# K-Means 聚类在我国专利申请分析中的应用

彭 剑 芳

(云南大学 情报与档案学系 昆明 650200)

**摘 要** 专利的申请量和拥有量是衡量一个国家或地区科技水平高低的重要指标,对专利信息的分析也是一项有意义的研究。借助数据挖掘工具 SPSS Clementine 11.1,使用 K-Means 聚类模型,在对我国 2009 年各省市区专利申请及授权的基本情况进行分析的基础上,对比各地区的社会经济状况,总结出了专利活动与经济社会发展之间的一些规律,并提出了一些建议,为研究专利对经济社会发展的促进作用提供参考资料。

**关键词** 数据挖掘 聚类分析 K-Means 聚类 专利申请分析

**中图分类号** G434 **文献标识码** A **文章编号** 1002-1965(2010)0044-04

知识经济时代,知识产权已经成为企业竞争的重要手段和武器,成为国家核心竞争力的重要组成部分。专利作为知识产权的重要组成部分,在企业的发展中起着至关重要的作用。“根据界知识产权组织的统计,专利文献中包含了世界上 95% 的研发成果。如果能够有效地利用专利情报,不仅可以缩短 60% 的研发时间,还可以节省 40% 的研发经费。”[1] 但是,随着发明创新的日益复杂和申请数量的不断增长,以及受各地区间经济、科技、地理等因素的影响,我国的专利活动呈现出很大的区域差异性,使得各项专利不能得到有效的利用和管理,这不仅制约了专利活动的进一步发展,也在一定程度上制约了经济的发展,因此,对专利信息进行分析研究也就显得十分必要。本文在对我国 2009 年各省市区专利申请及授权的基本情况进行分析的基础上,对比各地区的社会经济状况,总结出了专利活动与经济社会发展之间的一些规律,并提出了一些建议,为研究专利对经济社会发展的促进作用提供参考资料。

## 1 样本资料

本研究旨在对我国 2009 年的专利发展状况进行分析,所谓的专利是指受法律规范保护的发明创造,它是指一项发明创造向国家审批机关提出专利申请,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,包括发明专利、实用新型专利和外观设计专利三种:

发明:是指对产品、方法或者其改进所提出的新的技术方案。

实用新型:是指对产品的形状、构造或者其结合所提出的适于实用的新的技术方案。

外观设计:是指对产品的形状、图案或者其结合以及色彩与形状、图案所作出的富有美感并适于工业上应用的新设计。

对专利信息的分析研究可以从多个角度进行分析,本文主要是从空间的角度对我国各省市区的专利活动情况进行分析,因此本文所要分析的数据主要是我国各个省市区在 2009 年度的专利申请量(如表 1 所示)和授权量(如表 2 所示),其中包括发明、实用新型和外观设计三种专利类型的数据。本文所使用的数据均来自中华人民共和国国家知识产权局网站。

表 1 2009 年各省市区专利申请量

| 地区 | 合计 | 发明 | 实用新型 | 外观设计 |
|---|---|---|---|---|
| 总计 | 877611 | 229096 | 308861 | 339654 |
| 北京 | 50236 | 29326 | 15424 | 5486 |
| 天津 | 19624 | 6367 | 8267 | 4990 |
| 河北 | 11361 | 2811 | 6478 | 2072 |
| 山西 | 6822 | 2422 | 2791 | 1609 |
| 内蒙古 | 2484 | 719 | 1266 | 499 |
| 辽宁 | 25803 | 7125 | 12633 | 6045 |
| 吉林 | 5934 | 2166 | 2912 | 856 |
| 黑龙江 | 9014 | 3384 | 4357 | 1273 |
| 上海 | 62241 | 22012 | 19650 | 20579 |
| 江苏 | 174329 | 31779 | 36122 | 106428 |
| 浙江 | 108482 | 15646 | 40364 | 52472 |
| 安徽 | 16386 | 4465 | 7065 | 4856 |
| 福建 | 17559 | 3842 | 7844 | 5873 |
| 江西 | 5224 | 1502 | 2439 | 1283 |
| 山东 | 66857 | 13983 | 32091 | 20783 |
| 河南 | 19589 | 4952 | 9912 | 4725 |
| 湖北 | 27206 | 6065 | 10579 | 10562 |
| 湖南 | 15948 | 4416 | 7075 | 4457 |
| 广东 | 125673 | 32247 | 39027 | 54399 |
| 广西 | 4277 | 1280 | 2091 | 906 |

