数字文献资源高维聚合模型研究

牛奉高 著

中国社会科学出版社

数字文献资源高维聚合模型研究

牛奉高 著

中国社会科学出版社

图书在版编目（CIP）数据

数字文献资源高维聚合模型研究/牛奉高著．—北京：中国社会科学出版社，2017.8

ISBN 978-7-5203-0782-6

Ⅰ.①数… Ⅱ.①牛… Ⅲ.①文献计量学—研究 Ⅳ.①G250.252

中国版本图书馆 CIP 数据核字（2017）第 181653 号

出 版 人 赵剑英
责任编辑 田 文
特约编辑 陈 琳
责任校对 张爱华
责任印制 王 超

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http://www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印 刷 北京君升印刷有限公司
装 订 廊坊市广阳区广增装订厂
版 次 2017 年 8 月第 1 版
印 次 2017 年 8 月第 1 次印刷

开 本 710×1000 1/16
印 张 16.75
字 数 241 千字
定 价 69.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社营销中心联系调换
电话：010-84083683

序

G. Salton 等人于 20 世纪 70 年代提出的向量空间模型（Vector Space Model，VSM）最初应用于著名的 SMART 文本检索系统，而今已成为重要的文档表示模型和相似性计算模型，广泛应用于文本分类和聚类、图像处理等诸多重要领域。VSM 的提出可谓是检索领域的一次革命。VSM 在被广泛应用的同时，也有诸多改进和扩展。本书在 VSM 的框架上，引入共现理论挖掘词间潜在语义，提出了共现潜在向量空间模型（Co-occurrence Latent Semantic Vector Space Model，CLSVSM），经检验在文档聚类中有较好的表现。进一步讨论了模型的适应性、基于语义核的约简、共现语义的扩展等问题，使其应用更加方便、快捷和准确。CLSVSM 的提出在一定程度上规避了对复杂背景知识的依赖，提高了效率，具有普适性。

本书共分七章。

第一章主要介绍了数字文献资源聚合的相关概念和理论基础。从检索角度出发，解释了文献聚合的内涵和外延，探索了其核心理论和技术，并论述了聚类方法在文献主题聚合中的重要性和可行性。

第二章分析了数字文献资源聚合的主要途径和方法，重点是文献的表示和向量空间模型（VSM）。随着海量文献的涌现，文献集的向量表示存在着高维稀疏的特征。此时，基于 VSM 的方法就难以实现理想的文献主题聚类。虽然有研究引入了语义向量空间模型，但是对背景知识依赖性强，建设成本高。能否构建一种较好的模型，正是本书所要解决

的主要问题。

第三章论证了潜在语义分析在文献主题聚合中的可行性并提出共现潜在语义向量空间模型（CLSVSM）。共现潜在语义对文献主题的相似性度量很有帮助，在广义向量空间模型和语义向量空间模型中都体现了共现潜在语义的重要性。因此，本文在 VSM 的基础上，通过挖掘并利用共现潜在语义构建了 CLSVSM。其最大的优势是可以较好地测度特征词（维度）与文献向量的相似性，从而提高了文献主题聚类的准确性，最终实现文献的主题聚合。

第四章主要是对 CLSVSM 模型的检验。经过详细的设计和实验对比，新模型显示出了较 VSM 提高很多的效果。更多的比较结果将在第六章呈现。由于目前没有理想的数据类可以直接使用，因此本文的数据集均为自己采集整理而成。文献主题在分类中的模糊性较大，数据集的质量很难保证，因此不能完全体现模型的效果。

