

科技文本的可视化分析研究

任智军, 朱东华, 谢菲

北京理工大学 管理与经济学院, 北京 100081

摘要 文章研究了科技文献中可视化分析技术,探讨了基于数据挖掘和可视化技术的科技文献数据分析方法,实现了从数据获取、数据挖掘、可视化分析为一体的工作流程,其可视化分析的成果可以使科技管理决策部门和广大的科研工作者迅速了解科技领域的现状和科学发展的规律,大力推动了我国科研工作进展。文章还对我国管理科学领域进行了实证研究分析。

关键词: 可视化分析; 数据挖掘; 科技文本; 管理科学

中图分类号: G202 ;C931

文献标识码: A

文章编号: 1009-3370 (2007)01-0012-06

一、引言

随着信息社会的迅速发展,科技文献管理与分析利用的发展处在一个比较尴尬的境地之中,一方面,文献的数据量随着时间的推移高速增长,越来越多的科技文献给我们的研究工作带来巨大的便利。另一方面,文献数据量的增长,却给文献的管理与分析利用带来很多的不便,理解这些海量数据所表明的含义变得非常困难,面对浩如烟海的电子资源无法准确、快速地得到需要的资源,更无法通过目前的分析方法迅速地了解某个领域的发展状况。

可视化分析是利用计算机图形学和图像处理技术,将数据转换成图形或图像在屏幕上显示出来,并进行交互处理的理论、方法和技术。可视化分析目的就是很好利用人类的这种对于可视化形式下模型和结构的获取能力,并且对这个结果进行分析解释。利用网络获取工具、获取数据,使用数据挖掘和可视化技术相结合的方法,解决科技文本数据量过大,无法进行快速和有效的交流的问题,同时可视化分析观察、筛选、发现和理解信息,发现隐藏在数据和信息背后的含义。可以说,科技文本可视化分析的研究和发展从根本上改变了我们表达和理解复杂数据信息的方式。

二、可视化分析的研究方法

快速了解特定的科技文献的发展现状,不但对我们高效地获取有价值的信息有重大意义,而且能

够使我们快速地了解国内外科技发展的前沿问题,同时它也可以帮助我们了解科学发展的规律,进而可以帮助我们把握科技前进的脉搏。这将成为不仅是科技管理决策部门而且是各个领域的研究者甚至企业的一种基本需求。

研究使用可视化分析的技术手段辅助文献管理面临的主要问题有:

(1)如何获得海量的科技文献信息,并且可以把集成在本地数据库信息进行合理的管理,并提供方便的用户接口进行操作;

(2)使用何种信息可视化的形式,通过科技文献的引文信息、作者信息、摘要信息、全文内容^[1-5]等反映出潜在的知识,发现事物之间的联系,真正起到信息可视化发现问题、说明问题的特征;

(3)使用何种理论和算法,保证在海量的信息中获得有价值的、具有代表性的内容,进而为信息可视化做数据准备,为发现科学发展规律提供保证。

可视化分析使用数据获取、数据挖掘、可视化分析等过程解决了以上的问题。由于科技文献信息散布在世界各大文献数据库中,如:ISI、EI、中国期刊网等,因此科技文献信息的获取对很多科研工作者存在一定的问题。针对网络中科技文献信息量巨大,数据分布在各地难以使用的问题,结合搜索引擎技术,获取数据后,然后采用数据预处理技术进行处理,改变可用于可视化分析的格式,然后在数据中使用可视化模型开展研究,发现有价值的信息。分析过程如图1所示。