收稿日期:2010-06-06

作者简介:**彭剑芳**(1986-),女,硕士研究生。

（续表 1）

| | | | | |
|---|---|---|---|---|
| 重庆 | 13482 | 3845 | 5503 | 4134 |
| 四川 | 33047 | 6260 | 11943 | 14844 |
| 贵州 | 3709 | 1336 | 1659 | 714 |
| 云南 | 4633 | 1637 | 1825 | 1171 |
| 西藏 | 195 | 71 | 72 | 52 |
| 陕西 | 15570 | 5858 | 5798 | 3914 |
| 甘肃 | 2676 | 1120 | 1075 | 481 |
| 青海 | 499 | 175 | 147 | 177 |
| 宁夏 | 1277 | 182 | 284 | 811 |
| 新疆 | 2872 | 662 | 1865 | 345 |
| 海南 | 1040 | 456 | 359 | 225 |

表 2　2009 年各省市区专利授权量

| 地区 | 合计 | 发明 | 实用新型 | 外观设计 |
|---|---|---|---|---|
| 总计 | 501786 | 65391 | 202113 | 234282 |
| 北京 | 22921 | 9157 | 10141 | 3623 |
| 天津 | 7404 | 1889 | 3988 | 1527 |
| 河北 | 6839 | 691 | 4515 | 1633 |
| 山西 | 3227 | 603 | 1967 | 657 |
| 内蒙古 | 1494 | 178 | 762 | 554 |
| 辽宁 | 12198 | 1993 | 8585 | 1620 |
| 吉林 | 3275 | 719 | 1931 | 625 |
| 黑龙江 | 5079 | 1142 | 3212 | 725 |
| 上海 | 34913 | 5997 | 13158 | 15758 |
| 江苏 | 87286 | 5322 | 21939 | 60025 |
| 浙江 | 79945 | 4818 | 25295 | 49832 |
| 安徽 | 8594 | 795 | 4226 | 3573 |
| 福建 | 11282 | 824 | 4939 | 5519 |
| 江西 | 2915 | 386 | 1515 | 1014 |
| 山东 | 34513 | 2865 | 22635 | 9013 |
| 河南 | 11425 | 1129 | 6630 | 3666 |
| 湖北 | 11357 | 1478 | 6285 | 3594 |
| 湖南 | 8309 | 1752 | 4218 | 2339 |
| 广东 | 83621 | 11355 | 27438 | 44828 |
| 广西 | 702 | 326 | 1506 | 870 |
| 重庆 | 7501 | 834 | 3274 | 3393 |
| 四川 | 20132 | 1596 | 6561 | 11975 |
| 贵州 | 2084 | 322 | 1234 | 528 |
| 云南 | 2923 | 476 | 1338 | 1109 |
| 西藏 | 292 | 7 | 40 | 245 |
| 陕西 | 6087 | 1342 | 3446 | 1299 |
| 甘肃 | 1274 | 227 | 809 | 238 |
| 青海 | 368 | 35 | 89 | 244 |
| 宁夏 | 910 | 52 | 267 | 591 |
| 新疆 | 1866 | 120 | 1260 | 486 |
| 海南 | 630 | 84 | 330 | 216 |

## 2　算法描述

本文所采用的是统计分析中的聚类分析方法。聚类就是按照某个特定标准(一般为距离准则)把一个数据集分割成不同的类或簇,使得在同一个簇内的数据对象的相似性尽可能的大,不在同一个簇中的数据对象的差异性也尽可能地大。“传统的统计聚类分析方法包括系统聚类法、有序样品聚类法、动态聚类法、模糊聚类法、图论聚类法、聚类预报法等。采用 k-均值、k-中心点等算法的聚类分析工具已被加入到许多著名的统计分析软件包中,如 SPSS、SAS 等。”[2] 本文使用的是 SPSS Clementine 11.1 统计分析软件,提供了 K-means 、Two-step 和 Kohonen 三种聚类方法。本研究采用的是 K-means 聚类模型。

