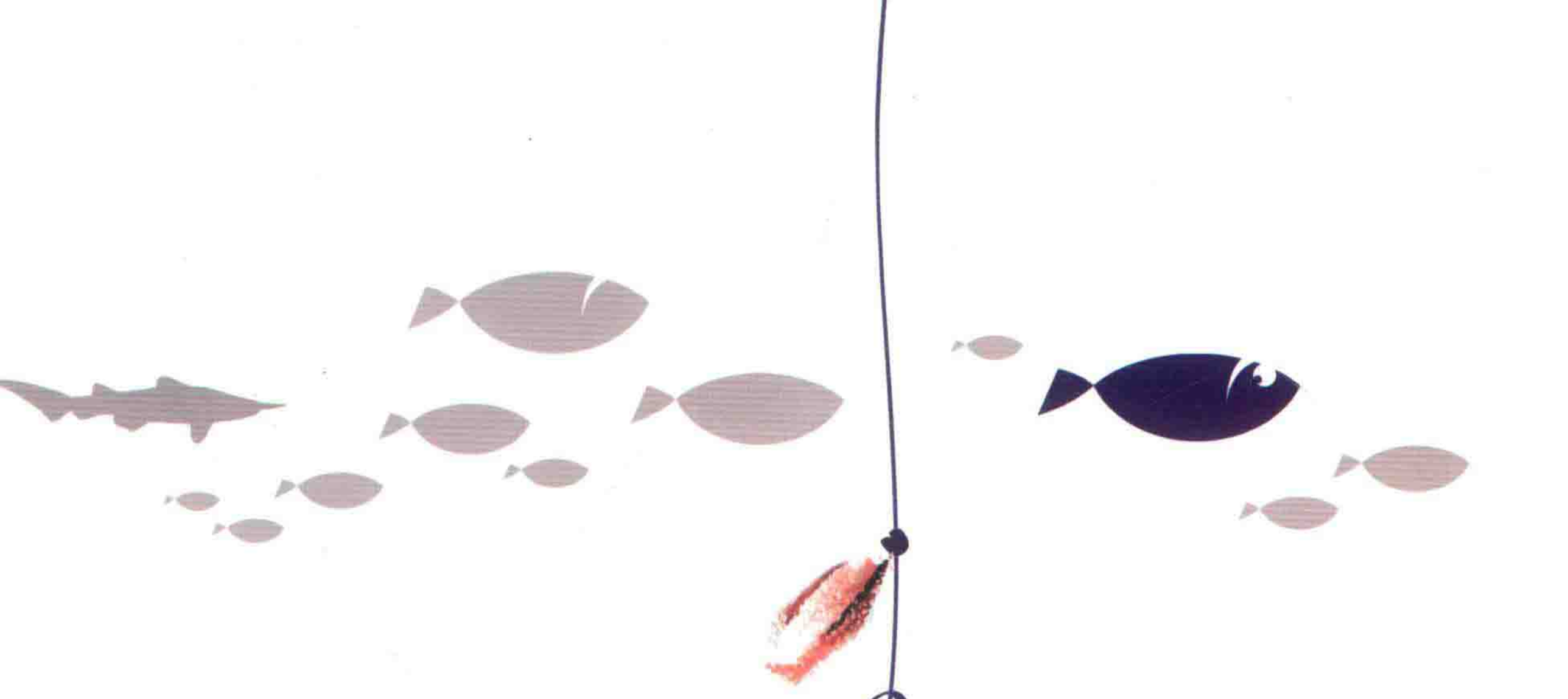

Web

文本挖掘技术

理论与应用

何慧　陈博　张莹　编著

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Web 文本挖掘技术理论与应用

何 慧 陈 博 张 莹 编著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

随着互联网和通信网的迅猛发展，网络文本成为信息的主要载体及人们生活中不可或缺的主要信息来源，文本挖掘技术的研究意义和实用价值越来越突出。另外，随着 Web 2.0 时代的到来，出现了越来越多的由用户创作的网络数字内容。用户数字内容的大量产生和传播使得短文本计算、Web 文本信息抽取、文本情感分析等逐渐成为 Web 文本挖掘研究的热点问题。本书从 Web 文本的信息抽取、聚类、分类、信息检索等技术出发，与读者分享作者多年的研究和开发经验。

