大数据
分析统计应用丛书

全国应用统计专业学位研究生教育指导委员会推荐用书

非结构化大数据分析

UNSTRUCTURED BIG DATA ANALYSIS

主编　李翠平

中国人民大学出版社

全国应用统计专业学位研究生教育指导委员会推荐用书

非结构化大数据分析

UNSTRUCTURED BIG DATA ANALYSIS

主编　李翠平

中国人民大学出版社
·北　京·

图书在版编目（CIP）数据

非结构化大数据分析/李翠平主编. —北京：中国人民大学出版社，2018.11
（大数据分析统计应用丛书）
ISBN 978-7-300-26297-0

Ⅰ.①非… Ⅱ.①李… Ⅲ.①数据处理-研究生-教材 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2018）第 222071 号

大数据分析统计应用丛书
非结构化大数据分析
主编　李翠平
Feijegouhua Dashuju Fenxi

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010－62511242（总编室）	010－62511770（质管部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京市鑫霸印务有限公司		
规　　格	185 mm×260 mm　16 开本	**版　　次**	2018 年 11 月第 1 版
印　　张	15 插页 1	**印　　次**	2018 年 11 月第 1 次印刷
字　　数	346 000	**定　　价**	36.00 元

大数据分析统计应用丛书编委会

总　序

一

统计学是收集、分析、展示和解释数据的方法性质的一门科学。信息技术的蓬勃发展，使统计在经济、社会、管理、医学、生物、农业、工程等领域有了越来越多、越来越深入的应用。2011 年 2 月，国务院学位委员会第 28 次会议通过了新的《学位授予和人才培养学科目录（2011）》，将统计学上升为一级学科，这为统计学科建设与发展提供了难得的机遇。

一般认为，麦肯锡公司的研究部门——麦肯锡全球研究院（MGI），在 2011 年首先提出了大数据时代（age of big data）的概念，并引起了全球广泛的反响。大数据是指随着现代社会的进步和信息通信技术的发展，在政治、经济、社会、文化等各个领域形成的规模巨大、增长与传递迅速、形式复杂多样、非结构化程度高的数据或者数据集。它的来源包括传感器、移动设备、在线交易、社交网络等，其形式可以是各种空间数据、报表统计数据、文字、声音、图像、超文本等各种环境和文化数据信息等。大数据时代是一个海量数据开始广泛出现、海量数据的运用逐渐普遍的新的历史时期，也是我们需要认真研究与应对的一个新的社会环境与社会形式。

大数据时代对统计专业的学生提出了更高的要求。他们不仅需要具有扎实的统计理论基础，并且要熟练掌握各种处理大数据和统计模型分析的计算机技能，还要懂得如何提出研究问题、如何判断数据质量、如何评价模型和方法，以及如何准确清晰地呈现分析结果。这对统计教育和人才培养提出了新的目标和方向。

二

顺应时势，在教育部全国应用统计专业学位研究生教育指导委员会推动下，由中国人民大学、北京大学、中国科学院大学、中央财经大学、首都经济贸易大学五所高校发起，集中统计学科、计算机学科、经济与管理学科的相关学院优势，依托应用统计专业硕士项目，组建了北京大数据分析硕士培养协同创新平台。2014 年 9 月首届实验班正式招生并开始授课。

实验班每年招收约50～60名学生，分别来自中国人民大学、北京大学、中国科学院大学、中央财经大学、首都经济贸易大学等院校。他们均是以优异成绩进入上述高校应用统计硕士项目的本科毕业生，对大数据分析有浓厚的兴趣，立志为大数据分析领域的发展做出贡献。

大数据分析硕士的培养是为了满足政府部门和企业等用人单位利用大数据决策的需求，其核心竞争力是快速部署从大数据到知识发现和价值的能力，培养方案与国际接轨，核心内容是面向大数据的统计分析和挖掘技术。经过前期的充分论证，大数据分析硕士培养方案确定了核心必修课与分方向的选修课。必修课的重点内容为统计学和计算机科学的交叉部分，侧重于培养从大数据到价值的实践能力，包括大数据分析必备的计算机基础技能、面向大数据分析的计算机编程能力、大数据统计建模和挖掘能力。每门必修课均配备了5人以上的教学团队，由包括国家“千人计划”入选者、长江学者、国家杰出青年基金获得者在内的在相关领域有较高造诣的中青年学者组成。

