WENBEN XINXI CHULI

文本信息处理

张世博　著

中国水利水电出版社
www.waterpub.com.cn

文本信息处理

张世博　著

中国水利水电出版社
www.waterpub.com.cn
·北京·

内 容 提 要

目前，大数据产业蓬勃发展，从而带动了人们对于非格式化文本数据的分析需求，本书全面、系统地介绍了文本信息处理的相关技术，包括分词、文本向量化、特征选择、文本相似度计算、文本分类、主题模型、情感计算等内容，并在若干综合性的章节中，设计了独到的模型算法，阐述了算法过程。所有章节都通过实例对过程做详细描述，并辅助以代码或伪代码实现，帮助读者理解，具有高度的可操作性和实用性。

本书内容新颖、层次清晰，适合高校教师、研究生、高年级本科生使用，也可供相关的软件工程师做参考。

图书在版编目(CIP)数据

文本信息处理 / 张世博著. —北京：中国水利水电出版社，2018.9

ISBN 978-7-5170-6926-3

Ⅰ. ①文… Ⅱ. ①张… Ⅲ. ①文字处理—信息处理 Ⅳ. ①TP391.1

中国版本图书馆 CIP 数据核字(2018)第 221673 号

书　　名	文本信息处理 WENBEN XINXI CHULI
作　　者	张世博　著
出版发行	中国水利水电出版社 (北京市海淀区玉渊潭南路 1 号 D 座 100038) 网址：www.waterpub.com.cn E-mail：sales@waterpub.com.cn 电话：(010)68367658(营销中心)
经　　售	北京科水图书销售中心(零售) 电话：(010)88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排　　版	北京亚吉飞数码科技有限公司
印　　刷	三河市元兴印务有限公司
规　　格	170mm×240mm　16 开本　13.5 印张　242 千字
版　　次	2019 年 2 月第 1 版　2019 年 2 月第 1 次印刷
印　　数	0001—2000 册
定　　价	65.00 元

前　言

网络上非结构化的文本数据越来越多，体现在新闻、微博、微信自媒体中，形式多种多样，非结构化数据将在未来所造的数据中占有很大的比例，文本信息的处理可以揭示文字之间很难或无法确定的重要相互关系，其属于自然语言处理范畴，让计算机处理和运用自然语言。本书重点讲解自然语言处理方面尤其是中文文本信息的处理过程、细节技术等，既包含了传统文本信息处理技术，也包括了某些最新的业界方法，还容纳了笔者的科研成果。

本书主要内容如下：

第 1 章，介绍了文本信息处理的意义、应用场景、应用现状，并总结了其存在的应用挑战。

第 2 章，介绍了常用的数学基础，本章不是罗列相关的数学概念、公式，而是从技术角度阐述如何分阶段、分步骤实现文本处理，在各个环节会涉及的数学知识范围。

第 3～4 章，介绍了中文信息处理的最基础要素分词和向量化方法，从维特比算法、序列标注及深度学习等多个技术途径讲解了分词的方法，介绍了文本向量化的概念和方法，并引入了在实际操作中的散列技巧。

第 5～6 章，介绍了特征选择和文本相似度的计算方法，并通过以 word2vec 为案例的方式给出相似度计算过程。

第 7～8 章，是较为综合性的章节，介绍了文本分类的方法，分为朴素贝叶斯分类和 fastText 分类方法，其中，fastText 分类是最新的基于深度学习的开源文本分类工具，文中用详细实例介绍了该过程。

