数字时代图书馆学情报学青年论丛

（第二辑）

基于多维尺度分析的潜在主题可视化研究

A Research on Underlying Topics Visualization Based on MDS Model

赵一鸣　著

WUHAN UNIVERSITY PRESS
武汉大学出版社

数字时代图书馆学情报学青年论丛

（第二辑）

本书的出版获得以下项目的资助：
国家自然科学基金青年项目（编号：71403190）
中国博士后科学基金特别资助项目（编号：2015T80840）
中国博士后科学基金面上项目（编号：2014M552090）

基于多维尺度分析的潜在主题可视化研究

A Research on Underlying Topics Visualization Based on MDS Model

赵一鸣　著

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

基于多维尺度分析的潜在主题可视化研究/赵一鸣著. —武汉：武汉大学出版社,2015.10

数字时代图书馆学情报学青年论丛. 第2辑

ISBN 978-7-307-17023-0

Ⅰ.基… Ⅱ.赵… Ⅲ.图书馆工作—可视化仿真—研究 Ⅳ.G250.7

中国版本图书馆CIP数据核字(2015)第248971号

责任编辑:王智梅　　责任校对:汪欣怡　　版式设计:马　佳

出版发行:**武汉大学出版社**　(430072　武昌　珞珈山)

(电子邮件:cbs22@whu.edu.cn　网址:www.wdp.com.cn)

印刷:湖北省荆州市今印印务有限公司

开本:720×1000　1/16　印张:12.25　字数:176千字　插页:1

版次:2015年10月第1版　2015年10月第1次印刷

ISBN 978-7-307-17023-0　定价:26.00元

前　言

潜在主题的发现，可以帮助人们迅速获得文本集的中心思想、内容特征，挖掘隐含的信息。如果能使用可视化方法提取文本集包含的潜在主题，将潜在主题表示在可视空间中，人们就可以直观、快速地获取文本集的中心内容，发现隐藏的知识结构和模式，发现潜在的规律特征，实现深层次的文本挖掘和知识发现。

本书报告了一个潜在主题发现方面的研究成果。研究的目标是“使用可视化方法表示、挖掘、呈现和解释文本集包含的潜在主题，展示不同层次和观测水平上的潜在主题与发现主题之间的关联，将潜在主题可视化方法应用于特定领域的文本知识发现”。

本书认为，可以找到一组在文本集中具有集聚关系的词条集合来表示潜在主题，将这种集聚关系抽象出来，就能得到从属于同一个主题的词条集合。为了将集聚关系抽象出来，使用词条在转置向量空间中的邻近关系表示词条在原始文本集中的集聚特性，有集聚关系的词条会在高维转置向量空间中相互邻近。由于高维空间不具有可以观测的几何结构，所以本书选择 MDS 可视化的方法将词条在高维空间中的邻近关系投影到人们可视的低维 MDS 空间图中，使用低维的空间对象结构来映射高维空间中的对象之间的关系和结构。由于保持了高维空间中的拓扑结构，从属于同一个主题的词条在低维可视空间中仍然相互邻近，在 MDS 空间图中形成一个个类团，每一个类团就是一个潜在主题。这个方法流程克服了共词分析

和数据库内容结构分析使用 MDS 进行空间聚类时对统计共现次数和必须事先选定种子词等步骤的依赖。

本书论证了使用词条集合表示潜在主题的原理、在转置向量空间中词的邻近关系表示集聚关系的原理、用多维尺度分析（MDS）将邻近关系投影到低维空间的原理，构建了使用 MDS 可视化方法挖掘并展示潜在主题的基本流程。

全书共分六章，主要围绕文本集中潜在主题的发现、表示、呈现和解释这一主线，逐层深入，主要分为原理论证、方法设计、实际应用三个模块，章节安排如下：

第一章为绪论。从实际工作对文本挖掘与知识发现的需求入手，引出了使用可视化方法来挖掘文本集中潜在主题的总体思路；从基于文本单元聚类的主题发现、基于共词分析的主题发现、基于数据库内容结构分析的主题发现和基于概率主题模型的潜在主题发现等方法进行了文献评述，并与本书的思路、方法和技术路线作了深入比较；接着提出了研究目标和拟解决的科学问题，报告研究方法和技术路线。

第二章为理论基础。本研究属于文本挖掘与知识发现的学科领域，所以，从研究的实际内容出发，着重介绍了与研究主题内容有密切关联的理论基础，包括文本的向量空间表示、文本的特征选择与提取、信息可视化、聚类知识发现，等等。