本书可作为计算机各专业高年级本科、研究生的教学参考书，也可以作为数据挖掘、机器学习等专业技术人员的实用工具书。

图书在版编目 (CIP) 数据

Web 文本挖掘技术理论与应用 / 何慧，陈博，张莹编著. —北京：电子工业出版社，2017.6

ISBN 978-7-121-29827-1

I. ①W… II. ①何… ②陈… ③张… III. ①数据采集—研究 IV. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 207471 号

策划编辑：冯小贝

责任编辑：周宏敏

印　　刷：北京七彩京通数码快印有限公司

装　　订：北京七彩京通数码快印有限公司

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱　　邮编：100036

开　　本：787×980　1/16　印张：7　字数：122 千字

版　　次：2017 年 6 月第 1 版

印　　次：2017 年 6 月第 1 次印刷

定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：fengxiaobei@phei.com.cn。

前　言

随着互联网和通信网的迅猛发展，网络文本成为信息的主要载体及人们生活中不可或缺的主要信息来源，文本挖掘技术的研究意义和实用价值越来越突出。另一方面，随着 Web 2.0 时代的到来，出现了越来越多的由用户创作的网络数字内容。用户数字内容的大量产生和传播使得短文本计算、Web 文本信息抽取、文本情感分析等逐渐成为 Web 文本挖掘研究的热点问题。

编著者长期从事自然语言处理的科研工作，在相关研究成果的基础上撰写了本书。全书共分 6 章，其中，第 1 章为概论；第 2 章为基于统计语言模型的短文本计算，属于 Web 文本聚类的研究问题；第 3 章为面向广告推荐和情感分析的 Web 文本信息抽取，属于 Web 文本信息抽取的研究问题；第 4 章为基于监督和文本情感分类，是 Web 文本挖掘的一个热点问题；第 5 章为文本观点检索研究；第 6 章为总结与展望。

第 1 章阐述了 Web 文本挖掘的研究背景和意义、文本挖掘的基本概念、文本分类、聚类、信息抽取与检索系统的结构和典型应用，以及文本挖掘领域亟待解决的问题。

第 2 章针对短文本包含字符少、文本语言不规范、文本数量巨大的特点，从统计语言模型的角度对短文本计算中的有效语言特征提取和选择、聚类等进行了研究和阐述。

第 3 章讨论了面向广告推荐和情感分析的 Web 文本信息抽取。由于传统的信息抽取任务仅面向命名实体识别、实体关系抽取、指代消解和事件探测四个方面，无法满足当前互联网上更多的信息抽取方面的技术需求。本章针对广告推荐中复合词抽取问题和用户产生内容的情感词抽取问题，结合当前主要的信息抽取技术，提出了相应的算法。

第 4 章对文本情感分类问题展开研究和阐述。对现有的监督学习和半监督学习方法进行介绍，并以音乐歌词的情感分类和电影评论的情感分类为例，讨论了情感分类系统的关键问题及其具体实现技术。

第 5 章介绍文本观点检索。以编著者 2008 年参加的 COAE2008 中的面向主题的中文文本观点检索任务为主线，介绍了参评系统 PRIS-SAS。在 COAE2008 数据集上的评测指标表明，我们设计的文本观点检索系统达到了较高的性能水平。

第 6 章总结了本书的知识要点，并展望了未来的发展前沿。

本书编著者为何慧、陈博和张莹。本书的顺利完成还得到了单位领导、老师、同事以及学生的大力帮助，在此一并致谢！此外，本书还引用了一些著作、论文和网上的相关资料，未能一一完全列出，对他们的相关工作表示敬意。

本书受到北京市青年英才项目“基于微博的舆情分析关键算法研究”（项目编号 YETP0706）、中央高校基金面上项目“短文本分析关键算法研究”（项目编号 2014MS21）、中央高校基金面上项目“基于深度学习的文本分析技术及其在电力大数据中的应用”（项目编号 2017MS072）、国家自然科学基金青年项目“网络资源的语义标识与分布式定位方法研究”（项目编号 61305056）等的支持，在此对国家自然科学基金委员会、北京市教育委员会和华北电力大学表示衷心的感谢。

由于编著者水平有限，加之涵盖的内容尚在迅速发展之中，本书难免存在不足、不当甚至错误之处，恳请同行及广大读者批评指正。

编著者

2017 年 4 月

目　录

第1章 概 论

随着计算机技术、通信网、互联网的迅速发展和日益普及，文本信息的快速积累使公司、政府和个人等用户在信息处理和使用中面临前所未有的挑战。一方面，Web 上每天都不断产生大量文本数据，这些文本资源中蕴含着许多有价值的信息和知识；而另一方面，由于信息产生的速度远远超过人们对信息的利用能力，使得人们在海量的信息面前无所适从，给广大用户带来时间、资金和精力的巨大浪费。因此，如何利用机器自动地处理海量的文本信息，并从中挖掘出有用的信息和知识成为了一个亟待解决的重大问题。文本挖掘正是为解决这个问题而产生的研究方向。

