

学校编码: 10384

分类号____密级____

学号: 24320131152448

UDC ____

廈門大學

硕士学位论文

面向特定领域的专利综合主题演化

模型研究

Research on Domain-Specific Patent Comprehensive
Topic Evolution Model

蓝婧

指导教师: 陈海山教授

专业名称: 软件工程

论文提交日期: 2016年 月

论文答辩日期: 2016年 月

学位授予日期: 年 月

指导教师: _____

答辩委员会主席: _____

2016年 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果,均在文中以适当方式明确标明,并符合法律规范和《厦门大学研究生学术活动规范(试行)》。

另外,该学位论文为()课题(组)的研究成果,获得()课题(组)经费或实验室的资助,在()实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称,未有此项声明内容的,可以不作特别声明。)

声明人(签名):

年 月 日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文(包括纸质版和电子版)，允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

()1. 经厦门大学保密委员会审查核定的保密学位论文，于
年 月 日解密，解密后适用上述授权。

()2. 不保密，适用上述授权。

(请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。)

声明人(签名)：

年 月 日

摘要

在科学技术迅猛发展的背景下，科学技术的生命周期变短，专利作为科学技术成果的一大体现方式其数量更是呈指数增长，专利分析的重要性也日益显现。仅靠人工识别的方法虽然能够保证质量，但是面对如此海量的专利数据往往会显得力不从心。如何结合专利分析技术，从海量的专利文本数据中识别出值得被关注的专利数据，分析技术趋势，进而制定技术研究战略显得尤为重要。

为了从技术层面分析某一特定领域专利结构及其发展演化情况，更好地把握领域内专利的发展历程、厘清发展脉络，本论文提出了一个基于文本和网络分析的专利综合主题演化模型，该模型由专利主题结构分析、专利主题结构演化分析和专利主题网络演化分析三个模块组成。对某一特定领域内的专利，分别从宏观和微观的角度来分析该领域专利主题的演化过程。宏观上关注不同时间中专利网络的拓扑结构，分析专利主题的网络结构的演化情况。微观上关注不同时间中专利主题簇之间的演化特征，判断是否存在主题融合、扩散、继承、新增以及消逝的演化状态，分析专利主题结构的演化过程。

本论文首先介绍了研究背景和专利分析领域的研究现状，明确了研究目标与内容。接着归纳了本论文研究内容所需要的相关技术及其理论基础。然后提出了基于文本和网络分析的专利综合主题演化模型，在随后的章节详细的介绍了该模型各个模块的各阶段，并且以新能源汽车领域为例，运用论文提出的模型，在各模块中进行了实证分析，验证模型的有效性。论文所构建的基于文本和网络分析的专利综合主题演化模型可在一定程度上揭示出某一特定领域内专利主题结构及其演化情况，实现了领域内技术发展的有效识别。在论文的最后，对论文的成果和不足进行总结。

关键词：专利；主题结构；演化模型

Abstract

With the rapid development of science and technology, science and technology life cycle becomes shorter, patent as a major embodiment of scientific and technological achievements, the number of patent literature is growing exponentially. The importance of patent analysis is also increasingly apparent. Relying on manual identification methods to analyze patent literature can guarantee the quality, but with such vast amounts of patent data, researchers' ability is less than their desire. How to use patent analysis technique to identify worthy data from vast amounts of data, then analyze the technology trends, and further tap its information to formulate technology research strategy have become an important part of patent analysis.

In order to analyze the structure and the development status of patent and clarify the development vein of a specific field from the view of technical level, a patent topic evolution comprehensive model based on text and network analysis is proposed in this dissertation. This model consists of three modules; there are patent topic structure analysis, structure evolution analysis of patent topic and network evolution analysis of patent topic. Analyzing the topic evolution from the macro view and micro view for a certain specific field of patent. At the macro, this dissertation concerns on topology of the patent network in different times, and analyzing the evolution of the topology of patent topic network. At the micro, this dissertation concerns on evolution between the topic clusters in different times, to determine whether exist the evolution state of topic fusion, topic diffusion, topic inheritance, topic emerge and topic eliminate, and analyzing the evolution of the structure of patent topic.

When it comes to the details, the research objectives and the current research situation based on the work of patent analysis field are introduced firstly. Then review the the key technology of this thesis research needs and propose the patent topic comprehensive evolution model. Subsequently introduce the modules' details of the model. Using the model on the field of new energy vehicles to do experiment in each module, to verify the validity of this model. At the end of the dissertation, summarize the achievements and shortcomings. the patent topic comprehensive evolution model based on text and network analysis which build in this dissertation, can reveal topic structure and its development path of a certain specific patent field to a certain extent,

and realizing the effective recognition of the evolutionary path in the field.

