指信息存在的媒体多种多样，包括文本、声音、视频等；多语言是指自然语言信息可以是多种语言；多文种是指数字化的信息存放在不同类型的文件中；多格式是指在同一种文件类型中，相同的信息可以以多种格式存放。原始的网络信息内容格式一般较为多样化，在进行内容分析前，需要对其进行预处理。

在众多的网络信息内容中，文本信息又占了很大的比重。文本信息是指用文本或带有格式标志信息的文本来存放的信息，如纯文本文件、HTML 文件及各种字处理器产生的文件等，其中又有自由文本(Free Text)和自然语言文本(Natural Language Text)之分。自由文本是指任何以文本形式存在的信息，包括程序源代码、数据等；自然语言文本则是指以文本形式存在的、主要是自然语言书写的信息。自然语言文本还可以由多种语言书写。以下约定，如果不作特别的说明，本书所说的文本是指中文的自然语言文本。

对文本信息的处理包括文本信息的分类、检索和浓缩等。目前在这几个方面的研究都取得了很大的进展，产生了许多可喜的成果。如上海交大纳讯公司由王永成教授主持开发的中英文自动摘要系统，在信息浓缩和抽取等方面的研究处于世界领先的地位，摘要的质量可以达到与手工摘要无明显差别甚至稍高的程度。但是，这些成果的研究大都是建立在比较理想的条件下。所谓的理想条件，是指所处理的文本信息的形式比较单一(大多是纯文本信息)，格式比较规范，文本中的一些特征信息比较清晰、容易识别等。而现实中的各种文本信息，形式多样化，格式不是都很规范，而且一些重要的特征信息比较模糊，这些可以称为文本信息的噪声和变形。噪声和变形的存在使处理文本信息非常困难，达不到预想的质量。在将实验室的研究成果产品化，推向市场的时候，就会面临这样一个问题：如何去除和减弱文本信息噪声和变形的影响。

这也是许多文本信息处理软件所遇到的一个共同的问题。为了便于交流使用，许多国家和地区都制定了不少信息发布的标准，但这些标准不可能包括信息发布的所有形式，而且即使是标准本身，因为各国所使用的媒体、语言、代码、控制符以及格式等都不一定相同，在信息交流中也会出现困难。为了方便对文本信息进一步的加工处理，全世界掀起了一个研究与开发“预处理器”的热潮。一般来说，网络信息内容预处理流程包括中文分词、去停用词、语义特征提取、特征子集选择、特征重构、向量生成和文本内容分析等几个步骤。下面将对这些步骤进行依次介绍。

3.1.1　中文分词

中文是以字为基本书写单位，单个字往往不足以表达一个意思，通常认为词是表达语义的最小元素。在汉语中，一句话的意思通过一段连续的字符串来表达，字符串之间并没有明显的标志将其分开，计算机如何正确识别词语是非常重要的步骤。例如，一条英文文本消息“I love this movie.”，其汉语意思为“我喜欢这部电影。”计算机处理过程中，可以依靠空格识别出 movie 是一个词，但不能识别的“电”和“影”是一个词，只有将“电影”切分在一起才能表达正确意思。因此，须对中文字符串进行合理的切分，可认为是中文分词。下面将分别对分词技术特点与分词系统作介绍。

(1) 中文信息处理首要解决的就是对文本内容进行分词。如何实现准确、快速的分词处理，是自然语言处理领域研究中的一个难点。当前主要的分词处理方法分为基于字符串匹配的分词方法、基于统计的分词方法和基于理解的分词方法。这三类分词技术代表了当前的发展方向，有着各自的优缺点。

基于字符串匹配的分词方法优点是：分词过程跟词典作比较，不需要大量的语料库、规则库，其算法简单，复杂性小，对算法作一定的预处理后分词速度较快。缺点是：不能消除歧义、识别未登录词，对词典的依赖性比较大，若词典足够大，其效果会更加明显。

基于统计的分词方法优点是：由于是基于统计规律的，因此对未登录词的识别表现出一定的优越性，不需要预设词典。缺点是：需要一个足够大的语料库来统计训练，其正确性很大程度上依赖于训练语料库的质量好坏，算法较为复杂，计算量大，周期长，但是都较为常见，处理速度一般。

基于理解的分词方法优点是：由于能理解字符串含义，对未登录词具有很强的识别能力，因此能很好地解决歧义问题，不需要词典及大量语料库训练。缺点是：需要一个准确、完备的规则库，依赖性较强，效果好坏往往取决于规则库的完整性。算法比较复杂，实现技术难度较大，处理速度比较慢。

(2) 常用的中文分词系统。中文分词技术是对汉语文本进行处理的基础要求，一直是自然语言处理领域的研究热点，目前已取得了很多成果，出现一大批实用、可靠的中文分词系统。其代表有：基于 Lucene 为应用主体开发的 IKAnalyzer 中文分词系统、庖丁中文分词系统，纯 C 语言开发的简易中文分词系统 SCWS，中国科学院计算技术研究所推出的汉语词法分析系统 ICTCLAS，哈尔滨工业大学信息检索研究室研制的 IRLAS，另外国内北大语言研究所、清华大学、北京师范大学等机构也推出了相应的分词系统。

林林总总的分词系统各有其特点，例如 IKAnalyzer 实现了以词典分词为基础的正反向全切分算法，更多的用于互联网的搜索和企业知识库检索领域；庖丁中文分词系统致力于成为互联网首选的中文分词开源组件，它追求分词的高效率和用户的良好体验；而简易中文分词系统 SCWS 目前仅用于 UNIX 族的操作系统；哈工大 IRLAS 主要采用 Bigram 语言模型，大大提高了对未登录词识别的性能。目前来看，表现最为抢眼的无疑是中国科学院研制 ICTCLAS，该分词系统综合性能十分突出，在国内外权威机构组织的多次公开评测中都取得优异成绩，已得到国内外大多数中文信息处理用户的支持。

3.1.2 停用词

停用词也称为功能词，与其他词相比，通常是没有实际含义的。在中文信息处理中，停用词一般是指在文本内容中出现频率极高或者极低的介词、代词、虚词以及一些与情感无关的字符。这些字符在中文信息研究中没有实际意义。若计算机对其进行处理，不但是没有价值的工作，还会增加运算复杂度，通常文本的停用词处理中可采用基于词频的方法将其除去。王素格与魏英杰构造 5 种不同的停用词词表作为候选特征依据，对汽车语料进行情感分类研究，考查对最终分类结果的影响，其结果表明，无停用词表，即全部作为候选特征与选用除了动词、副词、形容词的停用词表对情感分类的结果比较好。

3.2 语义特征抽取

根据语义级别由低到高来分，文本语义特征可分为亚词级别、词级别、多词级别、语义级别和语用级别。其中，应用最为广泛的是词级别。

3.2.1 词级别语义特征

词级别(Word Level)以词作为基本语义特征。词是语言中最小的、可独立运用的、有意义的语言单位，即使在不考虑上下文的情况下，词仍然可以表达一定的语义。以单词作为基本语义特征在文本分类、信息检索系统中工作良好，也是实际应用中最常见的基本语义特征。

在英文文本中以词为基本语义特征的优点之一是易于实现，利用空格与标点符号即可将连续文本划分为词。如果进一步简化，忽略词之间的逻辑语义关系及词与词之间的顺序，则文本将被映射为一个词袋(Bag of Words)，在词袋模型中只有词及其出现的次数被保留下来。图 3-1 为一个转换示例。

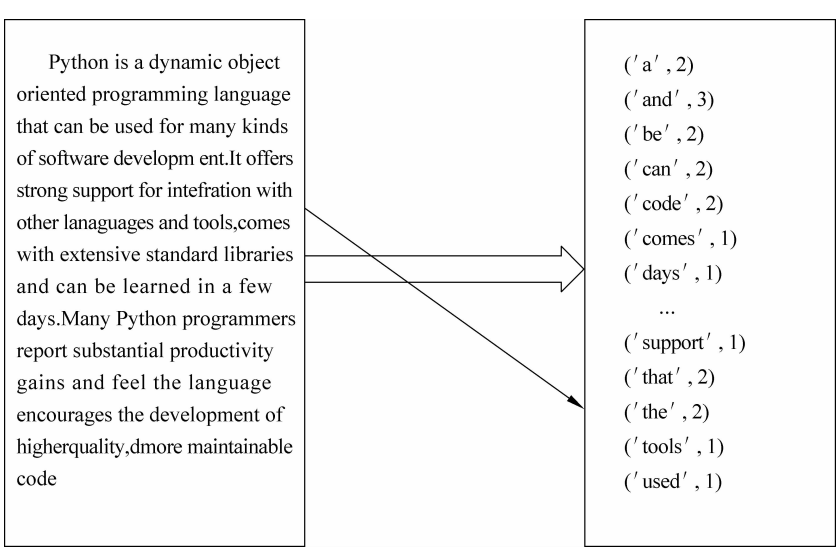

图 3-1 词袋模型

以词为基本语义特征会受到一词多义与多词同义的影响，前者指同一单词可用于描述不同对象，后者指同一事物存在多种描述形式。虽然一词多义与多词同义现象在普通文本信息中并非罕见，且难以在词特征索引级别有效解决，但是这种现象对分类的不良影响却较小，例如英文中常见的 book、bank 等词汇存在一词多义现象，在网络内容安全中判断一个文本是否含有不良信息时并不易受其影响。对使用词作为基本语义特征有较好分类效果，Whorf 曾经做过相关分析，认为在语言的进化过程中，词作为语言的基本单位朝着能优化反映表达内容、主题的方向发展，因此词汇有力地表示了分类问题的前沿分布。

当英文以词为特征项时，需要考虑复数、词性、词格、时态等词形变化问题。这些变化形式在一般情况下对于文本分类没有贡献，有效识别其原始形式并合为统一特征项，有利于降低特征数量，并避免单个词被表达为多种形式带来的干扰。

词特征可进行计算的因素有很多，最常用的有词频、词性等。

1. 词频

文本内容中的中频词往往具有代表性，高频词区分能力较小，而低频词或者未出现词常常可以作为关键特征词，所以词频是特征提取中必须考虑的重要因素，并且在不同方法中有不同的应用公式。

2. 词性

在汉语言中，能标识文本特性的往往是文本中的实词，如名词、动词或形容词等，而文本中的一些虚词，如感叹词、介词或连词等，对于标识文本的类别特性并没有贡献，也就是对确定文本类别没有意义。如果把这些对文本分类没有意义的虚词作为文本特征词，将会带来很大影响，从而直接降低文本分类的效率和准确率。因此，在提取文本特征时，应首先考虑剔除这些对文本分类没有用处的虚词；而在实词中，又以名词和动词对文本类别特性的表现力最强，所以可以只提取文本中的名词和动词作为文本的一级特征词。

