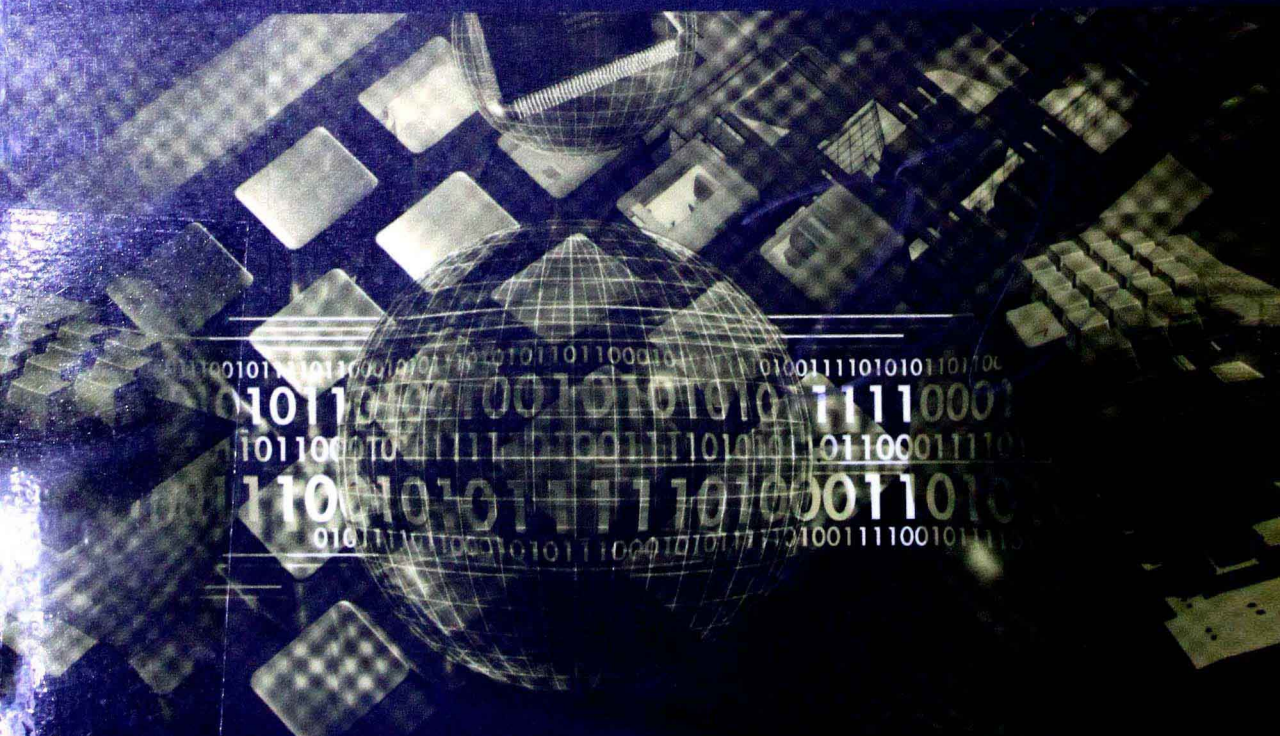
基于半监督与集成学习的文本分类方法

唐焕玲 ◎ 著



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

基于半监督与集成学习的文本分类方法

唐焕玲 著

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

内 容 简 介

文本分类技术广泛应用于新闻媒体、网络期刊文献、数字图书馆、互联网等领域，是人类处理海量文本信息的重要手段。

本书重点探讨了利用信息论中的评估函数量化特征权值的方法；基于权值调整改进 Co-training 的算法；利用互信息或 CHI 统计量构造特征独立模型，进行特征子集划分的方法；基于投票熵维护样本权重的 BoostVE 分类模型；融合半监督学习和集成学习的 SemiBoost-CR 分类模型。

其中特征选择和权值调整方法、基于特征独立模型划分特征子集的方法适用于文本分类，其他算法不仅适用于文本分类，对机器学习和数据挖掘的其他研究也有较大的参考价值和借鉴作用。