第 9 章，是关于文本摘要的内容，引入了笔者设计的基于句子评分的摘要模型。

第 10 章，介绍了主题模型，用于发现大数据文本下的主题倾向。以潜在狄利克雷分布为基础，通过分析句子结构，给出了针对具体文本数据的主题模型设计。

第 11 章，介绍了文本的情感倾向性计算方法。涵盖传统的分析方法和基于深度学习的分析方法。重点介绍了情感词库的自动扩充方法，以酒店

的评论文本为数据集，详细分析了数据集中的情感倾向性分析过程。

本书在内容上尽可能涵盖文本信息处理的各个环节，但受篇幅以及笔者水平的限制，很多重要的、前沿的方法未能覆盖，即便覆盖到的部分也仅是管中窥豹。

本书的撰写得到了北京市教委科技计划项目（KM201810017005）的资助。

自然语言处理发展极其迅速，针对文本信息的处理技术也层出不穷，很多科研机构和企业为此做了大量的基础工作和实际应用，笔者自认才疏学浅，书中错谬之处在所难免，请读者不吝告知，将不胜感激。

作　者

2017 年 5 月

目 录

第1章 引言

1.1 文本分析简介

1.1.1 文本分析的意义

在日常的产品和运营工作中，经常接触的数据分析方法、形式绝大部分是基于对数字的描述性分析，如销量情况、用户增长情况、留存情况和转化情况等，高级一些的数据分析方法有因子分析、聚类分析和回归分析等方法。

这些分析方法有一个共同点：都是跟数字在打交道，说得专业一点，就是基于对结构性数据（即行数据，存储在数据库里，可以用二维表结构来逻辑表达实现的数据）的分析，比如姓名、性别、年龄这些信息，以 Word、Excel 等形式呈现的数据。这种类别的数据比较好处理，只要简单地建立一个对应的表即可。从企业角度来说公司都有很多数据，传统意义上会认为只有阿拉伯数字叫作数据，比如企业的财务报表、经营状况、APP 每天日活等，除了这些之外，还有一些其他数据，比如文字型的数据：新闻内容、商品介绍、用户评论、企业内部各种各样的合同等，这些都是数据，其特点是以文本符号的形式存在。

目前，网络上非结构化的文本数据越来越多，体现在新闻、微博、微信自媒体等，形式多种多样，非结构化数据将在未来的数据中占有很大的比例[1]。作为一个尚未得到充分开发的信息源，非结构化数据分析可以揭示之前很难或无法确定的重要相互关系。所以，有必要对非结构性数据引起高度重视。非结构性数据是与结构性数据相对的一个概念，它包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像、音频和视频信息等，如图 1-1 所示。

无论是政府、企业还是个人都热切期望能从海量的文本中得到对自己有用的信息。要达到这个目的，凭人力费事、费力地逐条检索、逐条浏览是不现实的。

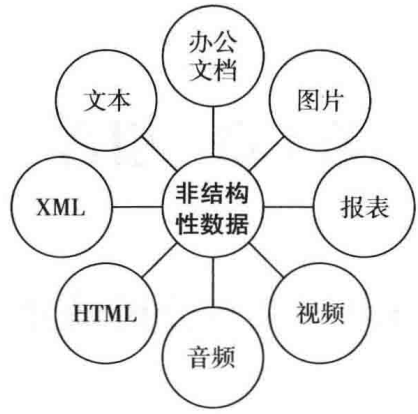

图 1-1　非结构性数据

企业希望通过数据挖掘技术提升效率，增加收入，降低成本，但是具体如何做？首先要把数据基础打好，比如尽可能地采集数据，较好地分析数据，把数据展示出来。现在很多挖掘还是人工用手工的规则和脚本实现，但是现在的信息化技术已经可以依靠计算机自动处理，并且做得更快、更好，减轻人的重复劳动，帮助企业提升效率。

网络上有非常多的数据，文本、图像、语音等类型的内容需要操作，识别归类和搜索。人工智能是把这两者联结在一起，让计算机自动完成从数据采集到识别搜索以及归类转化。“文本分析”或者“语义分析”是分析海量的非结构性文本信息数据，回答不仅是“是什么”的描述性分析，更多的回答“为什么”，即目标用户购买和使用产品的潜在动机和真实需求。

基于大数据的文本分析被广泛应用于各种行业来解决关键的知识性问题，例如从 CRM 数据、社交媒体、新闻网站和购物网站评论等渠道获取文本数据，再通过计算机自然语言处理，从而揭示出在任何非结构化文本信息中的人物、事件、时间、地点等内容，如图 1-2 所示，从而能够提供贯穿所有业务的全新层面的理解。