3. 文档、词语长度

一般情况下，词的长度越短，其语义越泛。通常，中文中较长的词往往反映比较具体、下位的概念，而短的词往往表示相对抽象、上位的概念。短词具有较高的频率和更多的含义，是面向功能的；而长词的频率较低，是面向内容的。增加长词的权重，有利于词汇进行分割，从而更准确地反映特征词在文章中的重要程度，词语长度通常不被研究者重视，但是在实际应用中发现，关键词通常是一些专业学术组合词汇，长度较一般词汇长。考虑候选词的长度，会突出长词的作用，长度项也可以使用对数函数来平滑词汇间长度的剧烈差异，通常来说，长词汇含义更明确，更能反映文本主题，适合作为关键词，因此需要将包含在长词汇中低于一定过滤阈值的短词汇进行过滤。所谓过滤阈值，就是指进行过滤短词汇的后处理时，短词汇的权重和长词汇的权重比的最大值如果低于过滤阈值，则过滤短词汇；否则，保留短词汇。

根据统计，两字词汇多是常用词，不适合作为关键词，因此对实际得到的两字关键词可以作出限制。例如，抽取 5 个关键词(本文最多允许 3 个两字关键词存在)。这样的后处理无疑会降低关键词抽取的准确度和召回率，但是同候选词长度项的运用一样，人工评价效果将会提高。

4. 词语直径

词语直径(Diameter)是指词语在文本中首次出现的位置和末次出现的位置之间的距离。词语直径是根据实践提出的一种统计特征。根据经验,如果某个词汇在文本开头处提到,在结尾处又提到,那么它对该文本来说将是个很重要的词汇,不过统计结果显示,关键词的直径分布出现了两极分化的趋势,在文本中仅仅出现了 1 次的关键词占全部关键词的 14.184%,所以词语直径是比较粗糙的度量特征。

5. 首次出现位置

Frank 在 Kea 算法中使用候选词首次出现位置(First Location)作为 Bayes 概率计算的一个主要特征,它被称为距离(Distance),简单地统计可以发现,关键词一般在文章中较早出现,因此出现位置靠前的候选词应该加大权重,实验数据表明,首次出现位置和词语直径两个特征只选择一个使用就可以了。例如,由于文献数据加工问题导致中国学术期刊全文数据库的全文数据,不仅包含文章本身,而且还包含了作者、作者机构及引文信息。针对这一特点,可以使用首次出现位置这个特征,尽可能减少由全文数据的附加信息所造成的不良影响。

6. 词语分布偏差

词语分布偏差(Deviation)所考虑的是词语在文章中的统计分布,在整篇文章中分布均匀的词语通常是重要的词汇。

3.2.2 亚词级别语义特征

亚词级别(Sub-Word Level)也称为字素级别(Graphemic Level)。在英文中比词级别更低的文字组成单位是字母,在汉语中则是单字。

英文有 26 个字母,每个字母有大小写两种形式。英文中大小写的区别并不在于内容方面,因此在表示文本时通常合并大小写形式,以简化处理模型。

1. *n* 元模型

亚词级别常用的索引方式是 n 元模型(n-Grams)。n 元模型将文本表示为重叠的 n 个连续字母(对应汉语情况为单字)的序列作为特征项,例如,单词 shell 的三元模型为 she、hel 和 ell(考虑前后空格,还包括_sh 和 ll_两种情况),英文中采用 n 元模型有助于降低错误拼写带来的影响:一个较长单词的某个字母拼写错误时,如果以词作为特征项,则错误的拼写形式和正确的词没有任何联系。若采用 n 元模型表示,当 n 小于单词长度时,错误拼写与正确拼写之间会有部分 n 元模型相同;另外,考虑到英文中复数、词性、词格、时态等词形变化问题,n 元模型也起到与降低错误拼写影响类似的作用。

采用 n 元模型时,需要考虑数值 n 的选择问题。当 $n<3$ 时,无法提供足够的区分能力(在此只考虑 26 个字母的情况);$n=3$ 时,有 $26^3=17\,576$ 个三元组;$n=4$ 时,有 $26^4=456\,976$ 个四元组。n 取值越大,可表示的信息越丰富,随着 n 的增大,特征项数目也以指数函数方式迅速增长,因此,在实际应用中大多取 n 为 3 或 4(随着计算机硬件技术的增长,以及网络的发展对信息流通的促进,已经有 n 取更大数值的实际应用)。仅考虑单词平均长度的情况,本文统计了一份 GRE 常用词汇表,7444 个单词的平均长度为 7.69;考虑到不同单词在真实文本中出现的频率不同,统计 reuters-21578(路透社语料库),平均长度为 4.98 个

字母；考虑到长度较短单词使用频率较高，而拼写错误词汇一般长度较长，可见采用 $n=3$ 或 4 可以部分弥补错误拼写与词形变化带来的干扰，并且有足够的表示能力。

2. 多词级别语义特征

多词级别(Multi-Word Level)指用多个词作为文本的特征项，多词可以比词级别表示更多的语义信息。随着时代的发展，一些词组也越来越多地出现，例如英文 machine learning、network content security、text classification、information filtering 等，对于这些术语，采用单词进行表示会损失一些语义信息，因为短语与单个词在语义方面有较大区别；随着计算机处理能力的快速增长，处理文本的技术也越来越成熟，多词作为特征项也有更大的可行性。多词级别中的一种思路是应用名词短语作为特征项，这种方法也称为 Syntactic Phrase Indexing，另外一种策略则是不考虑词性，只从统计角度根据词之间较高的同现频率(Co-Occur Frequency)来选取特征项，采用名词短语或者同现高频词作为特征项，需要考虑特征空间的稀疏性问题，词与词可能的组合结果很多，下面仅以两个词的组合为例进行介绍。根据统计，一个网络信息检索原型系统包含的两词特征项就达 10 亿项，而且许多词之间的搭配是没有语义的，绝大多数组合在实际文本中出现频率很低，这些都是影响多词级别索引实用性的因素。

3.2.3　语义与语用级别语义特征

如果我们能获得更高语义层次的处理能力，例如实现语义级别(Semantic Level)或语用级别(Pragmatic Level)的理解，则可以提供更强的文本表示能力，进而得到更理想的文本分类效果。然而在目前阶段，由于还无法通过自然语言理解技术实现对开放文本理想的语义或语用理解，因此相应的索引技术并没有前面的几种方法应用广泛，往往应用在受限领域。在自然语言理解等研究领域取得突破以后，语义级别甚至更高层次的文本索引方法将会有更好的实用性。

3.2.4　汉语的语义特征抽取

1. 汉语分词

汉语是一种孤立语，不同于印欧语系的很多具有曲折变化的语言，汉语的词汇只有一种形式而没有诸如复数等变化。另外，汉语不存在显式(类似空格)的词边界标志，因此需要研究中文(汉语和中文对应的概念不完全一致，在不引起混淆的情况下，文本未进行明确区分而依照常用习惯选择使用)文本自动切分为词序列的中文分词技术，中文分词方法最早采用了最大匹配法，即与词表中最长的词优先匹配的方法。根据扫描语句的方向，可以分为正向最大匹配(Maximum Match，MM)、反向最大匹配(Reverse Maximum Match，RMM)，以及双向最大匹配(MM)等多种形式。

梁南元的研究结果表明，在词典完备、不借助其他知识的条件下，最大匹配法的错误切分率为 169～245 字/次，该研究实现于 1987 年，以现在的条件来看，当时的实验规模可能偏小，另外，如何判定分词结果是否正确也有较大的主观性，最大匹配法由于思路直观、实现简单、切分速度快等优点，所以应用较为广泛，采用最大匹配法进行分词遇到的基本问题是切分歧义的消除问题和未登录词(新词)的识别问题。

为了消除歧义，研究人员尝试了多种人工智能领域的方法：如松弛法、扩充转移网络法、短语结构文法、专家系统法、神经网络法、有限状态机方法、隐马尔科夫模型、Brill 式转换法，这些分词方法从不同角度总结歧义产生的可能原因，并尝试建立歧义消除模型，也达到了一定的准确程度，然而由于这些方法未能实现对中文词的真正理解，也没有找到一个可以妥善处理各种分词相关语言现象的机制，因此目前尚没有广泛认可的完善的歧义消除方法。

未登录词识别是中文分词时遇到的另一个难题，未登录词也称为新词，是指分词时所用词典中未包含的词，常见有人名、地名、机构名称等专有名词，以及相关领域的专业术语，这些词不包含在分词词典中却对分类有贡献，就需要考虑如何进行有效识别。孙茂松、邹嘉彦的相关研究指出，在通用领域文本中，未登录词对分词精度的影响超过了歧义切分。

未登录词识别可以从统计和专家系统两个角度进行：统计方法从大规模语料中获取高频连续汉字串，作为可能的新词；专家系统方法则是从各类专有名词库中总结相关类别新词的构建特征、上下文特点等规则，当前对未登词的识别研究，相对于歧义消除来说更不成熟。

孙茂松、邹嘉彦认为分词问题的解决方向是建设规模大、精度高的中文语料资源，以此作为进一步提高分词技术的研究基础。

对于文本分类应用的分词问题，还需要考虑分词颗粒度问题。该问题考虑存在词汇嵌套情况时的处理策略，例如，“文本分类”可以看作是一个单独的词，也可以看作是“文本、分类”两个词，应该依据具体的应用来确定分词颗粒度。

2. 汉语亚词

在亚词级别，汉语处理也与英语存在一些不同之处。一方面，汉语中比词级别更低的文字组成部分是字，与英文中单词含有的字母数量相比偏少，词的长度以 2～4 个字为主，对搜狗输入法中 34 万条词表进行统计，不同长度词所占词表比例分别为两字词 35.57%、三字词 33.98%、四字词 27.37%，其余长度共 3.08%。

另一方面，汉语包含的汉字数量远远多于英文字母数量，GB 2312—1980 标准共收录 6763 个常用汉字（GB 2312—1980 另有 682 个其他符号，GB 18030—2005 标准收录了 27 484 个汉字，同时还收录了藏文、蒙文、维吾尔文等主要的少数民族文字），该标准还是属于收录汉字较少的编码标准。在实际计算中，汉语的二元模型已超过英文中五元模型的组合数量，即 $6763^2(45\ 738\ 169) > 26^5(11\ 881\ 376)$。

因此，汉语采用 n 元模型就陷入了一个两难境地：n 较小时（$n=1$），缺乏足够的语义表达能力；n 较大时（$n=2$ 或 3），则不仅计算困难，而且 n 的取值已经使得 n 元模型的长度达到甚至超过词的长度，又失去了英语中用于弥补错误拼写的功能。因此汉语的 n 元模型往往用于其他用途，在中文信息处理中，可以利用二元或一元汉字模型来进行词的统计识别，这种做法基于一个假设，即词内字串高频同现，但并不阻止词的字串低频出现。