大数据分析硕士培养协同创新平台是一个面向政府部门和企业等大数据分析人才需求单位开放的平台，目标是建成一个政产学研有机融和的协同创新平台。2014年5月19日平台成立大会就汇集了《人民日报》、新华社、中央电视台、中国移动、中国联通、中国电信、全国手机媒体专业委员会、SAS（北京）有限公司、华闻传媒产业创新研究院、北京华通人商用信息有限公司、龙信数据（北京）有限公司等，成为该平台的第一批实践培养和研发基地。在2014年9月开学典礼上又有中国科学院计算机网络信息中心、中国中医科学院、商务部国际贸易学会、国家食品安全风险评估中心、北京商智通信息技术有限公司、史丹索特（北京）信息技术有限公司、北京太阳金税软件技术有限公司、北京京东叁佰陆拾度电子商务有限公司、北京知行慧科教育科技有限公司、中关村大数据产业联盟、艾瑞咨询集团11家单位加入平台建设的联盟协作单位。实际部门的踊跃参与说明大数据分析人才培养的巨大发展空间。为了加强大学与实际部门专家的双导师制度，开学典礼上为第一届实验班专门聘任了26名实际部门专家担任硕士研究生指导教师。

2015年1月15日，大数据分析硕士培养协同创新平台联合京东、奇虎360、艾瑞咨询集团、华通人等多家公司举办了针对学生实习的宣讲会。会后组织学生到各相关部门进行有关数据挖掘、大数据分析的实习工作，学生们得到了锻炼。

为活跃学术氛围，拓展学生视野，大数据分析硕士培养协同创新平台组织了大数据分析学术系列讲座，邀请学界、业界相关人士交流分享学术、行业前沿的经验，共同推进大数据人才培养以及学术成果的转化。

三

迄今为止，五校联合大数据分析硕士实验班已经成功开展两届。在此基础上，课程组全体教师及时收集学生反馈意见，积极组织讨论，联合中国人民大学出版社，启动了“大数据分析统计应用丛书”的编写工作。

本套丛书第一期出版四本。《大数据分析计算机基础》着重介绍数据分析必备的计算机技能，包括Linux操作系统与shell编程，数据库操作与管理；面向大数据分析的计算机编程能力，我们重点推荐了Python语言。《大数据探索性分析》的内容包括大数据抽样、预处理、探索性分析、可视化以及时空大数据案例。《大数据分布式计算与案例》介

绍了单机并行计算以及 Hadoop 分布式计算集群，在此基础上介绍了 HDFS 文件管理系统以及 MapReduce 框架、各种统计模型的 MapReduce 实现，此外还介绍了处理大数据最常使用的 Hive，HBase，Mahout 以及 Spark 等工具。《大数据挖掘与统计机器学习》介绍了常用的统计学习的回归和分类模型、模型评价与选择的方法、聚类和推荐系统等算法，所有方法均配有 R 语言实现案例，支持向量机和深度学习方法给出了 Python 实现案例，最后是三个数据量在 10G 以上的大数据案例分析，所有的数据和程序均可下载。相信读者在学习本套丛书的过程中，数据处理与分析能力会得到锻炼和提高。

在丛书第一期的基础上，我们也在积极策划第二期，内容包括非结构化大数据分析、大数据统计模型、统计计算与统计优化方法等，希望可以涵盖更多的数据类型与统计方法。

该丛书面向的读者主要是应用统计专业硕士，也可以作为统计专业高年级本科生、其他专业的本科生、研究生以及对大数据分析有兴趣的从业人员的参考书，希望这套丛书可以为我国大数据分析人才的培养奉献我们的绵薄之力。

丛书编委会

前　言

非结构化数据是与结构化数据相对应的概念。要理解什么样的数据是非结构化的，先来看一下什么是结构化数据。结构化数据通常指具有固定格式的数据，例如，存放在关系数据库中的二维表格就是一种典型的结构化数据。这类数据由若干行和列组成。每一行表示一个对象，每一列表示对象的一个属性。例如，日常生活中经常使用的通讯录、工资单等就是这种类型的表格。

