第 5 章

网络信息内容过滤


理论讲解


实验讲解

5.1 网络信息内容过滤概述

5.1.1 网络信息内容过滤的定义

随着 Internet 的飞速发展和在世界范围的普及，越来越多的数据库和信息不断加入网络，网络上的各种信息正以指数级的速度增长，Internet 已经发展为当今世界上资料最多、门类最全、规模最大的信息库和全球范围内传播信息的主要渠道。Internet 主要以超文本的形式呈现给用户各种各样的信息，构成一个异常庞大的具有异构性、动态性和开放性的分布式数据库。然而，在 Internet 极大丰富用户信息量的同时，用户也面临着信息过载和资源迷向的问题。Internet 上的信息过于庞杂，而且具有不稳定和变动快的特点，缺乏一个权威机构对这些信息进行全面的整理和归类。这一方面给用户发现信息、利用信息带来了不便，另一方面，无序、庞大的信息世界和成千上万的超链接，又常常使用户在查找其所需信息时感到力不从心。

早期解决这个矛盾主要采用信息检索技术。所谓信息检索，也就是我们熟知的搜索引擎，是指对有序化知识信息的检索查找，本质上是一种“人找信息”的服务形态，每次检索时要求用户一次性提交一个或几个查询关键词。当时的搜索引擎虽然算法简单，但数据库容量小，其查找信息效率较高，从 1994 年 4 月 Web Crawler 搜索引擎在网上正式发布并开始服务以来，搜索引擎已经成为发展最快、最引人注目的网络服务之一。

当前，搜索引擎正经历着从“数量累积阶段”向“质量精炼阶段”的变革。随着 Internet 上的信息数量呈指数级增长，大量信息垃圾也混杂其中。如何向用户提供质量好且数量适当的检索结果，成为搜索引擎技术发展的方向之一。由于大多数搜索引擎的搜集范围是综合性的，它们的机器抓取技术是尽其可能地把各类网页“抓”回来，经过简单的加工后存放到数据库中备检；另外，搜索引擎直接提供给用户的检索途径大都是基于关键词的布尔逻辑匹配，返回给用户的就是所有包括关键词的文献。这样的检索结果在数量上远远超出了用户的吸收和使用能力，让人感到束手无策。这也就是现在经常谈论的“信息过载”“信息超载”现象。其实，这就是这一代搜索引擎的突出缺陷：缺少智力，不能通过“学习”提高自身的检索质量。

针对网络的日益普及和信息量的爆炸增长而导致的信息过载、信息污染等问题，网络信息过滤技术作为筛选信息、满足用户需求的有效方法应运而生。网络信息过滤是根据

用户的信息需求,运用一定的标准和工具,从大量的动态网络信息流中选取相关的信息或剔除不相关信息的过程。也就是在设置好过滤条件后,在运行过程中一旦触发条件则将有关的信息拒之门外,而其他信息可以进入。网络信息过滤技术的目的就是让搜索引擎具有更多的"智力",让搜索引擎能够更加深入、更加细致地参与用户的整个检索过程。从关键词的选择、检索范围的确定到检索结果的精炼,帮助用户在浩如烟海的信息中找到和需求真正相关的资料。现在,Internet 上已经有很多有关这方面的研究,包括已经部署运行的信息过滤系统。这些都表明了信息过滤技术对于网络发展和应用的重要意义。

相比于信息检索技术,网络信息过滤技术是一种更系统化的方法,用来从动态的信息流中抽取出符合用户个性化需求的信息;而传统的信息检索则是从静态数据库中查找信息。信息过滤系统检查所有的进入信息流并与用户需求进行匹配计算,只将用户需要的文档送给用户。相比于传统的信息检索模式,信息过滤技术具有较高的可扩展性,能适应大规模用户群和海量信息;可以为用户提供及时、个性化的信息服务,具有一定的智能和较高的自动化程度。而如何能够更有效、更准确地找到自己感兴趣的信息,滤除与需求无关的信息,真正做到"各取所需",一直是基于 Internet 的网络信息领域的核心问题。网络信息过滤技术正在被越来越多地应用于 Web 空间,并获得了长足的发展,成为研究和工程实践的热点区域。自 20 世纪 90 年代开始,相关主题的国际会议不断举行,有力地推动了网络信息过滤技术的不断完善和进一步深入。

5.1.2 网络信息内容过滤的原理

现有的网络信息内容过滤方法较多,从过滤的手段来看,可以分为基于内容的过滤、基于网址的过滤和混合过滤三种。基于内容的过滤是通过文本分析、图像识别等方法阻挡不适宜的信息;基于网址的过滤是对认为有问题的网址进行控制,不允许用户访问其信息;混合过滤是将内容过滤与网址过滤结合起来控制不适宜信息的传播。从是否对网络信息进行预处理来看,信息过滤可以分为主动过滤和被动过滤两种。主动过滤是预先对网络信息进行处理,如对网页或网站预先分级、建立允许或禁止访问的地址列表等,在过滤时可以根据分级或地址列表决定能否访问;被动过滤是不对网络信息进行预处理,过滤时才分析地址、文本或图像等信息,决定是否过滤。无论采用哪种过滤方法,一个最简单的网络信息过滤系统一般包括四个基本组成部分:信源(Information Source)、过滤器(Filter)、用户(User)、用户需求模板(Profiles)。图 5-1 是信息过滤系统的一个简单结构图。

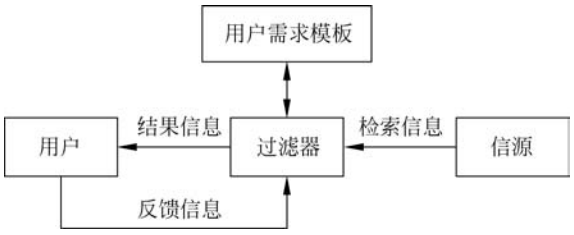

图 5-1 网络信息内容过滤基本原理

信源向过滤器提供信息，信息过滤器处于信源与用户之间，通过用户需求模板获取用户的兴趣信息，并据此检验信源中的信息，将其中与用户兴趣相关的信息递送给用户。反过来，用户也可以向信息过滤器发送反馈信息，以说明哪些信息的确符合他们的信息需求，通过这种交互行为使得过滤器不断进行学习，调整自身的过滤操作，进而能在以后提供更多更好满足用户兴趣的信息。

由于信息过滤的目的是向用户提供需要的信息。因此，网络信息内容过滤系统有以下最常见的特点。

(1) 过滤系统是为无结构化和半结构化的数据而设计的信息系统，它与典型的具有结构化数据的数据库系统不同。一个电子邮件就是半结构化数据的例子，它的头域有明确的定义，而它的正文却是半结构化的。

(2) 信息过滤系统主要用来处理大量的动态信息。非结构化数据这个词常用来作为它的同义词使用。一些多媒体信息系统包含图像、声音和视频信息。对于这些信息，传统的数据库系统没有进行很好的处理和表示。

(3) 过滤系统包含大量的数据。一些典型的应用基本上都要处理G字节以上的正文信息，其他媒介比这要大得多。

(4) 典型的过滤系统应用包含输入的数据流或是远程数据源的在线广播(例如新闻组、E-mail)。过滤也用来描述对远程数据库的信息进行检索，可用智能代理来实现。

(5) 过滤是基于对个体或群组的信息偏好的描述，也称为用户趣向。一般来说，这个用户趣向表示的是用户长久的信息偏好。

(6) 过滤是从动态的数据流中收集或去掉某些文本信息。

5.1.3　网络信息内容过滤的意义

网络信息内容过滤具有重要的现实意义和巨大的应用价值，主要体现在如下几个方面。

1. 改善 Internet 信息查询技术的需要

随着用户对信息利用效率要求的提高，以搜索引擎为主的现有网络查询技术受到挑战，网络用户的信息需求与现有的信息查询技术之间的矛盾日益尖锐，其不足主要有如下几方面。

(1) 在使用搜索引擎时，只要使用的关键词相同，所得到的结果就相同，它并不考虑用户的信息偏好和用户的不同，对专家和初学者一视同仁；同时，返回的结果成千上万、参差不齐，使得用户在寻找自己喜欢的信息时犹如大海捞针。

(2) 网络信息是动态变化的，用户时常关心这种变化。而在搜索引擎中，用户只能不断在网络上查询同样的内容，以获得变化的信息，这花费了用户大量的时间。因此，在现有情况下，传统的信息查询技术已经难以满足用户的信息需求，对信息过滤技术的研究日益受到重视，把信息过滤技术用于 Internet 信息查询已成为非常重要的研究方向。

2. 个性化服务的基础

个性化的实质是针对性，即对不同的用户采取不同的服务策略，提供不同的服务内容。个性化服务将使用户以最少的代价获得最好的服务。在信息服务领域，就是实现“信

息找人,按需要服务"的目标。既然是"信息找人",那什么信息找什么人就是关键。每个用户都有自己特定的、长期起作用的信息需求。用这些信息需求组成过滤条件,对资源流进行过滤,就可以把资源流中符合需求的内容提取出来进行服务。这种做法就叫作"信息过滤",信息过滤是个性化主动服务的基础。利用网络信息内容过滤技术有利于减轻用户的认知压力。它在为用户提供所需要信息的同时,着重剔除与用户不相关的信息,从而提高用户获取信息的效率;它根据用户信息需求的变化提供稳定的信息服务,能够节约用户获取信息的时间,从而极大地减轻用户的认知负担,起到减压阀的作用。网络信息过滤对个性化信息服务起到了巨大的推动作用。在个性化信息服务中,最重要的是收集和分析用户的信息需求。由于信息过滤的反馈机制具有自我学习和自我适应的能力,可以动态地了解用户兴趣的变化,因此可以越来越明确、具体地掌握用户的信息需求,从而为用户提供更有针对性的信息。在协作过滤系统中,还可以根据用户之间的相似性来推荐信息,从而有可能为用户提供新的感兴趣的信息,拓宽用户的视野。通过网络信息过滤,可以减少不必要的信息传递,节约宝贵的信道资源。

3. 维护我国信息安全的迫切需要

网络为信息的传递带来了极大的方便,也为机密信息的流出和对我国政治、经济、文化等有害信息的流入带来了便利。发达国家通过网络进行政治渗透和价值观、生活方式的推销,一些不法分子利用计算机网络复制并传播一些色情的、种族主义的、暴力的封建迷信或有明显意识形态倾向的信息。我国80%的网民在35岁以下,80%的网民具有大专以上文化学历,而这两个80%正是国家建设发展的主力军。所以,我国的信息安全问题已迫在眉睫,必须引起高度警惕和重视,而信息过滤是行之有效的防范手段。目前主要通过过滤软件及分级制度对来往信息尤其是越境数据流进行过滤,将不宜出口的保密或宝贵信息资源留在国内,将不符合国情或有害信息挡在网络之外,其中用得较多的为Internet接收控制软件和Internet内容选择平台(Platform for the Internet Content Selection,PICS)。

随着网络不良信息的泛滥,信息过滤作为解决不良信息问题的技术手段,更是受到社会各方面的广泛关注。过滤网络不良信息是信息过滤的重要的应用之一。通过分级类目、关键词、规则等描述用户的信息需求,以分级、URL地址列表、自动文本分析等方法来过滤不良信息,同时运用一些人工干预的方法提高信息过滤的效率,在保护网络用户尤其是未成年用户免受不良信息侵扰方面发挥了很好的作用。

4. 信息中介(信息服务供应商)开展网络增值服务的手段

信息中介行业的发展要经过建立最初的客户资料库、建立标准丰富档案内容和利用客户档案获取价值三个阶段。其中第一阶段和第三阶段的主要服务重点都涉及信息过滤服务。过滤服务过滤掉客户不想要的推销信息,信息中介将建立一个过滤器以检查流入的带有商业性的电子邮件,然后自动剔除与客户的需要和偏好不相符的不受欢迎的信息。客户可提前指定他们想经过过滤服务得到的信息或经过过滤服务排除出去的任何种类的经销商或产品。对于不受欢迎的垃圾信息,信息中介将会在客户得到之前把它们过滤掉。

利用网络信息过滤,可以对网络信息的流量、流向和流速进行合理的配置,使网络更加顺畅。而对于用户来说,信息过滤由于剔除了大量不相关信息的流入,因此可以避免塞

车现象。在网络环境下，尽量减少无效数据的传输对于节省网络资源、提高网络传输效率具有十分重要的意义。通过信息过滤，可减少不必要的信息传输，节省费用，提高经济效益。

5.2　网络信息内容过滤技术的分类

面对纷繁的过滤系统，按照单一的标准是无法准确区分的，下面按照如下三个标准对网络信息内容过滤技术进行分类。

5.2.1　根据过滤方法分类

1. 基于内容的过滤

基于内容的过滤(Content Based Filtering)又叫认知过滤，是利用用户需求模板与信息的相似程度进行的过滤，能够为用户提供其感兴趣的相似的信息，但不能为用户发现新的感兴趣的信息。在反馈机制的作用下，用户的信息需求处于循序渐进的变化过程中。基于内容的过滤首先要将信息的内容和潜在用户的信息需求特征化，然后再使用这些表述，职能化地将用户需求同信息相匹配，按照相关度排序把与用户信息需求相匹配的信息推荐给用户，其关键技术是相似性计算。其优点是简单、有效；缺点是难以区分资源内容的品质和风格，而且不能为用户发现新的感兴趣的资源，只能发现和用户已有兴趣相似的资源。

