第 3 章

网络信息内容处理技术

理论讲解

3.1 网络信息内容处理概述

实验讲解

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。互联网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，本章简述了网络信息内容在进行安全处理时涉及的通用方法或技术。

3.1.1 网络信息内容处理一般流程

一般来说，在网络信息内容处理中，被研究的业务对象——信息内容，也可被认为是数据，是整个过程的基础，它驱动了整个网络信息内容处理过程，也是检验最后结果和指引分析人员完成网络信息内容处理的依据和顾问。信息内容可以分为文本信息内容和数据信息内容，其处理流程非常类似，一般遵循图 3-1 所示步骤，整个过程中还会存在步骤间的反馈。网络信息内容处理过程并不是自动的，绝大多数工作需要人工完成。整个网络信息内容处理过程和一般的数据挖掘过程非常类似，60％的时间用在数据准备上。这说明了网络信息内容处理对数据的严格要求，而后续安全分析工作仅占总工作量的 20％～30％。

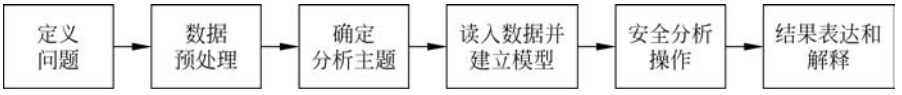

图 3-1 网络信息内容处理一般流程

网络信息内容处理过程主要包含 6 个步骤，各步骤的大体内容如下：

(1) 定义问题。首先明确定义将要解决的问题。信息内容安全分析者要熟悉所研究行业的数据和业务问题，缺乏这些，就不能够充分发挥数据挖掘的价值，很难得到正确的结果。模型建立取决于问题的定义，有时相似的问题所定义要求的模型几乎完全不同。清晰地定义业务问题，认清信息内容并分析目的，是数据挖掘的重要一步。挖掘的最后结果是不可预测的，但定义的安全问题应是有预见的。为了信息内容处理而处理带有盲目性，是不会成功的。

(2) 数据预处理。网络信息内容处理过程可以看作是一个“矿石精炼过程”,输入的是原始数据,输出的是“钻石”。数据预处理正是这个过程的核心。数据预处理阶段又可分为3个子步骤,即数据集成、数据选择、数据清洗。其中数据集成可将多文件或多数据库运行环境中的数据进行合并处理,解决语义模糊性问题、处理数据中的遗漏和清洗脏数据等。数据选择的目的是辨别出需要分析的数据集合,缩小处理范围,提高数据挖掘的质量,因此需要搜索所有与业务对象有关的内部和外部数据信息,并从中选择出适用于数据挖掘应用的数据。而数据清洗则是为了克服目前数据挖掘工具的局限性,提高数据质量,同时将数据转换成一个适用于特定安全分析算法的分析模型。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

(3) 确定分析主题。网络信息内容处理和数据挖掘一样,是经常需要进行回溯的过程。因此,没有必要在数据完全准备好之后才开始进行分析处理。因为随着时间的推移,分析处理所使用的数据、分组方式和数据清洗的效果等都将改变,并有可能改进整个模型。因此,在建立模型之前,需要了解研究主题的局限性,确定待研究的合适数据元素并决定如何进行数据操作等。

(4) 读入数据并建立模型。一旦确定要输入的数据之后,接着就是要用数据挖掘工具读入数据,并从中构造出安全分析模型。根据所选用的数据挖掘工具的不同,所构造出的数据模型也会有很大的差别。

(5) 安全分析操作。依照上述准备工作,利用选好的数据挖掘工具在数据中查找。这个搜索过程可以由系统自动执行,自底向上搜索原始事实以发现它们之间的某种联系,也可以加入用户交互过程,由分析人员主动发问,从上到下地找寻以验证假设的正确性。网络信息内容处理过程需要反复多次,通过评价数据挖掘结果不断调整网络信息内容处理的精度,以达到发现知识的目的。

(6)结果表达和解释。根据最终用户的决策目标对提取出的信息进行分析,把最有价值的信息区分出来,并通过决策支持工具提交给决策者。

一个网络信息内容处理工作的生命周期一般都包含上述6个阶段。这6个阶段顺序并非固定,可根据实际特定任务的产出进行前后调整。

3.1.2　文本信息内容预处理技术

在众多的网络信息内容中,文本信息占了很大的比重。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,如果不做特别的说明,本书所提及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

建立在比较理想条件下的。所谓理想条件,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并推向市场时,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加工处理,全世界掀起了研究与开发"预处理器"的热潮,并提出了很多网络信息内容预处理框架。

一般来说,网络信息内容预处理流程包括中文分词、停用词过滤、数据标准化等几个步骤,如图3-2所示。下面将对这些步骤依次进行介绍。

图3-2 文本信息内容预处理流程

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。"在计算机处理过程中,可以依靠空格识别出movie是一个词,但不能识别"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。下面分别介绍分词技术特点与常见的中文分词系统。

1) 分词技术特点

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:分词过程是跟词典做比较,不需要大量的语料库、规则库,其算法简单、复杂性小,对算法进行一定的预处理后分词速度较快。缺点是:不能消除歧义、不能识别未登录词,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，因此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，出现了一大批实用、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，另外国内的北大语言研究所、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的 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了优异成绩，已得到国内外大多数中文信息处理用户的支持。

2. 停用词过滤

停用词也被称为功能词，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。若计算机对其处理，不但是没有价值的工作，还会增加运算复杂度，通常在文本的停用词处理中可采用基于词频的方法将其除去。

停用词过滤一般都是基于对自然语言的观察，过滤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的与领域无关以及面向具体领域的停用词表，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的可用资源少一些。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。

除手工建立停用词表外，还可以采用统计方法，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,汉字中的"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. 数据标准化

网络信息内容中一部分内容有可能包含一些结构化数据。针对这些结构化数据的处理过程和前面文本内容的处理又有很大的不同,其中一个最关键的步骤是初始数据集的准备和转换。这个步骤与网络信息内容处理应用高度相关,但在大多数网络信息内容处理应用中,所给定的原始数据需要进行一定的转换操作,才能产生对所选的安全分析方法更有用的特征。转换操作如下:使用不同的方式计算,采用不同的样本大小,选择重要的比率,针对时间相关数据改变数据窗口的大小、移动平均数的变化等。

3.1.3 网络信息内容中数据信息的预处理

一般来说,网络信息内容中数据预处理阶段主要有两个中心任务,即数据标准化和数据平整化,下面分别进行介绍。

1. 数据标准化

一些安全分析方法,例如基于 n 维空间的点间距离计算的方法,可能需要对数据进行标准化,以获得最佳结果。测量值可按比例对应到一个特定的范围,如[-1,1]或[0,1]。如果这些值没有标准化,距离测量值将会超出数值较大的特征。数据的标准化有许多方法,这里列举3个简单有效的标准化技术。

