基于专利文献的技术演化分析方法研究*

方　曙　胡正银　庞弘燊　张　娴
中国科学院国家科学图书馆成都分馆　成都 610041

〔摘要〕在分析现有方法之不足的基础上，提出一种更完善的基于专利文献的技术演化分析方法：①采用分类号替代关键词作为专利文档聚类的基础；②采用基于语义的分类号－专利文档相似矩阵代替关键词－专利文档存在矩阵聚类；③采用更适合小样本聚类的系统聚类法。以石墨烯传感器技术为例，进行实证分析，绘制出石墨烯传感器技术层次语义网络图与技术演化图。研究结果显示，该方法可较好地应用于专利技术演化分析。

〔关键词〕语义分析　专利分析　技术演化　关键词语义网络
〔分类号〕G353.1

Study on the Method of Analyzing Technology Evolution Based on Patent Documents

Fang Shu　Hu Zhengyin　Pang Hongshen　Zhang Xian
The Chengdu Branch of National Science Library, Chinese Academy of Sciences, Chengdu 610041

〔**Abstract**〕This paper proposes an improved analysis method on the basis of an existing method of patent semantic analysis for emerging technologies. Compared with the existing method, this new method has the following improvements: ① using patent classifications instead of keywords as the basis of the patent documents clustering; ②using fuzzy similarity matrix of patent classifications-patent documents instead of keywords-patent documents existence matrix in clustering; ③ using hierarchical clustering analysis which is more suitable for small samples instead of k-Means clustering algorithm. Finally, the authors select the field of graphene sensor for empirical analysis, and draw semantic networks of technology keywords and technology evolution maps of graphene sensor. By comparison, the improved method can get more precise semantic network of technology keywords and clearer technology evolution map. The result shows that this new method can support better the patent analysis of technology evolution.

〔**Keywords**〕semantic analysis　patent analysis　technology evolution　keywords semantic network

1　引　言

技术演化分析是一种基于技术历史发展线索，描绘其发展历程，从中提炼出未来可能产生重大影响的新兴技术的方法。

专利文献是世界上最大的信息源之一，各国每年出版的专利文献占科技出版物的1/4，内容包含了世界科技信息的90%－95%[1]。相比其他信息资源，专利文献具有内容新颖、系统详尽、格式规范、分类科学等特点[2]，是一种理想的技术演化分析对象。

在利用专利语义信息，绘制技术关联图，分析技术演化趋势，进而预测新兴技术方面，中外学者进行了较广泛深入的研究。在专利语义分析方面，Nizar 等[3]通过对专利文献进行基于本体的语义标注来支持技术挖掘；姜彩红等[4]对中文专利摘要进行了语义抽取实验，为构建专利知识库提供语料基础。欧盟支持的 PATExpert 项目[5]是一个对专利文献的语义处理的实验型系统；Aureka 通过 Themescape 可对专利按主题聚类，并以地形图直观进行显示[6]。

在技术关联图方面，Yoon 和 Park[7]提出一种以关键词向量为基础，绘制专利网络关联图的分析方法；Christian 等[8]提出通过社会关系网络分析工具来对专利进行可视化分析。此外，通过文献引证关系来揭示技术之间的关联，也是一种常见的分析方法。

2　技术演化分析方法研究

2.1　Young 等[9]提出的分析方法

韩国学者 Young 等提出一种基于语义专利分析可

* 本文系中国科学院知识产权专项工作项目“中国科学院知识产权信息服务”（项目编号：Y11009）研究成果之一。
收稿日期：2011－02－21　修回日期：2011－06－09　本文起止页码：42－46　本文责任编辑：高　丹

视化的方法,用于新兴技术预测。方法流程为:关键词抽取、专利文档聚类、利用关键词在专利文档聚类中的层次分布关系分析技术演化,进而预测新兴技术。该方法具有操作简单、解读清晰、主题凝练充分等优点。