2. 协作过滤

协作过滤(Collaborative Filtering)又叫社会过滤，是利用用户需求之间的相似性或用户对信息的评价进行的过滤。对于价值观念、思想观点、知识水平或需求偏好相同或相似的用户，他们的信息需求往往也具有相似性。基于这一思路，通过比较用户需求模板的相似程度或者根据用户对信息的评价而进行的过滤，既可以为用户提供正感兴趣的信息，又可以提供新的感兴趣的信息。在这种系统中，用户的信息需求有可能呈现跃进式的变化。

协作过滤支持社会上个人间和组织间的相互关系，并将人们之间的推荐过程自动化。一个数据条款被推荐给用户，是基于他同其他有相似兴趣用户的需求相关。协作过滤推荐的核心思想是用户会倾向于利用具有相似意向的用户群的产品，因此，它在预测某个用户的利用倾向时是根据一个用户群的情况而决定的。可见，协作过滤法是找出一群具有共同兴趣的使用者形成社群，也就是有某些相似特性成员的集合，通过分析社群成员共同的兴趣与喜好，再根据这些共同特性推荐相关的项目给同一社群中有需求的成员。其优点是对推荐对象没有特殊要求，能处理非结构化的复杂对象，并且可以为用户发现新的感兴趣的资源，这种过滤类型对那些不是很清楚自己的信息需求或者表达信息需求很困难的用户非常重要。其缺点是存在两个很难解决的问题：一是稀疏性问题，即在系统使用初期，由于系统资源还未获得足够多的评价，系统很难利用这些评价来发现相似的用户；二是可扩展性问题，即随着系统用户和信息资源的逐渐增长，其可行性将会降低。协同过

滤方法只考虑了用户评分数据,忽略了项目和用户本身的诸多特征,如电影的导演、演员和发布时间等,用户的地理位置、性别、年龄等,如何充分、合理地利用这些特征,获得更好的推荐效果,是基于内容推荐策略所要解决的主要问题。

这两类过滤方法侧重不同,各有优点,综合使用这两类技术会给网络信息内容过滤带来更好的效果。

5.2.2 根据操作的主动性分类

1. 主动过滤

主动过滤(Active Filtering)系统主动为网络用户寻找他们需要的信息。这类系统可以在一个较大范围内或局部范围内帮助用户收集同用户兴趣相关的信息,然后主动从Web上为其用户推送相关的信息。因特网上所谓的"推送技术(Pushing Technology)"就是这个范畴内的应用。在有些主动信息过滤系统中,预先对网络信息进行处理,例如,对网页或者网站预先分级、建立允许或禁止访问的地址列表等,在过滤时可以根据分级标记或地址列表决定能否访问。这类系统有BackWeb。

2. 被动过滤

被动过滤(Passive Filtering)系统不对网络信息进行预处理,当用户访问时才对地址、文本或图像等信息进行分析,以决定是否过滤及如何过滤。这类系统是针对一个相对固定的信息源过滤掉其中用户不感兴趣的信息。例如信息源可以是用户的电子邮件、某些固定看的新闻组等,而主动型系统要主动地在可能的范围内寻找信息源。这类系统一般都是根据用户兴趣将信息源中新到的信息根据相关程度按从大到小的顺序排给用户,或根据某一门限值将系统认为用户不感兴趣的信息提前过滤掉。这类系统有GHOSTS、CiteSeer。

5.2.3 根据过滤位置分类

1. 上游过滤

用户需求模板存放在网络服务器端或者代理端上。一般来说,为了减小服务器端和客户端的负荷,过滤系统也可能处在信息提供者与用户之间的专门的中间服务器上,这种情况也叫作中间服务器过滤。中间服务器如同一个大型的网络缓存器,Internet信息内容只有经过它的过滤才能进入本地系统或局域网,而本地信息也要经过它的中转才能传递出去。服务器端采用隐含式方法获取用户信息需求,过滤系统通过记录用户的行为来获得用户的信息需求,如用户在指定页面的停留时间、用户访问页面的频率、是否选择保存数据、是否打印、是否转发数据等对信息项的反应都能作为用户兴趣的标志。一般上游过滤的优点是不仅支持基于内容的过滤,也支持协作过滤;缺点是模板不能用于不同的网络应用中,容易受到干扰的影响,所以这种方法通常用作下游过滤的补充。

2. 下游过滤

用户需求模板存放在客户端上,也称为客户端过滤。采用显式方法获取用户信息需求的过滤系统,通常要求用户填写一个描述他们兴趣领域需求的表或者要求用户根据提

供的特征项构造自身对特定领域信息需求的描述模型。用户根据自身需要设置一定的限定条件，将不感兴趣的信息排除在外。其优点是模板可用于不同的网络应用；缺点是只能实现基于内容的过滤。系统要求用户提供自身明确的信息，使系统能够把用户与用户原型模型相关联。所谓原型模型，是指一组用户的默认信息，将对用户原型模型上的隐含式推测与用户提供的明确知识相结合，可得到更好的表示用户信息需求的用户模板。

5.2.4　根据过滤的不同应用分类

网络信息内容过滤技术还可以根据过滤的不同应用进行分类，具体可分为如下几种类型。

1. 专门过滤软件

这是为过滤网络信息而专门开发的软件，一般要加载到网络应用程序中，根据预先设定的过滤模板扫描、分析网络信息并阻挡不适宜的信息。专门过滤软件又可以分为专用过滤软件和通用过滤软件两种。前者只能过滤某种网络协议的信息，如网页过滤软件、邮件过滤软件、新闻组过滤软件等；或者只能在某种网络应用中起作用，如儿童浏览器、儿童搜索引擎、广告过滤软件等。后者能对多种网络协议或应用起作用，如 NetNanny 可以过滤网页、电子邮件、网络聊天的信息，除此之外 NortonInternetSecurity 还可以过滤 ICQ、FTP 和新闻组的信息。目前用得比较多的是通用过滤软件。

2. 网络应用程序

有些网络应用程序如 Web 浏览器、搜索引擎、电子邮件、新闻组等附有过滤的功能，可以设置过滤不适宜的信息。如 IE 的内容分级审查功能，用户通过设置黑名单、白名单或组合使用各种支持 PICS 的分级标记进行过滤，该功能具有过滤成本低、使用方便的特点。典型的如浏览器端过滤，这种过滤方式使用存储一些已知的散布不良网站的 IP 地址、URL 地址的数据库，在浏览器进行访问时，将访问地址与数据库中的 IP 地址、URL 地址等信息进行匹配，如果浏览器需要访问的地址在数据库中是处于需要限制的内容，那么在浏览器请求访问的时候，对其进行限制，达到过滤的效果。过滤性能伴随数据库中的 IPP 地址、URL 地址数量以及准确性的提升而提升。

3. 其他过滤工具

如防火墙、代理服务器等，可以通过对源地址、目标地址或端口号的限制，防止子网不适宜的信息流出。运用 IP 地址或 URL 地址进行过滤有路由器端过滤方式。这种方式将过滤规则放置在路由器端，在路由器的“安全设置”的“IP 地址过滤”中可以设置 IP 地址、禁止访问的端口和协议等。使用路由器端的 IP 地址过滤，反应速度较快，可以对端口、协议等进行设置，可限制更多网站。但是路由器设置较为复杂，地址等一般不全面，不能普及。根据 IP 地址、URL 地址进行网页过滤是一种非常有效的手段，在 IP 库与 URL 库非常全面时，能够准确地识别需要过滤的网址。但是这种方式有一定的局限性，在当今网站层出不穷的情况下，缺少对于未知网址的发现，某些不法分子经常修改网址 IP 及端口设置，使用多级代理变换网址形式，对 IP 过滤造成了影响。

5.3 网络信息内容过滤的一般流程

1. 网络信息过滤的一般流程

为便于理解,首先给出网络信息过滤的一般流程,如图5-2所示。

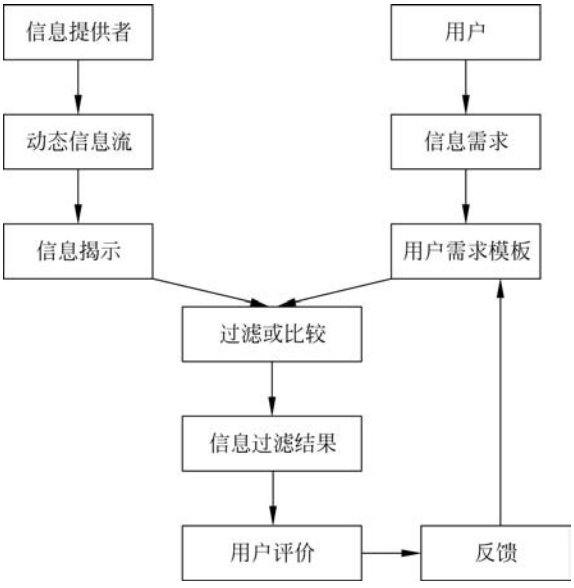

图5-2 网络信息过滤一般流程

从图5-2中可以看出,用户通过网络进行工作、学习、生活,从而产生了大量信息。用户的信息需求必须以计算机能够识别的形式揭示出来,这就是用户需求模板(Profile,也叫过滤模板)。对于用户需求模板,可以是正向的,也可以是反向的,也就是说既可以揭示用户希望得到的信息,也可以描述用户希望剔除的信息。在系统中,对动态的网络信息集不做预处理,只是当信息流经过系统时才运用一定的算法把信息揭示出来。匹配算法和用户需求模板的描述方法、信息的揭示方法是相互联系的,常用的匹配模型有布尔模型、向量空间模型、概率模型、聚类模型、基于知识的表示模型以及混合模型等,主要任务是剔除不相关的信息,选取相关的信息并按相关性的大小提供给用户。

为了提高信息过滤的效率,系统还根据用户对过滤结果的反应,即通过反馈机制作用于用户和用户需求模板,使用户逐渐清晰自己的信息需求,使得用户对需求模板的描述也会越来越明确、具体。图5-2中的反馈模块主要用于处理用户的反馈信息并依据反馈信息进一步精化用户模型,保存以便下一次用户注册登录时直接读取到精化后的模型。用户对返回的文档集进行评估,由系统根据这些反馈信息进一步修改用户兴趣文件,以利于下一次的过滤。在整个系统中,用户需求模板的生成、信息揭示、匹配算法和反馈机制是最为关键的部分。在实际应用中,往往会在这些关键部分进行必要的人工干预,如对动态的信息流做预处理、人工修改用户需求模板等。

2. 网络文本信息过滤模型

参考图5-2网络信息过滤的一般模型，可以创建一个基于Web的文本信息过滤模型，如图5-3所示。

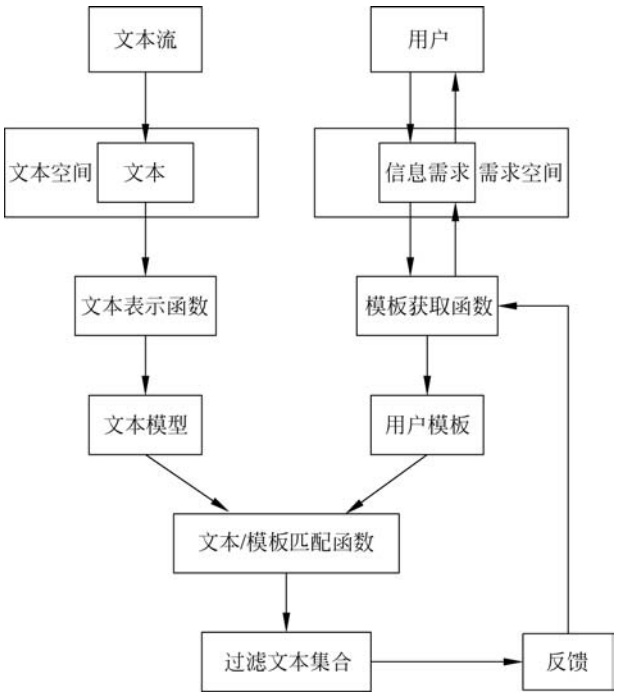

图5-3　网络文本信息过滤模型

从图5-3来看，文本信息过滤模型中主要包含文本表示模块、文本过滤匹配模块、用户(兴趣)模板生成模块、反馈模块等。其中，文本表示模块主要针对采集到的信息提取其中的特征信息，按照一定的格式来描述，然后作为输入信息传递给过滤匹配模块；用户模板生成模块是依据用户对信息的需求和喜好来生成，它根据用户提供的学习样本或主动跟踪用户的查询行为建立用户兴趣的初始模板，再根据用户反馈模块不断更新用户模板；文本过滤匹配模块就是将用户兴趣模板与信息表示模块中的信息分析表示的结果按照一定的算法进行匹配，并按照匹配算法决定将要传递给用户的相关信息项；用户得到文本过滤的结果后，对其进行评价并反馈给用户模块，用户模块通过不断跟踪学习用户兴趣的变化及用户反馈来调整甚至更改用户需求表达，以达到不断实现正确过滤无用信息的目的。以下简要介绍模型中各部分的主要技术。