1.1 研究的背景和意义

人类上下五千年丰富的文化遗产，大部分都随着文字以文本的形式流传下来。如今，随着 Web 技术的迅猛发展，网络上大量的信息以文本形式存储在文本数据库中，由来自各种数据源的文档组成，如 Web 页面、研究论文、书籍、数字图书馆、电子邮件等。在这个信息爆炸的时代，文本作为重要的信息载体之一，其数量正在以惊人的速度增长。中国互联网络信息中心(CNNIC)2016 年 1 月发布的《第 37 次中国互联网络发展状况统计报告》[1]中显示，截至 2015 年 12 月，我国网民规模达 6.88 亿，全年共计新增网民 3951 万人，增长率为 6.1%，我国互联网普及率达到 50.3%，中国网页数量为 2123 亿个，年增长 11.8%，其中静态网页数量为 1314 亿，占网页总数量的 61.9%；动态网页数量为 808 亿，占总数量的 38.1%。从网页内容上看，以文本居多，其次是图像、音频和视频网页，但相对比例仍旧不高。根据《2016 年中国网民权益保护调查报告》[2]，近半年，网民平均每周收到垃圾邮件 18.9 封、垃圾短信 20.6 条、骚扰电话 21.3 个。从以上统计数据中看，尽管互联网和通信网上信息的组成非常复杂，但文本信息依然占重要的比重。这是因为文本是信息的主要载体，而多数其他形式的信

息(图像、语音)均可以用文本进行标注。因此，面对如此庞大而且急剧膨胀的信息海洋，如何有效地组织和管理这些信息，并快速、准确、全面地从中找到用户需要的信息，是当前信息科学和技术领域面临的一大挑战。文本挖掘正是为解决这个问题而产生的研究方向。

首先区分文本(text)和文档(document)的概念。我们认为文档是一个较为广义的概念，其内涵涵盖各种文本组织形式，其外延包括文本文档、图像文档、视频文档及其混合组织形式。而这里所研究的对象——“文本”指的是文本文档(text document)，其内涵涵盖纯文本对象，其外延包括各种以纯文本内容为主的文本组织形式，包括 Web 网页、邮件/讨论组、短信、博客等[3]。

文本挖掘是一个跨学科的研究领域，它以文本为研究对象，涉及数据挖掘、模式识别、信息检索、自然语言处理等多个领域的内容。不同的研究者从各自的研究领域出发，对文本挖掘的含义有不同的理解，不同应用目的的文本挖掘项目也各有其侧重点。因此，对文本挖掘的定义也有多种，其中被普遍认可的文本挖掘定义为：文本挖掘是指从大量文本数据中抽取事先未知的、可理解的、最终可用的知识的过程，同时运用这些知识更好地组织信息以便将来参考[4]。文本挖掘的子任务涵盖了文本分类[5]、文本聚类[6]、文本检索[7]、信息抽取[8]等。

Web 文本挖掘就是从 Web 文档和 Web 活动中发现、抽取感兴趣的、潜在有用的模式和隐藏的信息的过程[9]。Web 文本挖掘与普通的平面文本挖掘有类似之处，但是，Web 文本又有其自身的特点。例如，通信网中的短信、互联网中即时聊天工具和聊天室产生的聊天记录等文本具有每条记录包含字符少，而文本数量巨大的特点；BBS、Weblog 等形式的网页越来越多地出现了带有个人情感色彩的文章、言论，这些由用户产生的文本包含大量不规范用语、网络流行语等。这些特点对传统文本挖掘的方法提出了新的任务和挑战。本书将对 Web 文本挖掘中的若干关键问题，包括对短文本、Weblog 网页的聚类、情感分类、信息抽取和检索展开讨论。

1.2 文本挖掘相关技术概述及研究现状

本书中讨论的 Web 文本挖掘的关键问题涉及文本分类、聚类、信息抽取和检索等技术。

1.2.1 文本分类概述及研究现状

在文本分类方面，最初人们希望通过经验和专业知识对事物进行定性分析，即由专业人员手工编写和维护分类规则来进行分类。这类系统的典型例子是CONSTRUE系统[10]。手工方法的缺点是构建自动分类器时必须要为领域专家获取的知识和知识工程师的知识表示之间架起桥梁。如果这种分类器被应用到完全不同的领域，则相应工作必须被推倒重来。

