技术路线图中未来技术词表构建方法研究

叶春蕾[1]　冷伏海[2]

[1](北京城市学院信息学部　北京 100094)

[2](中国科学院国家科学图书馆　北京 100190)

【摘要】利用文本挖掘技术，并结合科学计量、自然语言处理等方法，提出一种基于三重共现算法的技术路线图中未来技术词表构建方法，以揭示特定技术领域的未来技术发展方向和未来发展阶段水平特征，初步实现技术路线图中的未来技术分析目标。实验表明该方法能够在一定程度上支持技术路线图的未来技术分析研究。

【关键词】未来技术词表　技术路线图　三重共现

【分类号】G350

Building the Future – oriented Technology Thesaurus of Technology Roadmap

Ye Chunlei[1]　Leng Fuhai[2]

[1](Information Department, Beijing City University, Beijing 100094, China)

[2](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

【Abstract】The paper proposes a triple co – occurrence algorithm to construct the future – oriented technology thesaurus of technology roadmap based on text mining combining the method of scientometrics and natural language processing, which reveals the future – oriented technology development direction and level characteristics of special technical field and achieves preliminarily the target of future – oriented technology analysis of technology roadmap. The experiment shows that this method can support the future – oriented technology analysis of technology roadmapping to some extent.

【Keywords】Future – oriented technology thesaurus　Technology roadmap　Triple co – occurrence

1　引　言

目前，面向未来的技术分析已成为国内外许多战略科研机构的重要任务和核心能力之一。包括技术预见、技术预测、技术监测、技术竞争情报、技术路线图和技术评价等在内的各类未来技术分析研究活动受到广泛重视，许多国家的政府、机构和企业等组织在不同层面上开展了面向未来的技术分析研究和实践活动[1]。

作为一种预见方法，未来技术分析的主要目标是系统地生成有助于应对未来的挑战和危机的关于未来的知识。2008 年，Cagnin 等[1]对 FTA 的概念进行了明确的解释，并指出 FTA 主要是指如何构建未来可能的备选技术，以及如何从中优选出更具有价值的与未来技术有关的理论、方法和实践。目前，情景分析方法、技术路线图方法以及德尔菲方法是面向未来的技术分析主要研究方法[2]。本文以技术路线图方法为基础，利用文本挖掘技术，并结合科学计量、自然语言处理等方法，提出一种三重共现算法以构建技术路线图的未来技术词表和基于技术路线

收稿日期：2013 – 04 – 09

收修改稿日期：2013 – 05 – 02

图的未来可能的备选技术(即未来技术发展方向),并从中可以优选出具有更高价值的未来技术发展目标及其发展水平(即未来技术发展水平特征)。

2 文献综述

以美国兰德公司为代表的一些政策咨询机构积极开展未来科技发展的研究和预测工作,并形成了著名的德尔菲法和情景分析等方法。与此同时,由于技术路线图能支持公司或部门的计划和预测而获得研究人员和实践人员的更多注意。目前,技术路线图已被证明是在战略技术规划的背景下应对技术挑战、支持信息收集和决策制定的一种有效方法。作为一种重要的未来技术分析方法[3],国内外对其展开了深入的研究和探索。

Yoon 等[4]使用文本挖掘方法从产品手册和专利文件等材料中提取关键信息,用于识别现有的产品及其技术形态,并以此作为技术路线图的方法基础。Lee 等[5]对技术路线图方法进行了改进研究,利用文本挖掘技术从产品资料、科技文献和专利文献中抽取技术关键词,并利用组合、共词、网络分析等方法,以制作基于关键词的技术路线图,从而减少路线图对专家知识的依赖。刘兰等[6]将文本挖掘和技术路线图结合起来,通过挖掘隐含在科技信息中的知识和联系,并结合技术领域专家绘制技术路线图,以发现技术创新的机会。

技术路线图以简洁的图形、表格、文字等形式描述技术变化的步骤或技术相关环节之间的逻辑关系,能够帮助使用者明确技术领域的未来发展方向以及实现目标所需的核心技术,理清领域和核心技术之间的关系。刘细文等[7]指出,绘制技术路线图需要关注其关键组成要素,包括时间规划、层次关系、重要突破点等。因此,作为一种重要的辅助科技决策和管理的战略规划工具,技术路线图的内容是获得未来技术发展方向和发展水平特征的主要依据。