然而通过实践,发现这一方法存在以下不足:①文中通过计算关键词在专利文档聚类分组中的分布及其关系,来分析技术演化,但却采用关键词-专利文档存在矩阵作为聚类基础,有循环论证之嫌。此外,选定的关键词数量有限,难以作为聚类依据。②文中关键词-专利文档存在矩阵,通过计算关键词是否在专利文档中出现来进行赋值,如果出现,则为1,否则为0。这种赋值方式过于简单化,既没有考虑到关键词在专利文档中的分布特征,也没有考虑到关键词之间的语义特征,用来对专利文档聚类,效果较差。③文中采用k-Means算法聚类。k-Means作为经典聚类算法,要事先确定聚类的数目,进行分析次数较多。当观察值的个数较多时,采用"k-Means聚类分析法"较合适。而该方法分析专利数量一般较小,选择一些更直观的聚类算法,如"系统聚类分析法"将更方便准确[10]。

2.2 对Young等[9]分析方法的改进

针对以上不足,本文进行相应的改进,提出基于专利文献的技术演化分析方法,如图1所示:

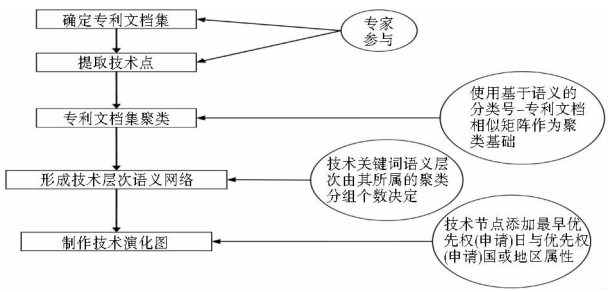

图1 基于专利文献的技术演化分析方法流程

2.2.1 确定专利文档集 与专家讨论,确定专利检索策略,进行专利检索,得到待分析专利文档集合。

2.2.2 提取技术点 从专利文档集中抽取技术点。请领域专家对自动抽取的技术点再进行人工筛选,最终确定专利文献涵盖的重要技术点,作为技术演化分析的基础。

2.2.3 专利文档集聚类 按技术相关性,对专利文档进行聚类。考虑到专利文献的特殊性,利用分类号对专利文档进行语义聚类。相对于关键词,分类号数量较多,特别是德温特专利数据库,其手工分类号(MC)由专家人工标引,较好反映专利技术属性,聚类效果更好,且作为关键词层次分析的基础,更具独立性。分类号与专利文档相似矩阵是聚类的基础,其计算不采用简单的存在与否进行判别,而是综合考虑分类号在专利文档中的分布特征及分类号之间的语义特征。相似矩阵算法参照时念云等[11]提出的计算文档与领域本体概念间关系算法,具体如下:

首先,考虑分类号在专利文档中的分布特征[11]:采用空间向量模型(VSM)描述专利文档与特征项之间的关系,将分类号作为一个特征项来计算。文档中特征项权重计算,通常考虑特征项在文档中出现频率、集合中文档总数、包含特征项的文档数等因素。特征项在文档集中出现的次数越高,包含该特征项的文档数目越少,那么该特征项对于该文档来说,其独特性就越强,其权重也就越高。公式如下[11]:

$$Weight[i,j] = f_{ij} \times log(N/n_i)/max_i(f_{ij}) \quad (1)$$

其中,f_{ij}:特征项i在文档j中出现的频率;N:集合中所有文档数;n_i:包含特征项i的文档数。

然后,考虑分类号之间的语义特征[11]:分类号之间有明确的上下位关系,可认为是一个领域本体。通过计算本体中词汇之间的相关性,可更准确描述特征项与专利文档之间的关系。对公式(1)进行修正,具体如下[11]:

$$NWeight[i,j] = f_{ij} \times log(N/n_i)/max_i(f_{ij}) + \sum_{m=1,m\neq i}^{L} weight[m,j] \cdot \theta_{im} \quad (2)$$

其中,L:文档j中特征项总数;θ_{im}:特征项i与特征项m之间的语义相似度。对领域本体中概念间的语义相似度计算方法,黄果等[12]进行了详细归纳,考虑到分类号特点,可采用基于距离的语义相似度计算模型,具体不作详述。

将式(2)得到的$NWeight[i,j]$作为分类号与专利文档之间的相似矩阵,以此为基础得到专利文档之间相似性矩阵。在专利文档相似矩阵的基础上,使用系统聚类法对文档进行聚类。

