doi:10.3772/j.issn.2095-915x.2018.04.008

基于文本挖掘的技术路线图构建研究

潘小换　蒋保建

中北大学　经济与管理学院　太原　030000

摘 要 针对传统的技术路线图设计方法耗时耗力、容易受到专家主观的限制等问题，通过对中文专利数据进行文本聚类，文本关联分析等过程，提出了将文本挖掘应用于技术路线图绘制的解决方案。该方案以中文专利数据为数据源，以文本挖掘为核心，结合社会网络分析和专家经验。综合运用定性分析和定量分析的方法，为技术路线图的绘制提供了一个新的思路。同时，本文以石墨烯在锂离子电池中应用领域为例，展开实证研究，绘制了石墨烯在锂离子电池应用领域的技术路线图，验证了本方法的科学性和有效性。

关键词： 文本挖掘；技术路线图；技术关键词；共现矩阵；聚类

中图分类号： F270; G306

开放科学（资源服务）标识码（OSID）

Construction of the Technology Roadmap Based on Text Mining

PAN Xiaohuan　JIANG Baojian

North University of China, School of Economics and Management, Taiyuan 030000, China

Abstract Because the traditional design of the technology roadmap is time-consuming and vulnerable to subjective restrictions, this paper used text mining technology through the text association analysis and text clustering of Chinese patent data to put forward the solution for roadmap drawing. The project took the Chinese patent data as the data source, utilized the text mining, and combined the social network analysis and expert experience. The method of qualitative analysis and quantitative analysis provided a new way for drawing the technology roadmap. Meanwhile, this research took the graphene in lithium ion battery

基金项目： 山西省高等学校哲学社会科学研究基地项目“山西省煤基低碳产业创新链技术预测研究”（2016325）。

作者简介： 潘小换（1992-），硕士研究生，主要研究方向：数据挖掘、创新管理，E-mail: zhengchenging@sina.com；蒋保建（1993-），硕士研究生，主要研究方向：数据挖掘、创新管理，E-mail: 729102673@qq.com。

doi:10.3772/j.issn.2095-915x.2018.04.008

application field as an example, draw the graphene lithium ion battery application technology roadmap to verify the correctness and effectiveness of this method.

Keywords：Text mining; technology roadmap; technological key words; co-occurrence matrix; clustering

0 引言

自上世纪七十年代，技术路线图首次被美国汽车行业应用，技术路线图作为一种综合管理工具被广泛用于技术管理中。技术路线图主要用于了解技术发展情况、识别重要技术，进行技术预测减少技术投资风险等。目前技术路线图的绘制主要以定性方法为主，比如德尔菲法、SWOT 分析法、情景分析法等，这些方法耗时耗力，绘制成本较高而且更新频率低。现代社会信息产生迅速，准确的把握技术发展的态势，预测技术发展的趋势，成为各国政府、企业和科研人员急需解决的问题。文本挖掘技术的发展使得快速处理海量数据成为可能。由于目前技术路线图绘制主要依靠定性方法，耗时耗力，而且容易受到专家主观因素的影响，所以本文尝试利用文本挖掘的相关方法，结合专家知识，综合运用定性分析和定量分析，构建技术路线图绘制模型，以期达到在保证科学性和有效性的基础上，减少专家依赖，简化技术路线图绘制流程的目的。

1 技术路线图研究综述

目前对于技术路线图没有一个统一的定义，例如 L.M.Branscomb 认为技术路线图是以科学知识和预测为基础，关于技术前景的共识 [1]。R.Phaal 认为，技术路线图将组织目标、技术资源和外界环境之间建立了一个动态联系 [2]。而本文认为技术路线图是指，利用简单图形，文字描述技术发展的脉络及技术之间的关系，帮助使用者技术发展情况、识别重要技术，进行技术预测减少技术投资风险。