在网络内容安全中，n 元模型也有重要的应用，对于不可信来源的文本，可以采用二元分词方法（即二元汉字模型），例如“一二三四”的二元分词结果为“一二”、“二三”和“三四”，这种表示方法，可以在一定程度上消除信息发布者故意利用常用分词的切分结果来躲避过滤的情况。

3.3　特征子集选择

特征子集选择从原有输入空间，即抽取出的所有特征项的集合，选择一个子集合组成新的输入空间。输入空间也称为特征集合。选择的标准是要求这个子集尽可能完整地保留文本类别区分能力，而舍弃那些对文本分类无贡献的特征项。

机器学习领域存在多种特征选择方法，Guyon 等人对特征子集选择进行了详尽讨论，分析比较了目前常用的 3 种特征选择方式：过滤（Filter）、组合（Wrappers）与嵌入（Embedded），文本分类问题由于训练样本多、特征维数高等特点，决定了在实际应用中以过渡方式为主，并且采用评级方式（Single Feature Ranking），即对每个特征项进行单独的判断，以决定该特征项是否会保留下来，而没有考虑其他更全面的搜索方式，以降低运算量，在对所有特征项进行单独评价后，可以选择给定评价函数大于某个阈值的子集组成新的特征集合，也可以评价函数值最大的特定数量特征项来组成特征集，特征子集选择涉及文本中的定量信息，一些相关参数定义如表 3-1 所示。

表 3-1　文档及特征项各参数含义

参数	含　义
N	训练样本数
n_{c_i}	c_i 类别包含的训练样本数
$n(t)$	包含特征项 t 至少一次的训练样本数
$\bar{n}(t)$	不包含特征项 t 的训练样本数
$n_{c_i}(t)$	c_i 类别包含特征项 t 至少一次的训练样本数
$\bar{n}_{c_i}(t)$	c_i 类别不包含特征项 t 的训练样本数
tf	所有训练样本中所有特征项出现的总次数
$tf(t)$	特征项 t 在所有训练样本中出现的次数
$tf_{d_j}(t)$	特征项 t 在文档 d_j 中出现的次数

很容易可知，参数间满足如下关系：

$$n = \sum_{i=1}^{k} n_{c_i} \tag{3-1}$$

$$n(t) = \sum_{i=1}^{k} n_{c_i}(t) \tag{3-2}$$

式(3-1)表示样本总数等于各类别样本数之和，式(3-2)表示只包含任一特征项 t 的样本集合，也满足类似关系。

$$n = n(t) + \bar{n}(t) \tag{3-3}$$

$$n_{c_j} = n_{c_i}(t) + \bar{n}_{c_i}(t) \tag{3-4}$$

式(3-3)表示 $n(t)$和$\bar{n}(t)$互补，式(3-4)表示这种关系也适用于任意给定文本类别。

$$tf = \sum_{i=1}^{\hat{m}} tf(t_i) \tag{3-5}$$

$$tf(t) = \sum_{j=1}^{n} tf_{d_j}(t) \tag{3-6}$$

式(3-5)和式(3-6)给出了 tf 和 $tf(t)$的计算方法。

利用这些参数,结合统计、信息论等学科,即可进行特征子集选择,最简单的方式是停用词过滤。

3.3.1 停用词过滤

停用词过滤(Stop Word Elimination)基于对自然语言的观察,存在一些几乎在所有样本中出现,但是对分类没有贡献的特征项。例如,当以词作为特征项时,英语中的冠词、介词、连词和代词等。这些词的作用在于连接其他表示实际内容的词,以组成结构完整的语句。

停用词词表可以手工建立,也可以通过统计自动生成,英语领域有手工建立领域无关和面向具体领域的停用词词表,一般停用词表中含有数十到数百个停用词,汉语的停用词表较英语可用资源少一些,对于特征项抽取时采用亚词级别的 n 元模型情况,应当先进行停用词过滤,然后再对文本内容进行 n 元模型构建,对于多词级别采用相邻词构成特征项的情况,也可先进行停用词去除。

除手工建立停用词词表外,还可采用统计方法,统计某一个特征项 t 在训练样本中出现的频率($n(t)$或 $tf(t)$),当达到限定阈值后,则认为该特征项在所有类别或大多数文本中频繁出现,对分类没有贡献能力,因此作为停用词而被去除。

针对具体应用还可以建立相关领域的停用词表,或者用于调整领域的无关停用词表。例如,汉字的"的"字,通常可以作为停用词,但在某些领域,有可能"的"字是某个专有名词的一部分,这时就需要将其从停用词表中去除,或调整停用策略。

3.3.2 文档频率阈值法

文档频率阈值法(Document Frequency Threshold)用于去除训练样本集中出现频率较低的特征项,该方法也称 DF 法。对于特征项 t,如果包含该特征项的样本数 $n(t)$小于设定的阈值 δ,则去除该特征项 t,通过调节 δ 值能显著地影响可去除的特征项数。

文档频率阈值方法基于如下猜想:如果一个作者在写作时经常重复某一个词,则说明作者有意强调该词,该词同文章主题有较强的相关性,从而也说明这个词对标识文本类别的重要性;另外,不仅在理论上可以认为低频词和文本主题、分类类别相差程度不大,在实际计算中,低频词由于出现次数过低,也无法保证统计意义上的可信度。

语言学领域存在一个与此相关的统计规律——齐夫定律(Zipf Laws),美国语言学家 Zipf 在研究英文单词统计规律时,发现将单词按照出现的频率由高到低排列,每个单词出现的频率 rank(t)与其序号 $n(t)$存在近似反比关系:

$$\mathrm{rank}(t) \cdot \mathrm{TF}(t) \approx C \tag{3-7}$$

中文也存在类似规律,对新浪滚动新闻的 133 577 篇新闻的分词结果进行统计,结果见图 3-2,其中 x 轴表示按照词频(特征项频率)逆序排列的序号,y 轴表示该特征项出现的次数。

这个规律说明,在训练样本集中大多数词低频出现(由于这一特点,这一语言规律也称为长尾(Long Tail)现象),解释了文档频率阈值法只需不太大的阈值,就能够明显降低维数的原因。另外,对于出现次数较多的项,有可能属于停用词性质,应当去除。因此,对于汉语

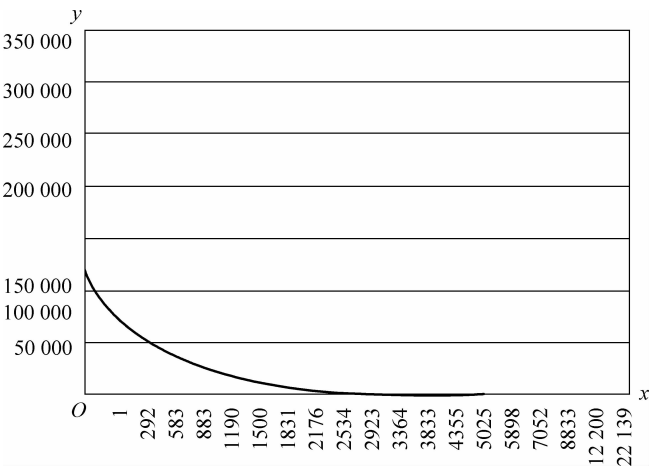

图 3-2　一个中文预料的齐夫定律现象验证

没有成熟的停用词词表，尤其对于网络内容安全相关的停用词表情况，单纯使用文档频率阈值法，会包含一些频率较高而对分类贡献较小的特征项。

3.3.3　TF-IDF

特征项频率——逆文本频率指数（Term Frequency-Inverse Document Frequency，TF-IDF）可以看作是文档频率阈值法的补充与改进。文档频率阈值法认为，出现次数很少的特征项对分类贡献不大，可以去除。TF-IDF 方法则结合考虑两个部分：第一部分认为，出现次数较多的特征项对分类贡献较大；第二部分认为，如果一个特征项在训练样本集中的大多数样本中都出现，则该特征项对分类贡献不大，应当去除。

一个直观的特例：如果一个特征项 t 在所有样本中都出现，这时有 $n(t) = n$，保留 t 作为特征，特征值采取二进制值表示方式时（特征出现时，特征值为 1；特征不出现时，特征值为 0），则该特征没有任何分类贡献，因为对应任一样本，该特征项都取 1，所以应当去除该特征。

第一部分可以用 TF(t)来表示，第二部分采用逆文本频率指数来表示，一个特征项 t 的逆文本频率指数 IDF(t)由样本总数与包含该特征项文档数决定：

$$\mathrm{IDF}(t) = \lg \frac{n}{n(t)} \tag{3-8}$$

第一部分和第二部分都满足取值越大时，该特征对类别区分能力越强，取两者乘积作为该特征项 TF-IDF 值：

$$\text{TF-IDF}(t) = \mathrm{TF}(t) \cdot \mathrm{IDF}(t) = n(t) \cdot \lg \frac{n}{n(t)} \tag{3-9}$$

一般停用词第一部分取值较高，而第二部分取值较低，因此 TF-IDF 等价于停用词和文档频率阈值法两者的综合。

3.3.4　信噪比

信噪比（Signal-to-Noise Ratio，SNR）源于信号处理领域，表示信号强度与背景噪音的

差值，如果将特征项作为一个信号来看待，那么特征项的信噪比可以作为该特征项对文本类别区分能力的体现。

信号背景噪声的计算，需要引入信息论中熵（Entropy）的概念，熵最初由克劳修斯在1864年提出并应用于热力学，1948年由香农引入信息论中，称为信息熵（Information Entropy）。其定义为：如果有一个系统 X，存在 c 个事件 $X=\{x_1,x_2,\cdots,x_c\}$，每个事件的概率分布为 $P=\{p_1,p_2,\cdots,p_c\}$，则第 i 个事件本身的信息量为 $-\lg(p_i)$，该系统的信息熵即为整个系统的平均信息量：

$$\text{Entropy}(X) = -\sum_{i=1}^{c} p_i \lg p_i \tag{3-10}$$

为方便计算，令 p_i 为0时，熵值为0（即0lg0），熵的取值范围是$[0,\lg c]$，当 X 以100%的概率取某个特定事件，其他事件概率为0时，熵取得最小值0；当各事件的概率分布越趋于相同时，熵的值越大；当所有事件趋于可能性发生时，熵取最大值 $\lg c$。根据熵的概念，定义特征项的噪声：

$$\text{Noise}(t) = -\sum_{j=1}^{n} P(d_j,t)\lg P(d_j,t) \tag{3-11}$$

式中，$P(d_j,t)=\dfrac{\text{TF}_{d_j}(t)}{\text{TF}(t)}$ 表示了特征项 t 出现在样本 d_j 中的可能性，特征项 t 的噪音函数取值范围为$[0,\lg n]$，当特征项 t 集中出现在单个样本内时，取得最小值0；当特征项 t 以等可能性出现在所有（n 个）样本中时，取得最大值 $\lg(n)$，这符合越集中在较少样本中，特征项为噪音可能性越小的直观认识，相应特征项 t 的信号值若用 $\lg\text{TF}(t)$ 来表示，可得信噪比计算公式：

$$\begin{aligned}\text{SNR}(t) &= \lg\text{TF}(t) - \text{Noise}(t)\\ &= \lg\text{TF}(t) + \sum_{j=1}^{n} P(d_j,t)\lg P(d_j,t)\end{aligned} \tag{3-12}$$

