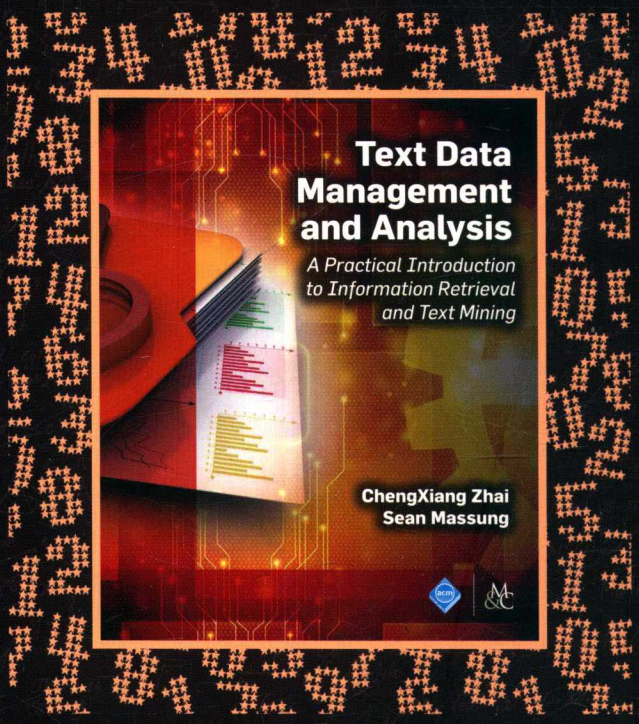
HZ BOOKS
华章IT

数据科学与工程技术丛书

文本数据管理与分析

信息检索与文本挖掘的实用导论

[美] 翟成祥（Chengxiang Zhai）
肖恩·马森（Sean Massung） 著
宋巍 赵鑫 李璐旸 李洋 等译
刘挺 审校



TEXT DATA MANAGEMENT AND ANALYSIS

A PRACTICAL INTRODUCTION TO INFORMATION RETRIEVAL AND TEXT MINING

机械工业出版社
China Machine Press

数据科学与工程技术丛书

TEXT DATA MANAGEMENT AND ANALYSIS

A PRACTICAL INTRODUCTION TO INFORMATION RETRIEVAL AND TEXT MINING

文本数据管理与分析

信息检索与文本挖掘的实用导论

[美] 翟成祥（Chengxiang Zhai）
肖恩·马森（Sean Massung） 著
宋巍 赵鑫 李璐旸 李洋 等译
刘挺 审校

机械工业出版社
China Machine Press

目 录

Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining

第一部分

Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining

概述和背景

第1章

Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining

绪　论

在过去的20年里，我们经历了在线信息的爆炸性增长。根据加利福尼亚大学伯克利分校2003年的一项研究："……世界每年产生1～2EB(exabyte，艾字节)的不同信息，这对于地球上的男人、女人和孩子来说，每人大约250MB(megabyte，兆字节)。各类印刷文档仅占总量的0.03%。"[Lyman等2003]

大量的在线信息是文本信息(即自然语言文本)。例如，根据上面引用的伯克利的研究："报纸每年发表25TB(terabyte，太字节或称兆兆字节)内容，杂志发表10TB内容……办公文档包含195TB内容。据估计，每年发送的电子邮件总数达到6100亿封，包含11 000TB信息。"当然，还有博客文章、论坛帖子、推文、科技文献以及政府文件等。Roe[2012]将电子邮件数量从2003年的6100亿封更新为2010年的107万亿封。根据IDC最近的一份报告[Gantz和Reinsel 2012]，2005～2020年，数字宇宙将增长300倍，规模达130EB～40000EB。

一般来说，各种类型的在线信息都是有用的，但由于以下原因，文本信息起着特别重要的作用，可以说是最有用的一种信息。

文本(自然语言)是人类知识最自然的编码方式。因此，大多数人类的知识都是以文本数据的形式编码的。例如，科学知识几乎都整理在科学文献中，而技术手册包含如何操作设备的详细说明。

文本是人们遇到的最常见的信息类型。事实上，一个人每天产生和消费的大部分信息都是文本形式的。

文本是最具表达能力的信息形式。它可以用来描述其他媒体，如视频或图像，甚至图像搜索引擎(如Google和Bing支持的图像搜索引擎)也经常依靠匹配图像周围的文本来检索"匹配"用户关键字查询的图像。

网络文本信息的爆炸式增长强烈需要能够提供以下两种相关服务的智能化软件工具，帮助人们管理和利用文本大数据。

文本检索。文本数据的增长使得人们无法及时消费数据。由于文本数据对我们积累的大部分知识进行了编码，因此通常不会被丢弃，从而导致大量文献数据的积累，这些文献数据现在超出了任何个人能够处理的能力范围，即便只是简单浏览。在线文本信息的快速增长也意味着没有人能够消化每天产生的所有新信息。因此，迫切需要开发智能文本检索系统，以帮助人们快速、准确地获取所需的相关信息。这种需求促进了近期网络搜索行业的迅猛发展。事实上，像Google和Bing这样的网络搜索引擎现在已经成为我们日常生活中不可或缺的一部分，每天都有数以百万计的查询。通常，在大量文本数据存在的地方，搜索引擎都是有用的(诸如桌面搜索、企业搜索或特定领域中的文献搜索，例如PubMed)。

