摘要

专利情报作为重要的信息资源，如果加以合理利用就可以提供相关的技术、经济、市场和法律等方面信息，从而为企业决策提供依据。近年来，新技术的不断涌现、技术更新换代频率的加剧，以及对于知识产权问题的日益重视，都促使专利信息以前所未有的速度迅猛增长。然而与信息的快速增长所不和谐的是专利分析方法的发展并不尽人意。

本论文以SOM（Self-Organizing Feature Map）方法为基础开展信息可视化的相关技术研究，尝试使用信息可视化、数据挖掘等技术手段进行专利信息的挖掘，从而提高专利分析水平，进一步发觉潜在信息。

本文根据专利信息本身的特点，提出了基于SOM的专利信息可视化模型，同时结合专利信息本身特点对SOM方法进行了改进，使其可以更好地挖掘隐含信息，然后通过中国专利信息的集成电路封装技术领域数据进行验证；接着本文根据作者提出的可视化图形生成原则，设计了可视化图形生成算法，并采用等高线图谱将挖掘的信息展现出来。最后本文对集成电路封装技术领域开展实证研究，在实证中对本文提出的模型和可视化算法全部进行了验证和分析。

关键词 专利情报，SOM方法、可视化数据挖掘、可视化模型

Abstract

Patent information could provide related technological, economic, market and law information as an important resource if it could be explored rationally, and it would be used as a base of company decision-making. Recently, patent information blooms at a unprecedented rate in response of new technology emergence, speeding up of new technology development, and more attention on intellectual property. Compared to information's fast-developing, however, the development of patent analysis method is not satisfying.

This paper is based on SOM method to make a related research on information visualization. We also apply technology such as information visualization and data mining to fulfill the mining of patent information, improve the patent analysis level, and explore the potential information.

According to the straits of patent information, we propose a SOM-based visual model, making some modifications to SOM methods considering the features of patent information, which could explore more hidden information, then we prove the method validated by data in IC encapsulation area of China patent; following, this research based on principled of generating visual graphics, designed a visual graphics generation algorithm, and present the excavated information with a contour map. At the end, this paper made a practical research in IC packaging technology, in which the model and algorithm would be validated.

Keywords: patent information; SOM; visual data mining; the module of visualization;

目录

图目录

表目录

1.绪论

1.1 选题背景

（1）专利信息作为情报资源日益重要。

随着经济的全球化发展，我国企业将面临着来自国内外的竞争压力。知己知彼，百战不殆，企业只有明确自身的地位和作用才能获得竞争的主动权。在此情况下，许多大公司、企业都在不断调整自己的技术路线[1]，重点关注高新技术的专利信息作用。

专利情报作为重要的信息资源，如果加以合理利用就可以提供相关的技术信息、经济信息、市场信息和法律信息，作为企业决策的依据。

（2）专利信息增长迅速，分析方法体系并不完善。

随着新技术的不断涌现、技术更新换代频率的加剧，以及对于知识产权问题的日益重视，专利的申请数量每年都在递增。仅2006年申请专利数就为357899件，2007年则申请为50223件。如何利用好这些庞大的资源，挖掘出企业所需的信息一直备受关注。

然后目前专利数据库丰富的同时，分析手段和决策支持的功能表现不足，对现有信息的利用不足，不能很好的发掘出数据的内在联系。以往的方法主要是建立在统计分析基础上。在专利数据到信息，情报和知识转化不足；没有动态更新的技术识别能力；关键技术，竞争技术，辅助技术等技术识别方法欠缺。

（3）目前专利分析中所面临的主要问题

在日益激烈的竞争中，为了获取竞争的主动权，企业必须明确自身在竞争环境中所处的地位与状态。反映在专利信息领域，人们的期望就不仅仅不停留在对专利表面信息的挖掘上。为了制定更加合理的专利战略上，企业需要对其在某领域的专利地位，以及竞争对手的状况有详细的认识。但是目前的专利分析方法对专利品质分类方面的挖掘并不完善，并不能满足企业对其的需求。

因此本文旨在结合实验室目前的专利数据库现状，从可视化的特点出发，对专利信息进行挖掘、揭示和说明，弥补传统分析方面的缺陷。

（4）信息可视化技术的研究现状

可视化技术(Visualization)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[10]。人们获取

和处理信息的效果与信息的展现方式密切相关。有人做过试验，如果把大量的数据排列成易于辨认的图案，人们可以在瞬间理解数亿比特的信息，大大提高了人们的认知率。这也说明信息只有通过有效的展现才能发挥他的本来意义[9]。

国内外的学者在可视化技术应用于科技信息知识发现方面做了大量的研究，陈超美1999年在“A Semantic-Centric Approach to Information Visualization”文中提出一种语义为中心的信息可视化方法，通过对个人文献集合以语义为中心的方法集中揭示信息空间的内在联系。在“Integrating Spatial, Semantic, and Social Structures for knowledge Management”文中提出一种基于虚拟现实表达语义结构的知识管理系统。 2001年“Visualizing a Knowledge Domain's Intellectual Structure”提出并实现了通过科技文献抽取模式进行作者引文分析，利用计算机图形技术将分析结果绘制为3D知识图景。在与印第安纳大学的Katy Börner合写的“Visualizing Knowledge Domains”一文中对包括降维技术、聚类分析与空间配置等在内的可视化关键技术做了系统的阐述。俄克拉何马州立大学电子与计算机工程系Steven Morris在2002年“DIVA: a visualization system for exploring document databases for technology forecasting”文中提出一种Database Information Visualization and Analysis system (DIVA)数据库信息可视化与分析系统，针对科技预测对文献和专利信息进行二维的可视化，该系统的过程模型是：获取文献、绘制文献图、聚类分析、关系挖掘和生成总结与趋势表达。美国亚里桑那大学的黄赞等2003年在“Longitudinal Patent Analysis for Nanoscale Science and Engineering: Country, Institution and Technology Field”文中运用内容图谱分析和引文网络分析等技术对纳米领域的专利进行分析并进行可视化表达。法国Jean-Charles Lamirel等2003年在“Intelligent patent analysis through the use of a neural network: experiment of multi-viewpoint analysis with the MultiSOM model”一文中着重阐述了神经网络方法在科学技术信息图谱中的应用，并将基于神经网络的复合自组织工具应用于信息分析和复合图形表达。如图1-1所示：

朱东华与Alan Porter在2001年“Automated extraction and visualization of information for technological intelligence and forecasting”文中指出科技信息管理的关键在于获取大量数据、快速处理和有效的结果表达，并在“Natural Language Knowledge Discovery: Cluster Grouping Optimization”文中对信息可视化在技术机会分析和技术预测中的应用做了深入的研究。

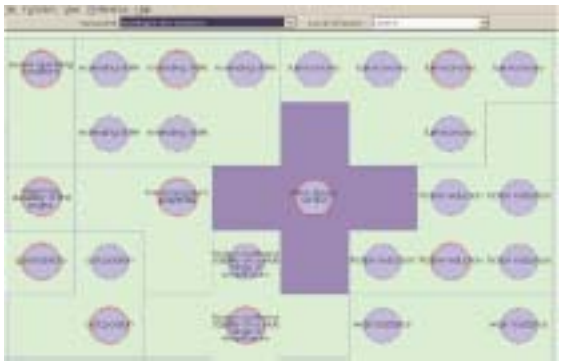

图 1- 1 2 维 10*10 节点的专利主题局部地形视图

1.2 研究意义

目前对专利的分析主要是集中在数理统计分析基础之上，结合专利分析方法的调研结合以往专利分析工作的经验，发现在现在的专利分析方法中，主要采用的方法有：

- 传统的统计方法
- 计量分析方法
- 技术生命周期分析法
- 技术矩阵分析法
- 引证分析法
- 关联分析法

这些方法对专利信息进行的挖掘。但是这些方法却有着固有的缺陷：在专利数据到信息，情报和知识转化不足；没有动态更新的技术识别能力；关键技术，竞争技术，辅助技术等技术识别方法欠缺等。

特别是在对技术组群的识别中，在对竞争技术的分析过程中缺乏有效的判定方式。在技术组群识别中，我们常采用关联和引用网络分析，但是这些方法本身是有其不可避免的缺陷的。关联分析法受分词准确与否的影响较大，但是目前存在的分词算法对于某些科技术语的抽取还是比较薄弱；引用分析受制于其引用专利的给出，在目前我们所得的专利数据库中只有美国专利对引用给出了比较详细的描述，这就限制了其对其他专利数据的利用，再次基于引用构造的技术组群更多的反应了技术的一个演变形式，对于相似技术的一个竞争合作关系并没有描述。

因此，在这种前提要，就需要我们来找寻另一种有效方式来填充其分析的空白。而通过对SOM方法改进以及应用到专利分析中后，可以提高专利分析的效率和正确性，

提高知识的准确性和可理解性；可以帮助企业理解和分析专利技术的之间的相互关系，以获得本行业或本企业的技术策略、技术热点领域、技术竞争态势、竞争企业情况等情报内容，为自身的决策提供十分重要的辅助参考，具有重大的研究和实用意义。

1.3 研究内容

本论文的主要内容是，通过利用SOM方法结合可视化数据挖掘这一技术对专利信息进行有效的处理和分析，找到大量专利情报背后隐藏的重要规律，获得企业或行业中专利技术策略、技术实力和技术特点等专利情报信息，以起到辅助决策的作用，如图1-2所示：

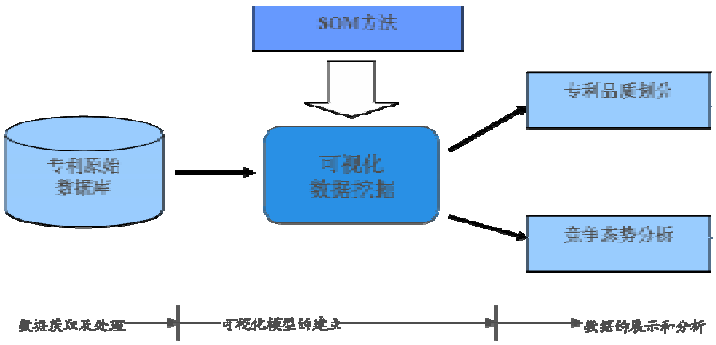

图 1- 2 基于 SOM 的可视化挖掘

具体研究内容有：对特定数据源专利数据的获取及预处理；SOM数据挖掘算法的应用及改进；可视化模型的设计和实现；以及某一领域的专利实证分析。

1.3.1 研究方法

本论文主要从科学技术角度，结合定性和定量分析方法，利用SOM技术，从宏观和微观层面对专利信息进行获取、分析、可视化和评价。在已有模型基础上对结果进行分析，得出无法从大量专利数据中直观得到的信息。

在实证方面，以集成电路的封装技术领域为例，结合北京理工大学知识发现与数据分析实验室专利分析平台的前期研究成果，集成获取该领域中专利信息，利用SOM可视化技术，对专利信息进行深层次的挖掘。

1.3.2 研究难点

- 专利数据的格式化和标准化。基于IPC分类的专利编号格式比较复杂，难于直接使用，所以必须进行格式化和标准化统一，并集成到专利数据库中，最后形成合理规范的业务数据集；同时，由于SOM是一种无教师自组织、自学习网络[30]，输入层的选择可以说对结果的产生具有非常重要的影响。因此，输入层变量的选择问题将是本次研究中的要点难点。
- 专业知识的淡化，规律的凸现。因为本研究旨在为企业、机构，或相关政府部门提供辅助决策的信息，然而专利的基础知识比较复杂和繁多，因此挖掘后的结果展示和分析应该尽量通俗易懂，简洁明了，所以必须考虑到对专利专业知识的淡化处理，而着重凸出规律性的结论。

由于本课题综合了目前先进的信息技术和专利技术进行探索性研究工作，所以在研究不断深入的过程必然会出现更多的潜在问题，对这些问题的有效解决也将成为本课题的重要研究内容。

1.4 论文结构及安排

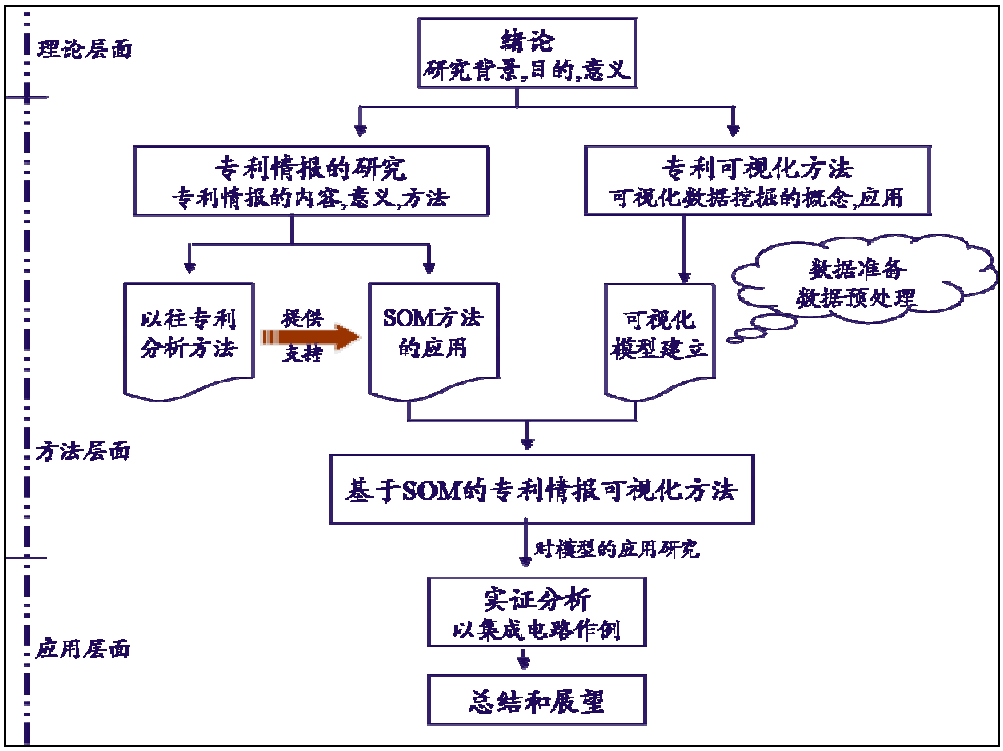

图 1- 3 论文框架

本论文主要从理论层面、方法层面和应用层面进行了划分。前两个部分主要来探

讨基于SOM技术的专利信息可视化方法，技术及以模型的探讨研究，最后一部分通过实证应用来验证方法的可行性。

第一部分，理论层面的研究，也即论文绪论的介绍。主要介绍了论文的选择依据以及研究意义；着眼于信息可视化技术以及专利情报分析领域的发展情况，提出论文的研究方法以及研究过程中的难点与重点。最后对论文研究提出简要框架。

第二部分，方法层面的研究，这也是整个论文的核心。通过对以往的专利分析工作的优缺点的比较，以及针对以往工作的不足，为了更好地利用专利情报挖掘隐含信息，本文提出了“ 基于SOM专利分析可视化模型 ”。然后在整体模型的基础上，对SOM可视化算法进行了研究，形成了论文的整个方法体系。

最后，实证分析，将已有模型应用到集成电路封装技术领域，对得到的结果进行分析，来验证本模型的正确性以及合理性。

2 相关理论研究

2.1 专利情报研究

专利是专利权的简称.它是指一项发明创造,即发明、实用新型或外观设计向果务院专利行政部门提出专利申请，经依法审查合格后，向专利申请人授予的在规定的时间内对该项发明创造享有的专有权[3]。

然而企业在专利申请上并不是一味的盲目[7]，会从市场、竞争前景、技术等方面进行考虑。也及在市场方面：它必须具有市场价值，也即企业需要通过从中来获得经济效益；竞争因素：如果与同类技术相比并不具有优势，形成不了足够强大的市场，那么也没有必要申请专利；技术因素：主要考虑该项技术创造仿制程度的难易。综上可知，专利情报是一种重要的信息，通过它可以获得相关技术、经济、市场方面的信息。本论文的可视化技术也是建立在专利的基础上的。

2.1.1 专利情报分析的意义

今天的社会是信息激增的社会。仅近三十年来的科学技术成果就超过以往人类历史两千年成果的总和。国外有的学者把信息同能源和材料并称为今日社会进步的三大技术支柱，把当今的社会称之为“ 情报社会 ”。以情报信息为对象，对其内容进行识别、整理、分析、综合、选择、推荐或加工出新的信息来服务于社会创造活动，这便是广义情报研究的任务[3]。情报研究是以当代科学技术的新成就为主要对象。判断这些成就的价值，发现问题启发思路，预告未来，提出建议。

文献情报研究是情报学的重要内容，属于情报服务中的高层次活动。它既包括文献研究，又在一定程度上涉及可行性分析和科学预测等领域[6]。文献情报研究作为一个以大量收集各种有关文献情报信息，经过加工、分析而提出有针对性的研究报告这样一种研究工作，与生产、科研、经济、政治等活动也有着非常密切的联系。毫无疑间，随着我国政治、经济、文化的发展，文献情报分析研究必将受到社会各界愈来愈大的重视，它的发展前景是十分广泛阔的。

2.1.2 专利分析流程

考虑到研究领域的不同、分析的目、要求不同，因此在研究过程中信息收集的范围，所采用的方法，以及最后结果的呈现方式都大相径庭。但是所有研究工作的分析流程都是大同小异的，具体图2-1所示：