信噪比取值范围为$[0,\lg\text{TF}(t)]$，仅当特征项 t 在全部（n 个）样本中均出现1次时，取得最小值0，表明这种情况下当前特征项是一个完全的噪音，没有任何分类贡献能力；当特征项 t 集中出现在一个样本内时，取得最大值 $\lg\text{TF}(t)$。

计算信噪比时未考虑样本所属类别。当特征项只出现在较少样本时，信噪比较高，如果这些文本基本属于同一类别，则表明该特征项是一个有类别区分能力的特征；如果不满足这种分布情况，则特征项的信噪比取值较大时也不表明其有较好的类别区分能力。

3.4 特征重构

特征重构以特征项集合为输入，利用对特征项的组合或转换生成新的特征集合作为输出，一方面，特征重构要求输出的特征数量要远远少于输入的数量，以达到降维目的；另一方面，转换后的特征集合应当尽可能地保留原有类别区分能力，以实现有效分类，与特征子集选择相比较，特征重构生成的新特征项不要求对应原有的特征项，新特征项可以是由原来单个或多个特征项经某种映射关系转换而成的。这种转换规则需要保存下来，以便于对新

的样本也进行同样的转换，得到该样本所对应特征重构情况的表示形式。

特征重构有基于语义的方法，如词干与知识库方法；也有基于统计等的数学方法，如潜在语义索引。

3.4.1　词干

由于英文存在词形变化情况，词干方法(Stemming)在英文文本处理中应用较为广泛，从分类角度考查，这些变化对类别区分贡献较小，因此词干方法的目的是将变化的形式与其原形式合并为单个特征项，从而有效降低特征项维数，英文中这些变化通常表现为词的后缀部分的变化，因此实际常用的解决方式是采用简单保留词前面的主体部分(去除后缀)，这样处理可以得到比较理想的结果，M. F. Porter 早在 1979 年就提出一种算法，并一直在其主页(http://wwwtartarus.org/～ martin/PorterStemmer/)上进行维护，先后完善了多种编程语言的实现。他对各种不同的词干算法进行了综述，并在原先基础上继续研究，认为进行词干处理对系统性能提高有限。

当采用 n 元模型作为特征项时，应当在构建 n 元模型前进行词干处理。

3.4.2　知识库

词干方法从词形变化方面进行降维，而知识库(Thesaurus)方法则从词义角度进行降维。自然语言中存在同义词和近义词现象，知识库可以构建这种关系的表达，以将其聚合在一起，从而实现降维。通常，知识库可以表示为一些词及这些词之间的关系。常用的关系有同义、近义方面，或者包含范围大小方面等关系。通用领域内研究较早、应用较为广泛的知识库，有面向英文的 WordNet(http://wordnet.princeton.edu/)与面向中文的“知网”(http://www.keenage.com/)。

知识库的构建往往需要手工建设，还需要维护更新，以便于添加新的、去除过时或修正错误内容等，以及根据具体的应用设定相应的各种映射规则。需要消耗大量人力，限制了知识库方式的自动实现程度与使用范围。

近年来，一种多人协作的写作方式 Wiki 发展迅速，Wiki 站点可以由多人(甚至任何访问者)维护，每个人都可以发表自己的意见，或者对共同的主题进行扩展及探讨，Wiki 指一种超文本系统，这种超文本系统支持面向社群的协作式写作，同时包括一组支持这种写作的辅助工具，以 Wikipedia(http://zh.wikipedia.org/)为代表的 Wiki 网站，已经达到相当数量的信息积累，不仅在更新速度、信息容量方面比以往的个人维护或专家集体创作的百科全书有明显优势，而且在信息质量方面也经受了实践的检验与认可。利用 Wiki 来辅助自然语言处理及文本分类，也有相关研究，它是知识库方式的新形势，且有较大的实际意义。

3.4.3　潜在语义索引

20 世纪 80 年代 M. W. Berry 和 S. T. Dumais 提出了一种新的信息检索模型：潜在语义索引(Latent Semantic Indexing，LSI)模型。该模型对利用向量空间模型(Vector Space Model，VSM)表示文本时遇到的困难问题进行回答，很快在信息检索、信息过滤、特征降维

等领域获得广泛应用，并有多种 LSI/SVD 实现。

VSM 将一篇文本表示为向量空间中的一个向量，不仅比复杂的语义表示结构易于实现，而且适合作为信息检索，用于机器学习领域的输入形式。因此，它作为文本表示的基础模型而得以广泛应用。然而 SVM 模型认为，各特征项之间独立分布（不相关），这一要求在自然语言领域内往往无法得到保证。以词为例，各个词之间并不是毫无关系，而是关系极为复杂（简单的，如存在一词多义和多词同义、近义现象），从理论上来说，若能将多义词按照不同含义分为多个特征项，将多个同义词合并为一个特征项，对于信息过滤和文本分类等应用会产生正面影响，在实际应用中，并不容易正确区分各种同义和多义现象，而且对于更复杂的词之间的关系，也没有简单的一分为多或多合为一的直观解决方法。可以说，这些是知识库方法面临的另外一个实用性限制。

LSI 模型则以大规模的语料为基础，通过使用线性代数中对矩阵进行奇异值分解（Singular Value Decomposition，SVD）的方法，实现了一种词与词之间潜在语义的表示方式，同时，克服了手工构建知识库耗费大量人力物力以及难以表达显式关系等缺点。

矩阵进行奇异值分解过程：设 $\boldsymbol{A}$ 是秩为 r 的 $m\times n$ 矩阵，则存在 m 阶正交矩阵（正交矩阵是指转置矩阵为自身逆矩阵的方阵）$\boldsymbol{U}$ 和 n 阶正交矩阵 $\boldsymbol{V}$，使 $\boldsymbol{A}$ 可分解为 $\boldsymbol{A}=\boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}}$，其中 $\boldsymbol{V}^{\mathrm{T}}$ 表示短阵 $\boldsymbol{V}$ 的转置矩阵；Σ 为对角矩阵，$\Sigma=\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r,0,\cdots,0)$，且有 $\sigma_1\geqslant\sigma_i\geqslant\sigma_r$。$\sigma_i(i=1,2,\cdots,r)$ 为矩阵 $\boldsymbol{A}$ 的奇异值。$\boldsymbol{U}$，$\boldsymbol{V}$ 的列向量，分别称为 $\boldsymbol{A}$ 的左、右奇异向量。

SVD 分解可以用于求解原矩阵 $\boldsymbol{A}$ 的近似矩阵。方法是选择一个 k 值（$k<r$），Σ 只保留前 k 个比较大的奇异值组成新的对角阵 Σ_k（保留奇异值从大到小顺序），$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 只保留前 k 列，分别记为 $\boldsymbol{U}_k$，$\boldsymbol{V}_k$，则通过计算 $\boldsymbol{U}_k\Sigma_k\boldsymbol{V}_k^{\mathrm{T}}$ 得到 $\boldsymbol{A}$ 的近似矩阵 $\boldsymbol{A}_k$，如图 3-3 所示。

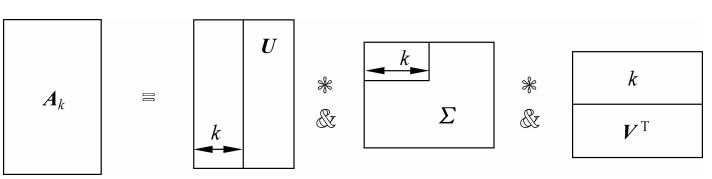

图 3-3 A_k 的计算示意图

新矩阵 $\boldsymbol{A}_k$ 是 $\boldsymbol{A}$ 的一个 k 秩近似矩阵，它在最小平方意义下最接近原矩阵，潜在语义索引认为 $\boldsymbol{A}_k$ 包含了 $\boldsymbol{A}$ 的主要结构信息，而忽略那些数值很小的奇异值，从而实现降维。对于文本分类问题来说，矩阵 $\boldsymbol{A}$ 表示特征项-样本矩阵，每一个列向量表示一个样本中各特征项的权重，行向量表示一个特征项在各文本中的权重，通过 SVD 分解，特征项-样本矩阵从 $\boldsymbol{A}$ 转换为 $\boldsymbol{A}_k$ 从而实现了降维，不仅去除了对分类影响很小的特征项，而且近似的特征项被合并。如同义词，在 k 维空间中有相似的表示，并且出现在相似文档中的特征项也是近似的，即使它们并未出现在同一个文档中，原向量空间模型中文档 $\boldsymbol{d}$ 经过 LSI 模型转换为 $\hat{\boldsymbol{d}}$，转换公式为

$$\hat{\boldsymbol{d}}=\boldsymbol{d}^{\mathrm{T}}\boldsymbol{U}_k\Sigma_k^{-1} \tag{3-13}$$

LSI 构造了特征项之间潜在的语义关系空间，下面以一个实例说明具体的计算过程，其训练数据来自 SIAM review 的一篇书评文章中的书名，如表 3-2 所示。

表 3-2　SIAM review 书评中所涉及书名

书编号	书　　名
B1	*A Course on IntegralEquations*
B2	*Attractors for Semigroups and Evolution Equations*
B3	*Automatic Differentiation of Algorithms: Theory, Implementation, and Application*
B4	*Geometrical Aspects of PartialDifferentialEquations*
B5	*Ideals, Varieties, and Algorithms-An Introduction to Computational Algebraic Geometry and Commutative Algebra*
B6	*Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*
B7	*Knapsack Problems: Algorithms and Computer Implementations*
B8	*Methods of Solving Singular Systems of OrdinaryDifferentialEquations*
B9	*Nonlinear Systems*
B10	*OrdinaryDifferentialEquations*
B11	*OscillationsTheory for Neutral DifferentialEquations with Delay*
B12	*OscillationsTheory of DelayDifferentialEquations*
B13	*Pseudodifferential Operations and NonlinearPartialDifferentialEquations*
B14	*Sinc Methods for Quadrature and DifferentialEquations*
B15	*Stability of Stochastic DifferentialEquations with Respect to Semi-Martingales*
B16	*The Boundary Integral Approach to Static and Dynamic Contact Problems*
B17	*The Double Mellin-Barnes Type Integrals and their Applications to Convolutions Theory*

其中有下画线的词，表明其至少在两本书的书名中出现过，去除只出现一次的低频词，组成特征项-文本矩阵，如表 3-3 所示。

表 3-3　16×17 维特征项-文本矩阵

特　征　词	文　本																
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17
Algorithms	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
Application	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Delay	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Differential	0	0	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
Equations	1	1	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
Implementation	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Integral	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Introduction	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
Methods	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
Nolinears	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
Odinary	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
Oscillation	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Partial	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
Problem	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
Systems	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0
Theory	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1