2.2.4 形成技术层次语义网络 计算聚类结果分组中包含的技术点。每一个分组用一个节点表示,如某一技术点出现在多个组中,那么将该技术点移到一个新的节点中,并调高其权重,用有向线段连接该节点与包含技术点的其他节点。对所有分组中的技术点进行相同的处理,进而形成技术层次语义网络。技术点出现在分组中的次数作为该节点的频率权重加以标记。

2.2.5 制作技术演化图 在形成的技术演化语义网络中,给每一个技术点节点添加最早优先权(申请)日与优先权(申请)国或地区项。技术节点的最早优先

权(申请)日与优先权(申请)国或地区定义如下:包含该技术点的专利文献集中,最早的专利优先权日或最早申请日及相应的专利申请国或地区。以最早优先权(申请)日作为横轴,技术点频率作为纵轴,绘制出初步技术演化历程图。

3 实证分析:以石墨烯传感器为例

石墨烯是一种新型的二维纳米材料,具有极高的电导率、热导率及出色的机械强度,是制作高灵敏度传感器的上佳材料。

3.1 绘制石墨烯传感器技术演化图

3.1.1 确定专利文档集 选取德温特专利数据库为分析检索数据源,专利检索策略如表1所示:

表1 专利检索式

编号	检索式
#1	S = (sensor * or transducer * or (sensing same (element * or devic * or unit * or organ * or apparatus * or system *)) or (sense same organ *) or Photosensor * or microsensor * or chemosensor * or multisensor * or hypersensor *) 数据库 = CDerwent, EDerwent, MDerwent 入库时间 = 所有年份
#2	TS = (graphene *) 数据库 = CDerwent, EDerwent, MDerwent 入库时间 = 所有年份
#3	#1 and #2 数据库 = CDerwent, EDerwent, MDerwent 入库时间 = 所有年份

以检索式#3进行检索,得到待分析专利文档80篇。

3.1.2 提取技术点 从专利文本域字段(标题、摘要、技术焦点)抽取关键词,请专家人工筛选,得到石墨烯传感器关键词,如表2所示:

表2 待分析关键词

编号	关键词	最早优先权(申请)国或地区、最早优先权(申请)日
1	carbon nanotube	JP 11 Dec 2002
2	silicon carbide substrate	US 09 Apr 2009
3	graphene sheet	JP 07 Feb 2003
4	graphene struct	US 25 Mar 2003
5	substrate	US 12 Sep 2001
6	graphene layer	US 03 Oct 2003
7	conductive material	US 25 Mar 2003
8	carbon fibers	DE 11 Jul 2002
9	carbon material	JP 11 Dec 2002
10	conductive polymers	DE 05 Nov 2004
11	graphite oxide	US 14 Oct 2005
12	carbon nano - tube array sensor	CN 22 Feb 2010
13	chemical sensor	US 13 Feb 2007
14	conductometric sensor film	US 18 Jul 2006
15	conductometric sensor manufacture	US 01 Nov 2005
16	graphene field effect transistor sensor	US 12 Sep 2007
17	magnetoresistive sensor	US 28 Feb 2008
18	molecular sensor	US 12 Sep 2007
19	transistor type sensor	FI 26 Feb 2009
20	chemical vapor deposition	US 25 Mar 2003

3.1.3 专利文档集聚类 利用MC对专利文档进行聚类。统计出80篇专利文档中共含有670种MC,按公式(2)得到MC - 专利文档相似矩阵。以此为基础,使用系统聚类法把专利文档分成11个大类。

3.1.4 形成技术层次语义网络 计算聚类结果分组中关键词分布情况,如表3所示:

表3 关键词分组

聚类组号	聚类分组中关键词编号
1	2, 3, 5, 6, 9, 19
2	1, 3, 4, 5, 7, 10, 11, 13, 16, 18, 20
3	1, 3, 5, 9
4	1, 5, 12, 13, 14
5	1, 5, 8, 10
6	1, 3, 4, 5, 6, 7, 17, 20
7	1, 3, 5, 6, 7, 9, 10, 20
8	1, 5, 6
9	1, 3, 6, 7, 8, 11
10	1, 3, 8, 11, 15
11	1, 3, 5, 6, 7, 11, 13

统计关键词出现在分组中的次数,结果如表4所示:

表4 关键词频率及最早优先权(申请)日

节点	最早优先权(申请)国或地区、最早优先权(申请)日	频次	关键词(编号)
1	JP 11 Dec 2002	10	carbon nanotube (1)
2	US 12 Sep 2001	9	substrate (5)
3	JP 07 Feb 2003	8	graphene sheet (3)
4	US 03 Oct 2003	6	graphene layer (6)
5	US 25 Mar 2003	5	conductive material (7)
6	US 14 Oct 2005	4	graphite oxide (11)
7	US 25 Mar 2003	3	chemical vapor deposition (20)
8	DE 11 Jul 2002	3	carbon fibers (8)
9	JP 11 Dec 2002	3	carbon material (9)
10	DE 05 Nov 2004	3	conductive polymers (10)
11	US 13 Feb 2007	3	chemical sensor (13)
12	US 25 Mar 2003	2	graphene struct (4)
13	CN 22 Feb 2010	1	carbon nano - tube array sensor (12)
14	US 09 Apr 2009	1	silicon carbide substrate (2)
15	US 01 Nov 2005	1	conductometric sensor manufacture (15)
16	US 12 Sep 2007	1	graphene field effect transistor sensor(16)
17	US 28 Feb 2008	1	magnetoresistive sensor (17)
18	US 12 Sep 2007	1	molecular sensor (18)
19	FI 26 Feb 2009	1	transistor type sensor (19)
20	US 18 Jul 2006	1	conductometric sensor film (14)

在表3和表4的基础上,基于关键词角度和改进的方法绘制技术层次语义网络图,见图2。

3.1.5 制作技术演化图 以最早优先权(申请)日作为横轴,关键词频次作为纵轴绘制出初步技术演化图,见图3。

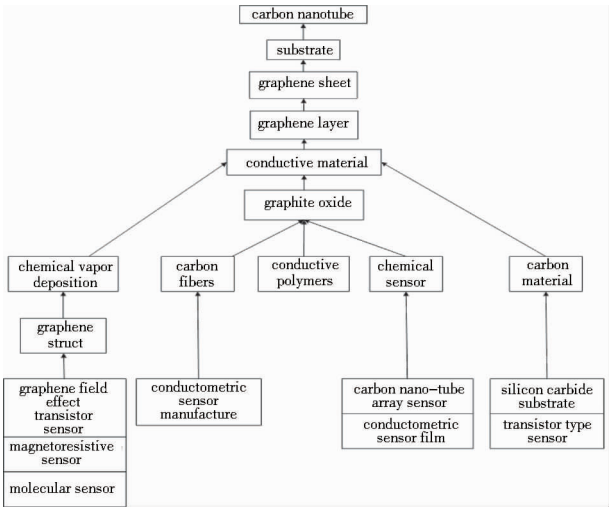

图 2 石墨烯传感器技术层次语义网络(改进的方法)

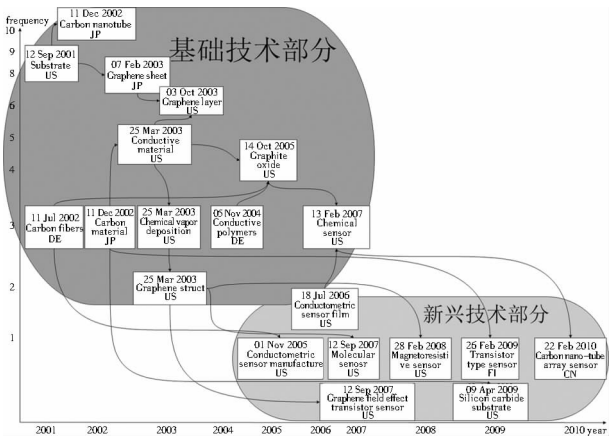

图 3 石墨烯传感器技术演化(改进的方法)

3.2 技术演化图解读

图 3 左上角所示技术点,时间出现较早,同时又在多个不同技术分组中出现,可认为是较早出现的基础技术领域,如:基底(substrate)、碳纳米管技术(carbon nanotube)、石墨薄膜(graphene sheet)。右下角所示技术点,则是最近才出现,且只在较少的技术分组中出现,可以认为是最新出现的较具体的技术发展方向。从图 3 右下角可以看出石墨烯传感器技术在朝以下几个方向发展:晶体管式传感器(transistor type sensor)、碳纳米管阵列传感器(carbon nano-tube array sensor)、碳化硅基底(silicon carbide substrate)。