可以看出，结构化数据具有固定的格式，看上去非常规整。这类数据可能是为了数据挖掘而特定收集的，在收集之初就设计好了格式；也有可能是经过了某个数据转换过程而得到的。对这类数据而言，行的增加比较容易，每增加一行相当于增加一个新的样本。而列的增加要困难得多，这要求对所有已存在的样本进行检查，并且为每一个样本的新属性添加测量值。

与结构化数据相反，非结构化数据指无固定格式的数据，例如，文本、网页、图像、视频、数据流、序列、社交网络、图结构等。这类数据也许存在着某种程度的内部结构，但是由于不具有固定的格式，所以通常称为非结构化数据。现有数据中绝大多数数据都是非结构化数据，而且随着时间的推移，非结构化数据的增长速度要远远超过结构化数据的增长速度。

传统结构化数据的统计分析已经较为普遍，但对于非结构化数据的分析，虽然已经引起业界和学界的广泛关注，但是仍有很大的探索空间。究其原因，主要源于非结构化数据的特点。与结构化数据相比，非结构化数据突破了结构定义不易改变和数据定长的限制，其数据存储和分析都更加复杂多样。

本书介绍了四种典型非结构化数据的分析和挖掘技术，分别是文本数据、社交网络数据、数据流数据和多媒体数据（包括图像、音频和视频），共 12 章。

第 1～5 章，主要介绍了文本挖掘的时代背景、文本挖掘与数据挖掘的关系、文本预处理、文本分类、文本聚类、话题检测、观点挖掘和情感分析等。第 6～10 章，主要介绍了社会网络的相关基本概念、常见统计属性、社区发现、个体社会影响力分析、链路预测、网络信息扩散等。第 11 章，主要介绍了数据流中的变化探测、直方图、聚类和分类等。第 12 章，主要介绍了图像、音频和视频数据的特征提取、内容检索、内容识别等。

为了便于读者学习，大部分内容除了理论讲解之外，还给出了相应的在大数据环境下的上机实践案例。例如，在文本挖掘部分，给出了在 Spark 环境下用朴素贝叶斯进行垃圾短信识别、用 LDA 模型进行话题检测的 Scala 语言实现代码，以及用 LIBSVM 实现的发债企业负面新闻识别系统的实现方案；在社会网络分析部分，用 Spark GraphX 实现了微博用户关系分析和个体社会影响力计算，给出了相应的 Scala 语言实现代码、边聚类社区发现算法的 C 语言实现代码，以及基于邻居相似度指标的链路预测算法和两种信息扩散计算过程的 R 语言实现代码。

本书第 1～5 章由李翠平、刘苗、卫斌合作撰写，第 6～10 章由马丽丽和孙怡帆合作撰写，第 11 章由贾金柱撰写，第 12 章由李锡荣和许洁萍撰写。全书由李翠平统稿校对。中国人民大学数据仓库与商务智能实验室的很多同学参与了本书的写作和研讨、实践案例设计、书稿校对等，他们是：王绍卿、赵衍衍、付岩松、葛昊、邵国栋、刘颖智、张超杰、李青华。作者在此对他们的支持和帮助表示诚挚的谢意。

最后，感谢北京五校联合（中国人民大学、北京大学、中国科学院大学、中央财经大学、首都经济贸易大学）大数据分析硕士培养协同创新平台的所有老师；感谢中国人民大学出版社的大力支持。

我们在编写本书的过程中，尽可能做到深入浅出，力求概念正确，理论联系实际。非结构大数据分析是一个应用很广的领域，发展非常迅速，但我们水平有限，书中一定存在许多不足之处，希望同行和广大读者不吝赐教，多提宝贵意见。

李翠平

目　录

第1章　文本挖掘概述

1.1　时代背景

随着计算机与网络技术的发展，一方面，科学、工程、商业计算等领域需要处理大规模的数据。例如，在高能物理、天文学、生物学和地球科学等领域，每年的数据规模都能达到若干 PB；另一方面，用户在生活工作中使用社交网络、新闻网站、办公软件等所创建的内容数据也呈爆炸式增长，例如，在谷歌、脸书、百度、新浪微博等应用中产生的数据量甚至能达到 EB 级。