20世纪90年代以来，随着信息存储技术和通信技术的迅猛发展，大量的文字信息开始以计算机可读的形式存在，而且其数量每天仍在急剧增加。在这种情况下，基于机器学习的文本分类逐渐取代了基于知识工程的方法，成为文本分类的主流技术。

文本分类是一个有监督的学习过程。它根据一个已经被标注的训练文本集合，找到文本特征和文本类别之间的关系模型，然后利用这种学习得到的关系模型对新的文本类别进行判断。我们可以更形式化地对文本分类过程进行描述。假设有一组文本概念类C和一组训练文本D。文本概念类和文本库中的文本可能满足某一概念层次关系h。客观上，存在着一个目标概念T，有：

$$T:D\to C \tag{1-1}$$

这里，T把一个文本实例映射为某一个类。对D中的文本d，$T(d)$是已知的。通过有监督地对训练文本集的学习，可以找到一个近似于T的模型H：

$$H:D\to C \tag{1-2}$$

对于一个新文本d_n，$H(d_n)$表示对d_n的分类结果。一个分类系统的建立或者说分类学习的目的就是寻找一个和T最相近似的H。即给定一个评估函数f，学习的目标应使T和H满足：

$$\mathrm{Min}\left(\sum_{i=1}^{|D|} f(T(d_i)-H(d_i))\right) \tag{1-3}$$

一般来讲，文本分类需要解决5个问题，或者说有5个步骤，如下所示。

(1)获取训练文本集

训练文本集选择是否合适对文本分类器的性能有较大影响。训练文本集应

该能够广泛地代表分类系统所要处理的客观存在的各个文本类中的文本。一般而言，训练文本集应是公认的经人工分类的语料库。

(2)建立文本表示模型

即选用什么样的文本特征和用怎样的数学形式组织这些文本特征来表征文本。这是文本分类中一个重要的技术问题。目前的文本分类方法和系统大多以词或词组作为表征文本语义的语言要素，表示模型主要有布尔模型和向量空间模型。

(3)文本特征选择

语言是一个开放的系统，作为语言的一种书面物化或者电子化的文本也是开放的。它的大小、结构、包含的语言元素和信息都是开放的，因此它的特征也是无限制的。文本分类系统应该选择尽可能少而准确且与文本主题概念密切相关的文本特征进行文本分类。

(4)选择分类方法

也就是说用什么方法建立从文本特征到文本类别的映射关系，这是文本分类的核心问题。常用的方法有朴素贝叶斯(Naïve Bayes，NB)[11~13]、k-近邻(k-Nearest Neighbor，kNN)[14]、类中心向量[15, 16]、回归模型[17]、最大熵模型(Maximum Entropy，ME)[18, 19]、支持向量机(Support Vector Machine，SVM)[20, 21]等。实际使用较多的是 k-近邻、最大熵模型和支持向量机，这几种方法分类效果不错，而且具有较强的稳定性。

(5)性能评估模型

即如何评估分类方法和系统的性能或者说分类结果。真正反映文本分类内在特征的性能评估模型可以作为改进和完善分类系统的目标函数。在文本分类中，使用什么评估参数取决于具体的分类问题。单标注分类问题(一个测试文本只属于一个类)和多标注分类问题(一个测试文本可以属于多个类)所使用的评估参数是不一样的。目前使用比较多的分类性能评估指标为查全率和查准率，这是来源于信息检索中的两个术语。