在技术路线图的绘制方法上，很多学者有不同的尝试。Gyungmi Jin 等提出了利用文本挖掘和质量功能展开（QFD）的方法去绘制技术路线图 [3]，但是该方法还是更侧重于定性分析，仅仅通过文本挖掘将技术关键词抽取出来，后续的工作全部依靠专家进行处理。张嶷等提出了基于文献计量学结合专家知识的技术路线图绘制方法，并尝试构建电动汽车领域的技术路线图 [4]，但是该方法并没有考虑技术关键词在文档中的分布情况。李牧南利用文本挖掘基本理论与工程管理知识对小微企业的技术路线图规划进行的研究 [5]，该方法需要依赖大量的专家判断。张颖采用 XML 的专利文档分析方法识别湖北省大重型数控机床关键技术，绘制出相关的技术路线图 [6]，但是该方法需要将专利数据转化为 XML 文档数据，然后再进行处理，转化过程需要耗费大量的人力。谭全元等提出了基于 WebGIS 的湖南省黑茶产业化技术路线图研制，包括基础资料的收集途径及加工方法、信息管理系统的设计和技术路线图的绘制 [7]。

但是该方法有很大的局限性，不适合大规模的推广。刘彤将多重关系专利网络应用到产业技术路线图的构建中，并以锂电池产业为例进行了实证分析[8]。该方法需要提前了解所研究行业的基本情况，不利于行业外的研究人员进行研究，适用范围小。李剑敏等在分析了基于情景分析的产业技术路线图的内涵和特点，在此基础上，依据产业技术路线图集成规划过程，提出了融合情景分析的产业技术路线图集成规划过程分析框架[9]，但是该方法容易忽略一些产业环境中不明显的但是对产业未来带来发展带来重大影响的环节，而且也没经过实证研究。

目前技术路线图的绘制主要还是依赖定性方法，虽然有不同的学者尝试将定量方法应用到绘制中，但是存在各种各样的问题。本文基于对现有技术路线图绘制方法的研究，提出一种基于文本挖掘的技术路新图构建方法，并结合社会网络分析，以期达到减少技术路线图绘制过程中对于专家经验的依赖，以及减少新兴产业技术创新和规划中的不确定性，从而提高技术路线图的可靠性和科学性，全面把握国家核心技术发展目标的目的。并且本文以石墨烯在锂离子电池中应用领域为例，展开实证研究，绘制了石墨烯在锂离子电池应用领域的技术路线图，验证了本方法的科学性和有效性。

2 基于文本挖掘的技术路线图绘制方法研究

2.1 研究思路

本文以中国国家知识产权局的专利数据（SIPO）为数据源，结合文本挖掘和社会网络分析，构建相关产业的技术路线图。本方法主要分成三个阶段，整体框架如图 2 所示，第一阶段技术主题词确定，主要是从专利的标题和摘要中获取技术主题词和技术主题词的时间维度；第二阶段主要是以社会网络分析为核心，确定技术之间的关联关系和技术主题聚类；第三阶段为技术路线图的绘制，这一阶段是本方法定性分析和定量分析的核心阶段，根据上两个阶段所形成的技术主题词和主题聚类，根据技术主题的 IPC 号，结合专家知识，绘制技术路线图。本方法最终形成的技术路线图样图如图 1 所示，图中横轴代表技术主题词第一次出现的时间节点，纵轴表示技术发展阶段，图中的技术主题是通过文本挖掘的方法从专利文献中获取的核心关键词。不同技术主题所形成的聚类主要借助社会网络分析进行确定。

2.2 实施步骤

1）数据源

本文所采用的数据源为中国国家知识产权局的专利数据（SIPO），因为专利数据能够准确的反映技术的特征。专利文献是商业经营、科学研究和技术发展最重要的信息来源之一，专利信息数量庞大，价值突出，来源广泛[10]。国家知识产权局中文专利数据被认为是中文科技工作者最直接最准确的数据源，对于中国科技能力的判定和预估未来发展方向具有非常重要的意义。