第五章则是对模型的应用。

第六章是对模型的深入讨论，旨在提高运算效率。其中重点是语义核的构建和三元共现的挖掘和使用，新模型在文献主题聚类中的精度有所提高。

第七章是全文总结。

全书在武汉大学邱均平教授指导下由本人撰写完成，在后期的增补和结构调整中我的硕士生张亚宇做了不少工作。

本书适用于信息检索、文本分析等相关领域的研究和应用参考。由于本人才疏学浅，书中难免疏漏，真心期望各位读者批评指正。

牛奉高

2017 年 3 月 6 日于山西大学渊智园

目　录

引　言

物质、能量和信息并称为当今人类社会三大基本资源，信息资源在人类社会经济发展中具有不可替代的地位。事实上信息与宇宙同在，信息的存在是客观的，但直到20世纪80年代以后信息资源才逐渐被人类重视，随着人们开始积极利用信息，加工信息以得到更有用的信息等等，信息才真正成为人类社会的重要资源。随着人类对信息认识的不断深入，其外延几乎无所不包，其中物质和能量不仅是信息的载体，自身也成为了信息。信息资源的重要性已经被提高到了无以复加的程度，正可谓是“没有信息，任何事物都没有意义”。

信息资源自20世纪开始得到了快速的增长，据估计其总量已经超过了之前人类历史的总和，已经成为人类社会的重要资源。特别是Internet的出现和普及，将我们置于信息的海洋之中。我们不再为信息资源的“缺”而担忧，但开始因为“多”而无从选择。处于信息时代的人们如置身于茫茫大海之上的小船，如遨游在太空的小尘埃，没有方向，充满了迷茫。然而人之所以为智人，就是人类认识世界的主观能动性，人类的发展史就是探索世界的发现史。马克思主义认识论所言极是：认识事物要抓住其本质。科学的本质就是告诉我们去认识一切事物的本质并利用之。面对浩瀚的信息资源，同样我们要剥离其他冗余和干扰，获取其精要。

一 研究背景和意义

（一）研究背景

1.“大数据”与“高维数据”并存对人们有效获取信息提出挑战

信息资源的形式是多样的，数据只是一种存在形式。人类文明的发展，产生了海量的信息，但随着科技的发展，这些信息才被大量的记录和保存下来。仪器观测数据、传感数据、实验数据、文献资料数据、统计数据、基因排列数据以及数据的数据等等，充斥在我们生活的周围，数量大，变化快，超出了我们的处理能力。因此我们称这个时代是“大数据时代”。

“大数据”作为互联网信息技术行业的流行词汇，大约从2009年开始迅速传播，几乎影响了每个行业和每个人。事实上，大数据一词早有提及，1980年著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。大数据（Big Data）在信息技术的视角下指的是所涉及的资料规模巨大到无法在合理时间内通过目前主流软件工具实现对数据信息的撷取、管理和处理，因此又可称为巨量资料。面对大量的信息，企业却无法通过它们实现企业经营决策的目的。在我们的研究学习中也会发现可选的资料和文献越来越多，反而让我们无从确定我们真正需要的。这种普遍的“信息过载”现象体现了大数据体量大（Volume）的特点。体量大是大数据的最主要特点，目前很多企业级的数据库体量规模在10TB以上，一些企业的多个数据集放在一起就已经是PB级的数据量了。更何况，据美国互联网数据中心的报道，互联网上的数据每年将增长50%，每两年便将翻一番，而且全世界90%以上的数据是最近几年才产生的。大数据所造成的信息爆炸现象要求实时提高处理水平，因此对传统工具和方法造成困难，也着实推动了技术的进步，比如存储速度的提高，同体积下存储量的快速提升，CPU数据处理速度的加速，显示技术的发展等等，这些技术都为了应对大数据时代而不断改进，这是大数据的第二

个特点——数据处理速度快（Velocity）。大数据的第三个特点是数据类型多样（Variety），一般有多种数据源、多种类型、多种结构和多种格式等，已超出了程序可以处理的结构化数据范畴，囊括了半结构化和非结构化数据。统计数据已不能满足人们对信息分析的要求，而需要自然语言处理和人工智能等方法进行文本挖掘、语义分析等更复杂的信息分析研究。大数据带来的巨量信息必然有很多冗余，我们可以不去注意它们吗？事实上，人类社会的很多规律和行为特征就可能隐藏在巨量信息中，比如社交网络数据、企业内容、交易数据等都是重要的数据源，因为其中包含着最真实的社会和人的特征，这就是大数据的第四个特征——真实性（Veracity）。大数据呈现的是第一手信息，而不是二手的，基于这些数据的分析结果就更能引起人们的兴趣，更能体现社会规律。除了真实性之外，还应该包括全面性。采用所有数据分析的方法，而不用随机分析法（抽样调查）这样的捷径，实际上是对经典统计方法的一种否定，这也超出了传统数据源的局限性。欣慰的是，计算机存储能力和运算能力的提高为分析大数据提供了一定的条件。近年来，围绕是否需要抽样有过很多讨论。但在实际应用中由于“大数据”的价值密度低抽样还是需要的。