Keywords: Patent; Topic Structure; Evolution Model

厦门大学博硕士学位论文摘要库

目录

第 1 章 绪论	1
1.1 研究背景	1
1.2 研究现状	2
1.3 研究目标与内容	3
1.4 论文组织结构	4
第 2 章 相关关键技术与理论基础	5
2.1 知识结构的构建技术	5
2.1.1 文献计量法	5
2.1.2 内容分析法	6
2.2 网络分析方法	7
2.3 演化分析方法	8
2.3.1 基于知识本身的演化分析	8
2.3.2 基于网络属性的演化分析	9
2.4 本章小结	10
第 3 章 专利综合主题演化模型	11
3.1 专利综合主题演化模型的研究思路	11
3.2 专利综合主题演化模型的模型框架	12
3.3 本章小结	13
第 4 章 专利主题结构研究	14
4.1 专利主题结构构建策略	14
4.2 专利数据预处理	15
4.3 专利主题结构构建	17
4.4 构建专利主题结构的聚类算法	20
4.5 专利主题结构的实证分析	21
4.5.1 新能源汽车领域介绍	22
4.5.2 数据预处理	22

4.5.3 领域专利主题结构构建	24
4.6 本章小结	31
第 5 章 专利主题结构演化研究	32
5.1 专利主题结构演化策略	32
5.2 主题簇演化特征定义	34
5.3 主题簇演化状态定义	36
5.4 专利主题结构演化实证分析	40
5.5 本章小结	45
第 6 章 专利主题网络演化研究	46
6.1 专利主题网络演化策略	46
6.2 专利主题网络分析指标	47
6.3 专利主题网络演化的实证分析	48
6.3.1 专利主题网络拓扑结构分析	48
6.3.2 专利主题网络演化分析	51
6.4 本章小结	54
第 7 章 总结与展望	55
7.1 总结	55
7.2 展望	56
参考文献	57
攻读硕士期间的研究成果	61
致谢	62

Contents

Chapter 1 Introduction	1
1.1 Background and the Topic	1
1.2 Research Status	2
1.3 Research Objectives and Contents	3
1.4 Outline of the Dissertation	4
Chapter 2 Key Technologies and Theories	5
2.1 Knowledge Structure Construction Technology	5
2.1.1 Bibliometrics	5
2.1.2 Content Analysis.....	6
2.2 Network Analysis	7
2.3 Evolution Analysis.....	8
2.3.1 Knowledge Based Evolution Analysis	8
2.3.2 Network Based Evolution Analysis.....	9
2.4 Summary.....	10
Chapter 3 Patent Comprehensive Topic Evolution Model	11
3.1 Research Ideas of Patent Comprehensive Topic Evolution Model.....	11
3.2 Model Framework of Patent Comprehensive Topic Evolution Model...	12
3.3 Summary.....	13
Chapter 4 Patent Topic Structure Analysis	14
4.1 Construction Strategy of Patent Topic Structure.....	14
4.2 Pre-process of Patent data.....	15
4.3 Patent Topic Structure Construction	17
4.4 Clustering Algorithm of Based on the Patent Topic Structure	20
4.5 Empirical Analysis of Patent Topic Structure	21
4.5.1 Field of New Energy Vehicles Introduction	22
4.5.2 Data Pre-process	22
4.5.3 Patent Topic Structure Construction.....	24
4.6 Summary.....	31
Chapter 5 Evolution Analysis of Patent Topic Structure.....	32

5.1 Evolution Strategy of Patent Topic Structure	32
5.2 Evolution feature Defination of Topic Cluster	34
5.3 Evolution State Defination of Topic Cluster	36
5.4 Empirical Analysis of Patent Topic Structure Evolution	40
5.5 Summary	45
Chapter 6 Evolution Analysis of Patent Topic Network	46
6.1 Evolution Strategy of Patent Topic Network.....	46
6.2 Analysis Index of Patent Topic Network.....	47
6.3 Empirical Analysis of Patent Topic Network Evolution.....	48
6.3.1 Topology Analysis of Patent Topic Network.....	48
6.3.2 Evolution Analysis of Patent Topic Network	51
6.4 Summary	54
Chapter 7 Conclusions and Future Work.....	55
7.1 Conclusions	55
7.2 Future Work	56
References	57
Publications	61
Acknowledgements	62