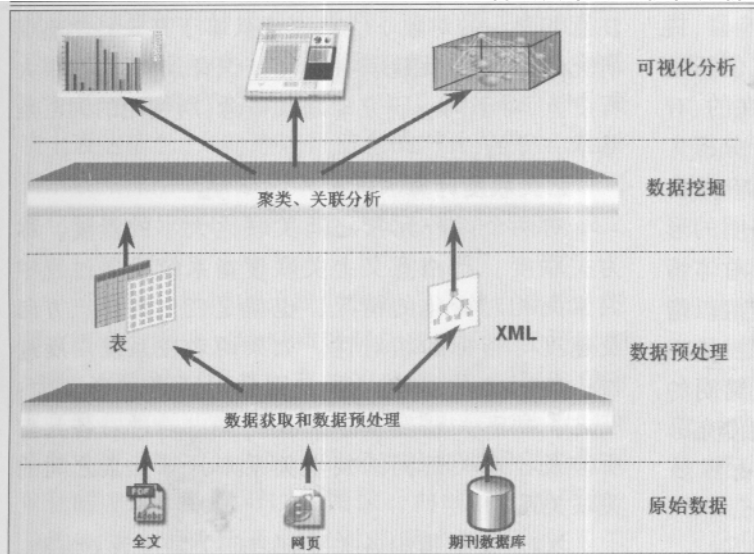


图1 可视化分析过程图

三、数据的获取与预处理

1. 科技数据获取

数据获取是可视化分析科技文本整个工作中重要的一步。动态、快速、便捷地获得数据，并把它存储起来是这一步的主要工作。首先我们把被获取源分为以下几个步骤：①利用搜索引擎搜寻。由于搜索引擎大都采用数据库来记录网站信息和网页信息，我们使用已经编制的科技搜索词典作为搜索关键词，在搜索引擎中键入查询对象名称。②采用人工查找的方法。网站设计了不让搜索引擎发现的安全措施，我们必须使用人工查找的方法发现数据源。获取后，集成在数据库中，整个过程如图2所示。

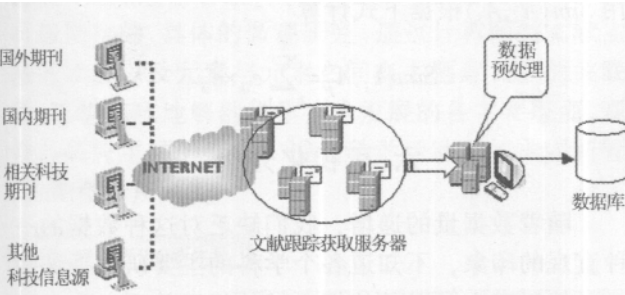


图2 文献数据跟踪获取功能图

2. 科技数据预处理

快速准确地识别并清除网页内的噪音内容是提高网页内容存储程序处理结果准确性的一项关键技术。有主题网页的提取方法是通过如下过程实现的，对于得到的科技信息网页中的HTML最常用的结构表示方法是构造网页的标签树。现有的标签树

构造DOM(Document Object Model)是一个常用的标签树构造工具,它可以将网页中的标签按照嵌套关系整理成一棵树状结构。整个网页内容提取过程为：①建DOM树；②设定网页提取规则；③依照设定好的格式，存储相关数据库根据格式化网页的格式特点，依照设定好的规则格式，存储相关的数据数据库。

