第一章　数据与方法说明

项目按数据分析和专家研判两部分来实施。在数据分析阶段，294 个备选工程研究热点及其核心论文的数据、备选工程开发热点分析用的 52 张学科组的专利地图及其核心专利数据，由科睿唯安提供；从专利地图解读出 318 个备选工程开发热点的部分在中国工程院（以下简称工程院）本项目组主持下由各领域课题组完成。在专家研判阶段，项目组在工程院各学部及“ 1+9 ”学术期刊的指导和协助下，组织院士专家通过问卷调查、开会研讨、分组解读，获得每个领域 10 个左右工程研究热点和 10 个左右工程开发热点以及 3 个左右工程研究焦点和 3 个左右工程开发焦点。

1　工程研究热点的遴选

本报告中反映的工程研究热点以科睿唯安的 Web of Science™ 核心数据库的期刊论文和会议论文（SCI、EI 收录）为原始数据，通过与工程院学部专业划分标准体系建立映射关系，获得每个领域组数据分析的基础；通过对每个领域被引频次位于前 10% 的高影响力论文的共被引聚类，关注其被引频次以及引用寿命，获得 294 个备选工程研究热点；进而通过专家研判获得 89 个工程研究热点。具体过程如下文所述。

1.1　论文数据的获取与预处理

在数据获取上，通过将 Web of Science 学科与工程院各学部专业进行匹配，获得每个领域对应的期刊和会议列表；领域组组织专家确认，最终获得各个领域组分析的数据源共计 12 822 种期刊和 28 312 个会议（具体见表 1.1.1）。此外，对于《Nature》等 58 种综合学科的期刊，采用单篇文章归类的方法，即根据期刊内单篇文章的参考文献主要归属的学科来定义这篇文章的领域学科组。

在此基础上，综合考虑期刊与会议差别和出版年等因素，对上述文献列表进行检索和数据挖掘，获得各领域组 2011 年至 2016 年发表的前 10% 高被引的论文，作为研究热点分析的原始数据集（表

表 1.1.1　各领域对应的期刊、会议数及其前 10% 高被引的论文数

序号	领域	期刊数	会议数	高被引的论文数
1	机械与运载工程	457	1 768	38 676
2	信息与电子工程	986	9 632	109 507
3	化工、冶金与材料工程	1 128	2 313	219 081
4	能源与矿业工程	226	785	440 641
5	土木、水利与建筑工程	359	512	28 384
6	环境与轻纺工程	1 003	605	93 524
7	农业	1 575	975	105 523
8	医药卫生	4 328	7 059	392 142
9	工程管理	755	681	32 927

注：各领域组间存在部分重复的期刊、会议和高被引的论文。

1.1.1），数据采集时间为 2017 年 2 月，即引用时间截至 2017 年 2 月。

1.2 论文数据的分析与备选研究热点的获取

基于研究人员相互引用而形成的研究主题间的关系网络，通过对 9 个领域高被引的论文进行共被引聚类分析[1]，获得每个领域的聚类主题。综合考虑各个聚类主题的总被引频次、篇均被引频次、平均出版年和“常被引论文”所占比例等指标获得每个领域组不少于 25 个的备选工程研究热点及 5 个关键词（由于部分备选热点主题在不同领域组中相似，为确保各个领域组至少包含独属本领域的 25 个备选工程研究热点，部分领域组备选热点数多于 25 个，具体见表 1.2.1）。

此外，由于每个领域组具有不同的学科特点和引用规律，有些领域施引文献的数量相对较少，因此从 9 个领域组中遴选的热点仅定义为本领域组中研究最为活跃的论文簇。

1.3 工程研究热点的遴选与研判

在通过数据挖掘分析获得 294 个备选工程研究热点的基础上，各领域组组织院士、专家研读核心论文，并归并相似热点，修正热点名称，而后以网络和纸质调查问卷的方式向广大院士、专家征求建议，获取专家对每个领域前 10 位的工程研究热点的投票意见。本次调查共有 9 个领域组 238 位院士和 632 位非院士专家参与，在问卷统计结果的基础上，经过专家研讨和学部常委会审核获得 89 个工程研究热点。

