安璐　李纲◎著

科研文化机构的主题特征可视化挖掘

中国社会科学出版社

安璐　李纲◎著

科研文化机构的主题特征可视化挖掘

中国社会科学出版社

图书在版编目(CIP)数据

科研文化机构的主题特征可视化挖掘/安璐，李纲著．—北京：中国社会科学出版社，2016.7

ISBN 978-7-5161-6989-6

Ⅰ.①科… Ⅱ.①安… ②李… Ⅲ.①图书馆工作—信息化—研究 Ⅳ.①G250.7

中国版本图书馆 CIP 数据核字(2015)第 251175 号

出 版 人 赵剑英
责任编辑 喻 苗
特约编辑 王福仓
责任校对 胡新芳
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

印刷装订 三河市君旺印务有限公司
版 次 2016 年 7 月第 1 版
印 次 2016 年 7 月第 1 次印刷

开 本 710×1000 1/16
印 张 18
字 数 262 千字
定 价 68.00 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

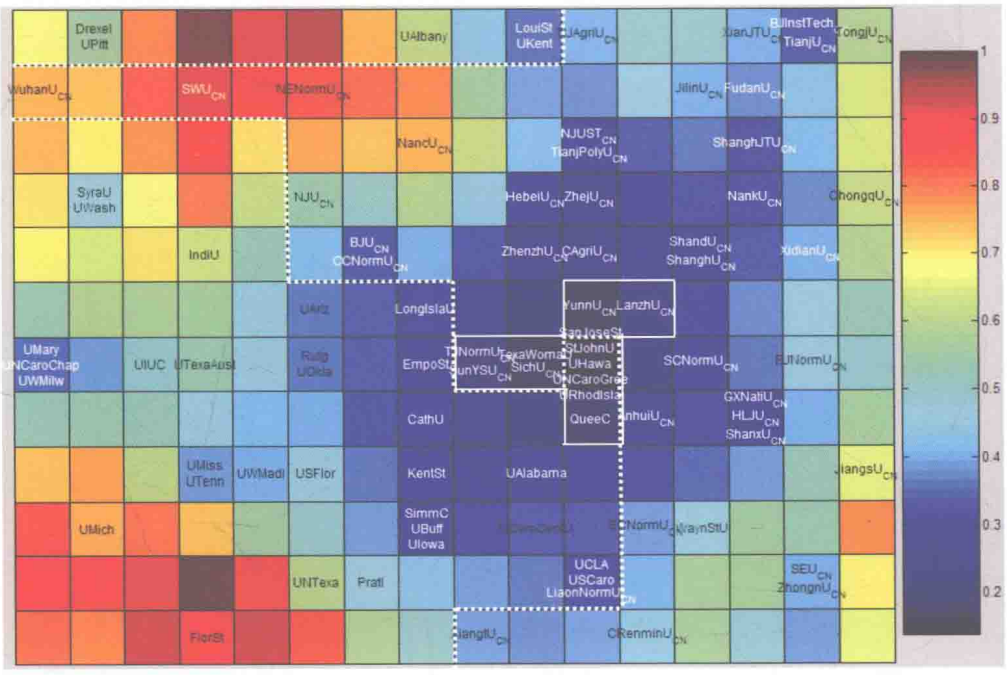

图 4—3　中美图情科研机构的 SOM 输出

注：图中的白色虚线为中美图情机构所映射位置在 SOM 输出中的边界。

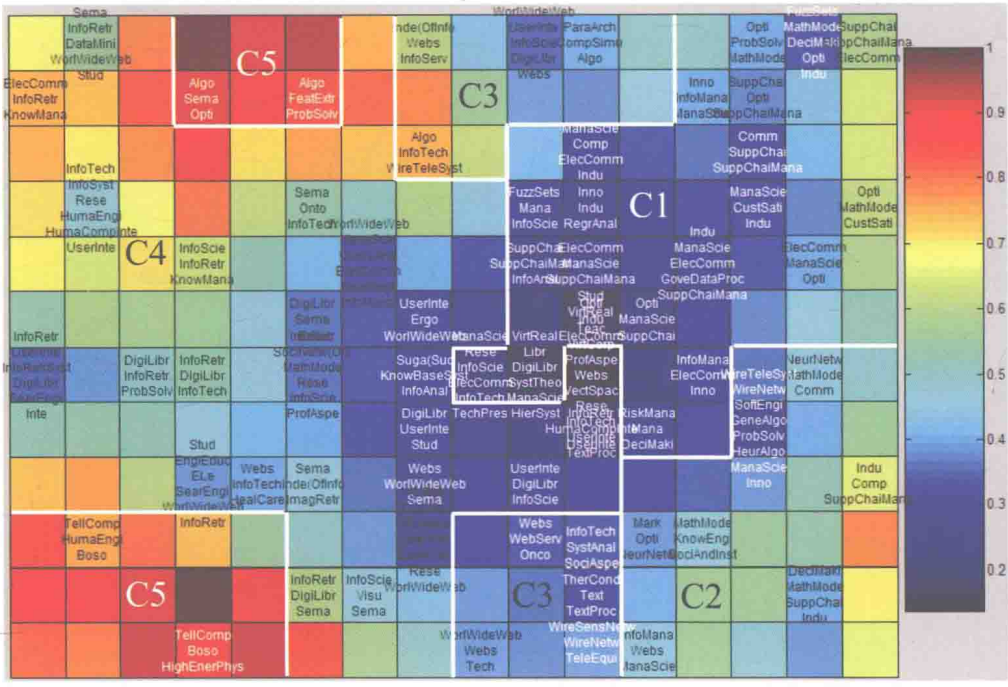

图 4—4　标注最频繁受控术语的 SOM 输出

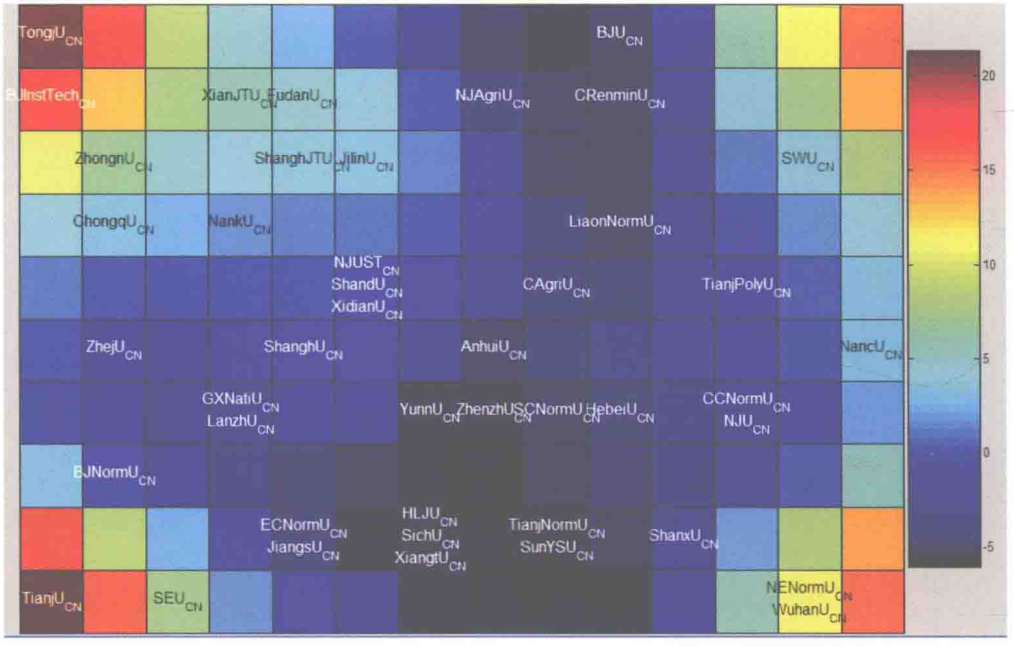

图 4—8　中国图情机构的前十个热点研究领域的综合成分图

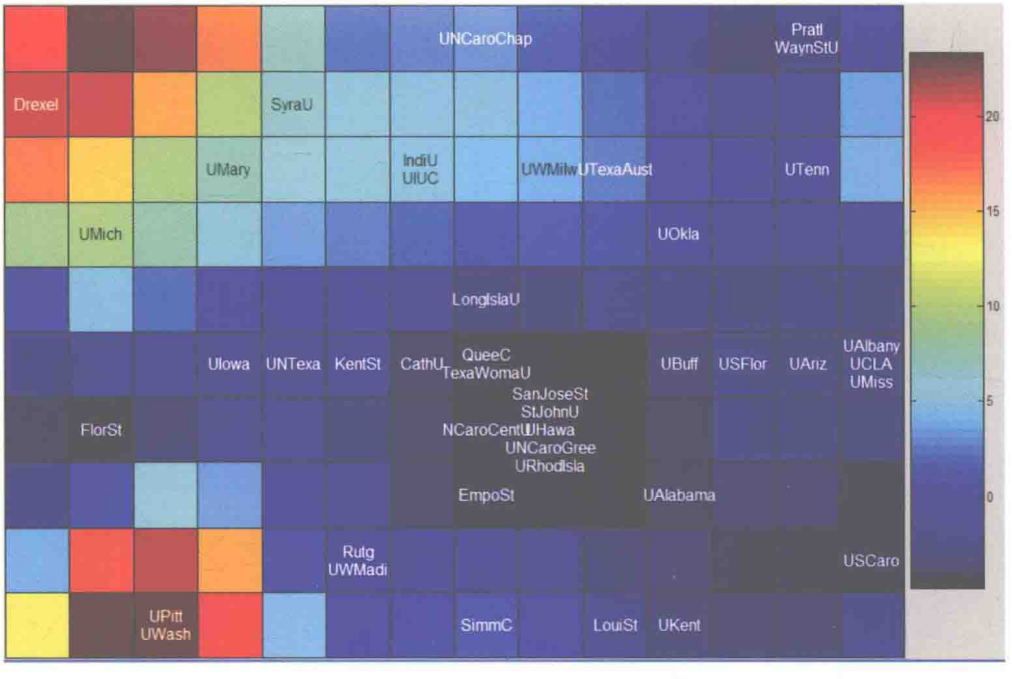