到目前为止,大多数研究将技术路线图作为一种面向未来技术分析的工具或方法进行优化,基本上没有将技术路线图作为研究对象对其文本内容进行深层次的情报分析,如构建揭示未来技术发展方向和发展水平的未来技术词表。

徐峰等[8]指出,情报分析方法是重要的面向未来的技术分析方法,可以与其他方法混合使用,以便能更好地反映未来技术的状态信息。目前,越来越多的研究人员将情报分析方法引入未来技术分析中,并取得了一些有价值的研究成果[9,10]。不管是技术形态关联的分析方法[11],还是非相关文献知识发现的分析方法[12],都是从方法的角度对未来技术分析进行探索和研究。

本文以技术路线图为蓝本,利用文本挖掘、科学计量分析、自然语言处理等方法与技术,结合技术路线图的文本结构,自动地对技术路线图全文进行深度扫描,构建技术路线图中的未来技术词表,旨在准确反映技术领域的未来技术发展方向和发展水平特征。

3 基于三重共现的未来技术词表构建方法

以构建技术路线图中的未来技术词表为主要研究内容。为了更好地实现词表的存储,本文定义一种三元组数据结构,包括时间特征词(Time)、核心技术关键词(Term)以及度量值(Value)三个维度指标,分别用来定义特定技术领域技术未来发展方向(由核心技术关键词定义)及其在未来发展的阶段(由时间特征词定义)中的发展水平特征(由度量值定义),表达了未来技术发展的三个维度。

根据技术路线图文档结构的特点,以三元组为词表项目的基本数据结构,以共现分析理论和方法为基础,提出基于三重共现的未来技术词表构建方法,具体构建过程如图 1 所示:

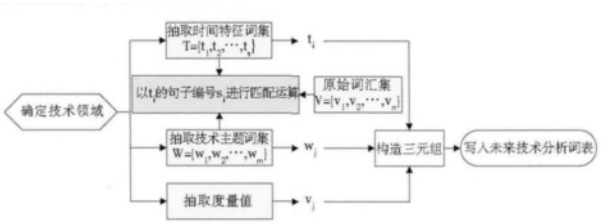

图 1 未来技术词表构建流程

(1)核心技术关键词的自动抽取

在技术路线图中,由核心技术关键词表征特定技术领域未来技术发展方向,核心技术关键词的自动抽取是实现未来技术词表构建的基础。具体抽取方法参见文献[13]。

(2)时间特征词的自动抽取

在技术路线图中,由时间特征词表征特定技术领域未来技术发展阶段,时间特征词的自动抽取是未来技术词表构建的核心。时间特征词的抽取工作相对比较简单,从分词的角度看,时间特征词具有明显的词性

特征;从抽取的角度看,时间特征词和领域关键词在一定窗口中存在共现关系。因此,根据时间特征词和领域关键词之间特定的语义关系,采用传统的共现算法,可以很好地完成时间特征词的抽取工作。

(3)时间特征词和技术关键词匹配算法

为了进一步获得特定技术领域未来技术发展阶段水平,需要获得该技术领域中未来技术方向(由核心技术关键词表示)在未来时间点(由时间特征词表示)的发展水平(由度量值表示)。由于一个技术领域包含多个核心技术关键词和多个时间点,而未来技术发展阶段水平度量值必须根据核心技术关键词和时间特征值匹配的结果再从原始词汇集中获得,因此,时间特征词、核心技术关键词之间的匹配运算是未来技术词表构建的关键。匹配算法的前提条件是选择一个确定的技术领域,该领域的内容是由词汇链的初始值标记。

具体的算法描述如下:

输入:领域关键词 fm、时间特征词集 T、核心技术关键词集 W、原始词汇集 V。

输出:具有较强语义关联度的词汇对 $W=\{fm,T,W\}$ 集合,其中 $t_i(t_i \in T)$ 和 $w_i(w_i \in W)$ 具有一定的语义关联度。

步骤如下:

①$i=0$;读取 t_i,s_i,其中 s_i 是句子编号。

②以 s_i 为条件在 V 中进行查找,结果返回与 t_i 在一定时间窗口共现的词汇集 V_i,V_i 是 V 的子集。

③对 V_i 和 W 进行交集运算,结果返回 V_i 和 W 共现的技术关键词集 W_i,W_i 是 W 的子集,如果 W_i 为空,则令 $s_i=s_i-1$,转至步骤②继续执行,直到 s_i 为空。

④结果返回 W_i 中等价指数 E 最大的技术关键词 w_i。

⑤$i++$,转至步骤①继续执行,直至 t_i 为空。

在使用匹配算法计算时间特征词 t_i 和技术关键词 w_i 之间的语义关联度时,要充分考虑两个主要影响因素:t_i 和 w_i 之间的等价指数 E,该指数主要考虑 t_i 和 w_i 之间关联强度的权重;t_i 和 w_j 同现范围,由于上述匹配算法中"一定窗口"体现 t_i 和 w_j 同现范围,当 s_i 与 w_j 所在的句子编号 s_j 相等时,则属于同句共现,否则属于临近度为 $n(n=|s_i-s_j|)$ 的句子共现,匹配参数的调整反映同现窗口的动态性,匹配算法的输出则是和 t_i 在最小范围同现的具有最大语义关联度的核心技术关键词 w_j。

(4)度量值的自动抽取

在技术路线图中,由度量值表征特定技术领域未来技术发展阶段水平,标志技术发展阶段水平度量值往往都是用数值型数据或部分数值型数据,抽取的目标对象是由原始文档构成的原始词汇集。在构造三元组时,度量值不是一个必要的条件,在技术路线图文档中,有些核心技术关键词没有明确的标志技术发展的度量值,在具体抽取操作时,可参照上述匹配算法。

4　实验及结果分析

使用 Java 和 SQL Server 2005 设计相应的测试系统完成图 1 所示的未来技术词表的构建,并对实验结果进行有效性分析。

4.1　实验数据源及过程

以美国 NASA 的"Draft Nanotechnology Roadmap(纳米技术路线图)"[14] 作为数据源和未来技术词表构建方法研究的实证对象,并进一步选取该文档中的"Lightweight Material(轻质材料)"技术领域作为实证领域,以本文提出的三重共现方法构建该技术领域的未来技术词表。

根据笔者前期的研究成果[13],可以确定包含该技术领域的备选技术关键词的目标词汇链。然后利用上述匹配算法进行时间特征词的抽取,时间特征词具有明显的词性特征,所获得的时间特征词主要包括 2013、2019、2022 以及 2030,这 4 个时间特征词分别表示"Lightweight Material"技术发展的 4 个时间阶段。

在技术关键词抽取的过程中,主要采用等价指数(E 值)作为语义关联强度的重要指标,从目标词汇链中抽取的部分核心技术关键词,以 E 值降序排列,如表 1 所示:

表 1　"Lightweight Material"技术领域核心技术关键词集

段落编号	关键词	频次	E 值
29	tensile strength	6	2.00
29	carbon fiber	18	0.91
29	produce fiber	2	0.67
29	intermediate modulus carbon fiber	2	0.67
29	modulus carbon fiber	2	0.67
29	carbon nanotube fiber	3	0.44
29	single wall carbon nanotube	3	0.44
29	wall carbon nanotube	3	0.44
29	improvement processing	3	0.44
29	nanotube method increase	1	0.33
29	nanotube – nano – tube interaction	1	0.33
29	carbonization gel spun	1	0.33
29	polyacrylonitrile pan	1	0.33
29	nanocomposite tensile strength	1	0.33
29	greater carbon fiber	1	0.33
…	…	…	…

由表 1 可以看出，“tensile strength(拉伸强度)”与“Lightweight Material”语义关联强度最大。

同时，采用词性分析、N - Gram 分词、C_value 术语识别等自然语言处理方法从技术路线图的全文中获得原始词汇集，该集合包括每个词汇所在的段落、句子、词汇次序、词性分析、频次以及 C_value 值等信息，原始词汇集的部分数据如表 2 所示：