2 工程开发热点的遴选

2017 年度项目共获得 8 个领域（工程管理领域组本年度不涉及工程开发热点相关工作）81 个工程开发热点。数据分析上，以科睿唯安的 Thomson Innovation 专利数据库为原始数据，通过建立德温特手工代码与工程院学部专业划分标

表 1.2.1 各领域共被引聚类结果统计

序号	领域	聚类主题数	前 10% 高被引的论文数	备选工程研究热点数
1	机械与运载工程	4 140	19 503	45
2	信息与电子工程	11 507	53 880	27
3	化工、冶金与材料工程	23 227	103 840	28
4	能源与矿业工程	4 553	20 208	33
5	土木、水利与建筑工程	2 971	14 435	23
6	环境与轻纺工程	10 239	46 846	39
7	农业	11 613	51 118	26
8	医药卫生	40 775	181 062	25
9	工程管理	3 442	15 084	29

[1] Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents. J Am Soc Inform Sci 1973;24(4):265–9.
Garfield E. New tools for studying the history of science. In: Essays of an Information Scientist: The Awards of Science and Other Essays, Vol: 11. Philadelphia: ISI Press; 1988. p. 20–1.
Garfield E. ABCs of cluster mapping. Part 1. Most active fields in the life sciences in 1978. In: Essays of an Information Scientist: The Awards of Science and Other Essays, Vol: 4. Philadelphia: ISI Press; 1980. p. 634–41.
Small H. Paradigms, citations, and maps of science: A personal history. J Am Soc Inf Sci Tec 2003;54(5):394–9.
Small H. Tracking and predicting growth areas in science. Scientometrics 2006;68(3):595–610.

准体系的映射关系，获得分析的基础数据；通过对 8 个领域 52 个学科组中被引频次位于各学科组前 10 000 的高影响力专利进行聚类，获得 52 张专利地图。专家研判阶段，通过专家对专利地图解读获得 318 个备选工程开发热点，通过问卷调查和专家研判遴选出 81 个工程开发热点。具体过程如下文所述。

2.1 专利数据的获取与分析

数据获取上，通过建立德温特手工代码与工程院学部专业划分标准体系的匹配关系，初步确定 8 个领域的专利数据检索范围；通过领域组专家对德温特手工代码的删减、补充和完善，确定 8 个领域 52 个学科组的专利检索式；通过 Thomson Innovation 专利数据库检索，获得专利分析的原始数据；通过综合考虑专利公开年和专利家族被引频次等指标，筛选获得每个学科组对应的前 10 000 篇高被引核心专利数据（专利公开年在 2011 年至 2016 年之间，专利引用时间截至 2017 年 2 月）。通过专利文本间的语义相似度，利用 Thomson Innovation 工具获得基于 10 000 篇高影响力专利的 52 张能快速直观呈现工程开发热点技术分布的 ThemeScape 专利地图（图 2.1.1）。

2.2 专利地图的解读

各领域组专家按照“以数据分析为基础，以专家研判为依据”的原则，参照 ThemeScape 专利地图，依照工程开发技术的聚焦程度和关联程度从高到低，在专利地图上确定每个学科组的备选工程开发热点位置。而后，各领域组通过对专利地图上热点位置所包含的高影响力专利的研阅，以及对关键词、专利公开量、公开国家、总被引频次和篇均被引频次等的统计分析，提炼技术开发热点、归并相似热点、确定热点名称，获得每个学科组的备选工程开发热点。如图 2.2.1 所示，土木、水利与建筑工程领域通过解读专利地图圈定了建筑学学科组的 4 个热点（热点 1：智能化建造；热点 2：装配式建筑；热点 3：绿色城市建筑；热点 4：采暖通风及空气调节），而后通过专家研判，并结合智能化建造和装配式建筑的技术特点，将装配式建筑作为智能化建造的一个分支技术，归并为一个备选工程开发热点。