(1) 文本表示。将Web中的有效文本信息内容提取出来，对于中文文本过滤来说，涉及中文分词、停用词处理、语法语义分析等过程。常用的方法是建立文本的布尔模型、向量空间模型和概率模型等。

(2) 用户模板的建立。用户模板空间常按照倒排索引的方式存储用户信息，建立用户模板的方式有建立关键字表和示例文本，而常用的技术有建立向量空间模型、预定义关键字、层次概念集和分类目录等。

(3) 用户模板与文本的匹配。最常用的方法有布尔模型、向量空间模型和概率模型。

(4) 用户反馈。用户反馈分为确定性反馈和隐含性反馈。确定性反馈指的是二元(是或否)反馈,另外还有分级打分的方法。利用这些反馈信息,应用机器学习方法,完善用户模板。

综合以上介绍分析,可以将网络文本信息内容过滤的工作概括为两个方面:一是建立用户需求模型,即用户模板,用于描述用户对于信息的具体需求,建立用户需求模型的主要依据是用户提交的关键词、主题词或示例文本;二是匹配技术,即用户模板与文本的匹配技术。简单地讲,文本过滤模型就是根据用户的查询历史创建用户需求模型,将信息源中的文本有效表示出来,然后根据一定的匹配规则,将文本信息源中可以满足用户需求的信息返回给用户,并根据一定的反馈机制,不断地调整改进用户需求模型,以期获得更好的过滤结果。从技术角度来看,文本信息过滤的关键技术是获得用户信息需求(用户模板的建立)和解决信息过滤算法,即信息过滤技术的研究应当集中在解决用户模板的表示及根据模板对文本流进行评价(ranking)的方法上。为提高信息过滤系统的性能,应加强对过滤匹配算法和用户模型的研究与实践。

3. 实例分析

下面将以 Websense 为例,介绍网络信息内容过滤的实际应用。Websense 是全球知名的过滤软件开发商,有 18000 多家公司、学校、图书馆和政府部门在使用 Websense 公司的过滤软件。软件主要用于企业网络管理,防止员工滥用网络,经过调整后也可用于网吧、图书馆等部门。软件由主数据库、Enterprise 应用程序、报表及三台用户机组成。Websense Enterprise 过滤系统如图 5-4 所示。

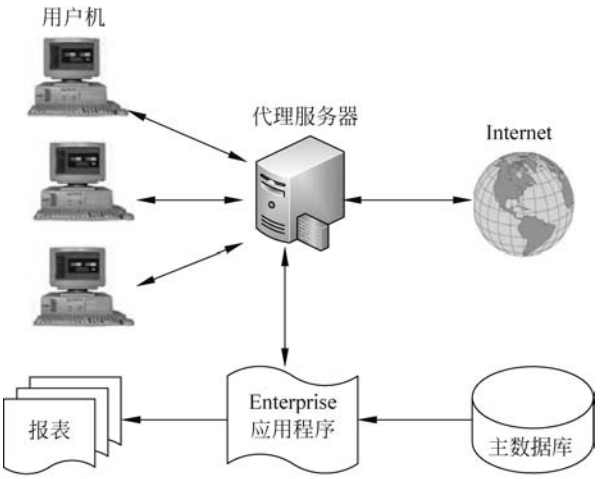

图 5-4 Websense Enterprise 过滤系统示意图

1) Websense 主数据库

Websense 主数据库存储了 400 多万个网站、10 亿个网页。这些网页涉及英、法、德、日、西等 44 种语言,根据不同的内容归入 Websense 分级体系的 31 个一级类目和 50 多个子类目中,号称是世界上最大最精确的采用自动和人工分级相结合的分级网址数据库。

主数据库安装在用户的代理服务器上，与 Enterprise 应用程序结合才能过滤网络信息。为了建立和维护这个庞大的数据库，Websense 公司有专门的工具收集网页。网页收集回来后利用自动分类器进行分级，对于分类器无法确定的类目再由人工分级，分级的结果保存在 Websense 的分级数据库中。用户代理服务器上的 Enterprise 应用程序每天都会自动从分级数据库中下载最新的内容，更新主数据库的记录。由于网络信息处于动态变化过程中，为了保证网页分级的有效性，Websense 有专门的工具定期回访网页，对内容有变化的网页进行重新分级。

2）Websense Enterprise 应用程序

Enterprise 应用程序是 Websense 过滤软件直接与用户交互的部分，也是整个系统的核心组成部分。它可以与防火墙、代理服务器整合，在 Windows NT/2000、Sun Solaris 和 Linux 系统中运行。它能够根据用户定制的过滤模板调用主数据库的数据过滤不适宜的信息，并将处理的结果传递给报表程序。由于 Websense 分级体系的类目众多而且周详，除了不良信息的类目外，还有许多类目是从防止员工滥用网络的角度而设立的，管理人员可以根据不同的用户、组、部门、工作站、IP 地址或网络设置不同的过滤模板，而且还可以为每一类目分别设置以下内容。

（1）时基限额。利用时基限额，允许用户在适当的时间内访问与工作无关的类目。例如，每天允许访问银行及购物站点的时间不超过 20 分钟。

（2）继续或延迟。用户可以选择“继续”浏览不允许的类目，或者选择“延迟”至在工作时间外浏览。

（3）设定时段。按类目设置过滤的时段。例如，每天的工作时间内禁止访问购物网站，而其他时间则可以访问。Websense Enterprise 应用程序可以通过白名单限制用户访问的范围，采用关键词列表阻挡不适当的内容，根据主文件名或扩展名进行过滤，还支持对网络聊天的限制。

5.4　网络信息内容过滤模型

从前面章节中可以看出内容过滤模型是网络信息内容过滤系统中的核心模块。在实际应用中，常用的过滤模型一般包括布尔模型、向量空间模型和神经网络模型。根据过滤系统的应用对象不同，其过滤效率也不同。下面将对这些模型做简要介绍。

5.4.1　布尔模型

布尔模型是基于特征项的严格匹配模型。首先建立一个二值变量的集合，这些变量对应着信息源的特征项。如果在信息源中出现相应的特征项，则特征变量取 True，否则特征变量取 False。查询是由特征项和逻辑运算符 AND、OR 和 NOT 组成的布尔表达式。信息源与查询的匹配规则遵循布尔运算的法则。根据匹配规则将信息源分为两类：相关类和不相关类。由于匹配结果的二值性，所以无法对结果集进行相关性排序。

布尔模型实现简单，检索速度快，易于理解，在许多商用的过滤系统中得到了应用。但是这种传统的布尔过滤技术也存在着一些不足之处。

(1) 原始信息表示不精确。布尔模型仅仅以特征项在原始信息中出现与否的布尔特性来表示原始信息,忽略了不同特征项对信息内容贡献的重要程度,容易造成结果的冗余。

(2) 基于布尔运算法则的匹配规则过于严格,容易造成漏检。严格且缺乏灵活性的布尔过滤规则往往会导致仅仅因为一个条件未满足的文档被漏检。

(3) 布尔模型匹配结果的二值性导致系统无法按结果信息的相关性大小为用户提供信息。

为了克服传统布尔模型的缺陷,人们对其进行了改造,引入了权重来表示特征项对文档的贡献程度,形成了所谓的加权布尔模型,即拓展的布尔模型(Extended Boolean Model)。

5.4.2 向量空间模型

向量空间模型已被人们普遍认为是一种非常有效的检索模型。它具有自然语言界面,易于使用。同样,向量空间模型也可以应用到信息过滤系统中。在以向量空间模型构造的信息过滤系统中,用户模板和原始信息均被表示成 n 维欧氏空间中的向量,用它们之间的夹角余弦作为相似性的度量。运用向量空间模型构造信息过滤系统主要包括四个方面的工作:

(1) 给出原始信息的向量表示;

(2) 给出用户模板的向量表示;

(3) 计算原始信息和用户模板之间的相似度,二者的相似度通常用原始信息向量和用户模板向量之间夹角的余弦值来衡量;

(4) 将与用户模板之间相似度大于给定阈值的原始信息提供给用户,并获得用户的反馈。

向量空间模型的优点在于将原始信息和用户模板简化为项及项权重集合的向量表示,从而把过滤操作变成向量空间上的向量运算,通过定量的分析,完成原始信息和用户模板的匹配。

向量空间模型的缺点在于存在信息在向量表示时的项与项之间线性无关的假设,在自然语言中,词或短语之间存在十分密切的联系,即存在“斜交”现象,很难满足假定条件,这对计算结果的可靠性造成一定的影响。此外,将复杂的语义关系归结为简单的向量结构,丢失了许多有价值的线索。因此,有许多改进的技术,以获取深层潜藏的语义结构,如潜在语义索引方法就是对向量空间模型的一种有效改进。

5.4.3 神经网络模型

神经网络模型(Neural Network Model)模拟人脑对信息的处理方式,用该模型过滤信息的基本思想是在其内部存储可行模式的整个集合,这些模式可被外部暗示唤起,即使“外部”提供的资料不足,也可以在其内部进行构造。当给系统输入一个文本的特征向量时,可通过神经网络存储的内部信息对此文本进行主题判断,即神经网络的输入为文本的特征向量,输出为用户给出的评价值。经过训练的网络模型通过将不同文本的特征向量

映射为大小不等的评价来实现主题区分的目的。

5.5　网络信息内容过滤的主要方法

分类是一个有指导的学习过程，也是网络信息内容过滤中的一个重要技术方法。其特点是根据已经掌握的每类若干样本（训练数据）的数据信息，总结出分类的规律，建立判别公式和判别规则。然后，当遇到待分类的新样本点（测试数据）时，只需根据总结出的判别公式和判别规则，就能确定该样本所属的类别。

实际上，基于内容的文本过滤在不考虑学习和自适应能力时是一个分类过程，如TREC中的Batch（自动过滤，结果不排序）和Routing（自动过滤，结果排序）过滤任务。其中，过滤的主题（用户需求）相当于分类的类别，过滤的检出准则相当于分类的判别规则，而判断某文档跟哪些主题相关的过程等价于判别文档所属的类别的过程。对于自适应过滤任务（Adaptive Filtering），其基本框架仍然是一个类似文本分类的判别过程。不同之处主要有两点：一是训练样本很少，几乎没有训练过程；二是在过滤过程中需要根据用户的反馈进行自适应学习，不断自我调整以实现边学习边提高的目的。后者是自适应过滤研究的重点，但是，作为核心的过滤算法仍然是一个分类算法。

过滤算法的选择是影响文本过滤效果好坏的重要因素。分类技术涉及很多领域，包括统计分析、模式识别、人工智能、神经网络等。由于过滤与分类、检索技术的共通性，上述领域的研究成果同样可以应用到网络信息内容过滤中。这些方法大致可以分为统计方法和逻辑方法。

5.5.1　统计方法

统计判别方法是统计分析领域的过滤和分类算法的总称，在网络信息内容过滤的实际应用中，常用的方法主要有向量中心法、相关反馈法（如Rocchio法）、K近邻（K-Nearest Neighbor，KNN）法、贝叶斯法、朴素贝叶斯（Naive Bayes）法和贝叶斯网络（Bayes Nets）、多元回归模型（Multivariate Regression Models）、支持向量机（Support Vector Machines）以及概率模型（Probability Model）等。

1. 向量中心法

向量中心法是建立在向量空间模型基础上的。该方法通过计算新到来的文档与表示过滤主题的用户兴趣（向量中心）之间的夹角余弦值：

$$\mathrm{sim}(D_1,D_2)=\cos\theta=\frac{\sum_{k-1}^{n}w_{1k}w_{2k}}{\sqrt{\left(\sum_{k-1}^{n}w_{1k}^2\right)\left(\sum_{k-1}^{n}w_{2k}^2\right)}} \tag{5-1}$$