表 2 “Lightweight Material”技术领域原始词汇集

段落编号	句子编号	词汇编号	词汇名称	词性分析	C_value
29	1	2	lightweight material	Adjs Nouns	2
29	1	10	specific strength stiffness	Adjs Nouns Nouns	1.58
29	1	17	carbon nanotube	Nouns Nouns	40.64
29	1	24	conventional carbon fiber	Adjs Nouns Nouns	1.58
29	1	25	carbon fiber	Nouns Nouns	16.91
29	1	28	composite cfrp	Adjs Nouns	1
29	1	31	various aerospace material	Adjs Nouns Nouns	3.17
29	1	32	aerospace material	Nouns Nouns	3
29	2	41	ultimate goal	Adjs Nouns	2
29	1	2	lightweight material	Adjs Nouns	2

由表 2 可以看出，原始词汇包含的信息很多，有 C_value术语强度指标、词性分析指标(本文主要采用名词词组)，更有重要的词汇语义指标，即该词汇所在的段落编号、句子编号以及词汇编号，用于进一步的共现分析。

以时间特征词集结合表 1、表 2 中的核心技术关键词集以及原始词汇集的数据为基础，采用匹配算法进行计算，最终获得“Lightweight Material”技术领域未来技术词表，如表 3 所示：

表 3 “Lightweight Material”的未来技术词表

词表项目	时间特征词	技术关键词	度量值
词表项目 1	2013	tensile strength、tensile strength factor、intermediate modulus carbon fiber	无
词表项目 2	2019	carbon fiber、porous carbon fiber property、intermediate modulus carbon fiber	TRL 6、30% lighter
词表项目 3	2022	carbon fiber、intermediate modulus carbon fiber、lightweight metal	TRL 6、30% lighter
词表项目 4	2030	carbon nanotube sheet、tensile strength、carbon fiber、nanocomposite tensile strength、carbonization gel spun、polyacrylonitrile pan	30% higher、40 - 60 GPa、50% greater

表 3 中显示的是临近度为 1(即同句和前后相差一句)的句子共现分析结果。临近度越小，所获得的技术关键词更精确，但是会存在共现缺失的现象；临近度越大，所获得的技术领域技术关键词范围更宽泛，词表的精确性将有所下降。因此，如何准确把握临近度的阈值需要参考领域专家意见并观察多次试验的效果才能确定。

4.2 实验结果分析

由表 3 可以看出，“Lightweight Material”技术领域在未来发展共有 4 个时间阶段，分别是 2013、2019、2022 以及 2030 年。每个时间阶段的发展方向在技术关键词中有明确的提示，同时在度量值中也有明确的技术发展水平特征。有了这样的词表作为支持，研究者很容易把握该技术领域未来发展方向和未来发展阶段水平特征。

为了对本文提出的三重共现构建方法进行评价，笔者结合领域专家的意见，采用人工方法对实验材料中的“Lightweight Material”技术领域建立未来技术词表，根据 4 个时间特征词相应地为词表构建 4 个表项。将人工构建的词表和表 3 的结果进行对比发现，两者的重合率为 75%。对不一致的词表项目(2022 年)进一步分析发现，人工构建的词表中包含的关键词为“carbon fiber reinforced polymer composite”，而表 1 中识别的关键词为“carbon fiber”。通过对实验过程回溯分析发现，自动抽取的候选关键词包括“carbon fiber”、“carbon fiber reinforced”、“polymer composite”三项，而最终列出的关键词“carbon fiber”是根据等价指数进一步筛选的结果。

5 结 语

本文根据技术路线图文本结构的特点，以包含时间特征词、核心技术关键词、度量值的三元组为词表项目的基本数据结构，以共现分析理论和方法为基础，提出一种基于三重共现的未来技术词表构建方法。实验表明，该方法能够比较理想地建立表征特定技术未来发展方向和发展水平特征的词表。根据分析可以看出，词表中的第三个表项的关键词不够准确，其主要原因是由于关键词自动抽取环节中提高准确率的同时降低了召回率，因此，在实际的应用中，可以考虑提供多关键词的方式提高关键词的召回率。但是，召回率的提高势必影响本词表中核心技术关键词的准确率，因此，要想获得更完整、更精确的未来技术词表，可以考虑对词表构建的各个环节做进一步的优化。

参考文献:

[1] Cagnin C, Keenan M, Johnston R, et al. Future – Oriented Technology Analysis [M]. Berlin Heidelberg: Springer – Verlag, 2008.

