

# 机器人产业技术机会发现研究

## ——基于专利文本挖掘

龚惠群, 刘琼泽, 黄超

(东南大学 经济管理学院, 江苏 南京 210096)

**摘要:**基于文本挖掘及核主成分分析方法,绘制了机器人产业技术的二维及三维技术空白专利地图,寻找机器人产业的技术空白点,通过专利空白点的平均被引频次、平均同族专利数量及密度三项指标,评估空白区域的重要程度,对于研究者进一步规划空白点技术的创新突破路径具有重要意义。

**关键词:**机器人产业;专利分析;文本挖掘;核主成分分析;技术机会

**DOI:**10.6049/kjbydc.2013041064

**中图分类号:**F406.67

**文献标识码:**A

**文章编号:**1001-7348(2014)05-0070-05

## 0 引言

当今时代,技术正以其特殊的力量推动着社会及经济发展,伴随着各种技术创新活动的兴起,企业获得了大量的技术机会。技术机会主要是指通过对某领域内已有技术横向和纵向发展趋势及相互关系的发掘,推断该领域可能出现的技术形态或技术发展点<sup>[1]</sup>。技术机会可以分为两类——内涵的技术机会和外延的技术机会。内涵的技术机会指现有技术的规范或性能有进一步改进的可能性;外延的技术机会指某项技术有转移到其它技术系统的可能性,转移到新系统中的技术或许能够比在原有系统中得到更加高效及有价值的应用<sup>[2]</sup>。

对于企业、高校及科研机构等主体来说,在研究了某项技术的机会后,才会决定是否启动新的技术创新项目。通过对潜在技术机会的挖掘和比较,并根据其差异进行选择,可以准确判断技术的未来走向。因此,技术机会可以为技术创新决策提供重要参考依据。

技术高速发展所带来的海量专利信息为技术机会的挖掘提供了可能。通过信息技术可以对已有专利技术进行深入研究与分析,寻找可能存在的技术机会,为技术创新和科技管理提供必要的决策信息支持。借助文本挖掘以及核主成分分析方法,本文以机器人产业

技术专利信息为研究对象,对其技术机会进行挖掘与分析。

## 1 核心方法

### 1.1 文本挖掘

数据挖掘是从海量数据中提取出未知并有潜在价值信息及知识的过程<sup>[3]</sup>。文本挖掘是数据挖掘的一种非常实用且有效的信息检索和分析技术,它能够对非结构化的文本形式进行处理,不仅可以由一定语言规则从文本文档中提取出相关主题和特征信息,还可以对大规模的文本文档集合进行处理,发现潜在的、有价值的模式和信息<sup>[4]</sup>。由于专利文件由非结构化文本格式构成,一般包含了重要的技术研究成果,篇幅基本较大,人工分析这些专利信息会占用大量的人力与时间。运用文本挖掘方法可以将专利文件转为结构化数据,用系统化方法从内容中提取关键词,帮助专利工程师、决策者或管理者快速进行专利分析,作出最有利的决策,减少风险。

### 1.2 核主成分分析法

通过对专利情报进行收集、整理,专利地图最终以图形、表格等方式呈现,通过分析专利地图中的信息,有助于把握技术研发重点并获得大量有价值的信息。

收稿日期:2013-08-13

基金项目:国家软科学研究计划项目(2011GXS5K095);国家科技支撑计划项目(2012BAH29F01);江苏高校哲学社会科学研究重点项目(2011ZDIXM049)

作者简介:龚惠群(1977—),女,湖南吉首人,东南大学经济管理学院博士研究生,东南大学经济管理学院讲师,研究方向为技术创新管理;刘琼泽(1988—),女,江苏南京人,东南大学经济管理学院硕士研究生,研究方向为技术创新管理;黄超(1977—),男,湖南衡阳人,博士后,东南大学经济管理学院副教授,研究方向为技术创新管理。

由于文本挖掘寻找出的关键词不能清晰显示技术的重要方向和战略意义,因而研究者需要借助专利地图把握各个专利之间的区别,并识别专利间复杂的联系。想要在专利地图中识别出技术空白点,需要根据每个专利的特点,将它们分布于专利地图的某个固定位置上。由于专利地图有可能包含多维度变量,形成的多维地图较为复杂,因而需要将关键词维度减少至易于解释的程度,例如二维或三维专利地图。

基于文本挖掘的专利地图是依据专利中关键词的相关性,将专利分布在所构建的三维或二维空间内。本研究没有采用聚类分析方法,而是改进已有的主成分分析方法,创新应用核主成分分析方法,其基本原理是通过核技巧将输入空间映射到某个高维空间,并在特征空间中实现主成分分析<sup>[5]</sup>。相较于一般的主成分分析方法,核主成分分析是一种非线性主分量分析方法,其实现更加简洁高效,可以最大限度地抽取指标信息。