2）分词方法改进

专利数据因为其专业性导致现有的分词算法并没有很好的分词质量，所以本文利用文本

doi:10.3772/j.issn.2095-915x.2018.04.008

挖掘技术，依据专利标题的特点，设计了提取专利主题词的方法。例如，专利标题为“本发明公开了一种使用寿命长的锂离子动力电池”，可以与“本发明公开了一种 * 的 *”这个模式进行匹配，得出其中的关键词为“使用寿命长”和“锂离子动力电池”。然后将这些抽取出的关键词作为词典，加入到现在的分词算法中。这样处理后，分词的准确率就会大大提高，同时也省去了寻找词典的步骤。在分词的过程中，程序可以在分词的同时获取每个关键词的时间维度。关键词的时间维度的确定依据关键词在专利文献中最早申请时间。

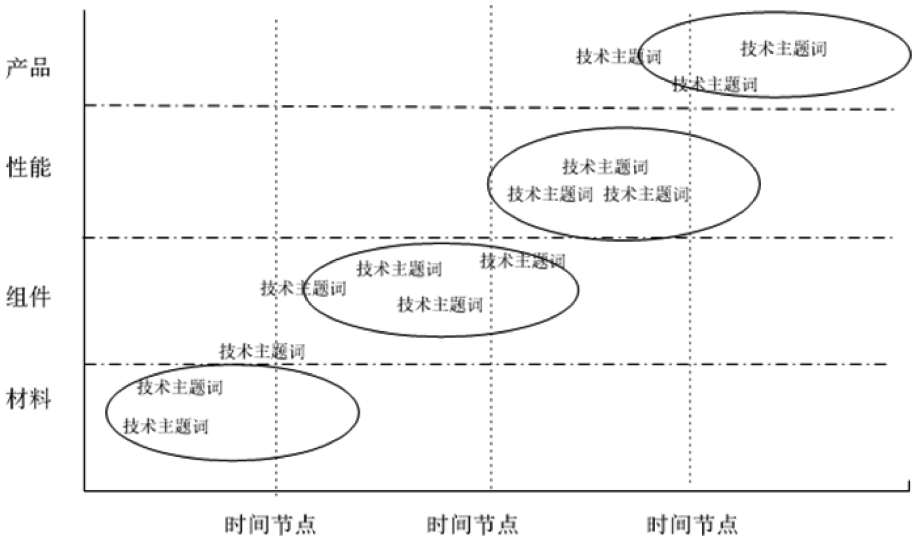

图 1 技术路线图样图

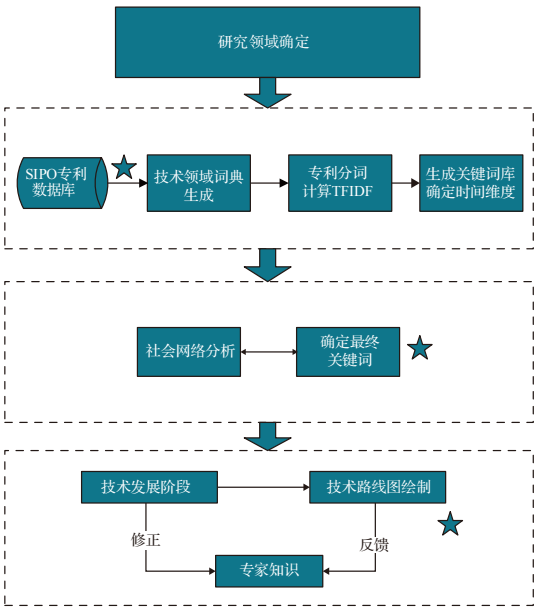

图 2 技术路线图绘制模型流程图

3）构建技术共现矩阵

首先利用 TF-IDF 算法 [11] 进行初步的技术关键词的筛选得到候选技术主题词，然后借助共词分析对技术主题词确定和技术主题词关联关系进行确定。在候选技术主题词的基础上构建技术主题词共现矩阵。技术主题词 kw1 和 kw2 在同一专利文档中出现，则数量加 1，这样能更好的反映技术主题词在文档中的共现关系，体现技术主题词的分布特征。利用 Jaccard 相似系数 [12] 对技术主题词共现矩阵进行标准化处理，获得技术主题词共现强度矩阵。Jaccard 相似系数，其计算公式（1）