1) 小数缩放

小数缩放也称小数定标规范化,是指通过移动属性值的小数位数,将属性值映射到[-1,1]之间,移动的小数位数取决于属性值绝对值的最大值。但仍然保留大多数原始数值。常见的缩放是使值在-1～1内。小数缩放可以表示为

$$v'(i)=v(i)/10^k \tag{3-1}$$

式中,$v(i)$是特征 v 对于样本 i 的值,$v'(i)$是缩放后的值,k 是保证$|v'(i)|$的最大值小于1的最小比例。

小数缩放方法在具体处理时首先在数据集中找出$|v'(i)|$的最大值,然后移动小数点,直到得出一个绝对值小于1的缩放新值。这个因子可用于所有其他的 $v(i)$。例如,数据集中的最大值为455,最小值是-834,那么特征的最大绝对值就是0.834,所有的 $v(i)$都用同一个因子1000($k=3$)。

2) 最小-最大标准化

假设特征 v 的数据范围为150～250,则前述的标准化方法将使所有标准化后的数据取值在0.15～0.25,整个取值范围堆积在一个极为狭窄的子区间中,不利于后续的特征提取操作。要使特征值在整个的标准化区间如[0,1]上获得较好的分布,可以用最小-最大标准化公式:

$$v'(i)=\frac{v(i)-\min[v(i)]}{\max[v(i)]-\min[v(i)]} \tag{3-2}$$

其中特征 v 的最小值和最大值是通过一个集合自动计算的，或者是通过特定领域的专家估算得到。这种转换也可应用于标准化区间[－1,1]。最小值和最大值的自动计算需要对整个数据集进行另一次搜索，但是计算过程较为简单。

3）标准差标准化

按标准差进行的标准化对距离测量值非常有效，但是把初始数据转化成了未被认可的形式。对于特征 v，平均值 mean(v)和标准差 sd(v)是针对整个数据集进行计算的。那么对于样本 v，可用下述等式来转换特征的值：

$$v(i)=\frac{v(i)-\text{mean}(v)}{\text{sd}(v)} \tag{3-3}$$

如果一个属性值的初始集合是 $v=\{1,2,3\}$，mean(v)＝2，sd(v)＝1，利用式(3-3)进行计算，则标准化值的新集合为 $v^*=\{-1,0,1\}$。

需要指出的是，数据标准化处理对几种数据挖掘方法来说很有用，如利用距离进行聚类等分析方法。但标准化并不是一次性或一个阶段的事件，除对信息内容处理当前阶段所需的数据进行转换和转变以外，还必须对后续阶段可能出现的新数据进行同样的数据标准化。因此，数据预处理时必须把标准化的参数和方法一起保存。

2. 数据平整化

在大型数据集中，通常有一些样本不符合数据模型的一般规则。这些样本和数据集中的其他数据有很大的不同，称为异常点。异常点可能是测量误差造成的，也可能是由于数据固有的可变性。例如，如果一个人的年龄在数据库中显示为－1，这个值显然不正确，这个错误可能是计算机程序中字段“未记录年龄”的默认设置造成的。另一方面，如果在数据库中，一个人的子女数为 25，这个数据是不同寻常的，需要检查其合理性，确认其是否为排版错误，还是真实情况。

在数据预处理过程中希望尽量将异常点对最终模型的影响减到最小或去除。异常点检测方法可以检测出数据中的异常观察值，并在适当时去除它们。出现异常点的原因有机械故障、系统行为的改变、欺诈行为、人为错误、仪器错误或样本总体的自然偏差。异常检测可以识别出系统故障和错误，以免它们逐渐累积，最终造成灾难性的结果。异常点检测也可被称为野点检测、反常检测、噪声检测、偏差检测等。异常点的存在可能导致机器学习方法失效，原因在于异常点可能会在数据模型中引入非正态分布或复杂性，从而很难(甚至不可能)以可行的计算方式找到准确的数据模型。因此，异常点的有效检测可以显著降低因异常数据做出错误决策的风险，并有助于识别、防止、去除由于恶意或错误行为带来的负面影响。

网络信息内容中数据平整化处理一般采用机器学习的方法。该方法是建立在数据集合中“正常”观测值要远远多于异常点的假设基础上，主要包含两个主要步骤：找出“正常”行为的规律和使用“正常”规律来检测异常点。例如在检测银行交易中的信用卡欺诈行为时，异常点检测揭示出其中的欺诈行为。但在大多数数据挖掘应用中，尤其是应用于大型数据集时，异常点并不是很有用，它们只是由于搜集数据时出现过失而产生的结果，而不是数据集的特征。

检测异常点，并从数据集中去除它们的过程可以描述为：从 n 个样本中选 k 个与其

余数据显著不同、例外或不一致的样本($k << n$)。异常点检测方法的主要类型有数据可视化方法、基于统计的异常点检测方法和基于距离的异常点检测方法。

1) 数据可视化方法

数据可视化方法是一种较为常见的异常检测辅助方法。该方法可根据待检测数据所处的维度,将其转换成用户易于理解、可直接观测的异常点分布图。在一维到三维场景中,异常点检测方法很有用,但在多维数据中其作用就降低很多,原因在于多维空间缺乏恰当的可视化方法。图 3-3 和图 3-4 给出了利用数据可视化方法针对二维样本和三维样本数据进行异常点检测的例子。该方法的主要局限是对数据维度有限制且过程非常耗时,对异常点探测具有一定的主观性。

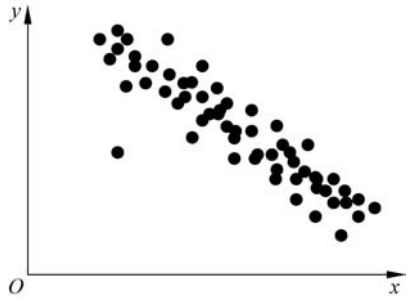

图 3-3 二维数据集中有一个无关的样本

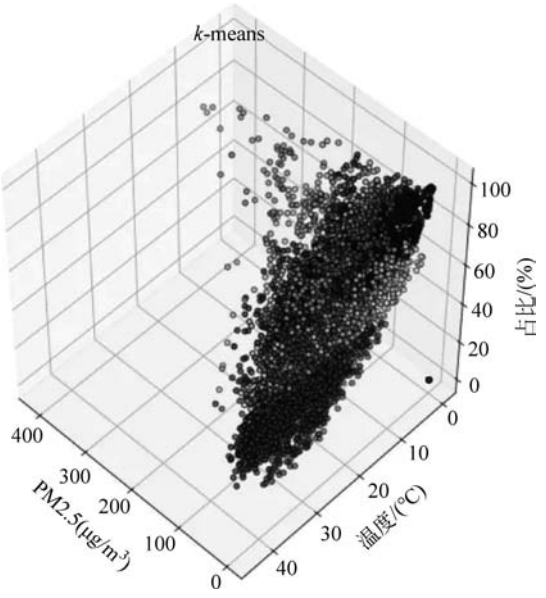

图 3-4 三维数据集中包含异常偏离的样本

2) 基于统计的异常点检测方法

基于统计的异常点检测方法是目前较为常用的检测方法,它可分为一元方法和多元方法。在早期工作中,一般采用一元方法,其检测效率都依赖一个假设:数据的基本分布是已知、相同且独立的,并且分布参数和异常点的期望类型也是已知的。在一元异常点探测中,如果样本没有被异常点污染,则通过样本均值和样本方差能很好地估计数据位置和数据模型。但在数据样本库受到异常点的污染后,这些参数就会背离目标,显著影响异常点检测的性能。最简单的一维样本异常点检测方法基于传统的单峰统计学,即假定值的分布已知,并确定好了基本的统计参数,如均值和方差。根据这些值和异常点的期望(预测)数目,就可以确定方差函数的阈值。所有阈值之外的样本都可能是异常点,如下例。

