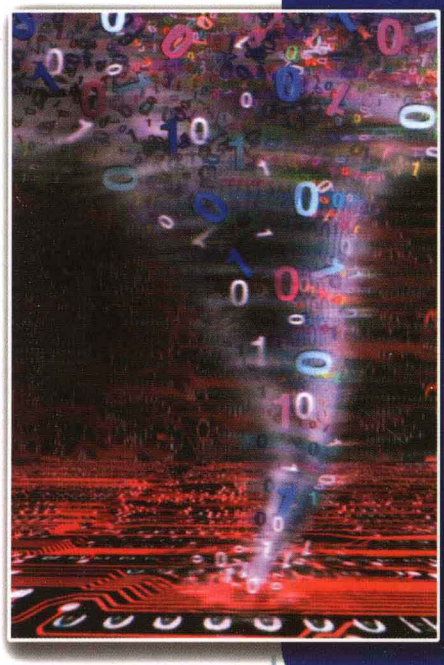
Springer

预测性文本挖掘基础



Fundamentals of Predictive Text Mining

〔美〕绍洛姆·韦斯
〔澳〕霓廷·因杜尔亚 著
〔美〕张潼
赵仲孟 侯迪 译

Sholom M. Weiss Nitin Indurkhya Tong Zhang

西安交通大学出版社
XI'AN JIAOTONG UNIVERSITY PRESS

Fundamentals of Predictive Text Mining

预测性文本挖掘基础

〔美〕绍洛姆·韦斯
〔澳〕霓廷·因杜尔亚　著
〔美〕张　潼

Sholom M. Weiss
T. J. Watson Research Center, IBM Corporation

Nitin Indurkhya
School of Computer Science & Engg. University of New South Wales

Tong zhang
Dept. Statistics, Hill Center Rutgers University

赵仲孟　侯　迪　译

西安交通大学出版社
Xi'an Jiaotong University Press

陕西省版权局著作权合同登记号　图字 25-2011-002 号

图书在版编目(CIP)数据

预测性文本挖掘基础/(美)韦斯(Weiss,S.),张潼,(澳)因杜尔亚(Indurkhya,N.)著;赵仲孟,侯迪译.
—西安:西安交通大学出版社,2012.10
ISBN 978-7-5605-4427-4

Ⅰ.①预… Ⅱ.①韦… ②张… ③因… ④赵… ⑤侯… Ⅲ.①数据采集-研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2012)第 142928 号

书　　名	预测性文本挖掘基础
著　　者	(美)绍洛姆·韦斯,(澳)霓廷·因杜尔亚,(美)张潼
译　　者	赵仲孟　侯　迪
出版发行	西安交通大学出版社 (西安市兴庆南路 10 号　邮政编码 710049)
网　　址	http://www.xjtupress.com
电　　话	(029)82668357　82667874(发行中心) (029)82668315　82669096(总编办)
传　　真	(029)82668280
印　　刷	西安建科印务有限责任公司
开　　本	700mm×1000mm　1/16　　印张 15.75
印　　数	0001～3000　　字数 247 千字
版次印次	2012 年 10 月第 1 版　2012 年 10 月第 1 次印刷
书　　号	ISBN 978-7-5605-4427-4/TP·569
定　　价	43.00 元

读者购书、书店添货、如发现印装质量问题,请与本社发行中心联系、调换。
订购热线:(029)82665248　(029)82665249
投稿热线:(029)82665397
读者信箱:banquan1809@126.com

推荐序

预测性文本挖掘基础(Fundamentals of Predictive Text Mining)为一本介绍文本挖掘基础的工具书,引导这个迅速发展的领域。本书结合许多不同之主题,跨越各个学科及研究平台(数据挖掘、机器学习、数据库和计算语言学等),理论与实务兼备。

文本挖掘(Text Mining)又可称为文字知识发掘(Knowledge Discovery from Text ,KDT)或是文件信息探勘(Document Information Mining),其应用了信息检索、信息萃取、计算语言学、自然语言处理、数据挖掘……等。文本挖掘特别着重于利用这些技术,自非结构或半结构的文字中发掘出先前未知、隐含而有用的信息,Dan Sullivan (2001)定义文本挖掘为"一种编辑、组织及分析大量文件的过程,为了提供特定用户特定的信息,以及发现某些特征及其间的关联"。相较于传统的数据挖掘,文本挖掘需要加上额外的数据选择处理程序,以及复杂的特征萃取步骤。文本挖掘整合了许多传统信息检索技术,包括了关键词萃取、全文检索、档自动分类、自动摘要等等,以提供文字处理更强大的功能。故本书为文本挖掘入门最佳的书。

本书在 Dr.Sholom M.Weiss & Dr.Nitin Indurkhya 以及 Dr.Tong Zhang 努力合作下,理论与实务兼容并包,不论对初学者还是深入研究者都是很好的入门教材及继续研究文本挖掘的经典参考教材。

