谢邦昌 朱建平 李 毅 著

厦门大学数据挖掘研究中心
厦门大学管理学院 MBA 中心

大数据丛书 3

文本挖掘技术及其应用

Text Mining and Its Application

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

谢邦昌 朱建平 李 毅 著

文本挖掘技术及其应用

Text Mining and Its Application

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

图书在版编目(CIP)数据

文本挖掘技术及其应用/谢邦昌，朱建平，李毅著. —厦门：厦门大学出版社，2016.3
ISBN 978-7-5615-5971-0

Ⅰ.①文… Ⅱ.①谢…②朱…③李… Ⅲ.①数据采集-研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2016)第 057608 号

出 版 人 蒋东明
责任编辑 吴兴友
装帧设计 王 琳
责任印制 吴晓平

出版发行 厦门大学出版社
社 址 厦门市软件园二期望海路 39 号
邮政编码 361008
总 编 办 0592-2182177 0592-2181253(传真)
营销中心 0592-2184458 0592-2181365
网 址 http://www.xmupress.com
邮 箱 xmupress@126.com
印 刷 厦门市明亮彩印有限公司印刷

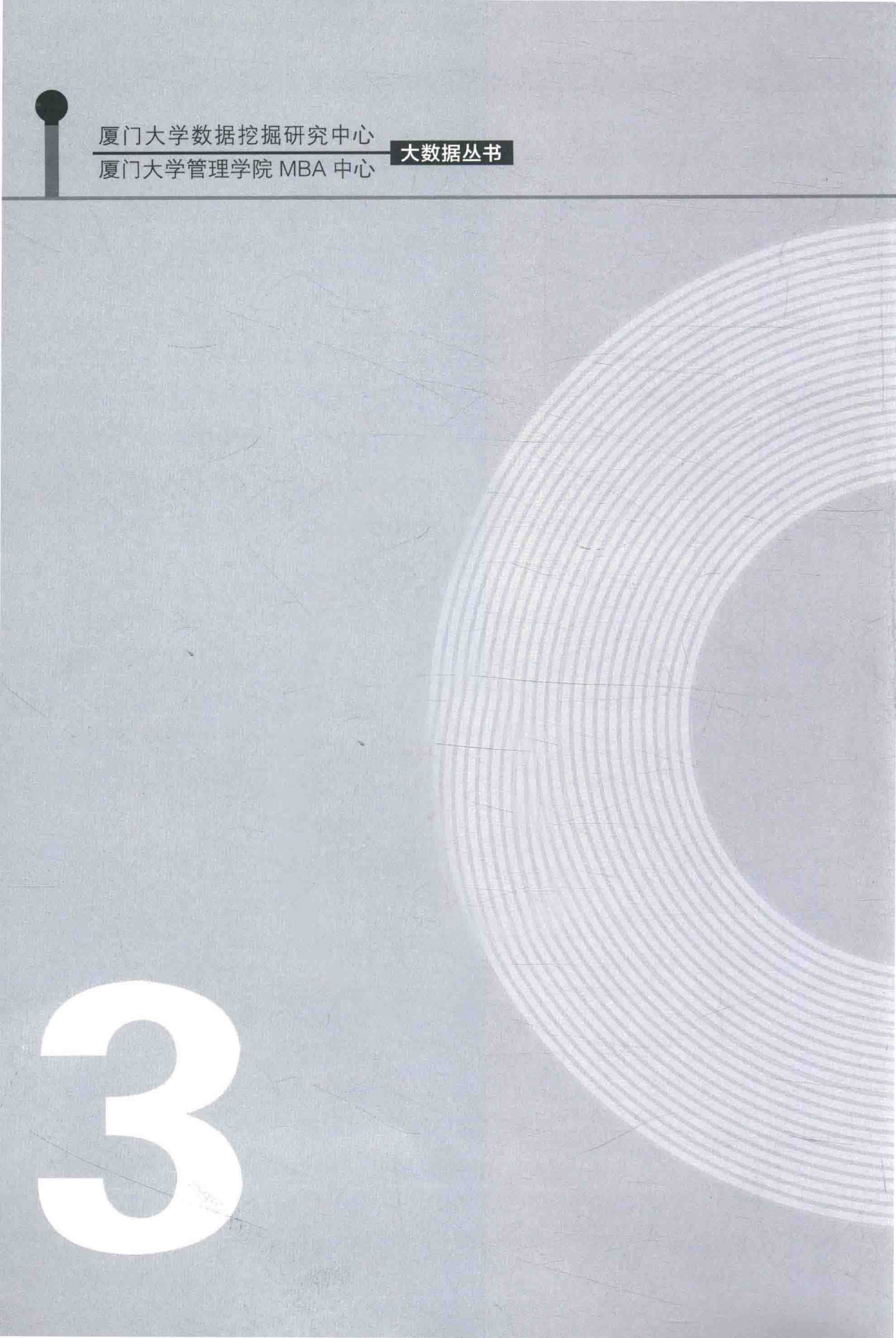
开本 787mm×1092mm 1/16
印张 10
插页 2
字数 280 千字
印数 1～3 000 册
版次 2016 年 3 月第 1 版
印次 2016 年 3 月第 1 次印刷
定价 32.00 元