或者向量内积

$$\mathrm{sim}(D_1,D_2)=\sum_{k-1}^{n}w_{1k}\cdot w_{2k} \tag{5-2}$$

来判断文档是否跟该主题相关。由于这种方法简单实用，因而在信息过滤、信息检索、文

本分类等多个领域得到了广泛应用。

2. 相关反馈法

Rocchio法是一个在信息检索中广泛应用于文本处理与过滤等业务的算法,它是一种基于相关反馈(Relevance Feedback)的、建立在向量空间模型上的方法。它用TF-IDF方法来描述文本,其中TF(w_i,d)是词w_i在文本d中出现的频率,DF(w_i)是出现w_i的文本数。该方法中可以选择不同的词加权方法、文本长度归一化方法和相似度测量方法以取得不同的效果。Rocchio法首先通过训练集求出每一个主题的用户兴趣向量,其公式如下:

$$\boldsymbol{C}_j = \alpha \frac{1}{|\boldsymbol{C}_j|}\sum_{\boldsymbol{d}\in \boldsymbol{C}_j}\frac{\boldsymbol{d}}{\|\boldsymbol{d}\|} - \beta \frac{1}{|\boldsymbol{D}-\boldsymbol{C}_j|}\sum_{\boldsymbol{d}\in \boldsymbol{D}-\boldsymbol{C}_j}\frac{\boldsymbol{d}}{\|\boldsymbol{d}\|} \tag{5-3}$$

其中,$\boldsymbol{C}_j$是主题的用户兴趣,α、β反映正反训练样本对$\boldsymbol{C}_j$的影响,$\boldsymbol{d}$是文本向量,$\|\boldsymbol{d}\|$是该向量的欧氏距离,$\boldsymbol{D}$是文本总数。

若以余弦计算相似度,则判别文本$\boldsymbol{d}$是否跟主题$\boldsymbol{C}_j$相关的公式为

$$\begin{aligned} H_{\text{TF-IDF}}(\boldsymbol{d}) &= \operatorname{argmax}\cos(\boldsymbol{C}_j,\boldsymbol{d}) = \operatorname{argmax}\frac{\boldsymbol{C}_j}{\|\boldsymbol{C}_j\|}\cdot\frac{\boldsymbol{d}}{\|\boldsymbol{d}\|} \\ &= \operatorname*{argmax}_{\boldsymbol{C}_j\in\boldsymbol{C}}\frac{\sum_{i=1}^{n}\boldsymbol{C}_j^{(i)}\boldsymbol{d}^{(i)}}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{C}_j^{(i)})^2(\boldsymbol{d}^{(i)})^2}} \end{aligned} \tag{5-4}$$

其中,n为每个文档的特征项(词)的个数。式(5-4)中忽略了$\boldsymbol{d}$的长度,因为它不影响argmax的结果。Rocchio法实现起来较为容易,但是它需要事先知道若干正负样本,受训练集合的影响较大,有时会导致性能下降。

3. *K* 近邻法

K 近邻法的原理也很简单。给出未知相关主题的文本,计算它与训练集中每个文本的距离,找出最近的k篇训练文档,然后根据这k篇来判断未知文本相关的主题。可以选择出现在这k个邻居中相关的文本与未知文本的相似度,值最大的主题就被判定为未知文本相关的主题,这就是最近邻法。最近邻法不是仅仅比较与各主题类均值的距离,而是计算和所有样本点之间的距离,只要有距离最近者就归入所属主题类。为了克服最近邻法错判率较高的缺陷,*K* 近邻法不是仅选取一个最近邻进行判断,而是选取k个近邻,然后检查它们相关的主题,归入比重最大的那个主题类。

4. 贝叶斯法

(1) 朴素贝叶斯法。朴素贝叶斯算法在机器学习中有着广泛的应用。其基本思想是在贝叶斯概率公式的基础上,根据主题相关性已知的训练语料提供的信息进行参数估计,训练出过滤器。进行过滤时,分别计算新到文本跟各个主题相关的条件概率,认为文本跟条件概率最大的主题类相关。其计算公式如下:

$$P(C_j \mid d;\hat{\theta}) = \frac{P(C_j \mid \hat{\theta})P(d_i \mid C_j;\hat{\theta}_j)}{P(d_i \mid \hat{\theta})} \tag{5-5}$$

式(5-5)中,等式右边的概率均可根据训练语料运用参数估计的方法求得。朴素贝叶斯法

是在假设各特征项之间相互独立的基本前提下得到的。这种假设使得贝叶斯算法易于实现。尽管这个假设与实际情况不相符，但实际应用证明，这种方法应用于信息过滤中是比较有效的。

(2) 贝叶斯网络。Heckerman 和 Sahami 分别提出了对贝叶斯网络的改进方法。贝叶斯网络的基本思想是取消纯粹贝叶斯方法中关于各特征之间相互独立的假设，而允许它们具有一定的相关性。K-相关贝叶斯网络是指允许每个特征有至多 k 个父节点 f，即至多有 k 个与之相关的特征项的贝叶斯网络。朴素贝叶斯则是贝叶斯网络的一个特例，也被称为 0-相关贝叶斯网络。

5. 多元回归模型

多元回归模型运用了线性最小平方匹配(Linear Least Square Fit)的算法。通过求解输入-输出矩阵的线性最小平方匹配问题，得到一个回归系数矩阵作为过滤器。具体来讲就是求出一个矩阵 $\boldsymbol{X}$ 使得 $\|\boldsymbol{E}\|_F=\left(\sum_{i=1}^{N}\sum_{j=1}^{i}e_{ij}^2\right)^{1/2}$ 最小，其中 $\boldsymbol{E}=\boldsymbol{AX}-\boldsymbol{B}$。在信息过滤中 $\boldsymbol{A}$ 是输入矩阵，是训练集文本的词-文本矩阵(词在文本中的权重)，$\boldsymbol{B}$ 是输出矩阵，是训练集文本的文本-相关主题矩阵(主题在文本中的权重)。求得的矩阵 $\boldsymbol{X}$ 是一个关于词和主题的回归系数矩阵，它反映了某个词在某一主题类中的权重。在过滤过程中，用相关主题未知的文本的描述向量 $\boldsymbol{a}$ 与回归系数矩阵 $\boldsymbol{X}$ 相乘就得到了反映各个主题与该文本相关度的矩阵 $\boldsymbol{b}$。相关度最大的主题就是该文本所相关的主题。

6. 支持向量机

支持向量机算法是 Vapnik 提出的一种统计学习方法，它基于有序风险最小化归纳法(Structural Risk Minimization Inductive Principle)，通过在特征空间构建具有最大间隔的最佳超平面，得到两类主题之间的划分准则，使期望风险的上界达到最小。支持向量机在文本分类领域得到了比较成功的应用，成为表现较好的分类技术之一，其主要缺点是训练过程效率不高。N. Cancedda 等人将这种方法用于解决自动信息过滤问题，同样取得了较好的效果。

7. 概率模型

概率模型是 Stephen Roberson 等人提出的信息检索模型，该模型同样可以用于信息过滤。其主要特点是认为文档和用户兴趣(查询)之间是按照一定的概率相关，因而在特征加权时融入了概率因素，同时也综合考虑了词频、文档频率、逆文档长度等因素。

5.5.2　逻辑方法

逻辑方法就是研究怎样学习主题过滤规律的方法，该方法认为知识就是过滤。逻辑方法比较适应于具有离散变量的样本。对于连续性的变量，常常采用一些离散化的手段把它们转化成离散值。传统的逻辑方法主要包括基于覆盖的 AQ 家族算法，以信息熵为基础的 ID3 决策树算法以及基于 Rough 集理论的学习算法。

1. ID3 决策树(Decision Tree)算法

ID3 是 Quinlan 于 1986 年提出的一种重要的归纳学习算法，在机器学习中有广泛的应用，它从训练集中自动归纳出决策树。在应用时，决策树算法基于一种信息增益标准来

选择具有信息的词,然后根据文本中出现的词的组合判断相关性。决策树有以下三个特点:

(1) 使用一棵过滤决策树表示学习结果;

(2) 决策树的每个节点都是样本的某个属性,采用信息熵作为节点的选择依据;

(3) 采用了有效的增量学习策略。

2. AQ11 算法

AQ11 使用了逻辑语言来描述学习结果。整个学习过程就是一个逻辑演算过程:

$$E_p \wedge \neg E_N = (e_1^+ \vee e_2^+ \cdots e_k^+) \wedge \neg (e_1^- \vee e_2^- \cdots e_m^-)$$
$$= (e_1^+ \wedge \neg e_1^- \wedge \neg e_2^- \wedge \cdots \wedge \neg e_m^-) \vee \cdots (e_k^+ \wedge \neg e_1^- \wedge \neg e_2^- \wedge \cdots \wedge \neg e_m^-) \tag{5-6}$$

其中,$e_1^+ \in E_p$ 表示正例样本集合中的一个正例样本,$e_1^- \in E_N$ 表示反例样本集合中的一个反例样本,然后使用分配率和吸收率对式(5-6)进行简化。

3. 基于 Rough 集理论的逻辑学习算法

Rough 集是波兰数学家 Pawlak 提出的一种不确定性知识的表示方法,后来被人们用作数据约简。数据约简是指去除那些对于过滤不起作用的元素,分为只删除属性值的值约简,以及可以删除整个属性的属性约简。数据约简可以在保持相关主题一致的约束下大大简化样本数据,最终使用很少的几条逻辑规则就能描述过滤规则。

5.6 网络信息内容过滤典型系统

本节针对互联网中信息需求个性化的特点,首先介绍一种多 Agents 信息过滤系统模型。接下来,从中文网页信息内容过滤系统的需求分析出发,讨论基于文本匹配的过滤系统的设计实现。

5.6.1 基于多 Agents 的过滤系统

由于 Internet 信息空间的分布性、异构性,人们对信息的需求体现出个性化的特征。本节介绍一种采用智能 Agents 技术的多 Agents 信息过滤系统模型,该模型借助 5.5 节介绍的过滤算法对系统检索得到的结果进行信息过滤,按照用户需求过滤掉无关信息,重视用户反馈,以用于进一步优化用户的检索;同时,建立个性化知识库,该知识库可使得检索过滤系统能够自学习用户兴趣,为信息过滤自动化过程提供事实依据,增强自动检索功能。

1. 智能 Agents 技术特点

智能 Agents 是一种计算机程序,它在计算机系统中的执行功能类似于现实世界的 Agent。软件 Agent 是一个处于某种环境并作为环境一部分持续自主运行的实体,它感知环境并作用于环境,执行自己的议程或目标序列以影响其将来可以感知到的东西。在充满分布性、异构性的 Web 信息空间中,人工智能方法特别是智能代理(Agent)技术,为基于 Internet 的信息过滤系统提供了一种智能化的信息获取和访问手段,是实现人机交互学习和信息收集、过滤、聚类以及融合的较好方法,尤其是应用在智能信息方面,以及实

现对传统信息检索系统的智能化接口的封装上有较好的效果。智能信息 Agent 具有以下五个特性。

(1) 综合性(Integrated)：Agent 必须支持一个易懂、相容的界面。

(2) 表达性(Expressive)：Agent 必须接受和理解不同形式的查询。

(3) 意图性(Goal-oriented)：Agent 必须知道“什么时候”和“如何完成”一个目标任务。

(4) 合作性(Cooperative)：Agent 必须同用户进行合作。

(5) 用户化(Customized)：Agent 能够适应不同的用户。

正是由于智能 Agents 的这些特性，许多组织和研究采用它来提高网上信息检索的能力。需要说明的是，本书介绍的基于多 Agents 的智能信息过滤系统并不给出各个 Agents 的具体形式定义和实现，对专门 Agents 技术的研究已经超出了本书的范畴。我们的主要目的是在现有 Agents 技术的基础上，利用 Agent 的特性，给出一个个性化的基于多 Agents 技术的智能信息过滤系统模型，以便从智能性、主动性、扩充性、易维护性等方面弥补现有智能信息过滤系统中的不足，提高检索速度和精度，帮助人们最大限度地发现自己感兴趣的问题。

2. 多 Agents 智能过滤系统中知识库的建立

多 Agents 智能过滤系统的核心是知识库的建立，建立过程一般需要三个表，分别用来存放学习得到的三种知识：①主题词、相关词和过滤词表；②用户个性化文件表；③检索结果数据表(WWW 资源表)。

在基于关键词的检索过程中，通常会遇到关键词的内涵和外延不够明确的问题，为此，引入了主题词和关联词的概念。主题词是指关键词，关联词是指与主题词相关的词，是对主题词的补充。关联词分为限制性关联词和近似性关联词，关联词典就是这些关联词的有机结合。在关联词典中存放的就是主题词和与之对应的关联词。例如，对于我们研究的智能 Agent 而言，主题词是 Agent，其相似的关联词是“智能代理”，限制性关联词是“人工智能”。可见，近似性关联词就是与原主题词内涵相同的词汇，限制性关联词就是对原主题词外延加以限制的词汇。而过滤词表示的是用户对与此词相关的信息不感兴趣的词。用户提交主题词和过滤词后，系统会构造包含主题词、关联词和过滤词的布尔表达式。在上例中，用户提交主题词 Agent 和过滤词“硬件”后，系统会给出如下的布尔表达式：

$$(((\text{Agent} \vee \text{智能代理}) \wedge \text{人工智能}) \wedge ! \text{ 硬件})$$

其中 $\wedge$ 表示“与”，$\vee$ 表示“或”，! 表示“非”。

采用关联词典的优点在于：

(1) 用户界面友好。采用关联词典，用户不必适应各种搜索引擎的关键词搜索界面和由此带来的不便，只要输入主题词和过滤词，系统就能给出各个搜索引擎的查询词，供其调用。