K-means 又称快速聚类,是由 Macqueen 于 1967 年提出的。其目的是:把样品聚集成 K 个类的集合,要求同一类中样品彼此相似,而不同类间的样品差异较大。K 的大小是事先确定好的。其基本思想是:把每个样品聚集到其最近形心(均值)类中去。K-means 算法步骤如下:a. 从原始数据中选取 K 个点作为初始的 K 个聚类中心。通常由 Clementine 自动选。本研究中由于数据较少,仅 31 条记录,因此手动设置为 6,即 K=6。b. 各样本到 K 个聚类中心的距离,把样本归到离它最近的那个聚类中心所在的类。c. 新形成的 K 个类的均值作为新的聚类中心,重复步骤 b 重新聚类。d. 重复步骤 c,直到到达最大迭代次数或前后两次迭代之间的差异小于制定阀值,聚类过程结束。

## 3　聚类分析

本研究借助统计软件 SPSS Clementine 11.1,使用聚类分析中的 K-Means 分析模型,对 2009 年各省份的专利授权数据进行分析,得到如下结果(见表 3 和表 4):

表 3　专利申请数据聚类结果

| 地区 | 合计 | 发明 | 实用新型 | 外观设计 | 类别 |
|---|---|---|---|---|---|
| 北京 | 50236 | 29326 | 15424 | 5486 | 聚类-1 |
| 天津 | 19624 | 6367 | 8267 | 4990 | 聚类-3 |
| 河北 | 11361 | 2811 | 6478 | 2072 | 聚类-3 |
| 山西 | 6822 | 2422 | 2791 | 1609 | 聚类-3 |
| 内蒙古 | 2484 | 719 | 1266 | 499 | 聚类-3 |
| 辽宁 | 25803 | 7125 | 12633 | 6045 | 聚类-3 |
| 吉林 | 5934 | 2166 | 2912 | 856 | 聚类-3 |
| 黑龙江 | 9014 | 3384 | 4357 | 1273 | 聚类-3 |
| 上海 | 62241 | 22012 | 19650 | 20579 | 聚类-1 |
| 江苏 | 174329 | 31779 | 36122 | 106428 | 聚类-2 |
| 浙江 | 108482 | 15646 | 40364 | 52472 | 聚类-4 |
| 安徽 | 16386 | 4465 | 7065 | 4856 | 聚类-3 |
| 福建 | 17559 | 3842 | 7844 | 5873 | 聚类-3 |
| 江西 | 5224 | 1502 | 2439 | 1283 | 聚类-3 |
| 山东 | 66857 | 13983 | 32091 | 20783 | 聚类-4 |
| 河南 | 19589 | 4952 | 9912 | 4725 | 聚类-3 |
| 湖北 | 27206 | 6065 | 10579 | 10562 | 聚类-3 |
| 湖南 | 15948 | 4416 | 7075 | 4457 | 聚类-3 |
| 广东 | 125673 | 32247 | 39027 | 54399 | 聚类-5 |
| 广西 | 4277 | 1280 | 2091 | 906 | 聚类-3 |

（续表3）

| | | | | | |
|---|---|---|---|---|---|
| 重庆 | 13482 | 3845 | 5503 | 4134 | 聚类-3 |
| 四川 | 33047 | 6260 | 11943 | 14844 | 聚类-3 |
| 贵州 | 3709 | 1336 | 1659 | 714 | 聚类-3 |
| 云南 | 4633 | 1637 | 1825 | 1171 | 聚类-3 |
| 西藏 | 195 | 71 | 72 | 52 | 聚类-3 |
| 陕西 | 15570 | 5858 | 5798 | 3914 | 聚类-3 |
| 甘肃 | 2676 | 1120 | 1075 | 481 | 聚类-3 |
| 青海 | 499 | 175 | 147 | 177 | 聚类-3 |
| 宁夏 | 1277 | 182 | 284 | 811 | 聚类-3 |
| 新疆 | 2872 | 662 | 1865 | 345 | 聚类-3 |
| 海南 | 1040 | 456 | 359 | 225 | 聚类-3 |