2.3 工程开发热点的遴选与研判

在解读专利地图的基础上，各领域组组织不同学科的院士、专家召开专题研讨会，对备选工程开发热点进行归并、命名优化、热点解释和指标计算等，

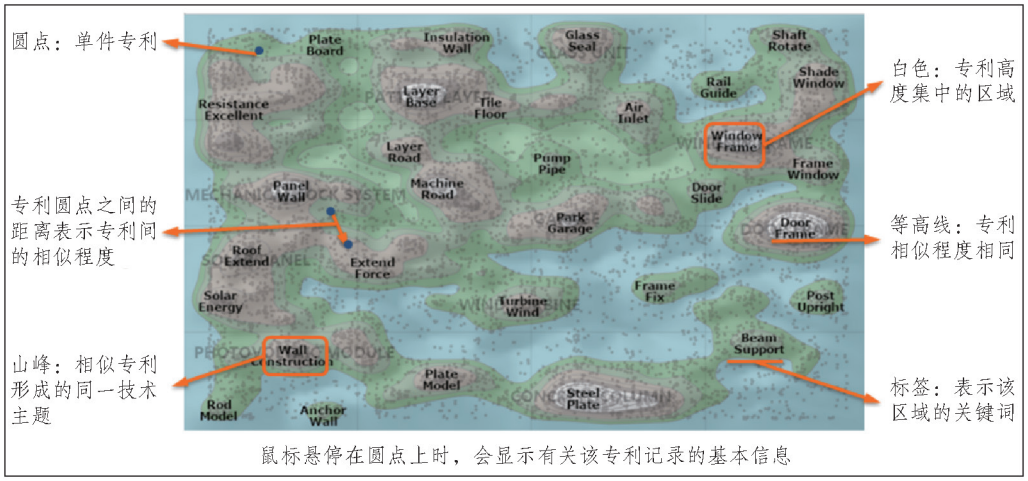

图 2.1.1 ThemeScape 专利地图的基本样式（以建筑学学科组为例）

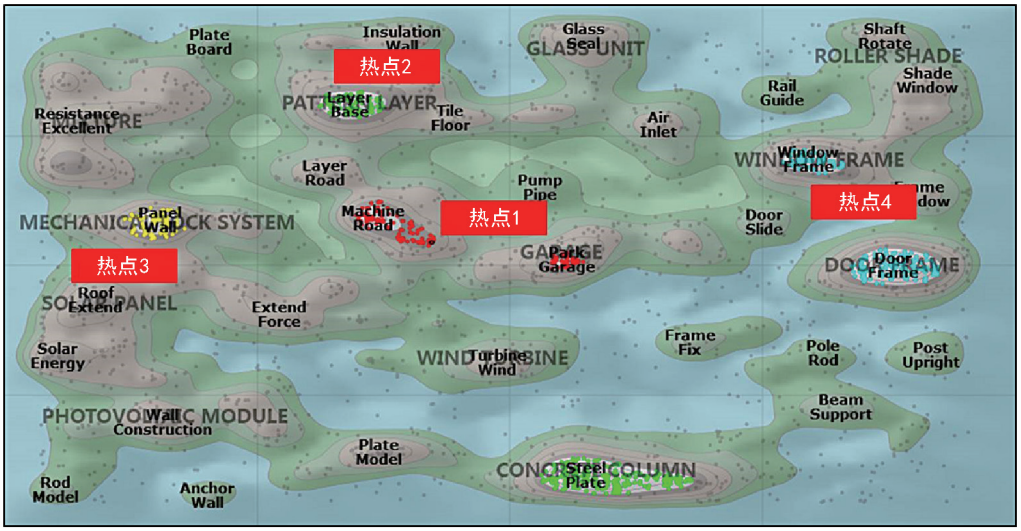

图 2.2.1 土木、水利与建筑工程领域建筑学学科组热点分布

最终获得 8 个领域 318 个备选工程开发热点。然后通过问卷调查（问卷调查实施过程见第 1.3 节），获取广大院士、专家对每个领域前 10 位的工程开发热点的意见。结合问卷统计结果，各领域组通过组织专家研判和学部常委会审核获得 81 个工程开发热点。