$$J(A,B)=\frac{A\cap B}{A+B-(A\cap B)} \quad \text{（公式 1）}$$

在公式（1）中，J(A,B) 为技术主题词 A 和技术主题词 B 的共现强度，A 和 B 分别表示技术主题词 A 和 B 出现的次数，A ∩ B 表示技术主题词 A 和技术主题词 B 共现的次数。Jaccard 系数的数值范围为 0~1，数值越大表示关键词之间的关联程度越大。

4）技术关联图谱生成

将技术主题词关联强度矩阵导入到 Ucinet 中，利用 Netdraw 进行绘图，由于此时候选技术主题词数量非常多，网络密度非常大，可视化效果不强。需要进一步调整阈值，来提高可视化效果同时去掉不重要的关键词。当我们使用不同的阈值时，网络的密度不一样，但是网络的大致结构不会发生改变，也就是说，阈值的大小不影响聚类的结果，只影响聚类中节点和联系的数量 [13]。阈值调整完成后，得到清晰明了的图像，邀请领域专家或者通过邮件的方式对候选技术主题词进行评价商定，确定最终技术主题词。此时技术主题词数量比之前候选技术主题词的数量少了很多，领域专家可以根据自己的专业背景很快的对现在的技术主题词进行甄别判定，最终确定领域技术主题词，并获取技术主题词关联图谱。

5）文本聚类分析

技术主题词确定后，然后使用 Girvan-Newman 算法对技术关联图谱进行技术聚类分析。通过该算法能够准确的识别相似的技术，对技术主题词在技术路线图中定位起到辅助作用。Grivan-Newman 算法通过移除网络中的边来将整个网络划分为合适的社团结构，运用此算法可以将原来的网络分割成为任意数目的社团 [14]。社团结构作为社会网络拓扑结构的重要方面，对其研究有着重要的应用价值。社团发现既可以使人们从社团结构的整体功能得到其中个体在网络中的作用，又可以从整体上把握整个网络的结构和未来走向。

6）关键技术识别

我们采用社会网络分析中的中介中心度作为测量关键技术的重要指标。社会网络中的中介中心度（betweenness centrality）是指网络中行动者作为中介者的能力，也就是说行动者对于资源的控制程度 [15]。在技术网络中，“桥”也被称为“切开点”或者“结构洞”，往往是连接技术和新兴技术的关键节点。“桥”代表的技术往往具有集成、带动和辐射作用，能够培育新的增长点 [16]。因此，本方案采用中介中心度指标来探测关键技术及其重要程度。中介中心度大的关键词节点，在整个技术网络中起到枢纽的作用，往往是其他技术的基础，连接着不同的子网络，这样的技术主题词可以被认为是整个技术网络的关键技术。

7）技术路线图的绘制

本步骤主要任务是根据前几步获得的信息绘制技术路线图，本步骤以定性方法为主，主要依赖于领域专家的知识。首先，根据技术主题词和技术主题词在专利信息中首次出现的时间，确定技术主题词在时间维度的定位。其次，利用技术主题词的主 IPC 号，确定技术主题词在技术路线图纵轴的定位，完成初始技术路线图的绘制。初始技术路线图的绘制主要以专利数据为支撑，保证数据的客观性。根据之前所确定的技术主题词聚类所属的不同的技术阶段和每个技术主题词所处的时间维度，将每个技

doi:10.3772/j.issn.2095-915x.2018.04.008

术关键词绘制到技术路线图中的相应的地方。绘制完成后可以发现，属于同一聚类的技术主题词相对集中，词之间的距离比不同的聚类中的词要近。

技术路线图绘制是本方案定性分析与定量分析的关键部分。完成初始技术路线图后，要结合领域专家进行技术路线图的修正。通过邀请领域专家召开数次相关的专家讨论会，对初始技术路线图进行反馈和修正，形成最终版技术路线图。

