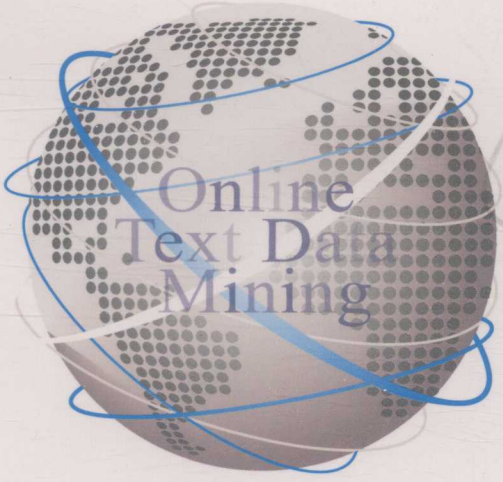
全方位解读大数据背景下的文本分析核心算法
助力打造下一代AI

在线文本数据挖掘

算法原理与编程实现

刘通◎著



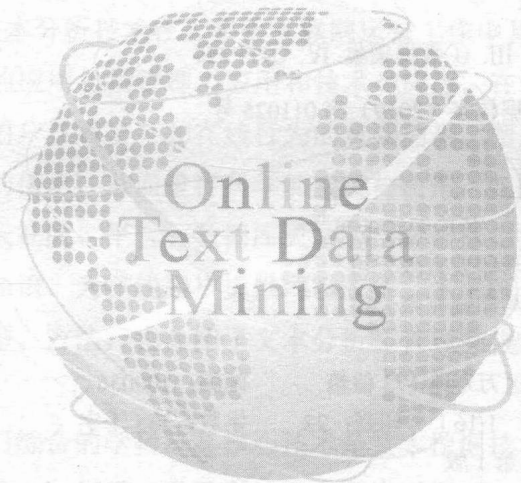
要点涵盖5大基础领域、6大应用场景
120个**模型**展示语义分析，技术场景化、具象化
精练150多个**图表**，快速掌握核心原理
通过60多个经典**案例**，高效引导商业实践
基于百万级语料分析**数据**，实战分析全面精准

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

在线文本数据挖掘

算法原理与编程实现

刘通◎著



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

内 容 简 介

为了满足大数据环境下网络运营与管理的需求，本书详细而系统地介绍了有关文本分析的核心技术与方法。本书基于统计分析、数据挖掘、机器学习等计算机技术，介绍了如何对在线环境的文本内容进行建模与分析，同时介绍了文本分析技术的具体应用场景。本书并非是纯粹的技术类书籍，而是一本教授读者如何更好地应用技术的实践手册。

本书分为 13 章，内容主要包括 3 个方面：①文本分析概要，包括概述、预备知识；②文本分析的基础类方法，包括文本建模、文本分类、文本聚类、序列标注；③文本分析的应用类方法，包括信息检索、文本摘要、口碑分析、社交网络分析、深度学习与 NLP、实证研究。

本书内容丰富、详略得当，结构清晰、系统。阅读本书需要读者具备一定的统计学知识和与数据挖掘相关的基础知识。本书特别适合对文本分析技术感兴趣的学生、科研工作者，以及数据分析类职业的工作人员阅读和参考。

图书在版编目（CIP）数据

在线文本数据挖掘：算法原理与编程实现 / 刘通著. —北京：电子工业出版社，2019.8
ISBN 978-7-121-35632-2

Ⅰ. ①在… Ⅱ. ①刘… Ⅲ. ①数据采集 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 011025 号

责任编辑：张　毅　　特约编辑：田学清
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：720×1000　1/16　印张：22　字数：388 千字
版　　次：2019 年 8 月第 1 版
印　　次：2019 年 8 月第 1 次印刷
定　　价：88.00 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。

本书咨询联系方式：（010）57565805。

作者简介

刘　通

上海交通大学安泰经济与管理学院管理科学与工程专业博士，现任国内知名金融机构算法工程师。作者曾在管理学领域和计算机科学领域的国内核心期刊上发表过诸多与文本挖掘应用技术相关的重要研究成果。此外，作者在就读博士期间，将技术理论有效结合实践，采用文本挖掘方法对用户在线医疗平台上的医生选择和医疗在线服务购买的决策行为进行了深入研究，对如何改进在线平台的盈利模型并提高市场收益提出有价值的管理学见解。作者还曾参与过上海交通大学与华为技术有限公司合作的分布式计算平台的深度学习算法选型研究，比较了包括Theano、Tensorflow、Deeplearning4j在内的不同深度学习平台在分布式环境下的性能与算法兼容情况。