厦门大学出版社
微信二维码

厦门大学出版社
微博二维码

厦门大学数据挖掘研究中心
厦门大学管理学院 MBA 中心

大数据丛书



3

前 言

“如何来认识这个世界?”通过感知来认识世界上的客观物质存在。互联网时代的到来,各个领域对大数据的认知逐渐清晰,使得我们做这本书的方向越来越明确,就是想尽可能把客观存在的信息纳入社会经济统计分析中,使得原先社会经济中很多定性分析转变为定量分析成为可能,或者成为最有力的补充。

文字是记录信息的图像符,它是记录人类社会不可或缺的工具之一。常常说“字如其人,人如其字”,我们做了大胆的假设:“如果收集到某人的全部文字资料,基于大数定理的思想,就能刻画出这个人必然而非偶然的特征”,将这假设横向延伸开来,对于某社会问题的全数据集(数据,文字,音像等)都收集起来进行处理,我们相信必然可以很精准找到这个问题客观存在的规律。原先很多人抱怨在社会经济科学中量化方法不准,其中很大原因之一,就是收集“数据”太有限了。

本书主要框架是基于跨行业数据挖掘标准流程(CRISP-DM)这一知识发现(KDD)过程模型展开,其内容主要包括互联网数据(来自新浪微博、Facebook 和 Twitter)的收集、准备、建模、评估和实施,并利用 R 软件和 Microsoft SQL Server 软件在实务案例中进行文本挖掘。本书通过实战案例整合大量文本挖掘中各种碎片化经验,提供相关的程序代码和技术实践支持,让读者阅读之后就能亲自动手做,而且做出来就可以解决许多实际问题。

本书是山西财经大学李毅副教授在台湾辅仁大学访学期间,联合台湾辅仁大学硕士研究生欧宇丰、高翊玮、高彦钧、陈誉晏完成的部分研究成果,谢邦彦博士提供大量实际案例。台北医学大学谢邦昌教授和厦门大学朱建平教授对本书进行了修改和总纂,台湾医学大学邓光甫教授审阅了全书。

本书在编写和出版过程中,得到了厦门大学

数据挖掘研究中心、台湾医学大学大数据研究中心、厦门大学管理学院 MBA 中心、山西财经大学统计学院和厦门大学出版社的支持，陈丽贞和吴兴友同志为本书的组稿、编辑做了大量的工作，在此表示衷心感谢！编写本书的一个初衷就是“抛砖引玉”，不同学科也有类似研究领域，如自然语言处理、文献计量学等有着大量丰富成果。我们求知若渴，限于作者的学识水平有限，书中不妥之处在所难免，恳请广大读者批评指正。