技术路线图的绘制阶段主要以领域专家为核心，初步的技术路线图完成以后，要经过数次专题研讨会等形式充分征求专家的意见，反复的对初始的技术路线图进行修正，获得最终版的技术路线图。通过结合专家知识和专利信息保证整个方法的科学性和客观性。

3 实证研究

石墨烯作为一种新型的单原子层二维碳材料，具有超大的比表面积，优异的导热、导电、力学性能以及化学稳定性等，可作为锂离子电池电极材料中的导电添加剂，对负极材料进行掺杂改性或者直接作为负极材料进行应用[17]。2015 年 9 月，国家制造强国建设战略咨询委员会发布的《中国制造 2025》重点领域技术路线图里明确未来十年我国石墨烯产业的发展路径，总体目标是“2020 年形成百亿元产业规模，2025 年整体产业规模突破千亿元”。因为国家政策和前景看好等原因，石墨烯产业即将进入快速发展阶段，石墨烯在锂电池领域的应用也更有研究价值。所以本文以石墨烯基锂离子电池为例，构建技术路线图，为企业研发指明方向，同时验证本方法的有效性和科学性。

本文以 SIPO 专利数据库为数据源，咨询了相关领域专家，确定了检索策略，利用 C# 程序一共获取了 5409 条专利数据（2008 年值 2016 年 12 月 1 日）。经过数据清洗，去除无效的数据后剩余 4613 条数据。通过自己写的 python 程序，构建技术领域的词典，然后利用 jieba 分词进行分词。利用 TF-IDF 算法选出每个专利文档的关键词，根据关键词在文档的共现关系形成共现关系矩阵，按照公式（1）对技术主题词共现矩阵进行标准化处理形成关键词共现程度矩阵，将数据导入到 Netdraw 中，经过验证，设置其阈值为 0.011，可以看到明显的分组情况，如图 3 所示。进过与领域专家的反复确定，获得技术关键词如表 2 所示。

通过图 3，可以看出石墨烯锂离子电池领域围绕性能、负极形成两个大的聚类，同时围绕着材料和氧化石墨烯形成第三个大的聚类。通过三个聚类间的潜在关系，在与领域专家进行充分的交流之后，得出这与技术路线图的纵坐标相吻合，表明了石墨烯锂离子电池领域从技术材料研究向技术产品发展的必然趋势。

基于表 2 中石墨烯锂离子电池领域技术主题词表及其技术主题词首次出现的时间，参考图 3 所获得的技术关键词之间的关联关系，并与领域专家进行邮件和相关的研讨会形式的讨论，完成技术主题词在时间维度和技术发展阶段的定位，形成石墨烯锂离子电池领域技术路线图初稿。然后通过数次研讨会对初稿进行反馈和修改，最终形成石墨烯锂电池技术路线图如图 4 所示。