采用核主成分分析方法将多个指标转化为少数综合指标即主成分来进行降维,势必会损失专利文件中包含的部分信息,但它最大程度保留了原始信息,并获得了比原始变量更加优越的性能,使每一个专利在生成的二维及三维专利地图上获得了准确的定位,使研究者能够更加清晰地获得隐含的技术信息。

## 2 基于专利文本挖掘的技术机会发现研究框架

### 2.1 专利文件检索及关键词矩阵构建

在专利数据库中确定规范的检索式,收集目标产业技术的专利文件并进行合理筛选,此时的专利文件为非结构化的文本格式,不适合构建关键词矩阵。因此,首先需要将文本文件转化为结构化数据,再采用文本挖掘的方法筛选高频关键词并形成关键词矩阵。在进行专利文本挖掘时,选取专利文件中的题目和摘要两部分内容作为文本挖掘的对象,以保证覆盖足够多技术领域的关键词,从而有效解决关键词区分度不高的问题。

形成关键词列表之后,检查每一个关键词在每条专利文献中是否存在,这样就形成一个关键词矩阵,这个矩阵以关键词 $(1, 2, \dots, n)$ 为横轴,专利文献 $(1, 2, \dots, m)$ 为纵轴。如果第 $j$ 个关键词在第 $i$ 个专利文件中存在,那么矩阵中点 $(i, j)$ 的数值就用关键词 $j$ 在专利文献 $i$ 中出现的次数 $p(p \geq 1)$ 来表示;如果第 $j$ 个关键词在第 $i$ 个专利文件中没有出现,则该点数值计为0。至此,就形成了一个数值不等的专利文献和关键词矩阵。

### 2.2 专利地图绘制

采用核主成分分析方法绘制专利地图,在减少关键词矢量个数的同时,最大程度地保留主成分代表的专利信息。在用 Matlab 软件进行核主成分分析时,得

到了3个核主成分构成的三维专利地图,当这3个主要成分占全部变量80%以上的信息时,就符合构建专利地图的条件。将此三维专利地图分别投影到每两个核主成分构成的平面上,最终形成3个二维专利地图。

### 2.3 技术空白点识别及评价

空白点即专利地图上被众多现有专利围起来的专利密度较小区域,该区域尚未有人或很少有人申请专利,但其本身具有广阔的发展空间,也为周边专利技术的发展提供了动力。这些区域代表的专利技术具有很强的技术创新潜力,企业或机构若成功开发出此空白区域的技术,必将享有先动优势并快速占领市场。通过分析周边专利技术,可以为空白区域的技术创新与突破提供参照,有利于探索发现技术创新点。

在识别出几个具有发展潜力的技术空白点后,需要对每个空白点进行重要性分析,以发掘出未来最有研发价值的技术机会。选取三项指标作为空白点重要程度的评价标准,分别为毗邻专利的平均被引频次、平均同族专利数及空白点密度。平均被引频次是指某一空白区域所有周边专利被其它专利引用次数的平均值。由于行业内被引用次数较多的专利往往是比较重要的专利,而且极有可能是该行业的核心专利及关键技术<sup>[7]</sup>,因而毗邻专利的被引频次平均值是技术空白点重要程度的体现。平均同族专利数是指某一空白区域所有周边专利同族专利数量的平均值。同族专利指不同国家授予同一项技术的不同专利号。在国外申请专利需要翻译和法律帮助等费用,比在本国申请昂贵得多,专利权属机构只会将那些未来最具有商业价值的发明在多国申请专利,以便保护今后投资和产品输出的独占权<sup>[6]</sup>。因此,同族专利规模会反映某一项专利权的覆盖率及潜在价值,是评价空白区域重要性的主要指标。空白点密度用所有周边专利的数量与空白区域面积之比表示,它的数值越大表明此区域周边专利分布越密,相互之间的竞争越激烈,则空白区域在未来可能有更多技术创新机会,因而该指标用来衡量技术竞争程度。

## 3 实证研究

机器人技术是现代科学与技术交叉和综合的体现。由于先进机器人的发展水平是一国综合科技实力的体现,许多先进工业国家已经将机器人技术列入本国21世纪高技术发展计划。考虑到机器人对我国未来经济建设及新兴领域发展的战略性作用,准确评估目前国内技术水平,研究机器人产业的技术机会,对于推动我国机器人产业整体实力发展具有十分重要的意义。