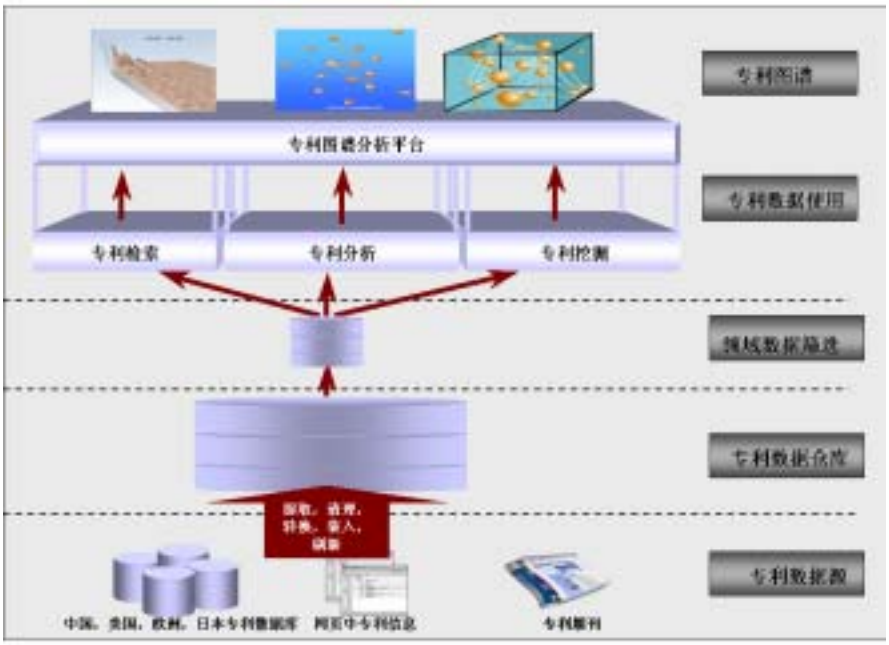

图 2- 1 专利分析流程

(1) 针对分析工作的目的，首先确定监测需求，初步选取监测对象。利用Spider技术，对下载服务器下达获取指令，服务器将相关数据进行获取，形成本地化数据库。

(2) 对于获取的数据，采用数据预处理技术进行有效数据提取，将其进行数据集成，形成特定领域的情报监测数据集，为监测分析做准备。

(3) 根据研究的领域，进一步根据前期调研获取的关键词或IPC进一步对所得的数据仓库进行筛选，最终获得所需的领域数据集。

(4) 在已获得领域数据集基础上，运用技术组群识别、关键技术识别、数据挖掘技术等技术对监测数据集和所拥有的本地化数据库进行分析。

(5) 在上部获得的分析结果基础上，采用信息可视化技术对所得结果进行可视化图谱展示，并攥写分析报告。其中可视化是用户对分析结果理解的基础，也是整个报告的核心。图谱的展现方式，直接关系着用户对整个结果的了解程度，也是本论文研究的核心。

2.1.3 基于 SOM 的专利情报分析

（1）以往分析方法

根据专利信息的特点，目前常用的专利分析方法可以简单划分为：定性分析和定量分析两大类。定量分析是指对专利文献所表现出来的具体显著特征进行的统计分析，也即对专利信息中所固有的元素如申请人，发明人，申请日期等进行简单的统计分析，或进一步的进行一些组合矩阵分析。常用的方法有：传统的统计分析、计量分析、技术生命周期分析、关联分析等。相对于定量分析，定性分析则表现的更加宏观，主要是从“质”的角度来把握技术发展的动向，从而进行情报预测，因此一般用于获取行业技术发展动向等方面的信息。目前我们知识发现与数据挖掘实验室在该领域常用到的分析方法为：功效矩阵、技术矩阵分析。

这些方法主要是建立在统计分析基础上，对专利信息进行的挖掘。但是这些方法却有着固有的缺陷：在专利数据到信息，情报和知识转化不足；没有动态更新的技术识别能力；关键技术，竞争技术，辅助技术等技术识别方法欠缺等。特别是在对技术组群的识别中，在对竞争技术的分析过程中缺乏有效的判定方式。在技术组群识别中，我们常采用关联和引用网络分析，但是这些方法本身是有其不可避免的缺陷的。关联分析法受分词准确与否的影响较大，但是目前存在的分词算法对于某些科技术语的抽取还是比较薄弱；引用分析受制于其引用专利的给出，在目前我们所得的专利数据库中只有美国专利对引用给出了比较详细的描述，这就限制了其对其他专利数据的利用，再次基于引用构造的技术组群更多的反应了技术的一个演变形式，对于相似技术的一个竞争合作关系并没有描述。

（2）基于SOM的专利分析

正是考虑到以往专利分析方法在获取技术组群，研究相关领域的竞争态势方面的缺陷，以及考虑到数据挖掘中的聚类算法对寻找相似项目群的可行性（如已知一个顾客数据集，确认一个具有相似购买行为的顾客子群），将聚类分析方法的引入到我们专利分析中是完全可能的。

聚类的算法主要有K-Means、Single-Link和SOM算法,这些算法同时也是划分聚类算法、层次聚类算法和基于模型的聚类算法的典型代表.它们有一个共同的特征就是建立在距离或者说相似度计算的基础之上。

通过对这几种算法优缺点的描述，本论文决定立足于SOM算法来对专利信息可视

化进行研究。

- SOM方法简介

SOM（自组织映射网络）也称为Kohonen网络，是由芬兰学者Teuvo Kohonen于1981年提出的。该网络是一个由全连接的神经元阵列组成的无教师自组织、自学习网络[29]。Kohonen认为，处于空间中不同区域的神经元有不同的分工。当一个神经网络接受外界输入模式时，将会分为不同的反应区域，各区域对输入模式具有不同的响应特征[29]。在论文第五章的基于SOM的可视化模型中对SOM方法将有详细的说明。

- SOM方法引入的原因

SOM自组织神经网络作为基于模型的聚类算法的代表，是试图优化给定的数据和某些数学模型之间的适应性。这样的方法经常是基于这样的假设：为每一个簇假定了一个模型，然后寻找数据与给定模型的最佳拟合。它可把任意高维输人数据变换到二维的网格映照图，并保持一定的拓扑有序性[31]。类别相似的神经元靠得比较近，类别不相似的神经元分得比较开，因而它在数据分析中具有独特而又强大的功能，并已在数据分类、知识获取、过程监控、故障识别等领域中得到了广泛应用。而k-means算法由于其结果的不确定性以及在孤立点方面的欠缺（因为在我们的分析中，孤立点有时并不是某些噪声数据，而很可能是某项新兴的技术或其他）都会对我们的分析工作产生阻碍。

因此，我们通过它来对专利数据进行技术组群分析是非常恰当的。

- SOM方法的可行性

一个神经网络接受外界输入模式，将分成不同区域，各区域中邻近神经元通过交互作用，相互竞争，自适应地形成对输入模式的不同响应检测器，相邻神经元常常识别同一类别模式[31]。因此，Kohonen网络在功能上通过网络中神经元间的交互作用和相互竞争，模拟了大脑信息处理的聚类功能；这对于我们研究技术的族群中的技术竞争与合作是非常有用的。

通过对SOM方法应用的进一步研究发现，SOM方法已经在评价产业竞争力、股票市场的预测等领域得到了广泛应用。而我们在对专利的分析中，拟采用专利指标作为其分析的输入样本。这些样本数据从类型上应该与产业竞争指标等具有同构性，因此将SOM方法方法引入到专利分析中来，是具有可行性的。

- 引入SOM方法预期达到的效果

基于SOM的专利情报可视化数据挖掘可以提高专利分析的效率和正确性，提高知

识的准确性和可理解性；可以帮助企业理解和分析专利技术的之间的相互关系，以获得本行业或本企业的技术策略、技术热点领域、技术竞争态势、竞争企业情况等情报内容，为自身的决策提供十分重要的辅助参考，具有重大的研究和实用意义。

2.2 专利情报可视化方法综述

关于可视化的概念目前并没有一致的说法。比较公认的有赫斯特教授提出的：“可视化就是使用空间图像和图表来表示和描述信息，充分利用人们对可视模式快速识别的自然能力,进而改变我们的认知技能”[41]。科学可视化、数据可视化、信息可视化是可视化领域的三个分支。而本论文着重研究的是信息可视化。

所谓信息可视化,就是利用计算机支撑的、交互的,用抽象数据的可视表示,来增强人们对非物理抽象信息的认知[42]。信息可视化与数据可视化相比,信息可视化并不仅仅满足于数据可视化的方式对数据进行存储、传输、检索及分类显示,信息可视化更注重对其进行高层次分析,以发现数据间的联系和变化规则。在信息可视化中,显示对象主要是多维数据,研究重点是设计和选择什么样的显示方式才能便于用户了解庞大的多维数据及它们的关系。

2.2.1 信息可视化

(1)与数据挖掘的关系

信息可视化使用计算机支持的、人机交互的、视觉表现的方式，是对抽象数据认识的放大。通过数据挖掘的方法和手段，将海量的数据之间的关系、重要的信息和特别的内容通过图形和图像的方式展现出来。其主要的功能是反映和揭示大量数据之后的隐藏信息，将难以形象表达的抽象数据设计成为更加容易理解的表现形式。

数据挖掘和知识发现与信息可视化密不可分，它们都可以认为是可视化过程中的数据分析和提取的过程；当然信息可视化也可以认为是数据挖掘和知识发现的一种表达形式。这主要取决于整个研究工作的重心在那个部分。从这种观点出发，数据挖掘和知识发现与信息可视化息息相关，在本文的研究中，它们具有较多的重叠，因此很多地方本论文将许多功能模块都看作为信息可视化的部分。

信息可视化作为文献信息分析的一种手段，将数据挖掘、知识发现等技术综合运

用到整个系统中。将信息对象进行综合、抽象、概念化、知识化，从而更方便简洁地实现可视化，发现文献之间地关系以及文献作者的信息等。

（2）信息可视化作用

信息可视化充分地利用了我们认知系统的能力,将信息转化为图像的形式,非常方便地增强了我们的记忆和识别能力。信息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。目前信息可视化主要应用在如下几个方面:多维数据可视化,如人口普查、顾客群、销售业绩等;数字图书馆,如书籍、照片、WWW网页、报纸和杂志等;复杂文档,如专利文献、年报、软件、论文信息等;网络,如互联网的远程连接和使用、电子电路、组织结构图等;统计和分类系统,如专利数据、物种分类、硬盘数据等在信息可视化的实际应用中,它的具体作用可概括为如下几点:使海量数据变得简洁有条理、多角度地展示信息、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[12]。美国伊利诺斯大学学者托马斯说,与科学计算可视化不同,信息可视化和数据挖掘,它们“内部含有让人喜欢的物理特征”,信息可视化解决了如何和信息资源之间进行对话的问题[14]。信息可视化有诸多的优点,下面将结合可视化在技术监测中的应用来探讨可视化技术的应用以及优势。

（3）可视化分析流程

如图2-2所示，参照一些以往关于可视化数据挖掘的分析步骤结合本论文的研究，我们将可视化分析分为了：数据采集、数据预处理和数据分析三个阶段。

其中数据采集是可视化挖掘的首要步骤，也是整个分析的基础，是整个可视化的数据获取准备阶段，它主要包括了数据源的选择和原始数据的存储两个方面。在数据源选择方面，需要明确所研究领域所涉及的数据（比如，我们需要研究一些国防方面的文献信息，我们所需的数据源可以选择AD报告、AIAA报告等），以及数据源的更新情况，原始数据的结构构成，这些都为可以高效、准确、完整的获得数据提供基础。数据源选择好之后，我们可以利用北京理工大学知识发现与数据分析实验室自行研制的专利/文献提取软件进行信息获取，进一步形成本地化的原始数据存储，以方面后续工作的进行。

数据预处理阶段：将上部所获得的本地原始数据根据用户的确定的研究领域对原始信息进行初步筛选，选择出待挖掘的数据集；接着在上述数据集基础上展开清洗工作，剔除冗余、噪音数据；在此基础上，对所获得的数据进行进一步的格式转换以便

提高挖掘的效率；最后将所获得的数据装载入等待挖掘的数据库中，至此数据预处理阶段结束。数据预处理主要是为提高分析的质量进行的，尽量使得分析不受噪声词的干扰。

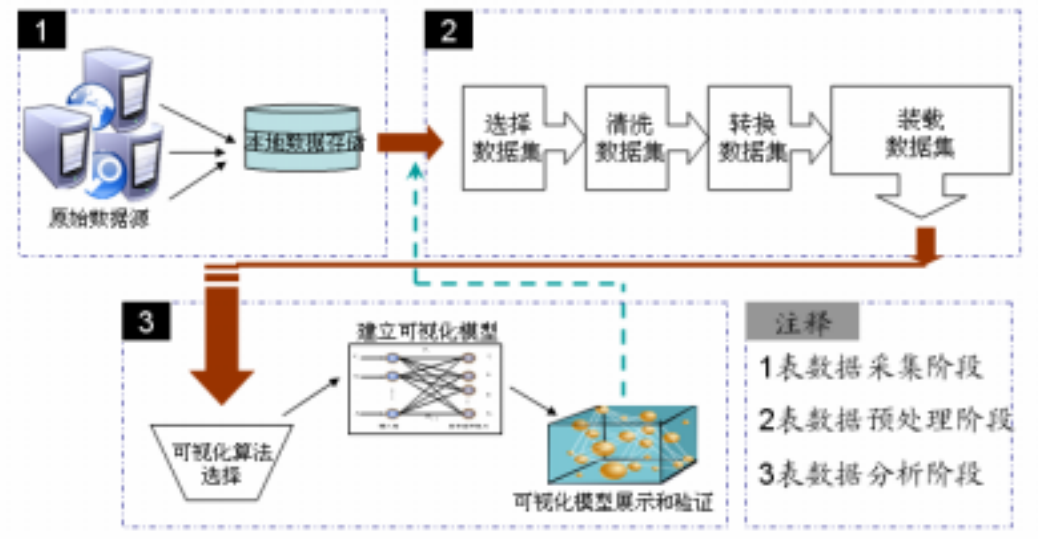

图 2- 2 可视化挖掘流程

最后的数据分析阶段是整个可视化挖掘的核心。其中的可视化模型的建立以及算法的选择都直接影响到了分析结果的准确性和可信度；而最后可视化的展示则也影响到用户对整个结果的理解程度。因此整个模型的建立既要将尽可能多的信息以人类视觉容易理解的形式展现，又要保证对用户的易理解性。

但是，考虑到数据预处理和数据分析不可能一步到位，而是一个不断循环往复的过程。当可视化所展现的结果不满足用户的需求或者分析有误时，就需要返回到数据初始阶段，对所得数据进行进一步检验，看是否仍存在噪声数据的干扰，或是数据清洗方法有误等等，接着对于分析的模型算法进行进一步的验证。另外，随着对一个领域认识的不断加深，以及对分析要求的不断提高，可视化挖掘的步骤也将不断重复。

2.2.2 专利信息可视化可行性研究

（1）专利信息特点

- 数据获得的可行性

目前专利信息的检索在很多地方是公开的，如中国国家知识产权局网

（http://www.sipo.gov.cn/sipo/zljs/default.htm），美国专利局专利专利检索（http://www.uspto.gov/patft/index.html）等等。因此专利信息的获取还是比较方便和简单可行的。

- 数据量庞大，涉及领域广泛

首先，专利设计的范围广发，从日常用品的制造到核能、人造卫星等高端技术都有专利文献的涉足。据不完全统计，从1985年迄今，在我国境内申请的专利数量为2884724，且相比美国、欧洲等地，中国的专利申请量并不是很大，由此可以看出，专利信息量之庞大。

- 信息包含全面，可供利用率强

专利信息除了详细的描述了某一技术细节之外，还包含了具体申请状态，类比划分（IPC），目前的法律状态等等信息。相较其他技术期刊，其内容包含的全面程度是远远所无法比拟的。

- 反映前沿技术发展

专利信息是受国家政策保护的，任何实体使用对方专利信息所涉及的技术都是需要支付对方专利费的。同时，大多数国家对于专利申请采取的政策都为先申请制，因此申请人都力争在技术方案构成后，抢先申请专利。这也使得了专利信息成为了新技术报道的最快途径。

- 规范统一，便于使用

专利文献据按国际统一格式出版，采用统一的著录项目识别代码、统一的国家名称代码和统一的国际专利分类系统，这些统一使得对专利信息的收集、使用带来了便利。

同时，知识产权已成为各个国家综合实力的重要体现，使各国在经济全球化过程中维持其竞争优势的有力武器。而很多国家已经明确提出本国的知识产权战略，并把其作为振兴本国经济、增强国际竞争力的战略措施。同时，考虑到专利信息在竞争中所起的作用不断加强，各国政府或是专利信息咨询机构纷纷投入大量人力、财力等建立自己的专利数据库，以便为社会提供高层次的专利信息服务。如中国专利局提供的中国专利的免费检索，主要有中国专利文摘数据检索；有美国专利商标局提供的美国专利免费检索，包括全文数据库和专利文摘数据库等；另外还有欧洲专利库提供的专利全文免费检索等等。

（2）专利信息可视化研究

目前，很多公司都对专利信息的获取、分析、可视化等方面都进行了深入研究，如恒和顿数据科技优先公司开发的“世界专利全文数据库”，通过互联网为广大用户提供专利信息检索、浏览等服务；并在数据管理基础上，开发出了“专利分析系统”系统，对数据进行进一步的挖掘分析，此外东方灵顿等公司在专利分析方面也有很大的贡献。

然而，面对如此“海量”的数据（无论从空间还是时间纬度来讲），传统的分析处理手段很难挖掘到真正的专利情报，人们也无法真正理解和利用这些潜在的信息。因此，数据挖掘技术可以弥补这些传统数据分析方法方面的不足，从而提供一种更加有效的解决方案。同时，可视化技术可以以友好的界面来呈现挖掘出的结果，这就使得基于可视化理念的系统可以更好的深入更多的人群中去，而不仅仅局限在少数专家手中，从而更好的发挥作用，

2.2.3 可视化方法研究

（1）可视化的实现策略研究

本论文的可视化是基于专利信息基础上的，属于信息可视化范畴。信息可视化的对象只具有语义属性，任何对语义关系进行空间排序是作为信息可视化来完成的[43]，其主要研究的是多维数据集，这些数据集的可视化的主要难点是在多维空间中同时表达各维属性的分布，也就是建立反映多维数据集的全局图像。因此，实现信息可视化必须有特殊的策略。

- 多维几何投影

信息可视化主要面对的是多维数据集的可视化，就要同时把各维属性投影在二维平面上，以反映各矢量间的关系及分布。信息可视化采用的投影策略主要有三种：一种是分别建立多维数据集中每两个属性之间的投影关系，有多幅关系图反映属性的全局关系，这种实现策略的典型实例是散乱图矩阵（Scatterplot Matrices）；第二种是把各维属性的数据轴同时布局在二维平面上，建立各数据矢量的图示，以反映数据集的关系特征，这种实现策略的典型实例是平行坐标（Parallel Coordinates）表示法；另一种是采用多维数据降维的方式显示，如：多维尺度变换（Multidimensional Scaling，MDS）、自组织特征映射的网络（Self-organized Feature Maps，SOMS）等方法降维。

- 数据映射技术

信息可视化面对的多维数据库存放的常常是海量数据集，数据可能有数万到数十万的数据项，如果要将这些数据以可视的方式表达出来，就要通过各种方式建立各数据项的各个属性与屏幕上相应显示元素的映射，对于此类海量数据向屏幕上映射的最有效的方式是面向象素（Pixel-oriented）的映射。