用户创建的这些内容数据，有很大一部分是用自然语言表达的文本数据。

这些文本数据包括微博、电子邮件、政府报告、学术文献、网页、会议纪要等。它们具有丰富的语义，蕴含了人们对自然界的认识，代表人们对不同事物的观点和偏好，具有极大的挖掘价值。例如，网络购物之后，消费者通常会发表对某个产品的评论，其他消费者会根据这些评论来决定是否选择购买该产品，厂家则可以根据这些评论来对产品加以改进。

文本数据是一种典型的非结构化数据。由于缺乏类似结构化数据那样固定清晰的模式结构，计算机很难理解文本数据的语义，也很难对其进行自动化处理。几十年来，人们一直在探索如何能让计算机精确地理解自然语言，同时自动地对文本数据进行分析。但到目前为止，这个愿望还没有完全实现。尽管如此，人们还是研究并提出了很多基于规则和统计的自动化文本分析和挖掘技术，本章主要对这些技术进行介绍。

1.2　文本挖掘与数据挖掘

文本挖掘是数据挖掘的一种特殊形式。首先来看什么是数据挖掘。数据挖掘是从大

量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。简单地说，数据挖掘是从大量数据中提取或挖掘知识的过程。数据挖掘要处理的原始数据可以是任何类型的，比如可以是表格、文本、网页、图像、视频、数据流、序列、图结构，等等。而文本挖掘主要指基于文本这种特定类型的原始数据进行的数据挖掘。因此，文本挖掘可以看作数据挖掘的一个分支。

文本挖掘有时也称为文本数据挖掘、文本知识发现、文字探勘、文本分析等。虽然文本挖掘的定义和数据挖掘有相似之处，但由于文本数据具有非结构化的特点，缺乏机器可理解的语义，传统的数据挖掘技术并不能直接应用于文本数据，即使可以使用，也需要建立在对文本预处理的基础之上。也就是说，首先需要将非结构化的文本数据转换为结构化的数据。

1.2.1 从非结构化数据到结构化数据

文档是典型的非结构化数据，文本挖掘通常在文档集合上进行。由于传统的数据挖掘算法大多以结构化数据作为输入，不能直接在原始的文档形式上进行，所以在进行文本挖掘时需要首先将文本转换为结构化数据，然后才能使用传统的数据挖掘算法进行模式的挖掘。通常将这个从非结构化数据到结构化数据的转换过程称为文本预处理。

文本预处理的主要任务是将文本转换为二维表格。由于二维表格中的每行代表一个对象（或者叫样本），因此，进行转换的时候可以将每一篇文档看作一个样本，转换成二维表格中对应的一行（也可看作一个向量）。如何将一篇文档转换成一个向量呢？需要经历如下三步：

第一步，分词，即将一个文档切分成一个一个的词。例如，将文档“我喜欢文本挖掘课程”划分成“我 喜欢 文本 挖掘 课程”5 个词。英文等文档由于词和词之间有空格作为分界符，分词相对容易。中文文档由于词和词之间并没有分界符存在，分词相对困难得多。分词属于自然语言处理技术范畴，目前已经提出了多种分词算法，如基于字符串匹配的分词算法、基于统计的分词算法、基于知识理解（规则）的分词算法、基于字标注的分词算法等。基于这些算法，人们也开发出了一些现成的分词工具，如中科院计算技术研究所开发的汉语词法分析系统 ICTCLAS 等。

第二步，去除停用词、取词根。经分词程序处理后的文本句子变成了词＋空格＋词的表现形式。其中许多语气助词、副词、连词等虚词虽然出现频率很高，但并无实际使用意义，如“的”“地”“得”“着”“了”“过”等，需要将其删除。这类词在文本预处理中称为停用词（stop word）。另外，英文里面有些经过变形的词，要将其还原成词根。例如，类似于“studying”“studied”这样的词，要将其还原成“study”。这一步通常称为取词根。

第三步，文本表示，即将每一篇文档表示成一个向量。前面提到，二维表格是一种典型的结构化数据表现形式。文本预处理的目标就是要将非结构化的文本数据表示为类似二维表格等结构化数据的表示形式。其中每个文档相当于二维表格中的一行，文档集合中经分词及去停用词取词根处理后的每个词取词根作为二维表格中的一个属性列。用作列属