如果给定的数据集用 20 个不同的值描述年龄特征:

```
年龄 = {3,56,23,39,156,52,41,22,9,28,139,31,55,20, - 67,37,11,55,45,37}
```

那么,相应的统计参数:

均值＝39.9

标准差＝45.65

如果选择数据正态分布的阈值：

阈值＝均值±2×标准差

那么，所有在[－54.1,131.2]区间以外的数据都是潜在的异常点。年龄特征还有一个特性：年龄总是大于零，于是可进一步把该区间缩小到[0,131.2]。在上例中，根据所给的条件，有3个值是异常点：156,139和－67。那么可以断定，这3个都是输入错误(多输入一个数字或“－”号)的概率很高。

这种一元异常点检测方法存在的主要问题在于预先假设了数据的分布，而在大多数现实案例中，数据分布是未知的。

3) 基于距离的异常点检测方法

基于距离的异常点检测方法能够直观指出远离数据分布中心的样本。该方法一般需要使用特定的距离度量值来完成。马氏(Mahalanobis)距离值法是常见的基于距离的异常点检测法所使用的量度，该方法通过分析样本内部属性之间的依赖关系并进行比较分析检测。马氏法依赖多元分布的估计参数。如给定 p 维数据集中的 n 个观察值 x_i(其中 $n \gg p$)，用 $\bar{\boldsymbol{x}}_n$ 表示样本平均向量，$\boldsymbol{V}_n$ 表示样本协方差矩阵，其中

$$\boldsymbol{V}_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}}_n)(x_i - \bar{\boldsymbol{x}}_n)^{\mathrm{T}} \tag{3-4}$$

每个多元数据点 $i(i=1,2,\cdots,n)$ 的马氏距离用 M_i 表示，则

$$M_i = \left[\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}}_n)^{\mathrm{T}} V_n^{-1}(x_i - \bar{\boldsymbol{x}}_n)\right]^{1/2} \tag{3-5}$$

于是，马氏距离很大的 n 维样本就被看作异常点。许多统计方法要求，数据特有的参数表示以前的数据知识。但此类信息常无法获得，或者计算成本很高。另外，大多数现实世界中的数据集并不遵循某个特定的分布模型。

基于距离的异常点检测技术在实现时，并没有预先假设数据分布模型。但它们的计算量呈指数型增长，因为它们要计算所有样本之间的距离。计算的复杂性依赖数据集的维数 m 和样本的数量 n，常常表示为 $O(n^2 m)$。因此，非常大的数据集往往没有合适的方法检测异常点。另外，数据集存在密集区域和稀疏区域时，该定义还会出问题。例如，维数增加时，数据点会散布在更大的空间中，其密度会减小，这样凸包就更难识别，这就是所谓的维数灾难。

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、信息检索系统等任务中工作良好,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 3-5 为一个转换示例。

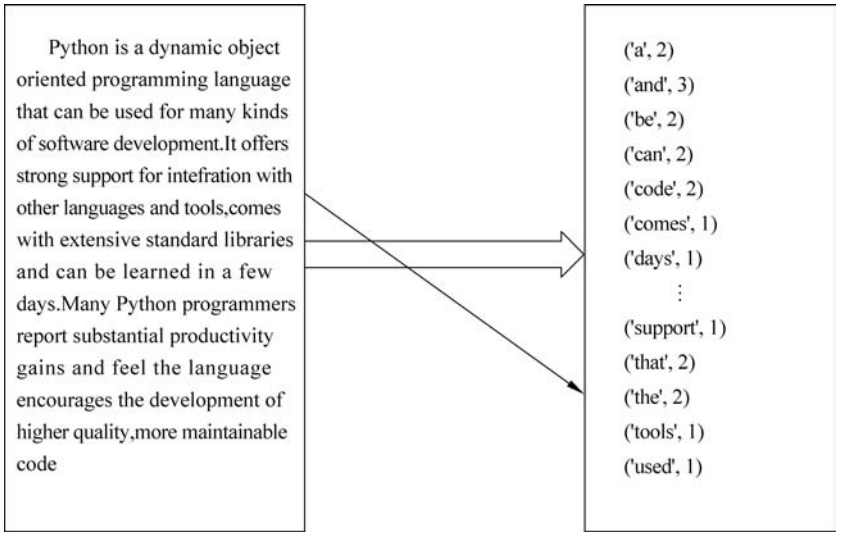

图 3-5 词袋模型

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。短词具有较高的出现频率和更多的含义，是面向功能的；而长词的出现频率较低，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做出限制。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 $26^3=17576$ 个三元组;$n=4$ 时,有 $26^4=456976$ 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(Bidirectional Maximum Match，BMM)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：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、隐马尔可夫模型、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，而且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，当前对未登录词的识别研究，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。该问题考虑了存在词汇嵌套情况时的处理策略,例如,“文本分类”可以看作一个单独的词,也可以看作“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标准共收录了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27484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$6763^2(45738169)>26^5(11881376)$。

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”“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 特征子集选取方法

3.3.1 特征子集选择概述

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
$\overline{n}(t)$	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
$\overline{n}_{c_i}(t)$	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
$\mathrm{tf}_{d_j}(t)$	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：

$$n=\sum_{i=1}^{k} n_{c_i} \tag{3-6}$$

$$n(t)=\sum_{i=1}^{k} n_{c_i}(t) \tag{3-7}$$

式(3-6)表示样本总数等于各类别样本数之和，式(3-7)表示只包含任一特征项 t 的样本集合，也满足类似关系。

$$n=n(t)+\overline{n}(t) \tag{3-8}$$

$$n_{c_i}=n_{c_i}(t)+\overline{n}_{c_i}(t) \tag{3-9}$$

式(3-8)表示 $n(t)$ 和 $\overline{n}(t)$ 互补，式(3-9)表示这种关系也适用于任意给定文本类别。

$$\mathrm{tf}=\sum_{i=1}^{\hat{m}} \mathrm{tf}(t_i) \tag{3-10}$$

$$\mathrm{tf}(t)=\sum_{j=1}^{n} \mathrm{tf}_{d_j}(t) \tag{3-11}$$

式(3-10)和式(3-11)给出了 tf 和 tf(t)的计算方法。

利用这些参数，结合统计学、信息论等学科，即可进行特征子集选择。

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),美国语言学家齐夫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 $n(t)$(式 3-12 中未出现)存在近似反比关系:

$$\text{rank}(t) \cdot \text{TF}(t) \approx C \tag{3-12}$$

中文也存在类似规律,对新浪滚动新闻的 133577 篇新闻的分词结果进行统计,结果见图 3-6,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。