- 层次技术

信息可视化面向的多维数据库中所存数据常常具有层次特征。例如，数据库中存放的某一机构、某类产品信息常按树状存放。在信息可视化过程中，要用图形表达出树状结构。传统的树状图表达数据量很大的复杂的层次结构是十分困难的，所以要引入便于图形表达的方法，典型的树图（Treemap）表示。

- 聚类分析

信息可视化的主要目标是通过可视化图像挖掘数据集中潜在的模式及知识，而聚类分析（Clustering Analysis）则是数据挖掘过程中的一种重要方式。聚类是指把数据项中相似对象聚焦为一个对象组。聚类分析在模式识别、图像处理、数据分析中都具有重要作用。信息可视化无论用什么策略实现，在最后形成的图像中，都应该是表现出聚类关系的直观图像，以从中发现有用的决策信息。

- 关联分析

信息可视化的对象数据之间存在着千丝万缕的联系，通过对事物的属性之间的关联分析（Association analysis），进一步可以帮助我们发现事物间的关联规则。这些关联的规则是信息可视化表示重要的关系表示，也是数据挖掘的重要的成果表示，同时也是模式识别的重要数据依据。因此它被广泛应用于信息可视化的数据分析过程中。

在本次可视化研究中，主要采用的是聚类分析中的SOM方法进行专利信息的研究。

（2）可视化算法研究

目前常见的网络图形可视化展示的算法主要有如下两种：

- Pathfinder Network(PFNET)

PFNET是根据经验性的数据，对不同概念或实体间联系的相似或差异程度做出评估，然后应用图论中的一些基本概念和原理生成的一类特殊的网状模型。它对不同概念或实体间形成的语义网络进行表达，从一定程度上模拟了人脑的记忆模型和联想式思维方式，主要应用于认识心理学和人工智能等研究方面。在一般变换情况下，PENET

具有一定的稳定性，并且通过对PENET的分析，可以对不同的概念、实体进行分层和聚类。

- Self-organized Feature Maps（SOMS）

SOM网络为一种自组织特征映射的网络，是由Kohone提出的一种人工神经元网络，本论文的可视化算法是以此为前提的，对于该方法在第三章有详细的描述，在此不再繁述。

另外还有最小生成树（Minimum Spanning Tree，MST）、多维尺度变换（Multidimensional Scaling ，MDS）等算法。各种算法都具有自己的优势和不足，主要是考虑可视化的对象侧重点，选择不同算法。

（3）可视化工具研究

SENTINEL是哈利斯公司的一个集合了n克获取引擎和有效文本矢量查询的可视化系统[20]。该可视化工具允许用户进行交互的文本三维聚类。美国Drexel大学的陈朝美教授采用VRML2.0对论文间的关联相似度和引用网络进行了可视化研究。在本研究中提出的Pathfinder算法能够很好的进行共引关系的分析[21]。

美国西北太平洋国家实验室研究开发了SPIRE可视化系统[22]。SPIRE系统是二维的可视化界面，所以相关的对象之间距离非常的近，并且该系统还提供了用户同数据进行交互的工具。在SPIRE中使用了两种的可视化模式。在采用星状分布模式时，文献以散点的形式出现。这种方式能够进一步深入到散点分布的某个部分中去，并且还提供了一些归纳总结的工具。在主题地图分布模式下，文献的主题以山峰和沟壑的形式出现。山峰的高度体现了一定文献领域中该主题的出现强度。

美国Sandia国家实验室开发的VxInsight[23]系统于SPIRE类似，也是提供了两种可视化的展示模式。但是所不同的是VxInsight采用了多维可视化展示，这使得信息的许多内在结构和联系能够在多维水平上进行分析。此外，VxInsight还能够在静态展示的基础上,进行视图的放缩，即可以在主题地形图上任意的进行放大缩小的操作，以获得更加详细的特征信息。

国内的学者也在可视化领域作出了一定的成绩。复旦大学计算机系自行设计实现了DMVisualMiner数据分析平台。该平台采用构件的设计方法，利用插件的概念增强了系统的可扩展性，设计并实现了基于XML的模型表示方法，使得DMVisualMiner 能够和预言模型系统集成，并能在网络环境下发布 [24] 。但是，相对于国外大量成熟的商业可视化系统，我们还需要作出更大的努力。

3 基于 SOM 的可视化模型

专利情报指的是定向、定时传递的有用的专利信息。专利情报不断公开，技术内容便不断更新。一般我们把不断公开的情报看成是在各个不同地点公开的互不相关的点情报。如果通过某种规则，将这些点情报利用聚类等挖掘算法进行操作，可以将这些点情报转换为面情报，进而可以使得我们更方便直观地掌握某特定领域技术、企业的竞争态势关系，发展和应用。因此一个有利的专利可视化方法模型对于挖据出潜在的信息有着不可代替的作用。

本章是在对SOM方法进行详细研究的基础上，结合专利信息本身的特性以及预期挖掘结果的期望提出了基于SOM可视化模型，并对模型的主要环节进行了详细阐述。

3.1 SOM 方法

近年来，人工神经网络因其强大的处理机制、任意函数的逼近能力以及自组织和自适应能力在建模、决策等广泛领域得到了广泛的应用。

自组织特征映射网络也称为Kohonen网络，或者称为Self-Organizing Feature Map(SOM)网络，它是由芬兰学者Teuvo Kohonen于1981年提出的。该网络是一个由全连接的神经元阵列组成的无教师自组织、自学习网络。Kohonen认为，处于空间中不同区域的神经元有不同的分工。当一个神经网络接受外界输入模式时，将会分为不同的反应区域，各区域对输入模式具有不同的响应特征[30]。

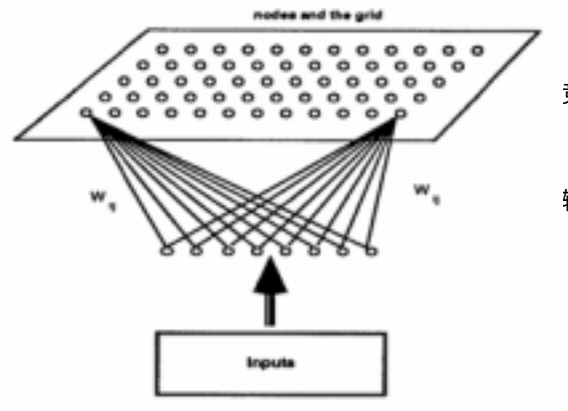

图 3- 1 SOM 网络结构

SOM 神经网络由输入层和竞争层组成。输入层由 N 个神经元组成，竞争层由 M 个神经元组成。为可视化表示结果，常将 M 表示为一个二维平面阵列 M=s*t ，当然，一维或多维也是允许的。SOM 网络的一个典型特征就是可以在一维或二维的处理单元阵列上，形成输入信号的特征拓扑分布，因此 SOM 网络具有抽取输入信号模式特征的能力。

SOM 网络模型由以下 4 个部分组成：

1.处理单元阵列。用于接受事件输入，并且形成对这些信号的“ 判别函数 ”

2.比较选择机制。用于比较“ 判别函数 ”，并选择一个具有最大函数输出值的处理单元。

3.局部互连作用。用于同时激励被选择的处理单元及其最邻近的处理单元。

4.自适应过程。用于修正被激励的处理单元的参数，以增加其对应于特定输入“ 判别函数 ”的输出值。

假定网络输入为 $X \in R^n$，输入神经元 i 与输出单元的连接权值 $W_t \in R^n$，则输出神经元 i 的输出 O_i 为 $O_i = W_i X$

网络实际具有相应的输出单元 k，该神经元的确定是通过“ 赢者通吃 ”的竞争机制得到的，其输出为：$O_i = \max\{O_i\}$

将上式可修正为

$$O_i = \delta[\psi_i + \sum_{t \in S_T} r_k O_t - O_k], \psi_i = \sum_{j=1}^{m} w_{ij} x_j, O_k = \max\{O_i\} - \xi$$

其中，w_{ij} 为输入神经元 i 和输入神经元 j 之间的连接权值。x_i 为输入神经元 i 的输出。$\delta(t)$ 为非线性函数，即

$$\delta(t) = \begin{cases} 0 & t < 0 \\ \delta(t) & 0 \le t \le A \\ A & t > A \end{cases}$$

ξ 为一个很小的正数，r_k 为系数，它与权值及横向连接有关，S_i 为处理单元 i 相关的处理稽核，o_k 称为浮动阙值函数。

SOM 网络学习算法

(1)初始化。对 N 个输入神经元到输出神经元的连接权值赋予较小的权值。选取输出神经元 j 个“邻接神经元”的集合S_j。其中，$S_j(0)$表示时刻 t=0 的神经元 j 的“邻接神经元”的集合，$S_j(t)$表示时刻 t 的“邻接神经元”的集合。区域$S_j(t)$随着时间的增长而不断缩小。

(2)提供新的输入模式 X

(3)计算欧式距离d_j，即输入样本与每个输出神经元 j 之间的距离：

$$d_j = \|X - W_J\| = \sqrt{\sum_{i=1}^{N}[x_i(t) - w_{ij}(t)]^2}$$

并计算出一个具有最小距离的神经元j^*，即确定出某个单元 k，使得对于任意的 j，都有$d_k = \min_j(d_j)$。

(4)给出一个周期的领域$S_k(t)$

(5)按照下式修正输出神经元j^*及其“邻接神经元”的权值：

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)[x_i(t) - w_{ij}(t)]$$

其中，η为一个增益项，并随时间变化逐渐下降到零，一般取

$$\eta(t) = \frac{1}{t} \text{ 或 } \eta(t) = 0.2[1 - \frac{t}{10000}]$$

(6)计算输出o_k：

$$o_k = f(\min_j \|X - W_j\|)$$

其中 f(.)为一般 0-1 函数或其他非线性函数

(7)提供新的学习样本来重复上述学习过程

3.2 可视化模型的提出

SOM神经网络可以在结构上模拟大脑皮层中神经元二维空间点阵的结构，并在功

能上通过网络中神经元间的相互作用和相互竞争，模拟了大脑信息处理的聚类功能、自组织和学习功能。目前SOM方法已经广泛应用到分类问题中。

目前，随着搜索技术的发展，我们可以方便快捷的获取大量的专利信息，然而获得的这些专利文献往往都是基础的、未处理的原始的非结构的数据。正是考虑到数据处理及清洗方面的问题，本文将专利情报研究与可视化数据挖掘技术结合起来，提出了基于SOM方法的专利情报可视化模型，如下图所示：

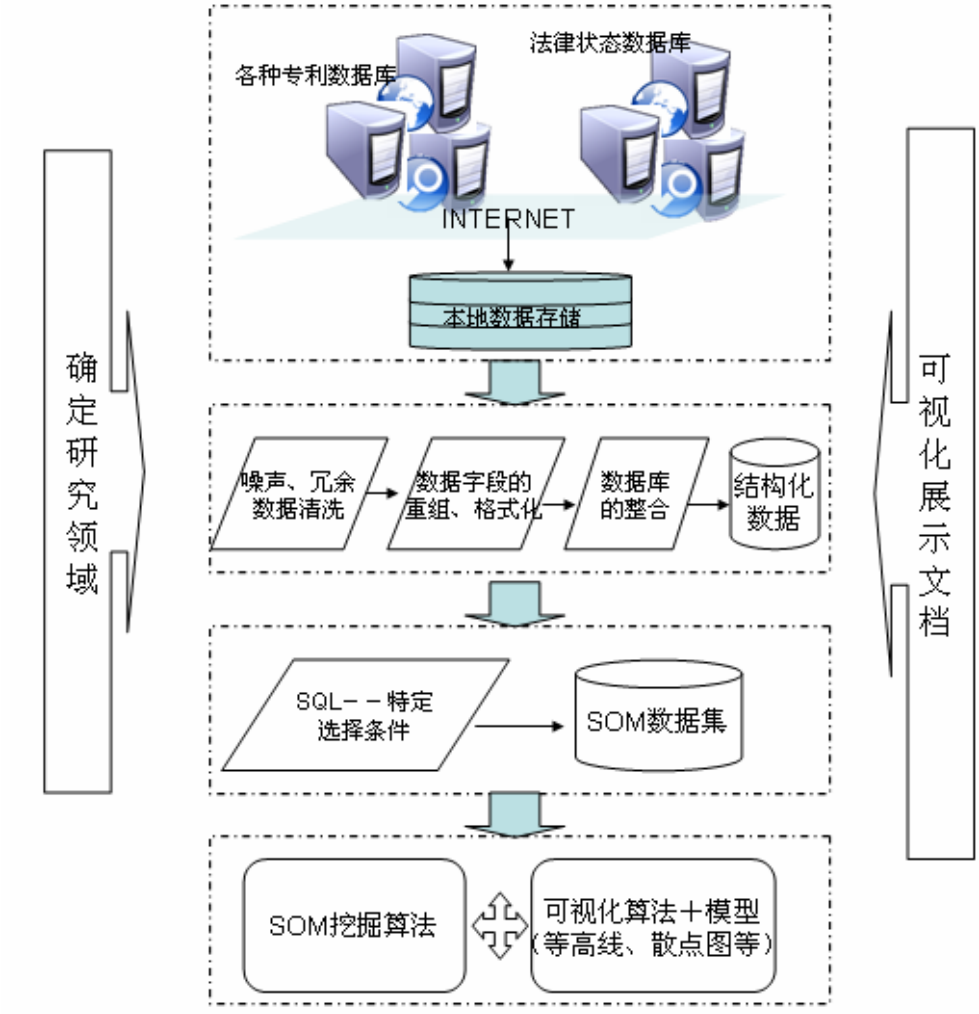

图 3- 2 基于 SOM 的可视化模型

同传统的专利情报分析研究过程一样，明确研究领域所需要设计的专利情报是首要条件。

（1）数据获取：本研究主要利用的中国专利和法律状态的web数据库。Web内容抽取的核心是从web页面包含的非结构化或半结构化信息识别用户所感兴趣的数据，并惊奇转化为结构化的格式。本文主要采用spider技术将网页自动下载至本地数据

库，然后利用正则表达式等方法对非结构化的网页数据信息进行分次处理，自动转存到SQL SERVER数据库中。

（2）数据清洗、转换、整合：采用IPC等技术去除噪声数据的影响，得到纯度较高的数据集；根据研究的具体要求，派生出其它的属性（如，在数据的发明人字段包含每个发明人的信息，而我们所关注的仅为发明人个数或是第一发明人的具体信息，为此需要采用相应的计算手段派生出相应的属性）。此外，由于本论文涉及专利数据库和相应法律状态两个数据源，因此需要根据数据的关联性对数据库进行重新设计，然后根据数据库字段需求将结构化的数据导入倒数据库中进行整合，形成一个整合的数据库。

（3）训练数据集的提取：（2）中生成的整合的数据库更多的是文本信息，而在我们具体的可视化过程中，需要的输入层是基于数字基础上的，因此需要借助一定的方法形成适合SOM方法的数据集。

（4）SOM可视化挖掘：是整个论文的核心部分，本论文也对其具体的可视化部分进行了详细的设计，如下所示：

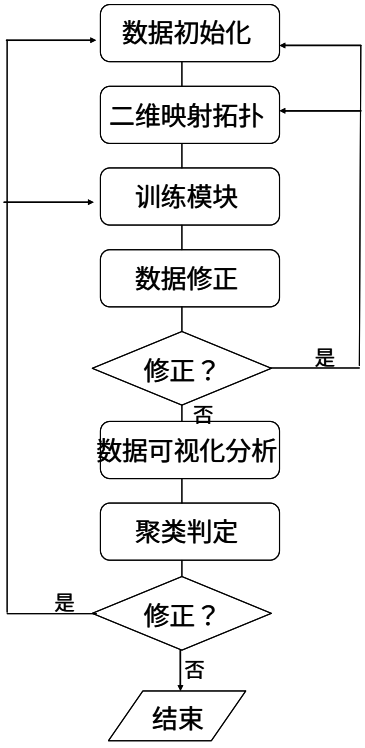

图 3- 3 可视化流程设计

在生成的数据集基础上，本论文将整个可视化模型划分为6个模块，具体描述如下：

第一个模块：数据初始化模块，为了减少数据中数据中不同属性绝对值大小对网

络计算的影响，需对训练样本的各属性进行归一化预处理，同时初始化权值。

第二个模块：二维映射拓扑模块，即设定输出层的形状和规模，通常根据数据的规模大小进行设定。

第三个模块：训练模块，通过训练模块对样本进行聚类，获得样本的二维映射图。

第四个模块：数据修正模块：对产生的一些异常点进行分析，判断其是否是噪声数据，如果是则重新调整1，2模块，否则进入下一环节。

第五个模块：数据可视化分析模块，通过此模块对网络训练后获得的巨雷各种特征进行可视化分析。

第六个模块：聚类判定模块：对聚类类别数进行最佳分类数判定。通过判定对聚类结果进行进一步认识和再分析（比如：剔出异点，数据属性修正等）。

在最后一个聚类判定模块中，采用了Davies-Bouldin聚类判定法，其基本原理如下：

$$f = \frac{1}{c}\sum_{k=1}^{c}\max\{\frac{S_C(Q_K)+S_C(Q_I)}{d_{cm}(Q_k,Q_i)}\}$$

其中，$S_C(Q_K)$，$S_C(Q_I)$分别为聚类后的数据中第k与第i个聚类的类内样本间距，c为分类数。通过Davies-Bouldin聚类判定法可以获得不同分类数的f值，f值最小时的分类数为最佳分类数。

3.3 本章小结

本章的主要目的是提出基于SOM的专利信息可视化模型，为以后的专利分析提供理论基础。首先，对SOM的组成及学习过程进行研究；然后结合专利信息特有的性质，提出了专利分析的可视化模型，并对SOM的可视化分析流程进行了阐述。

4 可视化模型研究

根据前章提出的可视化模型，本章主要针对模型中涉及的主要方法进行研究，以期实现预期的目标。在对方法的描述中，根据可视化模型的特点，本章从数据处理和可视化方法两方面对模型涉及的主要方法进行描述。数据处理方面：主要阐述了数据获取、预处理和训练集生成方面的主要方法；而在可视化层面着重描述了可视化过程的实现。

4.1 数据处理方法研究

数据处理部分主要包括数据获取、数据预处理以及相应可视化训练集的生成等部分。这些共同构成了信息可视化的基础。

4.1.1 数据准备

（1）数据获取

专利数据的获取是本论文进行一下研究的基础。动态、便捷的获取并转化数据是这个阶段的主要任务。为了分析能够长期进行，并且实现数据动态监测跟踪，远程数据本地化是工作的基础。对于数据获取本文主要采用了如下两种方法：利用Spider搜寻。由于搜索引擎大都采用的数据库来记录网站信息和网页信息，因此使用科技搜索词作为搜索关键词来查询对象名称；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，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4-1所示。