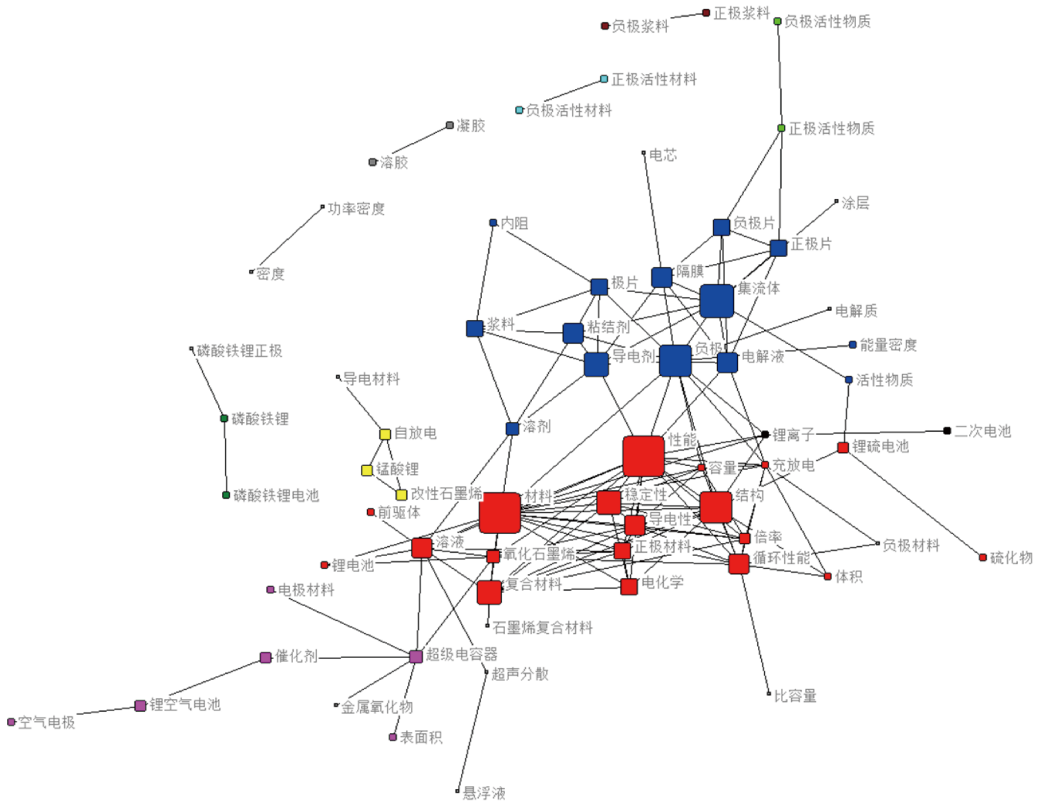

图 3 技术关主题词关联图谱

表 2 技术主题词表

技术关键词									
1	倍率	15	电芯	29	活性物质	43	磷酸铁锂正极	57	体积
2	比容量	16	二次电池	30	极片	44	硫化物	58	涂层
3	表面积	17	动力电池	31	集流体	45	锰酸锂	59	稳定性
4	材料	18	负极	32	浆料	46	密度	60	性能
5	超级电容器	19	负极材料	33	结构	47	内阻	61	悬浊液
6	充放电	20	负极活性材料	34	金属氧化物	48	能量密度	62	循环寿命
7	催化剂	21	负极活性物质	35	空气电极	49	凝胶	63	循环性能
8	导电材料	22	负极浆料	36	锂电池	50	前驱体	64	氧化石墨烯
9	导电剂	23	负极片	37	锂空气电池	51	热处理	65	粘结剂
10	导电性	24	复合材料	38	锂离子	52	容量	66	正极活性物质
11	电化学	25	改性石墨烯	39	锂离子动力电池	53	溶剂	67	正极活性材料
12	电池材料	26	隔膜	40	锂硫电池	54	溶胶	68	正极浆料
13	电解液	27	功率密度	41	磷酸铁锂	55	溶液	69	正极片
14	电解质	28	活性材料	42	磷酸铁锂电池	56	石墨烯复合材料	70	自放电