本书的编写和出版得到了国家社会科学基金重大项目“大数据与统计学理论的发展研究”(13&ZD148)的资助。

作者

2015 年 12 月

CONTENT

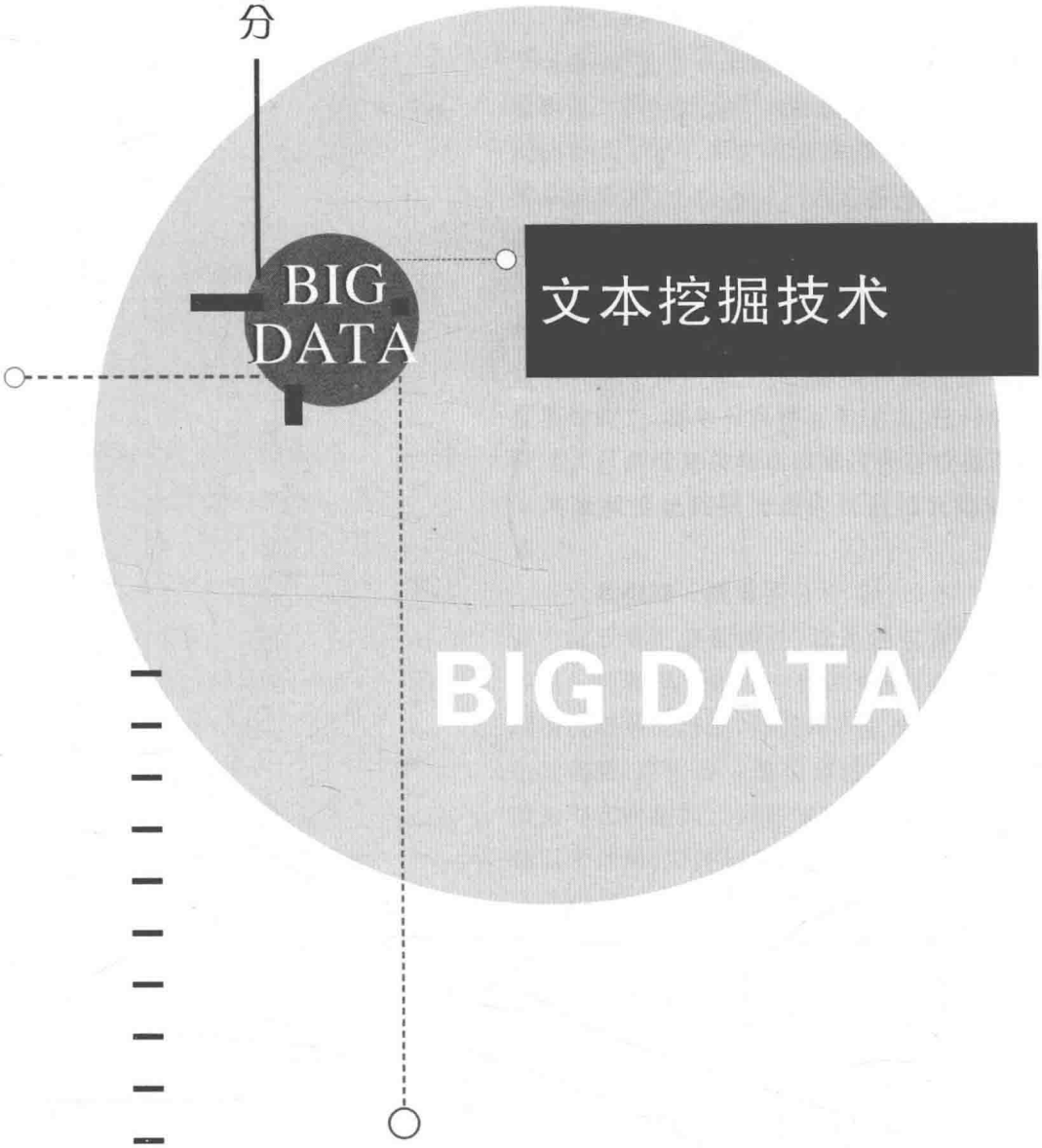
BIG DATA

第一部分 文本挖掘技术

第二部分 文本挖掘:以 R 软件为例

第三部分　文本挖掘之 SQL Server 2014

第一部分



文本挖掘技术

第一章

绪 论

大数据(big data)一词最早出现在1997年迈克尔·考克斯的论文中,直到2008年9月,《科学》杂志发表论文“Big Data:Science in the Petabyte Era”,Big Data一词开始广泛传播。“Big data”就字义上翻译是“大资料”或称为“巨量数据”“海量数据”“大数据”等,许多学者有不同的定义,亚马逊大数据科学家John Rauser将其定义为“任何超过一台计算机处理能力的庞大数据量”;Informatica中国区首席产品顾问但彬认为“大数据=海量数据+复杂类型的数据”,其规模和复杂程度超过了常用技术按照合理的成本和时限捕捉、管理以及处理这些数据集的能力;维基百科将其描述为“任何由于规模庞大且高度复杂而难以通过现有数据库管理工具或者传统数据处理应用进行处理的数据集。”

大数据和一般数据分析有什么不一样的地方?除了有更多数据外,就是还包含非结构化数据。随着Web 2.0时代的到来,电子文件呈现指数级数的增长,每天均产生庞大文件数据,包括消费、广告等一般信息或者是社会、经济、政治等实时新闻。一旦文件暴增到数以百计或数以千计时,文件与文件之间毫无关联,庞大的文件仍属一堆数据,要在短时间内阅读或是查询某一主题信息,将很困难,因而无法抓住及时信息或机会。网络媒体有别于传统媒体,每个用户都可以制造、生产信息,网络上的信息量比美国国会图书馆还多了N^N倍,这些资料都不是整理好的资料,甚至大多不是结构化数据。若想从海量的网络信息中及时准确地获取有巨大潜在价值的信息,就要求有大规模文本信息的自动处理与分析技术,需要文本挖掘技术。

1.1 整合文本挖掘与数据挖掘

文本挖掘(text mining,TM),又称文本数据挖掘(text data mining,TDM)或文本知识发现(knowledge discovery in texts,KDT),是一种跨领域的应用,其应用了信息检索、信息提取、计算语言、自然语言处理、数据挖掘技术等,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。文本挖掘整合了许多传统信息检索技术,包括关键词提取、全文检索、文件自动分类、自动摘要等等,以提供更强大的文字处理功能。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,为了向特定用户提供特定的信息,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,以及复杂的特征提取步骤。