前　言

在大数据时代，数据的价值开始被推上各行各业的舞台。人们更注重从海量的数据中挖掘感兴趣的信息，以实现丰富的技术应用，进行科学的管理决策。在互联网环境中，数据的分析与利用尤为重要，尤其是数值类型数据的分析和文本类型数据的分析。其中，文本类型数据的分析比一般数值类型数据的分析复杂，文本类型数据是大数据 4V 特征的具体体现，其相关技术也更具难度。尽管如此，文本类型数据在整个网络中的信息占比仍十分庞大，且对用户的各种在线交互、活动及购买行为也有着不容小觑的影响。因此，网络中的文本类型数据具有十分重要的分析价值。本书将重点对当今文本类型数据的重要分析技术进行详细、系统的介绍。

在应用方面，文本分析技术在大多数互联网运营工作中具有重要的实践意义。基于文本分析技术的应用包括管理类应用和技术类应用。在管理类应用中，文本分析可以有效提取用户在线交互和在线行为的重要信息，帮助管理者更好地掌握用户、产品、市场的信息，从而进行科学的建模与决策；在技术类应用中，文本分析可以充分从在线社区、平台、数据库大量的文本数据中提取、解析、创造用户感兴趣的信息与知识，为在线用户提供内容服务。本书既介绍了与文本分析密切相关的理论、模型、方法，也介绍了文本分析在管理类应用与技术类应用等具体场景中的实现。

文本分析是一门综合的学科，其核心技术是文本挖掘技术。文本挖掘技术与传统的数据挖掘技术一脉相承，是数据挖掘在语言学领域中的应用。从事文本分析的数据分析者不仅需要掌握丰富的数据处理、建模及挖掘方法，还需要掌握语言学知识、社会学知识，也需要充分理解语言产生的背景、应用和使用语言信息的用户对象。文本数据比一般的数值数据更容易体现人类的感情与行为，其相应的技术也具备更高的智能化程度，因此，在任何领域，掌握文本分析技术对数据分析者来说都是一个不小的挑战。

近些年，随着整个信息社会对文本数据重视程序的提升，以及计算机软硬件技术的飞速发展，文本分析领域的研究成果形成井喷式爆发。由于篇幅所限，本书虽然无法全面讲解文本分析的所有前沿技术，但是仍然尽可能地将所有经典的、有代表性的研究成果展现给大家，使从事文本分析的工作人员、科研人员及文本分析技术的爱好者能够高效而系统地对整个文本分析领域有一定的了解。阅读本书后，希望读者能够具备基于文本分析技术的能力，从而解决工作中的各种文本分析问题，并能深刻地认识到文本分析为互联网领域及整个社会带来的实践价值。

本书特色

1. 内容丰富，系统全面，详略得当

本书内容涵盖了当前大部分主流的文本分析技术与方法，笔者按照自身的知识体系对其进行了细致的归纳与梳理，并由浅入深地向读者进行了系统的介绍。本书内容详略得当，突出了知识的重点、难点。书中内容依托于数据
技术，但不拘泥于技术本身，在介绍相关技术理论时注重向读者教授
法及思维方式，帮助读者掌握技术的核心理念，从而使读者做到灵活应
入思考、举一反三、即时实践。

2. 行文通俗易懂，随意而不失严谨，有利于读者快速吸收理解

本书在介绍知识时，尽可能地用通俗易懂的语言对技术细节进行描述，而
是生硬地对学术文献中的定义、规范和公式进行搬运。对于很多技术难点，笔者均赋予了自身的思考和感悟，并用生动而接地气的语言进行了转述。

本书中所有方法和理论都具有翔实可靠的学术依据，是科学而严谨的，所介绍的方法和技术也都得到了学术上的广泛认可和接受。本书还在特定的位置附注了关键知识点的学术来源，以供感兴趣的读者进一步进行知识的补充、考证。

3. 图文并茂，配备实例，有趣生动

本书虽是一本技术类书籍，但在排版风格上力争做到图文并茂，以增加读者的阅读兴趣，提高读者对于知识的理解效率。一图胜千字，本书中很多文本分析中重要的技术流程采用了示意图的表述方式，这可以有效地对知识点进行串联与总结。

此外，对于很多分析方法，本书还介绍了其具体应用场景，以及具体技术实现。这样，读者不仅掌握了知识的核心理念，根据具体实例也知道了如何运用知识。本书在知识结构上，可大致分成基础篇和应用篇，基础篇重点讲述理论方法，而应用篇偏向于知识在具体场景中的技术实现。本书在知识点设计方面更加生动灵活，有效地保证了文本分析技术的落地与推广。