表4　专利授权数据聚类结果

| 地区 | 合计 | 发明 | 实用新型 | 外观设计 | 类别 |
|---|---|---|---|---|---|
| 北京 | 22921 | 9157 | 10141 | 3623 | 聚类-1 |
| 天津 | 7404 | 1889 | 3988 | 1527 | 聚类-3 |
| 河北 | 6839 | 691 | 4515 | 1633 | 聚类-3 |
| 山西 | 3227 | 603 | 1967 | 657 | 聚类-3 |
| 内蒙古 | 1494 | 178 | 762 | 554 | 聚类-3 |
| 辽宁 | 12198 | 1993 | 8585 | 1620 | 聚类-3 |
| 吉林 | 3275 | 719 | 1931 | 625 | 聚类-3 |
| 黑龙江 | 5079 | 1142 | 3212 | 725 | 聚类-3 |
| 上海 | 34913 | 5997 | 13158 | 15758 | 聚类-1 |
| 江苏 | 87286 | 5322 | 21939 | 60025 | 聚类-2 |
| 浙江 | 79945 | 4818 | 25295 | 49832 | 聚类-2 |
| 安徽 | 8594 | 795 | 4226 | 3573 | 聚类-3 |
| 福建 | 11282 | 824 | 4939 | 5519 | 聚类-3 |
| 江西 | 2915 | 386 | 1515 | 1014 | 聚类-3 |
| 山东 | 34513 | 2865 | 22635 | 9013 | 聚类-4 |
| 河南 | 11425 | 1129 | 6630 | 3666 | 聚类-3 |
| 湖北 | 11357 | 1478 | 6285 | 3594 | 聚类-3 |
| 湖南 | 8309 | 1752 | 4218 | 2339 | 聚类-3 |
| 广东 | 83621 | 11355 | 27438 | 44828 | 聚类-5 |
| 广西 | 2702 | 326 | 1506 | 870 | 聚类-3 |
| 重庆 | 7501 | 834 | 3274 | 3393 | 聚类-3 |
| 四川 | 20132 | 1596 | 6561 | 11975 | 聚类-3 |
| 贵州 | 2084 | 322 | 1234 | 528 | 聚类-3 |
| 云南 | 2923 | 476 | 1338 | 1109 | 聚类-3 |
| 西藏 | 292 | 7 | 40 | 245 | 聚类-3 |
| 陕西 | 6087 | 1342 | 3446 | 1299 | 聚类-3 |
| 甘肃 | 1274 | 227 | 809 | 238 | 聚类-3 |
| 青海 | 368 | 35 | 89 | 244 | 聚类-3 |
| 宁夏 | 910 | 52 | 267 | 591 | 聚类-3 |
| 新疆 | 1866 | 120 | 1260 | 486 | 聚类-3 |
| 海南 | 630 | 84 | 330 | 216 | 聚类-3 |

对表3和表4进行分析，我们可以得出以下结果：a. 北京、上海两地，专利活动整体水平最高，专利申请数量和授权数量都比较高，且发明所占比例高。b. 广东、江苏、浙江的专利活动水平也比较高，专利数量位居前列，但主要是外观设计，发明专利所占比例不是很高。c. 山东、天津等地，专利申请数量和授权数量较高，但发明专利所占比例低，主要是以实用新型为主。d. 湖南、四川、安徽等地，专利申请数量和授权数量居中，发明专利所占比例较低。e. 西藏、青海、宁夏等地，专利活动整体水平较差，数量少，且发明专利所占比例极低。

在聚类结果的基础上，根据以上分析可以将31个省市区分为以下五类：

表5　最终分类结果

| 类别 | 第一类 | 第二类 | 第三类 | 第四类 | 第五类 |
|---|---|---|---|---|---|
| 地区 | 北京、上海 | 广东、江苏、浙江 | 山东、湖北、天津、河北、辽宁、河南 | 福建、重庆、四川、陕西、安徽、湖南、 | 甘肃、青海、宁夏、新疆、海南、西藏、贵州、广西、云南、江西、吉林、山西、黑龙江、内蒙古 |