本书适合研究方向为文本挖掘、机器学习的硕士、博士研究生及相关专业技术人员学习和参考。

图书在版编目（CIP）数据

基于半监督与集成学习的文本分类方法 / 唐焕玲著. —北京：电子工业出版社，2013.8
ISBN 978-7-121-21256-7

Ⅰ. ①基… Ⅱ. ①唐… Ⅲ. ①文字处理—研究 Ⅳ. ①TP391.1

中国版本图书馆 CIP 数据核字（2013）第 188126 号

责任编辑：张 京
文字编辑：薄 宇
印 刷：三河市鑫金马印装有限公司
装 订：三河市鑫金马印装有限公司
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：900×1 280 1/32 印张：5.875 字数：205 千字
印 次：2013 年 8 月第 1 次印刷
定 价：29.00 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。

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：（010）88258888。

前　言

文本分类（Text/Document Categorization）是指按照预先定义的主题类别，通过一定的学习机制，在对带有类别标签的训练文本进行学习的基础上，给未知文本分配一个或多个类别标签的过程。文本分类技术广泛应用于新闻媒体、网络期刊文献、数字图书馆、互联网等领域，是人类处理海量文本信息的重要手段。数据挖掘技术在信息检索、邮件过滤、Web 个性化服务等领域的成功应用均在一定程度上依赖于准确的文本分类技术。因此，文本分类技术的相关研究一直是近年来国际学术界的研究热点。

本书对文本分类的关键技术进行了概述，阐述了基于半监督学习和集成学习的国内外相关研究，重点对基于半监督学习和集成学习的文本分类方法进行了深入探讨。

本书的第 1 章介绍了研究背景、文本分类及其面临的问题，阐述了基于半监督学习和集成学习的文本分类方法的研究意义和国内外研究现状。第 2 章对文本分类的关键技术进行了概述，主要包括文本预处理、文本的表示、特征选择、分本分类方法、实验数据集及分类模型的评估方法。第 3 章分析了特征选择存在的问题，采用信息论中的评估函数量化特征的重要性，调整特征的权值，提出 TEF-WA 权值调整技术；分析比较了文档频率、信息增益（Information Gain，IG）、期望交叉熵（Expected cross Entropy）、互信息（Mutual Information，MI）、χ^2 统计量（CHI）、文本证据权（Weight of Evidence for Text，WET）和几率比（Odds Ratio）等多种评估函数及实验结果。第 4 章分析了半监督学习中的代表方法 Co-training 算法，提出了利用 TEF-WA 技术对 Co-training 改进的算法 TV-SC 和 TV-DC，通过评估两个基分类器之间的差异性，可间接评估两个特征视图的独立性，并通过实验证明了所提方法的有效性。第 5 章针对 Co-training 方法的独立性假设问题，提出

了利用互信息（MI）或 CHI 统计量评估特征之间的相互独立性的方法，构造了一种特征独立模型（MID-Model）。基于该模型提出了特征子集划分方法——PMID 算法，以便把不存在自然划分的一个特征集合划分成两个独立性较强的子集，进而提出了改进的半监督分类算法——SC-PMID 算法。并且对由 PMID 算法划分得到的两个特征子集之间的独立性进行了理论论证。第 6 章分析了集成学习算法 AdaBoost 算法不能有效提升 Naïve Bayesian 分类器的原因，提出了基于投票信息熵和多视图的 AdaBoost 改进算法——BoostVE 算法，采用基于投票信息熵的样本权重维护新策略，能有效提高 Naïve Bayesian 文本分类器的泛化能力。理论分析证明改进的 BoostVE 算法的最小训练错误上界优于 AdaBoost。第 7 章基于半监督学习和集成学习，提出了置信度重取样的 SemiBoost-CR 分类模型，给出了基于最大差距和基于相似近邻两种置信度计算方法。实验表明利用少量标注样本和大量未标注样本，SemiBoost-CR 分类模型能够明显提升 Naïve Bayesian 文本分类器的性能指标。第 8 章介绍了采用 VC++ 6.0 实现的中英文文本分类系统 SECTCS，阐述了 SECTCS 系统的原有的功能与新扩展的功能、总体结构、主要的用户界面及操作。