本书内容及体系结构

第 1 章　概述

本章详细谈论了大数据时代下互联网公司的机会与挑战，介绍了在线文本分析技术在网站运营中重要的战略性地位。本章还基于大数据背景，从 4V 角度介绍了文本分析的主要技术特征。本章内容可以帮助读者更好地了解在线文本分析总体的知识框架和体系。

第 2 章　预备知识

本章引入了与在线文本分析密切相关的理论知识。首先，介绍了文本挖掘的主要任务，并介绍了与其相关的一些重要理论知识，如文本语义分析与语法分析、文本的结构化分析与标准化分析。其次，介绍了机器学习的基本概念，阐述了机器学习与深度学习的关系。对于机器学习，本章涉及的技术要点主要包括概率图模型、判别式模型、产生式模型、机器学习模型求解，以及模型过拟合。

第 3 章　文本建模

本章介绍了文本分析的基本任务——文本建模，即科学而有效地将非结构化的文本类型数据转换为可以直接进行数据分析与挖掘的数值类型数据。本章介绍了文本建模的主要应用场景，并从语言学建模和统计学建模两个主要方面对相关技术进行了详细介绍。

第 4 章　文本分类

本章所讨论的文本分类方法主要是对文档对象进行分类。本章从文本分类的基本概念、应用场景及分类特征优化等方面对文本分类的技术进行了系统的介绍。本章介绍了三类重要的分类模型：朴素贝叶斯模型、向量空间模型、支持向量机模型。

第 5 章　文本聚类

本章介绍了对文档对象进行聚类描述的主要技术方法，主要涵盖了扁平式聚类和凝聚式聚类两大基本问题解决思路。本章还介绍了如何对聚类结果进行分析，以及对聚类的特征进行优化等相关内容。对特殊文本对象的聚类技术的介绍也是本章的重点内容，具体包括半监督聚类、短文本聚类及流数据聚类。

第 6 章　序列标注

序列标注是特殊的分类问题，很多文本分析任务都需要抽象成序列标注问题进行解决。本章介绍了当前三类重要的序列标注基础模型，即隐马尔可夫模型、最大熵马尔可夫模型及条件随机场。本章还介绍了各模型的主要特征、优点和缺点，并提供了具体的应用范例。

第 7 章　信息检索

本章介绍了如何根据用户的特定信息需求，从在线环境中有效地提取重要的文本对象并进行反馈。除了介绍信息检索的重要应用场景，本章还讨论了三类主流的模型方案：基于空间模型的信息检索、基于概率模型的信息检索、基于语言模型的信息检索。

第 8 章　文本摘要

本章介绍了如何基于已有文本内容对信息进行压缩，并从中提取有价值的、关键的文本要素。文本摘要技术包括关键词提取和关键句提取，前者是本章介绍的重点。本章还介绍了很多经典的对词汇的关键词进行量化评估的指标，同时介绍了当前主流的基于图模型的关键词提取算法。

第 9 章　口碑分析

本章介绍了如何从在线平台的用户评论文本数据中提取有价值的产品信息。一方面，本章讨论了如何通过词典或语料集合对在线评价对象进行提取；另一方面，本章介绍了如何在不同的粒度水平上挖掘用户对于产品或服务的情感态度。

第 10 章　社交网络分析

社交网络是重要的互联网应用场景。本章介绍了很多社交网络上的文本分析

任务及具体的技术方案，包括社交网络的虚拟社区发现、用户影响力分析、情感分析、话题发现与演化，以及信息检索。本章还介绍了如何将社交网络的多属性特征和图结构特征有机地结合到文本分析技术框架中。

第 11 章　深度学习与 NLP

本章介绍了当前热门的深度学习技术在文本分析中的应用。深度学习以神经网络为基础模型。本章分别介绍了基于多层感知器模型和循环神经网络的深度学习文本分析技术。对于循环神经网络，本章特别介绍了词嵌入模型和机器翻译技术。

第 12 章　实证研究

本章介绍了文本分析技术在互联网领域中的管理类应用，讲述了如何通过实证研究来挖掘在线平台上的用户行为，并结合研究结果有针对性地提供管理决策建议。本章还介绍了文本分析技术在互联网医疗中的具体应用，以真实的场景、数据为依托，为从事互联网运营相关工作的读者提供了有价值的解决问题的思路。

第 13 章　总结

作为结束语，本章简要回顾了全书的核心内容，并为文本分析领域的工作者提供了若干条有价值的实践经验。

本书读者对象

- 从事数据分析、文本分析相关职业的技术人员、网络运营人员；
- 所学专业与计算机技术、互联网技术、语言学相关的本科生及研究生；
- 计算机科学、自然语言处理等领域的大学教师及科研工作者；
- 其他对文本分析有兴趣爱好的人员。

目 录