文本挖掘。文本数据是人类为了交流而产生的，所以它们通常含有丰富的语义内容，并且通常包含有价值的知识、信息、观点和个人的喜好。它们提供了很多机会来发掘对于许多应用有用的各种知识，特别是关于人类意见和偏好的知识。这些知识通常直接在文本

数据中表达。例如，现在人们习惯于通过产品评论、论坛讨论和社交媒体文本等包含主观见解的文本数据来获取有关他们感兴趣的话题的观点，并优化各种决策任务，如购买产品或选择一项服务。同样，由于信息的巨大规模，人们需要智能的软件工具来帮助发现相关知识以优化决策或帮助他们更有效地完成任务。尽管支持文本挖掘的技术还没有像支持文本获取的搜索引擎那么成熟，但近年来在这方面已经取得了显著的进展，专业的文本挖掘工具已经在许多应用领域得到了广泛使用。

4

结构化数据采用定义良好的模式，使计算机处理起来相对容易，与结构化数据相比，文本没有明显的结构，所以计算机在上述智能软件工具开发过程中需要处理和理解文本编码的内容。目前的自然语言处理技术还没有达到使计算机能够精确理解自然语言文本的水平(这也是人类往往应该参与到处理过程的主要原因)，但是采用许多不同的统计和启发式方法来管理和分析文本数据已经在过去的几十年中得到了发展。它们通常非常健壮，可以用于对任何自然语言、任何主题的文本数据进行分析和管理。本书旨在对其中的一些方法进行系统介绍，重点介绍构建各种实际有用的文本信息系统所需的最有用的知识和技能。

上面讨论的两种服务(即文本检索和文本挖掘)在概念上对应于分析任何“大规模文本数据”过程中的两个自然步骤，如图 1-1 所示。尽管原始文本数据可能很大，但是具体的应用通常只需要少量最相关的文本数据，因此在概念上，任何应用的第一步应该是根据具体的应用或者决策去识别相关的文本数据，避免对大量不相关文本数据做不必要的处理。将原始大规模文本数据转换成规模更小但高度相关的文本数据的第一步通常是在用户帮助下利用文本检索技术来完成(例如，用户可以使用多个查询来收集所有相关文本数据以用于决策问题)。在这第一步中，主要目标是将用户(或应用程序)与最相关的文本数据连接起来。

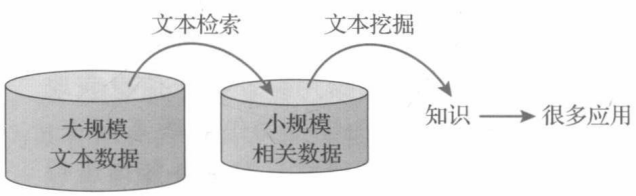


图 1-1　文本检索与数据挖掘是分析大规模文本数据的两项主要技术

5

一旦获得一个小规模的最相关文本数据，我们需要对文本做进一步的分析来帮助用户消化文本数据中的知识和模式。这是文本挖掘的一个步骤，其目标是从文本数据中进一步发现知识和模式，以支持用户的任务。此外，需要对任何发现的知识的可信度进行评估，所以用户通常需要返回到原始的文本数据中去获得用来解释所获得知识的上下文，并通过上下文信息验证知识的可信度。因此，作为主要用于文本获取的搜索引擎系统，也必须在任何基于文本的决策支持系统中提供知识来源。因此，这两个步骤在概念上是交错的，一个完整的智能文本信息系统必须在一个统一的框架中进行整合。

值得指出的是，在“大数据”的背景下，文本数据与其他类型的数据是非常不同的，因为它通常是由人类直接生成的，通常也意味着要被人类消费。相反，其他数据往往是机器生成的(例如通过使用各种物理传感器收集的数据)。由于人类可以比计算机更好地理解文本数据，所以人类参与挖掘和分析文本数据的过程绝对是至关重要的(比其他大数据应用程序更为必要)。如何最佳地将人与机器之间的工作分开从而优化人与机器之间的协作，以最少的人力来最大化其“智能组合”，是所有文本数据管理和分析应用中的一个普遍挑战。以上讨论的两个步骤可以被认为是文本信息系统协助人类的两种不同的方式：信息检索系统帮助用户从大量的文本数据中找到解决具体应用问题所需的最相关文本数据，从而有效地将大规模原始文本数据转换成可以被人类更容易处理的规模较小的相关文本数据；