图 4—9　美国图情机构的前十个热点研究领域的综合成分图

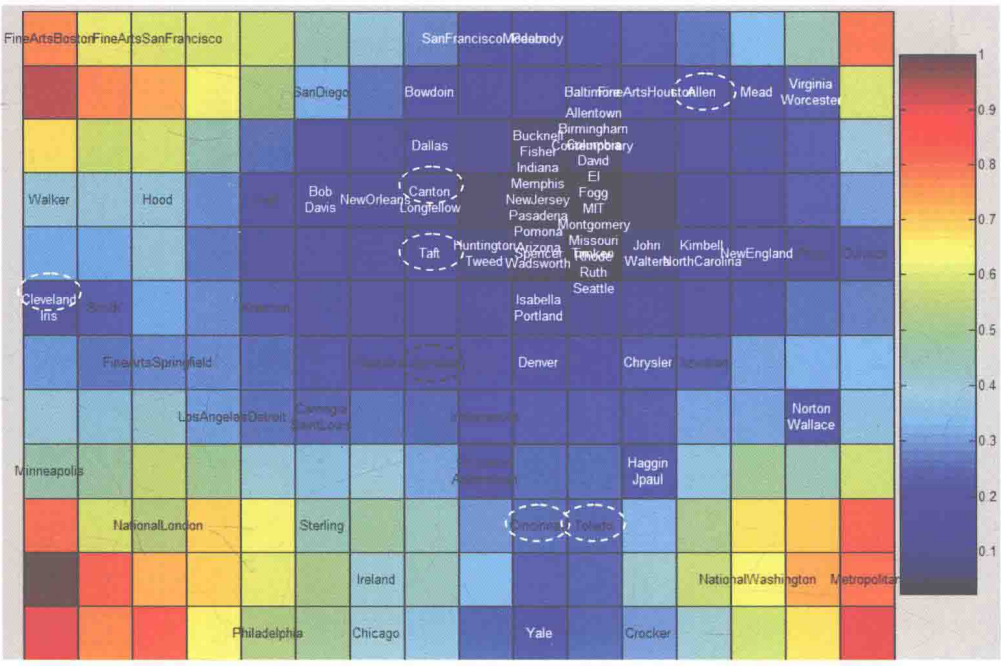

图 8—8（a） 博物馆涉及的艺术家 SOM 输出（博物馆缩写为标签）

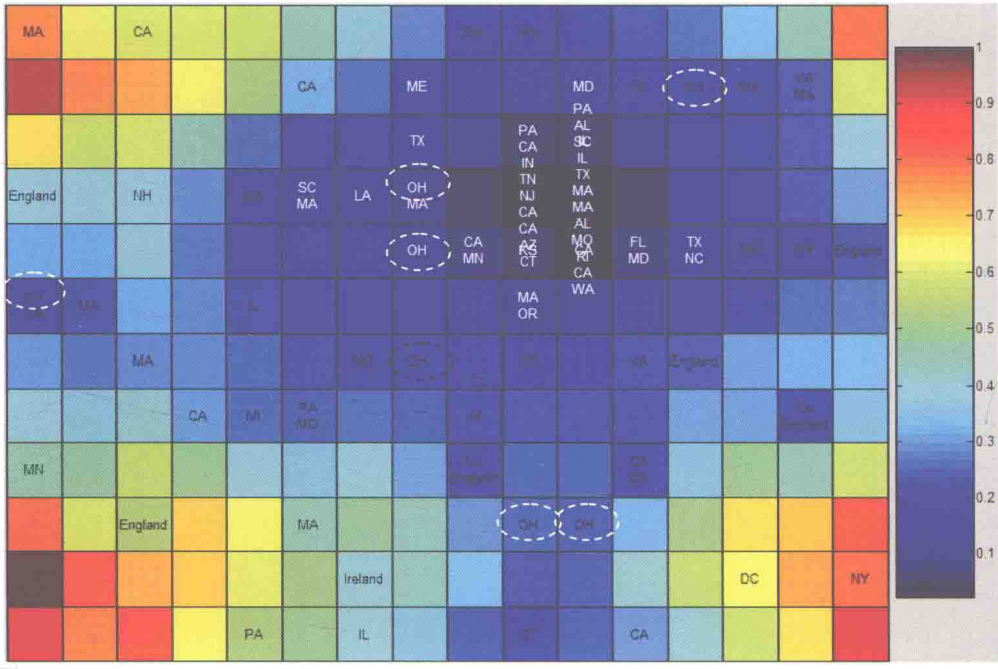

图 8—8（b） 博物馆涉及的艺术家 SOM 输出（国家或州缩写为标签）

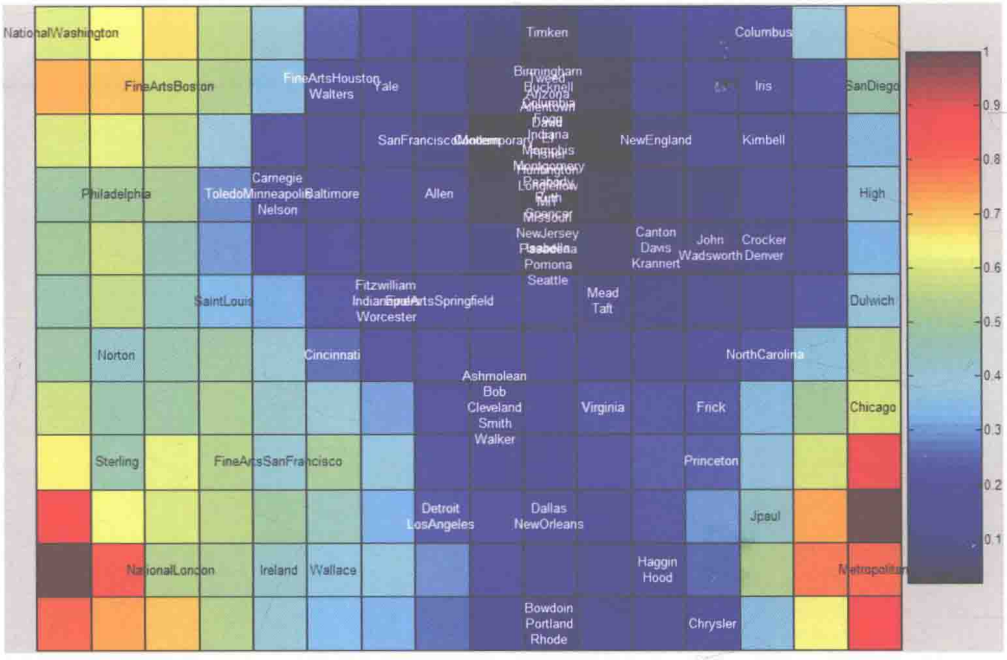