第三章是潜在主题可视化的基本原理和流程。本章论述了使用可视化方法表示、挖掘和展示的基本原理，包括用具有集聚关系的词条集合表示潜在主题的原理、用转置向量空间中的邻近关系表示集聚关系的原理、用 MDS 将邻近关系投影到低维空间的原理，重点分析了使用 MDS 可视化方法表示潜在主题的可行性和优势，构建了潜在主题可视化的整体流程，详细论述了具体步骤及每个步骤的基本原理。

第四章是潜在主题可视化的方法。根据潜在主题可视化的基本原理，针对实际应用中可能存在的局限和难点提出解决方案，改进、完善和优化潜在主题可视化的流程。本章解决的问题主要有：

为了突破可视空间只能展示有限个对象的缺点和克服 MDS 空间聚类结果在可解释性、可理解性方面的欠缺，将扎根理论的思想和部分方法融入潜在主题可视化中，一是在 MDS 进行降维之前引入开放式编码的环节，二是在得到 MDS 可视化结果之后返回原始文本集进行扎根性分析，对整个可视化流程进行了第一次重塑和优化；提出了领域情景、主题情景、上下文情景三个层次的情景模型，为可视化方法的改进提供了入口；为了能在不同观测水平上研究潜在主题、发现同一层次潜在主题之间的关联、解释主题的关联、还原更多的上下文情景、寻找新的潜在主题，设计了三个层次的应用方法，分别是：基于邻近矩阵、基于质心邻近矩阵、基于属性叠加邻近矩阵的可视化方法，并在加入了这些方法以后，完成了对可视化流程的第二次重塑和优化。至此，本书构建了一套完整的潜在主题可视化方法流程与策略体系。

第五章是潜在主题可视化在上市公司风险识别中的应用。使用潜在主题可视化的方法，以计算机应用服务业的 97 家上市公司招股说明书中关于“风险因素”的文字描述为目标文本集，进行上市公司风险识别的知识发现。研究结果表明：潜在主题可视化的方法体系成功挖掘、展示并解释了上市公司风险文本中不同层次的潜在主题及其内部结构，发现了潜在主题之间的关联，实现了多层次的知识发现。

第六章是总结与展望。对研究工作进行了全面总结，指出了研究中存在的不足和局限性，并就下一步的工作进行了展望。

本书主要是在笔者的博士论文和博士后研究基础上完成的。在此，首先衷心感谢我的博士导师黎苑楚研究员、博士联合培养导师张进教授、博士后导师马费成教授的精心指导和大力帮助。武汉大学出版社老师在本书出版的过程中也付出了大量辛勤的劳动。

本书的出版得到了相关项目的资助，它们分别是：国家自然科学基金青年项目（编号：71403190），中国博士后科学基金特别资助项目（编号：2015T80840），中国博士后科学基金面上项目（编号：2014M552090）。

本书参考了大量其他学者的研究成果，在此一并表示感谢。由于水平有限，书中难免有疏漏之处，恳请专家和读者们不吝赐教。

赵一鸣

2015 年 6 月于珞珈山

目　录

第1章 绪 论

本章从文本挖掘和知识发现面临的现实问题入手，提出使用可视化方法挖掘文本集潜在主题的思路。

内容安排如下：

1.1节阐述了使用可视化方法挖掘、发现、展示并解释文本主题的优势和意义。

1.2节从基于文本单元聚类的主题发现、基于共词分析的主题发现、基于数据库内容结构分析的主题发现、基于概率主题模型的潜在主题发现四个方面介绍了国内外研究现状，并与本书的研究思路进行了深入的对比和评述。

1.3节明确了研究目的与研究问题。

1.4节指出了研究方法与技术路线。

1.5节报告了本书的特色与创新点。

1.1 研究背景与意义

文本是重要的数据资源，是最天然的信息存储形式，包含着丰富的知识和模式。弗雷斯特咨询公司（Forrest Research）的统计资料指出，80%以上的数据以非结构化的文本形式存在，比如各种文

档、手册、网页、科技论文、研究报告、E-mail 等①。

以上市公司的信息资源为例。上市公司有信息披露的义务，需要发布大量生产与经营方面的文本信息，比如招股说明书、招股意向书、年报、重大事件公告等。这些都是针对上市公司展开研究的重要文字材料，而人们面对这些大规模文本信息时往往感到无所适从，要快速从中抽取出人们密切关心的、切实需要的信息和知识更是难上加难。普通人的阅读速度是 200～240 字/分钟②，而招股说明书的篇幅一般在 20 万字以上，读完一篇 20 万字招股说明书需要花费 667 分钟。即使用户只需要获取其关心的一部分信息，也将耗费大量的精力。比如，用户希望获得计算机应用服务业上市公司风险方面的信息，则需要阅读招股说明书中 496343 个字的文字内容，全部读完需要 2068 分钟（34.5 小时），且阅读一遍并不意味着能完全获取所需的重要信息。