图 1-1 为文本分类主要步骤的示意图。

显然，特征选择、分类训练和测试构成了一个循环。根据测试结果，调整特征选择和分类训练的参数，使得分类器具备最佳的分类效果。

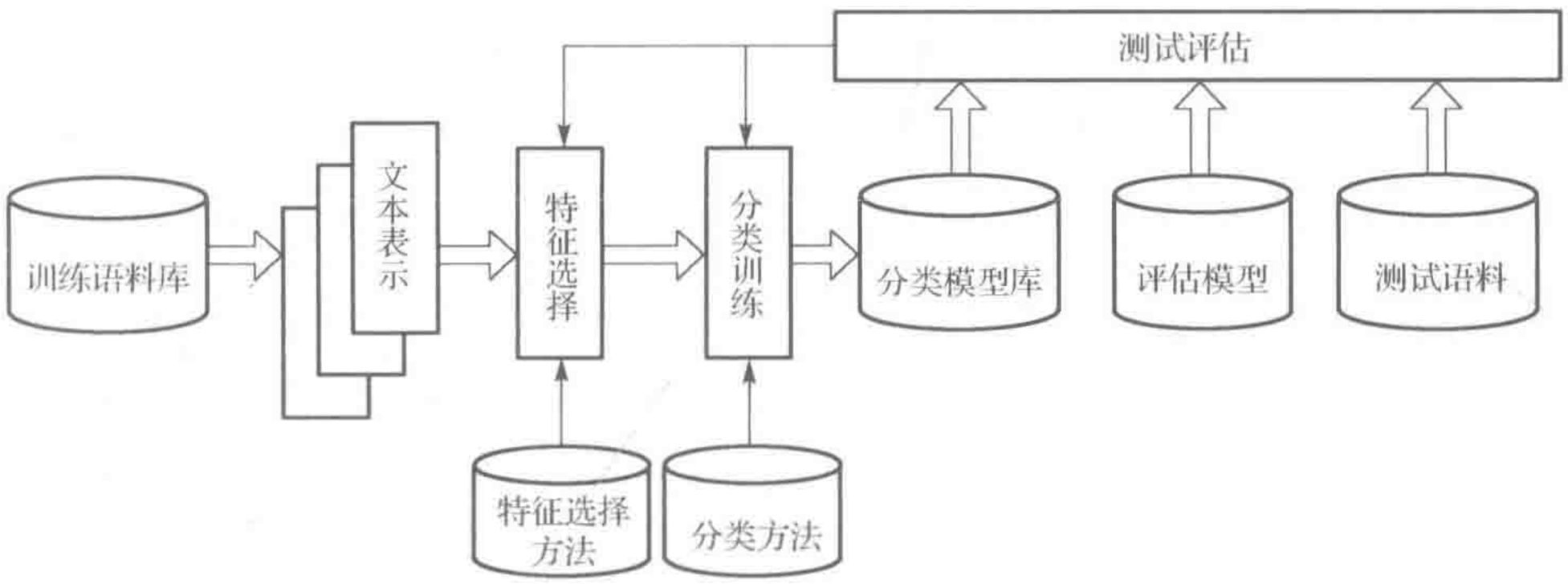

图 1-1 文本分类流程图

1.2.2 文本聚类概述及研究现状

将文本对象的集合分组成为由类似的文本组成的多个类的过程称为文本聚类。文本聚类是无监督学习过程。同一个簇中的文本彼此相似，不同簇中的文本相异。基于距离的聚类分析已经研究了许多年。许多成功的方法，如 *k*-means[22]、*k*-medoids[23]等，已经被加入到许多统计分析软件包或系统中，例如 S-Plus、SPSS、SAS 以及 MATLAB。随着信息化时代的到来，基于机器学习的文本聚类方法大行其道，包括平面划分方法、层次凝聚方法、基于 SOM 的方法、基于密度的方法、基于网格的方法、模糊聚类方法等已经得到了广泛的应用。