以时间为序,技术节点之间的连线在一定程度上反映了技术之间的演化趋势。后继技术节点与其前驱技术节点同属于某一个技术分组,前驱技术节点与后继技术节点之间的技术演化趋势大致有如下几种:

• 前驱技术经过分化成为多种后继技术,可认为:石墨结构(graphene struct)经过一定时间的发展,分化出分子传感器(molecular sensor)、磁阻传感器(magnetoresistive sensor)、石墨烯场效应晶体管传感器(graphene field effect transistor sensor)等后继技术。

• 几种前驱技术合并成某一种后继技术,可认为:碳纤维(carbon fibers)、导电材料(conductive material)、导电聚合物(conductive polymers)等前驱技术与其后继技术节点氧化石墨(graphite oxide)存在技术合并、归并关系。

• 前驱技术应用于后继技术当中,可认为:碳纳米管阵列传感器(carbon nano-tube array sensor)可作为化学传感器(chemical sensor)技术的一个具体应用。

通过专家对技术演化图的解读,并结合相关文献调研,可明确未来石墨烯传感器发展重点,对进一步专利分析有一定指导作用。

3.3 与 Young 等方法的对比

Young 等方法在聚类时,当分组数较大时,出现较多关键词丢失,不能分到任何一组中的现象;当分组数偏小时,又出现关键词属于绝大多数分组的情况。图 4、图 5 是聚类分组数为 6 时,使用原方法得到的关键词技术层次语义网络图和演化图,出现关键词(conductometric sensor manufacture)丢失及部分关键词属于绝大多数分类的现象。专家解读认为:图 5 中关键词之间语义层次与演化关系不甚清晰,且关键词之间存在较多无意义的关联;而图 3 中,关键词分类语义层次及演化更为精确,对进一步的专利情报分析有更具体的指导意义。

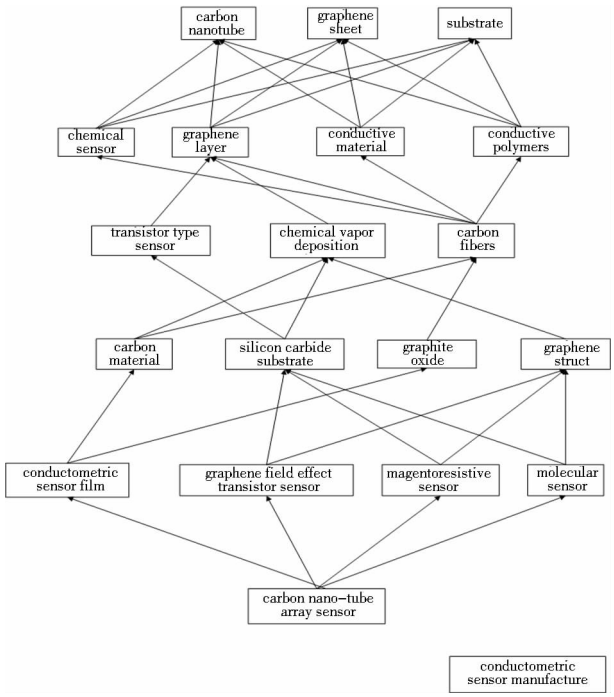

图 4 石墨烯传感器技术技术层次语义网络(Young 等的方法)

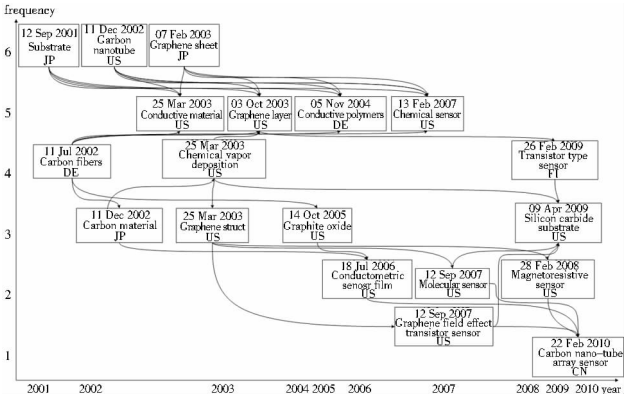

图 5　石墨烯传感器技术演化(Young 等的方法)