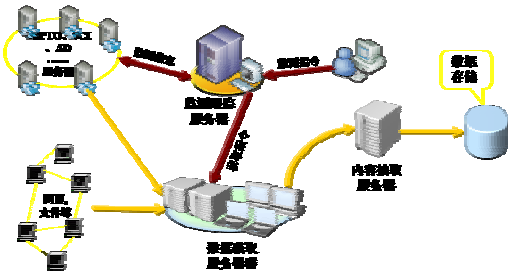

图4- 1 数据监测、集成流程

（2）数据抽取

本论文获取的信息都是可以从internet web页面中可以全部看到的，因此采用数据库记录网站信息和网页信息，再利用网页标签提取专利文献的要素。具体过程可以概括如下：首先，构建DOM（Document Object Model）树；其次，确定网页提取规则；然后按照设定好的格式提取网页信息。

由于我们获取到的原始专利信息是html形式的，而HTML最常用的结构表示方法就是构造网页的标签树，而现有的标签树构造DOM是一个常用的标签树构造工具，可以将网页中的标签按照潜逃关系整理成一颗树状结构。

- 中国专利信息提取

图4- 2 中国专利信息网络数据

上图是一个中国专利网路数据库的数据格式，途中的虚线框内的地方是需要提取存储在数据库中的。

首先根据掌握的专利数据库web页面的只是，分析所要处理的具体网页内；然后根据网页的结构树定制网页信息提取规则；最后则根据定制的规则便利整个结构树，将信息块填充倒数据库中，完成web信息的提取。

下图是采用正则表达式进行数据抽取的主要代码：

```
///提取专利号（好用）
static string distillPatNum(string strWebPage)
{
    Regex Pattern=new Regex(@"\bUnited\s*?States\s*?Patent\s*?Application\s*?</B></TD>(\s|.)*?<B>(?<PatNum>(\s|.)*?)</B
    Match m=Pattern.Match(strWebPage);
    if(m.Success)
    {
        return m.Groups["PatNum"].ToString().Trim();
    }
    else
    {
        return "NO";
    }

}
///提取第一发明人(好用)
static string distillPriInventor(string strWebPage)
{
    Regex Pattern=new Regex(@"\bUnited\s*?States\s*?Patent\s*?Application\s*?</B>(\s|.)*?</B>(\s|.)*?G>(?<PriInventor>
    Match m=Pattern.Match(strWebPage);
    if(m.Success)
    {
        return m.Groups["PriInventor"].ToString().Trim();
    }
    else
    {
        return "NO";
    }
}
///提取时间（好用）
static string distillTime(string strWebPage)
{
    Regex Pattern=new Regex(@"\bUnited\s*?States\s*?Patent\s*?Application\s*?</B>(\s|.)*?</B>(\s|.)*?</B>(\s|.)*?</B>(\
    Match m=Pattern.Match(strWebPage);
    if(m.Success)
    {
        return m.Groups["Time"].ToString().Trim();
    }
```

根据专利信息网页数据本身的结构确定相应的字段提取规则：在DOM树基础上根据各个字段的特点构造采用不同的构造的正则表达式匹配式对信息进行抽取。如：对于专利号的提取，在html文档中，专利号的一般表现形式为（United States Patent Application </td> 专利号），间隔符中间可能会存在空格之类的符号，因此在构造匹配表达式时加入了非数字符合(/s)，任意字符(*)，任意单个字符(?)等元素。

- 法律状态信息提取

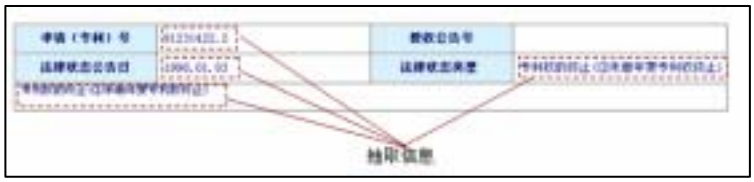

图4- 3法律状态信息网络数据

法律状态web数据与专利数据库最大的区别在与：法律状态的信息获取并不是已关键词为基础的，而是根据以获得的专利信息的专利号为检索条件查询其相应的法律状态数据。

获取方式也是以相应的DOM树结构为基础的。上图是法律状态数据需要提取的信

息，具体信息获取方法可参照下面的数据获取代码。

```
//核心：正则表达式取数据
if(dcompany.IsNull(0)==false)
{   //根据法律状态所特有的性质，进行替换算法
    string j=dcompany[0].ToString();
    j=j.Replace("\r\n","~~~~~~~");
    j=j.Replace("\n","~");
    j=j.Replace("~~~~~~~","\r\n");

    Regex r = new Regex(@"<td>.*</td>",
        RegexOptions.Compiled|RegexOptions.Multiline|RegexOptions.IgnoreCase);
    //正则匹配
    Match   m     =   r.Match(j);
    int   n=0;
    string[] a = new string[4];
    i=0;
    while   (m.Success)
    {
        n++;
        string str_result=m.Value.Replace("~","\n");
        str_result=str_result.Replace("<td>","");
        str_result=str_result.Replace("</td>","");
        str_result=str_result.Replace("</pre>","");
        str_result=str_result.Replace("<pre>","");
        //将匹配出的数据存加数组
        a[i]=str_result;
        m   =   m.NextMatch();
        i++;
    }
    //经过分词的数据入库处理
    string InsString = "insert into lawstatus(Record_No,Patent_NO,Issue_NO,Leg_State,Legstate_Date,
    SqlCommand InsertCmd1 = new SqlCommand(InsString,mycnn);
    try
    {
        InsertCmd1.ExecuteNonQuery();}
    catch(Exception ex)
    {
        MessageBox.Show("数据增加错误！" + ex.Message, "错误提示", MessageBoxButtons.OK, MessageBox.
    }
    addnode(a,dcompany[1].ToString(),dcompany[2].ToString());
```

4.1.2 数据预处理

数据预处理阶段主要完成数据清洗、数据属性的扩展、数据库的整合3个方面。数据清洗主要是消除噪音等错误信息，而属性扩展、数据整合则需要根据具体的研究需求来具体分析，具体的方法将在实证分析中给出，这里只是给出整体的方法描述。

（1）数据去重

根据专利信息的特殊性，即对于每条专利信息来讲，它的专利号都是唯一的，这也为我们提高了一条通过专利号去掉重复数据的方法，具体方法为将同一技术领域的不同关键词查询获取到的专利文献数据进行一次清洗，对前面已经存储过的数据不再进行存储。

（2）噪声数据、空值的处理

噪声数据也即通过关键词查询获取的特定领域数据其实并不属于该领域，因此如

果可视化挖掘之前没有对这些数据进行清洗则会对结果的正确性产生影响。本文对噪声数据的处理主要是采取IPC分类方法，

对于包含空值的数据来讲，一般方法为：a.该字段并不再待挖掘属性的范围，则无需处理；b.对于这条数据重新进行查询，以便可以补全内容；c.对于无法补全的值，可以用一个默认值来代替空值，通常这个值表示该数据的不确定性，并不影响最后的挖掘结果；d.将包含空值的数据剔除掉。

4.1.3 可视化训练集的生成

对于专利信息来讲，一般涉及所面临的更多是他的文本信息，如何将这些文本信息转化为适合SOM训练集的信息是本文首先需要解决的方面。然而训练集是进行可视化的基础，直接影响着可视化的效果，因此要求训练集体现具体的需求情况。因此，本文将训练集生成的方法放在实证分析中进行更为合理。

4.2 可视化结果实现

根据SOM算法的基本思想，本文将高维的信息通过投影为低维信息。本文对其整个设计如图4-1所示（其中关于SOM算法在3.1中已有详细的描述）：

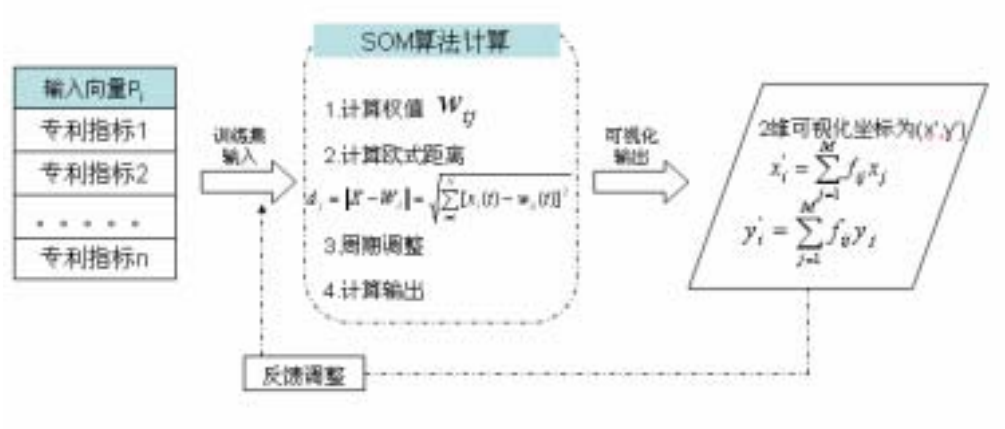

图 4- 4 基于 SOM 的可视化方法设计

对于SOM方法的计算在第3章提出可视化挖掘模型的同时给予了详细的介绍，下面本文将对可视化过程中存在的影响整个可视化结果的所涉及的重要层面给予说明及相应的解决方法。本次论文中所涉及的主要方面有：SOM方法所对应的专利信息输入

层的确认以及输出坐标的计算以及具体可视化方法的实现。

4.2.1 SOM 模型输入层的改进

对于专利信息来讲，我们涉及的更多的是他的文本信息，如何将这些文本信息转化为适合SOM训练集的信息是本文首先需要解决的方面。

通过http://search.sipo.gov.cn/等专利检索网站可以获取相关的专利信息，通过对获取信息的研究可以发现其主要包含专利申请人、发明人、专利名称、摘要、申请日期等等文本信息。而SOM方法本身要求输入的是一种量化的数字，也即输入变量Xi在各个维度上的一个量化值，因此解决数据的量化问题成为本论文开展的先决条件。

通过对专利分析方法的详细研究。本文决定以专利指标作为基本的输入维度。这是因为：专利指标是利用专利数据建立的一种科学技术指标，可以在宏观或微观的不同层面反应国家或企业的发明活动以及产出、科技产权的拥有量、技术发展水平及其在国际技术与经济竞争中的地位。因此应用专利指标作为群组划分的依据是非常必要的。

下面本文将着重描述主要应用的专利指标：

专利数量：通过专利数量可以进行各国不同时期、不同领域技术活动产出和谋求工业产权保护意向的比较。

专利成长率：测算专利数量成长随时间变化的百分率。

引证指标：测算的是一个专利被气候申请所引用的次数，当一个专利多次被后申请专利所引用就说明该项被引用专利技术在该产业具有很重要的技术先进性。

平均申请人数量：测算某领域国家或企业平均每件申请专利中的申请人数量，从而从侧面反映这个实体与其它的实体的关联情况。该值越高则反映出的实体间的联系越强。

研发机构投入/产出情况：主要衡量的是机构的研发团队的投入、产出情况。以平均每件专利的参加人数来描述。该值越高则表现的是机构的投入越大。现在的科技已经不能靠单兵作战来完成研发任务，只有依靠由多人组织的研发团队才能胜任。

专利授权率：体现公司申请的专利中具有实际应用价值的比率。

专利寿命分析：专利权人获得授权后，必须交纳费用才能维持其专利权。因此，

从经济的角度来讲，只有当维持专利所带来的收益大于其成本时，专利权人才会继续维护，这也使得我们可以通过专利的维持年限来侧面了解专利品质。本文对于专利寿命分析主要采用的是专利存活率指标，也即未实效专利占申请专利的百分比。

4.2.2 可视化图谱二维坐标计算

通过SOM这种方法，可以很方便的对企业的竞争态势，竞争领域的具体竞争情况有一个详细的认识。但是传统的SOM方式仅仅是将数据进行简单的分组，如何将这些差异和相似点展现出来也是本文关注的重点之一。经过对聚类方法的进一步研究，以及对其实现平台，如Matlab等观察，在这里本文将采用一种“距离映射法”来实现对其坐标的获取。

距离映射法指的是根据输入向量与神经元权向量的距离大小来定义两者的相似度[30]。采用这种方式的理论基础是因为SOM网络的竞争学习是根据输入向量与竞争层神经元权向量的欧式距离大小来评价的，而两个向量的相似程度与两者距离值也密切相关，距离越接近，向量越相似。另外，SOM网络通过对输入向量的反复学习，其权向量的空间分布能反映输入向量的统计特征。训练结束后，状态相同或相近的输入向量 $p_1, p_2, \ldots, p_N$ 在竞争层网络上所处位置和邻近，采用距离映射法在二维平面上得到的映射点 (x_i', y_i') 也相邻近。这也是本文选择其作为确定输出坐标方法的重要原因。

在距离映射法中，本文所采用的SOM竞争层网络结构为六边形网络结构，具体如下图所示：

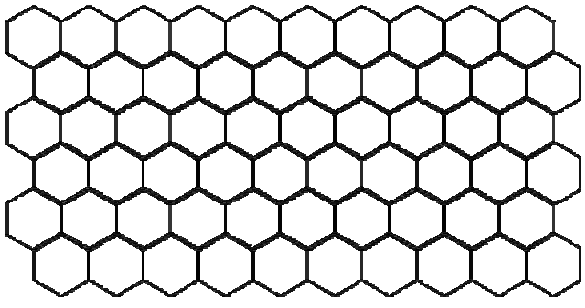

图4- 5六边形SOM网络结构

在此基础上，我们借助欧氏距离进行一系列的计算来获得最终的二维坐标，计算过程如下：

1．设有输入向量 $p_1, p_2, \ldots, p_N$，其中 $p_i \in R^n$ ；

2．在六边形网络结构中，网格点表示竞争层神经元。设共有M个神经元，其二维坐标为$(x_j, y_j), j = 1,...M$，对应权向量为$q_j, j = 1,....M$，且$q_j \in R^n$，对于输入量p_i，首先计算它与神经元j权值q_j的距离$d_{ij} = \sqrt{(p_i - q_j)^T (p_i - q_j)}$；

3．定义p_i与神经元j的相似度f_{ij}：

$$f_{ij} = \frac{\exp(-d_{ij})}{\sum_{k=1}^{M} \exp(-d_{ik})}$$

其中p_i与所有神经元的相似度之和为1。上式中相似度f_{ij}的定义为只需满足距离d_{ij}越小，f_{ij}越大，且d_{ij}可以为0即可。

4.获得映射点坐标$(x_i^{'}, y_i^{'})$：对所有相似度与对应神经元坐标值之积求和，得到p_i在二维平面上的映射点坐标$(x_i^{'}, y_i^{'})$；

$$x_i^{'} = \sum_{j=1}^{M} f_{ij} x_j \text{，} \quad y_i^{'} = \sum_{j=1}^{M} f_{ij} y_j$$

通过上面的步骤，可以得到经过SOM方法后的输出点的坐标，为下面的可视化展示提供了基础。

4.2.3 可视化图谱设计原则

事实上，数据挖掘不等同与数据查询，数据查询是向用户展示已经存储在数据库中已知的信息，而数据挖掘则是将一个事先并没有明确存储的数据记录之间不直观的关系展现出来。因此，将系统得到的结果转化为一个好的表达形式就成为一个很重要的问题。

一般情况下，可视化模型的展现要满足以下原则：

（1）表现简单直观：尽可能多的使用用户已经熟悉的视觉元素展现，这样才能让用户很快了解并挖掘分析方法或参数。

（2）良好的交互：使用户可以实时的对模型采取行动。

结合上述模型的挖掘后的特点结合上述原则，本文对专利可视化的模型设计和实现主要分为以下两个部分：

（1）宏观层面竞争体系可视化

从宏观层面讲，对于公司间、发明人间等的竞争态势在以往的分析方法中并没有很好的体现。本文希望通过一种比较合理的图形来展现其竞争关系，如等高线图----

处于同一平面的点展现了其研究的相关性，也就是说公司或个人间的竞争态势严重；而处于不同平面的则竞争相关性弱。

（2）微观层面竞争差异可视化

尽管宏观层面上的竞争关系得到了很好的体现，但是处于竞争关系的实体间具体什么领域竞争力强，什么地方较弱，则很难表现。为此，除了宏观层面的竞争态势分析外，本文还需要在微观领域方面进行进一步的研究。如，雷达图：不同的维度表示不同的领域，雷达图中的线条则表现的是具体的某个公司或发明人等信息，则在各个维度方面的竞争情况可以很清楚的表现出来。

在人类的神经系统及脑的研究中，人们发现：人脑的某些区域对某种信息或感觉敏感，如人脑的某一部分进行机械记忆特别有效；而某一部分进行抽象思维特别有效。这种情况使人们对大脑的作用的整体性与局部性特征有所认识。对大脑的研究说明，大脑是由大量协同作用的神经元群体组成的。而在大脑处理信息的过程中，聚类是极其重要的功能。

自组织语义映射网络算法SOM是一种聚类算法，它能根据其学习规则对输入的模式进行自动分类，即再在无监督的情况下，对输入模式进行自组织学习，通过反复地调整连接权重系数，最终使得这些系数反映出输入样本之间地相互关系，并在竞争层将分类结果表示出来。而可视化是SOM 用于数据分析的一个重要原因。SOM 通常用来对高维输入数据进行可视化分析，可把任意高维输入数据变换到二维的网格映照图，并保持一定的拓扑有序性。一个自组织映射可以用多种方式实现可视化。其可视化的目标是分析SOM 的输出层。通过对其输出层的分析，发现类别相似的神经元靠得比较近，类别不相似的神经元分得比较开，因而它在数据分析中具有独特而又强大的功能。

考虑到本论文分析的目的：在专利层面，公司、国家间的竞争合作关系。因此本次可视化采用了等高线图形。处于同一等高线领域的公司或国家在专利指标方面处于同一程度；而聚集在同一区域中的公司或国家则关系联系紧密，需要进一步分析挖掘其竞争、合作关系。