第1章 绪论

1.1 研究背景

在科学技术迅猛发展和经济全球化的时代背景下，技术、知识以及专利等无形的资产的价值已经被认为超过了有形资产的价值。在当前高度竞争的产业环境中，缺乏技术竞争力成为以技术为主导的公司失败的主要原因。然而，普遍没有公司能够承担独自发展所有技术带来的巨大投资和风险，这促使着这些技术型的组织与其他机构进行合作。所以，采用开放式的商业模式成为大多数企业的技术发展模式^[1]。在开放式的商业模式下，专利作为一个必要且强有力的法律工具，能够有效的保护知识产权。

专利是对发明授予的一种专有权利，发明是指提供新的做事方式或对某一问题提出新的技术解决方案的产品或方法。据世界知识产权组织 WIPO 发布的数据显示，专利信息包含了世界上 90%~95% 的技术信息，截止到 2016 年 3 月，世界上已经有 5110 万份专利文献^[2]。由于专利信息是公开的，研究人员经过整合和吸收很可能重新探索出一个新的有价值的专利发明，因此专利信息不仅保护知识产权，还在协助产品发明创新中发挥重要的作用。据文献[3]统计，有效地运用专利信息，可将研发时间缩短 60%，从而节省研发费用。

专利文献包括专利号、题目、摘要、国际专利分类号、申请日期、提交日期、申请人、发明人、专利权人等著录信息。其中包含了丰富的技术、商业知识，经分析后的著录信息的成果有助于决策者了解技术领域的发展现状和趋势、评估技术能力、分析技术的商业价值和预测新兴技术，从而支撑科技决策，进一步制定优先领域，从而推动科研与经济发展。

专利分析已被确认为一个探索关键技术和识别技术发展趋势的强大方式。专利研究在上世纪便已受到学术界的关注，同时学科交叉性显著，涉及学科广泛，包括法律、计算机、经济、管理等。在知识爆发性增长的今天，随着技术的生命周期变短，专利作为科学技术成果的一大体现方式其数量更是呈指数增长。仅仅依靠人工识别的方法虽然能有质量保证但是面对海量的数据仅靠人工往往会显得力不从心。如何从这样海量的专利文本数据中识别出值得被关注的的数据，对于把握先进的技术趋势，进而制定技术研究战略的确定具有重要的战略意义。

1.2 研究现状

专利文献是半结构化的文本信息，在利用专利文本和语义信息，借助自然语言处理、文本挖掘等技术进行专利挖掘和分析，随着相关技术的不断成熟，各国学者进行了一系列的相关研究，包括了专利术语提取、专利聚类、专利分类、专利网络分析以及专利时间序列演化分析等方面的研究^[4]。

专利术语提取主要是从专利中抽取重要的技术术语词。Yoon 等^[5]提出了一种运用 SAO 结构关键词分析法分析专利文献，提出了一种利用离群点预测新技术的方法。Tseng 等^[6]用开方拟和检验方法提取专利术语，进行专利技术。Noh 等^[7]提出了正交矩阵、k 均值聚类以及熵计算的关键词选择方法，提取得到专利术语。

专利聚类通过使用聚类方法，对专利文本、技术主题、申请人和国别等方面的信息进行聚类。邱洪华等^[8]将 Kmeans 算法用于形成语义网络，并在此基础上制作专利地图。Park 等^[9]人基于文献耦合和潜在语义分析，探索专利文献中潜在的研发合作伙伴。

专利文献中包含了对应专利的国际专利分类号(IPC)。专利分类主要目的是在特定领域中，设计适应 IPC 类目的分类器。Trappey 等^[10]基于 ANN 技术，实现了一个专利分类系统，以优化分类结果。Chen 等^[11]设计了一个三层分类的分类算法，将专利分类至 IPC 的子组级别，得到良好的分类效果。

因专利网络具有社会网络的特性，不少学者逐渐将专利分析与网络分析相结合，使用社会网络分析(SNA)的方法，分析专利网络中结点间的关系。Yoon 和 Park^[12]在专利主题分析的基础上，使用社会网络分析法分析专利之间的关系。Choi 等^[13]基于专利关键词分析方法，结合网络指数判定关键词网络特性以及网络发展趋势。张端阳等^[14]通过构建专利引用网络，使用 SNA 方法判断网络中各专利的重要性，寻找技术领域中的核心技术。刘倩楠^[15]建立了领域内相关专利文献的引文网络，并使用社会网络中的分析指标识别网络中的关键节点，从而绘制领域内的技术关键路径的演化图。

专利是不断发展和变化，即便是在同一领域，随着时间变化和技术的更替，领域内专利的研究热点会发生转移。因此专利具有演化的特性，通过专利时间序列演化分析，从而获得领域技术的发展趋势。方曙等^[16]提出了基于专利文献的技术演化分析方法，采用基于语义的“分类号—专利文档”相似性矩阵，聚类后构建技术层次语义网，形成技术演化图。Chen 等^[17]以引文分析为基础并融入文本