数字文献资源数量快速增长，降低了文献的查全率和查准率。同时，随着社会的进步，人们获取信息的要求逐步向知识发现的水平发展。因此，基于文献主题和内容的分析成为必然，这就要求对文献资源深度聚合研究。

现代数据信息的另一种特性是高维数据。高维数据是描述数据的特征维度很大。数据产生高维的原因是表征事物的全面性和事物间关联性的加强。比如描述一篇文献的属性常采用其元数据格式，如果将其每个属性视为一维，显然文献的属性是高维的。再如文本的描述采用词向量来表示，那么随着文本内容的增加，其表示向量的维度一般就会越来越高，面对大量的文本资源，其特征词的维度就更高。此外，在一些顾客资料数据库中，有 50 个指标；在进行网站的内容聚类时，会有 200—1000 个属性；生物学研究中的基因整列数据可以拥有超过 2000—5000

个属性。如果考虑时间序列数据，时间点作为维数的话就有很大的伸缩性。可见多维数作为研究事物的一种视角必然会带来数据的高维度。

究竟要多少维才算高维？以上也说明，数据维度与分析对象的特征有关系。陈建斌认为高于16维的数据就属于高维数据，聚类算法中所使用的索引在维数小于16时会有效发挥作用，但当维数大于20时，它们的性能会降低到顺序搜索的水平[①]。这样看来我们的很多研究都已经是面对高维数据了。

高维带来的困难是什么？随着维数的增加，很多维数是相关的，这给传统方法带来困难。如在聚类分析中只有少数的特征维对形成簇是有意义的，而大多数特征维是与其相关的，但是这些相关维的数据可能会产生大量的噪声而屏蔽真实的簇，使聚类算法失效。再者，随着维度的增加，表示事物的向量数据通常会变得更加稀疏，即0越来越多，有意义的数值相对较少，因为数据点可能大多数分布在不同的特征子空间中。当数据变得特别稀疏时，位于不同维的数据点间的距离区分度就会降低，如果距离度量失去了意义，聚类分析的结果就不可信。高维数据所带来的这些困难，被研究者称为“维度灾难”或“维灾”（curse of dimensionality）[②]。

综上，高维作为数据的统计特征成为描述事物的必然趋势，大数据作为时代特征之一对其分析意义非凡。如果数据集皆有两个重要属性，即高维的大数据，就需要专业的分析水准。而事实上现有的数据往往兼而有之：电信公司的客户资料是高维的而且数据量不断迅速增大，大型语料库的文本书档就是用成千上万的特征词表示，生物学研究的DNA微阵列数据在数以百万计的条件下提供关于数以千计的基因表达水平信息，等等。这些都需要我们提高数据挖掘的技术水平：大数据对数据存储和运算能力提出了挑战，而高维数据更是需要处理技术的改善。

---

① 陈建斌：《高维聚类知识发现关键技术研究与应用》，电子工业出版社2009年版，第3页。

② R. Bellman. On the Reduction of Dimensionality for Classes of Dynamic Programming Processes [J]. Journal of Mathematical Analysis and Applications, 1961, 3 (2): 358 - 360.