图 8—10（a） 博物馆的主题相似性 SOM 输出（以博物馆缩写为标签）

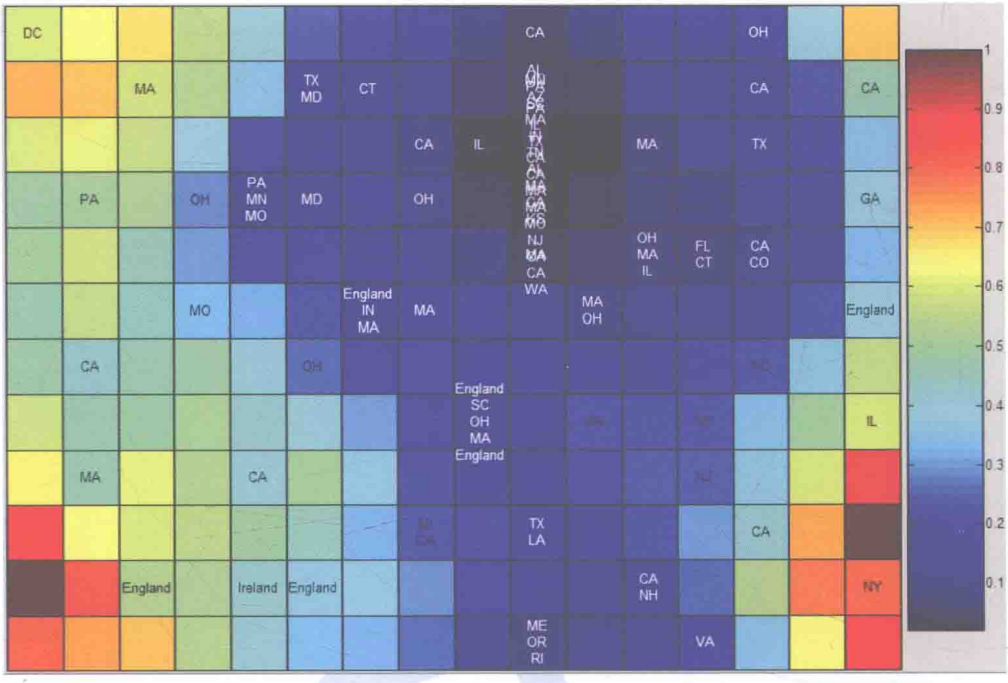

图 8—10（b） 博物馆的主题相似性 SOM 输出（以州或国家缩写为标签）

序

这是一本关于科研文化机构的可视化主题挖掘的书。它为探索机构之间的智力与主题联系提供了全面的方法、课题、实验与案例研究。本书的主要作者安璐博士在完成此书时正在我院做访问学者。我有幸阅读了本书的早期手稿，并与她讨论了其中的方法与实验。毫无疑问，这是一本及时的佳作。