4.2.4 SOM 可视化方法研究

（1）SOM算法

为了实现SOM算法在专利领域的应用，本文采用C#方法编写了如下几个重要的方

法：

A 从文件读入训练集数据：

作用：将训练集数据构造符合需要的节点数，为下一步的工作做准备。

```
public void ReadDataFromFile(string inputDataFileName)
{
    StreamReader sr = new StreamReader(inputDataFileName);
    string line = sr.ReadLine();
    int k = 0;
    for (int i = 0; i < line.Length; i++)
    {   if (line[i] == ' ') k++;   }

    inputLayerDimension = k;
    int sigma0 = outputLayerDimension;

    outputLayer = new Neuron[outputLayerDimension, outputLayerDimension];

    Random r = new Random();
    for (int i = 0; i < outputLayerDimension; i++)
        for (int j = 0; j < outputLayerDimension; j++)
        {   outputLayer[i, j] = new Neuron(i, j, sigma0);
            outputLayer[i, j].Weights = new List<double>(inputLayerDimension);
            for (k = 0; k < inputLayerDimension; k++)
            {  outputLayer[i, j].Weights.Add(r.NextDouble()); }
        }

    k = 0;
    while (line != null)
    {   line = sr.ReadLine();
        k++;          }
    patterns = new List<List<double>>(k);
    classes = new List<string>(k);
    numberOfPatterns = k;

    List<double> pattern;

    sr = new StreamReader(inputDataFileName);
    line = sr.ReadLine();

    while (line != null)
```

B 训练集位置坐标的计算

下面代码的含义是：通过将多维向量数据根据SOM算法生成相应的二维坐标值，为可视化进行奠定基础(下面的代码只是主要函数的截图，并不完全)。

```
public void StartLearning()
{
    int iterations = 0;
    while (iterations<=numberOfIterations && currentEpsilon > epsilon)
    {
        List<List<double>> patternsToLearn = new List<List<double>>(numberOfPatterns);
        foreach (List<double> pArray in patterns)
            patternsToLearn.Add(pArray);
        Random randomPattern = new Random();
        List<double> pattern = new List<double>(inputLayerDimension);
        for (int i = 0; i < numberOfPatterns; i++)
        {
            pattern = patternsToLearn[randomPattern.Next(numberOfPatterns - i)];

            StartEpoch(pattern);

            patternsToLearn.Remove(pattern);
        }
        iterations++;
        OnEndIterationEvent(new EventArgs());
    }
}
```

C 确定不同权值可视化的色彩分布

作用：由于本文采用的可视化方法图形是三维等高线，对于颜色的分布是主要的一个研究要点。

```
public System.Drawing.Color[,] ColorSOFM()
{
    System.Drawing.Color[,] colorMatrix = new System.Drawing.Color[outputLayerDimension, outputLayerDimension];
    int numOfClasses = NumberOfClasses();
    List<System.Drawing.Color> goodColors = new List<System.Drawing.Color>();
    goodColors.Add(System.Drawing.Color.Black);
    goodColors.Add(System.Drawing.Color.Red);
    goodColors.Add(System.Drawing.Color.Navy);
    goodColors.Add(System.Drawing.Color.Green);
    goodColors.Add(System.Drawing.Color.Yellow);
    usedColors = new List<System.Drawing.Color>(numOfClasses);
    usedColors.Add(goodColors[0]);
    int k = 0;
    int randomColor = 0;
    Random r = new Random();
    while (usedColors.Count != numOfClasses)
    {
        k = 0;
        randomColor = r.Next(goodColors.Count);
        foreach (System.Drawing.Color cl in usedColors)
            if (cl == goodColors[randomColor]) k++;
        if (k == 0) usedColors.Add(goodColors[randomColor]);
    }
    for (int i = 0; i < outputLayerDimension; i++)
        for (int j = 0; j < outputLayerDimension; j++)
            colorMatrix[i, j] = System.Drawing.Color.FromKnownColor(System.Drawing.KnownColor.ButtonFace);

    for (int i = 0; i < patterns.Count; i++)
    {
       Neuron n = FindWinner(patterns[i]);
        colorMatrix[n.Coordinate.X,n.Coordinate.Y] = usedColors[existentClasses[classes[i]]-1];
    }
    return colorMatrix;
}
```

（2）可视化算法

```
static COLORREF Colors[11] = { RGB(127, 127, 127), RGB(0, 0, 127), RGB(0, 0, 255), RGB(0, 127, 127), RGB(0, 255, 0),
                               RGB(127, 127, 0), RGB(255, 255, 0), RGB(0, 255, 255), RGB(255, 0, 255),
                               RGB(255, 0, 0) };

const int FLAT_HEIGHT = 600;
const int LEGEND_WIDTH = 200;

/////////////////////////////////////////////////////////////////////////////
// CContourView

IMPLEMENT_DYNCREATE(CContourView, CView)

CContourView::CContourView() : m_pImage(NULL), m_bImageReady(false), m_nZoom(3), m_nBaseWidth(200), m_nBaseHeight(200),
m_nNumOfLevels(11), m_bLegend(true), m_bValues(true), m_dThickMax(0), m_dThickMin(0)
{
}

CContourView::~CContourView()
{
        if(m_pImage)
                delete m_pImage;
}


BEGIN_MESSAGE_MAP(CContourView, CView)
        //{{AFX_MSG_MAP(CContourView)
        ON_COMMAND(ID_SHOWVALUES, OnShowvalues)
        ON_COMMAND(ID_SHOWLEGEND, OnShowlegend)
        ON_UPDATE_COMMAND_UI(ID_SHOWVALUES, OnUpdateShowvalues)
        ON_UPDATE_COMMAND_UI(ID_SHOWLEGEND, OnUpdateShowlegend)
        ON_COMMAND(ID_EDIT_COPY, OnEditCopy)
        ON_COMMAND(ID_FILE_SAVE, OnFileSave)
        //}}AFX_MSG_MAP
END_MESSAGE_MAP()

/////////////////////////////////////////////////////////////////////////////
// CContourView drawing

void CContourView::OnDraw(CDC* pDC)
```

定义图形的最后输出形式：通过静态变量colors对等高线输出色彩的控制，flat_height，legend_width是对图形结果输入高度和宽度的界限的设置，此外结合其他相应变量一起定义了输出的最后形式。

函数OnInitialUpdate()是本次研究的可视化图谱中的重要函数之一，主要进行结果的输出，其中调用了许多其他以定义函数，在这里并没有展现出来。

（3）SOM图形生成过程

在可视化展示方面，为了使结果更加直观，信息表述更为清晰易懂，本文采用了等高的形式来表示。等高线的选择主要是根据其特有的性质决定的：

- 位于同一等高线上的地面点，海拔高度相同。因此位于同一等高线区域中的国家从指标层面讲属于同一等级的，从视觉方面来讲，落在同一颜色的区域表示专利的质量等级相同。
- 在图廓内相邻等高线的高差一般是相同的，因此地面坡度与等高线之间的水平距离成反比，相邻等高线水平距离愈小，等高线排列越密；相邻等高线之间的水平距离愈大，等高线排列越稀。

下面以专利增长率和专利数量两个指标来描述专利可视化的过程，具体过程如下

图所示：

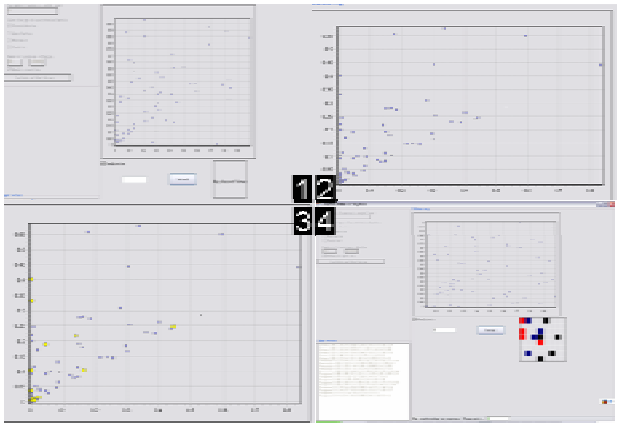

图4- 6 SOM图形生成过程展示

首先通过各指标的重要程度确定相应的权值 q_j ;然后根据归一化的输入量 p_i 和权值输入，计算它与神经元的距离：$d_{ij}=\sqrt{(p_i-q_j)^T(p_i-q_j)}$ （i表示目标神经元，j为除外的其他神经元）；接着定义神经元i与神经元j的相似度 f_{ij} （计算方法前面已有描述）；最后得到二维图形的坐标 $x_i^{'}=\sum_{j=1}^{M}f_{ij}x_j$ ， $y_i^{'}=\sum_{j=1}^{M}f_{ij}y_j$ 。根据神经元间的相似程度 f_{ij} 确定区域的颜色，相似程度越高则色彩越近，如上图聚类的描述，最终生成所需的等高线图形。

上图展现了SOM图形的生成过程：1>输入的训练集按照SOM算法进行相应的聚类计算；2>输出层二维坐标的计算；3>输出层的展现以及色彩的计算；4>等高线位置的定义及展现。

最后生成的图形效果如下所示：

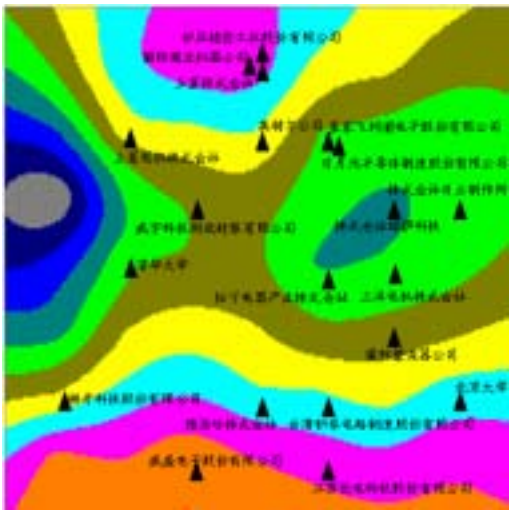

图4- 7可视化效果图

从上面的等高线图谱中，可以发现两个主要结论：（1）专利品质的分类：可以从色彩方面区分专利品质的类型，相同等级的品质处于同一等高线上也即颜色相同的区域，且等高线的颜色越明艳则表示专利的品质越好。（2）研究实体的关联程度：从指标层面来讲，聚集程度越紧密则关联性越强，通过对其进一步分析可以发现相关联的竞争合作关系，从而为企业制定相应战略提供依据。

4.3 本章小结

基于前一章提出的基于SOM的可视化模型，本章就其中涉及的主要方法进行研究。数据获取方面，主要利用北京理工大学知识发现与数据分析实验室自主开发的专利获取系统，获得所需专题数据；数据提取：利用DOM树原则通过C#.NET开发专利信息提出程序来获得结构化的数据集。可视化方法是本章的核心，主要就SOM方法在专利数据挖掘中所遇到的问题进行研究，最后通过C#.NET语言自行开发程序实现可视化挖掘。

5 实证分析：集成电路封装技术可视化挖掘

5.1 实证分析背景介绍

5.1.1 集成电路封装技术介绍

集成电路封装技术是一种将集成电路用绝缘的塑料、陶瓷或其他材料打包的技术。由于芯片必须与外界隔离，以防止空气中的杂质腐蚀芯片电路而造成电气性能下降，因此封装对于芯片来说至关重要。封装有时也指安装半导体集成电路芯片用的外壳，它不仅起着安放、固定、密封、保护芯片和增强导热性能的作用，而且还是沟通芯片内部世界与外部电路的桥梁——芯片上的接点用导线连接到封装外壳的引脚上，这些引脚又通过印刷电路板上的导线与其他器件建立连接。因此，对于很多集成电路产品而言，封装技术都是非常关键的一环。

5.1.2 选取集成电路封装技术的意义

集成电路行业作为高技术行业的典型代表，是一个强调“核心技术即企业生命”的行业，其发展速度之快、影响之深是其它高技术行业不能比拟的，现在集成电路产业已经成为其他高技术产业的基础，推动着其他产业的发展。所以，研究集成电路行业具有特殊的意义。

本文之所以以集成电路封装技术为例，除了以往拥有的实验室集成电路研究经验以外，还是因为集成电路封装技术的重要性，如下所述：

在中国集成电路产业的发展中，封装测试行业虽不像设计和芯片制造业的高速发展那样抢眼，但也一直保持着稳定增长的势头。特别是近几年来，随着国内本土封装测试企业的快速成长以及国外半导体公司向国内大举转移封装测试能力，中国的集成电路封装测试行业更是充满生机。2001年到2005年，国内封装测试业销售收入由142.9亿元扩大到344.9亿元，年均增长率达到19.3%。到2005年底，国内集成电路封测企业已经超过100家，从业人员超过3万人，年封装能力达到300亿只。从产业规模看，封测业是国内集成电路产业的主体，其销售收入一直占产业整体规模的70%以上，封装

测试行业销售收入的增长对国内集成电路产业整体规模的扩大起着极大的带动作用。近几年，随着IC设计和芯片制造行业的迅猛发展，国内集成电路价值链格局正在发生改变，其趋势是设计业和芯片制造业所占比重迅速上升，封测业比重则逐步下降，但封测行业仍占据国内集成电路产业的半壁江山。

5.2 数据获取

5.2.1 数据源选择

本文的研究是建立在中国专利和中国法律状态数据库基础上的，所以本次实证采用的数据为1985年至2007年的中国专利数据和相应的法律状态数据。专利数据的获取来自（http://www.sipo.gov.cn/sipo/zljs/default.htm）中国国家知识产权局，采用的信息检索方式为（名称+摘要）关键词检索；法律状态数据的获取是依托已获得的专利数据信息的专利号进一步检索其对应的法律状态信息。

5.2.2 专利信息获取与存储

本次研究的数据主要基于中国专利信息和法律状态数据，具体的获取方法在4.1.1数据获取中有详细的描述，这里不再繁述。为了更好的存储抽取出的信息，本文设计了中国专利本地化的数据表结构。设计的原则：需要提取的字段以及字段在html中的结构，专利数据库数据表设计如下表所示：

表格 5- 1 中国专利数据库设计

字段名称	类型	含义
ID	INT	数据表的主键
ApplyNO	nvarchar	专利号
ApplyName	nvarchar	专利名称
abstract	ntext	摘要信息
MainIPC	nvarchar	主IPC
patentee	nvarchar	专利权人
PatenteeCity	nvarchar	专利权人地址
inventor	nvarchar	发明人
appdate	datetime	申请日期
granteddate	datetime	专利授权日期

与中国专利数据库相比，法律状态数据库的字段设计除了抽取必须信息所需字段

外，另外增加了PatentNum、LegIndex两个字段。这是因为对于每条专利数据来讲，对应着若干条相应的法律状态信息，假设集成电路封装技术中国专利库中有n条相关信息，每条对应m条法律状态信息，则在进行数据库整合中对数据的查询需要进行n*n*m次，查询效率低下。而通过增加PatentNum（专利索引号）、LegIndex（法律状态索引号）两个字段后，查询次数减为（n+m）次，大大提高了整合的效率。

5.3 数据预处理

5.3.1 数据清洗

本文在数据获取阶段的原则是尽可能的全，这也导致了现实中的数据的不完全和不一致性，带有许多噪声数据，因此在完成了业务数据集的选择之后，在进行可视化挖掘之前必须保证数据的质量。

具体采取的步骤如下：

- 数据去重

根据专利号将同一技术领域的不同关键词查询获取到的专利文献数据进行一次清洗，对前面已经存储过的数据不再进行存储。

- 噪声数据清洗

本文对噪声数据的处理主要是采取IPC分类方法，解释如下：

以下面两条专利信息为例来阐述本文噪声数据处理方法：上面编号为20、24的两条数据是按关键词筛选出的，也就是说满足数据获取的条件；但是仔细观察发现，第20条数据的IPC为A01B（含义：农业或林业的整地；一般农业机械或农具的部件、零件或附件）显然不属于集成电路封装技术领域数据，而24条数据IPC为H01L（半导体器件；其他类目未包括的电固体器件）才是应该获取的数据内容。通过这种IPC的筛选进一步除掉噪声数据的干扰。

20	96196192	用于网格绿球...	一种用于具有...	A01B	PCD公司	0	詹姆斯·M·拉姆...
24	97107274	用于多芯片封...	多芯片封装体...	H01L	三星航空产业...	0	安志洙、柳在哲

- 空值数据处理

5.3.2 属性转换

经过前面一系列的数据处理工作后，数据库中现一存在大量搞质量的待挖掘的数据，但是要想转化为本文所需的信息，还需要经过一定的前期挖掘工作，也即在已有数据属性基础上发据出满足要求的信息的属性。这种数据格式的转化主要涉及了表级别的属性扩展和字段的逻辑转换。

（1）属性扩展

经过数据清洗之后，虽然得到了高质量的集成电路封装技术领域的干净数据，然后对于数据挖掘来讲仅仅依靠这些表面数据很难得到本文想要的结果，因此需要进行进一步的处理。表级别的属性转换则影响到了整个待挖掘的业务数据表的结果和数量，影响了可视化结果的呈现。

考虑到数据挖掘的目的，本文主要扩展了以下属性，扩展原因如下表所示：

表格 5- 2 扩展的属性

属性	原因
专利授权情况	只有经过授权的专利才享有国家专利政策的保护；也即专利的授权与否直接影响着专利的价值；对于公司等研究机构来讲，专利的授权情况也代表着机构对该领域的专利优势。
国家	专利权人所属国家：主要根据专利权人的地址获取。通过专利权人所属国家，可以看出各国之间的在领域研究情况以及研究的关联、竞争情况。
专利权人类型	主要分为：个人、公司、大学、研究所四类。通过属性分析可以看出领域的研究主题状况。
发明人个数	现如今是一个竞争合作的社会，几乎每一件专利都是集体创作的结晶。同时专利信息中发明人的数量也代表着研究机构对领域技术的一个投入情况。
申请人个数	申请人代表着机构之间的合作关系；如果企业间经常联合申请专利，则更好的表明了其间的战略伙伴关系。

（2）字段级别的格式转换

字段级别的格式转化影响着整个待挖掘业务数据集表中的字段的结构。为了提高可视化挖掘的效率，需要对一些字段的类型进行一定的格式转换。

5.3.3 数据库整合

图5- 1数据整合

本次所涉及的数据库主要是中国专利数据库和法律状态数据库，两者之间是以专利号为连接依据的。同时对于一条专利信息来讲，可能对应着多条法律状态信息，这是因为：法律状态信息是反应的是专利信息在时间点上的具体状态情况，专利的授权情况不是一成不变的，可能会随着时间点的变迁有所改变，这些对于数据整合来讲都是需要考虑的。