## 4　结果分析

根据图表5的最终分类，对照各地区的经济社会条件，总结出专利活动与社会经济发展等方面的关系和规律：

**4.1　专利活动的多少与当地经济社会状况的好坏密切相关**　2009年度专利活动最多的几个省份如：广东、浙江、上海、北京等，都是经济比价发达的地区。这些地区强大的经济实力以及经济发展的巨大需求都极大地促进了本地区专利活动的发展。反过来，专利创新尤其是发明专利又为本地区创造了巨大的经济效益，从而形成了一个良性循环。也正因为如此，这些省份的专利授权总量也位居前列。第一类的北京、上海和第二类浙江、广东都属于这种情况。而专利活动较少的第五类地区则往往是经济不发达或欠发达的地区，如西藏、青海、宁夏等地区，受经济、技术条件的限制，专利活动开展得比较少，由于缺乏技术创新，经济发展也相应比较缓慢。因此，专利创新与经济发展是一个互为因果，互相促进的过程。

**4.2　专利类型结构影响经济的发展**　由于各类型的专利对经济的贡献程度是不一样的，一般来说发明专利的创新程度最高，对经济的推动会比较大，而每个省份的专利类型结构又都有差异，因此，专利授权数量最高的地区并不一定就是经济最发达的地区。如：第二类的江苏、广东、浙江在专利数量上排在前三名，但其经济却不如排名第四和第六的上海、北京，究其原因，就是因为专利类型结构不一样，江苏2009年的专利授权中，有68%以上的专利都是外观设计，发明专利只占6%，而上海的发明专利占到了17%以上。因此，专利数量高且发明专利比例高的第一类地区北京、上海，相对于专利数量高但发明专利比例较低的第二类地区江苏、浙江而言，经济更为发达。

**4.3　专利结构特色与该地区发展特点相一致**　创新源自现实的需求，很多专利的发明创造都是为了更好地满足人们的需求，适应经济的发展，因此一个地区的发展模式和特点往往无形中影响了该地区的专利结构。第二类地区广东、江苏、浙江以发展食品、服饰等轻工业为主，因此，专利活动以外观设计为主，所占比例均超过了50%。第三类地区中的辽宁、天津等以发展重工业为主，因此实用新型以及发明创造所占比例较高。第一类地区的北京和上海是全国的文化、科技中心，经济实力也比价雄厚，因此，发明专利所占的比例

就较高,尤其是北京,发明专利所占比例竟达到了 40%。

4.4　**专利活动主体为企业,且主体地位在不断加强**　"创新是从发明创造到产业化的全过程,利用技术或知识产权增强竞争力、赢得市场是创新的意义和目的。"[3] 企业是市场的主体,也是创新的主力军,因而也是专利活动的主要从事者和利用者。我国知识产权局专利统计资料表明:"2009 年,国内企业申请占国内申请总量的比重达到 44.9%,比上年增长 3.7 个百分点,显示企业的创新主体地位不断强化。"[4] 在国内职务申请的 483051 件中,企业申请更是多达 394299 件,占 81.6%,同比增长 33.4%。除了企业之外,个人以及科研单位的专利活动也比较频繁,相比之下,一些大专院校的专利申请数量却比较,在上述的五类地区当中,几乎每一类地区的专利申请者都是企业和科研单位,大专院校及机关团体所占比例较少。

4.5　**专利授权率低**　专利申请量体现一个地区专利意识的强弱,专利授权率则体现专利申请的有效率,授权率高就说明有效率高,浪费的资源和重复劳动就少,反之亦然。从图表 1 和图表 2 的对比中,我们不难看出,我国目前的专利授权率还比较低,仅有 50% 左右,即便是专利水平较高的第一、二类地区,授权率也不高。北京 2009 年的专利申请量为 50236 件,而授权量只有 22 921 件,授权率只有 45.6%。这就意味着有一半以上的专利申请作废,这是对社会资源的极大浪费。