本书的研究工作得到了山东省高校智能信息处理重点实验室（山东工商学院）、国家自然科学基金项目（No.61073133，No.61175053，No.61272369，No.61272244）及山东省优秀中青年科学家科研奖励基金计划项目（S2010DX021）的资助，特此表示感谢。

唐焕玲

2013 年 3 月

目 录

第1章 绪 论

1.1 研究背景及意义

1.1.1 数据挖掘和文本挖掘

随着信息技术和网络技术的迅速发展，网络数据规模呈指数增长，Internet 已发展成站点遍布全球的巨大信息服务网络，包含了涉及许多领域的丰富的信息资源。面对内容异构的海量信息，传统的数据分析方法只能获得数据的表层信息，无法获得数据属性的内在关系和隐含的信息，难以适应需求的不断发展。数据挖掘和知识发现（Data Mining & Knowledge Discovery in Database，DM&KDD）是 20 世纪 90 年代兴起的一门信息技术领域的前沿技术，它是在数据和数据库急剧增长远远超过人们对数据处理和理解能力的背景下产生的。

数据挖掘（Data Mining，DM）是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中采掘出隐含的、先前未知的、对决策有潜在价值的知识和规则[1]。知识发现（Knowledge Discovery in Databases，KDD）指识别出存在于数据库中有效的、新颖的、具有潜在效用的、最终可理解的模式的非平凡过程[2]。数据挖掘是一个交叉学科领域，受多个学科的影响，包括数据库系统、统计学、机器学习、可视化和信息科学等。此外，依赖于所用的数据挖掘方法及可使用的其他学科的技术，如神经网络、粗糙集理论、知识表示、归纳逻辑程序设计或高性能计算。依赖于所挖掘的数据类型或给

定的数据挖掘应用，数据挖掘技术也可能集成空间数据、信息检索、模式识别、图像分析、信号处理、计算机图形学、Web 技术、经济、商业、生物信息学或心理学领域的技术[1]。

传统的数据挖掘技术，主要针对的是结构数据，如关系的、事务的、数据仓库的数据。随着数据处理工具、先进数据库技术及网络技术的迅速发展，大量的形式各异的复杂类型的数据（如结构化与半结构化数据、超文本与多媒体数据）不断涌现。因此数据挖掘面临的一个重要课题就是针对复杂数据的挖掘，这包括复杂对象、空间数据、多媒体数据、时间序列数据、文本数据和 Web 数据。

文本挖掘是数据挖掘领域的一个分支，在国际上，文本挖掘是一个非常活跃的研究领域。从技术上说，它实际是数据挖掘和信息检索两门学科的交叉。文本挖掘与传统数据挖掘的差别在于文本数据与一般数据的巨大差异。传统数据挖掘所处理的数据是结构化的，如关系的、事务的、数据仓库的数据。其特征数通常不超过几百个，而文本数据没有结构，转换为特征矢量后特征数将达到几万甚至几十万。所以，文本挖掘既采用了很多传统数据挖掘的技术，又有自己的特性[5-14]。

近年来随着 Internet 的大规模普及和企业信息化程度的提高，信息积累越来越多，Internet 已经发展为当今世界上最大的信息库。Internet 上的信息，绝大多数是以网页形式存放的，而网页的内容又多以文本方式来表示，传统的信息检索技术已不适应日益增长的大量文本数据处理的需要。如何快速、准确地从来自异构数据源的大规模的文本信息资源中提取符合需要的简洁、精炼、可理解的知识，就要用到文本知识挖掘。Internet 的发展极大地促进了文本挖掘的发展。