(2) 用户可以根据自己的需求生成不同的关联词典，从而满足个性化查询。其结构见表 5-1。

表 5-1 关键词表结构

字 段 名	说 明
keyWordID	关键字 ID
KeyWord	关键字
RelevantWord	关联词
FilterWord	过滤词

WWW 资源表存储从 WWW 上获取的站点信息,包括 Title、URL、文档主题内容、站点更新时间等,这些站点信息大多数是用户感兴趣的信息,这为进一步的信息过滤提供了本地资源。其基本结构见表 5-2。

表 5-2 WWW 资源结构表

字 段 名	说 明
PageID	页面 ID
SiteID	所属站点 ID
Title	页面标题
URL	页面地址
StoredPath	存储路径
Description	页面描述
UpdateTime	页面更新时间
AnalysisResult	页面分析结果

用户个性化文件表包含两个内容:一是保存了各个用户感兴趣的主题信息;二是保存了用户经常性的网络行为特征,例如用户经常搜索的关键词信息、经常访问的网站的信息、关键词的访问频率等。

3. 多 Agents 智能过滤系统的总体框图

图 5-5 给出了一种通用的多 Agents 过滤系统结构,按照功能的不同,将系统分成用户界面 Agent、兴趣管理 Agent、过滤查找 Agent、站点操作 Agent、搜索更新 Agent 和系统主控 Agent 六大部分。其中,用户界面 Agent 是用户和过滤系统的中介;过滤查找 Agent 接受用户的特征请求,对 WWW 资源库进行查找和过滤;兴趣管理 Agent 接收来自用户界面的反馈信息,对个性化文件库的信息进行修改;搜索更新 Agent 和站点操作 Agent 是面向网络操作的,搜索更新 Agent 按一定周期自动从 web 上获取信息补充到 WWW 资源库中,站点操作 Agent 直接面向资源系统或者站点获取信息,并将结果返回到用户界面 Agent;系统主控 Agent 负责多 Agents 之间的通信与协作。

下面将详细介绍系统主要模块的功能及采用的相关技术。

1) 用户界面 Agent

用户界面 Agent 是用户和过滤系统的中介,其主要功能包括三方面:一是实现信息导引,帮助用户确定自己需要的信息所在的领域,细化和规范查询要求;二是提供用户相关信息反馈窗口,记录用户对查找结果的满意程度;三是为用户提供注册登录界面,以便存储用户的个性化信息,这是用户兴趣管理的一部分,也是个性化服务的一个

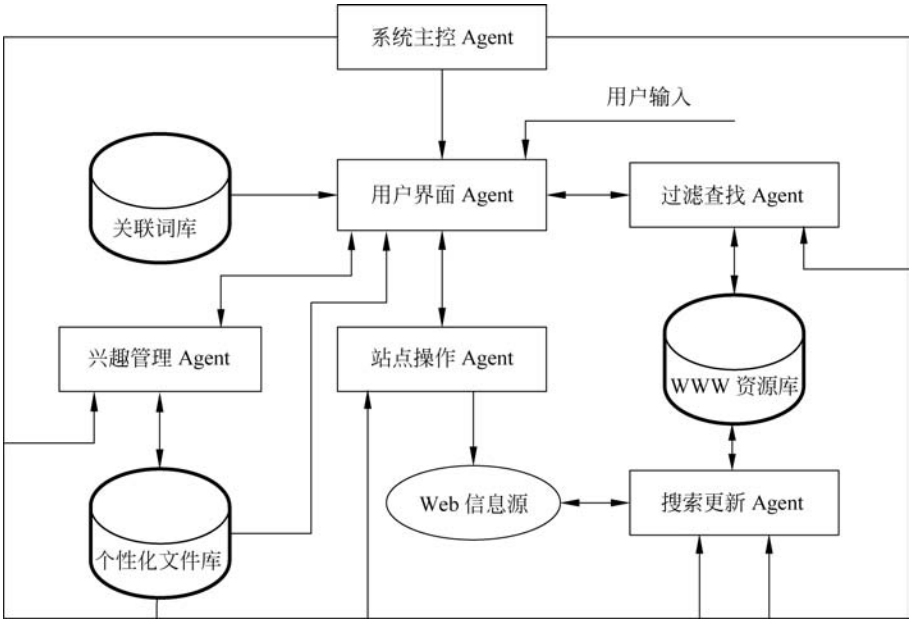

图 5-5　多 Agents 过滤系统的结构

特点。

其中，实现信息导引的关键技术是主题信息分类。对此，我们分别在知识库中建立了针对不同用户不同主题的个性化文件库和关联词库，用户界面 Agent 根据知识库对用户提交的查询请求给出最满意的表示方式。对反馈信息的描述一般采用等级化选择的返回方式，由用户对结果匹配的满意程度做出评价。

2）过滤查找 Agent

过滤查找功能是根据用户界面 Agent 的请求实现对 WWW 资源库的查找，并将查找结果反馈给用户界面。这里所涉及的技术是查找方式，单纯的关键词匹配查找是不够的，容易造成返回结果过多或定位不准确。这里充分利用了布尔模型和向量空间模型的优点，给出一种新的过滤算法，同时计算用户特征文件与检索文档的匹配度和相似度，从而为用户提供最能反映用户特征主题的过滤结果，前文已有详细介绍。过滤查找 Agent 返回的只是用户查找的中间结果，例如，相关站点 IP 地址和站点的主题内容等。由用户界面 Agent 返回中间结果给用户，并由用户人工选定后，再交给站点操作 Agent，由其直接从目标站点获取所需结果。

3）站点操作 Agent

站点操作 Agent 是直接与信息源进行连接获取信息的代理，可以在现有网络通信协议 TCP/IP 的基础上实现。技术关键在于 Agent 与相关系统之间的接口关系的确定。我们的方法是在 WWW 资源库中直接存储资源站点的绝对路径，这种方案与当前的网络数据获取方式是一致的，但前提是 WWW 资源库中获取数据的路径必须绝对正确，不能出现链接不上或链接错误的情况。

4) 兴趣管理 Agent

兴趣管理 Agent 与用户界面 Agent 以及个性化文件库相连,接受并存储用户界面 Agent 的反馈评价信息表,能对用户反馈意见进行统计分析,按一定的学习规则对个性化文件库 Profile 中特征词条的权重信息进行修改,同时根据用户要求设定兴趣监控站点。建立合理的权重更新修改规则是该 Agent 的技术重点,可以引入相关反馈技术(Relevance Feedback)和 Hopfield 神经网络的联想记忆学习功能进行处理。

5) 搜索更新 Agent

搜索更新 Agent 的主要功能是完成网上信息的自动获取,实时扩充和更新 WWW 资源库的内容,保证 WWW 资源库中的站点信息是实时的、正确的和有效的。关键技术有两点:一是多线程机制,提高检索速度;二是借助已有的搜索引擎实现自己的搜索目标。最常见的问题在于常用的搜索引擎用户接口一般为异构的,有其特定和复杂的连接方式和查询语法。针对这种状况,通用的解决方案是在搜索更新 Agent 模块中使用屏蔽接口转换技术,将搜索引擎的位置、接口等细节屏蔽起来,将用户的查询转换成不同的形式连接到不同的搜索引擎,同时将不同搜索引擎的返回结果处理成一致的形式,输入 WWW 资源库。此搜索更新 Agent 具有如下优点:

(1) 将用户的查找请求转换为若干个底层搜索引擎处理格式;

(2) 向各个搜索引擎发送查询请求,并统一其返回检索结果;

(3) 不需要建立庞大的索引数据库,也不需要使用复杂的检索机制,便于维护。

6) 系统管理模块

该模块分为系统初始化和系统设置两个子模块。系统初始化子模块在系统加载时自动启动,该模块处理过程包括连接数据源、打开数据库、启动自动网页监视后台进程、初始化程序界面、调出已写入注册表的系统初始化默认信息、恢复默认搜索引擎、恢复默认代理设置等。系统设置子模块用于重新设置代理和默认的搜索引擎等,所设置的内容写入系统配置表,当再次启动系统时,该配置将作为默认的系统参数配置。

7) 知识库管理模块

对用户长期没有访问的网站信息和主题兴趣,采用一定策略减少其权值,当权值低于预先设定的阈值时,将该网站信息或主题兴趣抛弃,这样可以避免随着时间的增加,数据库的内容无限增大,以达到对知识库进行动态管理和维护的目的,并且提高程序的运行速度。

5.6.2 基于文本匹配的过滤系统

本节从中文网页信息内容过滤系统的需求分析出发,讨论基于文本匹配的过滤系统的总体结构设计和模块划分,并对系统各模块的功能进行详细阐述。

1. 总体设计

系统采用后台程序和监控端相结合的结构。监控端负责网页信息的截获,并将其反馈给后台程序,接收后台程序的命令对网页重定向不做处理。后台程序负责网页信息的检测和判定,并将判定结果发送给监控端,同时,维护数据库更新并提供相关管理界面等。系统工作原理如图 5-6 所示。

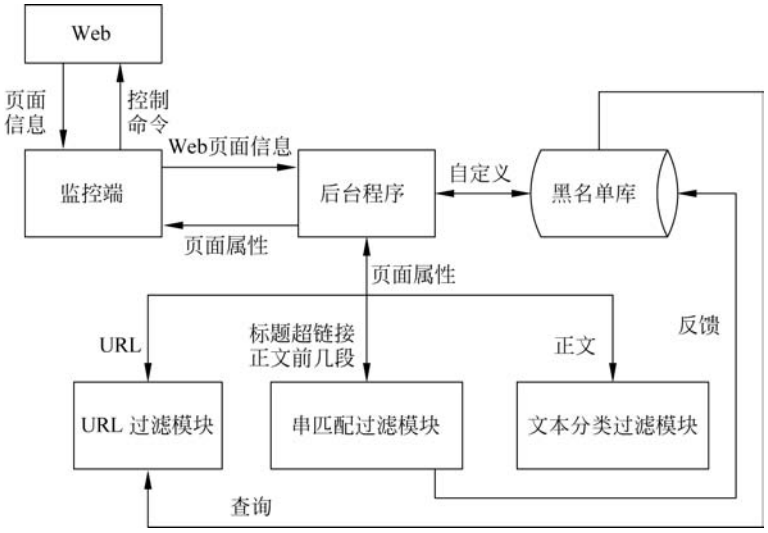

图 5-6　基于文本匹配的过滤系统工作原理

网页判定流程图如图 5-7 所示。系统对 IE 浏览器实时监控，当监控到用户有新的访问请求时，系统将用户访问的 URL 和对应的网页文本信息发送给后台程序，在没有接收到后台程序指令之前，屏蔽 IE 浏览器的显示。后台程序收到监控端的新数据后进行网页属性判定，根据网页的 URL 和网页文本信息判定网页性质，并发送判定消息到监控端。

系统的过滤方法采用 URL/IP 过滤和内容过滤相结合的方法。根据监控端发送到后台程序的网页信息，首先，判断该 Web 页面的 URL 是否在黑名单上，若网页在黑名单上则阻止用户访问，若不在则进入内容过滤模块，对文字图片进行分别处理得到 Web 页面的属性信息；接着，根据属性信息判断是否阻止用户访问，并且反馈给数据库，加入黑名单。由于图片的处理速度要慢于文字的处理速度，且很多情况下文本不良信息和图像不良信息会同时出现，因而采用先文本过滤后图片过滤的过滤策略，这样可以减少图片过滤模块的调用次数，从而提高系统的处理速度。当然，在系统配置允许的情况下，也可以将文本过滤和图片过滤并行处理。

网页文本过滤模块采用字符串匹配过滤和文本分类过滤两种过滤模式相结合的策略。首先依据敏感词库对网页文本信息一些特定的位置进行字符串检索，如果检索出敏感词汇，则判定为网页非法，发送判定消息给监控端；否则继续进入文本分类过滤检测，通过文本分类算法判定网页属性，并发送判定消息给监控端。对于判定非法的网页，须及时反馈 URL（空格）到黑名单库，当再次访问同一个网页时就不需要再进行文本过滤模块处理。由以上分析可知，系统采用三级过滤的策略，分别为 URL 过滤、字符串匹配过滤和文本分类过滤。过滤顺序按照处理速度进行排序：URL 本身长度很短，检测过滤只需要对比黑名单，处理速度最快；字符串匹配过滤在网页的一部分文本中检索敏感词汇，将文本内容和敏感词库进行对比，速度次之；文本分类算法计算复杂，耗时最长。三级处理中任意一级将网页判定为非法网页后，就不需要再进行接下来的判定，只有当网页判定为正常网页时才需要进行下一级的处理。这样的设计策略可以用最短的时间检测出不良