文字数据处理是信息的抽象提炼。这些数据其实是“一句话浓缩了很多内容”。文字数据的场景非常多，差别也很大。例如有的场景中用户的评论数据都是短短几十个字，也会有一些合同文本和法律文书，这些内容的字数则上千字和上万字。各种各样的长短文本，如果能够让计算机代替原来的人工进行自动化处理，便可以发挥很大的价值。在一些行业中，比如人事行业、法律行业、财务行业都有大量的资料，让计算机自动来分析这些文字资料，并自动来理解这些内容，这是非常有意义的事情。

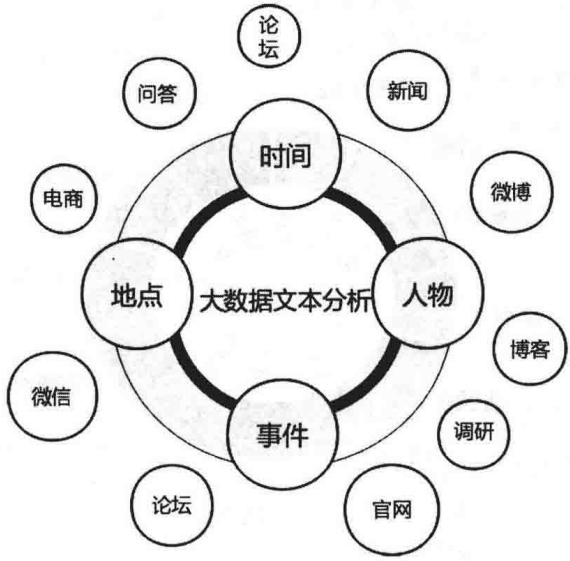

图 1-2 大数据文本分析提取的主要维度

1.1.2 文本分析的应用场景

从以下几点讲述海量文本分析的实际应用场景。

1. 开放式作答处理

大量问卷调研中有开放式问题，这些开放式的问题以电子文档的形式进行存储，使计算机进行文本分析成为可能，计算机可以在短时间内从数以万计的作答中提取出有价值的分析维度，如图 1-3 所示，实现对（潜在）用户需求的洞察。

2. 内容运营优化

（1）捕捉优秀作者的写作风格

对于一些初入新媒体运营岗位的人来说，研究和模仿某些知名自媒体作者的写作风格很有必要，学习他们的写作手法和套路可以使文案写作进步神速。

要想对这些优秀作者的行文风格进行深入研究，除了熟悉他们的行文脉络和篇章结构，更要熟稔其遣词造句上的风格，包括措辞特点、常用关键词和情感倾向等，在模仿中逐步形成自己的写作风格。