而文本挖掘应用系统可以帮助用户分析文本数据中的模式，以提取和发现对于完成任务或进行决策直接有用的、可操作的知识，从而为用户提供更直接的任务支持。

从这个角度，我们将本书所涵盖的技术分成两部分来匹配图 1-1 所示的两个步骤，然后再用一章讨论如何将所有的技术整合到统一的文本信息系统中。本书试图从实践的角度全面介绍信息检索和文本数据挖掘的主要概念、技术和思想。其中包括许多亲身实践的练
6 习，使用配套软件工具 MeTA 来帮助读者学习如何将信息检索和文本挖掘技术应用于真实世界的文本数据，并学习如何实验和改进一些有趣的应用任务的算法。本书可作为计算机科学方向的本科生和研究生以及图书馆和信息科学工作者的教材，也可作为从事分析和管理文本数据等相关工作的技术人员的参考书。

1.1　文本信息系统的功能

从用户的角度来看，文本信息系统(TIS)可以提供三种不同但相关的功能，如图 1-2 所示。

信息获取(information access)。这种能力使用户可以在需要时获取有用的信息。有了这个能力，文本信息系统可以在正确的时间连接正确的信息和正确的用户。例如，搜索引擎使得用户能够通过查询来获取文本信息，而推荐系统可以在发现可用的新信息项目时将相关信息推送给用户。由于信息获取
7 的主要目的是将用户与相关信息联系起来，提供这种能力的系统通常只对文本数据进行最小限度的分析，只需满足将相关信息与用户的信息需求匹配，而原始信息项目(例如，网页)通常以其原始形式交付给用户，但是也经常提供项目的摘要。从文本分析的角度来看，用户通常需要阅读信息项目来进一步消化和利用所传递的信息。

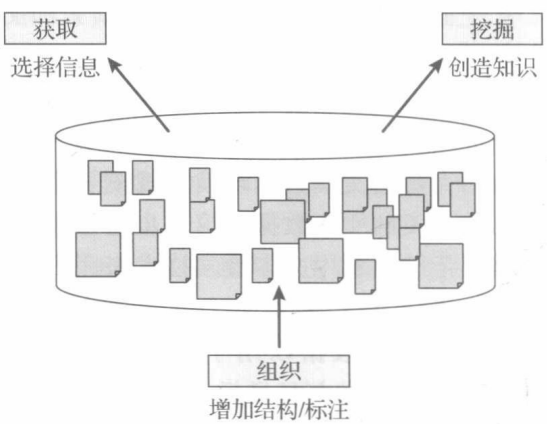


图 1-2　信息获取、知识获取和文本组织是文本信息系统的三个主要功能，文本组织对信息获取和知识获取起到支撑作用，而知识获取也常被称为数据挖掘

知识获取(knowledge acquisition)或文本分析(text analysis)。这种能力使得用户能够获得文本数据中蕴含的有用知识。如果没有对大规模的数据进行合成和分析，用户通常不容易获得这些知识。文本信息系统可以分析大量的文本数据以发现文本中隐藏的有趣模式。具有知识获取能力的文本信息系统可以被称为分析引擎。例如，搜索引擎可以将产品的相关评论返回给用户，分析引擎可以使用户直接获得关于产品的主要的正面或负面意见，并比较人们对多个类似产品的意见。提供知识获取能力的系统通常需要更详细地分析文本数据，综合来自多个文本文档的信息，发现有趣的模式，创造新的信息或知识。

文本组织(text organization)。此能力使系统能够用有意义的(主题)结构来注释一组文本文档，从而可以连接分散的信息，使用户可以根据该结构在信息空间中巡览。虽然这样的结构可以被认为是从文本数据中获得的“知识”，并且直接对用户有用，但是通常它们仅用于促进信息获取或知识获取，或者兼而有之。从这个意义上说，文本组织的能力在文本信息系统中起到了支持作用，使得信息获取和知识获取更加有效。例如，添加的结构可以允许用户使用结构上的约束进行搜索，或者根据结构进行浏览。考虑到结构的约束，结

构也可以用来进行详细的分析。

信息获取可以进一步分为两种模式：拉取和推送。在拉取模式下，用户主动从系统中“拉”出有用的信息，在这种情况下，系统是被动的，等待用户提出请求，然后系统用相关信息回应。当用户具有临时信息需求(即临时需要关于产品的意见)时，这种信息获取模式通常非常有用。例如，像 Google 这样的搜索引擎通常为用户提供拉取模式信息获取。 8
在推送模式下，系统主动向用户“推”(推荐)它认为对用户有用的信息。当用户具有相对稳定的信息需求(例如，一个人的爱好)时，推送模式常常工作良好；在这种情况下，系统可以“预先”知道用户的偏好和兴趣，从而能够向用户推荐信息而不需要用户采取主动。本书涵盖了两种信息获取模式。