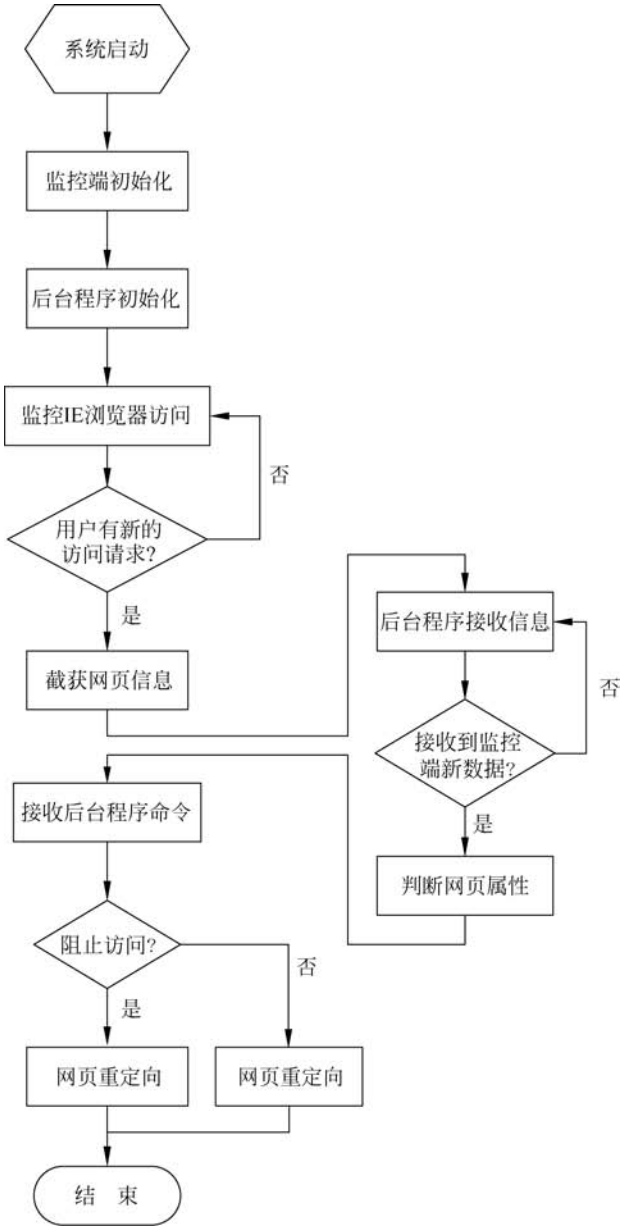

图 5-7 网页判定流程

网页,最大限度提高系统的效率,保证系统的实时响应。

2. 模块设计

中文网页过滤系统最关键的是过滤算法的设计和实现。系统总体设计采用三级过滤系统,将过滤系统分为三个主要的模块,分别是基于 IP/URL 的过滤模块、基于字符串模式匹配的过滤模块和基于文本分类技术的过滤模块。下面对各个模块详细设计进行说明。

1）基于IP/URL的过滤模块

基于IP/URL的过滤模块是三个过滤模块中的最上层，网页信息要首先经过该模块的处理。模块流程图如图5-8所示。从网页信息中提取出URL，然后在黑名单库中进行查询，若查询到该URL则表示网页包含不良信息，并予以阻止，否则不进行处理，进入后续模块的处理。基于IP/URL的过滤模块的所有操作都是以黑名单库为中心，围绕黑名单库进行的，由此可见模块的关键是黑名单库的设计，且黑名单库的设计好坏直接关系模块处理速度的快慢。黑名单数据库主要包含两个查询操作和三个更新操作。

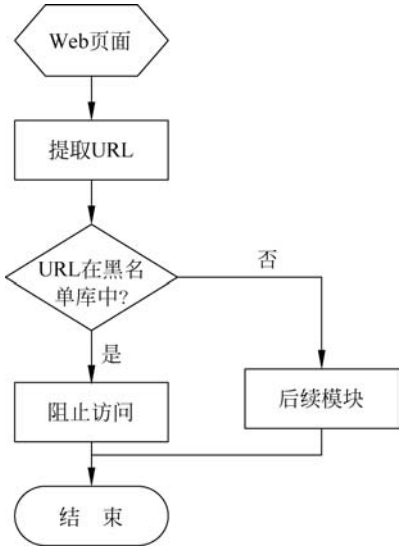

图5-8　基于IP/URL的过滤模块流程图

两个查询操作分别是：

（1）待检测网页URL的查询；

（2）用户自定义黑名单库的查询操作。

黑名单库还要接收三个更新操作，分别是：

（1）接收基于字符串模式匹配的过滤模块反馈信息；

（2）接收基于文本分类技术的过滤模块反馈信息；

（3）接收用户自定义操作，对黑名单库进行的添加和删除操作。

2）基于字符串模式匹配的过滤模块

网页文本的信息一般包含在标题、正文和超链接中。标题通常是网页内容的概括，一般情况下，当人们看到标题就可以知道文章大概讲述的内容，因此，标题中一般包含比较大的信息量，是检索敏感信息的重点。相比于标题，正文内容较长，但是重要的信息一般会在前几段出现，前几段如果不出现不良信息，则后面再出现不良信息的概率就比较小，因此，正文的前几段也是不良信息检索的重点。现在越来越多的网站通过超链接的形式嵌入到其他的网站中，而超链接中的文字一般会选择比较诱人且信息量大的文字，因此，这也成为检索的重点。由以上可以看出，从标题、正文前几段和超链接中检索出不良信息的概率比较大，应对其进行特殊处理。

基于字符串匹配技术的过滤模块流程图如图5-9所示。首先，模块得到用户将要访问的互联网Web页面，对Web页面进行分析，提取出标题、正文前几段和超链接；然后，初始化字符串模式匹配算法，通过敏感词库在标题、正文前几段和超链接中进行敏感词汇检索，若没有检测出不良信息，则对用户访问不加限制并进入后续模块的处理，一旦检索出敏感词汇则阻止用户访问，同时将网页的URL信息反馈给黑名单库。

基于字符串匹配的过滤模块采用AC-BMH作为其核心算法，这主要是由于基于字符串匹配的文本过滤有两个特点：一是主要针对中文文本过滤；二是敏感词库中的词语一般较短。这两点都使得拥有好后缀规则的应用较少，起主导作用的是坏字符规则。因此，针对这种大字符集上的应用采用AC-BMH算法，只使用坏字符规则对算法进行优

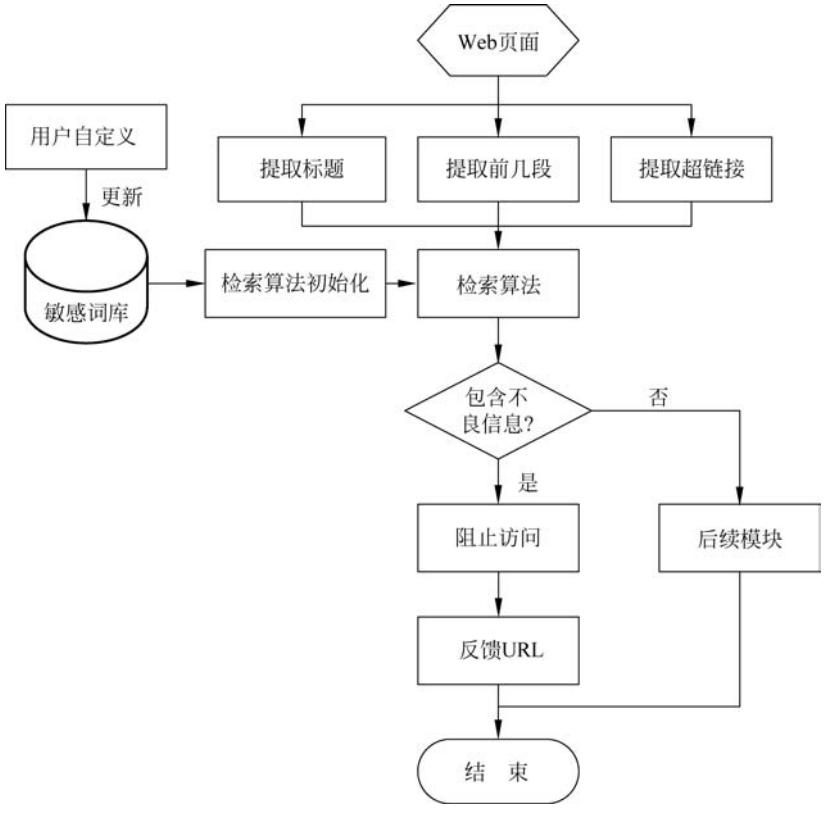

图 5-9 基于字符串匹配技术的过滤模块流程图

化,提高效率。敏感词库的建立,是通过对大量文章中词频的统计,选出最能代表敏感文章的词语。高频率词汇通常是文章中的常用语,如“我们”“开始”等,这些词汇在所有文章中出现的频率都很高,因而不能代表文章的类别;低频词包含信息很少,也不能反映文章的类别;最能表达文本属性的一般是文章中的中频率词汇。通过词频统计选出最能体现文本属性的中频词,将这些词加入敏感词库,也可以通过人工手动添加作为补充,同时,也可以为特定用户、特定的过滤添加不同的词库。敏感词库在 AC-BMH 算法初始化时通过 Init_tree 函数读取并添加到模式树中,初始化时对词库顺序没有要求,依次读取敏感词库的每一个词汇。处理过程中没有复杂的处理和其他数据的出现,因而在这里仅采用了普通文本形式来存储敏感词库。

3) 基于文本分类技术的过滤模块

基于文本分类技术的过滤模块是三层过滤中最后一层,当前两种过滤策略都将网页判定为正常网页时才进行该模块的处理。文本分类技术将待检测文本自动分类,具体到中文网页过滤的应用中是一种二文本分类,文本只有合法和非法之分,没有类别的区分。该模块的数据处理对象是网页文本的正文部分,通过分类模型判定文本的分类属性,依据分类属性进行过滤。该模块涉及整个正文部分的检测,数据处理量大,分类模型计算复杂,因而整体速度偏慢。模块首先得到用户将要访问的互联网 Web 页面,提取出正文内

容，然后对正文进行预处理，得到分类器可以识别的文本数据，然后通过分类计算得到Web页面的属性判定，网页归为正常网页则允许访问，若归为不良网页则阻止访问，同时将网页的URL信息反馈给黑名单库。

基于文本分类的过滤模块选取支持向量机算法作为模块的核心算法。主要原因有以下三点：

(1) 中文网页过滤的处理对象是单个的Web页面，一般来讲页面比较小，而支持向量机算法对小样本分类时速度快、分类准确率高；

(2) 训练样本库只包含支持向量的样本，训练出来的分类模型占用空间少；

(3) 支持向量机是一种原生的两类分类算法，很适合网页过滤。

支持向量机文本分类算法分为训练过程和识别过程。训练过程是对训练样本库训练得出分类模型的过程。训练样本库中的数据均是已确定分类属性的有代表性的文本，其质量好坏关系到分类模型的质量，进而影响系统识别过程的准确性。训练样本库中的不良文本要涵盖暴力、色情和反动等多个方面的文本，正常文本要包含政治、经济、科技、生活等全方位的文本。这样的样本库才最有代表性，也最能突出两类文本各自的特点，训练出来的分类模型的准确率和实用性才会更好。由于没有标准库，只能是从网络手动搜集一些样本库资源，尽可能做到准确详尽。

5.6.3　基于朴素贝叶斯算法的垃圾邮件过滤

随着互联网的迅速发展，网络改变了人们传统的通信方式。电子邮件因为其方便快捷而被人们广泛接受和使用。但是现今垃圾邮件问题日益泛滥严重，邮件系统的安全和可靠性依然是人们关注的焦点。根据中国网络不良与垃圾信息举报受理中心2016年的数据显示，中国网民平均每周收到的垃圾邮件达12封，每年收到的垃圾邮件总计3700亿封。垃圾邮件严重干扰了正常的互联网秩序，研究并设计有效的垃圾邮件过滤器具有非常重要的现实意义。白名单、行为监控、黑名单以及关键字过滤等是目前常用的垃圾邮件过滤技术，但这些过滤技术缺乏自适应性，面对内容多变的垃圾邮件其过滤效果并不够理想。针对这一问题，面向信息内容的朴素贝叶斯过滤器不仅具有自适应性，算法复杂度低、分类精度高等优点，而且也可以根据用户需求进行个性化过滤。本节介绍了一种基于朴素贝叶斯算法的垃圾邮件过滤方法。

1. 原理概述

1) 贝叶斯定理

贝叶斯定理描述的是两个不同的事件A、B，A为条件B发生的概率与B为条件A发生的概率之间的关系。贝叶斯公式可表示为

$$P(A \mid B)=\frac{P(A)P(B \mid A)}{P(B)} \tag{5-7}$$

在式(5-7)中，P为事件发生的概率。$P(A)$称为先验概率(Prior Probability)，即在B事件发生之前，对A事件概率的一个判断。$P(A \mid B)$称为后验概率(Posterior Probability)，即在B事件发生之后，对A事件概率的重新评估。$P(B|A)/P(B)$称为可能性函数(Likelihood)，这是一个调整因子，使得预估概率更接近真实概率。所以条件概

率可以理解为

后验概率 = 先验概率 × 调整因子

这就是贝叶斯推断的含义,也就是首先预估一个先验概率,然后加入实验结果,观察这个实验到底是增强还是削弱了先验概率,由此得到更接近事实的后验概率。如果可能性函数>1,意味着先验概率被增强,事件 A 发生的可能性变大;如果可能性函数=1,意味着 B 事件无助于判断事件 A 的可能性;如果可能性函数<1,意味着先验概率被削弱,事件 A 发生的可能性变小。

贝叶斯公式的意义在于它反映了导致一个事件发生的若干"因素"对这个事件发生的影响分别有多大。例如考虑如下问题:

已知某种疾病的发病率是0.001,即1000人中会有1个人得病。现有一种试剂可以检验患者是否得病,它的准确率是0.99,即在患者确实得病的情况下,它有99%的可能呈现阳性。它的误报率是5%,即在患者没有得病的情况下,它有5%的可能呈现阳性。现有一个病人的检验结果为阳性,请问他确实得病的可能性有多大?