在本研究中,以机器人产业技术相关专利为对象进行实证分析。通过检索式:主题=(robot\*) AND 德温特手工代码=(T06-D07B OR X25-A03E OR

X25 - A03F OR U11 - F02A1 OR T01 - J07D1 OR T01 - S03 OR T01 - J08F OR X27 - D04R OR T01 - J07B OR T01 - J30A OR T01 - J05B4P), 在 DII 专利数据库中检索到机器人技术专利 20 331 条。由于专利数据数量相对较大且有些专利缺乏代表性, 故使用 Histcite 软件在当前主题数据库中选取被引用次数排名前 1% 的专利约 200 件, 作为专利文本挖掘的分析对象。为了提取机器人领域足够的专利信息, 选取专利文件中的题目和摘要两部分内容作为结构化文本进行文本挖掘。

### 3.1 关键词矩阵构建

将整理出来的 200 个专利文件的题目及摘要构成的文本文件, 按照单词出现频率高低得到高频词汇列表; 处理同义词并进行频次合并计算, 如将 Control 和 Controlling 进行同义词合并, 解决单词重复问题, 形成原始的关键词列表; 进行人工挑选二次挖掘之后, 删除 of、and 等与技术本身相关度不高的词语, 取出在题目和摘要中出现频率大于 60 次的高频词, 形成了最终的关键词列表, 共包含 30 个关键词, 如表 1 所示。

表 1 关键词列表

关键词	出现频次	关键词	出现频次	关键词	出现频次
Robot	993	Movement	135	Sensor	103
System	360	Operation	146	Point	142
Control	500	Object	113	Industrial	56
Position	314	Tool	113	Computer	68
Arm	270	Process	122	Joint	128
Data	135	Motion	103	Path	78
Wafer	183	Image	102	Effector	67
Unit	151	Mobile	79	Surgical	71
Method	113	Time	88	Motor	96
Device	129	Force	88	Axis	60

统计这 30 个关键词在 200 个专利的题目与摘要中分别出现的次数, 形成以下关键词矩阵(本文仅列出部分), 如表 2 所示。

表 1 关键词矩阵(部分)

专利	robot	system	control	position	arm	data	wafer	unit	...
1	4	1	3	1	4	0	0	0	...
2	4	3	4	4	1	0	0	0	...
3	3	4	2	2	3	1	0	0	...
4	1	5	0	0	0	0	0	0	...
5	8	5	10	2	2	0	0	0	...
6	4	2	1	0	3	0	0	3	...
7	5	0	4	1	0	0	0	0	...
8	4	0	1	1	0	0	0	0	...
9	3	4	6	1	5	0	0	0	...
10	9	1	6	0	2	0	0	0	...
...	...	...	...	...	...	...	...	...	...

### 3.2 核主成分分析

将关键词矩阵输入 Matlab 软件, 应用核主成分分析方法绘制专利地图。根据核主成分结果, 可以构建出许多不同的专利地图。为了使专利地图一目了然,

一般提取两个或三个核主成分绘制二维或三维专利地图。图 1 是由结果排名前三位的核主成分构建得到的专利地图, 其中的每一个点代表一项专利。由于每一个核主成分关联到非常多的关键词, 很难总结出其含义, 因而将不再分析每个核主成分的含义。

### 3.3 技术空白点识别

将图 1 的三维专利地图分别投影到 3 个平面上, 分别得到基于  $KPC_1$  和  $KPC_2$ 、 $KPC_1$  和  $KPC_3$ 、 $KPC_2$  和  $KPC_3$  构建的二维专利地图。从专利地图上识别出专利密度较低、分布较稀疏的区域, 将这些区域周边的专利相连, 形成 6 个红色线条围成的多边形区域, 这些区域即是要寻找的专利技术空白点, 如图 2、图 3、图 4 所示。