拉取模式还包括两种互补的方式让用户获得相关信息：查询和浏览。在查询的情况下，用户通过(关键字)查询指定信息需求，系统将该查询作为输入并返回估计与查询相关的文档。在浏览的情况下，用户简单地沿着将信息项目链接在一起的结构进行巡览并逐渐地获得相关信息。由于查询也可以被看作是一步即导航到一组相关文档，很显然，浏览和查询可以自然地交织。事实上，网络搜索引擎的用户通常交错进行查询和浏览。

从文本数据中获取知识通常是通过文本挖掘过程来实现的。文本挖掘可以被定义为挖掘文本数据以发现有用的知识。数据挖掘社区和自然语言处理(Natural Language Processing，NLP)社区都开发了文本挖掘的方法，但两个社区对这个问题的看法往往略有不同。从数据挖掘的角度来看，我们可能将文本挖掘视为挖掘一种特殊的数据，即文本。遵循数据挖掘的总体目标，文本挖掘的目标自然会被视为发现和提取文本数据中的有趣模式，其中可能包括潜在主题、主题趋势或异常值。从 NLP 的角度来看，文本挖掘可以被看作是理解自然语言文本的一部分，将文本转化为某种形式的知识表示，并基于提取的知识进行有限的推理。因此，一个主要的任务是执行信息抽取(information extraction)，它的目标是识别和提取所涉及的各种实体(例如人员、组织和位置)及其关系(例如谁与谁见面)。实际上，任何文本挖掘应用都可能涉及模式发现(即数据挖掘角度)和信息抽取(即 NLP 角度)。信息抽取丰富了文本的语义表示，使得模式发现算法能够生成语义上更有意义的模 9
式，而不是直接处理文本的字或字符串表示。由于我们的重点是介绍可以用于各种文本数据的具有通用性和健壮性的技术，这些技术不需要太多的人力；而信息抽取技术更具有语言相关性，并且一般需要很多人工参与，所以我们在本书中普遍采用数据挖掘的角度。然而需要强调的是，信息抽取是任何文本信息系统的重要组成部分，它试图支持更深入的知识发现或语义分析。

文本挖掘的应用可以被分类为直接应用和间接应用。直接应用中被发现的知识将被用户直接消费，而间接应用中发现的知识不一定直接对用户有用，但可以通过提供更好的支持间接地帮助用户获取信息。知识获取也可以基于发现了什么知识来进一步分类。然而，由于“知识”涉及的范围广泛，所以不可能使用少量的类别来覆盖所有的形式。尽管如此，我们仍然可以在本书中找出几个常见的类别。例如，可以发现的一种知识类型是一组隐藏在文本数据中的主题或子主题，它们可以作为文本数据中主要内容的简明摘要。另一种可以从用户生成的主观性文本中获得的知识是关于某个主题的观点的总体情感极性。

1.2 文本信息系统的概念框架

从概念上讲，文本信息系统可能由几个模块组成，如图 1-3 所示。

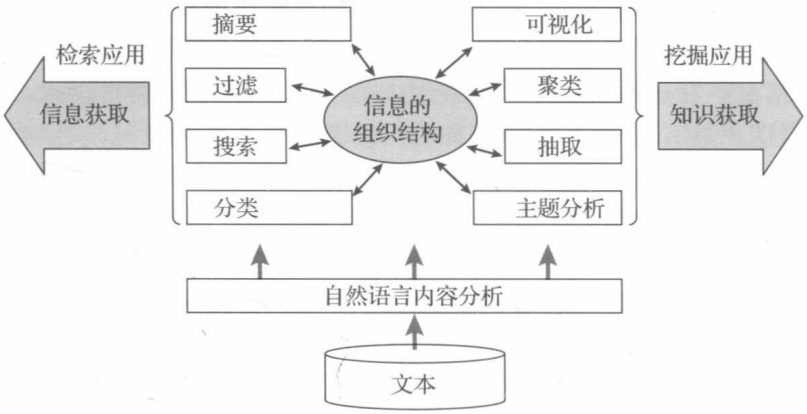


图 1-3　文本信息系统的概念框架

首先，需要基于自然语言处理技术的内容分析模块。该模块允许系统将原始文本数据转换为更有意义的表示，以便在搜索引擎中可以更有效地与用户的查询匹配，在文本分析中可以更有效地进行处理。目前的 NLP 技术主要依赖于统计机器学习，以有限的语言知识作为辅助，进行不同深度的文本数据理解；浅层技术是健壮的，但更深层的语义分析只适用于非常有限的领域。一些文本信息系统能力(如摘要)会比其他能力(如搜索)需要更深的 NLP。大多数文本信息系统使用非常浅的 NLP，其中文本将被简单地表示为“词袋”，词是表示的基本单位，并且文本中词的顺序被忽略(尽管保留了词频)。然而，也可以使用更复杂的表示，可以基于识别出的实体、关系或其他更深层的文本理解技术。