具体如图3所示，这是一个格式化数据库摘要的数据格式，分析后发现其中带虚线的地方是我们应当存储在数据库中的，所以这个规则应当包含虚线包括的所有部分。

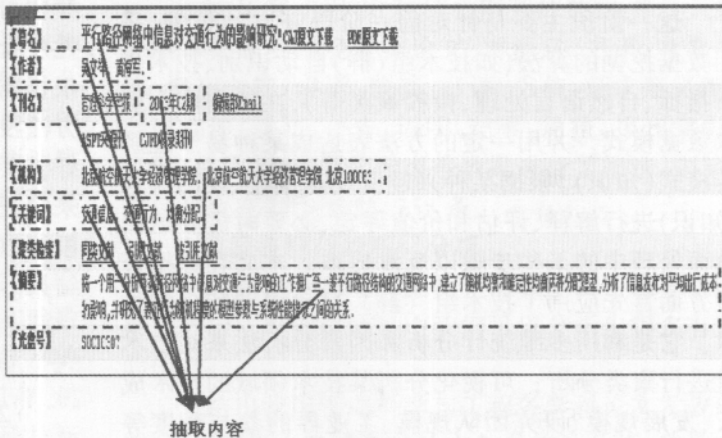


图3 可视化分析数据抽取图

四、可视化分析

可视数据挖掘是指用可视化技术从大的数据集中发现隐含的和有用的知识。数据挖掘离不开可视化技术，目前大多数数据挖掘应用中都结合了一些可视化方法,但这种结合是松散的,主要表现在可视化技术主要是用于数据挖掘结果的可视化，即将挖掘后得到的知识以数据分布图、曲线(面)、三维立方体、散列图、决策树等形式表示出来^[6],但是用户在其中并未涉及数据挖掘的过程，这样容易对最后结果的可信度产生怀疑。可视数据挖掘旨在将数据挖掘的过程可视化，用户通过可视的交互手段调用后台的数据挖掘算法，由用户所感兴趣的数据得到相应的数据分析,而不仅仅是数据查询。可视数据挖掘实现了数据可视化^[7]和数据挖掘之间的一种更紧密的结合。

可视化数据挖掘是可视化技术与数据挖掘技术相结合的产物。可视化数据挖掘是KDD(Knowledge

Discovery in Databases) 过程中的一个特定步骤,提供了用户与计算机之间的一个通讯接口,以便帮助用户从数据库或数据仓库中发现未知的、潜在的、有使用价值的信息的方法、理论和技术^[8]。对于分析人员来说,将数据挖掘整个过程进行可视化具有非常高的价值和重要性。将抽象的信息以一种简明的形式呈现出来,从而给用户一个关于分析结果总体情况的概念。这样,也可以给分析人员更多的自信以指导下一步的工作。通过将数据挖掘过程用可视化方式呈现出来,从而帮助分析人员以一种形象、简明的方式掌握知识萃取和决策分析的过程,并让你充分地融入其中。

五、数据挖掘

这一阶段主要是针对前一阶段获取的数据,采用数据挖掘的算法(如技术组群)自动识别、技术创新指标、自然语言处理、模糊聚类等),在数据库中提取数据模式,并用一定的方法表达成某种易于理解的模式(知识)。根据某种兴趣度度量,对发现的模式(知识)进行解释、评估和价值评定,必要时需要返回前面处理中的某些步骤以反复提取。本文在数据分析方面首先应用了技术组群)自动识别和分类技术,它是利用多维统计分析和因素分析法来对技术组进行聚类分析;可视化分析某技术领域技术成长、发展规模、研究团队规模、工业界的参与程度等情况,研究主题的变迁为科学活动和技术创新管理提供支持。

1. 共生计算

首先,在本文研究中,我们通过对技术关键词的共生关系(Terms Co-occurrences)计算来识别、确定一组文献内部所包含的技术组群)。假定我们有 n 篇文献,这 n 篇文献包含有 m 个技术关键词,则我们就建立了 $\{n \text{ 篇文献} \times m \text{ 技术关键词}\}$ 的关联矩阵 X , X 矩阵中,文献 D_i 的关键词 $Term_j$ 的权值,用布尔代数值表示,当 $Term_j$ 在 D_i 文献中出现时取 1,否则取 0, n 是文献组内包含文献的总数, m 是文献组内所有关键词总数。基于这个 $\{\text{文献} \times \text{关键词}\}$ X 矩阵,我们进一步得到 $\{\text{关键词} \times \text{关键词}\}$ 共生的关联矩阵 T

$$T = X^T \cdot X$$

技术共生关系分析是通过反映文献主题内容的词进行关联性或相异性定量分析来研究文献内在联系和科学结构的一种方法,其基本出发点是:①科学研究的热点是由一系列在内容上密切相关的研究课题和概念组成的,这些热点是众多科学研究人员