[2] Kreibich R, Oertel B. Futures Studies and Future – oriented Technology Analysis Principles, Methodology and Research Questions [EB/OL]. [2013 – 03 – 02]. http://berlinsymposium.org/sites/berlinsymposium.org/files/foresight_final_draft_formatted_sfr_111015_2.pdf.

[3] Porter A. Technology Futures Analyses: New Methods [EB/OL]. [2013 – 03 – 02]. http://www.cgee.org.br/arquivos/ib02.pdf.

[4] Yoon B, Phaal R, Probert D. Morphology Analysis for Technology Roadmapping: Application of Text Mining [J]. *R&D Management*, 2008, 38(1): 51 – 68.

[5] Lee S, Lee S, Seol H, et al. Using Patent Information for Designing New Product and Technology: Keyword Based Technology Roadmapping [J]. *R&D Management*, 2008, 38(2): 169 – 188.

[6] 刘兰, 赵新力, 李艳. 基于文本挖掘和技术路线图的技术创新机会发现 [J]. *中国软科学*, 2007(6): 102 – 110. (Liu Lan, Zhao Xinli, Li Yan. The Discovery of Technology Innovation Opportunity Based on the Text Mining and the Technology Roadmap [J]. *China Soft Science*, 2007(6): 102 – 110.)

[7] 刘细文, 柯春晓. 技术路线图的应用研究及其对战略情报研究的启示 [J]. *图书情报工作*, 2007, 51(6): 37 – 41. (Liu Xiwen, Ke Chunxiao. The Applications of Technology Roadmap and Its Enlightenment to Strategic Intelligence Research [J]. *Library and Information Service*, 2007, 51(6): 37 – 41.)

[8] 徐峰, 冷伏海. 面向未来的技术分析概念、方法与应用研究进展 [J]. *情报学报*, 2010, 29(3): 539 – 544. (Xu Feng, Leng Fuhai. Concepts, Methods and Development of Future – oriented Technology Analysis [J]. *Journal of the China Society for Scientific and Technical Information*, 2010, 29(3): 539 – 544.)

[9] 徐峰. 专利技术形态分析方法优化研究 [D]. 北京: 中国科学院研究生院, 2010. (Xu Feng. Research on Modified Method of Patent Technology Morphology Analysis [D]. Beijing: Graduate University of Chinese Academy of Sciences, 2010.)

[10] 王林. 技术形态关联分析方法优化研究 [D]. 北京: 中国科学院研究生院, 2012. (Wang Lin. Research on Key Issues in Morphology Association Analysis [D]. Beijing: Graduate University of Chinese Academy of Sciences, 2012.)

[11] 冷伏海, 王林, 王立学. 基于文本挖掘的形态分析方法的关键问题 [J]. *图书情报工作*, 2012, 56(4): 27 – 30. (Leng Fuhai, Wang Lin, Wang Lixue. Key Issues in Morphology Analysis Based on Text Mining [J]. *Library and Information Service*, 2012, 56(4): 27 – 30.)

[12] 张云秋. 非相关文献知识发现的方法改进研究 [D]. 北京: 中国科学院研究生院, 2008. (Zhang Yunqiu. Improvement on the Methods of Disjoint Literature – based Knowledge Discovery [D]. Beijing: Graduate University of Chinese Academy of Sciences, 2008.)

[13] 叶春蕾, 冷伏海. 基于词汇链的路线图关键词抽取方法研究 [J]. *现代图书情报技术*, 2013(1): 50 – 56. (Ye Chunlei, Leng Fuhai. Study on the Keyword Extraction from Roadmap Based on the Lexical Chains [J]. *New Technology of Library and Information Service*, 2013(1): 50 – 56.)

[14] Meador M A, Files B, Li J, et al. Draft Nanotechnology Roadmap: Technology Area 10 [R]. National Aeronautics and Space Administration, 2010.

(作者 E – mail: yechunlei@mail.las.ac.cn)