对表 3-3 所表示的特征项-文本矩阵进行奇异值分解，只保留最大的两个奇异值($k=2$)，得到 $\boldsymbol{U}_k$，Σ_k，为

$$\boldsymbol{U}_k=\begin{cases}0.0159 & -0.4317\\ 0.0266 & -0.3756\\ 0.1785 & -0.1692\\ 0.6014 & 0.1187\\ 0.6691 & 0.1209\\ 0.0148 & -0.3603\\ 0.0520 & 0.1120\\ 0.1503 & 0.1127\\ 0.0813 & 0.0672\\ 0.1503 & 0.1127\\ 0.1785 & -0.1692\\ 0.1415 & 0.0974\\ 0.0105 & -0.2363\\ 0.0952 & 0.0399\\ 0.2051 & -0.5448\end{cases}\quad \Sigma_k=\begin{pmatrix}4.431 & 40 & 0\\ 0 & 0.275 & 82\end{pmatrix}$$

以信息检索方面的应用为例，一个查询 $\boldsymbol{q}$ 为 Application Theory，对应原始向量空间模型为 $\boldsymbol{q}=[0\ 1\ 0\ 000\ 000\ 0\ 0\ 0\ 0\ 0\ 0\ 1]$，利用查询 $\boldsymbol{q}$ 从原来的 17 本书中查询相关书的问题可以转化为如下问题：即认为查询 $\boldsymbol{q}$ 也是一本书(或者说是书名，因为例子中以书名代表书的内容)，任务就转换为判断有哪些书和 $\boldsymbol{q}$ 比较近似。根据式(3-13)进行降维，结果为 $\hat{\boldsymbol{q}}=\boldsymbol{q}^{\mathrm{T}}\boldsymbol{U}_k\Sigma_k^{-1}=[0.0511,-0.3337]$。至此，就完成了 $\boldsymbol{q}$—$\hat{\boldsymbol{q}}$的降维过程，然后根据余弦相似度即可计算和各文档之间的相似程度。

LSI 模型有着良好的降维性能，对特征项之间的潜在关系有着优秀的表达能力，这是 LSI 的优点所在。LSI 模型也存在一些在应用时需要注意的不足之处，如转换结果不直观、矩阵分解运算量大、动态更新需重新运算等。随着 LSI 相关研究的深入，部分不足正逐渐得以解决，如奇异值分解的并行算法有助于实现更大规模的矩阵奇异值分解。

3.5 向量生成

上述特征项抽取及特征选择环节回答了文本表示的一个基本问题：选择适合作为表示文本的特征项集合；而向量生成(Vector Generation)环节回答了文本表示的另一个基本问题：给这些特征项赋予合适的权重，与向量生成相关的一些参数定义：设共有 m 项($t_1,\cdots,t_m$)特征，对给定样本 $\boldsymbol{d}$，由每一个特征出现的频率次数组成特征频率向量 $\boldsymbol{DT}_F=(\boldsymbol{TF}_d(t_1),\cdots,\boldsymbol{TF}_d(t_m))^{\mathrm{T}}$，其中 $\boldsymbol{TF}_d(t_i)$表示特征 t_i 在样本 $\boldsymbol{d}$ 中出现的次数，向量生成环节研究在此基础上的权重向量 $\boldsymbol{d}=(w(\boldsymbol{d},t_1),\cdots,w(d,t_m))^{\mathrm{T}}$。

Salton 认为，一个样本中某特征项的权重由局部系数、全局系数和正规化系数 3 部分组成。即

$$w(\boldsymbol{d},\boldsymbol{t})=\frac{w_1(\boldsymbol{d},\boldsymbol{t})w_g(\boldsymbol{t})}{w_n(\boldsymbol{d})}$$

3.5.1　局部系数

局部系数(Local Component)$w_l(\boldsymbol{d},\boldsymbol{t})$,表示特征 $\boldsymbol{t}$ 对当前样本 $\boldsymbol{d}$ 的直接影响,一般认为在样本 $\boldsymbol{d}$ 中一个特征 $\boldsymbol{t}$ 出现的次数越多,则 $\boldsymbol{t}$ 对 $\boldsymbol{d}$ 的影响越大,常用局部系数方式见表 3-4。

表 3-4　常用局部系数

简记	计 算 方 法	说　明
n	$\boldsymbol{w}_l(\boldsymbol{d},\boldsymbol{t})=\boldsymbol{TF}_d(\boldsymbol{t})$	n 表示无转换(No Conversion)
b	$\boldsymbol{w}_l(\boldsymbol{d},\boldsymbol{t})=\begin{cases}1,\boldsymbol{TF}_d(\boldsymbol{t})>0\\0\end{cases}$	二进制值表示(Binary Term Indicator)
m	$\boldsymbol{w}_l(\boldsymbol{d},\boldsymbol{t})=\dfrac{\boldsymbol{TF}_d(\boldsymbol{t})}{\boldsymbol{TF}_d(t_{\max})}$	$t_{\max}$ 表示样本 $\boldsymbol{d}$ 中单个特征出现最多的次数
a	$\boldsymbol{w}_l(\boldsymbol{d},\boldsymbol{t})=\dfrac{1}{2}+\dfrac{1}{2}\dfrac{\boldsymbol{TF}_d(\boldsymbol{t})}{\boldsymbol{TF}_d(t_{\max})}$	增大(Augment)m 方式结果,m 方式的变形,由[0,1]至[0.5,1]
l	$\boldsymbol{w}_l(\boldsymbol{d},\boldsymbol{t})=\begin{cases}1+\lg\boldsymbol{TF}_d(\boldsymbol{t}),\boldsymbol{TF}_d(\boldsymbol{t})>0\\0\end{cases}$	对数(Logarithm)运算

3.5.2　全局系数

全局系数(Global Component)$\boldsymbol{w}_g(\boldsymbol{t})$考虑特征 $\boldsymbol{t}$ 在整个训练样本中的重要性,包含特征 $\boldsymbol{t}$ 的文档数较少时,特征 $\boldsymbol{t}$ 比较有分类区分能力,应给予较大权重。常用全局系数方式见表 3-5。

表 3-5　常用全局系数

简记	计 算 方 法	说　明
t	$\boldsymbol{w}_g(\boldsymbol{t})=\log\dfrac{\boldsymbol{n}}{\boldsymbol{n}(\boldsymbol{t})}$	即 TF-IDF 中 IDF
p	$\boldsymbol{w}_g(\boldsymbol{t})=\log\dfrac{\bar{\boldsymbol{n}}}{\boldsymbol{n}(\boldsymbol{t})}$	$\bar{\boldsymbol{n}}=\boldsymbol{n}-\boldsymbol{n}(\boldsymbol{t})$,$t$ 方式的变形
n	$\boldsymbol{w}_g(\boldsymbol{t})=1$	不考虑全局因素

3.5.3　规范化系数

规范化系数(Normalization Component)用于调节权重的取值范围,一种常见的方式是将所有的权重向量的取值范围映射到[0,1]区间。常用规范化系数方式见表 3-6。

表 3-6　常用规范化系数

简记	计 算 方 法	说　明
n	$\boldsymbol{w}_n(\boldsymbol{d})=1$	不考虑规范化系数
s	$\boldsymbol{w}_n(\boldsymbol{d})=\sum_{i=1}^{m}\boldsymbol{w}_1(\boldsymbol{d},t_i)\boldsymbol{w}_g(\boldsymbol{d},t_i)$	单个样本的所有权重之和调节为 1
c	$\boldsymbol{w}_n(\boldsymbol{d})=\sqrt{\sum_{i=1}^{m}(\boldsymbol{w}_1(\boldsymbol{d},t_i),\boldsymbol{w}_g(\boldsymbol{d},t_i))^2}$	单个样本所有权重的平均和为 1

3.6 文本内容分析

虽然可以不断提高文本表示模型的效率，但每个文本都是由大量的特征所组成的这一事实导致文本表示维数会达到数十万维的大小，对将要进行的文本内容分析可能带来灾难性的计算时间指数增长，而产生的特征子集分类结果与小得多的特征子集相近。因此，减少文本特征的维数至关重要。本节分别从语法、语义和语用三个方面进行文本内容分析，为展开文本内容安全应用研究打好基础。

3.6.1 文本语法分析方法

文本语法分析(Text Grammar Analysis)是指通过语言模型或语法模型来处理文本的过程，包括隐马尔科夫(Hidden Markov Model，HMM)词性标注、最大熵(Maximum Entropy，ME)命名实体识别和 N 元语法模型(N-gram)等。

1. HMM 模型词性标注

当马尔科夫模型中的状态对于外界来说不可见的时候，就转换成隐马尔科夫模型(HMM)。一般来说，HMM 是一种随机模型，适合非常随机序列，具有统计特性，可以用于处理多个不同平稳状态过程中的随机转移。HMM 是一个双重随机过程，其中的一重随机过程是描述基本的状态转移，而另一重随机过程是描述状态与观察之间的对应关系。HMM 适合序列标注问题，即给定一个观察序列 $X=\{x_1,x_2,\cdots,x_m\}$，求出最适合这个观察序列的标记序列 $Y=\{y_1,y_2,\cdots,y_m\}$，使得条件概率 $p(Y|X)$最大。HMM 中，条件概率通过贝叶斯原理变换后求得

$$p(X \mid Y)=\frac{p(Y)p(X \mid Y)}{\sum_{Y}p(Y)p(X \mid Y)} \tag{3-14}$$

在序列标注任务中，X 是一个给定的观察序列，式(3-14)中的分母对所有的 X 相同，因此可以不予考虑，同时应用联合公式可得

$$Y^{*}=\underset{Y}{\operatorname{argmax}}\,p(Y \mid X)=\underset{Y}{\operatorname{argmax}}\frac{p(X)p(Y \mid X)}{p(X)}=\underset{Y}{\operatorname{argmax}}\,p(X,Y) \tag{3-15}$$

即隐马尔科夫模型实质上是求解一个联合概率。式(3-15)中编辑序列 Y 即可作为一个马尔科夫链，进一步对式(3-15)应用乘法公式：

$$\begin{aligned}p(x_{1,m},y_{1,m})&=\prod_{i=1}^{m}p((x_i,y_i) \mid x_{1,i-1},y_{1,i-1})\\&=\prod_{i=1}^{m}p(x_i \mid x_{1,i-1},y_{1,i})p(y_i \mid x_{1,i-1},y_{1,i-1})\end{aligned} \tag{3-16}$$