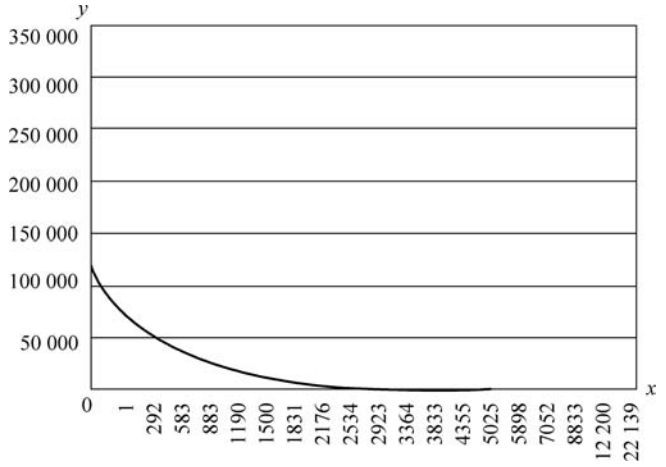

图 3-6 中文语料的齐夫定律现象验证

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,对于汉语没有成熟的停用词表,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 法

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。TF 背后隐含的假设是,查询关键字中的单词应该相对于其他单词更加重要,而文档的重要程度也就是相关度,与单词在文档中出现的次数成正比。例如,Car 这个单

词在文档A中出现了5次，而在文档B中出现了20次，那么TF计算就认为文档B可能更相关。

具体来说，第一部分可以用TF(t)表示某关键词在文档中出现的频率，可简单理解为词出现的次数越多，则该词重要程度越高；第二部分采用逆文本频率指数(Inverse Document Frequency，IDF)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，一个词在其他文档中出现的次数越多，分母就越大，取对数的值就越小，说明这个词在所有文章中的重要程度就越小：

$$\mathrm{IDF}(t)=\log\frac{n}{n(t)} \tag{3-13}$$

第一部分和第二部分都满足取值越大时，TF-IDF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：

$$\text{TF-IDF}(t)=\mathrm{TF}(t)\cdot\mathrm{IDF}(t)=n(t)\cdot\log\frac{n}{n(t)} \tag{3-14}$$

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　信噪比法

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声的差值，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年由香农引入到信息论中，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$-\log(p_i)$，该系统的信息熵即为整个系统的平均信息量：

$$\mathrm{Entropy}(X)=-\sum_{i=1}^{c}p_i\log p_i \tag{3-15}$$

为方便计算，令p_i为0时，熵值为0(即0log0)，熵的取值范围是$[0,\log c]$，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$\log c$。根据熵的概念，定义特征项的噪声：

$$\mathrm{Noise}(t)=-\sum_{j=1}^{n}P(d_j,t)\log P(d_j,t) \tag{3-16}$$

式(3-16)中，$P(d_j,t)=\dfrac{\mathrm{TF}_{d_j}(t)}{\mathrm{TF}(t)}$表示特征项$t$出现在样本$d_j$中的可能性，特征项$t$的噪声函数取值范围为$[0,\log n]$，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$\log n$，这符合越集中在较少样本中，特征项为噪声可能性越小的直观认识，相应特征项t的信号值若用$\log\mathrm{TF}(t)$来表

示,可得信噪比计算公式:

$$\begin{aligned}\mathrm{SNR}(t) &= \log\mathrm{TF}(t) - \mathrm{Noise}(t)\\ &= \log\mathrm{TF}(t) + \sum_{j=1}^{n} P(d_j,t)\log P(d_j,t)\end{aligned} \tag{3-17}$$

信噪比取值范围为[0,logTF(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声,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 logTF(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,则在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 网络信息内容安全分析

网络信息内容安全分析是一门交叉学科,融合了数据库、人工智能、机器学习、统计学等多个领域的理论和技术。下面简单介绍网络信息内容常见的安全分析方法。

3.4.1 网络信息内容安全分析方法概述

网络信息内容安全分析中采用了很多结构或非结构化数据,并使用了大量数据挖掘理论和方法。由于数据挖掘应用领域十分广泛,因此产生了多种数据挖掘的算法和方法,如关联分析、聚类分析等。这些方法的效果依赖于网络信息内容分析中样本数据的分布情况。因此,应针对具体的挖掘目标和应用对象来选择不同的安全分析方法。目前具有代表性的安全分析方法一般有以下几类。

1. 概念描述

概念通常是对一个包含大量数据的数据集总体情况的描述。概念描述就是通过汇总、分析和比较与某类对象关联的数据,对此类对象的内涵进行描述,并概括这类对象的有关特征。这种描述是汇总的、简洁的和精确的,当然也是非常有用的。概念描述分为特征性描述和区别性描述。前者描述某类对象的共同特征,后者描述不同类对象之间的区别。生成一个类的特征性描述只涉及该类对象中所有对象的共性;生成区别性描述则涉及目标类和对比类中对象的共性。该功能在网络信息内容安全分析中可以用于从一堆新闻或博客信息中抽象提取出某个热点事件或话题。

2. 分类分析

分类刻画了一类事物,这类事物具有某种意义上的共同特征,并明显与不同类事物相区别。分类分析就是通过分析示例数据库中的数据,为每个类别做出准确的描述,或建立分析模型,或挖掘出分类规则,然后用这个分类规则对其他数据库中的记录进行分类。从机器学习的观点来看,分类技术是一种有指导的学习,即每个训练样本的数据对象已经有类标识,通过学习可以形成与表达数据对象和类标识间对应的知识。目前已有多种分类分析模型得到应用,主要有神经网络方法、Bayesian 分类、决策树、统计分类、粗糙集分类、SVM 方法、覆盖算法等。在数据挖掘中这些方法均遇到数据规模的问题,即大多数方法

能有效解决小规模数据库的数据挖掘问题，但当应用于大数据量的数据库时，会出现性能恶化、精度下降的问题。分类分析可用于恶意代码检测分析或入侵检测中的恶意流量分析。

3. 聚类分析

聚类是把一组个体按照相似性归成若干类别，其目的是使得属于同一类别的个体之间的差别尽可能小，而不同类别上的个体间的差别尽可能大。聚类结束后，每类中的数据由唯一的标志进行标识，各类数据的共同特征也被提取出来，用于对该特征进行描述。提高聚类效率、减少时间和空间开销，以及如何在高维空间进行有效数据聚类是聚类研究中的主要问题。聚类分析的方法很多，如 k-平均算法、k-中心点算法、基于凝聚的层次聚类和基于分裂的层次聚类等。采用不同的聚类方法，对于相同的记录集合可能有不同的划分结果。

分类和聚类技术不同，前者总是在特定的类标识下寻求新元素属于哪个类，而后者则是通过对数据的分析比较生成新的类标识。由于聚类可使用无监督的数据集，因此可应用于网络信息内容安全分析中的舆情分析。