数据整合过程具体如下：

（1）规划整合后的数据格式，构建所需数据表，主要从两个库的数据库格式和需要分析的内容考虑构建。

（2）将经过数据清洗后的中国专利数库信息逐条查询，根据查询出的专利信息根据专利号查询相关的法律状态信息：a.若查询出0条记录，则相关法律状态的数据直接填空值；b.若查询出的数据大于1条，则根据公告时间情况将最新的法律状态信息更新入库；c.若查询出的数据只有一条，则直接更新入库。

（3）根据（2）所得的数据，更新数据，则得到了更新好的数据库。

5.4 训练集生成

经过5.3的处理已经得到了干净的可以进行数据分析的集成电路封装技术领域数据集，但是由于SOM方法要求的特殊性，需要将领域数据集进行进一步的计算以生成

符合可视化要求的训练集。

5.4.1 国家层面 SOM 训练集的生成

通过针对投入此产业技术的竞争国家进行相关分析，可看出领先国家和潜在进入者。但以往的分析仅仅是从专利的申请量角度出发，以数量为基础确定该领域内国家的实力情况，而这显然是不充分的。同时专利的法律状态数据也蕴含着大量不容忽视的信息，不仅在专利引进、专利转让和组建专利联盟中起着至关重要的作用，对其进行分析还可以获取许多有价值的情报，比如可以通过专利权的维持年限来了解专利的品质与质量等，常见的法律状态信息有公开、实质审查的生效和授权等。因此本文结合专利本身的特性以及对于国家我们研究的侧重方面，我们选取了如下指标，具体说明如下表所示：

表格 5- 3 国家领域训练集生成原则

指标名称	计算方法	意义
专利数	各国每年申请专利数的总和	通过专利数量可以进行各个国家不同时期、不同领域技术活动产出和谋求工业产权保护意向的比较。专利数的高低可以反应企业或研究实体对于领域的重视及投入情况，也从一个侧面反应了企业在该领域的实力。
授权率	各年授权专利总数/总的专利数	体现公司申请的专利中具有实际应用价值的比率。只有经过授权的专利才享有国家对于实体的各种专利保护措施，因此授权率的高低体现了实体专利申请质量的高低。授权率越高，企业专利申请的质量越高。
平均申请人	专利的申请人总数/专利总数	测算某领域国家或企业平均每件申请专利中的申请人数量，从而从侧面反映这个实体与其它的实体的关联情况。该值越高则反映出的实体间的联系越强。
投入产出比	专利的发明人总数/专利总数	主要衡量的是机构的研发团队的投入、产出情况。以平均每件专利的参加人数来描述。该值越高则表现的是机构的投入越大。现在的科技已经不能靠单兵作战来完成研发任务，只有依靠由多人组织的研发团队才能胜任。
专利成长率	(某时期专利申请数 - 上一时期专利申请数)/上一时期专利申请数	测算专利数量成长随时间变化的百分率，可显现技术创新随时间变化是增加还是迟缓。
专利存活率	(申请专利总数 - 失效专利总	失效专利技术虽然失去了专利法的保护，对于研究实体来讲失去了获取利益的条件，这是因为不用支付昂贵的专利转让费，可以

	数)/申请专利总数	无偿地进行改进；因此只要当维持专利所带来的收益大于其成本时，专利权人才会继续维护以获取利润
专利平均维持年数	领域专利维持总年份/申请专利总数	专利权人获得授权后，必须交纳费用才能维持其专利权。因此，从经济的角度来讲，只有当维持专利所带来的收益大于其成本时，专利权人才会继续维护，这也使得我们可以通过专利的维持年限来侧面了解专利品质。因此专利的维持年数也反应了专利对于实体的重要程度。
权利人数	领域某国家申请专利中共有的申请实体个数	尽管专利申请量的高低反映了国家在该领域的技术水平，但是如果大量的专利申请集中在少数几家公司/企业，则势必会引起行业的垄断情况，因此该值越高则反应了行业的均衡发展；反之越小则越体现着垄断程度较高。
发明专利比例	申请的发明专利数/专利申请总数	专利法意义上所说的发明有特定的含义，发明是发明人的一种思想，是利用自然规律解决实践中特定问题的技术方案。我国专利法实施细则规定，专利法上所称的发明，是指对产品、方法或者其改进所提出的新的技术方案。主要包括产品发明和方法发明两类。由于发明是可以产生一种全新的产品或者方法的技术方案，是科技含量和创造性都较高的一种发明创造，因此，各国专利法都将发明作为专利保护的基本对象。因此发明专利比例越高，则表明创新力越强。

5.4.2 企业层面 SOM 训练集生成

利用专利数据对特定竞争对手进行各种竞争指标分析，可以找出领导公司、潜在竞争者、竞争对手进入或退出相关领域的动向等，了解主要竞争对手的专利申请趋势、技术构成和地域分布等。

由于国家和企业本质的不同，因此在指标选择方面也会有所侧重与不同。与国家层面的不同在于：缺少了权利人数，增加了独立申请率指标。由于分析是站在单一企业层面的，因此该指标失去了其意义；而增加独立申请率指标则是考虑到很多情况下，企业间都是联合申请专利的，而当企业是该专利唯一权利人时则享有了该专利所带来的所有收益，因此独立申请率也从侧面反应了企业在该领域的优势情况。计算如下：独立授权率 = 独立申请的专利数/申请的总专利数。

5.4.3 数据归一化

由于每个专利指标的衡量单位不一致，如专利数量以件数为单位，而授权率则以百分比为衡量标准，因此如果仅仅以上式所得的数据进行处理的话，势必带来各指标对于结果影响的不一致。

如下图形所示：

图 5- 2 未处理指标单位带来的影响

以美国、日本数据中的专利数和授权率为例说明：美国的专利数为104件，日本未80件，从数量上来讲美国是日本专利数的1.3倍；授权率方面，美国为0.23，而日本则达到0.41，是美国的将近两倍，如果对这两个指标的权重我们都设为0.1，也就是说两个指标对于最后结果的影响是一样的。抛开权重的计算公式，仅仅从上述对指标的分析来看，认为日本的总体专利质量应该是高于美国的。但是从权重的计算结果却并非如此：权重= 专利数*0.1 + 授权率*0.1，结果得出美国的专利质量要远远高于日本。

为了避免上述问题的出现，本文引入了归一化的概念。归一化是指原属性值表中不同指标的属性值的大小差别很大，为了直观，更为了便于各种多属性评估，需要把属性表中的数据归一化，即把表中数均变换倒[0,1]区间中去。

常用的归一化方法很多，归纳为：

（1）线性变换

设原始的决策矩阵为Y={y_{ij}}，变换后的矩阵记为Z={z_{ij}}，i=1,...,m，j=1,...,n。设 $y_i^{\max}$，$y_j^{\min}$ 分别是决策矩阵第j列中的最大值和最小值。

若i为效益型属性，则 $z_{ij} = y_{ij} / y_j^{\max}$；

若j为效益型属性，则 $z_{ij} = 1 - y_{ij} / y_j^{\max}$。

（2）标准0-1变换

为使每个属性变换后的最优值为1且最差值为0，可以进行0-1的标准变换。当j代表的是效益性指标（即值越大表示结果越好）时，令：

$$z_{ij} = \frac{y_{ij} - y_j^{\min}}{y_j^{\max} - y_j^{\min}}$$

当j代表成本型属性（值越大结果越差）时，令：

$$z_{ij}=\frac{y^{\max}-y_{ij}}{y_j^{\max}-y_j^{\min}}$$

（3）最优值为给定区间时的变换

对于属性既非效益型又非成本型的情况，这种属性不能采用前面介绍的两种处理。设给定的最优属性区间为$\left[y_j^0,y_j^*\right]$，$y_j^{'}$为无法容忍下限，$y_j^{''}$为无法容忍上限，则

$$z_{ij}=\begin{cases}\dfrac{y_j^0-y_{ij}}{y_j^0-y_j^{'}} & y_j^{'}<y_{ij}<y_j^0\\ 1 & y_j^0\le y_{ij}\le y_j^*\\ 1-\dfrac{y_{ij}-y_j^*}{y_j^{''}-y_j^{'}} & y_j^{''}>y_{ij}>y_j^*\end{cases}$$

（4）向量的规范化

无论成本型属性还是效益型属性，向量规范化均用下式进行变换：

$$z_{ij}=y_{ij}\Big/\sqrt{\sum_{i=1}^{m}y_{ij}^2}$$

这种变换也是线性的，它的最大特点是规范化后，各方案的同一属性值的平方和为1，因此常用于计算各方案与某种虚拟方案的欧式距离的场合。

（5）原始数据的统计处理

有时某个目标的各方案属性值往往相差极大，或者由于某种特殊原因只有某个方案特别突出。如果按一般方法对这些数据进行预处理，该属性在评价中的作用将被不适当的夸大，会使整个评估结果发生严重扭曲。为此可以采用类似于评分法的统计平均方法。做法之一是设定一个百分制平均值M，将方案集X中各方案该属性的均值定位于M，再用下式进行变换：

$$z_{ij}=\frac{y_{ij}-\bar{y}_j}{y_j^{\max}-\bar{y}_j}(1-M)+M\ ；$$

其中，$\bar{y}_j=\frac{1}{m}\sum_{i=1}^{m}y_{ij}$ 是各方案属性 j 的均值，m 为方案个数，M 的取值可在 0.5-0.75 之间。

（6）专家打分的预处理

有时某些性能指标很难或根本不能用适当的统计数据来衡量优劣，为了使评价尽可能客观、公正，通常要请若干个同行专家对被评价对象按指标打分，再用各专家打

分的平均值作为相应指标的属性并据此确定被评价对象的优劣。

假设被邀请的各位专家意见的重要性相同，则每个专家在评价中理应发挥同样的作用，但是对同一批被评价对象的同一指标，由于不同专家的打分习惯不同，所给分值所在区间往往会有很大差别。

比如，专家甲的打分范围再50到95之间，而专家乙的打分范围再75到90之间。如果不对专家所打出的原始分值进行处理直接计算平均值，则专家甲在评价中所起的实际作用将是专家乙的3倍。

为了改变这种无形中造成的各专家意见重要性不同的状况，使各位专家的意见再评价中起同样重要的作用，应该把所有专家的打分值规范化到相同的分值区间[M^0，M^*]。M^0和M^*的选值不同对评价结果并无影响，只要所有专家的打分值都规范到该区间就行。具体算法为：

$$z_{ij} = M^0 + (M^* - M^0)\frac{y_{ij} - y_j^{\min}}{y_j^{\max} - y_j^{\min}}$$

考虑到数据预处理的方便性与可行性，本文采用了（2）标准[0,1]变化方式对数据进行归一化处理。

5.5 SOM 方法可视化

5.5.1 国家层面分析

5.5.1.1 以往分析方法

以往对于国家的分析方法更多的采用了统计、指标等简单的分析方法，为能发掘出隐藏的内部信息。

具体的分析方法如下所述：

（1）专利数量方面

国家层面的专利数量分析主要是从量上对国家/机构的专利申请情况有个宏观的认识。该方法的局限在于，专利的申请数量虽然可以反应国家对于该领域的申请情况，但是申请量并不能全面的反应专利信息，仅仅依靠申请量得到的结论是片面的，也有可能是不准确的。

如下表是集成电路封装技术中国专利前十高产国家或地区专利数量申请表格。从表中可以得到的结论为：1>专利申请方面：除中国大陆外，美国申请数量最多，为104件，其次是日本，80件，远远领先于其他国家，韩国的优势也比较领先。2>专利授权方面：中国香港的授权率最高，为66.7%。这几个国家中，除了中国大陆发明专利的数量占总数比例较低，其它国家的专利中绝大部分都是发明专利。

表格 5- 4 前十高产国家/地区专利申请数量分布表

机构	专利总数	发明专利	授权专利	授权率
中国大陆	298	191	163	54.7%
美国	104	104	24	23.1%
日本	80	77	33	41.3%
韩国	40	40	11	27.5%
荷兰	8	8	0	0.0%
德国	6	6	0	0.0%
新加坡	4	4	1	25.0%
毛里求斯	4	4	0	0.0%
香港	3	2	2	66.7%
以色列	2	2	0	0.0%

（2）研究分布方面

当然除了上述简单的专利分布外，以往的分析中还会采用时间序列来研究历年的专利走势情况；结合IPC方法对于主要的领域的研究态势的探讨。尽管关注的对象不同，使用的方法不同，但是基本都是以数量统计为基础的，这样就造成了很大的局限性。

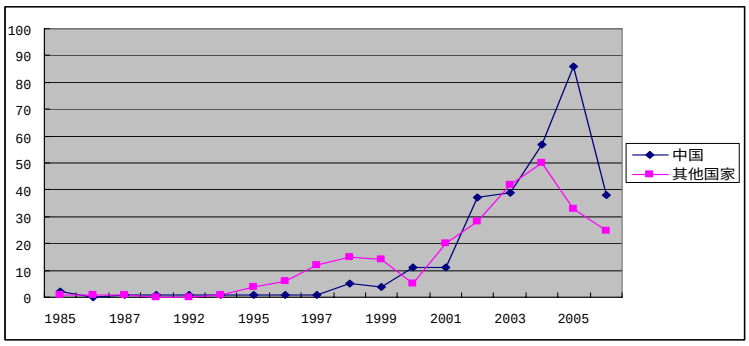

图 5- 3 前十高产国家/地区专利申请年份分布

年度趋势走势图一本的可视化展现方式是以折线图或柱状图表示。图形展示的优点可以概括为如下两个方面：1>从中可以清楚的看出何时是波峰，何时处于波谷，以

及增长或下降速度的情况，从何可以有针对性的进行在分析在处理。2>研究实体间的比较：折线图或柱状体可以一目了然的发现两者之间的实力的比较情况。

通过上面的前十高产国家/地区专利申请年份分布图得出如下结论：中国大陆的专利申请数量和国外的专利申请数量都呈递增趋势，除了看到各国的技术进步之外，还应该对这种现象引起足够的重视，国外企业在对我国采取专利进攻。

为了进一步了解各国的技术研发趋势，对各个国家/地区申请的专利展开主IPC和年份分析。下图和下表体现了IPC的分布情况，从中可以看出，H01L（半导体器件；其他类目未包括的电固体器件）在各国的申请数量都是较多的，在其他各方向上的研究重点则有所不同。

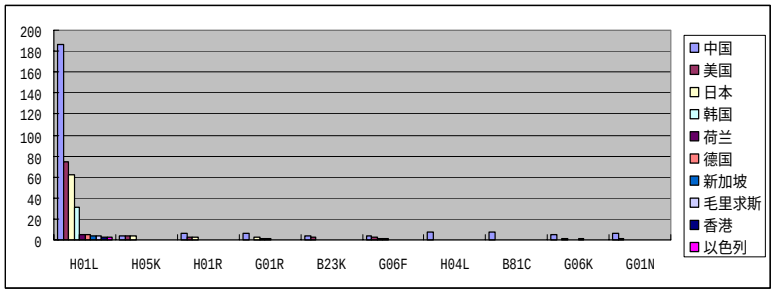

图 5- 4 专利国家分布图

此外对于技术的展现还可以去雷达图的形式展示，不同的节点可以表示不同的技术（以IPC来体现技术领域），以各个研究实体为折现表现在不同技术领域的发展状态。但是用雷达图来展现技术分布走势首先需要明确研究实体的选择，只有选择了合适的研究组群（处于相同的竞争态势组群或是合作关系组群），相应的研究比较才会有意义。因此，做雷达图的前提是对研究实体进行聚类分析。

(3)专利国家维持年份OLAP分析

联机分析处理(OLAP)是共享多维信息的、针对特定问题的联机数据访问和分析的快速软件技术。通过对信息(维数据)的多种可能的观察形式进行快速、稳定一致和交互性的存取，允许管理决策人员对数据进行深入观察。OLAP是商务智能数据分析的核心技术，是数据仓库应用的前端工具，不仅能进行数据汇总/聚集，建立多维度的分析、查询和报表，同时还提供切片、切块、下钻、上卷和旋转等数据分析功能，使分析人员从更快、更易使用的交互方式中获得信息。在本报告中,依托实验室的联机分析处理技术和数据仓库技术快速、灵活地对大量专利数据进行复杂的多维查询处理，并且以图形或者表格的形式直观表示查询分析的结果，实现对专利数据的深入分析。

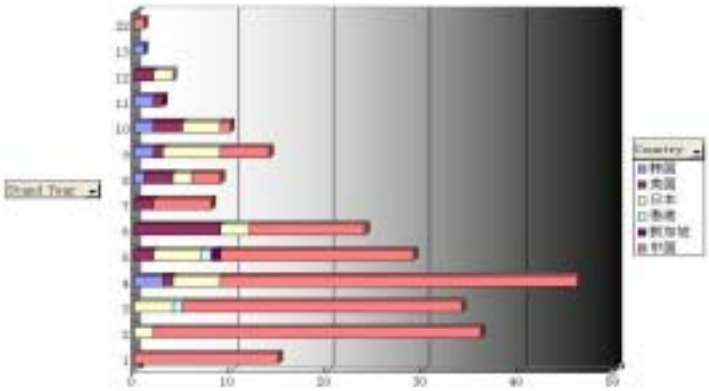

图 5- 5 高产国家维持年份分析图

通过对封装技术中国专利高产国家的专利维持年份进行OLAP分析，得到图3-3所示的结果，可以看出，维持年份较长的专利主要分布在韩国、美国和日本，维持年份为13年的1件专利是韩国的，维持12年的4件专利中，美国和日本各占2件，维持11年的3件专利中，韩国2件，美国1件，这几个国家的专利质量比较高，中国虽然数量最多，但是维持年份不高，说明中国封装技术的专利质量与其他国家还有一定的差距，有待提高。

专利国家维持年份OLAP分析与前面年度趋势分布相比，最大的特点是整合了两个维度的指标，即将专利数量在年份和国家两个方面展开，因此较之仅仅从时间维度来展现专利分布趋势所反应出的信息量更为深刻也更为全面。同时采用多维柱状图作为可视化的呈现方式。其优点在于：1>同一指标间的比较：可以根据不同色彩间的长度清楚的判断实力分布状况；2>不同指标见的比较：可以通过宏观方面来把握研究实体的指标分布状况；对于每个单一的研究实体，也即图中同一颜色表示的柱形在时间维度展开，则是折线图的变相表示。

