主题模型与文本知识发现应用研究

阮光册 著

华东师范大学出版社

主题模型与文本知识发现应用研究

阮光册 著

华东师范大学出版社

图书在版编目（CIP）数据

主题模型与文本知识发现应用研究/阮光册编著.
—上海：华东师范大学出版社，2018
华东师范大学新世纪学术著作出版基金
ISBN 978-7-5675-8375-7

Ⅰ. 主… Ⅱ. 阮… Ⅲ. 数据处理—研究
Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 231425 号

华东师范大学新世纪学术著作出版基金资助出版

主题模型与文本知识发现应用研究

著　　者　阮光册
组稿编辑　孔繁荣
项目编辑　夏　玮
特约审读　韩　蓉
装帧设计　高　山

出版发行　华东师范大学出版社
社　　址　上海市中山北路 3663 号　邮编 200062
网　　址　www. ecnupress. com. cn
电　　话　021-60821666　行政传真 021-62572105
客服电话　021-62865537　门市（邮购）电话 021-62869887
地　　址　上海市中山北路 3663 号华东师范大学校内先锋路口
网　　店　http://hdsdcbs. tmall. com

印 刷 者　昆山市亭林彩印厂有限公司
开　　本　787×1092　16 开
印　　张　15. 75
字　　数　282 千字
版　　次　2018 年 11 月第 1 版
印　　次　2018 年 11 月第 1 次
书　　号　ISBN 978-7-5675-8375-7/G·11533
定　　价　79. 00 元

出 版 人　王　焰

“人类正被数据淹没，却饥渴于知识”

(We are drowning in information but starving for knowledge)

——约翰·奈斯伯特（John Naisbett）

前　言

随着信息技术的快速发展，人们处在信息环境的变革之中。在数据泛滥的时代，一方面，人们意识到知识对社会和经济发展的作用越来越大；另一方面，人们获取有价值信息的代价也在不断提高。

在这种变化趋势中，知识发现作为一种工具，在知识管理和决策支持中体现了它特有的价值，并发挥着越来越重要的作用。知识发现可以针对特定的问题和需要，从杂乱无章的数据中发现对人们有价值的信息和智慧，借助技术方法解决人们的知识需求，帮助人们在数据中发现新的认知模式。

20 世纪末，知识发现作为一个新学科被人们所关注。它的理论意义在于情报分析研究的科学性，并丰富和完善信息研究的内涵。然而，随着信息技术的发展，知识发现技术也面临许多挑战，这需要我们花费更多的精力去研究和发展该学科。

面对复杂的、多变的知识需求，方法作为工具的价值不言而喻。本书将机器学习领域的研究成果引入情报分析，以主题模型作为知识发现方法的主体，借助其语义识别的能力，挖掘社会活动数据之间的联结；为研究复杂的数据关系和处理大量数据提供了一种新的研究思路和框架。文本以解决实际问题为基本出发点，将知识发现应用于不同场景，包括科技文献分析、新闻文本分析、网络用户生成内容分析等。书中内容的阐述、实例的选取、方案的提出，具有广义上的通用性。

本书是在大量实践基础之上完成的。2000 年，读硕士时，我在中科院上海有机化学研究所计算机实验室开始从事网络数据库的学习，在信息加工、系统化组织信息资源、提炼知识等方面进行了大量的实践。几年前，我开始研究语义挖掘在情报分析中的应用，以主题模型作为方法，对多种类型数据进行实践，并形成了一系

列的研究成果。在这个过程中，我逐步形成了非概论性的、结合交叉学科知识、以应用为主的研究思路。

本书得以问世，得到了多方面的支持和帮助。感谢华东师范大学新世纪学术出版基金的资助；感谢上海图书馆的张帆老师，为本书的实验提供了大量的数据资源；感谢上海图书馆的夏磊老师，为本书的实验环节提供了部分初稿；感谢华东师范大学信息管理系图书与情报 2016 和 2017 级硕士的数位同学在数据处理中给予的帮助。

本书提出了基于主题模型的知识发现研究框架，是一种新的尝试和探索，虽然有一定的实践作为基础，但仍需要进一步的检验、补充和完善。随着深度学习的发展，知识发现的研究方法必将进一步深化和扩展，也将会有新的研究思路和框架丰富这一学科研究的内涵。