文本挖掘（Text Mining，TM）：以计算语言学、统计数理分析为理论基础，结合机器学习和信息检索技术从文本数据中发现和提取独立于用户信息需求的文本集中的隐含知识。它是一个从文本信息描述到选取提取模式，最终形成用户可理解的信息知识的过程[6]。根据 KDD 的框架，结合文本挖掘的定义和特点，文本挖掘的过程示意图如图 1.1 所示。开始处是原始文本信息源，最终结果是用户获得的知识模式，经历了信息预处理→文本挖掘→质量评价三个主要阶段。

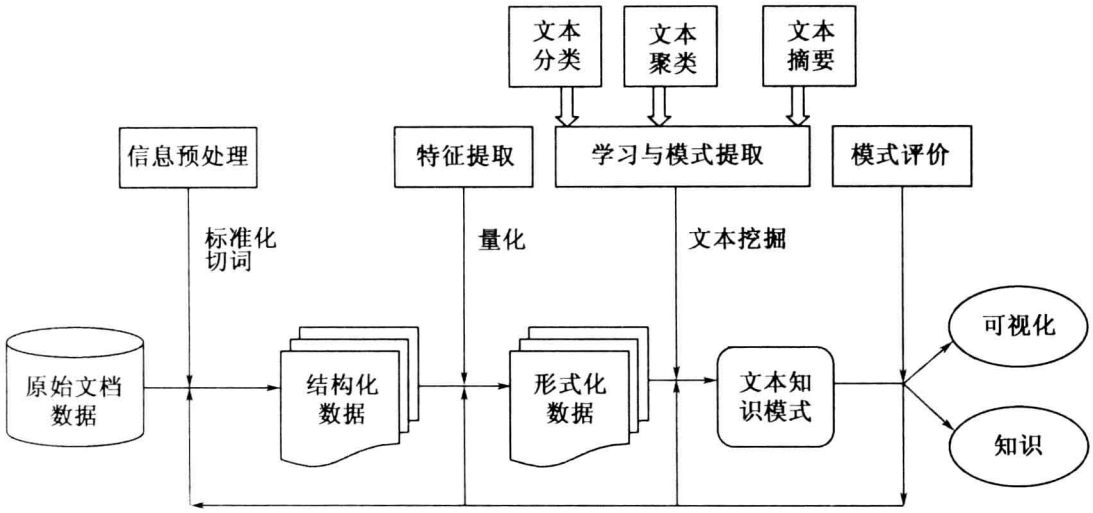


图 1.1　文本挖掘的过程示意图

1.1.2　文本分类及其面临的问题

文本分类（Text Categorization，TC）是文本挖掘中最重要的研究领域之一。对文本进行准确、高效的分类是许多数据管理任务的重要组成部分。数据挖掘技术在信息检索、邮件过滤和提供个性化的服务等方面，均在一定程度上依赖于准确的文本分类技术。

1．文本分类的定义

所谓文本分类，是指按照预先定义的主题类别 $C=\{c_1, c_2,\cdots, c_L\}$，通过一定的学习机制，在对带有类别标签的训练文本 $D=\{(x_1,y_1),\cdots,(x_N,y_N)\}$ 进行学习的基础上，给未知文本分配一个或多个类别的过程。其中 C 可以是并列的也可以分层次组织起来的。可以用一个目标函数 $\phi: D\times C \to \{T,F\}$ 来描述文本分类[1]，称 $\phi: D\times C \to \{T,F\}$ 为分类规则或假设或模型，对 $x_i \in D, c_j \in C$，$(x_i,c_j)\to T$ 表示 x_i 属于类别 c_j；而 $(x_i,c_j)\to F$ 表示 x_i 不属于类别 c_j。

这里，文本既可以是传统的纯文本，也可以是经过 HTML Parser 等网页

解析工具去掉网页标记、转换成纯文本的网页（Web 文档），网页可以看做是特殊的文本。Web 上的信息资源大多以 HTML 页面或 XML 页面的形式存在，与一般文本的表示不同，网上大量半结构化的文本及其之间的超链接提供了多于传统文本的有用信息，如标题、段落标题、超链接文字、链接及所用的字号等辅助信息为文本分类提供了更多的有用信息。根据 Web 网页的具体特性，一般可以从两个方面来选取网页特征项：① 通过提取网页内容中的关键词；② 利用网页中的有关标识符及其结构特征。Web 网页经过 HTML Parser 等网页解析工具，去掉网页标记，可转换成纯文本。