性的这些词也叫特征词。相关表示形式如图 1-1 所示。

	特征词 1	特征词 2	特征词 3	…	特征词 n
文档 1	权重 11	权重 12	权重 13	…	权重 1n
文档 2	权重 21	权重 22	权重 23	…	权重 2n
文档 3	权重 31	权重 32	权重 33	…	权重 3n
文档 4	权重 41	权重 42	权重 43	…	权重 4n
⋮	⋮	⋮	⋮	⋮	⋮
文档 m	权重 $m1$	权重 $m2$	权重 $m3$	…	权重 mn

图 1-1　文档集合的结构化表示形式

这种表示方式称为文本的向量空间模型，即用向量空间模型来表示文本。在向量空间模型中，图 1-1 中的每一个特征词称为向量空间模型中的一个维度，即文本集可以看作由一组特征词（特征词 1，特征词 2，特征词 3，特征词 4，…，特征词 n）组成的向量空间，每个文本文件可以看成这 n 维空间中的一个向量。权重可以根据不同的方法计算得出。最简单的向量空间模型是布尔模型，它将某个词在文档的出现与否作为权重的度量指标，词出现时权重为 1，没出现时权重为 0。例如，如下三个经过分词之后的文档所生成的布尔模型如图 1-2 所示。

文档 1：“我　喜欢　文本　挖掘”；

文档 2：“我　喜欢　信息　检索”；

文档 3：“我　是　一个　学生”。

	我	喜欢	是	文本	挖掘	信息	检索	一个	学生
文档 1	1	1	0	1	1	0	0	0	0
文档 2	1	1	0	0	0	1	1	0	0
文档 3	1	0	1	0	0	0	0	1	1

图 1-2　布尔向量空间模型示例

至此，我们又看到了熟悉的二维表格数据，也意味着，非结构化到结构化数据的转换完成。将转换之后的二维表格数据输入到经典的数据挖掘算法中，就能够得到想要的各种挖掘结果。当然，这里只是做了一个简单的示例，实际中需要考虑更多的问题。更详细的内容请参见本书第 2 章中的“分词技术”及“文本表示”两节。

1.2.2　文本挖掘的知识类型

如上所述，文本挖掘是从大量的、不完全的、有噪声的文本数据中提取隐含在其中的、潜在有用的信息和知识的过程。这些信息和知识是人们在用自然语言进行表达时蕴含在文本中的，不同的人将自己对自然界的认识、对事物的看法表达出来，存放在文本数据中。现在，需要从这些大量的多人产生的文本中将这些有用的知识挖掘出来。一般来说，需要对如下知识进行挖掘：

(1) 关于自然语言本身的知识。寻找关于自然语言的使用方法、词语搭配、同义词情况、俗语情况等的知识。

（2）关于文本内容的知识。寻找蕴含在自然语言中的关于某个特定的人或物的看法或认识，主要有文本分类和聚类、话题检测、文本摘要、文本关联分析。

（3）关于文本作者情感的知识，主要指通过文本内容来推断文本作者的观点或者情感倾向。

1.2.3 文本挖掘的过程

图 1-3 给出了文本挖掘过程的形象化表示。其主要步骤如下：

（1）文本获取：主要指从不同的数据源通过不同的方式获得文本挖掘所需要的文本数据。例如，可以首先通过搜索引擎技术，获得相关的结果文档，然后用这些结果文档作为文本挖掘的输入；或者，通过网络爬虫技术从互联网上爬取所需要的文本数据。有关搜索引擎和爬虫的技术本书未涉及，感兴趣的读者请参考相关书籍。

（2）文本预处理：首先对文本进行过滤，对文本的类型进行分类。而后对文本进行预处理，即进行分词、去除停用词、取词根、文本表示。最终将每一篇文档表示成一个向量，完成从非结构化数据到结构化数据的转换。

（3）文本挖掘：首先确定挖掘任务，然后选择或者设计合适的算法和工具进行挖掘操作。根据用户提供的指标，对挖掘出来的模式进行评估，并使用可视化的知识表示技术，向用户提供容易理解的数据模式（知识）。基本的文本挖掘任务通常包括：文本分类、文本聚类、观点挖掘、话题检测等，详细的内容请参照本篇后续第 3、4、5 章的内容。一些更高级的文本挖掘任务如问答系统、机器翻译等本书并没有涉及，感兴趣的读者请参考相关书籍。