式(3-16)中，$x_{1,i}=x_1,x_2,\cdots,x_i$，$y_{1,i}=y_1,y_2,\cdots,y_i$，$1\leqslant i\leqslant m$。式(3-16)给出了不作任何假设的理想化的序列标注的概率模型。序列标注的任务便是寻找一个最佳的标注序列$\hat{Y}$，使得式(3-16)最大，即

$$\hat{Y}=\underset{Y}{\operatorname{argmax}}\,p(Y \mid X)$$

$$= \underset{Y}{\operatorname{argmax}} \prod_{i=1}^{m} p(x_i \mid x_{1,i-1}, y_{1,i}) p(y_i \mid x_{1,i-1}, y_{1,i-1}) \tag{3-17}$$

式(3-17)虽然反映了理想状况下标注序列的模型，但是在求解该模型时需要估计的参数空间太大，无法完成操作。为此，隐马尔科夫模型作如下假设。

假设一：标注的 y_i 出现只和有限的前 $N-1$ 个标记相关，即 n-pos 模型：

$$p(y_i \mid x_{1,i-1}, y_{1,i-1}) \approx p(y_i \mid y_{1,i-1}) \approx p(y_i \mid y_{i-N+1}, y_{i-N+2}, \cdots, y_{i-1}) \tag{3-18}$$

如果 $N=2$，则是常用的一阶隐马尔科夫模型。

假设二：一个观察值 x_i 的出现不依赖于前面的任何观察值，只依赖于前面的标记，并进一步假设只和该观察值的标记 y_i 相关，即

$$p(x_i \mid x_{1,i-1}, y_{1,i}) \approx p(x_i \mid y_{1,i}) \approx p(x_i \mid y_i) \tag{3-19}$$

由式(3-18)和式(3-19)可以将一阶隐马尔科夫模型式(3-17)重写如下：

$$p(Y \mid X) = \prod_{i=1}^{m} p(y_i \mid y_{i-1}) p(x_i \mid y_i) \tag{3-20}$$

其中，$p(x_i \mid y_i)$被称为发射概率，$p(y_i \mid y_{i-1})$被称为转移概率。

隐马尔科夫模型有 3 个基本问题：

(1) 估值问题。假设已有一个 HMM，其转移概率和发射概率均已知。如何计算该模型产生某一个观测序列的概率。

(2) 解码问题。假设有一个 HMM 和它所产生的一个观察序列，决定最有可能产生这个观测序列的隐状态序列。

(3) 学习问题。怎样调整现有的模型参数，使其描述给定观察序列最佳，即使得给定观察序列概率最大。

对于以上 3 个问题的行为，衍生出了 5 个算法。这 5 个算法都是动态规划算法。在实际使用 HMM 模型的时候，模型的转移概率和发射概率的估计方式通常有两种：无指导的 Baum-Welch 重估算法(即 Forward-Backward 算法)和有指导的极大似然估计方法(MLE)。对于 HMM 进行序列标记而言，最后为了字节最好的一个标记序列，需要对所有可能的路径寻优，即解码。常用的解码方法是 Viterbi 算法。

2. ME 模型

最大熵(ME)模型是通过求解一个有条件约束的最优化问题来得到概率分布的表达式。假设现有 n 个学习样本$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$，其中 x_i 是由 k 个属性特征构成的样本向量 $x_i = \{x_{i1}, x_{i2}, \cdots, x_{ik}\}$，$y_i$ 是类别标记 $y_i \in Y$。所要求解的问题是：在给定一个样本 x 的情况下，其最佳的类别标记是什么。

最大熵的目标函数被定义如下：

$$H(p) = -\sum \tilde{p}(x) p(y \mid x) \log p(y \mid x) \tag{3-21}$$

式(3-21)即为条件熵，也就是说最大熵模型要求信息系统的目标状态的条件熵取得最大值，同时要求满足下述两个条件：

$$P = \{p \mid E_p f_i = E_{\tilde{p}} f_i, 1 \leqslant i \leqslant k\} \tag{3-22}$$

$$\sum_y p(y \mid x) = 1 \tag{3-23}$$

式中 f_i 是定义在样本集上的特征函数，$E_p f_i$ 表示特征 f_i 在模型中的期望值，$E_{\tilde{p}} f_i$ 表示特

征 f_i 在训练集上的经验期望值。两种期望分别定义如下：

$$\begin{cases} E_p f_i = \sum_{C,h} \tilde{p}(x) p(y \mid x) f_i(y,x) \\ E_p f_i = \sum_{C,h} \tilde{p}(y,x) f_i(y,x) = \frac{1}{N} \sum f_i(y,x) \end{cases} \tag{3-24}$$

$$f_i(y,x) = \begin{cases} 1 & \text{if } y = y' \text{ and } h(x) = \text{TRUE} \\ 0 & \text{else} \end{cases} \tag{3-25}$$

其中 $h(x)$为谓词函数，其类型的个数和系统特征模板的类型个数相等。通过对式(3-21)、式(3-22)和式(3-23)进行拉格朗日变换，求出满足条件极值的概率：

$$p(y \mid x) = \frac{1}{z(x)} \exp\left(\sum_i \lambda_i f_i(y,x) \right) \tag{3-26}$$

$$z(x) = \sum_c \exp\left(\sum_i \lambda_i f_i(y,x) \right) \tag{3-27}$$

λ_i 是特征 f_i 对应的拉格朗日系数，只能通过数值计算方法求得。在最大熵模型中，最多被使用的参数估计是 GIS (Generalize Iterative Scaling)算法，在实践中，为了计算方便，需要把指数形式变换为对数形式，所以最大熵模型也是对数线性模型的一种。

最大熵模型本身是分类模型，在解决序列标注问题时，需要辅以一定的搜索策略。最大的序列标注方法可采用顺序标注，即假设标注序列$\{t_1, t_2, \cdots, t_n\}$，则在利用分类方法标注 t_1 后，顺序标记 $t_2, t_3, \cdots, t_n$。然而这种标注方法往往没有考虑 t_{i+1}的变化对于 t_1 的影响。实质上，对于序列标注，若能考虑标注序列内部标记的影响，往往能够获得更好的标注效果。给定一个句子，包含 n 个词，分别为$\{w_1, w_2, \cdots, w_n\}$，一个对应的标注序列$\{t_1, t_2, \cdots, t_n\}$的条件概率为

$$p(t_1, \cdots, t_n \mid w_1, \cdots, w_n) = \prod_{i=1}^{n} p(t_i \mid h_i) \tag{3-28}$$

其中 h_i 是第 i 个词 w_i 所对应的上下文环境。从式(3-28)可以看出，处理序列标注问题，可以枚举出对应句子的所有标注序列的候选，并且将输出的概率值最大的一个标注序列作为答案。常见的搜索算法主要有 Viterbi 算法，另外就是 Beam Search 算法。Beam Search 算法其实质是一个宽度优先搜索(Breadth First Search)。为了避免搜索过程中的组合爆炸问题，对每一步后续的所有候选中，只有前 K 个最优的候选进行扩展，其他的通过剪枝处理掉。

3. N-gram 模型

N-gram 模型是目前各种统计计算方法中应用最普遍且效果最好的基于离散 Markov 的模型。n 取 2 和 3 时分别叫 Bi-Gram 和 Tri-Gram。N-Gram 统计计算语言模型的思想是：一个单词的出现与其上下文环境(Context)中出现的单词序列密切相关，第 n 个词的出现只与前面 $n-1$ 个词相关，而与其他任何词都不相关，设 $W_1 W_2 \cdots W_n$ 是长度为 n 的字串，则字串 W 的似然度用方程表示如下：

$$p(W) = p(W_i \mid W_{i-n+1} W_{i-n+2} \cdots W_{i-1}) \tag{3-29}$$

式(3-29)表明，在 N-Gram 中，每一个词出现的概率仅仅与前面 $n-1$ 个最近词有关，根据离散 Markov 模型的定义可知，它相当于 $n-1$ 阶 Markov 模型。当 $p(W)$的值超过一定的阈值时，表明这 n 个字的结合能力强，可以认为它们是一个词。

根据大数定理，可以通过统计大量训练（学习）样本中字串 $W_{i-n+1}W_{i-n+2}\cdots W_{i-1}W_i$ 的出现次数 $f(W_{i-n+1}W_{i-n+2}\cdots W_{i-1}W_i)$来计算。

$$p(W_i \mid W_{i-n+1}W_{i-n+2}\cdots W_{i-1}) \approx \frac{f(W_{i-n+1}W_{i-n+2}\cdots W_{i-1}W_i)}{\sum_{W_i} f(W_{i-n+1}W_{i-n+2}\cdots W_{i-1}W_i)} \tag{3-30}$$

不难看出，为了预测词 W_n 的出现概率，必须知道它的前面所有词的出现概率。从计算上来看，这种方法太复杂了。如果任一词 W_i 的出现概率只同它前面的两个词有关，问题就可以得到极大的简化。这时的语言模型叫作 Tri-gram 模型。

$$p(W) \approx p(W_1)p(W_2 \mid W_1)\prod_{i=3,\cdots,n} p(W_i \mid W_{i-2}W_{i-1}) \tag{3-31}$$

符号概率 $\prod_{i=3,\cdots,n} p(W_i \mid W_{i-2}W_{i-1})$ 表示连乘。一般来说，N 元模型就是假设当前词的出现概率只与同它前面的 $N-1$ 个词有关。重要的是，这些概率参数都是可以通过大规模语料库来计算的。比如 3 元概率有

$$p(W_i \mid W_{i-2}W_{i-1}) \approx \frac{\text{count}(W_{i-2}W_{i-1}W_i)}{\text{count}(W_{i-2}W_{i-1})} \tag{3-32}$$

式中 count()是词频函数，表示一个特定词在整个语料库中出现的统计次数。

统计语言模型有点像天气预报中使用的概率方法，用来估计概率参数的大规模语料库好比一个地区历年积累起来的气象记录。例如，用 3 元模型来进行天气预报，就如同是根据前两天的天气情况来预测当天的天气情况。天气预报虽然没有做到百分之百准确，但是其高效的预测已经成为实用的生活助手。因此，采用 3 元统计模型实现词频统计是一种常用的方法。

3.6.2　文本语义分析方法

文本语义分析（Text Semantic Analysis）是将句子转化为某种可以表达句子意义的形式化表示，即将人类能够理解的自然语言转化为计算机能够理解的形式语言，做到人与机器相互沟通。语义分析解决的是句中的词、短语直至整个句子的语义的问题，通过语义分析找出语义、结构意义及其结合意义，从而确定语言所表达的真正含义或概念。语义分析方法包括词义消歧、信息抽取和感情倾向性分析内容。

1. 词义消歧

词义消歧（Word Sense Disambiguation）是对多义词根据上下文给出它所对应的语义编码，该编码可以是词典释义文本中该词所对应的某个义项号，也可以是义类词典中相应的义类编码。词义消歧在自然语言处理的许多方面都有很重要的用途。汉语多义词（歧义词）在词典中只占总词语量的 10％左右，大约有 8000 个多义词。目前词义消歧的主要对象是多义实词，主要是名词、动词、形容词三大类，其中，动词在实词词义消歧中占有特殊地位。