5.5.1.2 SOM 方法国家层面分析

本文的SOM算法的输入层采用了专利数、授权率、平均申请人数、投入产出比、专利成长率、专利存活率、专利平均维持年数、权利人数、发明专利比例共九个指标，从专利数、授权情况、研发实力投入情况、逐年的发展状况到专利的维持情况等等，全面的反应了专利的性质，从而解决了单一指标信息反应片面的问题。同时，以往的国家层面的专利研究主要集中在矩阵统计层面，因此国家层面专利评价也就集中在了专利数量方面，对于国家层面专利质量的划分并不完善。而在本文的可视化中每一个国家/地区都有不同的属性值（专利指标值）来反映，具有相同的属性值或者相似属

性值的国家/地区自然聚在一起，形成了聚类的基本工作。采用SOM神经网络运算速度快，非监督学习客观性大，非线性问题求解能力强，可以避免主观确定个指标权重的盲目性，也可以避免有不同操作者带来权重赋值的不稳定性，从而对领域专利质量进行客观、合理的分类，也为各国的专利发展指明了方向。

基于前面5.4生成的SOM训练集，利用本文的可视化算法，生成的国家层面的专利分析图谱如下所示：

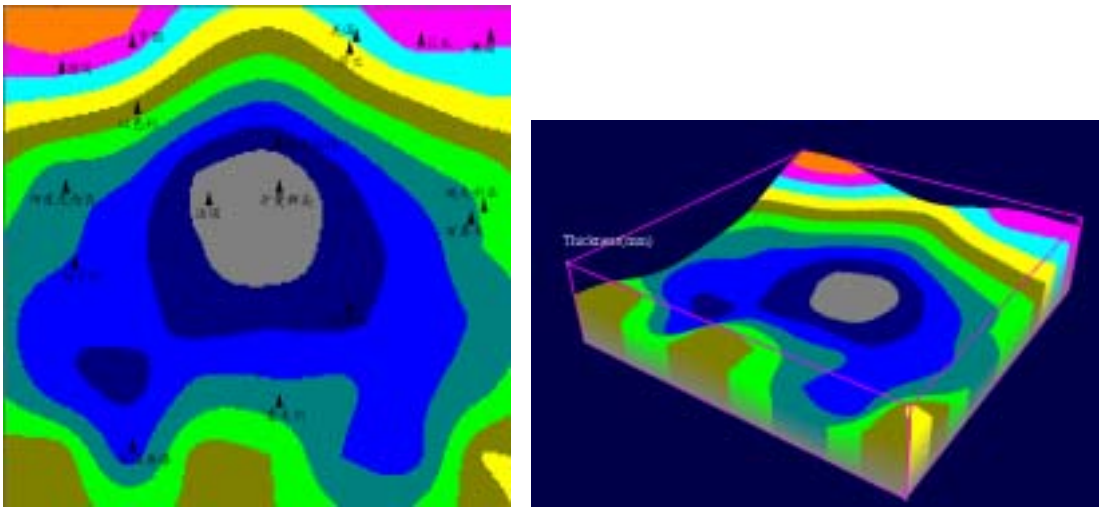

图5- 6 SOM国家层面专利分析

图中颜色越鲜艳表示的专利质量越高，即灰色区域的专利质量<（弱与）深蓝色区域<（弱与）浅蓝色区域<（弱与）墨绿区域<（弱与）草绿区域<（弱与）黄褐色区域<（弱与）明黄色区域<（弱与）紫色区域<（弱与）红色区域。为了更加清晰的显示专利区域的等级情况，本文还绘制了三维的可视化图形。在三维的等高线图上，可以一目了然的看出分布情况。

由于不同国家在中国专利申请的实力相差较大，因此其数据的分布比较松散。将颜色相近的两色归为一类，则国家层面的专利质量分为了5类：

第一类：中国、德国、日本、韩国、美国；

第二类：荷兰、以色列；

第三类：澳大利亚、百慕大、印度尼西亚；

第四类：匈牙利、新加坡、中国香港、毛里求斯；

最后一类：法国、开曼群岛；

这些只是在中国专利申请的情况一个划分，与调研的情况基本类似，只是对于第

三类的划分有些许出入：这主要是因为澳大利亚、百慕大、印度尼西亚这些国家申请的专利数仅为一件，因此其授权率、专利维持率等指标就会相应的较大，造成SOM计算的权值较高，难以完全正确的反映相应状况，这也是本文的SOM方法不可避免的问题。

5.5.1.3 分析方法对比小结

从以上的分析中已经可以发现相较于以往的分析方法，SOM的可视化分析具有如下特点：

- 专利的多指标体现：本文SOM方法输入层囊括了专利数、授权情况、研发实力投入情况、逐年的发展状况到专利的维持情况等等很多方面，全面的反应了专利的性质，从而解决了单一指标信息反应片面的问题。
- 数据挖掘的思想的体现：以往的专利分析主要是集中在矩阵统计分析基础上基于数量的加减乘除的计算。而SOM算法是一种无监督的自组织学习的聚类算法，更好的体现了数据挖掘的思想。
- 专利质量等级的划分：通过SOM方法可以将各国在某领域的实力展现出来，从而为进一步的分析奠定基础。

对于专利的分类，除了上述方法外，还可以采用四象限分析方法（该方法借助管理学领域的市场分析9象限演变而得），也即x轴、y轴分布选用两个重要指标（x轴坐标展现的是公司实力；y轴指标可以反应行业的发展）来表示，根据指标值来确定各点的位置，然后根据专家意见确定划分区域的界限，形成下面展现的四象限图形。

通过一些人为的因素，划分为领先区域、潜在区域、落后区域、活跃区域四个部分。领先区域：两个指标的值都比较高，出去该区域的国家则处于领域发现的领先区

域，掌握着领域的主要技术；活跃区域：表示研究实体在该领域投入的比较高，然而在行业的发展一般，对于这类实体需要改进自己的技术发展，以期提升行业实力；潜在领域：表示研究实体在行业领域的发展程度较好但是演示实体的相对投入不足，只要加大投入实力比较容易进去行业领先状况。最后一部分则是落后领域：研究实体在两个方面的数值都比较低，处于行业的落后水平。

按照上述的四象限分析，本文对集成电路封装技术领域也做了相应的分析，具体如下所示：

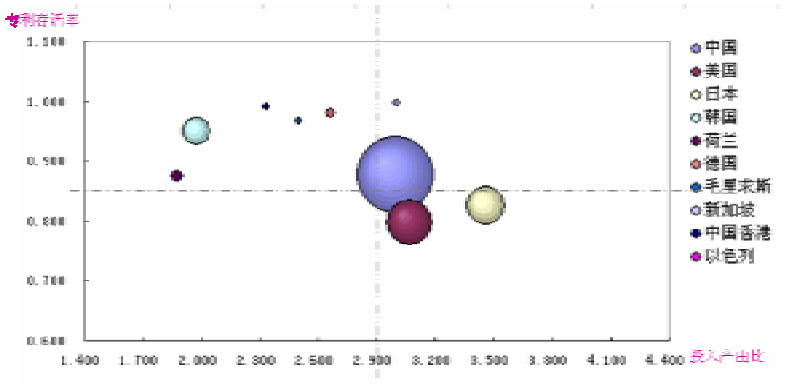

图5- 7国家层面的四象限图

本文以投入产出比表示x轴，y轴采用的则是专利存活率。投入产出比代表了企业的投入情况，而专利存活率也可以代表行业的需求状况，另外球的大小代表了申请专利数的多少。从上面可以发现国家层面的专利情况都主要集中在潜在区域；而中国在自己本土从申请量来看处于绝对优势，而且也处于领先地位。

然而这四象限的确定需要认为的干预，去确定界限，同时仅仅依靠两个指标也是有失偏颇的。而这些都是SOM方法可以避免的。

5.5.2 企业层面分析

对于企业层面的分析，以往的分析方法像前面国家的分析类似采用的还是以统计、指标等为主的简单方法，未能发掘出隐藏的内部信息。而本文采用的SOM（自组织语义映射）方法以专利指标为输入层（指标覆盖了技术、发展、数量指标等各个方面），聚类输出的结果更加有说服力。

5.5.2.1 以往分析方法

（1）数量分析

下表展示了集成电路封装技术中国专利高产机构分析结果。从表中可以看出，威盛电子股份有限公司的专利总数最多，有 23 件，虽然授权率很高，但是其中发明专利较少。国际商业机器公司专利总数排名第二，且全部为发明专利。

表格 5- 5 前十家高产机构申请专利分布表

机构	专利总数	发明专利	授权专利	授权率
威盛电子股份有限公司	23	8	19	82.6%
国际商业机器公司	20	20	8	40.0%
矽品精密工业股份有限公司	18	18	7	38.9%
英特尔公司	16	16	3	18.8%
江苏长电科技股份有限公司	16	9	8	50.0%
三星电子株式会社	12	12	4	33.3%
日月光半导体制造股份有限公司	12	11	1	8.3%
株式会社日立制作所	11	11	6	54.5%
台湾积体电路制造股份有限公司	10	8	3	30.0%
三菱电机株式会社	8	8	6	75.0%

在前十家高产机构中，有五家为外国公司，包括国际商业机器公司、英特尔公司等，这表明国外公司在集成电路封装技术上具备相当的研发实力。

（2）多维 OLAP 分析

为了发现各高产申请人的研发重点及其专利质量，我们利用 OLAP 对申请人、维持年份和主 IPC 进行分析，结果如下图所示。从图中可以看出，维持年份在 10 年以上的高产申请人均为国外公司，维持年份 13 年的 1 件专利属于三星电子株式会社，所属类别为 G01R(测量电变量；测量磁变量)，可见三星在该领域的专利质量和技术实力较高。维持年份 12 年的 4 件专利中，株式会社日立制作所 2 件，英特尔公司 1 件，国际商业机器公司 1 件，而且这 4 件专利均属于 H01L(半导体器件；其他类目未包括的电固体器件)，同时我们可以看到，维持年份在 7 年以上的 19 件专利中，除了三星公司的 1 件维持年份 13 年 G01R 专利以外，其余的 18 件专利都属于 H01L 的类别，可见各公司对 H01L 的研究是非常重视的，且该类别的专利质量是相对较高的，这与 H01L 的技术发展趋势一致，是出现专利最早的技术。

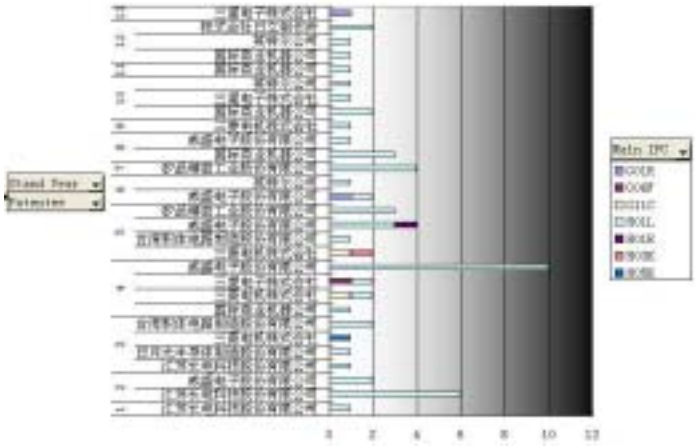

图 5- 8 封装技术中国专利申请人-维持年份-主 IPC OLAP 分析

从图中可以看出，在中国的各企业中，威盛电子股份有限公司的专利质量是相对较高的，有维持年份 8 年的专利 1 件，维持 6 年的专利 2 件，5 年的 4 件，4 年的 10 件，且大部分属于 H01L，其次是矽品精密工业股份有限公司，有维持年份 7 年的专利 4 件，且均为 H01L，可见，中国封装的先进公司研发的重点是与国际接轨的，但是专利质量与国外先进水平还有一定差距，需要提高研发水平。

5.5.2.2 SOM 方法企业层面分析

同国家层面一样，公司层面的SOM算法的输入层也采用了专利数、授权率、平均申请人数、投入产出比、专利成长率、专利存活率、专利平均维持年数、权利人数、发明专利比例共九个指标。然而，一直以来公司层面的专利分析一直是集中在专利数量统计方面，对于公司层面质量的划分并不完善。对于公司层面来讲，这种划分比国家层面的划分更为重要，因为对于公司层面的分析更加看重期间的竞争合作态势分析。而SOM方法除了对专利层次的划分外，还可以将具有相关联系的研究实体给聚集到一起，通过对其进一步研究来发现其竞争或合作态势关系。

由于企业的研究实体较多，为了简单本文选用了前二十家作为研究的对象。

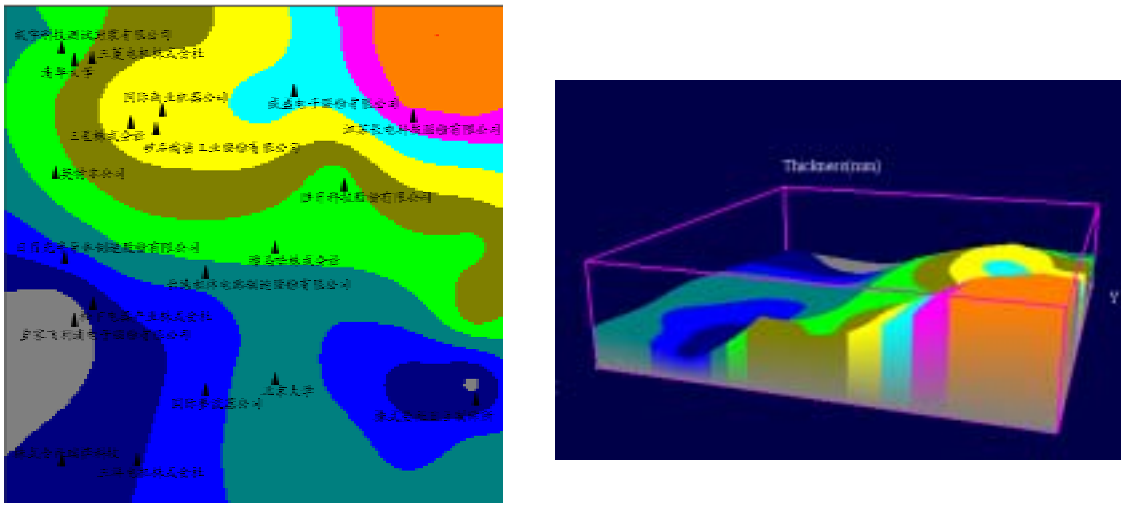

图5- 9 SOM企业层面分析图

图中颜色越鲜艳表示的专利质量越高，即灰色区域的专利质量<（弱与）深蓝色区域<（弱与）浅蓝色区域<（弱与）墨绿区域<（弱与）草绿区域<（弱与）黄褐色区域<（弱与）明黄色区域<（弱与）紫色区域<（弱与）红色区域。为了更加清晰的显示专利区域的等级情况，本文还绘制了三维的可视化图形。在三维的等高线图上，可以一目了然的看出分布情况。通过立体图还可以发现所有点的分布都集中在一个环形圈里，则表明彼此间的联系比较紧密；如果存在不同的波峰，则不同峰就会代表不同的组群。

从图中可以看出，经过SOM聚类算法后，公司在专利层面的级别可以分为一下五种：

第一类：威盛电子股份有限公司，江苏长电科技股份有限公司；

第二类：国际商业机器公司，三星株式会社，矽品精密工业股份有限公司，三菱电机株式会社；

第三类：威宇科技测试封装有限公司，清华大学，英特尔公司，胜开科技股份有限公司，雅马哈株式会社，台湾积体电路制造股份有限公司，北京大学；

第四类：日月光半导体制造股份有限公司，国际整流器公司，松下电器产业株式会社，三洋电机株式会社，株式会社瑞萨科技，株式会社日立制作所；

第五类：皇家飞利浦电子股份有限公司；

同时发现了指标层面的如下三个聚类：国际商业机器公司，三星株式会社，矽品精密工业股份有限公司；另一类，威宇科技测试封装有限公司、清华大学、威宇科技

测试封装有限公司；最后一类，松下电器产业株式会社，皇家飞利浦电子股份有限公司。对于这些聚类，需要进一步分析来确定具体的竞争合作关系。

下面就以际商业机器公司，三星株式会社，矽品精密工业股份有限公司这种聚类为例来进一步分析公司间的竞争态势关系。

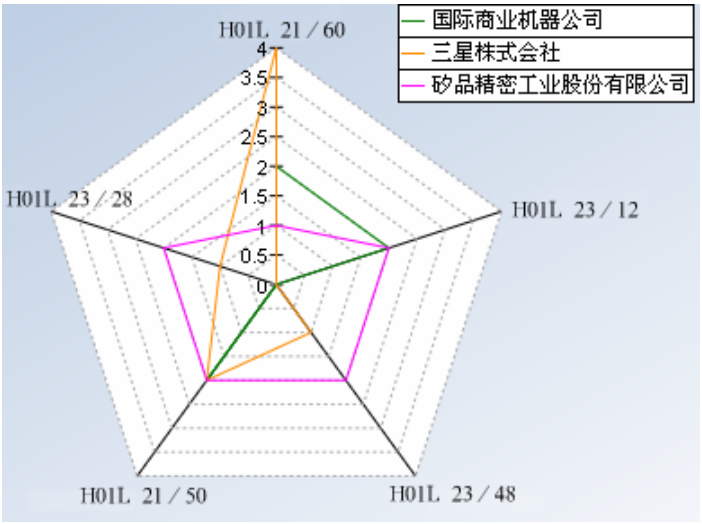

图5- 10公司间聚类分析

表格 5- 6 公司 IPC 分布表

企业/IPC	H01L 21/60	H01L 23/12	H01L 23/48	H01L 21/50	H01L 23/28
国际商业机器公司	2	2	0	2	0
三星株式会社	4	0	1	2	1
矽品精密工业股份有限公司	1	2	2	2	2

注：表中的各IPC含义为：H01L 21/60，引线或其它导电构件的连接，用于向或从器件传导电流；H01L 23/12，安装架，例如:不可拆卸的绝缘衬底；H01L 23/48，用于向或自处于工作中的固态物体通电的装置，例如引线，接线端装置(一般入H01R)；H01L 21/50，应用21/06至21/326中的任一小组都不包括的方法或设备组装半导体器件的；H01L 23/28，封装，例如:密封层，涂覆物(23/552优先)。另外，此处采用的IPC是在集成电路封装技术中国专利数据中研究比较集中的领域。

从上图中可以看出，三家公司在集成电路封装技术的研究侧重各有特点：三星株式会社主要的研究集中在H01L 21/60（引线或其它导电构件的连接，用于向或从器件传导电流），且相对另两家企业来讲占有绝对的优势。而国际商业机器公司则在H01L 21/60，H01L 23/12，H01L 21/50这三个领域均衡发展；矽品精密工业股份有限公司则是各个研究方面均有涉及，总体实力不容忽视。