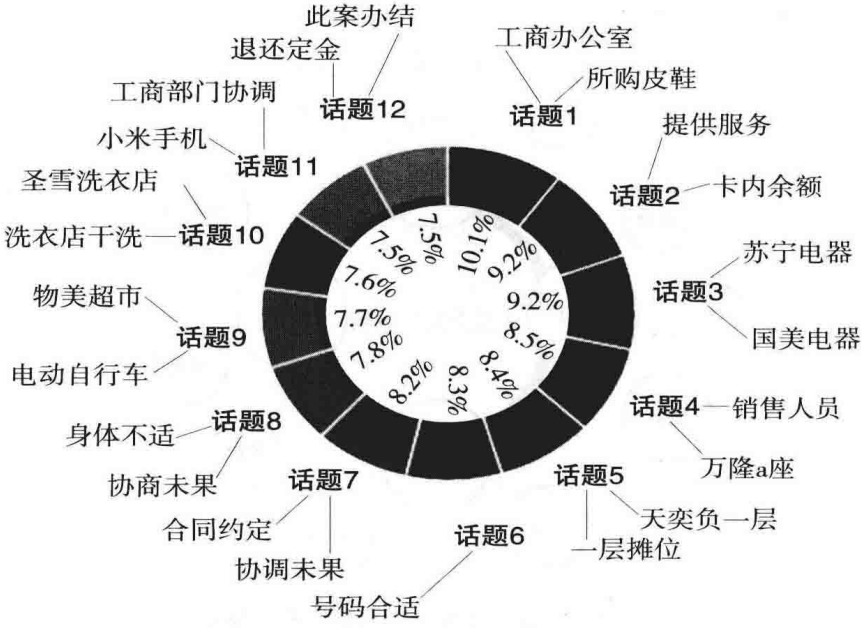

图 1-3　开放式问答题中提炼出的焦点话题

(2)新媒体热点采集、追踪及预测

基于大数据的文本分析能快速获取全网具有趋势传播的关键词，可以实时监测传播趋势，包括全面研究阅读数、评论数、分享量、传播趋势等，并且通过分析内容属性和成功原因，预测内容在未来的传播潜力。

在未来的媒体竞争中，媒体人需要转型变成“内容＋技术”的复合型人才，一方面发挥在内容创作中的人性的独立判断和分析，另一方面需要借助大数据分析技术提升文章的传播效果。

3. 口碑管理

基于大数据的文本分析能快速准确地识别出企业、品牌、产品自身及竞争对手在互联网上的口碑变化，深度挖掘文本数据价值，在消费者洞察、产品研发、运营管理、市场营销、品牌战略方面，为管理决策提供科学依据。

4. 舆情监测及分析

利用基于大数据的文本分析，可以清晰地知晓事件从始发到发酵期、发展期、高涨期、回落期和反馈期等阶段的演变过程，分析舆情的传播路径、传播节点、发展态势和受众反馈等情报。这一应用越来越受到各方面的重视。

5. 了解用户反馈

通过基于大数据的文本分析，企业可以用正确的方式阅读用户散落在

网络上的“声音”，企业可以直接读懂自己用户的想法，挖掘出用户对于产品或服务的情绪和态度。比如，大数据文本分析可以回答如下问题：

用户喜欢的是产品的哪一方面？

比起其他公司的产品，客户是否更倾向他的产品？

这些偏好会随着时间发展和变化吗？

1.2 技术发展历程

上述的各种应用都属于自然语言处理中的大范畴，关于自然语言处理，学术界有两个派别：

1. 理性派

认为所有语言其实都有潜在内生结构，都是有内在的语法。

2. 经验派

认为只要完成某一个功能就可以了，计算机完全不需要理解人说什么。

早期人工智能刚刚提出来，符号主义流行。在60年代的时候用了很多的词典和符号规则做自然语言的处理，但是后来发现这样翻译走不通。70～80年代，在语法规则的基础上，加上了语言模型，当时很多语言专家做自然语言处理时遇到非常严峻的挑战，因为语言不是特别严格的模型。例如：汉语特别灵活，很多时候甚至没有规则可言。

90年代开始，统计学习模型异军突起，现阶段看到大量自然语言处理的应用都是基于统计学习的模型。所应用的大数据方法也是因为已经积累的文本数据非常多，每天在各种平台上看到、写下的文字数据都可以成为计算机训练的语料，通过训练能让计算机发现语言的规律。

2010年以后，随着学习越来越深度化，知识图谱变得非常流行，它带有结构，目前很多主流方法是两者相结合，统计学习方法加上一些结构，能够更好地理解、处理文字内容。

1.2.1 文本结构解析的三个层次

现在流行的方法从结构的角度来说分三个层次：一是词汇级，二是句法级，三是篇章级。词汇级有很多具体的模块开发，结构分析包括句子结构之间的关系等。在汉语文本里面单个的字表现很弱，两个字或者三个字才构

成一个有表达力的词。比如“公司”是一个词，但是拆出来，“公”没有表达能力，“司”也没有表达能力。组词之后是造句，很多句法构成了一篇作文。同样，让计算机来阅读文字从结构角度来说是相似的，先让计算机了解字、词，然后理解句子的意思，最后理解整篇文章每个段落的含义。