注意和跟踪的对象。②热衷或从事于某一科学热点研究的科学家,无论其社会 and 知识背景如何,在很大程度上,对于同一研究课题和概念,所使用的词汇是基本一样的。

2. 关联度计算

对两个成分元素之间关联性大小的量度,称为关联度。严格意义上关联度指系统发展过程中因素间相对变化的情况,也就是变化大小、方向及速度等指标的相对性。如果两者在系统发展过程中相对变化基本一致,则认为两者关联度大;反之,两者关联度就小。我们提出通过对各元素所涉及技术关键词的共生关系来计算元素之间的关联关系。

①文献组内各关键词之间关联度计算方法。我们可以使用一般聚类分析方法中常用的一些关联度计算方法,如关联系数、距离系数等;也可以根据需要,自己定义关联性测度方法。关联性测度方法没有严格的标准。具体问题中哪种测度方法最好,要用聚类和关联分析的结果是否符合实际情况来验证。文献组中 $Term_i$ 和 $Term_j$ 之间的关联度计算根据以下公式,其中 T_{ik} 是关键词共生计算归一化的结果。

$$\text{Sim}(Term_i, Term_j) = \sum_{k=1}^m t_{ik} x_{jk}$$

②文献组中其他主体之间的关联度计算根据以下公式。假定文献组中包含有 n 个学科或地区,这 n 个学科或地区包含有 m 个文档,则我们就建立了 $\{n \text{ 个研究机构} \times m \text{ 科技文本}\}$ 的关联矩阵 E 。在 E 中, n 是学科或地区的总数, m 是文献组内所有文献的总数。文献组中学科和地区 A_i 和 A_j 之间的引用关系度用 $\text{Sim}(A_i, A_j)$ 根据下式计算

$$\text{Sim}(A_i, A_j) = \sum_{k=1}^m e_{ik} e_{jk}$$

六、可视化分析

随着数据量的递增,我们缺乏对这种数据的一种直观的印象,不知道各个学科的主要研究领域和主要的研究人员以及他们之间的相互关系;我们不能快速地找到有价值的、权威的文章。因此我们有必要对科技文献管理的手段进行进一步的分析,以便更好地把握科学的发展规律。国外对科技文献管理手段的研究从未停止过,通过对这个领域的调研,发现比较有影响力的几种手段都是在一定的程度上结合应用了数据挖掘、知识发现和信息可视化等技术^[9-12]。因此结合数据挖掘、知识发现和信息可

视化等信息前沿技术是辅助科技文献管理的有效手段,尤其是信息可视化技术具有使海量数据变得简洁有条理、多角度的展示信息,可以在不同的细节层次上表达信息,方便地进行信息的对比与交互,能够揭示数据深层的联系和规律的优势。因此把信息可视化的技术与科技文献管理的结合是将来的解决之道^[13]。

可视化分析可以让我们观察、操纵、查找、漫游、探索、筛选、发现和理解信息,而且还可以和大量的信息进行快速、有效的交流,以便于发现隐藏在数据和信息背后的含义。因此可视化分析的研究和发展从根本上改变了我们表达和理解复杂数据信息的方式。数据可以用图像、曲线、二维图形、三维体和动画来显示,并可对其模式和相互关系进行可视化分析。可视化为我们发现规律、辅助决策、解释现象提供了有力的工具。在信息可视化的实际应用中,它的具体作用总结可概括为如下几点:使海量数据变得简洁有条理,多角度的展示信息,可以在不同的细节层次上表达信息,方便地进行信息的对比与交互,能够揭示数据深层的联系和规律^[2]。

科学领域的任何一项新技术都不是孤立的,它可能包含有基础研究部分、应用研究部分、实验数据,或者是继承技术、派生技术和横向关联技术等。监测和分析文献组内的技术组群、作者组群、机构组群、主题变迁组群的关联关系,定量地对科技活动的主导因素、潜在因素、优势组(群)和时间序列变化进行分析,实质上应是科技活动系统分析、预测、决策的基础。