信息，提出了将专利文献集按时间窗口切割进行聚类，再计算不同技术主题之间的关联，从而绘制出技术主题演化的过程。

1.3 研究目标与内容

专利结构是专利知识长期形成的客观存在，是专利知识内在逻辑结构的集中体现，某领域的专利结构可以反映领域内专利知识的逻辑关系。同时专利知识是在不断发展变化的，从而专利结构也在不断发生演化。通过分析专利结构的演化情况，可以更好的把握领域内知识的发展历程，厘清发展脉络，更快的获得领域内的创新和发展趋势，探测和追踪技术前沿和研究热点。因此，研究如何科学、准确地揭示专利的知识结构及发展演化具有重要意义。

本论文通过专利挖掘的方法构建某一特定领域内的专利主题结构，使用主题网络演化分析、主题结构演化分析的方法，实现从技术层面分析领域内专利主题的发展和演化情况。对领域内的专利，从微观和宏观两个角度对专利主题进行演化分析，宏观上通过分析专利主题结构在不同时间片中网络拓扑特征，分析专利主题网络的演化情况；微观上，通过分析不同时间片中专利主题簇的演化情况，判断主题簇之间主题融合、扩散、继承、新增以及消逝等不同的演化状态，分析专利主题结构的演化过程。

本论文将从以下几个方面展开研究工作：

1. 专利主题结构构建分析。使用文本挖掘的方法构建某一领域内专利的主题结构，为后续的主题演化分析做铺垫。主要包括了专利数据预处理、构建主题共现矩阵、使用聚类算法对主题共现矩阵进行聚类的处理流程，并阐述专利主题结构的构建过程。

2. 专利主题结构演化分析。研究专利主题结构的演化和发展情况。通过分析不同时间中的专利主题结构，判断主题簇间的演化情况，寻找主题簇间的主题融合、主题扩散、主题继承、主题新增以及主题消逝的演化状态，展现专利主题结构的演化过程。

3. 专利主题网络演化分析。通过分析不同时间中的专利主题网络的相关指标，分析网络的拓扑结构以及网络的演化过程，并绘制专利主题的网络演化图。

论文以新能源汽车领域为例，使用德温特专利数据库中的该领域的专利文献作为数据源，使用实证法对论文模型进行验证和实证分析。并借助查阅文献资料和调研相关课题的进展，分析实验结果。

1.4 论文组织结构

论文分为七章，论文首先介绍专利分析的意义和研究目标等研究背景，接着对该领域的研究现状进行综述，再简要介绍本论文提出的专利综合演化分析模型。随后对本论文提出模型的各个阶段进行详细介绍和实证分析。

论文各章内容的安排如下：

第一章，绪论。主要介绍了本论文的研究背景和专利分析领域的研究现状，明确本论文的主要研究目标和内容。

第二章，相关关键技术及其理论基础。介绍了本论文研究内容所需要的相关技术及其理论基础。主要包括了知识结构的构建技术、网络分析方法和演进分析方法。

第三章，专利综合主题演化模型。按照实验流程分别详细阐述了基于文本和网络分析的专利综合主题演化模型的研究思路和模型框架。

第四章，专利主题结构分析。包括专利主题结构构建策略、专利数据预处理、专利主题构建和构建专利主题结构的聚类算法。并以新能源汽车领域专利作为实验数据，构建并可视化该领域的专利主题结构。

第五章，专利主题结构演化分析。描述了专利主题结构的演化策略、定义了专利主题簇的演化特征和演化状态。并以新能源汽车领域专利作为实验数据，对该领域的专利主题结构的演化情况进行实证分析。

第六章，专利主题网络分析。描述了专利主题网络的演化策略、定义了专利主题网络的分析指标。并以新能源汽车领域专利作为实验数据，对该领域的专利主题网络的拓扑结构和主题网络的演化情况进行实证分析。

第七章，总结与展望。总结了本论文的主要工作及其优缺点，并针对本论文存在的不足对以后的改进工作进行了展望。

第2章 相关关键技术与理论基础

2.1 知识结构的构建技术

2.1.1 文献计量法

文献计量法是以科学知识或技术文献作为研究主体,通过研究文献的外部特征、数量关系等分布的规律,使用计量的方法构建科学知识的内在逻辑结构。引文分析(Citation Analysis)是建立在加菲尔德(Garfield)的科学引文索引的理论上的基础上的,通过使用各种数学、统计学以及逻辑方法,对科技期刊、专利、著作等科学文献的引用或被引情况进行分析,以便揭示其数量特征和规律和追溯知识或技术的演化情况,达到评价和预测科学发展趋势的目的。引文分析是文献计量的最广泛使用的方法之一^[18]。引文分析法依赖于科学文献之间的相互引证与被引证关系,这种引证关系继而形成引文网络^[19],同时这种关系可扩展到作者之间、机构之间、领域之间甚至国家之间。在引文网络中存在大量的共引关系和文献耦合。