以内容分析为基础，文本信息系统中有多个组件以不同的方式帮助用户。以下是管理和分析文本信息的一些常见功能。

搜索(search)。接收用户查询并返回相关文档。文本信息系统中的搜索组件通常称为搜索引擎。网络搜索引擎是最有用的搜索引擎之一，它使用户能够有效和高效地处理大量的文本数据。

过滤/推荐(filtering/recommendation)。监督传入的数据流，确定哪些项目与用户的兴趣相关(或不相关)，然后向用户推荐相关项目(或者过滤掉不相关的项目)。根据系统是否侧重于识别相关项目或不相关项目，这个组件可以被称为推荐系统(其目标是向用户推荐相关项目)或者过滤系统(其目标是过滤掉非相关项目，允许用户只保留相关项目)。文献推荐器和垃圾邮件过滤器分别是推荐系统和过滤系统的典型例子。

10
~
11

分类(categorization)。将文本对象划分到一个或多个预定义类别，其中类别可根据应用程序而变化。文本信息系统中的分类组件可以用各种有意义的类别对文本对象进行注释，从而丰富了文本数据的表示，进一步提升了文本分析的效率和深度。类别也可用于组织文本数据，便于文本访问。将文章分类为一个或多个主题类别的主题分类器和将句子分类为正面、负面或中性的情感极性的情感标注器都是文本分类系统的具体例子。

摘要(summarization)。对一个或多个文本文件生成一个简要的内容摘要。摘要减少了人们消化文本信息的负担，也可以提高文本挖掘的效率。生成摘要的组件称为摘要器。新闻摘要和意见摘要都是摘要器的实例。

主题分析(topic analysis)。提取并分析给定文档集合的主题。主题直接促进了用户对文本数据的理解，并支持浏览文本数据。当与相关的非文本数据如时间、地点、作者等元数据相结合，主题分析可以产生许多有趣的模式，如主题的时间趋势、主题的时空分布和

作者的主题概况。

信息抽取(information extraction)。从文本中提取实体、实体之间的关系或其他“知识单元”。信息抽取组件可以构建实体关系图。这种知识图有多种用途，包括支持导航(沿着图的边和路径)以及进一步应用图挖掘算法去发现有趣的实体关系模式。

聚类(clustering)。发现相似文本对象(例如术语、句子及文档等)的群组。聚类组件在帮助用户探索信息空间方面起着重要的作用。它使用经验数据来创建有意义的结构，这对于浏览文本对象和快速理解大型文本数据集都非常有用。它对识别无法与其他对象聚集的异常对象也是非常有用的。

可视化(visualization)。以可见的方式显示文本数据中的模式。可视化组件对于吸引人们参与发现有趣模式的过程非常重要。由于人类非常善于识别视觉模式，所以将各种文本挖掘算法产生的结果可视化有很大需求。 12

以上各项也是本书后面将要讨论的主题的纲要。具体来说，第二部分内容讨论文本数据获取，其中将介绍搜索和过滤；第三部分内容讨论文本分析，将介绍分类、聚类、主题分析和摘要。本书没有介绍信息抽取，因为我们希望重点介绍容易应用于任何自然语言文本数据的一般方法，而信息抽取通常需要特定语言的相关技术。由于本书着重于算法，所以可视化也没有涵盖。但是，必须强调的是，信息抽取和可视化是与文本数据的分析和管理密切相关的重要话题。对这些技术感兴趣的读者可以在本章最后的书目说明中找到一些有用的参考资料。

1.3　本书结构安排

本书分为四部分，如图 1-4 所示。

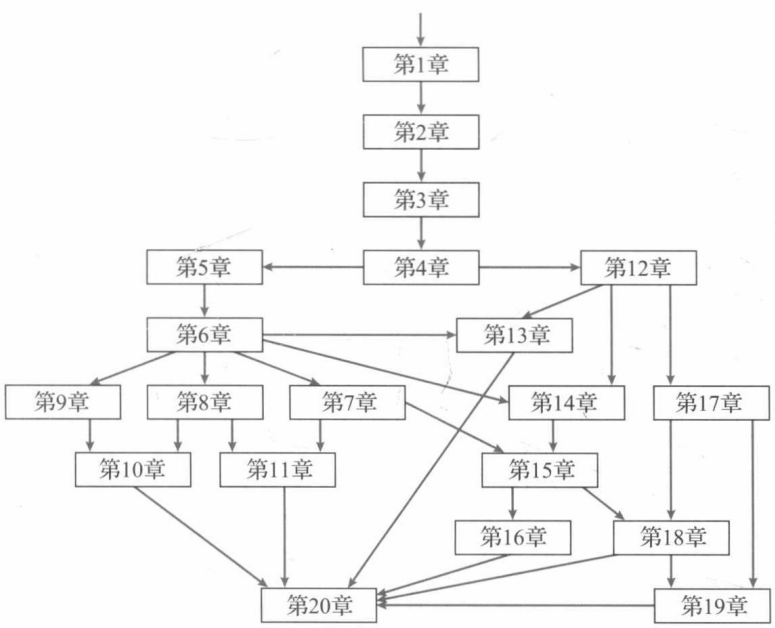


图 1-4　各章之间的依赖关系

第一部分　概述和背景。本部分由前 4 章组成，提供了本书概述和背景知识，包括了读者可能不熟悉的但对理解某些书籍内容而言必备的基本概念，以及这本书中用于练习的 MeTA 工具包的简介。本部分还简要概述了自然语言处理技术。自然语言处理是所有文本