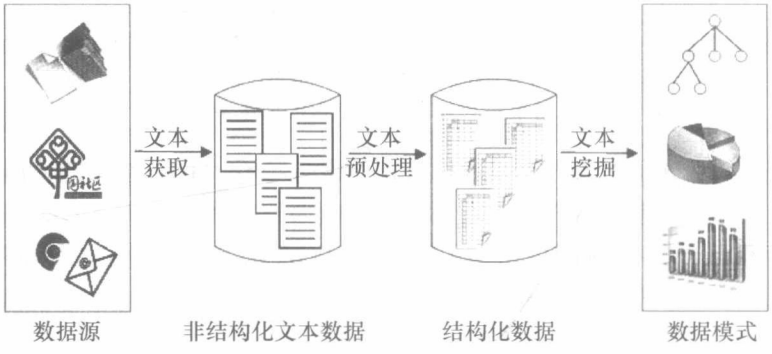

图 1-3 文本挖掘的过程

1.2.4 文本挖掘的特点

文本挖掘有很多独特之处，主要表现在：（1）文本挖掘处理的是大规模的文本集合，而不是一个或少量的文本文档。（2）文本挖掘发现的知识是隐藏在大量文本文档中的，是新的、以前未知的模式或关系。（3）文本挖掘抽取的知识是以真实世界为基础的、具有潜在价值的，是直接可用的，它或者是某个特定用户感兴趣的，或者对于解答某个特定问题是有用的。（4）文本挖掘处理的是大规模的文本库，其挖掘算法复杂度较高。（5）文本数

据有大量的噪声和不规则的结构，因此文本挖掘算法应具有很强的鲁棒性。

1.2.5 文本挖掘的应用

文本挖掘是应用驱动的，在商务智能、科学研究、舆情监测、自动问答、情感分析、信息安全、生物信息处理、垃圾邮件过滤、自动简历评审等方面都有广泛的应用。下面就几个重要领域展开说明。

一、商务智能

文本挖掘可以帮助企业快速搜集和整理自己企业或同类型竞争企业的商业信息，从而更好地辅助企业决策。例如，企业可以对网络购物平台的商品评价进行文本挖掘。网络购物平台的商品评价文本中包含了用户对于某个或某类商品的情感倾向和观点（支持、反对、喜欢、讨厌等）。对商家或服务提供商而言，这些观点和态度可以帮助企业改进商品缺陷，有针对性地提高服务质量，更高效地完成产品的升级换代和企业的规划安排。

二、科学研究

在科学研究中，及时获取文献中的知识是很必要的。研究人员通常需要及时了解和把握所在研究领域的最新研究热点和未来研究趋势。文本聚类技术可用于发现现有研究文献中的热点话题。通过对不同时期的文献的追踪研究，绘制出不同领域的技术发展路线图，或者通过对不同地区的文献的研究，概括出研究热点的地域分布及应用领域等，从而帮助研究者发现技术创新机会。

三、舆情监测

舆情是指公众对于现实社会中各种现象、问题，所表达的信念、态度、意见和情绪表现的总和，是实现社会调控管理不可忽视的制约力量。网络的普及以及 Web2.0 的出现，为社会大众表达对国家政策和各种社会问题的情绪与态度提供了开放的渠道，从而形成了各种各样的网络舆情信息。网络舆情成为观察民意焦点指向的一个风向标。通过对来自不同渠道的文本数据进行分析和归纳，可以全面系统地了解某一时期或某一地区的社会舆情情况，从而有利于决策者做出正确的决策。

四、自动问答

自动问答系统最突出的特点是允许用户用自然语言句子进行提问，系统会自动分析用户的提问，然后通过反问即人机交互的方式，准确地辨识用户的意图，并为用户直接返回所需要的答案。和搜索引擎相比，用户不需要将自己的问题分解成关键字，可以把整个问题直接交给问答系统。问答系统通过对问题的理解，结合自然语言处理技术，能够直接提交给用户想要的答案。因此，问答系统比传统的搜索引擎方便、快捷和高效。

文本挖掘的应用领域远不止以上四个方面，本书只简单罗列以上几类应用，相信读者可以发现更多的文本挖掘的应用案例。

习题

1. 请简述文本挖掘与数据挖掘的联系与区别。
2. 如何将非结构化数据转换为结构化数据？请简述其过程。
3. 文本挖掘的知识类型主要有哪些？