解:假定 A 事件表示得病,那么 $P(A)$ 为 0.001。这就是先验概率,即没有做试验之前,我们预计的发病率。再假定 B 事件表示阳性,那么要计算的就是 $P(A|B)$。这就是后验概率,即做了试验以后,对发病率的估计。

根据条件概率公式:

$$P(A \mid B)=\frac{P(B \mid A)P(A)}{P(B)} \tag{5-8}$$

用全概率公式改写分母:

$$P(B)=P(A)P(B \mid A)+P(\overline{A})P(B \mid \overline{A})$$

由题设可知:$P(B|A)=0.99$,$P(B|\overline{A})=0.05$ $P(\overline{A})=0.999$。

将数字代入得到结果,$P(A|B)$ 约等于 0.019。也就是说,即使检验呈现阳性,病人得病的概率,也只是从 0.1%增加到 2%左右。这就是所谓的"假阳性",即阳性结果完全不足以说明病人得病。

利用贝叶斯定理构造的决策方法是在所有相关概率都已知的情况下,考虑如何基于这些概率和可能的期望损失来选择最优分类的方法。现假设有 N 种可能的类别$\{C_1, C_2, \cdots, C_N\}$,且存在样本 $x \in \{x_1, x_2, \cdots, x_N\}$,需要将样本 x 分为相应的类别,则可以定义基于后验概率 $P(C_i|x)$ 将某一样本 x 分类为 C_i 所产生的期望损失:

$$R(C_i \mid x)=\sum_{j=1}^{N}\lambda_{ij}P(C_j \mid x) \tag{5-9}$$

其中 λ_{ij} 表示将真实类别为 C_j 分类为 C_i 所产生的损失。利用贝叶斯定理来分类的目标是:寻找能够最小化全局风险的准则 h,h 应为

$$h(x)=\operatorname{argmin}R(c \mid x), \quad c \in \{C_1, C_2, \cdots, C_n\}$$

即在每个样本 x 上都选择能使得期望损失 R 最小的类别 c,此时为所得到的贝叶斯分类器的性能上限。

利用贝叶斯定理最小化期望损失相当于利用有限的训练样本尽可能准确地估计后验概率 $P(c|x)$。

2）朴素贝叶斯分类器

给定类标号 y，朴素贝叶斯分类器在估计类条件概率时假设属性之间条件独立。条件独立假设可形式化地表示如下：

$$P(X \mid Y=y)=\prod_{i=1}^{d} P(X_i \mid Y=y) \tag{5-10}$$

其中每个属性集 $X=\{X_1,X_2,\cdots,X_d\}$包含 d 个属性。

在深入研究朴素贝叶斯分类法如何工作的细节之前，先介绍条件独立概念。设 X、Y 和 Z 表示三个随机变量的集合。给定 Z，X 于 Y 条件独立，如果下面的条件成立：

$$P(X \mid Y,Z)=P(X \mid Z) \tag{5-11}$$

条件独立的一个例子是一个人的手臂长短和他的阅读能力之间的关系。可能手臂较长的人阅读能力也较强。这种关系可以用另一个因素解释，那就是年龄。小孩子的手臂往往比较短，也不具备成人的阅读能力。如果年龄一定，则观察到的手臂长度和阅读能力之间的关系就消失了。因此，可以得出结论，在年龄一定时，手臂长度和阅读能力二者条件独立。

X 和 Y 之间的条件独立也可以写成式(5-12)：

$$\begin{aligned} P(X,Y \mid Z)&=\frac{P(X,Y,Z)}{P(Z)}=\frac{P(X,Y,Z)}{P(Y,Z)}\times\frac{P(Y,Z)}{P(Z)} \\ &=P(X \mid Y,Z)\times P(Y \mid Z)=P(X \mid Z)\times P(Y \mid Z) \end{aligned} \tag{5-12}$$

有了条件独立假设，就不必计算 X 的每一个组合的类条件概率，只需对给定的 Y，计算每一个 X_i 的条件概率。后一种方法更实用，因为它不需要很大的训练集就能获得较好的概率估计。

分类测试记录时，朴素贝叶斯分类器对每个类 Y 计算后验概率：

$$P(Y \mid X)=\frac{P(Y)\prod_{i=1}^{d} P(X_i \mid Y)}{P(X)} \tag{5-13}$$

由于对所有的 Y，$P(X)$是固定的，因此只要找出使分子 $P(Y)\prod_{i=1}^{d} P(X_i \mid Y)$ 最大的类就足够了。一般来说，可以估计分类属性的条件概率，也就是对分类属性 X_i，根据类 y 中属性值等于 X_i 的训练实例的比例来估计条件概率 $P(X_i=x_i|Y=y)$。

朴素贝叶斯分类法使用两种方法估计连续属性的类条件概率。

(1) 可以把每一个连续的属性离散化，然后用相应的离散区间替换连续属性值。这种方法把连续属性转换成序数属性。通过计算类 y 的训练记录中落入 X_i 对应区间的比例来估计条件概率 $P(X_i=x_i|Y=y)$。估计误差由离散策略和离散区间的数目决定。如果离散区间的数目太大，则会因为每一个区间中训练记录太少而不能对 $P(X_i|Y)$做出可靠的估计。相反，如果区间数目太小，有些区间就会含有来自不同类的记录，因此失去了正确的决策边界。

(2) 可以假设连续变量服从某种概率分布，然后使用训练数据估计分布的参数。高斯分布通常被用来标识连续属性的类条件概率分布。该分布有两个参数，均值 μ 和方差 σ^2。对每个类 y_j，属性 X_i 的类条件概率等于：

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}} \tag{5-14}$$

其中,参数 μ_{ij} 可以用类 y_j 的所有训练记录关于 X_i 的样本均值 $\bar{x}$ 来估计。同理,参数 σ^2 可以用这些训练记录的样本方差 s^2 来估计。

2. 总体设计

本节目标是基于朴素贝叶斯分类方法设计一个快速精准的垃圾邮件过滤系统。贝叶斯过滤器是一种统计学过滤器,建立在已有的统计结果之上。所以,必须预先提供两组已经识别好的邮件,一组是正常邮件,另一组是垃圾邮件。在本节中用这两组邮件,对过滤器进行训练。这两组邮件的规模越大,训练效果就越好。一般训练使用的邮件规模为正常邮件和垃圾邮件各 4000 封。

在垃圾邮件过滤设计中,相比于速度,精准度是更为重要的考量因素。因为相比于漏拦截的垃圾邮件,将正常邮件误分为垃圾邮件会对用户造成更大的麻烦。因此本系统对将正常的邮件误分为垃圾邮件会有比较高的要求。系统采用模块化的设计,可分为:邮件预处理、训练数据集以及垃圾邮件过滤等模块。其中,邮件预处理模块的功能是通过分析邮件的格式,将邮件的头部以及正文部分解析出来,并且运用分词工具对邮件内容进行分词。训练数据集模块的功能是将预处理过的已知邮件类别的邮件进行训练,目的是提取出正常邮件和垃圾邮件的分词数据,并将其写入数据库。垃圾邮件过滤模块的功能是根据每个特征词的相关权值检测待分类邮件的类别,图 5-10 是系统总体的架构图。

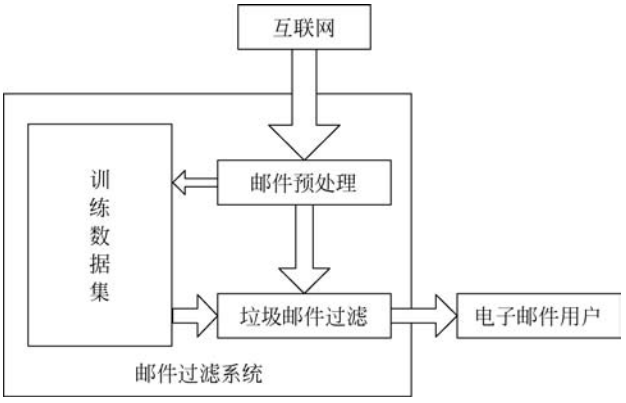

图 5-10 垃圾邮件过滤系统总体的架构图

垃圾邮件过滤其实就是运用朴素贝叶斯分类模型对垃圾邮件进行分类。朴素贝叶斯分类模型如图 5-11 所示。考虑到朴素贝叶斯分类器针对的一般是离散型数据,因此需要对邮件内容进行分词处理,提取分词后的一些关键单词或词语作为特征项进行后续过滤处理。

利用贝叶斯定理进行垃圾邮件的训练过程其实较为简单。主要是解析所有邮件,提取每一个词,然后计算每个词语在正常邮件和垃圾邮件中的出现频率。例如假定“广告”这个词,在 4000 封垃圾邮件中,有 200 封包含这个词,那么它的出现频率就是 5%;而在

4000封正常邮件中，只有2封包含这个词，那么出现频率就是0.05%（如果某个词只出现在垃圾邮件中，可假定它在正常邮件的出现频率是1%，反之亦然，这样做是为了避免概率为0。随着邮件数量的增加，计算结果会自动调整）。

具体来说，按照朴素贝叶斯分类过滤模型，垃圾邮件分类问题可转化为下面的问题。即假如已有 m 个邮件样本集合$\{S_1, S_2, \cdots, S_m\}$，其中邮件的类型分为垃圾邮件和正常邮件，现在有一封新的邮件，需要确定它到底是垃圾邮件还是正常邮件，可分为如下步骤：

（1）首先获取样本数据集，例如从网上爬取一些不同类型的邮件数据，其中垃圾邮件和正常邮件的数目已经确定。

（2）将所有的邮件利用分词算法进行分词处理，并按照垃圾邮件和正常邮件两个方面分别进行标记，得到垃圾邮件的词频表和正常邮件的词频表。

（3）将待检测邮件进行分词处理，并分别计算出每个分词在不同类别的邮件中出现的概率。

（4）根据已计算的各个分词在不同类别邮件中的权值，利用概率公式，分别得到在此样本数据集中该邮件是垃圾邮件的概率和是正常邮件的概率并进行对比。通过以上的几个步骤，就可以根据朴素贝叶斯分类器来判断待检测邮件是否是垃圾邮件。

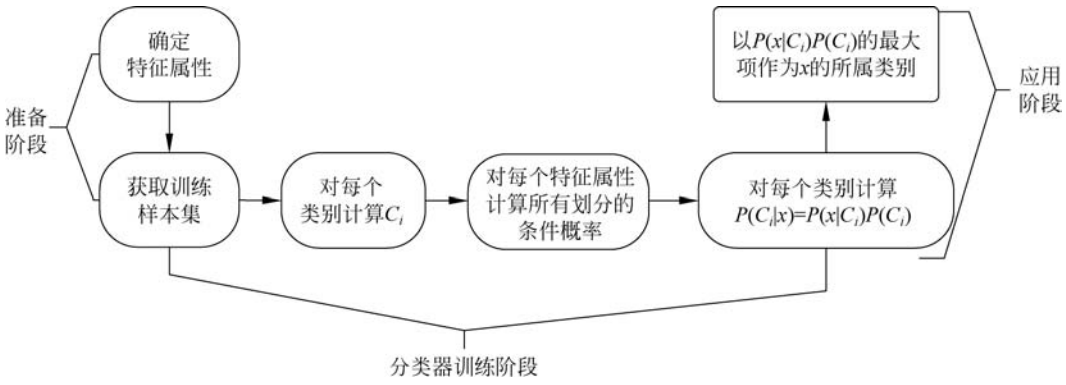

图5-11　朴素贝叶斯分类模型

从图5-11可以看出，训练集学习训练过程是整个检测过程的核心，也就是先根据训练集计算某一类已知文本分类的先验概率，得到计算结果的后验概率后，对后续收到的新的文本类型进行分析预料。在已知的分类概率条件下，得到待检测邮件属于某一类别的概率值，并取其中最大值，将该文本归类到最大值的那一类中。