数据分析应用为理解文本数据和获取有信息的文本表示所必需的技术。

第二部分　文本数据获取。本部分由第5～11章组成，涵盖了支持文本数据获取的主要技术。本部分对信息检索基本技术进行了系统讨论，包括将检索任务作为面向查询的文档排序问题(第5章)，构成搜索引擎排名函数设计基础的检索模型(第6章)，反馈技术(第7章)，检索系统的实现(第8章)以及检索系统的评价(第9章)。它涵盖了迄今为止最重要的信息检索应用——网络搜索引擎(第10章)，其中引入了用于分析文本数据中的链接以提高文本对象排名的技术，并简要介绍了利用监督机器学习结合多个特征的排序学习技术。这部分的最后一章(第11章)介绍了推荐系统，它提供了信息获取的“推送”模式，这不同于典型的搜索引擎支持的“拉取”模式的信息获取(即需要用户主动查询)。

第三部分　文本数据分析。本部分由第12～19章组成，涵盖了各种分析文本数据的技术，以方便用户消化理解文本数据，并发现文本数据中有用的主题或其他语义模式。第12章从数据挖掘的角度给出了文本分析的概述，可以将文本数据看作是由作为世界的“主观感受器”的人类生成的数据；这个观点使我们可以在更一般的数据分析和挖掘的背景下看待文本分析问题，并且有助于讨论文本和非文本数据的联合分析。接下来的多个章节，涵盖了许多非常有用的通用技术，可以不依赖或仅依赖最少的人工来分析文本数据。具体而言，第13章讨论了用于发现文本数据中的词汇单元之间的两种基本语义关系的技术，即聚合关系和组合关系，这可以被看作是发现用于生成文本数据的自然语言知识(即语言知识)的示例。第14章和第15章分别介绍了两种紧密相关的技术，即文本聚类和文本分类，用于生成有意义的结构或注释并与其他无组织的文本数据相关联。第16章讨论了文本摘要，能够帮助人们快速消化文本信息。第17章详细讨论了用于发现和分析文本数据中主题模式(即主题模型)的一类重要概率方法。第18章讨论分析文本数据中表达的情感和观点的技术，这些技术是在分析人们所产生的文本数据的基础上发现关于他们的偏好、观点和行为等知识的关键。最后，第19章讨论了文本和非文本数据的联合分析，因为在大数据分析中使用尽可能多的数据来获得知识和智能是有益的，许多应用程序均需要这样的技术。

13～14

第四部分　统一的文本管理和分析系统。最后一部分由第20章组成，我们试图说明本书中讨论的所有技术如何在概念上集成到一个基于操作器的统一框架中，从而有可能在一个统一的文本管理和分析系统中实现，用于支持各种不同的应用。本部分还作为进一步扩展MeTA的路线图，为各种应用提供有效和普遍的高级支持，并讨论如何将MeTA与许多其他相关的现有工具包整合起来，特别是与搜索引擎系统、数据库系统、自然语言处理工具包、机器学习工具包和数据挖掘工具包相结合。

由于我们试图从实践的角度来看待所有的话题，书中的大多数概念和技术的讨论都是非形式化的、直观的。为了满足一些读者可能有兴趣更深入了解某些主题的需求，本书附录对几个重要的主题提供了详细和严格的解释。

1.4　如何使用本书

15

由于我们希望涵盖的主题范围非常广泛，必须在覆盖范围的广度和深度之间进行许多折中。在进行这种折中的时候，我们选择强调文本数据挖掘的基本概念和实用技术的覆盖，代价是不能涵盖许多先进技术的细节，并且在很多章节的最后给出一些参考文献帮助有兴趣的读者更多地了解这些先进的技术。我们的希望是，在阅读本书的基础上，你将能够通过自己或通过其他资源了解更先进的技术。我们还选择了涵盖更多的文本管理和分析

的一般技术，并支持适用于任何自然语言的任何文本的技术。我们讨论的大多数技术都可以在没有任何人力或仅需要最小人力的情况下实施，这与文本数据的一些更详细的分析不同，特别是使用自然语言处理技术。这样的“深度分析”技术显然是非常重要的，对于一些我们想深入理解文本细节的应用来说，这确实是必要的。然而，在这一点上，这些技术往往是不可扩展的，往往需要大量的人力。在实践中，结合这两种技术将是有益的。