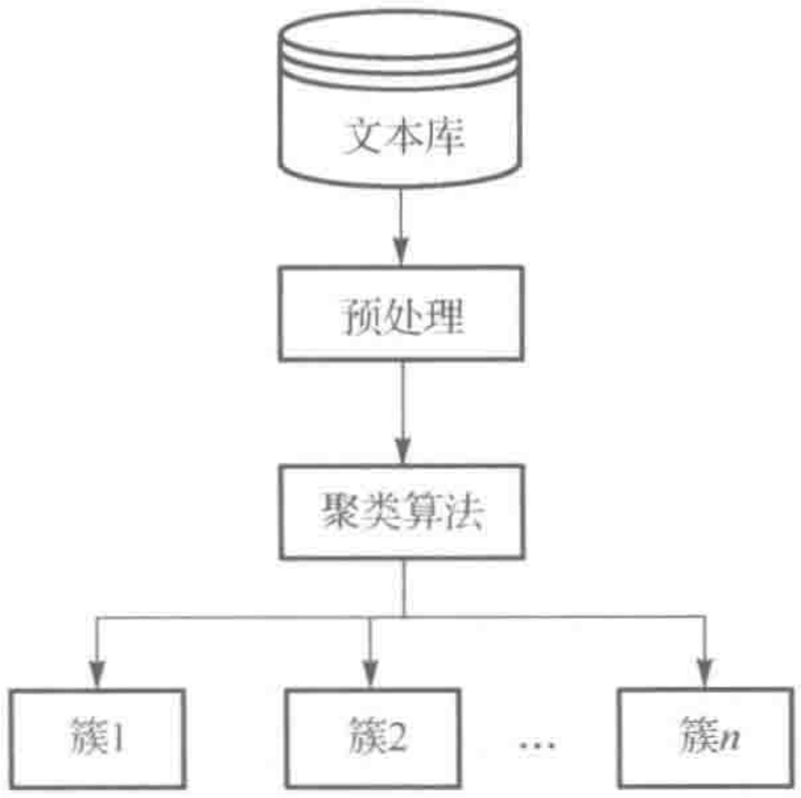

图 1-2 典型的文本聚类模型

典型的文本聚类模型如图 1-2 所示。文本聚类中，首先对文本库中的文本进行预处理得到文本表示，然后对文本表示利用各种聚类算法将其聚集成簇。聚类的指导原则是追求较高的类内相似度和较低的类间相似度。

文本聚类已经被广泛应用在很多地方，如在信息检索系统中用以提高信息检索的效率、组织搜索引擎返回的结果、帮助用户浏览超大规模的文本数据、生成 Web 文本的分类层次树、帮助用户管理和组织个人 Email、电子文档[24~26]等。

文本聚类同时又是一个非常难的问题。一方面是因为它没有任何预知信息，对所要划分的类别信息也是未知的，因而难以处理。另一方面，聚类算法和所要解决的问题密切相关。即有多少种具体问题，相应地便会有多少种为此而开发的聚类算法[27]，因此很难对不同的聚类算法进行客观、公正、科学的评价。但是在最近的文本聚类研究中，一个比较重要的趋势是，人们希望脱离原先基于语法层次的相似性聚类，得到能够理解文本内容的聚类方法。基于概念的文本聚类，以及最近基于语义、语用层次所做的研究[28~31]都是这一思想的体现。

1.2.3 信息抽取概述及研究现状

信息抽取(Information Extraction，IE)是指从一段文本中抽取指定的事件、事实等信息，形成结构化的数据并填入一个数据库中供用户查询使用的过程。

IE 最早开始于 20 世纪 60 年代中期，从自然语言文本中获取结构化信息，这被看作是 IE 技术的初始研究，它以两个长期的、研究性的自然语言处理项目为代表，即美国纽约大学的 Linguistic String 项目[32]和耶鲁大学的 Roger Schank 及其同事在 20 世纪 70 年代开展的有关故事理解的研究[33]。

从 20 世纪 80 年代末开始，信息抽取研究蓬勃开展起来，这主要得益于消息理解系列会议(Message Understanding Conference，MUC)的召开。MUC 的显著特点并不是会议本身，而在于对 IE 系统的评测[34]。各届 MUC 吸引了许多来自不同学术机构和业界实验室的研究者参加 IE 系统竞赛，每个参加单位根据预定的知识领域，开发各自的 IE 系统处理相同的文档库。官方评分系统对各家的结果进行统一评分。MUC 会议对 IE 这一研究方向的确立和发展起到了巨大的推动作用，MUC 定义的 IE 任务的各种规范以及确立的评价体系已经成为 IE 研究事实上的标准。

除此之外，正在推动 IE 研究进一步发展的动力主要来自美国国家标准技术研究所(NIST)组织的自动内容抽取(Automatic Content Extraction，ACE)评测会议[35]。这项评测研究的主要内容是自动抽取新闻语料中出现的实体、关系、事

件等内容，即对新闻语料中实体、关系、事件的识别与描述。

在国外，IE 的研究已经在某些特定领域达到实用化，目前已经有不少 IE 技术产品。具有代表性的 IE 系统包括：SRV 系统[36]、STALKER 系统[37]、PALKA 系统[38]等。中文信息抽取方面的研究起步较晚，目前主要的研究工作集中在对中文命名实体识别方面和基于 Web 方面的信息抽取，在设计实现完整的中文信息抽取系统方面还处在探索阶段。

其中，台湾大学参加了 MUC-7 中文命名实体识别任务的评测[39]。Intel 中国研究中心的 ZHANG 和 ZHOU 等人在 ACL-2000 上演示了他们开发的一个抽取中文命名实体以及这些实体间相互关系的信息抽取系统[40]，该系统利用基于记忆的学习算法获取规则用以抽取命名实体及它们之间的关系。