doi:10.3772/j.issn.2095-915x.2018.04.008

图 4 列出了自 2008 年以来全国石墨烯锂离子电池领域的研究热点和核心技术，展示了近几年来石墨烯锂离子电池技术发展宏观趋势和技术细节。由于石墨烯特殊的性能优势，使得石墨烯在电池领域的应用十分广泛。从图中可以看出来，在材料方向，这个方向最早于 2008 年开始发展，一直持续到现在。这几年主要是研究石墨烯复合材料的研究，主要包括金属氧化物与石墨烯的复合材料、石墨烯与硫化物的复合材料、石墨烯与磷酸铁锂的复合材料等。主要是利用石墨烯的良好的力学、热学、电学等优异的性能来改善电化学性能。在组件方向，主要始于 2009 年，2009 年和 2010 年是组件方向集中兴起的时间。主要是利用石墨烯复合物作为新的正极、负极或者电池的其他组成部分，用来提升电池的容量或者稳定性等性能。在电池性能方面主开始于 2009 年，在 2009 年到 2010 年主要研究石墨烯锂离子电池的稳定性等方面，但是随着研究的深入从 2010 开始关注于电池功率密度等方面。在产品方面，大约在 2010 年开始出现石墨烯相关的电池，一直持续到 2016 年，都有产品研发出来。这些产品的研发都与石墨烯复合材料的研发息息相关，而且，这些产品相比于传统的非石墨烯电池具有非常大的优势。这些优势主要体现于延长电池寿命，缩短充电时间，提高能源转化率方面。进过文献调研与领域专家的论证，图 4 基本上与政府工作报告和相关咨询公司等权威结果基本相同，有效的验证了本方法的有效性和科学性。