1.2.2 确保文本分析效果的要素

1. 针对特定应用场景定制语言模型

虽然用的都是汉语或英语，但在不同的场景需要的方法有很大不同。例如：让计算机自动提取合同文本信息，自动判断合同文本中关联的要素和法律风险，这些文本都有一定的、潜在的语法结构。在作具体的专家文本判别时，需要建立这些具体的行业文本的知识库。

评论分析是目前很多企业应用的领域。很多企业每天会收到网上用户留下的成千上万条评论意见，甚至其中有一些是竞争对手的情报信息和评论信息。比如说手机行业分析用户评论意见时，通常评论有大量的省略和简称，小米手机第六代通常说米 6，计算机没有专业领域知识很难能像人一样解读这句话。

另外，口语和书面语的分别处理方式也不同，书面语是常写在内部文件中，但是通常弹幕、网络评论都是口语表达。比如说“杯具”、“稀饭”都不是吃的东西。

2. 持续的学习能力，确保泛化能力得到提升

机器学习的好处是可以通过反复迭代，实现持续学习、持续提升的效果。在文本挖掘中很多企业的挖掘都是依照规则的方法，但长期来看这种方法泛化能力或自主学习能力不够。通过机器学习以及用算法提升算法的能力来提升挖掘的效果是计算机处理模块时很重要的能力。

1.2.3 应用类型划分

计算机不像人一样可以阅读文字，计算机很多时候是输入一段字库，输出相应的结构。一边是编码，一边是解码。

文本挖掘基础应用的类型可以分为三大类：

1. 抽取

计算机想要自动解析文本，需要能够识别很多关键要素。例如，当计算

机阅读一份法律合同文书时，能够识别里面的判决书编号、被告人、辩护人、判决依据等，并能够从文本中提取出这些要素进行结构化处理。对于很多文本密集的行业，抽取这件事情很有价值。

2. 划分

比如，企业拿到大量客户的网上评论意见，需要知道这些意见哪些是好的、哪些是坏的，不同的意见需要后续给哪个部分负责处理，这些是典型评论意见观点的识别和观点划分的应用。

3. 合成

计算机写作将会是未来比较热门的行业。目前的写作还是以模板为主，比如基于一些合同模板把要素填写进来。但未来希望除了模板外，计算机还可以帮助人们修改、润色文章。甚至可以摆脱模板的方式，通过“阅读”大量的文字来实现机器写作。

1.3 应用现状

上面提到的抽取、划分和合成可以对文字进行很多处理，在满足企业的一些应用需求后，还可以进一步延伸。比如，每天都在用的搜索和推荐都是进一步的应用。

搜索是典型的自然语言处理的应用。它的核心技术有两部分，其一是对文本语义的深入理解，第二是解决搜索时间的性能问题。通常索引资料库很大，可能有上千亿的内容，在搜索的过程中不需要计算机一个一个找，而是用零点几秒解决响应的问题。这些需要用特殊的数据结构来完成。

另外，在搜索时如何让计算机帮助人来匹配更多优质资源，其实需要做更多语义的延伸。同一句话不同的人可以用不同的语言方式来表达。计算机帮助人做语义的扩展需要了解词和词、句子和句子之间的关系，才能更好地做语义之间理解的功能。

除搜索之外，个性化推荐也是语义理解中重要的应用。做内容和人的连接时，更好地完成用户画像需要分析出哪一个人之前看过这些内容，它的语义如何。文本挖掘技术在提升企业的运营质量方面发挥了很大作用，很多电商、新闻门户网站都开启了个性化推荐功能，它在帮助用户提升点击率、留存以及关键指标上都有着明显的效果。