20 多年来，我作为信息可视化与知识图谱的研究者，阅读并评审了无数关于文本挖掘与信息可视化的项目与应用。应用文献计量学与可视化方法和技巧来揭示学术文献、作者、引文网络、社会网络、网站、研究团队甚至国家的主题特征的例子有很多。本书所涵盖的内容有颇多新颖之处。在本书中，一个国家的科研机构并不是作为一个整体来看待。相反，从比较两国的研究主题，到对于新兴或热点研究主题的研究，再到研究领域的演化，作者突出了每所机构的作用，详细分析其主题特征。读者就好像配备了一部显微镜，能够更好地理解每所被调查的机构在每个主题上的现状与趋势。

本书最突出的特点在于，作者在书中所描述的众多研究课题中如何选择分析单元。作者持续探索了科研文化机构及其所涉及的主题术语之间的联系。机构与主题术语均为分析单元。两者都可以作为信息网络中的节点或联系。众所周知，新的分析单元的选择与处理通常会带来新领域的研究。文献计量学始于引用模式作为分析的焦点。当焦点从文献引文转移到作者共引时，新的作者智力关系的分析就诞生了。当焦点转移到网络时，我们就有了网络信息计量学，

转移到科学技术时，就有了科学计量学。虽然所有这些领域具有相似的研究方法与工具，但是它们的产生与应用有很大的区别，每一个领域都提供了一种理解复杂信息网络的新方式。本书的焦点是科研文化机构。当揭示出机构之间的主题关系模式时，我们会从中学到什么？不同科研机构与领域的创新趋势或模式是什么？科研机构在对新兴研究主题的参与度上有何相似性与差异？本书回答了这些问题。

本书对于分析方法，尤其是利用自组织映射进行数据分析方面做出了显著的贡献。自从 1991 年，我首次将自组织映射引入信息检索以来，许多研究者在数据分析与信息检索的不同领域都应用了自组织映射方法。本书应用自组织映射方法来分析科研机构的研究主题，以及艺术博物馆所拥有的法国艺术品的艺术家与主题。作者应用一种有趣而新颖的自组织映射输出方式，称为综合成分图来识别对热点主题做出显著贡献的科研机构。自组织映射方法使用户能够观察对象（本书中的科研机构与博物馆）在属性（对应于研究主题、艺术家与主题）上的全面分布，而不仅仅是将对象聚类。

本书讨论的另一种新方法是如何区分研究主题的新颖性与热度，以及如何量化科研机构对新兴与热点研究主题的贡献度。虽然许多研究者提出探测或识别新兴和热点主题的解决方案，但是研究科研机构对新兴和热点主题的贡献度是一种崭新且有启发性的视角。作者提醒我们，科研机构对新兴和热点主题的贡献度能够比新兴和热点主题本身提供更加细致且有用的洞察力，即使不投入更多的注意力，至少也值得投入同等的注意力。

在实践方面，本书对于希望提升其信息管理实践的文化机构尤其有帮助。本书介绍了一种全面的方法论，可视化地分析艺术博物馆的藏品数量、主题和艺术家的构成。作者探索了欧洲和北美的 90 多所博物馆的 7000 多件法国艺术家的作品，构建了基于博物馆的艺术家或主题结构向游客推荐博物馆的机制。

整体上，本书既包括深刻的理论探索，也拥有丰富的实验结果。所提出的方法对于分析和可视化地显示科研文化机构的主题特征

是可靠且有用的。我向可视化主题分析、科研文化机构管理、信息可视化、知识图谱与知识发现领域的学生与研究者推荐此书。

林夏博士
信息学教授
德雷塞尔大学计算与信息学学院
美国宾夕法尼亚州费城
2015 年 5 月

安璐译

前　言

科研文化机构是国家科技创新与文化交流的主体。随着科学技术的发展与人类知识的积累，科研文化机构所产出的学术文献及收藏的知识存储日益丰富。为了更好地掌握科研文化机构的科学产出与文化收藏的主题特征，我于2011年申报了国家社会科学基金项目“科研组织的研究领域可视化挖掘研究”（项目号：11CTQ025），并和李纲老师、余传明老师等人一起开展了相关研究。2014年我到美国德雷塞尔大学计算与信息学学院做访问学者，有幸与著名的信息可视化与知识图谱专家林夏教授一起共同研究科研文化机构的主题特征可视化挖掘。