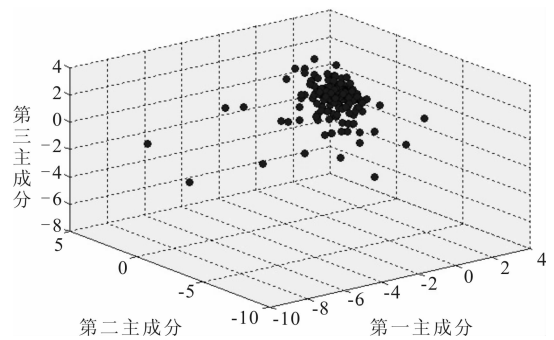


图 1  $KPC_1$ 、 $KPC_2$  及  $KPC_3$  构建的三维专利地图

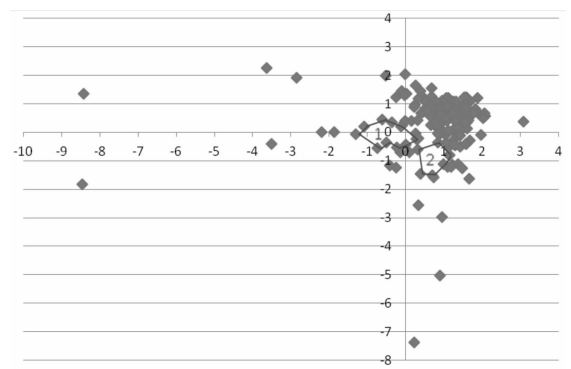


图 2  $KPC_1$  和  $KPC_2$  构建的专利地图空白点

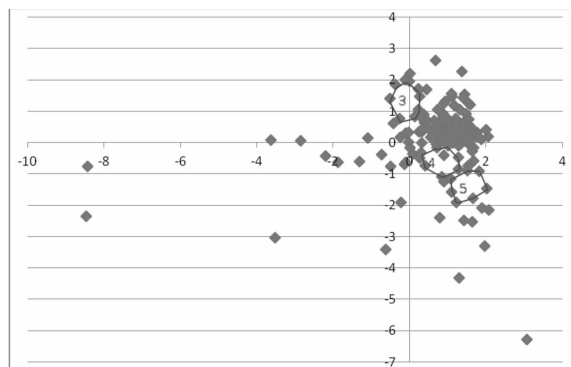


图 3  $KPC_1$  和  $KPC_3$  构建的专利地图空白点

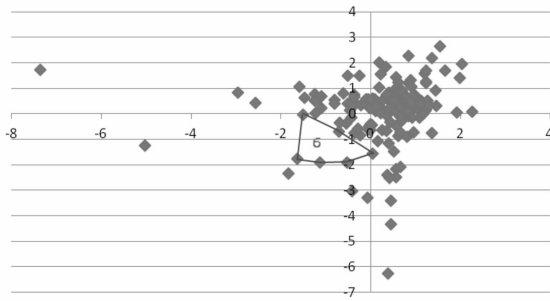


图 4 KPC<sub>2</sub> 和 KPC<sub>3</sub> 构建的专利地图空白点

对每个空白区域毗邻的专利进行统计,以便进行后续空白点重要性分析,如表 3 所示。

### 3.4 技术空白点评价

按照方法论阐述的方法,对每个专利空白点的平均被引频次、平均同组专利数量及密度分别进行计算,进行重要性分析,结果如表 4 所示。这 3 个指标为判定专利空白点的创新性和发展潜力提供参考。当某个空

白区域所有指标的值均相对较大时,说明该区域具有较大技术发展空间,应受到研究者的重点关注。

从表 4 的各项数据可以看出,空白区域 3 的各项指标都为最大值,因而空白区域 3 是机器人产业技术领域最具有发展潜力的专利空白点,最有可能成为新技术突破点,应当受到重视并进行持续深入研究。空白区域 1 及空白区域 5 的各项指标数值较高,因而具有一定发展潜力,可以进行深入分析与研究。空白 2、4 和 6 的各项指标数值均较小,表示其发展前景较差,不具有继续研究价值。

综上所述,空白区域 3 是未来机器人产业技术最需关注、最具有研究与发展潜力、蕴含技术机会最有价值的区域。表 5 列出了空白区域 3 所有相邻专利的专利号及标题。通过分析相邻专利所代表的技术,研究者可以对目前已有技术中还未突破的技术难点及还未受到重视的开发区域进行重点关注,以促进机器人产业技术进一步发展。

表 3 空白区域毗邻专利

空白点	1	2	3	4	5	6
毗邻专利	WO9516396-A	WO200176830-A	EP456103-A2	EP1068837-A	US5432417-A	US5049797-A
	WO9403113-A	EP597637-A	US5236432-A	WO2003036324-A	US5022695-A	WO9735241-A
	US6385509-A	WO9601977-A	WO200033726-A	US5400244-A	EP856457-A	WO9740435-A
	US5236432-A	US2002120362-A1	WO9614967-A	US6166504-A1	WO9710080-A	US7849629-N
	US4979949-A	US5049797-A	US5083073-A	EP820918-A	WO2003061917-A	EP440202-A2
	US5083073-A	EP1011035-A	US2002068992-A1	US5737500-A	EP440202-A2	WO2003061917-A
	US4999553-A		WO9912082-A	WO200002808-A	WO9740435-A	
	US7849629-N		EP1464452-A2			
	EP778611-A					
	EP880337-A					