在基于 Web 的信息抽取方面，国内的研究如基于信息抽取的 Web 查询系统的设计与实现[41]、基于 DOM 的 Web 信息提取[42]、基于结点语义关系的信息抽取[43]、基于多层模式的多记录网页信息抽取[44]、基于 Ontology 的 Web 页面信息抽取[45]、对 Web 页面表格的信息抽取[46]等，这些都是充分利用了 Web 这种半结构化和它的 HTML 标记语言特征来达到抽取信息的目的。另外，北京大学计算语言所对中文信息提取也进行了较早的探讨。

1.2.4 文本检索概述及研究现状

信息检索(Information Retrieval，IR)研究如何从海量的信息资源中找出满足用户信息需求的信息子集，它涉及信息的表示、组织、存储、访问以及搜索等问题[7]。文本检索(Text Retrieval)与图像检索、音频检索一样都是信息检索的一部分。它主要研究如何从给定的无结构或半结构化文档集中找到与用户查询相关的文档子集，并依据相关度排序把检索结果返回给用户。如图 1-3 所示是典型的文本检索系统。

文本检索模型可以用一个四元组$\{D,Q,F,R(q_i,d_j)\}$来表示，其中：

- D是文本集中的文本逻辑表示；
- Q是用户信息需求(查询)的逻辑表示；
- F是一种文本与查询之间关系的模型；
- $R(q_i,d_j)$是排序函数，其函数值反映文档$d_j \in D$和查询$q_i \in Q$的相关程度。