随着计算机设备及网络技术的蓬勃发展和快速普及,许多传统的信息处理方式因此而改变,大量原本是以书面方式存在的文本信息,被转换成电子文本的形式来储存及传递,而这些文本中极可能隐藏着许多有用的宝贵知识。但是,当信息的产生和传递效率加速提升时,也产生了信息爆炸的现象,然而,传统信息检索方式无法有效地帮助使用者分析和了解大量的文本数据,许多试图从文本中获取知识的研究便因此而产生。

文本挖掘与数据挖掘的不同在于,数据挖掘所处理的数据属性是结构化的数据,而文本挖掘则是处理半结构或非结构化的数据,例如电子邮件、网页或是社群网站发文等,文本挖掘经常使用自然语言处理、统计分析、概率模式、机器学习等技术,探讨概念提取(concept extraction)、文本摘要(text summarization)、信息过滤(information filtering)、命名实体的标注或辨识(named entities tagging or identification)、意见分析(opinion analysis)、关系探索(relation discovery)、情绪分析(sentiment analysis)、文本分类(text classification)、文本聚类(text clustering)等议题。

目前在医学、法律、商业、工程、计算机等诸多领域已有广泛应用。

(1)安全类 (security application)

在此类别中最重要的应用当属 ECHELON,它可以经由卫星通信、大众交换电话网络及微波来截听全球的电话、传真及电子邮件。另外,在 2007 年欧盟智能型部门也开发了一套分析系统 OASIS(Overall Analysis System for Intelligence Support),用来追踪组织型交易犯罪。

(2)生物医学类 (biomedical applications)

在此类中第一个最有名的应用是 PubGene,它是一个搜索引擎,取自 MEDLINE (Medical Literature Analysis and Retrieval System Online) 这个含有大量生命科学及生物医学文献的数据库。它用生物医学的关键词组织成一个图形网络,用视觉方式来展现关键词与文献数据间可能的关联,目前此技术已商用化了。另外一个应用是 GoPubMed,