3. 模块设计

1）邮件预处理模块设计

现在大多数的邮件系统都是MIME标准，邮件中包含了各种格式、各种数据类型的内容。如图5-12所示，邮件预处理流程一般包含邮件解析、邮件文本分词、邮件特征词提取等模块。系统在进行垃圾邮件检测分类之前需要获取邮件主体内容，并

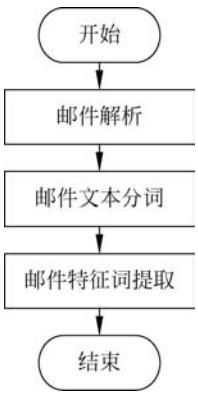

图5-12　邮件预处理模块的流程图

去除其中不能解析的非文本内容,留存容易解析的邮件头以及正文等内容。不同于英文文本,中文文本中词与词之间缺少间隔,计算机无法直接提取中文的词语。因此在特征提取之前需要将中文文本分成一堆易于解析的词组,即对中文邮件实行去停留词等特殊处理,去除邮件内容中许多没有语义的语气词、连接词等,并提取出邮件内容的特征项进行后续过滤模型构建。

其中邮件解析模块主要是解析 MIME 编码,根据 Content-Type 字段的值来对邮件内容的类型进行判断,然后根据判断结果对其解码,最后提取到所需的标题和正文的数据。如果一些垃圾邮件为了伪装成正常邮件额外加入无关的信息,那么在邮件预处理中就必须对这些混淆信息进行删除。例如,如果对应解码类型的字段 Content-Type 值为 text/*,这就说明正文的格式是纯文本,可直接对正文进行提取,而如果正文格式是 multipart/*的,则需要先将其分为多个对象,再对每个对象直接获取其文本。

邮件文本分词模块主要将邮件文本分成可处理的特征词。由于中文文本中字与词之间没有明确的界限,有些句子本身就存在歧义,使得人工进行中文分词较为困难。目前国内对中文分词的研究已经比较成熟,较著名的如中国科学院计算技术研究所推出的 NLPIR 汉语分词系统(又称 ICTCLA),其主要功能包括中文分词、词性标注和命名实体识别等,分词的速度与精度都比较高。在垃圾邮件过滤中,可通过调用 ICTCLA 分词库来实现中文分词。

2) 邮件训练模块设计

邮件训练模块是垃圾邮件过滤系统的核心,而其中最重要的部分是朴素贝叶斯模型的建模。在本节原理概述部分,已经知道,朴素贝叶斯模型是贝叶斯模型的一种特殊情况,由于贝叶斯定理中的先验概率是所有属性上的联合概率 $P(x|c)$,而各个属性之间相互联系,使得先验概率求取困难。朴素贝叶斯模型将每个属性之间视为独立条件,将计算方法转变为

$$P(c \mid x)=\frac{P(c)}{P(x)}\prod_{i=1}^{N}P(x_i \mid c) \tag{5-15}$$

其中 N 为属性数量,在此假设下,所需要寻找的准则 h 转变为

$$h(x)=\operatorname{argmin} P(c)\prod_{i=1}^{N}P(x_i \mid c) \tag{5-16}$$

式(5-16)为朴素贝叶斯模型分类表达式。

在垃圾邮件过滤系统中,通过已给定的训练集,以特征词之间独立作为前提假设,学习从输入到输出的联合概率分布,再基于学习到的模型,输出 X 求出使得后验概率最大的输出 Y。

在垃圾邮件分类系统中,对于收到的一份邮件,其出现的单词集为(word_1,word_2,…,word_n),计算该单词集出现情况下该封邮件可能是垃圾邮件(Spam)或正常邮件(Ham)的后验联合概率为

$$\begin{aligned}P(\text{Spam} \mid \text{word}_1,\text{word}_2,\cdots,\text{word}_n)&=\frac{P(\text{word}_1,\text{word}_2,\cdots,\text{word}_n \mid \text{Spam})P(\text{Spam})}{P(\text{word}_1,\text{word}_2,\cdots,\text{word}_n)}\\ P(\text{Ham} \mid \text{word}_1,\text{word}_2,\cdots,\text{word}_n)&=\frac{P(\text{word}_1,\text{word}_2,\cdots,\text{word}_n \mid \text{Ham})P(\text{Ham})}{P(\text{word}_1,\text{word}_2,\cdots,\text{word}_n)}\end{aligned} \tag{5-17}$$

这两个条件概率分布是一个 n 维空间向量的联合概率，如果每个特征值有 t 种取值，那就说明可能的情况一共有 t^n 次，求解该问题是一个 NP 难问题。

为了能够进一步简化问题，可假设：特征项之间是相互独立的，这也就是朴素贝叶斯的思想。这样 n 维的联合概率问题被简化成 n 个分离的概率乘积，计算复杂度也就降到了易于计算多项式级别。此时，使用类集合 $Y=c_k$ 表示邮件类型，那么该封邮件可能是垃圾邮件(Spam)或正常邮件(Ham)的后验联合概率可变成式(5-18)：

$$P(Y=c_k \mid \text{word}_1,\text{word}_2,\cdots,\text{word}_n)=\frac{P(\text{word}_1 \mid Y=c_k)P(\text{word}_2 \mid Y=c_k)\cdots P(\text{word}_n \mid Y=c_k)P(Y=c_k)}{P(\text{word}_1,\text{word}_2,\cdots,\text{word}_n)} \tag{5-18}$$

由于特征属性集即单词集$(\text{word}_1,\text{word}_2,\cdots,\text{word}_n)$是不变的，因此比较后验概率时，只需要比较式(5-18)的分子即可，也就是说，如果用类集合 $Y=\{c_1,c_2\}$ 分别表示垃圾邮件和正常邮件，y 为输出类标记，那么最后类输出为

$$y=\arg\max_{c_k} P(Y=c_k)\prod_{j=1}^{n} P(\text{word}_j \mid Y=c_k) \tag{5-19}$$

那么，最后的问题就集中在如何估计 $P(Y=c_k)$ 和 $P(\text{word}_j \mid Y=c_k)$，不同的朴素贝叶斯模型对这些参数有不同的求解方法，常用的有三种：伯努利模型、多项式模型和高斯模型。其中一般使用伯努利模型，这里重点介绍伯努利模型分类器。

伯努利分布也就是 0-1 分布。对于邮件过滤系统，可以理解为同时进行多个不同的伯努利试验，即满足多元伯努利分布。

邮件的特征属性集$(\text{word}_1,\text{word}_2,\cdots,\text{word}_n)$中，对于每一个单词 word_i，用 $\text{word}_i \in \{0,1\}$来表示这个单词是否出现。因为特征之间是相互独立的，所以多元伯努利也就变成了各个伯努利分布的连乘积。因为是伯努利分布，所以一个特征(单词)无论出不出现，都有一个概率。如果记单词 word_i 在类 $Y=c_k$ 下出现的概率为 $P(\text{word}_i \mid Y=c_k)$，根据伯努利分布可得

$$P(\text{word}_1,\text{word}_2,\cdots,\text{word}_n \mid Y=c_k)=\prod_{i=1}^{n} P(\text{word}_i \mid Y=c_k)\text{word}^i+(1-P(\text{word}_i \mid Y=c_k))(1-\text{word}^i) \tag{5-20}$$

在式(5-20)的基础上，可得

$$P(Y=c_k \mid \text{word}_1,\text{word}_2,\cdots,\text{word}_n)=\frac{P(Y=c_k)\prod_{i=1}^{n} P(\text{word}_i \mid Y=c_k)\text{word}^i+(1-P(\text{word}_i \mid Y=c_k))(1-\text{word}^i)}{P(\text{word}_1,\text{word}_2,\cdots,\text{word}_n)} \tag{5-21}$$

同样地，在计算输出类时，对不同的类的条件概率，只需要比较式(5-21)的分子即可。

根据极大似然估计，对于先验概率 $P(Y=c_k)$，其极大似然估计为

$$P(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N} \tag{5-22}$$

邮件过滤系统中,y_i 表示当前邮件,$I(y_i=c_k)$表示判断当前邮件是否属于 c_k,属于即为1,N 为邮件总数,那么式(5-22)的实际含义为类 $Y=c_k$ 样本所占比例。

对于条件概率 $P(\text{word}_i|Y=c_k)$,结合拉普拉斯平滑,其极大似然估计为

$$P(\text{word}_i \mid Y=c_k)=\frac{\sum_{i=1}^{N} I(\text{word}_i \mid Y=c_k)+1}{\sum_{i=1}^{N} I(y_i=c_k)+2} \tag{5-23}$$

式(5-23)最终的含义为类 $Y=c_k$ 下包含单词 word_i 的文件数占类 $Y=c_k$ 样本的比例。

3) 邮件过滤模块设计

邮件过滤模块功能是根据训练模块已计算的各特征项的条件概率和权值等参数,对待分类邮件类型进行判断并过滤。如图5-13所示,邮件过滤模块主要通过设定计算待分类邮件分属不同类别的概率,并基于预先设定好的阈值进行邮件过滤。

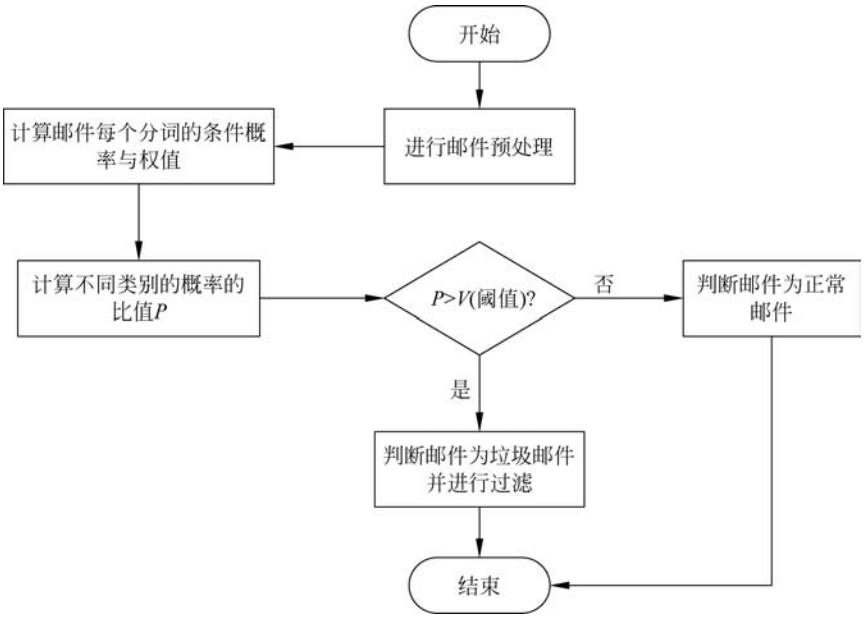

图5-13 邮件过滤模块设计流程图

基于伯努利模型构建邮件过滤模型时,可直接调用 sklearn 库实现,使用的是 BernoulliNB 方法,如下所示:

```
BernoulliNB(alpha = 1.0,binarize = 0.0,fit_prior = True,class_prior = None)
alpha: 拉普拉斯/Lidstone 平滑参数,浮点型,可选项,默认 1.0
binarize: 将数据特征二值化的阈值,< = binarize 的值处理为 0,> binarize 的值处理为 1
class_prior: 类先验概率,数组大小为(n_classes,),默认 None
fit_ prior: 是否学习先验概率,布尔型,可选项,默认 True
```

通过调用 fit 函数和 predict 函数就可以实现训练和测试的功能,如下所示:

```
clf = BernoulliNB(alpha = 1.0, binarize = 0.0, class_ prior = None, fit_prior = True)
```

```
clf.fit(X, Y)
#用下面的语句实现对测试集的分类并输出分类结果
print(clf.predict(X[2]))
```

5.7　本章小结

网络信息过滤技术能够有效、准确地找到用户感兴趣的信息，为用户提供及时、个性化的信息服务，真正做到“用户所需”。近年来，网络信息过滤技术获得了长足的发展，正在被越来越多地应用于 Web 空间，并成为研究和工程实践的热点。本章对网络信息内容过滤技术展开论述，介绍了网络信息过滤的原理，概述了网络信息过滤系统的主要类型，深入描述了网络信息内容过滤模型，分析比较了不同过滤模型，并对其中的关键技术做了重点研究。最后还给出了几种典型的信息内容过滤系统介绍。

习　　题

1. 网页内容过滤有哪些应用？目前主要有哪些方法？

2. 简单描述字符串匹配过滤算法。

3. 试描述网络信息内容过滤系统的基本框架。

4. 简要描述网络信息内容过滤的主要方法。

5. 简单比较统计和逻辑方法的异同和优缺点。

6. 两个一模一样的碗，一号碗有 30 颗水果糖和 10 颗巧克力糖，二号碗有水果糖和巧克力糖各 20 颗。现在随机选择一个碗，从中摸出一颗糖，发现是水果糖。请问这颗水果糖来自一号碗的概率有多大？