表 4 空白区域重要性分析结果

指标	空白区域 1	空白区域 2	空白区域 3	空白区域 4	空白区域 5	空白区域 6
平均被引频次	38.70	25.50	44.50	18.86	20.57	16.83
平均同族专利数	6.40	6.33	7.88	5.57	7.71	7.17
密度	10.77	9.35	11.53	10.64	10.37	3.33

表 5 空白区域 3 的毗邻专利(专利号及标题)

专利序号	专利标题
EP456103-A2	Image directed robotic system for precise surgery-uses manipulating arm with surgical tool, optical tracking system and processor to monitor tool position relative to bone.
US5236432-A	Robot-aided surgical tool positioning unit-includes mounting carrying template, manually manipulated for correct orientation w. r. t. patient's bone
WO200033726-A	Image shifting system for telerobotic system has remote operator controls alternately controlling tools or camera
WO9614967-A	Planning method for planning motion of robot in free space in sheet metal bending station-involves building model of working space and forming proposed robot movements that are checked against model for collisions against obstacles
US5083073-A	Calibration of vision guided robot-includes slit light sequence to direct light beam onto target and producing correction values
US2002068992-A1	Relative position determining system between robot and workpiece support, includes controller which moves sensor close to pattern and detects position of each transition point on pattern
WO9912082-A	Industrial robot inspection system calibration method
EP1464452-A2	Robotic system controlling method for vehicle industry, involves operating controller by time coordinated motion instructions with unique label, to produce mutual robotic motion with common relative velocity profile

## 4 结语

本研究提出了基于专利文本挖掘及核主成分分析的专利地图构建方法,分析了发现机器人技术潜在技术机会的方法。其核心原理是准确识别专利地图中未被开发的具有发展潜力的空白区域,通过对这些空白区域毗邻专利的深入分析,对其技术创新机会的潜力大小和重要性程度进行比较评估,发掘其潜在技术突破信息,进一步明确空白区域的含义及其未来发展潜力。相比传统的聚类分析与主成分分析,文章所采用的核主成分分析更大程度地保留了原始专利信息,提高了后续研究的准确性。

对于文章挖掘出的具有进一步发展潜力的空白区域,可以在对周边专利技术进行分析研究的基础上,采用合适的路径突破这些新兴技术。对此,提出3条原创技术突破路径:①以基础理论研究突破为基础,进行应用研究和技术开发的“技术主导型”突破路径;②以市场创新和产品创新为基础,引导企业开展研发工作的“市场创新+产品创新拉动型”突破路径;③以公共产品技术创新为基础,探索企业技术创新的“公共产品技术创新带动型”突破路径<sup>[8]</sup>。

## 参考文献:

- [1] 李辉,乔晓东.基于科技文献的技术机会分析方法初探[J].情报杂志,2007(5):74-76.
- [2] 张妍.技术创新过程中的技术机会研究[D].沈阳:东北大学,2008:1-122.
- [3] 温志宏,赵鹏.浅析企业数据仓库与练级分析处理及挖掘技术的应用[J].太原师范学院学报:自然科学版,2004(2):37-39.
- [4] 王卫平,郭长旺.基于文本挖掘的企业竞争情报系统[J].现代情报,2004(9):188-189.
- [5] 王和勇,姚正安,李磊.基于聚类的核主成分分析在特征提取中的应用[J].计算机科学,2005(4):64-66.
- [6] 黄超,刘琼泽,仲伟俊.基于专利分析的机器人产业技术情报研究[J].情报杂志,2012(11):99-105.
- [7] 王庆民.专利信息的情报功能和专利情报分析[J].现代情报,2007(7):223-225.
- [8] 黄超,龚惠群,刘琼泽,等.我国重大原创技术突破的主要路径和政策措施研究[J].科技进步与对策,2012,29(10):105-109.

(责任编辑:侯慧波)

## The Research on Discover Robotics Industry Technology Opportunities Based on Patent Text Mining

Gong Huiqun, Liu Qiongzhe, Huang Chao

(School of Economics and Management, Southeast University, Nanjing 210096, China)

**Abstract:** The two-dimensional and three-dimensional technology gaps patent maps of robotics industry technology were drawn based on patent text mining and kernel principal component analysis, then the technology blank spots of robotics industry were found. The importance of the blank area was evaluated through three indicators includes the average cited frequency of patent blank point, the average number of patent families and the density, which has great significance for researchers to further develop the innovation breakthrough path of blank spots technology.

**Key Words:** Robotics; Patent Analysis; Text Mining; KPCA; Technology Opportunities