谢邦昌 敬上,于 2012 年 9 月
台湾辅仁大学商学研究所博士班 所长
台湾辅仁大学统计资讯学系暨应用统计所 教授
中华资料采矿协会 理事长

译者序

电脑的普及，使得越来越多的文档以数字化形式出现，互联网的发展使得这些数字化文档得以快速传播。如今海量电子化文档的规模仍在急剧增长，这些电子化文档（例如研究报告、学术论文、在线文献库、E-mail、Web 页面、公司内部公告、会议纪要等）包含了大量的信息，也是重要的知识源。但是很多情况下，由于文档的数量非常庞大、缺乏组织整理并且格式多种多样，导致人们无法充分利用这些数据。这就要求开发自动化的大规模文本数据分析技术，来帮助我们处理和利用这些文档。文本挖掘就是研究如何从这些文件中提取信息，也就是关注分析非结构化自然语言文字的过程。作为一个新的正在迅速成为热点的研究领域，它致力于从文本数据中发现新的事实和知识，帮助我们获得有价值的信息。

预测性文本挖掘研究涉及不同学科的论题，如数据挖掘、机器学习、模式识别、信息处理、数据库和计算语言学等，目的是寻找隐藏在文本数据之中具有预测功能的模式。目前对于预测问题的研究比较广泛，并且已经获得很好的算法，但这些算法专门处理结构化的数值数据，而文本文件往往是非结构化数据，这就要求研究新的方法处理文本信息。

呈现在读者面前的这本《预测性文本挖掘基础》，是近年来国际上少见的专门介绍预测性文本挖掘技术的入门性教材和指南，原文作者总结了多年来在这一领域研究的大量优秀成果。作为入门级读物，论述由浅入深、理论结合实践、语言风趣、样例详实，详细地分析和总结了该领域的研究现状及未来发展。深入地讨论了文档分类、信息检索、聚类与组织文档、信息提取、基于 Web 的数据源、预测与评价等方面提出的问题。时值互联网和云计算蓬勃发展时代，为关注海量非结构化信息处理的学习者了解文本挖掘研究前沿打开了一扇窗口。

本书前言、摘要、第 2、4、6、8 章以及封底、附录及索引等由赵仲孟副教授

负责翻译，第1、3、5、7、9章等由侯迪副教授负责翻译。赵仲孟副教授负责对全书内容进行整理、修改和校审。参与本书翻译和修订工作的有王文科、沈辉、曹志、张新峰、辛蓉、岳鹏、张丹、杨丁福、赵静、赖文清等硕士研究生，特别是王文科同学对本书修订做出了巨大贡献。李颖编辑也给予了宝贵的修改意见，她大量专业而细致的工作保证了本书的质量。在此向他们表示衷心感谢！

本书适用于大学高年级本科生或者研究生，既可作为计算机科学、软件工程、信息技术等专业相关课程的教材，也可作为文本挖掘爱好者及工程研究人员的参考书。

限于译者的经验和水平，书中难免有诸多纰误与不足，欢迎读者通过电子邮件提出您的宝贵意见！

赵仲孟、侯迪
zmzhao@mail.xjtu.edu.cn
2012年夏于西安

前　言

五年前，拙作《文本挖掘：非结构化数据预测方法》付梓。那时，我们认为读者是有文本挖掘经验的业内人士，后来学生也在老师的指导下拿它当教材来用。《文本挖掘：非结构化数据预测方法》是最早尝试去收集整理预测性文本挖掘的相关资料之一，它包含了丰富的内容。自出版至今，在众多老师的帮助之下，我们收到了大量学生读者的积极反馈。近年来，互联网为我们提供了越来越多的数据资源，随之而来文本挖掘领域也出现了大量有趣新颖的技术，考虑到这些新的内容，再版实有必要。一年前，出版商希望《文本挖掘》再版，以使它更好地成为一本教科书。所以我们改进了许多章节，添加了新的能够反应当前技术发展的内容，同时也增加了习题和总结部分。

所谓预测问题，即寻找隐藏在数据之中的具有预测功能的数据，目前对于预测问题的研究已经广泛展开，并且已经得到了很多的算法。这些算法专门处理结构化的数值数据，用统一的衡量方法分析被采样的数据。然而，文本往往是非结构化数据，显而易见，文本和数值数据是不同的，这就要求用新的方法处理文本信息。然而我们认为，预测问题有通用的解决方法，无论是对于结构化数据还是非结构化数据而言。文本可以转化成数值格式，例如对于一个词而言，我们可以用存在(1)或者不存在(0)来表示它。用于数据挖掘的预测方法已经被证明同样可以很好地应用于文本预测。当然，二者之间也有一些关键性的区别。评价技术也必须适应数据发布的时序，并且改变评估错误的方法。因为我们要处理的数据是文档，所以要选择一些更特殊的适应分析文本的方法。更重要的是，还需要考虑到我们要处理的数据很可能是高维的、成千上万的单词和文档，但是原理不变。