1.3.1 文本分析工具

1. 商业化工具

近年来，国内外文本挖掘技术发展较快，许多技术已经进入商业化阶段。各大数据挖掘工具的提供商也都推出了自己的文本挖掘工具。这些工具除具备常规的文本挖掘功能（如数据预处理、分类、聚类和关联规则等）外，针对庞大的、非结构化数据都能作出较好的应对，支持多种文档格式，文本解析能力强大，大部分支持通用数据访问，但是价格都十分昂贵。由于各提供商的专注领域或企业背景不同，工具的定位和适用性也有所不同。以目前市面上比较流行的 10 款商业文本挖掘工具[2]为对象，针对其不同点进行简要的分析比较，如表 1-1 所示。

表 1-1 商业文本挖掘工具

工具名称	提供商	工具简介
Intelligent Miner for Text	IBM	挖掘结果展现能力较强，系统具有可扩展性，但是缺乏统计方法，限制了其本身的挖掘能力。在连接除 DB2 以外的数据库时，需要安装中间件。图形界面不友好且操作复杂，适合专业人员
Text Miner	SAS	算法齐全，360°数据视图展示。提出 SEMMA 方法论。用户界面灵活友好，但是操作复杂，分析结果难以理解，适合专业人员
Text Mining	IBM SPSS	提出 Crisp-DM 方法论。图形界面非常友好，易于操作，支持脚本功能，应用领域广泛且维护和升级成本较低。但是缺少最新的统计方法，且分析结果与其他软件的交互性较弱
IDOL	Server Autono-my	基于贝叶斯概率论和香农信息论。工具性能较高，支持 SOA，提供完全可配置的监控。但是系统的维护与管理缺乏相应的图形化应用界面，且工作过程中没有相关报告输出
Darwin	Oracle	通过 ODBC 访问数据，提供 wizard 引导用户构建模型。可扩展性较高，模型能够作为 C、C++和 Java 代码导出并集成于其他应用，用户界面友好。但是工具的适用面窄，市场份额较小；数据展示需要额外的工具，交互性差
SQL Server	Microsoft	基于 OLAP，利用数据源系统对数据进行清洗、转换和加载。挖掘功能集成于 SQL Server 系列产品中，易于使用。但是由于算法不足，解决问题有限，适合中小型业务

2. 开源工具

目前开源文本挖掘工具较多，如表 1-2 所示是主流开源文本挖掘工具。

表 1-2 主流开源文本挖掘工具

工具名称	开发者	开发语言
Weka	新西兰怀卡托大学	C/C++
GATE	谢菲尔德大学自然语言处理研究小组	Java
ROST CM	武汉大学 ROST 团队	C++
Open NLP	Apache	Java
LIBSVM	台湾大学林智仁团队	Java、Matlab、C #、Ruby、Python、R、Perl、Common LISP、Labview
Mallet	马萨诸塞大学 Andrew Mc Callum 团队	Java
Orange	斯洛文尼亚卢布尔雅那大学计算机与信息科学学院人工智能实验室	C++

Weka 以算法全面得到了许多数据挖掘工作人员的青睐，LIBSVM 是 SVM 模式识别与回归的工具包，ROST CM 在各大高校应用面非常广，对中文的支持最好。ROST 是由武汉大学沈阳博士 ROST 虚拟学习团队研发的一款内容挖掘软件，可以对数字化的材料进行组织、标引、检索和利用，具有海量性、智能性和客观性等特点，通过定量分析和定性分析的结合，ROST 文本挖掘软件能从数字化的材料中归纳出具有说服力的普遍性结论。ROST 文本挖掘软件可以对各类文本进行词频、聚类、分类、情感等分析。

大部分商业文本挖掘工具都对多语言、多格式的数据提供了良好的支持，且数据的前期处理功能都比较完善，支持结构化、半结构化和完全非结构化数据的分析处理。开源文本挖掘工具一般会有自己固有的格式要求，国外开源文本挖掘工具对中文的支持欠佳，而且大部分开源工具仍然停留在只支持结构化和半结构化数据的阶段。商业文本挖掘工具的分类、回归、聚类和关联规则算法普遍都较开源文本挖掘工具齐全，包含了目前主流的算法，只是每个工具在算法的具体实现上存在差异。同时，前