我们预想的读者(可能重叠)主要有三类。

学生。本书专门提供使用真实文本挖掘工具和应用程序的实践经验。如果单独使用，我们建议先阅读第1～4章，以便对本书的前提知识有一个很好的理解。第1～3章将帮助你熟悉和理解未来章节所必需的概念和词汇。第4章介绍每个章节的练习中使用的配套工具包MeTA。我们希望练习和章节描述为你自己的文本挖掘项目提供灵感。MeTA提供的代码是一个良好的技术支撑，让你更专注于你的贡献。

如果在课堂上使用，教师可以选择采取几个逻辑流程。作为必备知识，我们假定一些概率和统计方面的基本知识以及用C++或Java等语言编程的能力已经具备。MeTA是用现代C++编写的，但是一些练习只要通过修改配置文件就能完成。

教师。本书涵盖了具有逻辑性、彼此关联的主题，便于与各种课程体系结合。例如，本书的第一部分和第二部分可以作为本科的信息检索导论课程，重点介绍搜索引擎如何工作。练习假定学生具备基本的编程经验以及一些概率以及统计学的数学背景。也可作为通识课程，概览全书作为“文本数据挖掘”的入门，而跳过第二部分中更具体的搜索引擎实 16
现和特定于Web的应用程序的章节。另一种选择是将所有部分作为研究生教科书的补充内容，其中仍然强调实用的编程知识，可以结合每章中的参考文献阅读。研究生的练习可以要求实现MeTA的参考文献中提及的很多方法。

每章末尾的练习除了题目，还给学生提供了一个功能强大但易于理解的文本检索和挖掘工具包。在一个以编程为中心的课堂上，强烈鼓励使用MeTA练习。编程作业可以通过在每章中选择练习的子集来创建。由于工具包的模块化特性，可以通过扩展现有系统或实现目前MeTA默认不包括的其他众所周知的算法来创建新的编程实验。最后，学生可以使用他们通过练习学到的MeTA组件来完成一个更大的最终编程项目。将不同的语料库与工具包结合使用会产生不同的项目挑战，例如评论摘要与情感分析。

从业者。大多数行业读者很可能会用这本书作为参考，我们也希望这本书可以为你的工作带来启发。与学生用户的建议一样，阅读最初的3章就可以得到本书的大部分内容，然后可以选择一个与目前兴趣相关的章节深入研究或更新知识。

由于MeTA中的许多应用程序可以简单地通过配置文件使用，我们预计它是一个无须编程即可快速处理数据集并获得基线结果的工具。

每章的章后练习可以被认为是手头特定任务的默认实现。可以选择将MeTA纳入你的工作，因为它使用宽松的免费软件许可证。事实上，它拥有麻省理工学院和伊利诺斯大学/NCSA双重许可。当然，我们仍然鼓励并邀请读者分享对MeTA的修改、扩展和改进，为了所有读者的利益着想，这些改进不应作为专利。

不管你的目标是什么，我们希望这本书对你有帮助。我们感谢你提出的意见和建议，从而改进这本书。谢谢阅读！ 17

书目说明和延伸阅读

在信息检索(IR)领域已经有多部优秀的教科书。由于信息检索方面的研究历史悠久，

而且在20世纪60年代已经做了大量基础性的工作，即使van Rijsbergen[1979]、Salton和McGill[1983]和Salton[1989]等一些非常古老的书籍在今天也非常有用。另一个有用的早期书籍是Frakes和Baeza-Yates[1992]。更近的包括Grossman和Frieder[2004]、Witten等[1999]和Belew[2008]。最近的是Manning等[2008]、Croft等[2009]、Büttcher等[2010]、Baeza-Yates和Ribeiro-Neto[2011]。与这些书相比，本书对信息检索主题有更广泛的认识，并试图涵盖文本检索和文本挖掘的内容。虽然现有的一些关于IR的书籍也涉及一些话题，如文本分类和文本聚类(我们将其归类为文本挖掘)，但是没有一本书包含了对主题挖掘和分析的深入讨论，对于文本挖掘这是非常有用的一类重要技术。现有的关于IR的书籍似乎缺少推荐系统，我们将其视为另外一种支持用户进行文本获取的方式，从而与搜索引擎相辅相成。更重要的是，本书更为系统化地处理所有这些主题，并通过统一的概念框架来管理和分析大型文本数据。本书还尝试通过提供许多练习的配套工具包来尽量减小算法中的抽象解释和实际应用之间的差异。希望了解更多关于IR研究历史和早期重要里程碑的读者应该关注Sparck Jones和Willett[1997]中的阅读材料。