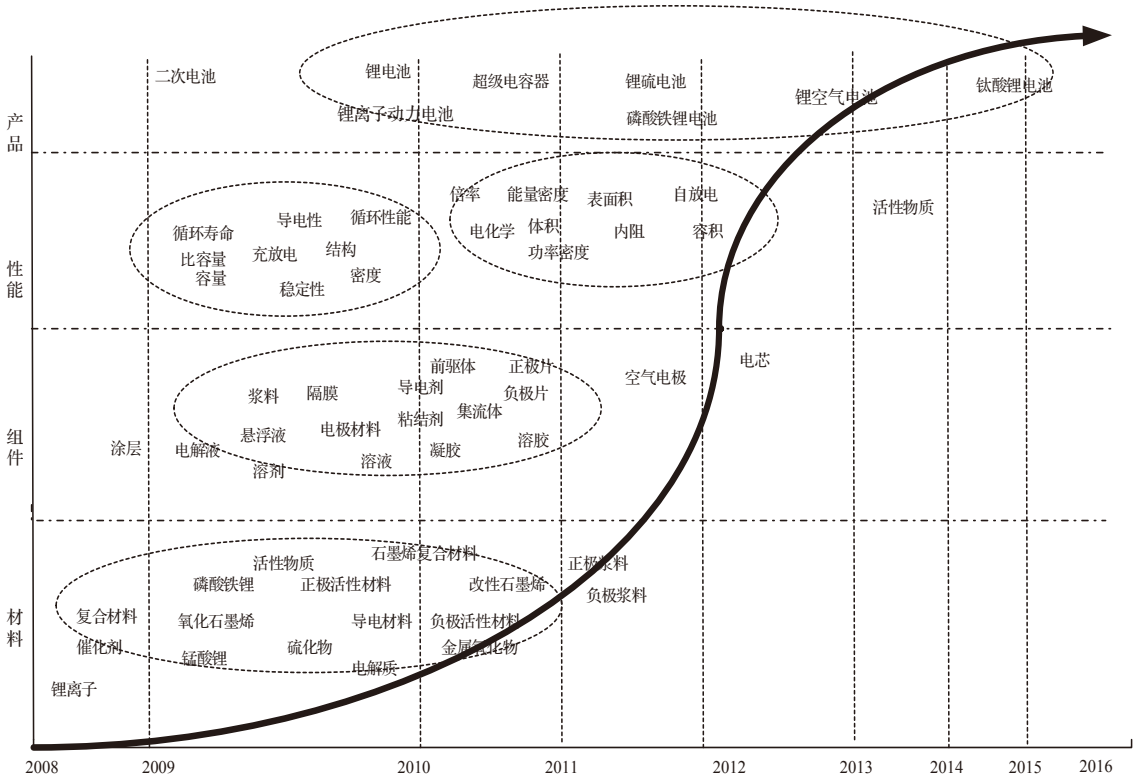

图 4 石墨烯锂离子电池技术路线图

5 结论

综上所诉，本文以文本挖掘为核心，结合社会网络分析和专家知识，综合运用定性分析和定量分析，提出了一种技术路线图的构建方法，并且以石墨烯锂离子电池为例，构建相关领域的技术路线图。在确保方法模型专家权威性的同时，提升了客观数据分析的科学性与有效性，为企业研发提供指导方向。现阶段本文的数据来源比较单一，主要以SIPO数据库为主，未来可以扩展到CNKI等文献数据库，来进一步提升数据的有效性。

参考文献

[1] 长城战略研究所. 路线图与企业自主创新企业研究报告[R]. 北京：长城战略研究所, 2005.

[2] Phaal R, Farruch C J P, Probert D R. Technology roadmapping—A planning framework for evolution and revolution[J]. Technological Forecasting & Social Change, 2004, 71(1–2):5–26.

[3] Jin G, Jeong Y, Yoon B. Technology-driven roadmaps for identifying new product/market opportunities: Use of text mining and quality function deployment[J]. Advanced Engineering Informatics, 2015, 29(1):126-138.

[4] 张嶷, 汪雪锋, 郭颖,等. 基于文献计量学的技术路线图构建模型研究[J]. 科学学研究, 2012, 30(4):495-502.

[5] 李牧南, 黄芬. 面向科技型小微企业技术路线图规划的技术挖掘工程研究[C]. 全国科技评价学术研讨会. 2015.

[6] 张颖. 湖北大重型数控机床产业技术路线图研究[D]. 武汉: 华中科技大学, 2010:45-67.

[7] 谭全元, 胡双发, 龚巨成. 基于WebGIS的湖南省黑茶产业化技术路线图研制[J]. 湖南农业科学, 2016(1):86-90.

[8] 刘彤, 侯元元, 吴晨生. 多重关系专利网络分析方法在产业技术路线图的应用[J]. 情报杂志, 2015(3):65-70.

[9] 李剑敏, 余婉浈. 融合情景分析的产业技术路线图集成规划过程研究[J]. 科技进步与对策, 2017, 34(3):56-61.

[10] 文庭孝. 专利信息计量研究综述[J]. 图书情报知识, 2014(5):72-80.

[11] 周源, 刘怀兰, 杜朋朋,等. 基于改进TF-IDF特征提取的文本分类模型研究[J]. 情报科学, 2017, V35(5):111-118.

[12] 杨家慧, 刘方爱. 基于巴氏系数和Jaccard系数的协同过滤算法[J]. 计算机应用, 2016, 36(7):2006-2010.

[13] 张宪义. 基于专利共类分析的技术领域关联研究[D]. 大连: 大连理工大学, 2013.

[14] 朱小虎, 宋文军, 王崇骏,等. 用于社团发现的Girvan-Newman改进算法[J]. 计算机科学与探索, 2010, 04(12):1101-1108.

[15] 王海龙, 和法清, 丁堃. 基于社会网络分析的专利基础技术识别——以半导体产业为例[J]. 情报杂志, 2017, 36(4):78-84.

[16] 张宪义. 基于专利共类分析的技术领域关联研究[D]. 大连: 大连理工大学, 2013.

[17] 龚佑宁, 黎德龙, 张豫鹏,等. 石墨烯及其复合材料在锂离子电池负极材料中的应用[J]. 材料导报, 2015, 29(7):33-38.