可见，依靠传统人工阅读的方法获取信息，不仅费时费力，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高低更多地取决于“阅读者”的受教育水平、知识结构、工作经验等外部因素，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集内部的各种关联和模式。

如何帮助人们快速获取、处理和利用这些特定领域文本集合中的知识，在充分理解的基础上获得文本集合的内在关系和隐含信息？如何将文本内容提取出来，用相对直观、简短的方式向用户呈现？如何将复杂的高维文本数据转化为人们可以直接观测的可视化图形，进而发挥人们的思维判断能力？

这些都是文本挖掘与知识发现需要面对的现实问题，具有重要的现实意义。信息可视化技术可以很好地解决这些问题。

一幅图胜千句话，人处理图形图像等视觉信息的效率远高于文

① 韩客松，王永成．文本挖掘、数据挖掘和知识管理——二十一世纪的智能信息处理［J］．情报学报，2011，18（1）：100-104.

② 东尼·博赞．快速阅读［M］．北京：中信出版社，2009：25.

本，根据 Zeki 的研究①，人类的视觉大脑皮层中的 4 个平行系统会同时工作来处理视觉输入，这种并行处理机制让人对图形图像的感知能力大大超过对线性、抽象文本的感知能力。如果可以把文本集的主要内容呈现在二维或三维的可视空间中，则可以更高效地发挥人的认知能力去理解和探索复杂的文本对象。

使用可视化技术展示文本集的内容，可视空间中的对象可以是文本、段落、词条，也可以是概念、主题。相比而言，主题可以代表文本或文本集的中心思想，更能精练地反映文本集的主要内容，对文本集有更强的概括能力。

文本集必定包含着若干个主题，也可以说包含了若干类“含义”。文本、词条及词条的频率是可以观测的，而文本集包含多少主题、包含哪些主题、哪些词条属于同一个主题等信息则不可直接获取（Unobserved），文本集中混合的主题（一类相关的词所表示的含义、语义）是隐藏的、不能观测的，因此称为“潜在主题”。

潜在主题可以表示文本集的主要内容，如果能使用可视化方法提取文本集包含的潜在主题，将潜在主题表示在可视空间中，把大量的文本内容转化成可视的图形图像，人们就可以直观、快速地获取文本集的中心内容，并能够发现隐藏的知识结构和模式、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，使用可视化方法挖掘、发现、呈现并解释文本集中的潜在主题具有以下意义：

（1）丰富了文本挖掘与知识发现的方法体系，允许用户以图的方式浏览文本集的概貌和细节

可视化是计算的一种方法，它将数据信息转换成几何形态，通过对研究对象相互关系的空间展示，使传统的线性文本结构转变为可视化的立体结构，丰富了科学发现的过程并提高了人们洞察复杂

① Zeki S. *A Vision of the Brain* [M]. London: Blackwell Scientific Publications, 1993.

和潜在知识的能力，是一种为研究人员提供潜在信息的方法①。

本书使用 MDS 对相关单词的空间聚类功能提取并表示潜在主题，通过多层次的方法设计和策略选择，使研究者可以根据兴趣和研究需要在不同观测水平上发现潜在主题及其之间的联系，可以在不同层次上发现潜在主题，并解释潜在主题之间的关系。不仅可以提高读者的阅读效率，使用反映主题的最核心的词条表述原文的意思，还去除了冗余的、非关键的信息，提供简洁、清晰、直观的文本主题视图。本书可以实现三个层次的知识发现：一是将文本集包含的高频词条分成若干个具有内在紧密联系的小类，发现新的细分主题；二是为类与类之间的关系判断提供线索；三是可以发现每一个主题下包含的主要内容，进而揭示文本集的真实含义。

将文本集包含的潜在主题表示在可视空间中，把大量的文本内容转化成可视的图形图像，通过空间展示把传统的线性文本结构转变为可视化的立体结构，将更加符合人们获取信息的视觉偏好，大大提高知识获取的效率，而且能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）将文本单元集聚发现主题的粒度细化到了词条的层面，可以实现面向知识单元的主题展示

通过文本聚类发现主题只能将一个文本与一个主题建立联系②，将文本单元发现主题的粒度细化到文本片段、自然段、自然句的层面，可以发现文本中的多个细分主题。若要更深入地揭示主题的语义内容，有必要将主题发现的表示对象细化到词语的层面，因为词是文本中最小的语义单元。而且，使用词条的聚类发现主题，不仅可以覆盖文本内更多的主题信息，还可以用词条来表示主题和解释主题。本书将文本单元集聚发现主题的粒度细化到了词条的层面，可以实现面向全文、面向内容、面向知识单元的潜在主题

① Angelika Zartl, Edgar Schiebel. The Combination of Content Maps by Co-Word Analysis [J]. *Proceedings of SPIE*, 2002, 359 (3): 359-367.