本书共分九个章节。第一章导论部分概述了国内外关于科研文化机构的主题特征可视化挖掘的相关研究与不足，提出了本书的研究内容。第二章论述了科研文化机构的主题特征可视化挖掘的理论基础，包括科研文化机构的主题识别与分析、新兴主题与热点主题的识别与分析、科研文化机构的主题演化以及信息可视化的理论基础。第三章论述了科研文化机构主题特征可视化挖掘的方法论，包括潜在狄利克雷分配模型、聚类分析、非相关文献的知识发现以及社会网络分析等方法如何应用于科研文化机构主题特征的可视化挖掘，以及CiteSpace，VOSviewer，Sci^2，Treemap和主题图等工具与科研文化机构的主题特征可视化挖掘之间的关系与应用。

第四章是科研机构研究领域的可视化比较研究。以中美图情科研机构为例，以EI数据库及其为文献分配的受控术语为数据来源与主题线索，利用自组织映射（SOM）方法对两国图情科研机构的研究领域进行可视化比较，识别出蜂群机构、枢纽机构与里程碑机

构，根据科研机构的主题特征，确定国内外潜在合作机构，利用SOM和多维标度（MDS）对被调查的科研机构进行主题聚类，对两国图情科研机构的热点与特色主题进行比较，并利用自定义的综合成分图（CCP）揭示了对热点主题做出主要贡献的科研机构。

第五章是科研机构对新兴主题的贡献度可视化分析，建立了一种区分主题的新兴程度并进行加权的方法，构建了科研机构对新兴主题的贡献度计算并可视化显示的方法，识别中美图情科研机构的新兴主题，计算并利用树图可视化地显示各科研机构对新兴主题的贡献度大小。

第六章是科研机构对热点主题的贡献度可视化分析，建立了一种区分主题的热点程度并进行加权的方法，构建了科研机构对热点主题的贡献度计算并可视化显示的方法，识别中美图情科研机构的热点主题，分别利用SPSS工具和MDS方法对热点主题进行层次聚类与多维尺度分析，计算并利用树图可视化地显示各科研机构对热点主题的贡献度大小，并结合第五章的分析结果，将被调查的科研机构按照新兴与热点程度划分为四个类别。

第七章是科研机构研究领域的演化研究，以中国知网为数据来源，将最近15年的数据划分为三个时间段，利用树图方法可视化地分析了不同时期国内图情科研机构的论文数量、热点主题以及各科研机构的研究侧重点的变化。

第八章是博物馆藏品特征的可视化挖掘。以Getty研究所的法国艺术家博物馆数据库为数据来源，利用树图工具、三维投影等方法可视化地提示了被调查的博物馆的藏品数量、藏品主题数量、构成与特征、艺术家的数量、专长与作品材质、尺寸特征，利用自组织映射方法对博物馆藏品的艺术家构成与主题特征进行了可视化分析，分别建立了面向艺术家偏好与主题偏好的博物馆游览推荐机制。

第九章是总结与展望，归纳了科研文化机构主题可视化挖掘的研究意义与用途，总结所建立的科研文化机构主题可视化挖掘的方法体系与应用，指出研究的不足之处以及未来的研究方向。

本书由李纲老师制定内容框架并审定全稿，我负责全书的大部分内容撰写与修改。李纲老师、余传明老师和杨书会参加了第四章

关于科研机构研究领域的可视化比较研究，林夏老师、董丽、潘青玲、张馨文参加了第五章关于科研机构对新兴主题的贡献度可视化分析，秦佳佳参加了第六章关于科研机构对热点主题的贡献度可视化分析，郭雨橙参加了第七章关于科研机构研究领域的演化研究。

本书系国家社会科学基金项目“科研组织的研究领域可视化挖掘研究”和国家自然科学基金项目“大数据环境下基于领域知识获取与对齐的观点检索研究”（项目号：71373286）的研究成果之一。感谢美国德雷塞尔大学 Philly Codefest 2015 委员会为本书的研究提供 Getty 研究所的法国艺术家数据库的数据。

安璐

2015 年 3 月于美国费城

Foreword

This book is about visual topical mining of research and cultural institutions. It provides a comprehensive coverage of methods, projects, experiments, and case studies on the exploration of intellectual and topical connections between and among institutions. The main author of the book, Dr. Lu An, was a visiting scholar in my institution while she was completing this book. I was fortunate to read early manuscripts of the book and discussed with her about the methods and projects described in the book. I have no doubt that this is a timely and excellent piece of work.