文本挖掘的主题也被许多书籍所覆盖(例如，Feldman和Sanger[2007])。这本书与那些书的主要区别是我们强调文本挖掘和信息检索的整合，相信任何文本数据应用系统都必须让人参与其中，而搜索引擎是任何文本挖掘系统的基本组成部分，以支持两个关键功能：1) 数据来源——帮助将大的原始文本数据集合转换成更小但更相关的文本数据集，从而可以使用文本挖掘算法有效地分析；2) 知识起源——帮助用户验证文本挖掘算法用于发现知识的原始文本文章。因此，本书更全面地涵盖了开发大型文本数据应用所需的技术。

本书的重点是介绍易于应用到任何自然语言的任何文本数据的通用和健壮的算法，这
18 些算法往往不需要或仅需要最少的人力。因此，一个不可避免的代价是它缺乏对文本挖掘中个别关键技术的覆盖，特别是信息抽取(IE)技术。我们决定不覆盖IE技术是因为它经常是语言相关的，而且需要并不简单的人工工作。另一个原因是，许多IE技术依赖于监督机器学习方法，这在许多现有的机器学习书籍(例如参见[Bishop 2006]和[Mitchell 1997]中已有很好的介绍。有兴趣了解更多关于IE技术的读者可以从查阅书籍[Sarawagi 2008]和评论文章[Jiang 2012]开始。

从应用的角度来看，由于强调对模型和算法的覆盖，本书缺少的另一个重要话题是信息可视化。然而，既然每个应用系统都必须具有用户友好的界面，以便用户与系统进行最佳的交互，所以那些对开发文本应用系统感兴趣的读者肯定会发现学习更多关于用户界面设计的内容是有用的。可以参考Hearst [2009]作为入门，它对信息可视化有较为详细的介绍。

最后，由于强调广度，本书没有深入介绍任何组件算法。为了了解一些主题的更多信息，读者可以进一步阅读自然语言处理的书籍(例如，Jurafsky和Martin[2009]，Manning和Schütze[1999])、关于IR的高级书籍(如Baeza-Yates和RibeiroNeto[2011])以及关于机
19 器学习的书籍(如Bishop[2006])。你可以在每章末尾的书目说明中找到与特定主题相关的
~ 更具体的阅读建议。
20

背　景

本章主要涉及学习本书其他内容需要掌握的背景知识。对于已经熟悉这些基本概念的读者，可以选择性地跳过部分小节，也可以跳过整章内容。首先介绍一些概率论和统计学的基本概念，这些概念在本书的大部分算法和模型中均有涉及。接着概述文本挖掘应用中经常用到的一些信息论概念。最后一节介绍机器学习的基本思想和问题设置，尤其是有监督的机器学习，它对文本分类和基于文本的预测非常有用。总体来说，机器学习被广泛应用到了很多信息检索和数据挖掘任务。

2.1　概率和统计基础

正如本章以及其他章节的内容所述，概率和统计模型在文本挖掘算法中发挥着重要作用。本节将为读者提供足够的背景信息和专业概念，以便理解本书后面章节所涵盖的概率和统计方法。

概率分布是对某个概率空间 Ω 中事件可能性赋值的一种方法。例如，假设概率空间是一个六面骰子，每一面都有不同的颜色，那么 $\Omega=\{$红，橙，黄，绿，青，蓝，紫$\}$，掷一次骰子并观察颜色就是一个事件。

我们通过为所有可能的事件定义一个概率分布来量化掷骰子的不确定性。假设有一个无偏的骰子，掷出任意一种颜色的可能性都是 $\frac{1}{6}$，或者说大概 16%。这样，可以将概率分布表示成一个概率的集合：

$$\theta=\left\{\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}\right\}$$

其中第一项对应 p(红色)$=\frac{1}{6}$，第二项对应 p(橙色)$=\frac{1}{6}$，以此类推。但是如果骰子有偏，情况会是什么样呢？可以用一个不同的概率分布 θ' 来表示相应事件的可能性：

$$\theta'=\left\{\frac{1}{3},\frac{1}{3},\frac{1}{12},\frac{1}{12},\frac{1}{12},\frac{1}{12}\right\}$$

在这种情况下，相比其他颜色，红色和橙色被认为更容易被掷出。注意样本空间 Ω 和用来量化不确定性的概率模型 θ 之间的差别。在文本挖掘任务中，通常会利用与 Ω 有关的一些信息来估计 θ，不同的估计方法将决定一个概率模型的准确性和可用性。

现在引入下面的记号：

$$x\sim\theta$$

它表示随机变量(random variable) x 取自(服从)概率分布 θ。随机变量 x 对应 Ω 中每个元素的概率都由 θ 确定。例如，如果有 $x\sim\theta'$，那么 x 是红色或橙色的概率为 $\frac{2}{3}$。

在文本应用任务中，样本空间 Ω 通常为文本语料库的词汇表 V。例如，词汇表可以是

$$V=\{\text{a, and, apple, …, zap, zirconium, zoo}\}$$

我们用一个概率分布 θ 对文本数据进行建模。因此，给定一个单词 w，可以将其对应的概