② Blei D. M., Ng A. Y., Jordan M. I.. Latent Dirichlet Allocation [J]. *The Journal of Machine Learning Research*, 2003, 3 (3): 993-1022.

挖掘、呈现和解释。

（3）从原理上克服了共词分析和数据库内容结构分析应用于知识发现的障碍

共词分析和数据库内容结构分析（DT）都是在统计词对共现次数的基础上，对相关词条进行聚类。而词条之间的集聚关系则是共现现象产生的根源，本书正是使用集聚关系来描述这种词与词相互联系的现象，把词条的集聚关系提取并呈现在可视空间中，使在意义上关联的词条在可视空间中集聚成潜在主题。

共词分析和数据库内容结构分析（DT）构建共现矩阵是基于布尔模型，如果两个词条共现一次，则计数一次；而本书则是基于向量空间模型，通过将词条表示在转置的向量空间中，使用词条在高维空间中的邻近关系表示集聚现象。

数据库内容结构分析（DT）将共词分析应用到全文，但需要事先选定种子词。由于潜在主题的特性，人们不一定能事先观察到文本集中隐含的主题，所以，事先选定种子词的做法限制了对新主题及隐含主题的发现。本书克服了对统计共现次数和事先选取种子词的依赖，因此可以发现更多隐藏的主题和知识模式，实现“提供‘认识论层次的价值’，发现现有知识中‘出人意料’的联系或问题，最终可能会产生新的知识”①。

（4）弥补了主题模型在潜在主题展示方面存在的不足，可以直观地揭示潜在主题之间的关联

以 pLSA、LDA 及其扩展模型为代表的主题模型在潜在主题的展示方面显得略为不足，大多只是以列表的形式给出了潜在主题包含的词条，用户很难理解单个词条在上下文中的含义②。这种呈现方式只能得到潜在主题的词条构成，观察不到主题之间、词条之间的关联，而且所有的潜在主题都是平行关系，不能挖掘出主题之间

① 朱幼平．论信息化对经济增长的影响［J］．情报理论与实践，1996（5）：35-46.

② Blei D. M., Lafferty J. D.. Visualizing Topics with Multi-Word Expressions［J］. arXiv：0907. 1013v1［stat. ML］, 2009.

上下位的层次关系。不同于 LDA 等主题模型使用参数估计、三层贝叶斯网络来提取潜在主题的方法，本书使用另一种思路，即直接使用可视化的方法提取潜在主题。

（5）融入了扎根理论的潜在主题可视化可以发现词汇关联中有意义的知识模式①，为进一步研究提供研究线索和启发。

本书首次将共词聚类分析和信息可视化技术引入上市公司行业分析，通过旋转、调整观测角度等人机互动方式，揭示研究对象的空间邻近关系，发挥人的智力判断，并发现知识内容上的关联，形成一系列假设，以便专业研究人员进一步证实，推进新知识的产生，为科研工作人员提供新的思路和启发。

1.2 国内外研究现状

本书通过关联词条的空间聚类来发现并表示主题的思路来自“文本单元聚类形成主题”的启发，并将用于聚类的文本单元的粒度细化到词的层面，不仅可以用于发现主题，还可以用于表示和解释主题；本书使用多维尺度分析（MDS）对文本集中的关联词条进行空间聚类，而关键词共词分析也经常使用 MDS 对有共现关系的关键词进行空间聚类和可视化，DT 方法将共词分析应用至全文，与本书的方法比较相似，但也有本质区别；概率主题模型的方法也是用一组关联词条表示文本中的潜在主题，对本书的整体构思起到了重大的启发作用，但本书是直接使用可视化的方法提取词条的集合来表示主题。

下面对这几个与本书有密切关联的领域、对本书的创作有重大启发的文献进行综述。

① 王曰芬，宋爽，熊铭辉．基于共现分析的文本知识挖掘方法研究［J］．图书情报工作，2007，51（4）：66-70.