文本分类的过程一般包括文本预处理、特征约简、训练与分类、分类结果的评价和反馈等过程，如图 1.2 所示。

本书研究的基于半监督学习和集成学习的文本自动分类方法，既适用于纯文本的分类，也可应用于网页的分类。在后续的章节，没有特别说明时，文本泛指纯文本和经过预处理后的 Web 文档。

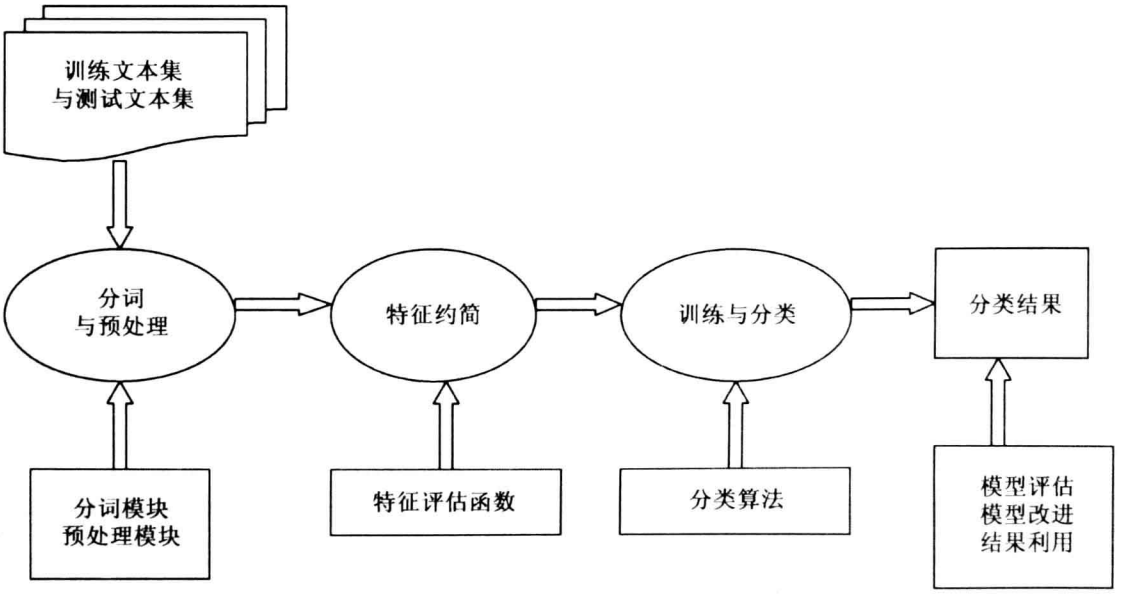


图 1.2　文本分类的一般过程

2. 文本分类的发展

文本分类是信息检索与数据挖掘领域的研究热点。1960 年 Maron 在 Journal of ASM 上发表了有关自动分类的第一篇论文，标志着文本分类技术

的诞生[8]，而 H. P. Luhn 在这一领域进行了许多开创性的研究工作。随后许多著名的情报学家如 K. Sparch、G. Salton 及 R. M. Needham 等都在这一领域进行了卓有成效的研究[9-11]。

文本自动分类在国外大体经历了三个发展阶段：

第一阶段（1960—1964），主要进行文本自动分类的可行性研究；

第二阶段（1965—1974），进行自动分类的实验研究；

第三阶段（1975 至今），自动分类进入实用化阶段，新方法和新系统层出不穷。

我国的自动分类工作也经历了这三个阶段，只是起步较晚。1981 年侯汉清先生在国内首次对文本自动分类进行探讨，此后国内一些科研院所也相继开展了文本分类研究，在分类理论和应用特别是中文文本分类方面取得了众多成果。2007 年侯汉清教授承担的国家社科基金项目“基于知识库的网页自动标引和自动分类研究”结题。