As an information visualization and knowledge mapping researcher for more than twenty years, I have read and reviewed numerous projects and applications of text mining and information visualization. There are many examples of employing bibliometric and visualization methods and techniques to reveal the thematic features of academic litertures, authors, citation networks, social networks, Web sites, research groups and even nations. What is covered in the book is something new and novel. In this book, the research institutions of a nation were not considered as a whole. Instead, the role of each institution was highlighted and their thematic features were analyzed in detail, from the comparison of research topics between two nations, to the studies on emerging or salient research topics, and to the evolvement of research fields. Readers will get a better understanding of the currency and trend of each investigated research institution on each topical theme as if they were equipped with a microscope.

The most distinguished feature of this book is how the authors chose

the unit of analysis for many of the research projects described in the books. The authors consistently explored the connections between research and cultural institutions and subject topic terms that they involved. Both the institutions and the topic terms are the unit of analysis. Both can be treated as nodes or links in the information network. As we know, selection and treatment of new unti of analysis will often lead to new fields of studies. Bibliometrics started when citation patterns are the focus of analysis. When the focus was moved from document citations to author co-citations, a new field of author intellectual relationship analysis was born. Moving to the Web, we have webometrics, and to sciecce and technology, we have scientometrics. While all these fields share similar research methods and tools, their achievements and applications are very different, each providing a new way of understanding complex information networks. In this book, the focus is research and cultural institutions. What will we learn when patterns of topical relationships among institutions are revealed? What are the trends or patterns of innovation in different research institutions and different domains? What are the similarities and differences among research institutions in term of their engagement to new research topics? This book has answers to these questions.

The book also makes significant contributions to the analysis methods, in particular the use of Self-Organizing Map for data analysis. Since I introduced the Self-Organizing Map technique to the field of information retrieval in 1991, a lot of researchers employed the SOM technique in many different aspects of data analysis and information retrieval. In this book, the SOM technique was employed to analyze the research topics of research institutions and the artists and subjects of French art works held by art museums. An interesting novel SOM display named Compound Component Plane was also applied to identify the research institutions that made significant contributions to the salient subjects. The SOM technique enables users to observe the comprehensive distribution of objects (research institutions and museums in this book) among attributes (research topics, art-

ists and subjects correspondingly) instead of merely clustering objects.

Another new approach discussed in the book is how to differentiate the novelty and salience of research topics and how to quantify the contributions of research institutions to emerging and salient research topics. While many researchers propose solutions to detect or identify emerging and salient research topics, it is a new provoking perspective to study the contributions of research institutions to emerging and salient research topics. The authors remind us that the contributions of research institutions to emerging and salient research topics provide more fine-grained and useful insights than the emerging and salient research topics themselves and are at least worthy of equivalent if not more attention.

On the practical side, the book is particularly useful for cultural institutions which want to improve their practice of information management. The book introduces a comprehensive methodology to visually analyze the quantities, subjects and artists of the collections held by art museums. More than seven thousand art works by French artists in over ninety notable musenms in Europe and North America were explored, and mechanisms of recommending art museums to the tourists were constructed either based on the artist or subject structure of the museums.

On the whole, this book contains both insightful theoretical exploration and rich experimental results. The proposed methods are sound and useful for analysis and visualization of the topical features of research and cultural institutions. I would recommend this book to students and researchers in the fields of visual topical analysis, research and cultural institution management, information visualization, knowledge mapping, and knowledge discovery.

Xia Lin, Ph. D.
Professor of Informatics
College of Computing and Informatics
Drexel University
Philadelphia, Pennsylvania, USA
May, 2015

目　录