利用机器学习理论进行词义消歧的方法可以分为两种：有指导方法和无指导方法。这种划分的依据基于该方法是否利用了手工标注语料。有指导的词义消歧模型需要事先对训练语料进行歧义标注，而无指导的方法没有此要求。在有指导词义消歧方面，刘亚涛等人提出了一种基于义原同现有频率的汉语词义无指导消歧方法。

1）有指导的词义消歧

词义消歧需要根据上下文语境来确定正确的词义，这是一个典型的分类问题。设词条 w 有 n 个词义 $\{S_1, S_2, \cdots, S_n\}$，上下文语境为 C，词义消歧的任务就是根据上下文 C 来确定正确的词义 S'：

$$S' = \text{argmax}\ P(S_i/C) \tag{3-33}$$

因此在现有指导的词义消歧中，很多机器学习方法用于其中，如贝叶斯分类器、决策树和决策表算法、最大熵模型以及支持向量机等。特征选择也是对有指导的词义消歧中的重要步骤，特征选择就是在一定的上下文语境 C 中选择最有效的消歧特征。词义消歧研究中用到的上下文特征主要是以下 4 个层面：话题、词汇、句法和语义。

话题层面的消歧特征主要是用于一定上下文中的词来表示，即词袋（Bag of Words，BOW）。词汇层面的消歧特征主要有局部词（LW）、局部词性（POS）、局部（CON）等。话题层面和词汇层面的消歧特征来自于句子的表层信息，只需要进行基本的词语切分和词性标注即可方便地获得，而且也可以得到较高的消歧准确率，可称为词义消歧的基本特征。有指导词义消歧的研究中一般都要使用这两类特征，只有在具体运用时会稍有变化，例如词袋是否包括虚词等。

句法层面的消歧特征主要是句法结构信息。词义消歧常用的句法信息包括：是否带有主语、主语的中心词；是否带有宾语、宾语的短语类、宾语的中心词；是否带有 VP 类补语；是否在句法关系的基础上加上了语义类信息。有研究表明，将人工标注的语义角色（Semantic Role）用于词义消歧时，消歧准确率在句法特征的基础上又提高了约 3%。句法特征和语义特征确实可以提高词义消歧准确率，但需要付出的前期劳动却是巨大的。句法特征的获取需要一个高效的句法分析器，语义特征的获取需要一个高效的稳定语义角色标注器。另外，高效的句法分析器和语义角色标注器一定程度上又依赖于高效的词义标注器。

2）无指导的词义消歧

为解决消歧知识获取瓶颈的问题，无指导的词义消歧方法需要从无人工标注的资源中挖掘可用于词义消歧的信息。那么，具体需要什么信息？这些信息从哪里来？如何才能得到这些信息？这些都是无指导方法必须要考虑的问题。

从词义消歧任务的实际效果来看，无指导方法的性能较有指导及半指导方法的性能要差。但是由于其无须人工标注的训练语料，在性能提高到一定程度的时候却更有希望能够进行大规模应用。

无指导方法所获得知识的来源大体有：单语料库、双（多）语料库、词典以及 Web 等。目前无指导方法已经逐渐体现出多种知识源合用的趋势，特征是单独利用词典的无指导方法已经不多见。无指导的消歧方法依据所用资源大致可以分为 4 种：自动聚类词义辨析的方法、自动获取标记语料的方法、双语料法及基于 Web 的方法。从各类无指导词义消歧方法的分析中可以发现，由于首要问题是如何从含“隐性知识”的知识源中得到“显性知识”，而后再针对“显性知识”进行利用，因此，该类方法最关键的问题是知识获取及利用方法。

3）词义消歧算法

一般认为，词语的不同意义在句法组合上会显现差异，当今的词汇语义研究主要根据词语的句法分布来分析词义。本小节采用《现代汉语语法信息词典》进行词义消歧，该词典以复杂特征集为形式手段、以词类为纲，描述了词语不同意义的组合特征。例如，动词“保管”

的属性特征描述如表 3-7 所示。

表 3-7　《现代汉语语法信息词典》中“保管”的属性特征描述

词语	同形	释义	体谓准	动趋	动介	着了过	重叠	aabb	备注
保管	①	保藏,管理	体	趋	在	着了过	ABAB		～粮食
保管	②	担保,有把握	谓						～甜

“词语、同形、体谓准……”都是属性名(Attribute),“保管、①、谓……”是相对应的属性值(Value)。表 3-7 清晰地展示出了“保管①”和“保管②”在句法组合上的差异,借此差异可正确分辨出同形。例如下面的句子:

这份资料你先保管着,下午再交。

“保管①”的属性“着了过=着了过”,“保管②”的属性“着了过=否”,由此可判定例句中是保管①。对于一个词条的多个同形条目,同一个属性字段相异的取值即构成同形词之间的区别特征(Distinguish Features)。例如,对于“保管”,“着了过=着了过”构成“保管①”区别于“保管②”的一个属性特征,“体谓准=谓”构成“保管②”区别“保管①”的一个属性特征。词语 W 可区分为 n 个同形 $S_1,S_2,\cdots,S_n(n>1)$,同形 S_i 用复杂特征集来描述:

$$S_i\begin{bmatrix} f_1 = v_1 \\ f_2 = v_2 \\ \vdots \\ f_m = v_m \end{bmatrix}(m \geqslant 1) \tag{3-34}$$

词语 W 的不同同形 S_i,S_j 存在相同的属性特征 f_k,设 $S_i(f_k=v_{ki})$,$s_j(f_k=v_{kj})$,若 $v_{ki}\neq v_{kj}$,则称 $f_k=v_{ki}$ 是对 S_i 的区别特征,对应的 $f_k=v_{ki}$ 是 S_j 对 S_i 的区别特征。

基于词条语法属性的词义消歧的基本思路是:检查待消歧的目标多义词所在的上下文是否满足字典中特定同形的属性特征约束,若满足,则确定为该同形的意义。上下文语境是词义消歧的知识来源,语境范围的选取会影响到消歧的效率。本小节以多义词所在句子作为上下文语境范围,词义消歧算法描述如图 3-4 所示。

算法 WSD:词义消歧算法

输入:待消歧的词条

输出:消歧后的词条

① 依据《现代汉语语法信息词典》,对每一个多义词 W,比较不同同形的属性特征进而找出相互之间的肯定性区别特征,对每一个同形 S_i,以 $f_k=v_{ki}$ 的形式列出其肯定区别特征,对每一个多义词 W 生成一个属性特征文件 W_Lex_Rule(如上文“保管.txt”);

② 定位目标多义词 W,以句子范围作为上下文语境 C;

③ 对 W 的不同同形赋值 $S_i\cdot$ Score=0;

④ 检索文件 W_Lex_Rule,提取同形 S_i 的肯定性区别特征,判断 W 所在的上下文 C 是否满足约束条件,若满足,则 $S_i\cdot$ Score=$S_i\cdot$ Score+1;

⑤ 若文件 W_Lex_Rule 中属性特征列表非空,则重复④;

⑥ Score 取最大的同形 S_i 为标注结果。

图 3-4　词义消歧算法 WSD

2. 信息抽取

信息抽取(Information Extraction,IE)最早是在 Frump 系统背景下提出的,后来得到了美国政府资助的 MUC(Message Understanding Conference)系列会议的支持。

信息抽取是自然语言处理领域的重要研究方向之一,其研究内容包括实体识别(Named Entity Recognition,NER)、术语自动识别(Term Extraction Automatically,TEA)和关系抽取。命名实体识别包括中国姓名、中国地名、组织机构、英译名的自动辨识,即是通常说的未登录词的自动辨识问题。胡文敏等提出了一种基于卡方检验的汉语术语抽取方法:先从网络上下载语料,然后使用改进的互信息参数抽取结构简单的合串,并在此基础上进一步使用卡方检验结合子串分解方法抽取具有复杂结构的合串。AIRS 2008 会议上介绍了一种上下位关系(hyponymy 或 IS-A)自动获取的方法。该方法基于两个假设:一是相同的术语类型具有相似的上下文;二是两个术语如果具有上下位关系,则可被相似属性的名词和领域动词所描述。

信息抽取有两个特点:一是想获得的知识可以通过相对简单和固定的模板或带有槽的框架来进行描述;二是文本中只有一小部分信息需要填入模板或框架,其他的都可以被忽略。最简单的信息抽取是实体抽取,没有框架,只有实体类型。

图 3-5 给出了信息抽取过程的示意图。其中,信息抽取引擎的输入是一组文本,引擎通过使用一个统计模块、一个规则模块或者两个的混合进行信息抽取。IE 引擎的输出是一组从文本中抽取的标注过的框架,即填好的一张表。目前,从文本中可以抽取到以下 4 种基本类型的元素:

(1) 实体。实体是文本中的基本构成模块,如人、公司、地址等。

(2) 属性。属性是所抽取实体的特征,如人的年龄、头衔、组织的类型。

(3) 关系。实体之间存在的联系即为事实,如公司与员工之间的雇佣关系、两个公司之间的关联关系等。

(4) 事件。事件是实体的行为或实体因为兴趣而参加的活动,如参加一次有组织的旅游、两个公司之间的合并、一次突发意外等。

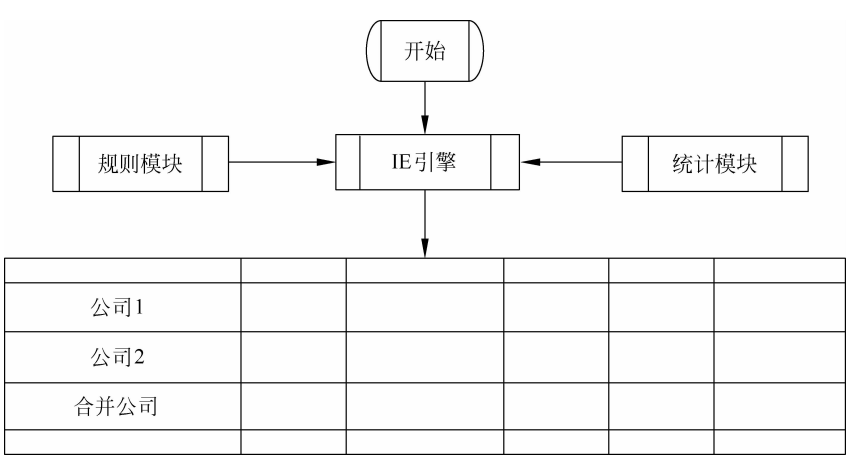

图 3-5 信息抽取过程示意图

3. 情感倾向性分析

文本情感倾向性分析，就是对一篇文章进行情感色彩判断。具体来说，就是对说话人的态度（或称观点、情感）进行分析，即对文本中的主观性信息进行分析。由于立场、出发点、个人状况和偏好的不同，民众对生活中各种对象和事件所表达出的信念、态度、意见和情绪的倾向性必然存在很大的差异。在论坛、博客等网络媒体上，这种差异表现得尤为明显。