它是整合式搜索引擎，目标族群主要是生物医学的研究者，这是个不错的文献搜索引擎，相当方便。

(3)软件应用类 (software and applications)

一些软件公司如 IBM 及微软正在研发文本挖掘的软件，来加速挖掘及分析程序的运行。有一些公司则专注于搜索与索引领域。

(4)营销类应用 (marketing applications)

文本挖掘已经开始被运用到营销中，尤其是客户关系管理(customer relationship management)。学者 K.Coussement 和 Van den Poel 运用文本挖掘来系统性分析客户的流动率(customer attrition)。

(5)学术类应用 (academic applications)

文本挖掘对于在学术界中握有大型信息数据库，并且为了快速获取信息而要做索引的出版商来说尤其重要，因此 Nature 及 NIH 两个重要的出版商分别提出 OTMI (open text mining interface，文本挖掘开放界面)及 DTD(document type definition，文档类型定义)，使系统可以运用语意线索响应包含在文档中的特定问题。许多学术机构也纷纷投身到文本挖掘领域中，例如曼彻斯特及利物浦就合作创立了文本挖掘国家中心，提供定制化的工具及研究设备。这个中心是由 JISC 及韩国的研究委员会所共同创办，主要是在生物学及生物医学科学领域，目前已扩展到社会科学。另外，在美国加州大学伯克利分校信息系已开发了一套 BioText 系统，以协助生物医学研究者进行领域分析。

文本挖掘主要的定义为从非结构化的文字中发掘出有用的或是有趣的片段、模型、方向、趋势或规则，图 1-1 是文本挖掘的概略流程示意图。一般数据挖掘(data mining)的方法，只适用于结构化的关联表格数据，无法直接运用到非结构化的文本数据上，而一些已发表的文本挖掘研究中结合了信息提取(information retrieval)、自然语言处理(natural

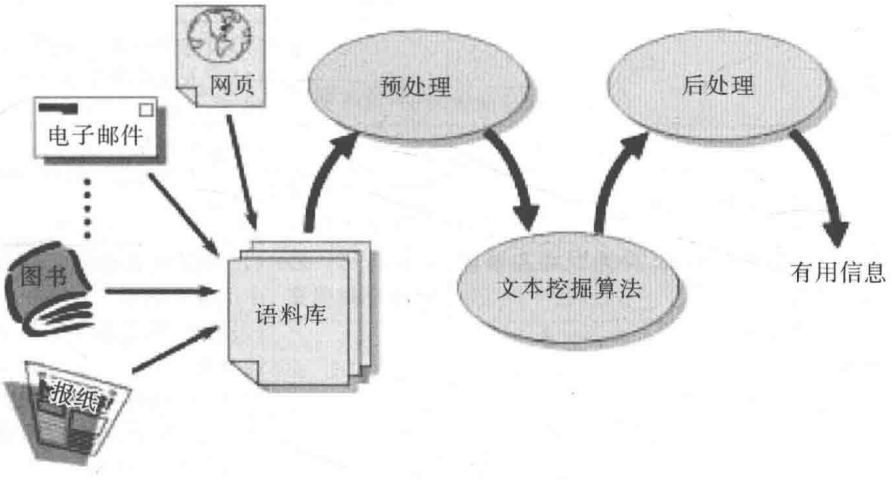


图 1-1 文本挖掘示意图

language processing)和数据挖掘的技术,试图从文档数据中找出重要的术语(term)或词组(phrase)、术语间的关联强度(association degree)或是分类和推论规则(classification or prediction rule)。虽然不同的研究者对文本挖掘的定义不尽相同,但目的都在于帮助使用者高效地从大量文本性数据中整理出有用的信息,而其中的关联信息便是本书分析的主要对象。