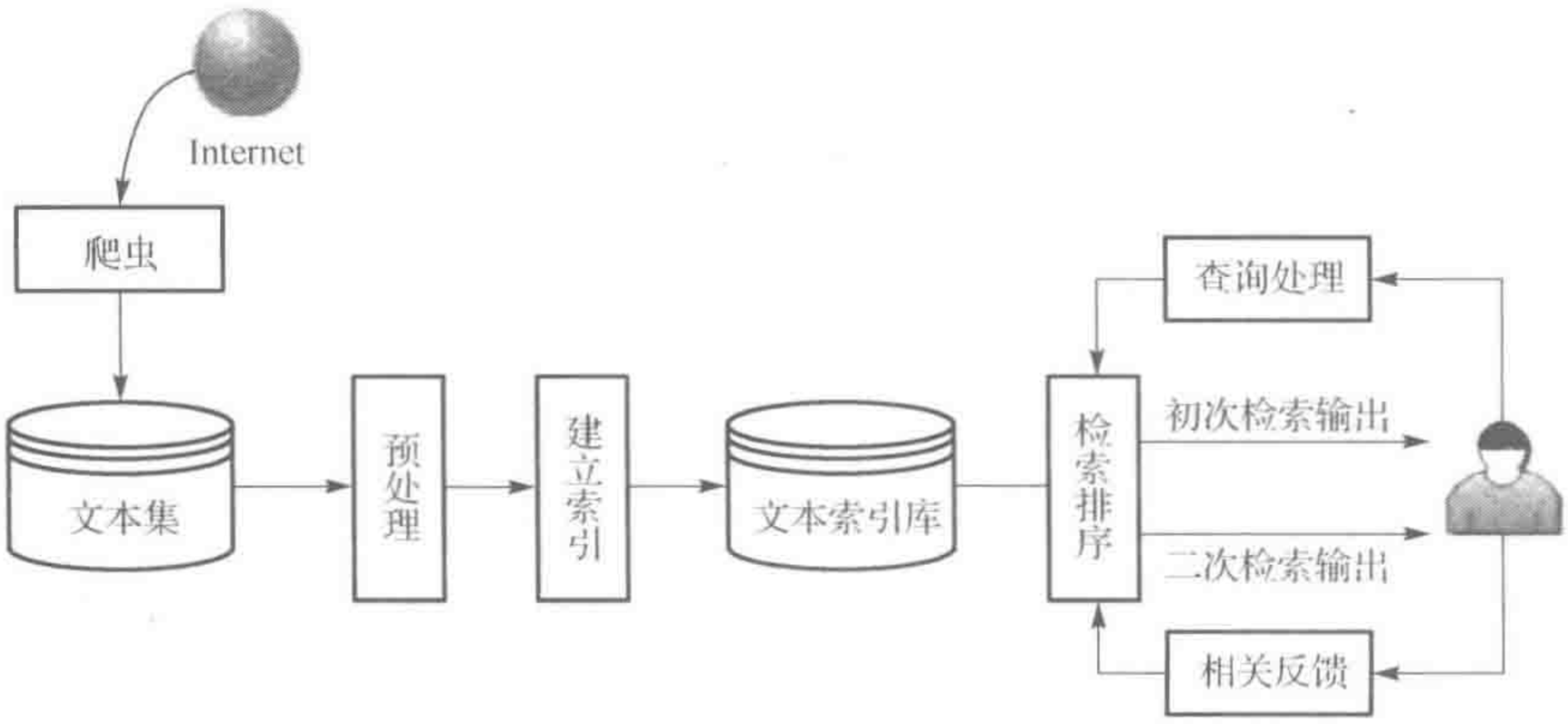

图 1-3　典型的文本检索系统

根据 D、Q、F、R 的不同定义，可以建立不同的文本信息检索模型，其中最常用的有布尔模型[7]、向量空间模型[47, 48]、概率模型[49, 50]和语言模型[51~53]等。比较经典的索引模型包括签名文件、倒排文件法、Patricia Tree 和互关联后继树等。

随着互联网上信息资源的惊人增长，越来越多的网民选择使用搜索引擎来查找自己所需要的资讯。搜索引擎的核心技术是信息检索，因此，信息检索，尤其是文本信息检索技术受到越来越多的关注。目前，信息检索领域最著名的国际会议之一是从 1992 年开始，由美国标准与技术研究所和美国高级研究规划局每年举行一次的文本检索会议(Text REtrieval Conference，TREC)[54]。该会议主要致力于超过百万文献的大型测试集的实验研究。在每年的 TREC 会议上，都会设计一系列信息检索中的专项任务(TRACK)，提供统一的评测标准和评测数据，对参加评测的系统进行评测。并设立论坛，让与会的各研究团体通过这些参考实验来对其检索系统进行比较以及交流学术成果。另一个重要会议是 NTCIR 会议(NII-NACSIS Text Collection for IR Systems)[55]。NTCIR 是日本国立情报学研究所针对亚洲语种(现包括中文、日文、韩文等)的文本信息检索、跨语言检索和相关的文本处理技术如文本摘要、文本抽取等进行评测的研究组织。

1.3　文本挖掘领域亟待解决的问题

如前所述，文本的分类、聚类、信息抽取、检索都有很大的应用价值。然而，Web 2.0 技术驱动下的用户创作内容、分享内容的网络信息模式的出现与普

及，给传统的文本挖掘带来巨大的挑战。随着研究的逐步深入，一系列制约技术发展和应用的关键问题也逐渐暴露出来。只有解决了这些关键问题，才能进一步推动 Web 文本挖掘技术的发展。下面介绍本书所关注的几个关键问题。

1. 短文本特征稀疏，用语不规范，数量巨大，语义计算困难

短文本特指那些长度非常短，一般在 200 个字符以内的文本形式，如常见的通过移动通信网络发送的手机短消息，通过即时通信软件发出的即时消息，在线聊天室的聊天记录，论坛中帖子的标题，网络日志的评论，互联网新闻的评论等[56, 57]。

短文本方式深刻改变了亿万中国人的沟通方式和生活习惯，各种形式的短文本已经成为中国各阶层普遍接受的信息沟通渠道和情感交流手段。形形色色的短文本语料中包含了网民对当前社会各种现象的立场和观点，话题涉及政治、经济、军事、娱乐、体育、卫生、科技、个人生活等各个领域。因此，短文本的研究对于语言演变与进化、话题检测与跟踪、舆情预警与疏导等方面都有重要价值。然而，短文本独特的语言特征导致短文本语言计算的困难，一般来说，短文本具有如下特点：

- 稀疏性。单条短文本长度非常短，样本特征稀疏。短文本通常只有几十字节大小，仅包含几个到十几个词典词语，很难准确抽取有效的语言特征。
- 数量庞大。短文本实时发送，实时接收，数量异常庞大。我国大陆地区每天就有大约 19.17 亿手机短信息发出，即时通信软件发出的即时消息数量更为庞大，这些都要求短文本的语言计算必须具有很高的处理速度。
- 不规范性。短文本表达简洁，用语极不规范，错误拼写、不规范用语及噪声非常多。不规范缩略语(如“brb”——be right back，“lol”——laughing out loud，“ft”——faint，“886”——再见)，不规范翻译(如“奋特”——faint，“晒”——share)，不规范网络用语(如“玉米”——李宇春迷，“笔迷”——周笔畅迷，“酱紫”——这样子)等在短文本语料中都非常普遍。

传统的文本挖掘技术已无法有效地对短文本进行处理[58]，因此，如何抽取短文本的有效特征成为短文本挖掘亟待解决的问题。