3 工程研究焦点与工程开发焦点的遴选和解读

在 89 个工程研究热点和 81 个工程开发热点的范围内，通过讨论研判、学部常委会审议等方式，从发展前景、受关注程度等角度遴选出需要重点解读的 29 个工程研究焦点和 26 个工程开发焦点。进而组织各个焦点相关专业专家对工程研究焦点和工程开发焦点的现状、发展趋势和技术重点进行详细解读。例如，通过专家研判，研讨焦点含义、发展趋势和国际研究态势；通过客观数据分析，佐证揭示全球主要国家或地区、机构在此焦点的实力与贡献度，并探讨了国家或地区间、机构间的合作关系与研发布局。

4 术语解释

被引频次：被引频次指文献被给定的出版物——本项目指被科睿唯安收录的出版物引用的次数（SCI 和 SCI-E 引用），而不是所有的被引记录。

出版物：科睿唯安著录的文献类型包括社论、会议摘要、会议论文、书评和其他的研究型文章。本项目中的论文分析部分包括经过同行评议的公开发布的研究型期刊论文、综述和会议论文，不包含特种文献。专利分析主要为全球德温特专利引文索引（Derwent Patent Citation Index，DPCI）数据库中的专利（发明）数据及其被引专利和引证专利，不包含国防专利。

高被引的论文（高影响力论文）：在本项目中，考虑到出版年和期刊主题分类，被引频次在前 10% 的论文定义为高被引的论文。由于不同出版方式的引用率不同，会议论文和期刊论文分开处理。

聚类主题：对高被引的论文进行共被引聚类分析获得的一系列主题和关键词的组合。

核心论文：在本项目中，与 294 个备选工程研究热点相关联的高影响力论文即为核心论文。

高被引专利（高影响力专利）：本项目中，高被引专利指年均引证专利次数在 DPCI 数据库中排在前 10 000 篇的专利。

常被引论文：在本项目的分析中，引文速度排在前 10% 的论文被遴选为常被引论文，并会考虑

出版年和期刊的学科类别等因素。本报告中用常被引论文占比间接反映论文的引用寿命。

引文速度：引文速度是一定时间内衡量累计被引频次增长速度的指标。在本项目的分析中，每一篇文献的引文速度是从发表的月份开始，通过记录每个月的累计被引频次来计算。

被专利引用的文献：专利会引用其他专利或非专利文献（NPR）如科技文献。在本项目的分析中，被专利引用的非专利文献将通过 DPCI 数据库获取。

学科交叉系数：学科交叉系数是衡量论文的学科分布紧密或分散程度的指标，取值范围为 [0,1]。值为“ 1 ”表示这组论文平均分布到所有学科类别。值越接近 1，表明越是多学科的论文集合。值为“ 0 ”表示该论文集合中只有一个学科类别。学科交叉系数基于信息熵理论，计算公式如下：

$$-\left(\sum_{i=1}^{n} p_i \log_2 p_i\right) \Big/ \log_2 n$$

其中，p 为在学科 i 中所占的论文比例，n 为所有的学科个数。

专利地图：通过分析专利文献中的语义相似度，利用科学的统计分析方法对专利的文本内容进行精细剖析整理，并以地图形式进行可视化展现，是形象地反映某一行业或技术领域的整体面貌的主题全景图。

工程院学部专业划分标准体系：包含工程科学技术（含农、医）的各学部所涵盖的专业领域，按照《中国工程院院士增选学部专业划分标准（试行）》确定。