我们将文本挖掘看成是综合了不同领域的任务，而不再仅仅是依赖于语言学和计算机科学的“自然语言处理过程”，也不再是“区别于其他机器学习的搜索引擎。”我们的视野更加开阔，我们支持你个性化的分析、学习数据，无论

是文本还是数值。因特网上有大量的大文本集合，其中蕴藏了海量等待挖掘的信息，可惜我们今天的语言处理技术还不足以胜任。有一些学者在试图通过发现语义理解的本质来解决文本挖掘问题，我们现在仅仅试图在大容量的文档数据集中寻找可复用的模式。

我们讨论的大多数内容都经过文本挖掘方法的严格验证和一些著名分析方法的应用。本书提供了丰富的实例和软件环境，并且介绍了大量的有实际学术研究价值的内容。本书追求实践，同时也包含了广泛的对文本挖掘有帮助的内容，既涵盖了预测学习方法，又引出了信息检索、搜索引擎和聚类等技术。如果读者可以跟随这本书，使用我们提供的软件动手分析每一实例，相信一定可以大有收获。

随着分析方法的高度发展，文本预测挖掘的应用领域会日新月异。本书总结了笔者多年的经验，并且提供最新的工具和技术，让读者更好地学习和实践。

读者：

本书适用于 IT 开发人员，管理人员，同时也适用于计算机专业的研究生。了解数据挖掘方面的背景知识有助于更好地阅读本书，但这并不是必须的。一些讨论高级概念的章节所涉及的数学知识都附有解释和说明，绝大多数章节对于具有研究分析精神的读者而言都易读易懂。如果你致力于此领域的研究，那么本书将开拓你的视野；如果你志在文本挖掘开发，那么你能从我们推荐的算法和我们举出的实例中有所斩获。我们提供的软件环境要求用户熟悉命令行编程和编辑配置文件。

说明：

这本书所包含的内容已经成功地用于教学工作，从为期一周的短期课程到十二周的全学期课程，都可以使用本书。如果是短期课程，本书所涉及的数学分析内容可以跳过。书中的练习题已经经过了几年的测试。每一章的后面都有与之配套的章节总结、习题。

另外，本书所举的大量例子和图表都可以在 data-miner. com 上免费下载。

其他一些网络资源：

Data-Miner 有限公司为本书的读者提供了一份免费软件，该软件实现了本书介绍的很多算法，可以从 data-miner. com 上下载。大量实例的 Linux scripts 也可以从这里下载。详情请参考附录 A。

致谢：

感谢我的同事 Fred Damerau，他是我的良师益友，他也是本书的共同作者。他不仅和我们共同完成了本书的写作，还为我们的项目做出了杰出贡献，特别是他在语言学方面的造诣，给我们带来了难以言表的巨大帮助。本书中第 7 章中的一些例子基于我们之前出版的《文本挖掘》一书，我们也要感谢那本书的共同作者：Chidanand Apté，Radu Florian，Abraham Ittycheriah，Vijay Iyengar，Hongyan Jing，David Johnson，Frank Oles，Naval Verma，Brian White. Arindam Banerjee 为本书提供了大量建议。书中的习题设计参考了 Statistics. com 中我们定期开设的文本挖掘课程里的内容。最后感谢我们的编辑，Wayne Wheeler，Ann Kostant 和 Wayne Yuhasz，感谢他们的支持和帮助。

美国，纽约
Sholom Weiss and Tong Zhang
澳大利亚，悉尼
Nitin Indurkhya

目 录

第1章

文本挖掘概述

1.1 文本挖掘有什么特别之处?

你缺乏数据吗?很显然不太可能。由于计算机的普及,目前大部分数据 P 1
都以电子表格存储。我们使用计算机系统交易股票、使用文本编辑软件进行写作、在线购买商品,这些行为都离不开计算机。越来越多的纸上交易现在都在朝着无纸化方向发展,因此未来我们将会拥有以电子表格形式存在的"海量数据"等待我们分析。

数据挖掘建立在收集和存储海量数据基础上,它致力于寻找隐藏在数据之中的有用模式。现如今,数据挖掘已不是什么方兴未艾的新兴技术,虽然在业界远未普及,但是,数据挖掘领域理论技术已经高度发达,对于某些问题解决也几近成熟。

我们很想说"只需要提供给我数据,我就能告诉你数据隐藏的模式"。不幸的是,现有数据挖掘方法多数会要求被分析数据格式高度结构化,因此必须对于原始数据进行预处理。最好原始数据本身就是高度结构化的,否则我们必须对原始数据进行格式转换。

数据挖掘依赖于我们的经验。对于致力于做预测数据挖掘的专家而言,他们所处理的数据往往是数值型的,这些人每天和数字打交道。而"文本挖掘者"却从未期望得到一系列有序数字。他们的目标是寻找到一系列文本文件集合,并且这些集合内容是可读的,同时意义明显。

数据挖掘和文本挖掘之间的第一个区别:前者针对数字而后者专注于文