通过技术关键词共生关系分析方法对文献组内的技术组群、作者组群、机构组群、主题变迁组群进行深层次的、具体的微观研究。通过计算同类文献主题关联度以及元素与元素之间在主题意义上的关联性,能够细致地解剖科学技术发展的各主要层面,观察各研究主题的变迁,进行科学活动和技术创新管理(如图4所示)。

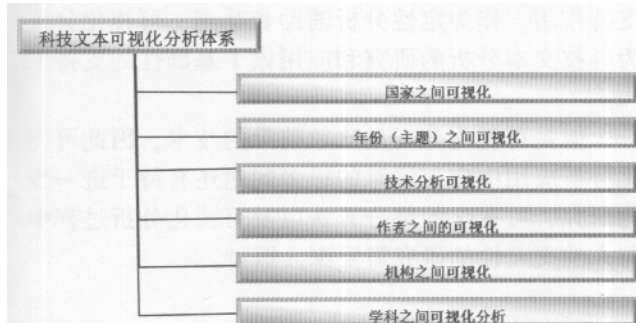


图4 科技文本可视化分析体系图

七、实证分析

管理科学是一门研究人类管理活动规律及其应用的科学,是一门横跨自然科学与社会科学两大领域的综合性交叉科学。

1. 数据获取

本研究采用网络数据自动采集系统,中文按照管理类77种期刊的中文期刊名,分别对每一期刊1994~2003年期间的全部科技论文的中文题目、中文关键词和中文摘要等信息进行了数据采集和预处理,共包含244 278条本地数据库记录。通过数据对检索到的文献的阅读与比较,对国内外研究的热点问题进行了综合比较分析。

2. 数据预处理

本次研究使用20种核心期刊的数据,并且只统计学术类和实践经验类论文,对消息、书讯、首次文摘、报刊文摘、讲话、访谈、书评、书信、征求意见稿、会议综述、通知、公告、说明、纪要、简报、报刊索引、目录、总目次、编辑部记事等数据进行了筛选、剔除,最后剩余有效数据25 755条。

3. 数据挖掘与可视化分析

使用数据挖掘中的聚类、关联分析得出分析结果,然后从可视化结果中我们可以分析出很多有用的信息。

总结出我国管理科学与工程各学科论文发表的情况,根据以上各学科的论文数量统计分析,我们对结果进行排序,得出国内管理科学与工程重点领域方向,排序结果见图5。

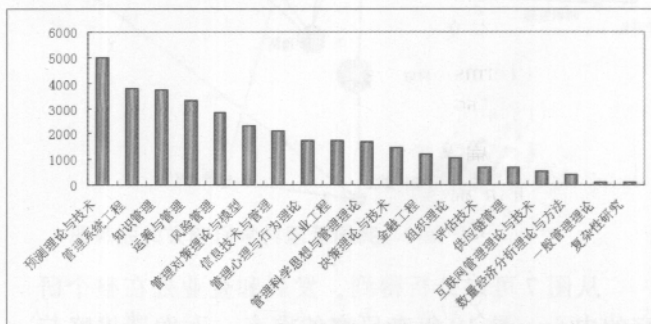


图5 我国管理科学与工程各学科论文统计排序

从图5中可以看出,历年论文数量相对较多的学科包括预测理论与技术、管理系统工程、知识管理、运筹与管理、风险管理等。说明这些学科领域可以被看作是研究的主流领域,也是近几年国内管理科学与工程研究的主要内容。