限于我的学识水平，书中的遗漏和不足在所难免，还望读者不吝赐教。

阮光册

2018 年 7 月

于华东师范大学

目　录

图目录

表目录

第1章　绪论

1.1　课题背景

随着信息技术的飞速发展，人类社会不再为信息资源的“缺”而担忧，而开始转而为信息资源的“过多”而困扰。处于信息时代的人们如置身于茫茫数据海洋之上的小舟，迷失了方向，充满了迷茫。正如 Lazer 2009 年在 *Science* 发文提到的，人类正面临信息超载（information overload）的问题[①]。面对这些近乎灾难的数据资源，信息带给用户的不再是优越感，而是对其使用的迷茫。2011 年，分析调研机构 IDC 发布的数字宇宙研究报告（Digital Universe Study）——《从混沌中提取价值》（“Extracting Value from Chaos”）显示，全球互联网上的数据每年将增长 50%，每两年便将翻一番，而目前全世界 90%以上的数据是最近几年才产生的。另一项统计表明，Facebook 每天要新增 32 亿条评论、3 亿张照片，信息量达 10 TB；Twitter 每天新增 2 亿条微博，约有 50 亿个单词，比纽约时报 60 年的词语总量还多一倍，信息量达 7 TB；对淘宝而言，一天意味着千万量级交易，1.5 PB 原始记录……[②]

如果说，20 年前，互联网应用的普及方便了人们获得信息；10 年前，搜索引

① D. Lazer, A. Pentland, L. Adamie, et al. “Computational Social Science”, *Science*, 2009, Vol. 323, Issue. 5915, pp. 721 - 723.

② 数据来源：https://www.aliyun.com/zixun/content/2_6_299379.html。

擎、网络爬虫技术使得互联网变成了一个巨大的数据库；那么，当前社会化网络的应用则不仅改变了人们的生活方式，更改变了企业的运营模式和科研的研究范式。信息资源的重要性已经被提高到无以复加的程度，正所谓没有知识，任何事物都没有意义。目前 Google 等公司处理的海量语料库就如同一个人类社会的实验室，如何开发和利用这些信息资源，成为摆在人们面前的一个新的研究课题。

在众多的信息资源中，我们所面对的文本信息越来越多。文本是一种重要的数据资源，是最天然的信息存储形式，包含着丰富的知识和模式[①]，也是最为普遍和应用最广的一种信息形式。有数据显示，在组织中，有 80%的信息是以文本的形式存在的，而且大多是非结构化的数据。人们在面对这些文本信息时往往感到无所适从，要快速从中抽取出我们所关心的、切实需要的信息和知识更是难上加难。依靠人工阅读的方法获取信息，不仅费时费力，而且得出的结论掺杂了过多的主观因素，结论的准确性及质量更多地取决于“阅读者”的受教育水平、知识结构、主观认识等外部因素，不能完全客观地还原文本的真实信息，更难以发现隐藏在文本内部的各种关联和模式。

此外，大量产生的网络文本信息也为我们快速地获取知识带来了更大的困难。网络文本数据不仅形式上多样，如博客、新闻、BBS、问答社区产生的文字信息等，而且结构上也更为复杂，一般由非结构化的数据（如文本）和半结构化的数据（如 HTML 文档）构成。

面对这些非结构化文本信息，传统的基于关系数据库和数据仓库技术的数据挖掘，对非结构化、半结构化的文本信息而言，有些力不从心[②]。如何帮助人们快速获取、处理和利用这些文本集合中的知识，在充分理解的基础上获得文本集合的隐含信息和内在关系？如何将复杂的、高维度的文本数据转化为低维语义的形式？如何将文本内容提取出来，用相对直观的、简短的、有利于人们理解的形式呈现给用户？这些都是文本知识发现需要面对的现实问题。