4　结　语

本文对 Young 等提出的利用专利可视化技术预测新兴技术方法进行了改进。重点是:采用分类号替代关键词作为专利文档聚类的基础,避免了循环论证的缺陷;采用基于语义的分类号 - 专利文档相似矩阵替代关键词 - 专利文档存在矩阵作为聚类的基础,该矩阵既考虑分类号在专利文档中的分布特征,又考虑分类号之间的语义关系;采用更适合小样本聚类的系统聚类分析法代替需要多次迭代的 k-Means 聚类算法,聚类过程更清晰简单。最后对石墨烯传感器进行实证分析,绘制其层次语义网络图与技术演化图。

通过对比及专家解读显示本文的方法对分析专利技术演化进程更具有指导意义。下一步的工作将是使本方法更加标准化,并应用到其他技术领域,以进一步检验本方法的有效性。

致谢:中国科学院合肥智能机械研究所副所长刘锦淮研究员,中国科学院苏州纳米技术与纳米仿生研究所、中国科学院"百人计划"研究员程国胜,四川大学吕弋教授等,对本研究给予了大力支持与诚挚帮助,特此致谢!

参考文献:

[1] 吕祥惠,仇宝艳,乔鸿. 基于本体的专利知识发现体系研究. 计算机与信息技术,2008(7):43 - 46.

[2] 王朝晖. 专利文献的特点及其利用. 现代情报,2008(9):151 - 152,156.

[3] Nizar G, Khaled K, Rose D. Supporting patent mining by using ontology-based semantic annotations//Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Washington:IEEE Computer Society,2007:435 - 438.

[4] 姜彩红,乔晓东,朱礼军. 基于本体的专利摘要知识抽取. 现代图书情报技术, 2009(2): 23 - 28.

[5] Leo W, Sören B, Barrou D,etc. PATExpert: Semantic processing of patent document. [2010 - 04 - 05]. http://www.vis.uni-stuttgart.de/ger/research/pub/pub2006/samt06 - wanner.pdf.

[6] 陈燕,邓鹏,李芳. AUREKA 信息平台介绍. 中国发明与专利,2007(5):63 - 64.

[7] Yoon B, Park Y. A text-mining-based patent network:Analytic tool for high-technology trend. The Journal of High Technology Management Research, 2004,15(1):37 - 50.

[8] Christian S, Adam B, Reinhard S. Visualizing patent statistics by means of social network analysis tools. World Patent Information, 2008, 30(2):115 - 131.

[9] Young G, Jong H, Sang C. Visualization of patent analysis for emerging technology. Expert Systems with Applications, 2008, 34(3):1804 - 1812.

[10] 吴元奇, 冯荣扬. 聚类分析计算方法的理论及结果比较. 湛江海洋大学学报,2002, 22(1):57 - 63.

[11] 时念云,杨晨. 基于领域本体的语义标注方法研究. 计算机工程与设计,2007,28(24): 5985 - 5987.

[12] 黄果,周竹荣,周亭. 基于领域本体的语义相似度计算研究. 计算机工程与科学,2007,29(5): 112 - 117.

〔作者简介〕方　曙,男,1957 年生,中国科学院国家科学图书馆副馆长、成都分馆馆长,博士,博士生导师,发表论文 70 余篇;胡正银,男,1979 年生,高级工程师,硕士,发表论文 10 余篇;庞弘燊,男,1983 年生,博士研究生,发表论文 10 余篇;张　娴,女,1973 年生,副研究员,硕士,发表论文 40 篇。

(上接第 127 页)

[15] 汪传雷,谭星,郑红军. 基于 ACSI 的电信内容增值服务的用户满意度模型及其影响因素研究. 图书情报工作,2009,53(11):48 - 52.

[16] 曹兴中. 基于客户生命周期的 ACSI 模型修正及应用研究[学位论文]. 哈尔滨:哈尔滨工业大学,2006:3 - 5.

[17] 林卉. ACSI 模型的因果关系检验研究. 统计与决策,2005(2):22 - 23.

〔作者简介〕叶凤云,女,1980 年生,讲师,博士研究生,发表论文 7 篇。
汪传雷,男,1970 年生,教授,博士后,硕士生导师,发表论文 60 余篇,出版专著 1 部。