表 1-1 文本挖掘与数据挖掘对比

	文本挖掘	数据挖掘	交互应用的程序
资料分析	(1)主题词库建立与管理 (2)索引建立与管理 (3)文本自动分类及其结果验证 (4)文本自动聚类	分类及其结果验证	(1)Data Joiner (用于异类数据整合及数据清洗) (2)基本统计分析及可视化 (3)报告功能:报告聚类/分类/同质群组等结果 (4)个人化呈现
数据挖掘	(1)自动聚类(聚类结果须以至少具备三层以上的分类树来表达) (2)关联性分析 (3)动态查询并计算类别及来源之间符合度/笔数,提供项目流程(内含相关算法),让系统可自动找出文件中的重要主题内容:Baysian Graph 推演及归纳树 HyperGraph 自动摘要		
查询功能及使用接口	数据搜寻通过分类树或词组相关概念		查询结果可视化 个人化呈现
档案专家知识库	建立知识地图		案例式推理与归纳法则
决策支持模块		关系型法则	(1)汇整使用者的一连串查询 (2)个人化决策支持平台 (3)决策模拟 (4)风险评估
渐进式数据挖掘模块	所有功能均为渐进式	智能型协同整合系统(CIS)	
知识探索模块	相似文档侦测,并建立相互参照表 产生索引文档	使用者可以针对特定阈值聚类	(1)以图形表达公文在时间序列中的分布情形 (2)特定部门或时期的重要人事物分析 (3)统计分析各部门所处理的文档形态、数量、内容及其他参与机关等信息 (OLAP)

1.2 基础技术

数据处理与专业领域人员，在处理文本相关信息时会依据自身的专业审慎从所搜集的文本数据中筛选出自己所需的信息，并设定变化因子，控制与掌握变异，进行资料的分析比较。分析人员根据事件的差异，分析事件的发生概率，并从累积的已有知识中验证事件的发生情况与处理模式，以作为应变的根据，并随时监控处置程序与结果，依据新的处置结果所产生的证据，校正未来的处置程序，建立新的预测与警示程序。

文本研究与分析的流程往往都依赖于所累积的知识资产与专家知识；计算机科技与人工智能，可协助专家进行自动化的分析研究（见图 1-2）。

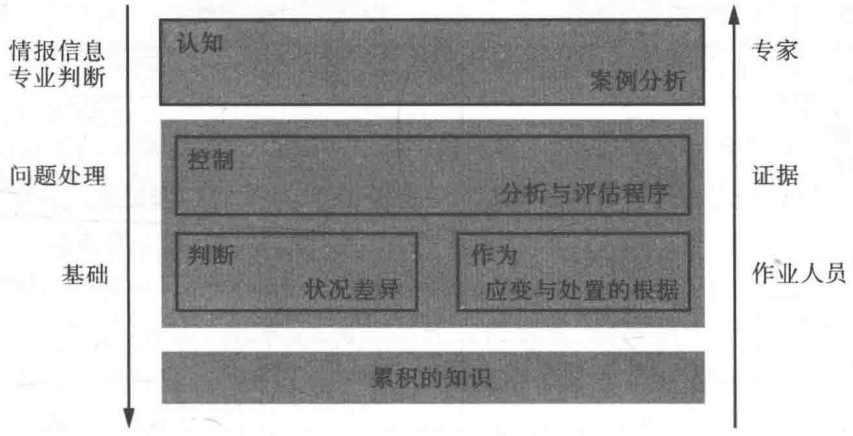


图 1-2 文本研究与分析的流程

文本挖掘的基本技术如表 1-2 所示。

表 1-2 文本挖掘技术列表

技 术	理 论	方 法	图 示
法则归纳法	相依理论 (Meo,2000)	关联分析 (Agrawal,Imielinski & Swami,1993;Chiang 2002)	贝叶斯网络 (Buntine,1993) 延展图 双曲图
自动分类	Rocchio 相似度计算 (Joachims,1997)	文件分类 (Sebastiani,2002) →SVM	
自动摘要	关联分析的关键 内容提取(Lie,1998)	DBS 综合式摘要 (Nomoto,2002)	
自动聚类	概念分析 (Ganter&Wille,1999)	高维图切割 (Lin,Chiang,2004)	高维图 (Berger,1979)