4. 关联分析

关联分析的目的是找出样本数据集中属性值之间的联系，形成关联规则。为了发现有意义的关联规则，需要给定两个阈值：最小支持度和最小可信度。在这个意义上，挖掘出的关联规则就必须满足最小支持度和最小可信度。关联规则是 1993 年由 R. Agrawal 等人提出的，然后扩展到从关系数据库、空间数据库和多媒体数据库中挖掘关联关系，并且要求挖掘出通用的、多层次的、用户感兴趣的关联规则。随着应用和技术的发展，近年来对网络信息内容安全分析的技术提出了更新的要求，如大数据开源数据情报的在线挖掘以及如何提高挖掘大型安全数据库的计算效率、减小 I/O 开销、挖掘定量型关联规则等。

5. 时间序列分析

时间序列分析中的相似模式发现分为相似模式聚类和相似模式搜索两种。相似模式聚类是将时间序列数据分隔成等长或不等长的子序列，然后用模式匹配的方法进行聚类，找出序列中所有相似的模式。相似模式搜索是指给定一个陌生子序列，在时间序列中搜索所有与给定子序列模式最接近的数据子序列。时间序列分析主要应用于话题跟踪检测中检测出热点话题，并依据时间序列进行后续跟踪挖掘处理等场景。

6. 偏差分析

偏差分析包括分类中的反常实例、例外模式、观测结果对期望值的偏离以及量值随时间的变化等，基本思想就是对数据库中的偏差数据进行检测和分析，检测出数据库中的一些异常记录，它们在某些特征上与数据库中的大部分数据有着显著不同。通过发现异常，可以引起人们对特殊情况的格外关注。异常包括的模式有：出现在其他模式边缘的奇异点；不满足常规类的异常实例；与父类或兄弟类不同的类；观察值与模型推测出的期望值有明显差异的例子等。偏差分析方法主要有：基于统计的方法、基于距离的方法和基于偏移的方法。孤点数据的发现可以应用于网络信息内容安全分析入侵检测中的异常检测等领域。

3.4.2 分类分析方法

网络信息内容安全分析中,分类分析是较为常见的方法。分类分析就是确定对象属于哪个预定义的目标类。分类问题是一个普遍存在的问题,有许多不同的应用。例如,根据电子邮件的标题和内容检查出垃圾邮件,根据核磁共振扫描的结果区分肿瘤是恶性的还是良性的。

分类技术(或分类法)是一种根据输入数据集建立分类模型的系统方法。分类法主要包括决策树分类法、基于规则的分类法、粗糙集理论、支持向量机和朴素贝叶斯分类法。这些技术都使用一种学习算法(Learning Algorithm)确定分类模型,该模型能够很好地拟合输入数据中类标号和属性集之间的联系。学习算法得到的模型不仅要很好地拟合输入数据,还要能够正确地预测未知样本的类标号。因此,训练算法的主要目标就是建立具有很好的泛化能力的模型,即建立能够准确预测未知样本类标号的模型。

图3-7展示了解决分类问题的一般方法。首先,需要一个训练集(Training Set),它由类标号已知的记录组成。使用训练集建立分类模型,该模型随后将运用于检验集(Test Set),检验集由类标号未知的记录组成。

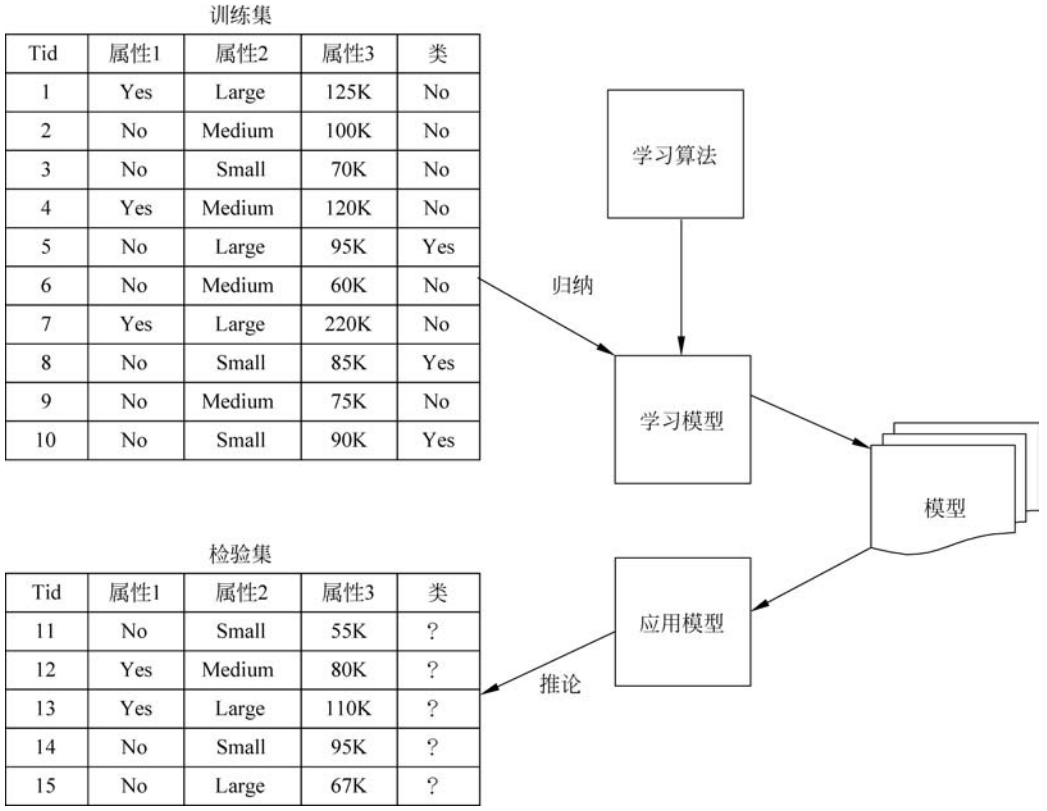

图3-7 建立分类模型的一般方法

分类模型的性能根据模型正确和错误预测的检验记录计数进行评估，这些计数存放在称作混淆矩阵(Confusion Matrix)的表格中。表3-2描述了二元分类问题的混淆矩阵。每个表项 f_{ij} 表示实际类标号为 i 但被预测为类 j 的记录数，例如，f_{01} 代表原本属于类0但被误分为类1的记录数。按照混淆矩阵中的表项，被分类模型正确预测的样本总数是($f_{11}+f_{00}$)，而被错误预测的样本总数是($f_{10}+f_{01}$)。

表3-2　二元分类问题的混淆矩阵

实际的类	预测的类	
	类=1	类=0
类=1	f_{11}	f_{10}
类=0	f_{01}	f_{00}

虽然混淆矩阵提供衡量分类模型性能的信息，但是用一个数汇总这些信息更便于比较不同模型的性能。为实现这一目的，可以使用性能度量(Performance Metric)，如准确率(Accuracy)，其定义如下：

$$\text{准确率}=\frac{\text{正确预测数}}{\text{预测总数}}=\frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}} \tag{3-18}$$

同样，分类模型的性能可以用错误率(Error Rate)来表示，其定义如下：

$$\text{错误率}=\frac{\text{错误预测数}}{\text{预测总数}}=\frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}} \tag{3-19}$$