## 5　建议

专利的申请量和拥有量是衡量一个国家或地区科技水平高低的重要指标,可以从一个侧面反映一个国家或地区的创新能力、科技水平和市场化程度,衡量该国家或地区的科技产出和知识创新[5]。通过以上分析,我们可以得出我国专利发展的大致情况,专利申请量和授权量持续增长,企业成为专利活动的主体,各地区专利活动水平参差不齐,发明专利所占比例小,专利结构有待进一步优化。面对这样一种情况,我们该如何促进专利活动的开展,并充分利用专利促进经济社会的发展?在此,笔者提出以下几点建议:

5.1　**要进一步增强专利意识,促进专利申请活动的发展**　"2009 年我国共受理专利申请 976,686 件,同比增长 17.9%。其中,受理国内申请 877611 件,占总量的 89.9%,同比增长 22.4%;"[6] 可以说在 2009 年,我国的专利申请保持了平稳较快增长,尤其是在金融危机席卷全球的环境下更是难能可贵,但不容忽视的是,这样的增长速度是建立在庞大的人口基数上的,若按人均量来算,与美国、日本等发达国家仍然有很大的差距。特别是在我国广大的经济欠发达的中西部地区,专利意识淡薄,专利申请活动寥寥无几。针对这种情况,不仅要在这些地区普及专利知识,更要加强当地的科技、文化、经济基础设施建设,提高其进行专利研发的能力和条件。政府及各机构还可以适当地出台奖励政策,对研发出专利的人员进行奖励或者提供优惠政策,调动创新的积极性。

5.2　**开展专利合作**　专利合作的开展可以从多方面进行,如:地区间的合作,专利活动整体水平较高的地区可以帮助水平较低的地区,向其提供相应的资金、技术、设施等援助,也可以相互间进行互补。这种合作可以通过政府间的合作来实现,也可以通过企业、学校、机关团体间的合作来实现,通过合作实现互补。除了地区间的合作,企业、学校、机关团体间也可以进行合作。企业有雄厚的资金,学校有足够的人力资源、科研单位的设施比较齐全,通过彼此间的合作,不但可以实现优势互补,节省成本和时间,也可以优化专利申请人的结构,促进学校、机关团体专利申请活动的开展。

5.3　**进一步优化专利类型结构**　虽然都属于专利,但发明专利是技术水平较高的发明创造,与实用新型及外观设计专利相比,更能体现科学技术水平及发展。但就目前来看,我国的专利申请中,发明专利所占的比例较小,仅二十多个百分点,而外观设计说占比例最高,其次是实用新型。这样的类型结构是不太利于经济社会发展的。造成这种现象的原因除了发明创造数量本身少之外,也有可能是没有正确处理好实用新型专利和发明专利的关系,将一些技术水平高的发明创造申请了实用新型专利。因此,审批部门在进行申请的时候也要注意把关,严格分辨类型。

5.4　**提高专利授权率**　没有授权的专利申请是没有法律效力的,也不能产生经济社会效益,只有经过授权的专利申请才是合法的,有效的。造成专利授权率低的原因主要是出现了重复创造,或者是达不到授权的标准。要解决这个问题,专利申请人必须事先进行充分的专利信息查询和分析,避免出现重复劳动,而且一定要保证专利具备新颖性、实用性和创造性,避免因达不到授权标准而不予授权。

**参 考 文 献**

[1] 王晓琳. 中国历年专利数据统计分析研究[J]. 科技情报开发与经济,2008,18(2):77

[2] 张国华. 中国城镇居民消费结构的聚类分析[D]. 南京:南京财经大学,2008

[3] 国家知识产权局规划发展司. 专利统计简报[EB]. 中华人民共和国国家知识产权局网站(http://www.sipo.gov.cn/sipo2008/ghfzs/zltjjb/201001/P020100122519350133217.pdf),2010-01-12

[4] 国家知识产权局规划发展司. 专利统计简报[EB]. 中华人民共和国国家知识产权局网站(http://www.sipo.gov.cn/sipo2008/ghfzs/zltjjb/),2009-12-30

[5] 周秀会,王志辉. 中国专利申请情况分析[J]. 情报科学,2001,19(1):110

(责编:白燕琼)