1. 共引关系

共引(Co-citation)关系发最早是由 Henry Small 提出的,当两篇或两篇以上的文献被其他文献同时引用,那么这些文献之间就具有共引关系^[20]。通常来说,文献之间共引关系越高,则主题相似度也高,研究的方向和内容也越接近。一组文献间的共引关系继而形成了共引网络,可通过可视化的方式表示某研究领域的科学结构。在共引网络中,被共同引用的文献通常在主题上都具有一定的相似性,各节点间的距离不同,可以通过相关度的计算来衡量节点间的距离,从而辨识节点间在内容主题上的相关程度,可以分析出各节点主题的亲疏关系^[19]。

共引关系通过一组共引文献的相互引用,为反映研究前沿提供知识基础。潘教峰等^[21]通过 ESI 论文作为数据源,对高被引论文的同被引聚类,分析揭示了科学研究结构。Takeda 等^[22]构建了一个光学领域文献的引文网络,使用拓扑聚类方法,发现了五个核心主题,并分别从这五个核心主题中探索出新兴的主题。Chen Chaomei 等^[23]以 1996 年至 2008 年出版的信息科学领域内的 12 种期刊中的文献作为实验数据,提出了一种作者共引、文献共引、共引簇演变的多角度的共引分析模型,分析和解释共引簇的结构和动态变化。Lee M R 等^[24]通过建立从

1995年至2010年的知识管理领域共10974篇文献的共引关系，构建寻径网络图谱和战略技术图谱，分析知识管理研究领域的演化情况。

2. 文献耦合

文献耦合(Bibliographic Coupling)最早是由 Kessler 于 1963 年提出的，是指两篇文献共同引用了一篇或多篇文献，其引用文献的数量多少来度量相似度，称为耦合强度^[25]。当两篇文献引用多条相同的文献时，可以推断它们间就具有某些相似性和联系。

通过对学科之间的文献进行耦合分析，可以挖掘出不同学科之间的关联性与交叉演化情况；通过对某一学科领域进行文献耦合分析，可以得到该领域中的知识结构与领域研究现状。肖明等利用文献耦合方法构建数字图书馆领域的知识结构，并使用可视化工具展示了该领域的主题文献之间的耦合关系^[26]。杜辉^[27]将文献耦合关系分为文献引文耦合与文献关键词耦合两种形式，结合这两种形式挖掘学科领域的文献，从而构建知识结构。Tseng 等^[28]使用期刊中所有论文间的相同引用文献的数量衡量期刊间的相似性，使用聚类算法对期刊文献耦合网络进行聚类，从而识别学科领域。Ni C 等^[29]使用期刊中共享相同的引用文献的发文作者的数量衡量期刊间的相似性，使用聚类算法对期刊作者耦合网络进行聚类，从而构建相关领域的知识结构。

文献耦合度量的是引用文献之间的关系，所以其耦合关系是固定不变的，因此基于文献耦合的分析方法揭示了稳定且不变的科学结构关系。共引关系度量的是被引文献之间的关系，共引关系随着文献增加不断发生变化，并非稳定不变的。基于共引关系的分析方法可以揭示动态的知识结构，随着时间的发展，其知识结构发生变化，从一定程度能够更加准确的反映科学结构的发展变化情况^[30]。尽管共引模式被用于分析文献研究工作连接和协作的问题，但是这种模式关注点着重于全局而非局部。

2.1.2 内容分析法

在当前的研究环境下，科学文献的范围、目标以及研究范围和重点都与文献本身记载的内容相关，例如文献的题目、关键词、摘要等。公开发表的文献有着较为严格的著录规范体系，包括正文以及相关对文献信息描述的特征项。不同的特征项描述文献不同的特征，例如题目是对文献的主题的简要表达；关键词则更是对主题的提炼；摘要是对文献正文的高度概括。基于这些科学文献特征项提炼

Degree papers are in the “[Xiamen University Electronic Theses and Dissertations Database](#)”.

Fulltexts are available in the following ways:

1. If your library is a CALIS member libraries, please log on <http://etd.calis.edu.cn/> and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to etd@xmu.edu.cn for delivery details.