目前国外开展文本自动分类研究比较著名的机构包括卡耐基梅隆大学（CMU）、麻省理工学院（MIT）、加州大学伯克利分校、康奈尔大学、马里兰大学、微软剑桥研究院、微软亚洲研究院、IBM 研究中心、卡耐基集团等。国内比较活跃的单位有清华大学、北京大学、复旦大学、上海交通大学、哈尔滨工业大学、东北大学、北京邮电大学、中国科学院（计算所、软件所、计算机语言信息工程研究中心）、南京大学、纳讯科技公司、西风网站等。此外国内外还有大批研究机构和公司也在从事同类研究。

从文本分类使用的方法上说，主要有：① 20 世纪 80 年代基于知识工程和专家系统的文本分类模式；② 20 世纪 90 年代逐渐成熟的基于机器学习的文本分类方法，更注重分类器的模型自动挖掘和生成及动态优化能力，在分类效果和灵活性上都比之前基于知识工程和专家系统的文本分类模式有所突破，成为相关领域研究和应用的经典范例。

文本分类的主要方法包括决策树（Decision Tree）、K 近邻算法（K Nearer Neighbors Classifier）、线性分类器（Linear Classifier）、回归模式（Regression Models）、简单 Bayesian 网络（Bayesian belief Networks）、规则学习算法（Rule Learning Algorithms ）、BP 神经网络（BP Neural Networks）、归纳学习技术（Inductive Learning Techniques）、支持向量机（Support Vector Machine，SVM）等[15-33]。20 世纪 90 年代出现了基于集成学习的分类方法，即组合多个分类

器以克服单个分类器的不足，有效提高了分类的精度，已成为一个研究热点，Boosting 和 Bagging 方法是其中的代表算法[34-38]。

3. 文本分类面临的问题

文本分类技术能够较好地解决大部分具有数据量相对较小、标注比较完整及数据分布相对均匀等特点的应用问题，但是对大规模应用仍受到很多问题的困扰，目前面临着诸多挑战[33]，本书主要讨论以下两点。

（1）标注瓶颈问题

传统的有监督分类算法（Supervised Categorization Algorithm）需要提供足够的已标注训练样本（Labeled Data），但是已标注的训练样本集的建立需要专家知识，费时费力，代价高，获取困难，制约了分类模型的建立，致使许多实际问题的研究无法开展，这就是所谓的标注瓶颈问题。然而，互联网上存在大量未标注样本（Unlabeled Data），获取相对比较容易。因此，利用大量的未标注样本结合少量的标注样本的半监督分类（Semi-supervised Categorization）的研究引起了学术界比较广泛的关注。

针对标注瓶颈问题，基于半监督学习（Semi-supervised Learning）的分类方法是比较有效的解决方法[39-41] [47-56]。半监督学习在文本分类、图像分类、邮件过滤、机器翻译、主题词识别、词性标注等方面都有广泛的应用。基于半监督学习的分类称为半监督分类（Semi-supervised Categorization 或 Semi-supervised Classification），其研究重点在于使用大量的未标注（Unlabeled）样本，结合少量的标注（Labeled）样本训练生成分类器，提高分类器的性能指标。

半监督分类的研究虽然已经取得了不少成果，但是也存在许多值得探讨和亟待解决的问题。例如，半监督分类算法的应用存在一定的约束条件；半监督分类的精度还有待提高；半监督分类算法往往要付出大量的迭代代价，计算复杂度比较高等。如何提高半监督分类的精度、降低计算复杂度，是值得研究的问题。基于半监督学习的文本分类方法的研究，在理论和实践上都是比较有意义的研究方向。