这些都为企业进一步分析其竞争对手以及制定整体发展策略提供技术依据。

5.5.2.3 分析方法对比小结

通过公司层面的进一步分析，可以发现SOM方法除了体现了专利的多指标性；体现了数据挖掘的思想；对专利质量进行了很好的分类外，还可以将通过对专利指标聚类来进一步研究公司间的竞争合作关系，这些都是以往分析方法所无法比拟的。

5.5.3 SOM 方法的对比分析

（1）SOM方法与其他聚类方法对比分析

为了更好地体现SOM方法在专利信息挖掘中的优点，此处采用层次聚类法来进行对比分析。层次聚类是实际工作中使用最多的一次方法，其含义为：开始时每个样品各看成一类，将距离最近的两类合并；重新计算新类与其他类的距离，再将距离最近的两类合并……，以此类推，直至所有的样品都合并为一类。聚类分析方法最常用的还有K-MEAN分析，但是它适用与大样本数据，且需要事先指定分类数，并不适合分析小量数据。

考虑到很多的统计软件都已经整合了该聚类方法，因此本文采用SPSS软件进行了该领域的层次聚类分析；为了与SOM方法相比较，此次选用了SOM的训练集作为层次聚类法的数据集。

- 国家层面层次聚类分析

如下的冰柱图反映了层次聚类的全过程，是我们分析聚类结果和判断最优分类数的依据，图中“X”纵向排列代表了冰柱的融合过程。

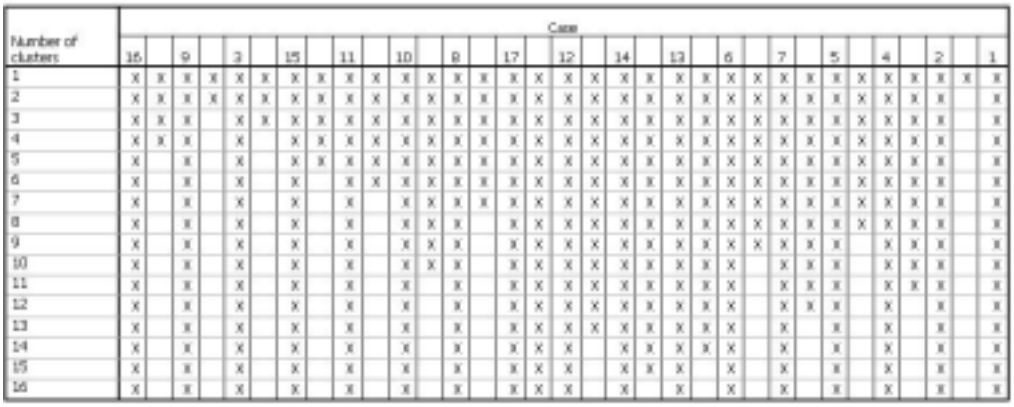

图5- 11 层次聚类冰柱图

注：（case中的1=中国；2=美国；3=日本；4=韩国；5=荷兰；6=德国；7=毛里求斯；8=新加坡；9=中国香港；10=以色列；11=意大利；12=印度尼西亚；13=匈牙利；14=开曼群岛；15=法国；16=澳大利亚；17=百慕大）

从图中已经可以看出各种聚类情况，但是为了更好的表现聚类结果，采用下面的层次聚类树形图来表现聚类的结果。

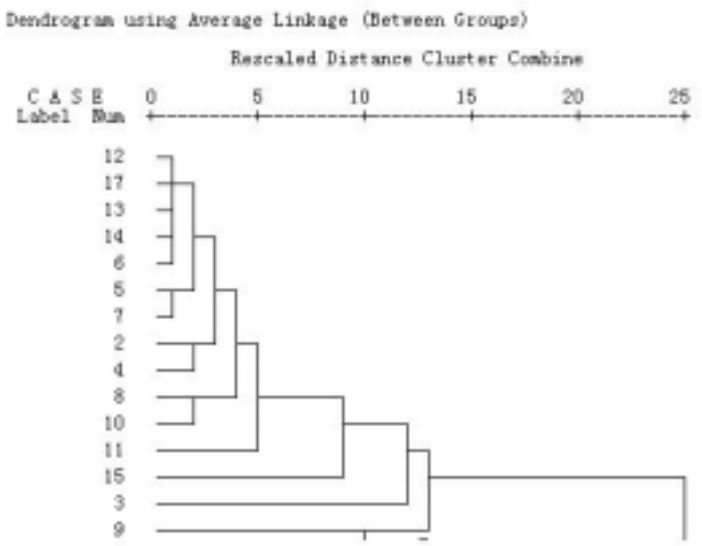

图5- 12 国家层面聚类分析树形图

从树形图中可以看出不同的聚类数量的各种结果聚类情况。假设分为4个不同的组别，用下表表示聚类的结果：

表格 5- 7 国家层次聚类的结果分析

Case	4 Clusters	Case	4 Clusters	Case	4 Clusters
中国	1	毛里求斯	2	匈牙利	2
美国	2	新加坡	2	开曼群岛	2
日本	3	中国香港	4	法国	2
韩国	2	以色列	2	澳大利亚	4
荷兰	2	意大利	2	百慕大	2
德国	2	印度尼西亚	2		

上述的结果显示国家的聚类太过集中，主要分布在第2个等级层面，而其他分布比较弱，同时单从结果来看百慕大之申请一件专利却与美国等专利实力较强的国家划分为同一等级是有待商榷的，这是因为：首先层次分析法并没有体现不同指标所起的影响及没有反应指标的权重情况，其次由于聚类方法本身没有体现考虑专利数据特点等原因导致的，而这些是可以在本文改进的SOM方法中避免的。

➢ 企业层面层次聚类分析

在国家层面的聚类分析中已经对层次聚类法进行了描述，因此这里只是将聚类的结果展现出来进行对比分析。

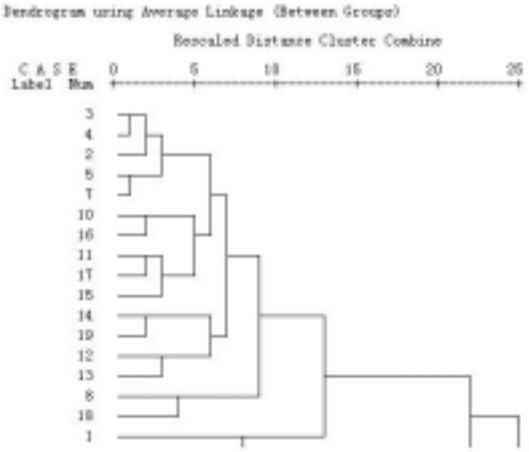

图5- 13 企业层面聚类分析树形图

假设分为4个不同的组别，则从树形图中可以看出聚类的结果：

表格 5- 8 国家层次聚类的结果分析

Case	4 Clusters	Case	4 Clusters
威盛电子股份有限公司	1	三菱电机株式会社	2
三星株式会社	2	三洋电机株式会社	2
国际商业机器公司	2	株式会社瑞萨科技	2
矽品精密工业股份有限公司	2	皇家飞利浦电子股份有限公司	2
英特尔公司	2	国际整流器公司	2
江苏长电科技股份有限公司	1	雅马哈株式会社	2
日月光半导体制造股份有限公司	2	北京大学	2
威宇科技测试封装有限公司	2	清华大学	2
株式会社日立制作所	3	松下电器产业株式会社	2
台湾积体电路制造股份有限公司	2	胜开科技股份有限公司	4

上述的结果显示企业的聚类结果太过集中，主要分布在第2个等级层面，而其他分布比较弱，而且专利实力比较差的企业与实力强的企业划分为同一等级存在一定的不合理性，原因在前已有描述。这些都是层次分析法所无法避免的。

因此通过如上分析，可以看出通过改进的SOM的可视化挖掘在专利应用中有着绝对的优势。

5.5.4 本章小结

通过同以往专利分析方法进行对比，可以发现SOM方法可以实现以往专利分析方法所无法触及的方面，简单总结如下：

- 专利的多指标体现：本文SOM方法输入层囊括了专利数、授权情况、研发实力投入情况、逐年的发展状况到专利的维持情况等等很多方面，全面的反应了专利的性质，从而解决了单一指标信息反应片面的问题。
- 数据挖掘的思想的体现：以往的专利分析主要是集中在矩阵统计分析基础上基于数量的加减乘除的计算。而SOM算法是一种无监督的自组织学习的聚类算法，更好的体现了数据挖掘的思想。
- 专利质量等级的划分：通过SOM方法可以将各国在某领域的实力展现出来，从而为进一步的分析奠定基础。
- 企业间联系的挖掘：由于每个研究实体都有不同的属性值（专利指标值）来反映，具有相同的属性值或者相似属性值的实体自然聚在一起，形成了聚类。而聚在一起的企业则有着千丝万缕的关系，或竞争或合作，这些都可以作为进一步分析的切入点；再通过雷达图的形式将企业在各个领域方法的竞争情况一目了然的呈现，从而为企业提供了情报分析的数据支持。

正是由于SOM可视化方法的在专利应用中的诸多优点使得本文对其的研究非常有意义。

6 总结与展望

6.1 本文的研究意义

（1）理论意义：本文主要进行的是研究方法的探索，相对于传统的专利分析方法和可视化研究，本文的研究方法可以更加直观、明了的展现专利质量的划分和公司间竞争合作态势的划分。

（2）实践意义：由于专利信息的重要性越来越收到企业和行业的重视，而可视化数据挖掘在专利信息层面的应用可以使得专利信息的获取更为方便、及时，分析也更加智能化，并且呈现方式简单明了易懂，因此为相应领域中企业的研究策略制定提供了依据。

6.2 本文的创新点

本论文从专利信息分析入手展开了专利信息可视化工作，并建立了可视化模型和图形生成算法，同时在数据获取、集成方面也展开了相应的研究工作。在文献的研究上，开展了一个完整的工作。最后结合实证分析来验证模型的可行性及正确性。

概括起来，本文的创新点主要包括：

- 将SOM方法引入到专利分析层面中来。并根据专利信息的特点，对SOM方法进行了相应的改进。
- 在专利信息可视化方面，应用等高线作为SOM输出结果的呈现方式，使得专利的划分情况以更加简便、清洗、完整的展现。

6.3 下一步工作

（1）寻找和改进数据挖掘算法，设计多样化的可视化模型；

（2）SOM输入层的进一步改进，使得可以更全面的反应专利信息；

（3）可视化交互功能的研究，如OVER-VIEW AND DETAIL等功能的加入。

致谢

随着时间的流逝，转眼两年年的硕士学习生涯将告结束。在此，仅已“基于 SOM 的专利信息可视化研究”为自己的学习生涯划一个圆满的句号。这篇论文的完成与导师朱东华教授的悉心指导是分不开的。感谢他在我工作遇到困惑时的指点，感谢他对我论文撰写的悉心指导，在这里向朱老师表示衷心的感谢！

同时，毕业论文也是对研究生阶段所学知识的综合运用，在这里我还要感谢管理与经济学院的各位老师，感谢各位老师两年来的谆谆教导和精心培养，没有您们的辛勤培养与教导，今天的毕业论文也是无法顺利完成的。

另外，还要感谢北京理工大学知识发现与数据分析实验室的师兄师姐，感谢他们在这段时间对我的照顾。他们在整个论文攥写过程中给予我的指导和帮助，在此向他们表达我最真诚的谢意！

最后，谨向百忙之中抽出宝贵时间的评审本论文的专家、老师致以最诚挚的谢意。

攻读硕士期间发表的学术论文

[1] 谢菲，朱东华，任智军. 面向国防文献情报的技术监测分析研究. 情报杂志，2006.11

[2] 任智军，朱东华,谢菲. 基于技术监测的国防文献情报分析研究. 北京理工大学学报（自然科学版）管理科学与工程专刊 2006.10

[3] 任智军，朱东华,谢菲.专利引用分析方法研究专利.商业时代.2007.05

[4] 任智军，朱东华,谢菲.科技文本的可视化分析研究.北京理工大学学报（社会科学版）2007.1

[5] 任智军，朱东华,谢菲.商业银行专利情报分析研究.情报杂志.2007.05

参考文献：

[1] 吴贵生. 技术创新管理[M]. 北京:清华出版社, 2000:12-35.

[2] 中国专利局专利法研究所. 专利法研究[M]. 北京:专利文献出版社, 1992：3-15.

[3] 张永嘉,魏铁进. 专利与专利情报教程[M]. 北京:航空工业出版社,1990 :12-15.

[4] 余世银,乐嘉锦,张侃. 数据挖掘可视化研究[J]. 东华大学学报(自然科学版) ,2001 ,27(2):102-106.

[5] Kevin W.Boyack ,Brain N.Wylie,George S.Davidson,David K.Johnson. Analysis of patent databases using Vxinsight[R]. McLean,VA:CIKM2000,2000.

[6] 王崇德. 文献计量学教程[M]. 天津:南开大学出版社,1990:15-20.

[7] 彭爱东. 企业专利情报分析研究[D]. 南京:南京大学,2000.

[8] Mogee, M.E. Patent analysis methods in support of licensing[R]. Denver: the Technology Transfer Society Annual Conference,1997.

[9] 陈建军,于志强,朱昀. 数据可视化技术及其应用[J]. 红外与激光工程, 2001,30(5):339-342.

[10] 刘勘,周晓峥,周洞汝. 数据可视化的研究与发展[J].计算机工程,2002,28(8).

[11] H.Y Lee ,H.L Ong. Visualization Support for Data Mining[J]. IEEE Expert,1996,11(5):69.

[12] 朱建秋,蔡伟杰(译),Tom Soukup,Ian Davidson(著). 可视化数据挖掘－数据可视化和数据挖掘的技术与工具[M].北京:电子工业出版社,2004:8-21.

[13] 黄志澄. 给数据以形象 给信息以智能--数据可视化技术及其应用展望[J].电子展望与决策,1999(6):3-9.

[14] Mihael Ankrst, Jiawei Han. Visual Data Mining[R]. Canada: Simon Fraser University,2000.

[15] Schwander,P. An evaluation of patent searching resources: comparing the professional and free on-line databases[J]. World Patent Information, 2000, 22:147-165.

[16] Albert, M.B., Yoshida, P.G., and van Opstal, D. Global patenting trends[J]. CHEMTECH ,1999,29(2):47-58.

[17] Gupta, V.K. & Pangannaya, N.B. Carbon nanotubes:bibliometric analysis of patents[J].

World Patent Information,2000,22:185-189.

[18] Margolis R.M. & Kammen, D.M. Evidence of underinvestment in energy R&D in the United States and the impact of Federal policy[J]. Energy Policy ,1999,27: 575-584 .

[19] 基于 SOM 神经网络的中国区域文化产业的竞争力评价[J].中国科技论文在线

[20] Fox, K. L., Frieder, O., Knepper, M. M. & Snowberg, E. J.SENTINEL: A multiple engine information retrieval and visualization system[J]. Journal of the American Society for Information Science,1999, 50(7):616-625.

[21] 飞思科技产品研发中心(著). 神经网络理论与 MATLAB7 实现[M].北京:电子工业出版社,2005-3

[22] Wise, J. A. The ecological approach to text visualization[J].Journal of the American Society for Information Science,1999, 50(13):1224-1233.

[23] Davidson, G. S., Hendrickson, B., Johnson, D. K., Meyers,C. E. & Wylie, B. N. Knowledge mining with VxInsight: discovery through interaction[J]. Journal of Intelligent Information Systems,1998, 11: 259-285.

[24] 钱肖鲁，朱建秋，朱扬勇. DMVisualMiner：一个可视化数据挖掘分析平台[J].计算机工程,2003,29(z1):148-150.

[25] 吴星玮, 饶培伦.运用自组织特征映射文本挖掘算法分析中国人类工效学研究状况[J].

[26] Singhal, M.; Payne, D.A.; Weir Lipton, M.S.; Webb-Robertson, B.-J.M. Enabling proteomics discovery through visual analysis. IEEE Engineering in Medicine and Biology Magazine. May-June 2005: 50-57.

[27] Mizoguchi, F. Anomaly detection using visualization and machine learning. Enabling Technologies: Infrastructure for Collaborative Enterprises, 2000. (WET ICE 2000). Proeedings. IEEE 9th International Workshops on, 2000:165 – 170.

[28] George G. Robertson, Jock D. Mackinlay, Stuart K. Card. Cone Trees: animated 3D visualizations of hierarchical information. Conference on Human Factors in Computing Systems1997: 189 – 194.

[29] 唐贤伦,仇国庆,李银国,曹长修. 基于粒子群优化和ＳＯＭ网络的聚类算法研究[J].华中科技大学学报（自然科学版）,2007，5：31-33

[30] 朱家元,张恒喜,虞健飞. 在数据挖掘中基于 SOM 网络的数据分析可视化设计[J]

[31] 蔡元萃,陈立潮. 聚类算法研究综述[J]. 科技情报开发与经济: 2007 年第 17 卷第 1 期

[32] 张红云,石阳,马垣. 数据挖掘中聚类算法比较研究[J]. 鞍山钢铁学院学报:2001 年 10 月,364-369

[33] 周志勇,袁方,刘海博. 用聚类-分类模式解决聚类问题[J]. 广西师范大学学报:自然科学版: 第 25 卷第 2 期,127-131

[34] 朱莉. 数据挖掘与企业竞争优势[J]. 情报杂志 2003 年第 11 期:43-44

[35] 宇传华. SPSS 与统计分析[M]. 北京：电子工业出版社, 2007.2

[36] 曹雷.面向专利战略的专利信息分析研究[J]. 科技管理研究，2005,(3):97~100

[37] 暴海龙,朱东华.专利情报分析方法综述[J]. 北京理工大学学报(社会科学版)，2002,10(4):91~93

[38] 戚昌文，邵洋：市场竞争与专利战略[M]. 华中理工大学出版社，1995

[39] 彭爱东.一种重要竞争情报——专利情报的分析研究[J]. 情报理论与实践，2000,(3):196~199

[40] 党倩娜.专利分析方法和主要指标[J].上海企业技术创新信息网

[41] http://www.sims.berkeley.edu/course/is247/s02/Lectures.html

[42] 赵焕方，朱东华. 信息可视化在技术监测中的应用[J]. 情报杂志,2005(12):46-48

[43] Nahum D.Gershon, Stephen G. Eick. Information Visualization[J]. IEEE Computer Graphics and Application, 1997(7-8): 29-31