大多数分类算法都在寻求这样一些模型，当把它们应用于检验集时具有最高的准确率，或者等价地，具有最低的错误率。通过估计误差有助于学习算法进行模型选择(Model Selection)，即找到一个具有合适复杂度、不易发生过分拟合的模型。模型一旦建立，就可以应用到检验数据集上，预测未知记录的类标号。

测试模型在检验集上的性能是有用的，因为这样的测量给出模型泛化误差的无偏估计。在检验集上计算出的准确率或错误率可以用来比较不同分类器在相同领域上的性能。然而，为了做到这一点，检验记录的类标号必须是已知的。

3.4.3　聚类分析方法

聚类是指根据数据对象之间的相似性，把一组数据对象划分为多个有意义组的过程，每个组称为类或簇(Cluster)，同一个簇内的数据对象之间具有较高的相似性，不同簇内的数据对象之间相差则较大。与分类不同的是，聚类目标所要求划分的类别是未知的，且聚类数据对象中没有关于类别特征的数据，其划分簇的过程不是以包含类别的数据对象为指导，而是根据数据对象的特征来进行的。以此为基础的聚类分析对于数据理解及数据处理都有着重要的作用。数据理解用来分析和描述类或概念上有意义的、具有共同特征的对象组，而聚类分析是研究自动地发现潜在的类或簇的技术。在许多领域中有着大量基于数据理解的聚类分析应用，以下是一些常见的例子。

在数据处理方面，聚类分析一般用于汇总、压缩、发现最近邻等处理。聚类分析提供由个别数据对象到数据对象所指派的簇的抽象。此外，一些聚类技术使用簇原型(即代表

簇中其他对象的数据对象)来刻画簇特征。这些原型可以用作大量数据分析和数据处理技术的基础。

(1) 汇总。许多数据分析技术,如回归和PCA,都具有$O(n^2)$或更高的时间或空间复杂度(其中n是对象的个数)。因此,对于大型数据集,这些技术不切实际。然而,可以将算法用于仅包含簇原型的数据集,而不是整个数据集。依赖分析类型、原型个数和原型代表数据的精度,汇总结果可以与使用所有数据得到的结果相媲美。

(2) 压缩。簇原型可以用于数据压缩。例如,创建一个包含所有簇原型的表,即每个原型赋予一个整数值,作为它在表中的位置。每个对象用与它所在的簇相关联的原型的索引表示。这类压缩称作向量量化(Vector Quantization),并常常用于图像、声音和视频数据,此类数据的特点是:①许多数据对象之间高度相似;②某些信息丢失是可以接受的;③希望大幅度压缩数据量。

(3) 发现最近邻。找出最近邻可能需要计算所有点对点之间的距离。通常,可以更有效地发现簇和簇原型。如果对象相对地靠近簇的原型,则可以使用簇原型减少发现对象最近邻所需要计算的距离的数目。直观地说,如果两个簇原型相距很远,则对应簇中的对象不可能互为近邻。这样,为了找出一个对象的最近邻,只需要计算到邻近簇中对象的距离,其中两个簇的邻近性用其原型之间的距离度量。

聚类技术的基本类型一般包括划分法、密度法、层次法、网格法和模型法。同样的数据集采用不同聚类方法,其聚类结果也往往不相同。甚至采用相同类型的聚类算法,选用不同参数,结果也很不一样。实际应用中,聚类结果好坏不仅仅取决于算法的选择,同时取决于业务领域的认识程度。聚类用户需要深刻了解所选用的聚类技术,而且要知道数据收集的细节和业务领域知识。对聚类数据了解越多,用户越能成功地评估数据集的真实结构。

划分法(Partition Method)聚类把一个包含n个数据对象的数据集分组成k个簇($k\leqslant n$)。每一个簇至少包含一个数据对象,且每一个数据对象属于且仅属于一个簇。对给定k,基于划分法的聚类算法首先给出一个初始的分组方法,随后通过反复迭代重新分组,使得每一次重新分组好于前一次分组。评判分组好与差的标准是:同一簇中的数据对象相似度越高越好,不同簇中的数据对象相异度越高越好。为计算一个簇内所有数据对象的相似度,需要为簇指定一个原型(簇中心、代表对象),簇内所有对象的相似度为簇内其他对象与原型之间相似度之和。因此,又称划分法为基于原型的聚类算法。划分法的代表算法有k-means、k-medoids、k-modes、PAM(Partition Around Medoid)等。由于划分法基于与原型的距离进行分组,因此一般只能发现圆形或球形的簇。

与基于簇原型和相似度的划分法不同,密度法(Density-Based Methods)聚类基于密度定义分组数据对象。密度法中,首先根据用户给定参数,计算每个数据对象的密度大小,并以此区分低密度区域和高密度区域,前者将后者分隔,每个高密度区域中的数据对象则可构成一个聚类或簇。密度法的代表算法有DBSCAN(Density-Based Spatial Clustering of Application with Noise)、OPTICS(Ordering Points to Identify the Clustering Structure)和DENCLUE(Density Based Clustering)。密度聚类法可以克服基于距离的聚类只能发现圆形或球形的簇的缺点,聚类结果也不要求每个数据对象都划分

到某个簇中。图 3-8(b)给出了密度法聚类的结果示意图。图中，虚线包围的区域为高密度区域，共有 4 个，即密度法聚类识别该数据集有 4 个簇。没有被虚线包围的其他区域为低密度区域。

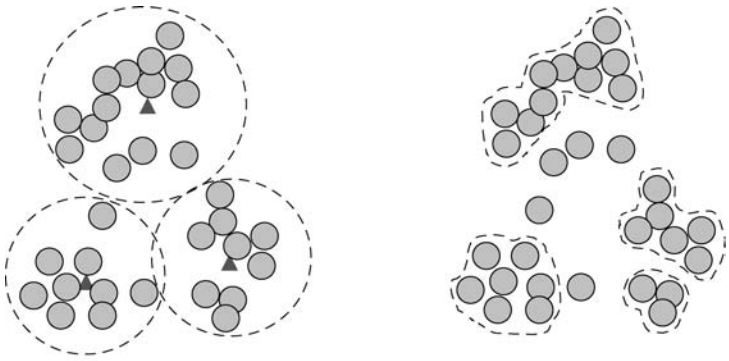

图 3-8　同样的数据集采用划分法和密度法聚类结果

3.4.4　关联分析法

“数据海量，信息缺乏”是很多行业在数据爆炸过程中普遍面对的尴尬，如今对信息的获取能力，决定了在前所未有的激烈竞争环境中的决策能力。如何挖掘出数据中存在的各种有用的信息，即对这些数据进行分析，发现其数据模式及特征，然后可能发现某个客户、消费群体或组织的金融和商业兴趣，并可以观察金融市场的变化趋势。如何有效地获取信息，是每个人、每个组织面临的难题。信息是现代企业的生命线，如果一个“节点”既不提供信息也不使用信息，也就失去了存在的价值。关联分析(Association Analysis)用于发现隐藏在大型数据集中的令人感兴趣的关联关系，描述数据之间的密切度。所发现的模式通常用关联规则(Association Rule)或频繁项集的形式表示。