对文本信息的处理需要采用科学的方法进行，科学的本质就是要求我们去认识一切事物的本质并加以利用。面对浩瀚的文本信息资源，需要我们剥离冗余和干

---

① 赵一鸣：《基于多维尺度分析的潜在主题可视化研究》，武汉：武汉大学出版社，2015。

② 余肖生，周宁，张芳芳：《基于可视化数据挖掘的知识发现模型研究》，中国图书馆学报，2006 年第 5 期，第 44—46 页，第 56 页。

扰，获取其精要，即认识事物要抓住其本质。伴随着机器学习、知识抽取、人工智能等技术的飞速发展，人们在使用文本信息资源开发和利用方面正面临着新的挑战和机遇。

文本知识发现是数据挖掘的延伸，其处理对象也从结构化的数据延伸到非结构化、半结构化的文本数据。随着“大数据”时代的到来，从有限的结构化数据中获取的知识已不足以满足需要，大量的半结构化或非结构化的数据需要分析，因此挖掘这些文本信息中的知识显得尤为重要。文本知识发现的目的是从无序的信息中发现潜在的、未被识别的、有价值的知识模式。文本知识发现的结果有利于消除“数字鸿沟”，有利于用户“知识获取”，有利于信息资源的重组。

1.2 研究意义

文本知识发现的研究是情报学及相关学科研究的重点领域，具有特别重要的理论和现实意义。

1.2.1 理论意义

文本集合一般包含有若干个“含义”，也可以说包含有若干个主题。主题可以表示文本的主要内容，提取文本包含的主题，将主题所包含的知识以易于理解的形式呈现给用户，将有助于发现隐藏在文本中的知识结构和模式，并发现潜在的规律特征，实现深层次的文本挖掘和知识发现。

总的来说，使用主题模型方法在挖掘、发现、解释文本集合中的潜在知识具有如下的理论意义：

1. 丰富了文本知识发现的方法体系

主题模型是一种文本语义生成模型，它实现了文本信息在语义层面上的降维表示，通过挖掘“文本—主题—词项”的相互关系，使文本结构上升为主题空间，提

高了人们对文本潜在知识的挖掘能力，也丰富了知识发现的过程。

本书使用主题模型对科技文献文档、网络新闻、UGC[①] 文本进行主题挖掘，并提取主题之间的关联等信息，通过多角度的方法设计和策略选择，发现文本集合在不同层次上的潜在知识，通过可视化的方式展现并解释潜在知识之间的内在关联。我们的研究，不仅可以提高用户对文本知识的深入理解，还可以去除文本集合中冗余的、非关键的信息，提供简洁、清晰、直观的文本知识架构，而且可以实现三个层次的知识发现：一是识别文本包含的有价值的潜在知识；二是为挖掘文本集合中知识之间的关联提供线索；三是可以发现文本中所描述知识的具体内容，进而揭示文本集的真实含义。

通过主题模型将文本的含义表示在语义空间层面，把大量文本内容转化为方便用户理解的主题，将更便于人们获取信息，大大提高知识获取的效率，而且能够挖掘出一些依靠传统阅读难以获取的系统性知识和隐性知识。

2. 从原理上克服了传统文本知识发现的不足

传统的文本知识发现是在统计文本中词对共现次数的基础上，对相关词项进行聚类。这种方法的理论依据是：词项的共现现象是产生关联的根源。然而，共现关系经常会出现高频孤立词、关键词之间缺乏语义联系等问题。本书将使用主题模型识别文本内词与词相互联系的现象，将主题看成是词项的概率分布，通过词项在文本级的共现信息抽取出语义相关的主题集合，将词项空间中的文档变换到主题空间，得到文档在低维空间中的表达。

布尔模型是传统知识发现中词项矩阵的基本处理方法，其原理是：如果两个词项共现一次，则计数一次；而主题模型则是基于概率统计学原理，将每个主题表示成一个多项式分布，对文本表达的内容进行抽象和浓缩，揭示隐藏在文本背后的语义信息。因此，每个主题是基于文本内容的潜在知识模式，更有利于揭示文本的内在知识。主题模型克服了对统计共现次数的依赖，可以发现更多隐藏的主题和知识模式，发现现有知识中“出人意料”的联系，最终可能会产生新的知识。

---

① 互联网术语，全称为 User Generated Content，也就是用户生成内容的意思。UGC 的概念最早起源于互联网领域，即用户将自己原创的内容通过互联网平台进行展示或者提供给其他用户。