2. 信息组织、知识发现和知识服务需要资源聚合

资源聚合是通过全方位、多层次、多手段对资源的集成整合，而通过资源多属性或多层次的聚类是实现资源聚合的方法之一。既可以通过前期集成，后期聚类形成，也可以直接通过聚类得到聚合的形式。而所有的过程都应该是面向信息组织、知识发现和知识服务。

信息组织需要资源聚合的手段。面对浩瀚的信息资源，全面的整合相关资源无疑是信息分析的理想前提。随着计算机的普及和网络的盛行，信息获取和信息组织主要依赖检索手段，正如百度和谷歌成为主要的信息搜索和组织工具。从信息组织的角度而言，信息聚合主要是指内容的聚合，其中最重要的就是内容聚合器（content aggregator）。内容聚合器是一个从不同线上资源收集网络内容并重用或重新销售的个体或组织。根据不同的目的，内容聚合器可以分为两种：一种是简单的从网站收集不同的源信息；另一种是为了客户需要收集和分发内容资源聚合的知识发现功能[①]。关键词或特征词匹配是最流行的信息检索手段，但已不能满足人们对信息组织的需要。要获得更相关的结果，就需要深度聚合，即通过语义在内容层面的检索和聚类。这样才能实现信息更加有序的组织和利用，实现信息熵的下降。

知识发现需要资源聚合的技术。知识发现是从数据集中识别其精华以及新的模式的过程，基于数据库的知识发现（Knowledge Discovery in Database，KDD）是知识发现研究的主题和热点[②]。从概念范畴来看，数据挖掘是知识发现的一个阶段，当然也有人认为两者是等同的。而聚类作为数据挖掘的重要方法，自然是知识发现的重要手段。在众多信息资源中检索出相关的信息，再从相关的信息中挖掘出新的知识是知识发现的一般流程。人类的肉眼可以简单识别少量三维及以下的数据特征，并从中发现新的知识，但是面对大数据和高维数据就无能为力了。聚类

① 刘明辉、张志平、张新民：《网络资源聚合方法探析》，《机械管理开发》2008 年第 23（5）期。

② 陈建斌：《高维聚类知识发现关键技术研究与应用》，电子工业出版社 2009 年版，第 3 页。

分析通过挖掘数据集中的共性集，提供新的数据集特征和知识。比如通过关键词的聚类对文献资源的聚合，可以发现研究主题群和研究热点等。

知识服务需要资源聚合的方法。在信息爆炸、知识充斥的时代，用户需要更准更快地获取到需要的知识，因此需要一种提供知识服务的媒介。首先从众多信息资源中检索出相关的信息，经过信息组织和知识发现等手段的处理，最后推送给用户。目前图书馆作为知识服务的载体，正在向数字图书馆的方向发展，也会提供更加全面智能的推送服务。聚类分析秉承“物以类聚”的思想，和价值高的信息相似或相近的信息也是高价值的，而和无用的信息相近或相似的信息自然也是无用的，是智能服务的主要方法。

馆藏资源的深度聚合已经成为前沿的研究热点，而馆藏资源又以数字文献资源为主。数字文献资源有节约实体空间、便于维护、便于检索等诸多有利方面，因此比较盛行。基于数字文献资源的知识组织、知识发现和知识服务成为研究的重点。

3. 信息资源分布的“长尾”现象不可回避

齐普夫（Zipf）定律是资源分布规律的重要描述，其体现为一种幂律（Power Laws）特征。比如词频分布，即词频从高到低排列就是典型的幂律分布。齐普夫定律可分为两个定律，即高频词分布定律和低频词分布定律[①]。在实际研究和应用中，更多关注了高频词的分布。因此有高频词共现分析，以及在此基础上的聚类分析，来解释研究热点等。低频词虽然频次低但是分布广，体量大，它们的作用不能被完全抹杀，这就是词频分布的长尾现象。

根据维基百科，“长尾”（The Long Tail）一词的提出是描述诸如亚马逊和 Netflix 之类网站的商业和经济模式。严格讲，长尾分布是概率统计中重尾分布的一个分支（另一支是次指数分布）。长尾分布又称厚

① 邱均平：《信息计量学（五）第五讲　文献信息词频分布规律——齐普夫定律》，《情报理论与实践》2000 年第 5 期。