关联规则的概念产生于 1993 年，由 Agrawal、Imielinski 和 Swami 提出。其一般定义如下：令 $I=\{i_1,i_2,\cdots,i_m\}$表示一个项集。设任务相关的数据 D 是数据库事务的集合，其中每个事务 T 是项的集合，使得 $T\subset I$。每一个事务都有一个标识符，称为 TID。设 A 是一个项集，事务 T 包含 A，当且仅当 $A\subseteq T$。关联规则是形如 $A\Rightarrow B$ 蕴含式，其中 $A\subset I$，$B\subset I$，并且 $A\cap B=\varnothing$。

如果 D 中包含 $A\cup B$(即集合 A 和 B 的并或 A 和 B 二者)的比例是 s，则称关联规则 $A\Rightarrow B$ 在事务集 D 中的支持度为 s，也可以表示成概率 $P(A\cup B)$。如果 D 中包含 A 事务的同时也包含 B 的比例是 c，则称关联规则 $A\Rightarrow B$ 在事务集 D 中具有置信度 c，它可以表示为条件概率 $P(B|A)$。即

$$\text{support}(A\Rightarrow B)=P(A\cup B) \tag{3-20}$$

$$\text{confidence}(A\Rightarrow B)=P(B\mid A) \tag{3-21}$$

支持度和置信度是描述关联规则的两个重要概念，支持度用于衡量关联规则在整个数据集中的统计重要性。简单来说，支持度度量的是在所有行为中规则 A 和 B 同时出现

的概率。置信度用于衡量关联规则的可信程度,即置信度度量的是出现 A 的情况下 B 出现的概率。如对于购物篮分析,挖掘支持度的意义就是"购买 A 商品,也购买 B 的人数/全部销售订单";置信度就是"购买 A 商品,也购买 B 的人数/所有包含商品 A 的销售订单"。

同时满足最小支持度阈值(min_sup)和最小置信度阈值(min_conf)的规则称为强关联规则。一般来说,只有支持度和置信度较高的关联规则才可能是用户感兴趣、有用的关联规则。在本书中采用 0～100%之间的值表示支持度和置信度值。

项的集合称为项集。包含 k 个项的项集称为 k 项集。例如集合{computer,antivirus_software}是一个 2 项集。

项集的出现频率是包含项集的事务数,简称项集的频率、支持度计数或计数。式(3-20)定义的项集支持度有时称作相对支持度,而出现频率称作绝对支持度。如果项集 I 的相对支持度满足预先定义的最小支持度阈值(即 I 的绝对支持度满足对应的最小支持度计数阈值),则 I 是频繁项集。频繁 k 项集的集合通常记作 L_k。

由式(3-21),有

$$\text{confidence}(A \Rightarrow B) = P(B \mid A) = \frac{\text{support}(A \cup B)}{\text{support}(A)} = \frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)} \tag{3-22}$$

式(3-22)表明规则 $A \Rightarrow B$ 的置信度容易从 A 和 $A \cup B$ 的支持度计数推出。即一旦得到 A、B 和 $A \cup B$ 的支持度计数,导出对应的关联规则 $A \Rightarrow B$ 和 $B \Rightarrow A$,并检查它们是否是强关联规则。这样一来,挖掘关联规则的问题可以归结为挖掘频繁项集。

一般说来,关联规则的挖掘可以看作以下过程。

(1) 根据最小支持度找出数据集 D 中所有的频繁项集:根据定义,这些项集的每一个出现的频繁性至少与预定义的最小支持度计数 min_sup 一样。

(2) 由频繁项集产生强关联规则:根据定义,这些规则必须满足预先给定的最小支持度和最小置信度阈值。

关联规则的原理看似很简单,但实际运用的时候,就会发现存在很多问题,想从浩瀚的记录集中,挖掘一条有意义的关联规则,如果仅从支持度和置信度两个度量指标进行评估和选择强弱,会发现在个别情况下推荐的规则效果非常差。由于第二步的开销远低于第一步,所以挖掘关联规则的总体性能由第一步决定。第一步是关键,它将影响整个关联规则挖掘算法的效率,因此,关联规则挖掘算法的核心问题是频繁项集的产生。

从大型数据集中挖掘频繁项集的主要挑战是这种挖掘常常产生大量满足最小支持度(min_sup)阈值的项集,当 min_sup 设置得很低时尤其如此。这是因为如果一个项集是频繁的,则它的每个子集也是频繁的。一个长项集将包含组合个数较短的频繁子集项。例如,一个长度为 100 的频繁项集$\{a_1, a_2, \cdots, a_{100}\}$包含 $C_{100}^1 = 100$ 个频繁 1 项集 a_1, $a_2, \cdots, a_{100}$,C_{100}^2 个频繁 2 项集$\{a_1, a_2\}$,$\{a_1, a_3\}, \cdots, \{a_{99}, a_{100}\}$,以此类推。这样,频繁项集的总个数为

$$C_{100}^1 + C_{100}^2 + \cdots + C_{100}^{100} = 2^{100} - 1 \approx 1.27 \times 10^{30} \tag{3-23}$$

这对于任何计算机而言,项集的个数都太大,无法计算和存储。为了克服这一困难,引进

闭频繁项集和极大频繁项集的概念。

如果不存在真超项集[①] Y 使得 Y 与 X 在 S 中有相同的支持度计数，则称项集 X 在数据集 S 中是闭合的。如果 X 在 S 中是闭合的和频繁的，则项集 X 是 S 中的闭频繁项集。如果 X 是频繁的，并且不存在超项集 Y 使得 $Y \subset X$ 并且 Y 在 S 中是频繁的，则项集 X 是 S 中的极大频繁项集(或极大项集)。

设 C 是数据集 S 中满足最小支持度阈值 min_sup 的闭频繁项集的集合，令 M 是 S 中满足 min_sup 的极大频繁项集的集合。假定有 C 和 M 中的每个项集的支持度计数。注意，C 和它的计数信息可以用来导出频繁项集的完整集合。因此，称 C 包含了关于频繁项集的完整信息。另一方面，M 只存储了极大项集的支持度信息。通常，它并不包含其对应的频繁项集的完整的支持度信息。用下面的例子解释这些概念。

闭频繁项集和极大频繁项集。假定事务数据库只有两个事务：$\{a_1, a_2, \cdots, a_{100}\}$，$\{a_1, a_2, \cdots, a_{50}\}$。设最小支持度计数阈值 min_sup=1。有两个闭频繁项集和它们的支持度，即 $C=\{\{a_1, a_2, \cdots, a_{100}\} : 1; \{a_1, a_2, \cdots, a_{50}\} : 2\}$。只有一个极大频繁项集：$M=\{\{a_1, a_2, \cdots, a_{100}\} : 1\}$(不能包含 $\{a_1, a_2, \cdots, a_{50}\}$ 为极大频繁项集，因为它有一个频繁超集 $\{a_1, a_2, \cdots, a_{100}\}$)。与上面相比，确定了 $2^{100}-1$ 个频繁项集，数量太大，根本无法枚举。