从图6中我们可以分析得到,在北京地区和管理科学有关的论文是最多的,而且就发表文章的类型、研究方向、研究进度的情况来看,北京、福建、广

东的研究是比较接近的,相对而言,黑龙江与大家的
研究不是很一致,而且数量也是比较少的。

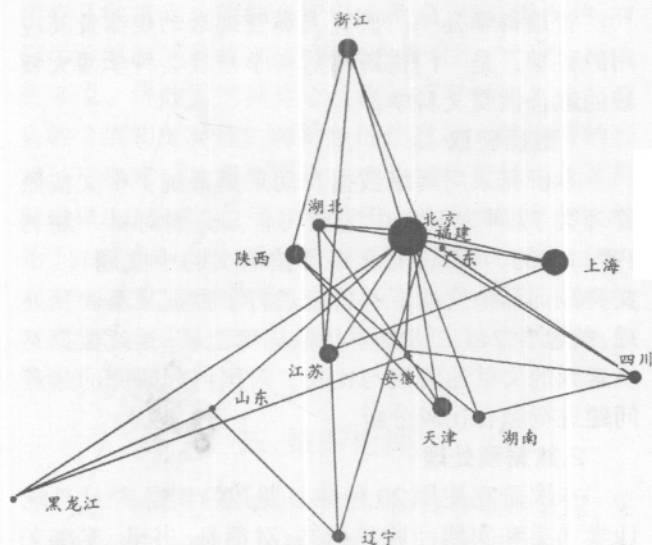


图6 我国管理科学地区关联图

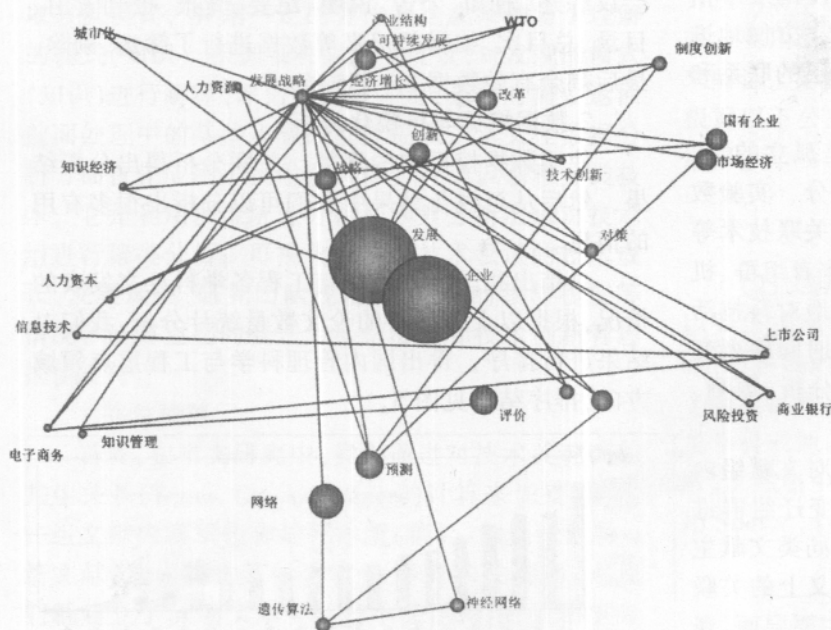


图7 我国管理科学研究方向关系图

从图7可以分析得到,发展和企业处在整个研究的中心,是10年来研究的重点,而发展战略与每一个词都紧密相关,说明其他与之相关的研究方向比较多。从中还可以看到,那些与其他管理连接不太密切,但是几个研究方向的研究有较强相关性的,如图7中电子商务与组织管理密切相关,说明研究组织管理与电子商务之间关系的文章非常多。

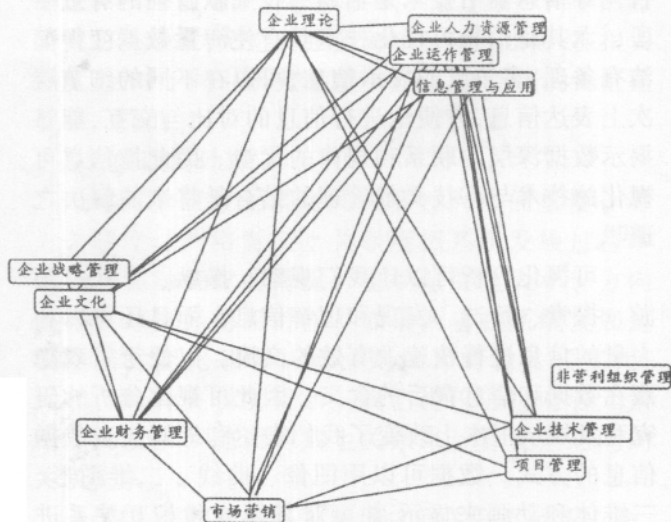


图8 我国管理科学工商管理学科关系图

从图8我们发现,在这些企业人力资源管理、企业运作管理、信息管理与应用等工商管理小学科的研究是比较集中的,而非盈利性组织的研究与大家的关联度都是0,所以本次研究得出结论,是否应当把非盈利性组织放在工商管理中是值得思考的。