1.2.1 基于文本单元聚类的主题发现

通过文本聚类发现主题是使用词袋法把文本表示在向量空间中，通过计算文本在向量空间中的相似度对文本进行归类，并为每一类文本概括一个主题，进而对文本内容进行归类或重新组织，帮助人们分析和理解高维数据，发现潜在的知识①②。与本书相关的研究有：计算奇异值分解之后的文本向量之间的相似度，在一定程度上实现了基于语义的文本聚类和主题发现③；基于主题地图的方法进行文本聚类④；使用非线性映射的可视化技术（MDS）实现文本聚类⑤，等等。

为了发现文本中的多个子主题，很多学者使用文本片段聚类进行主题发现、主题抽取，比如 Salton 等⑥、Yaari Y.⑦ 及孔庆莘等⑧，以自然段落作为基本单元，将自然段落表示为词向量，并用词汇相似度和层次聚合性聚类识别文档的层次结构，一定程度上实

① 史忠植．知识发现（第二版）[M]．北京：清华大学出版社，2011.

② Pons-Porrata A., Berlanga-Llavori R., Ruiz-Shulcloper J.. Topic Discovery Based on Text Mining Techniques [J]. *Information Processing & Management*, 2007, 43 (3): 752-768.

③ 王国勇，徐建锁．TCBLSA：一种中文文本聚类新方法 [J]．计算机工程，2004，30 (5)：21-22.

④ 吴江宁，田海燕．基于主题地图的文献组织方法研究 [J]．情报学报，2007，26 (3)：323-331.

⑤ 杨峰，周宁，吴佳鑫．基于信息可视化技术的文本聚类方法研究 [J]．情报学报，2006，24 (6)：679-683.

⑥ Salton, Gerard, Amit Singhal, Chris Buckley, and Mandar Mitra. *Automatic Text Decomposition Using Text Segments and Text Themes* [C]. *In Proceedings of the Seventh ACM Conference on Hypertext*, *ACM*, 1996: 53-65.

⑦ Yaari Y. *Segmentation of Expository Texts by Hierarchical Agglomerative Clustering*. arXiv preprint cmp-lg/9709015, 1997.

⑧ 孔庆莘，刘宗田，廖涛．基于概念获取的多文档主题划分研究 [J]．计算机科学，2008，35 (5)：131-133.

现了文本子主题的层次分割；胡珀、何婷婷①在其研究中，认为作者会将表达中心思想的主题反复表述，由于表达方式各异，不同主题往往散落在文本的不同段落中，进而通过文本段落的自适应聚类发现文本的潜在主题。

由于该方法以词条作为向量空间的维度来表示段落群、段落、句子，所以其表示文本主题的粒度最小只能细化到句子的层面，而不能细化到语义单元（词条）层面。也就是说，文本单元聚类只能发现主题，却不能很好地在词条层面表示主题、解释主题。

有学者以词为聚类单元进行了主题提取。他们认为一个词的含义与其上下文中共现的词密切相关，也即语言环境有关，如果两个词的共现词或语言环境非常相似，可以认为这两个词彼此相似②③。袁里驰④在研究中使用互信息来衡量词的语言环境是否相似，定义词的相似度，进而根据相似度进行词的聚类分析，Ayad等⑤将这种方法称为基于语境相似性计算的词聚类。在此基础上，陈炯等⑥提出了一种基于词聚类的汉语文本主题抽取方法，其操作方法是在文本集中选取若干种子词，并提取与种子词具有强共现关

① 胡珀，何婷婷．基于自适应聚类的文本潜在主题的自动发现［J］．郑州大学学报：理学版，2007，39（2）：92-95.

② Finch S. P.，Cbater N.．*Bootstrapping Syntactic Categories*［C］. Proceedings of the 14th Annual Conference of the Cognitive Science Society of America Bloomington，IN，1992：820-825.

③ Dagan Ido，Shaul Marcus，Shaul Markovitch. Contextual Word Similarity and Estimation from Sparse Data［J］. *In Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*，*Association for Computational Linguistics*，1993：164-171.

④ 袁里驰．一种基于互信息的词聚类算法［J］．系统工程，2008，26（5）：120-122.

⑤ Ayad H.，Kamel M.．*Topic Discovery from Text Using Aggregation of Different Clustering Methods*［C］. In Advances in Artificial Intelligence：15th Conference of the Canadian Society for Computational Studies of Intelligence，Canada：Springer，2002：161.

⑥ 陈炯，张永奎．一种基于词聚类的中文文本主题抽取方法［J］．计算机应用，2005，25（4）：754-756.