（2）分类方法本身存在局限性

分类技术在其发展过程中出现了许多经典的分类方法，如决策树方法、Naïve Bayesian 学习方法、神经网络（Netware Net）、K 近邻法（K Nearest Neighbor）、支持向量机（Support Vector Machine，SVM）等，由于受到分类方法本身的局限性，这些经典方法的性能指标在原有基础上很难进一步提高[3][15-33]。因此，基于集成学习的分类方法，即组合多分类器来提高分类的精度成为学术界比较关注的另一个研究方向。

集成学习（Ensemble Learning）是一种机器学习范式，多个学习器的单独决策被以某种方式组合起来（通过加权或无权重投票）解决同一个问题[34-38]。集成学习技术已经在行星探测、地震波分析、Web 信息过滤、生物特征识别、计算机辅助医疗诊断等众多领域得到了广泛的应用。在文本挖掘领域，集成学习技术可用于文本分类、文本过滤等；在网络挖掘方面，集成学习技术在网页分类、信息检索和网络用户行为分析——偏好排序方面都有应用。由于集成学习技术可以有效地提高学习系统的泛化能力，因此成为国际机器学习界的研究热点，并被国际权威 T. G. Dietterich 称为当前机器学习四大研究方向之首。

鉴于文本分类面临的这两种挑战，利用少量标注样本（Labeled Data）和大量的未标注样本（Unlabeled Data）的半监督分类（Semi-supervised Categorization）和组合多个分类器来提高分类性能的集成学习（Ensemble Categorization）是近年来模式识别和机器学习领域的研究热点，也是本书研究的主要内容。

❑1.2 国内外相关研究

1.2.1 半监督学习

半监督学习是模式识别和机器学习研究领域的热点之一，在文本分类、图像分类、邮件过滤、机器翻译、主题词识别、词性标注等方面得到了广泛的应用。根据半监督分类算法的实现方式，现有的典型方法大致分为五种：

基于生成模型（Generative Model）的半监督分类方法[39-43]、自训练方法（Self-Training）[44-46]、协同训练方法（Co-Training）[47-56]、基于图的半监督分类方法[57-59]、直推式支持向量机方法（Transductive SVM，TSVM）[60-64]等。

1. EM 算法

Dempster、Laird 和 Rubin 提出的 EM（Expectation-Maximization）算法[40]是一种对不完整数据进行最大化参数估计的迭代算法。结合标注文本和未标注文本的信息进行半监督学习，未标注文本的类别可以看成缺失的值。Nigam 结合 EM 算法和 Naïve Bayesian 算法，从标注文本和未标注文本中学习，改进了 Bayesian 分类器的分类效果[41-42]，是基于生成模型（Generative Model）的半监督分类方法的典型代表。

初始时，只使用标注样本集，建立 Naïve Bayesian 分类器，然后重复交替执行 E 步骤和 M 步骤，达到使似然函数增加的目的[41]。直观地，EM 试图在未标注样本的分布上建立最大可能的分类假设，EM 算法可以看成未标注样本在初始的标注训练样本“周围”的聚类。

标注文本集和未标注文本集组成混合模型（Mixture Model），当标注文本集的数目远远小于未标注文本集时，EM 的参数估计在很大程度上来源于未标注文本集。当未标注文本集上的无监督聚类学习产生的类别与标注文本集的类别不一致时，反而会降低分类的正确性。EM-λ 算法是 Nigam 在基本的 EM 算法上，引入了一个参数$\lambda(0 \leqslant \lambda \leqslant 1)$，以调整未标注文本在 EM 算法中的权重[39][42]。

2. Self-training 算法

Self-training 首先使用由少量的标注样本训练生成分类器对未标注样本分类，选出置信度高的未标注样本，加上预测的类标记，添加到训练样本集中重新训练，迭代这个过程。Self-training 又称 Self-teaching 或 Bootstraping。为了避免一个分类器强化自己的错误，通常当置信度低于某个阈值时就停止迭代，或者使用 Co-training 方法。Self-training 在主题名词的识别[44]、emotional 和 non-emotional 的对话分类[45]、图像检测[46]等应用领域取得了比较好的效果。