闭频繁项集的集合包含了频繁项集的完整信息。例如，可以从 C 推出：①$\{a_2, a_{45} : 2\}$，因为 $\{a_2, a_{45}\}$ 是 $\{a_1, a_2, \cdots, a_{50} : 2\}$ 的子集；②$\{a_8, a_{55} : 1\}$，因为 $\{a_8, a_{55}\}$ 不是 $\{a_1, a_2, \cdots, a_{50} : 2\}$ 的子集，而是 $\{a_1, a_2, \cdots, a_{100} : 1\}$ 的子集。然而，从极大频繁项集我们只能断言两个项集 $\{a_2, a_{45}\}$ 和 $\{a_8, a_{55}\}$ 是频繁的，但是不能断言它们的实际支持度计数。

3.4.5　安全分析常用算法

网络信息内容安全分析本质上属于数据挖掘应用的一种，其常用算法涵盖但不限以下多种。

1. 决策树方法

决策树表示形式简单，所发现的模型也易于为用户理解，是挖掘分类知识中最流行的方法之一。它利用信息论中的信息熵作为节点分类的标准，建立决策树的一个节点，再根据属性当前的值域建立节点的分支。决策树的建立是一个递归过程。在知识表示方面具有直观、易于理解等优点。最早的决策树算法是 ID3 方法，它对较大的数据集处理效果较好。在 ID3 的基础上，Quinlan 又提出了改进的 C4.5 算法。

2. 模糊集方法

模糊集方法是利用模糊集合理论对实际问题进行模糊评判、模糊决策、模糊模式识别和模糊聚类分析，是一种应用较早的处理不确定性问题的有效方法。系统的复杂性越高，模糊性越强。模糊集理论是用隶属度来刻画模糊事物的亦此亦彼性的。

在很多场合，数据挖掘任务所面临的数据具有同样的模糊性和不精确性，因此把模糊

---

① Y 是 X 的真超项集，即 X 是 Y 的真子项集，$X \subset Y$。换言之，X 中的每一项都包含在 Y 中，但是 Y 中至少有一个项不在 X 中。

数学理论应用于数据挖掘则顺理成章。使用模糊集方法可以对已挖掘的大量的关联规则的有用性、兴趣度等进行评判,也可用于分类、聚类等数据挖掘任务。

3. 神经网络方法

神经网络是指一类计算模型,它模拟人脑神经元结构及某些工作机制,利用大量的简单计算单元连成网络来实现大规模并行计算,它有并行处理、分布存储、高度容错、自组织等诸多优点,因此它是数据挖掘中的重要方法。近年来人们研究从训练后的神经网络中提取规则的方法,从而推动了神经网络在数据挖掘分类问题中的应用。神经网络的知识体现在网络连接的权值上,它是一个分布式矩阵结构;神经网络的学习体现在神经网络权值的逐步调整上。在数据挖掘中应用最多的是前馈式网络。它以感知器、反向传播模型、函数型网络为代表,可用于预测、模式识别等方面。

4. 粗糙集方法

粗糙集是一种刻画不完整性和不确定性的数学工具,能有效地分析和处理不精确、不一致、不完整等各种不完备信息,并从中发现隐含的知识,揭示潜在的规律。粗糙集的核心概念是不可区分关系以及上近似、下近似等。对于给定的一个信息表,粗糙集的方法是通过等价类的划分寻找信息表中的核属性和约简集,然后从约简后的信息表中导出分类/决策规则。对信息表进行属性约简,获得和原信息表具有相同信息分布的子表,提高了数据挖掘的效率,并且使得获得的知识更为简单、易于理解。属性约简是数据挖掘中数据预处理阶段的重要环节。

粗糙集理论具有良好的数学性质和可解释性,但在应用于实际数据时,还需要解决复杂度高、数据中的噪声等问题。

5. 统计分析方法

统计方法是从事物的外在数量上的表现去推断该事物可能的规律性,统计分析的本质是以数据为对象,从中获取规律,为人类认识客观事物,并对其发展趋势进行预测、决策和控制提供有效的依据。统计分析方法在数据挖掘中有许多应用,理论也最为成熟。常见的统计方法有回归分析、判别分析、差异分析、聚类分析、描述统计、相关分析和主成分分析等。

6. 生物智能算法

生物智能算法在优化与搜索应用中前景广阔,用于数据挖掘中,常把任务表示成优化或搜索问题,利用生物智能算法可以找到最优解或次优解。生物智能算法主要包括以下几个方面:

(1) 遗传算法。该方法是由 John Holland 于 1975 年提出的一种有效地解决最优化问题的方法,是一种基于生物进化理论的技术。其基本观点是"适者生存",用于数据挖掘中,则常把任务表示为一种搜索问题,利用遗传算法强大的搜索能力找到最优解,是一种仿生全局优化方法。遗传算法作用于一个由问题的多个潜在解(个体)组成的群体上,并且群体中的每个个体都由一个编码表示,同时每个个体均需依据问题的目标函数而被赋予一个适应值。遗传算法是多学科结合与渗透的产物,它广泛应用在计算机科学、工程技术和社会科学等领域。

(2) 蚁群算法。该方法由意大利学者 Dorigo M. 等在 20 世纪 90 年代初首先提出。

它是一种新型仿生类进化算法，是继模拟退火、遗传算法、禁忌搜索等之后的又一启发式智能优化算法。蚂蚁有能力在没有任何提示的情况下找到从巢穴到食物源的最短路径，并且能随环境的变化，适应性地搜索新的路径，产生新的选择。蚁群算法成功地应用于求解 TSP、二次分配、图着色、车辆调度、集成电路设计及通信网络负载等问题。

(3) 粒子群优化(PSO)算法。该方法是一种基于群体智能的随机优化算法，源于对鸟群或鱼群群体运动行为的研究。由于 PSO 算法概念简单、易于实现、调整参数少，现已广泛应用于许多工程领域。然而，粒子群优化算法具有易于陷入局部极值点、进化后期收敛慢、精度较差的缺点，为了克服粒子群优化算法的缺点，目前出现了大量的改进粒子群优化算法。

(4) 人工鱼群算法(AFSA)。该方法是李晓磊等于 2002 年提出的一种基于动物自治的优化方法，是集群智能思想的一个具体应用。它的主要特点是不需要了解问题的特殊信息，只需要对问题的解进行优劣的比较，通过各人工鱼个体的觅食、聚群和追尾等局部寻优行为，最终在群体中使全局最优解突显出来。该算法具有良好的求解全局极值的能力，收敛速度较快。

3.5　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、文本内容安全应用三个方面介绍了文本内容安全状态。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。文本表示介绍了布尔模型、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 3.4 节中，分别从分类分析、聚类分析以及关联分析三个方面介绍了网络信息内容常用的安全分析方法，为后续的网络信息内容处理提供解析方法及量化指标。本章内容重点是网络信息内容的预处理技术，难点是文本的安全分析。

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. 列举安全分析常用的方法有哪些？聚类和分类方法最主要的区别在于？