文本倾向性分析近年来已经成为自然语言处理中的一个热点问题。文本所蕴含的情感（Emotion）和观点（Opinion）皆是人物主观意愿的反映，情感表达人物自身的情绪起伏，如快乐、悲伤等；观点则表达人物对外界事物的态度，如赞成、反对等。其中，对于文本情感的研究正得到越来越多研究者的关注。在 ACL、SIGIR 等国际会议上，针对这一问题的文章已开始出现；而对于文本观点倾向性的研究，国外早已开展得如火如荼，这类文章在 WWW、CIKM、SIGHAN 等顶级会议上层出不穷；针对倾向性分析的国际评测也已经开展，例如 TREC Blog Track 以及 NTCIR 等。

识别出网页文本中的倾向性语言是正确开展网络舆情倾向性判断、屏蔽不良网页、维护网络安全的关键工作之一。本小节介绍网页情感倾向性分析的具体过程。该方法从中文网络舆情采集入手，借助中科院中文分词软件 ICTCLAS 完成中文分词，充分考虑网络舆情信息表达的复杂性与共享性，把网络舆情倾向性分析模块分解为词语情感倾向性分析、句子情感倾向性分析和篇章情感倾向性研究 3 个子模块，如图 3-6 所示。

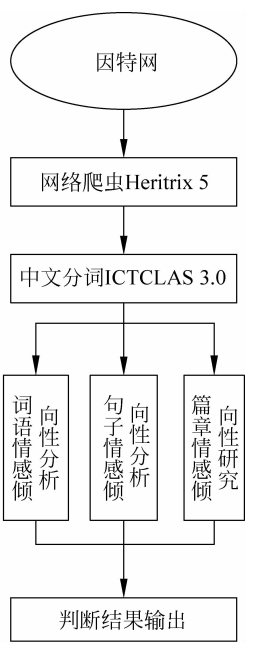

图 3-6　网络舆情情感倾向性分析模块结构

1）词语情感倾向性分析子模块

词语情感倾向性研究是倾向性研究工作的前提。具有情感倾向的词语以名词、动词、形容词和副词为主，也包括人名、机构名、产品名、事件名等命名实体。其中，除部分词语的褒贬性（或称为极性，通常分为褒义、贬义和中性 3 种）可以通过查词典[①]的方式得到之外，其余词语都无法直接获得。

词语情感倾向性分析包括对词语极性、强度（如“谴责”强度远超过“批评”）和上下文本模式的分析，分析甚至可以写入词典中。词语情感计算的方法有关键词测定（Keyword Spotting）、词汇类同（Lexical Affinity）、统计方法（Statistical Methods）、手工制作模式（Hand Craft Models）等。具体实现可归纳为以下三种。

（1）由已有的电子词典或词语知识库扩展生成情感倾向词典。如英文词语情感倾向词典 WordNet、中文词语情感倾向词典 HowNet。这种方法的种子词数量的依赖比较明显。

（2）无监督机器学习方法。这种方法以词语在语料库中的词频同现情况判断其联系紧密程度，与第（1）种方法相比，这种方法的噪声比较大。

① http://www.keenage.com

(3) 基于人工标注语料库的学习方法。首先对情感倾向分析语料库进行手工标注。标注的级别包括文档集的标注(即只判断文档的情感倾向性)、短语级标注和分句级标注。在这些语料的基础上,利用词语的共现关系、搭配关系或者语义关系,以判断词语的情感倾向性。这种方法需要大量的人工标注语料库。

2) 句子情感倾向性分析子模块

句子情感倾向性分析的处理对象是在特定上下文中出现的语句。其任务是对句子中的各种主观性信息进行分析和提取,包括对句子情感倾向性的判断,以及从中提取出与情感倾向性论述相关联的各个要素,包括情感倾向性论述的持有者、评价对象、倾向极性、强度,甚至是论述本身的重要性等。

通过对网络一些文章的分析提取,得到以下 16 个句子结构作为句子结构分析的模板库,参见表 3-8。

表 3-8 句子结构分析模板库

评价对象/s. +形容词/a. /名词/n.
评价对象/s. +副词/adv. +形容词/a. /动词/v.
评价对象/s. +副词/adv. +动词/v.
评价对象/s. +形容词/a. /动词/v. +转折连词/副词/adv. +形容词/a. /动词/v. 动词/v. +评价对象/s.
副词/adv. +动词/v. +评价对象/s.
评价对象/s. +否定词/d. +形容词/a. /名词/n.
评价对象/s. +否定词/d. +副词/adv. +形容词/a. /名词/n.
评价对象/s. +否定词/d. +副词/adv. +动词/v.
评价对象/s. +形容词/a. /动词/v. +转折连词/c. /副词/adv. +形容词/a. /动词/v.
否定词/d. +动词/v. +评价对象/s.
否定词/d. +副词/adv. +动词/v. +评价对象/s.
评价对象/s. +'是'动词/vs. +形容词/a. /名词/n.
评价对象/s. +副词/adv. +动词/v. +形容词/a. /名词/n.
评价对象/s. +否定词/d. +'是'动词/vs. +形容词/a. /名词/n.
评价对象/s. +否定词/d. +副词/adv. +动词/v. +形容词/a. /名词/n.

依据概率树分析后,为每种句式设置一种算法,并依照情感词进行初步的句子倾向性的判断。句子倾向性分析的步骤如下。

一是通过情感词库(含褒义词词库、贬义词词库)中的情感词定位含有情感词的句子,通过分词结果的词性调用,得到句子的情感程度。

二是初步情感判断完成以后,进行精细的分级程度判断,并依此为结果,得出句子的最终倾向值,具体实现步骤如下。

第一遍扫描序列,找到所有程度副词(类别为 2),将其程度值乘到模板中离其最近的一个 1 类词的程度值上(考虑到副词可能位于其中心词的前面或者后面,所以这里的“最近”是前后双向的查找,同时由于副词在前的情况比较多,所以前向查找的优先级高)。具体的处理是标注程度为 3 的因子为 1.5,程度为 2 的因子为 1,程度为 1 的因子为 0.5。

第二遍扫描序列,找到所有否定词(类别为 3),将其往后碰到的第一个 1 类词的褒贬性取反。

第三遍扫描序列,以转折词为单位将序列分成几个小部分,对每个小部分累加其 1 类词

的褒贬倾向值，然后按转折词类型的不同乘以转折词相应的权值（让步型如“虽然”，对位部分要减弱，因子为 0.7；转折型如“但是”，对应部分要加强，因子为 1.3）。

3）篇章情感倾向性研究子模块

如果说句子是点，篇章则是线。该模块的主要功能就是从整体上判断某个文本的情感倾向性，即褒贬态度。将篇章作为一个整体笼统地进行主观性分析，存在很大的局限性，其本质缺陷在于假设整体文本是针对同一个对象进行评论。而真实文本往往由包含多个对象，不同对象所涉及的观点、态度等主观性信息是有差异的。从另一面看，篇章内的对象总数仍是有限的，不足以支撑对于整体倾向性的处理。因此，本模块研究以篇章内情感倾向性论述的分析以及在大规模数据集上进行整体倾向性分析为主要研究内容。

设定一定的阈值，并对含有情感的句子值综合相加，得出篇章的情感色彩，完成文本倾向性分析。根据得出的网页文本情感阈与设定的阈值相比较的结果，将网页分为 4 级：恶性网页、消极网页、中性网页和积极网页，如图 3-7 所示。篇章情感倾向性分析算法如图 3-8 所示。

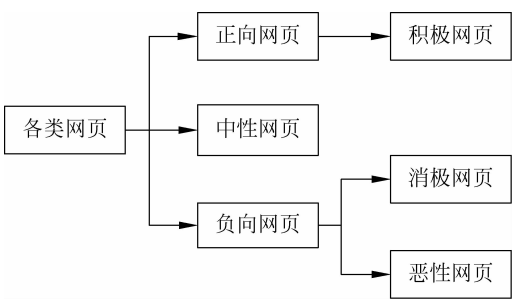

图 3-7　网页情感倾向性分类

```
Input:一篇待计算情感的文本/网页
Output:该文本/网页经计算后的情感结果(积极/消极/恶意)
for (int nc = 0;nc < ncount;nc++)
{
CString getpos (result[nc].sPOS) ;//得到文本全体词的词性
//wj 句号,全角:.半角:. ww 问号,全角:?半角:?
//wt 叹号,全角:!半角:! Ws 省略号,全角:…半角:…
if ( getpos == "wj"|| getpos == "wt"||getpos == "ww"||getpos == "ws")
    {
        finish = nc;
        CSentence cen ( result,start,finish,readtext) ;
        //调用 CSentence 中的函数
        //寻找句中第(int) ( ends - start ) /2 个词
        float g = cen.getpolarity (( int ) ( ends - start ) /2 );
        showresult = showresult + cen.MessageReturn;
        polaritysum += g;
        start = finish + 1;
        AllSentence.push_back ( cen );
    }
}
```

图 3-8　篇章情感倾向性分析算法

3.6.3 文本语用分析方法

语用学是一门研究如何用语言来达成一定目的的学科，即利用语用学进行文本分析，针对句子群（又称话题，Topic）开展高端分析，获取对文本内涵的掌握。话题是有因果关系的一些句子，它们必须连贯（Coherence），如例句 1；把可独立理解并且是良构的几个句子放到一起的结果，并不能保证获取的是话题，如例句 2。

例句 1：张玉把车钥匙弄丢了，她喝醉了。

例句 2：张玉把车钥匙弄丢了，她喜欢吃菠菜。

为完成文本因果关系提取，出现了话题检测与跟踪方法；为了完成互联网上不同文本信息内容自动分类，提出了文本分类器（也称为信息内容过滤）。话题检测与跟踪方法详见本书第 5 章；网络信息内容过滤方法详见本书第 4 章。

3.7 本章小结

本章介绍了网络信息内容的预处理技术，重点从文本预处理技术、文本内容分析方法、文本内容安全应用 3 方面介绍文本内容安全状态。文本预处理技术涉及中文分词技术、文本表示和文本特征提取，中文分词涉及机械分词法、语法分词法。文本表示介绍布尔模型、向量空间模型和概率模型等内容。文本特征提取给出了停用词过滤、文档频率阈值法、TFIDF 方法及信噪比的内容。在文本内容分析小节，分别从文本语法分析、语义分析以及语用分析 3 方面进行文本内容分析，从而为后续的文本处理提供量化的指标。本章内容重点是文本内容预处理技术，难点是文本语义分析。

习 题

1. 简述文本信息的语义特征。
2. 如何进行文本特征提取？
3. 词语情感倾向性分析有哪些方法？
4. 如何衡量特征抽取过程与选择过程所造成的信息损失？
5. 为什么要进行特征重构，常用的方法有哪些？