八、结 论

本文研究了科技文本的可视化分析,指出可视化分析作为科技文本分析重要的一部分,利用网络获取工具获取数据,引用数据挖掘和数据可视化技术分析科技文本,建立了以关联度为主要指标的科技文本可视化分析的新方法。并对管理科学的相关性进行了实证研究,得出了在1994~2003年我国重点研究领域和管理领域各省相关度等有意义的结论。这些将对我国管理科学研究主体及资助主体决策活动有较强的支持作用。作为定性分析辅助和补充,可视化分析为科技文本分析的研究和应用做了基础性的支持工作。

但是可视化分析是比较前沿的技术,因此可视化分析应用中应该注意的很多问题还有待于进一步的研究,可视化的部分算法以及可视化分析过程中与人的交互性的研究需要深入探讨。

参考文献：

- [1] 骆茹敏, 吴俊卿, 倪星源. 科研管理[M]. 北京: 中国学术出版社, 1982.
- [2] 毕克新, 郭东亮, 姚秀荣. 科技信息与现代科研管理[J]. 科研管理, 1995(3):16-19.
- [3] 靖培栋, 康仲远. 关于科技文献增长的数学模型[J]. 情报学报, 2003(1):90-96.
- [4] 沈玉春, 江子章. 科技管理[M]. 北京: 科学技术文献出版社, 1993.
- [5] 刘凯. 信息可视化概念的深入探讨[J]. 情报杂志, 2004(12):20-21, 25.
- [6] George G Robertson, Jock D Mackinlay, Stuart K Card. Cone trees: animated 3D visualizations of hierarchical information[J]. Conference on Human Factors in Computing Systems, 1997: 189-194.
- [7] Mizoguchi F. Anomaly detection using visualization and machine learning[J]. Proceedings. IEEE 9th International Workshops on 14-16, June 2000:165-170.
- [8] Chaomei Chen. Measuring the movement of a research paradigm[J]. Visualization and Data Analysis(VDA 2005), San Jose, CA, 2005(5669):63-76.
- [9] <http://www.sims.berkeley.edu/courses/is247/s02/Lectures.html>
- [10] Nahum D Gershon, Stephen G Eick. Information visualization[J]. IEEE Computer Graphics and Applications, 1997(7-8):29-31.
- [11] Tamara Munzner. Information visualization[J]. IEEE Computer Graphics and Applications, 2002(1-2):4-6.
- [12] Daniel A Keim, Ming C Hao, Umesh Dayal, Meichun Hsu. Pixel bar charts: a visualization technique for very large multi-attributes data sets[M]. Information Visualization, 2001:1-4.
- [13] Kwan- Liu Ma. User interfaces for managing and exploring scientific visualization data[J]. IEEE Computer Graphics and Applications, 2000(9-10):16-19.

The Visualization Analysis of Scientific Text

REN Zhi-jun, ZHU Dong-hua, XIE Fei

(School of Management & Economics Beijing Institute of Technology, Beijing 100081, China)

Abstract: This paper discusses visualization analytical technique and the data analysis method of scientific literature based on data mining and data visualization. The workflow of Data Getting, Data Mining and Analysis of visualization was applied. The result of Visualization analytical technique can be used to make technical decision department and scientific researcher rapidly know the present situation of Science & Technology and the rule of Science development, promoting the scientific research in our country. An example in the field of Management Science was provided in the last part of the paper.

Key words: visualization analysis; data mining; scientific text; management science

[责任编辑 孟青]