企业基础研发的知识挖掘与可视化研究

——以波音公司为例

岳洪江[1],刘思峰[2]

(1.南京审计学院 公共管理与绩效评估研究院,江苏 南京 210029;
2.南京航空航天大学 经济与管理学院,江苏 南京 210016)

摘 要:面向技术的基础研发是国际大企业研发环节之一,以计算机语言处理和文献计量学结合起来的科技文本挖掘为企业研发提供了一个很好的了解同业竞争者研发知识的工具手段。以国际大企业波音公司为例,挖掘其基础研发知识结构,并对其知识结构进行了可视化展示。

关键词:基础研发;科技文本挖掘;波音公司
中图分类号:F403.6 文献标识码:A 文章编号:1001-7348(2009)18-0136-04

0 前言

20世纪50年代以来,美、欧、日公司之间的竞争优势越来越依赖于长期的技术积累。特别是制造业,如果企业没有持续的技术创新与积累,就难于形成领先的新产品和新市场。如今世界研发投资的80%、技术创新的71%,均由世界500强企业所创造和拥有,62%的技术转让在世界500强企业间进行,技术创新已是全球500强的重要特征之一[1]。

上世纪80年代中期,许多日本大企业纷纷设立基础研发所,并不断增加投资。近百年世界产业发展的历史表明,真正起巨大推动作用的技术几乎都来自企业。如通讯领域中的贝尔实验室、汽车领域中的福特公司、航空领域中的波音和空客、化工领域中的杜邦和拜耳、机床领域中的西门子、计算机领域中的IBM、英特尔、微软等,都是自主技术创新的领军企业。在技术进步和创新中,企业具有重要作用。

研发是现代大企业持续发展的核心,在大企业技术创新体系中,面向技术的基础研发是整个研发环节之一。研发前通过各种不同媒介展现出来,包括各种科学与技术文本,主要有各种技术报告、学术论文、专利,以及其它媒体信息,还有研发的中试产品等。科技文献以每年6%的速度在增长[2]。在互联网时代可以很容易地获取各种各样的信息,但面对日益快速发展的科学技术工业,特别是非结构化的数据管理,仍是一件复杂和费时的工作。而科技文本挖掘(science and technology text mining)可以为企业研发提供一个很好的信息挖掘手段。

1 文本挖掘程序与方法

科技文本挖掘正是利用自然语言和统计技术从大量的科技文本中获得有价值的信息。通过大量的在线文献数据库、专利数据库、企业网站、各类金融报表,获得企业研发的信息。快速的技术跟踪要依靠互联网来有效挖掘这些数据库,可以获得某个研究领域的各类基础信息,如研究者、研究对象所发表的期刊、主要的研究国家、机构等。更主要的是可以获得该研究主体领域的知识图谱,如技术突破点以及技术之间的关系等。

科技文本挖掘可以分3步来完成,文本信息提取、文本信息处理和文本信息综合。文本信息提取界定要分析研究对象及其涵盖的范围, 将其从科技文本数据库中提取出来;文本信息处理主要是利用文献计量学和计算机语言程序以及统计分析技术对非结构化的文本数据进行排序、分类以及定量化图示;文本信息综合是通过专家对文本信息处理结果的分析,发现所需要的信息。

目前世界上常用的科技文本数据库主要有SCI(美国科学引文数据库)、EI(美国工程索引数据库)、MEDLINE(美国国家医学图书馆医学文献数据库)、INSPEC(英国机电工程师学会科学文摘数据库)、USPTO(美国专利局专利数据库)、ESP(欧洲专利局专利数据库)、PCI(美国德温特专利引文索引)。这些覆盖范围广的半结构化科技文献数据库大大提高了对全球范围内文本挖掘的广度和深度。

收稿日期:2009-03-24
基金项目:国家自然科学基金项目(70473037);河南省普通高校人文社科研究基地项目(STS060014)
作者简介:岳洪江(1973-),男,辽宁沈阳人,蒙古族,博士,南京审计学院副教授,研究方向为科技管理与科学计量学;刘思峰(1955-),男,河南平舆人,博士,南京航空航天大学经济与管理学院院长、教授,博士生导师,研究方向为灰色系统理论和科技管理。

随着人们对文本挖掘的重视，国内外有许多种文本挖掘计算机程序，其中主要用于科技文本挖掘的非商业软件主要有：SITKIS（http://users.tkk.fi/~hschildt/sitkis/）、ARROW SMITH（http://kiwi.uchicago.edu/）、BIBEXCEL（http://www.umu.se/inforsk/Bibexcel/index.html）、BIBTECHMON (http://www.arcs.ac.at/S/ST/BibTechMon)，另外，还有ISI公司的商用软件 Thomson Data Analyzer。上述软件都是针对大型文献数据库而设计的。

文本挖掘是企业研发人员、研发管理者、研发管理和资助管理部门以及企业研发竞争情报人员获得有用信息的手段。能够在企业制定战略计划时提供全面的指示，也能在产品开发阶段提供指导，在产品开发过程中导向新的性能和创新，也可以带来新技术突破的机会以及新研究领域的开辟。

科技文本挖掘增强了人们对全球技术文献的了解，通过非相关文献之间的融合，可以找到新的突破和创新点。可以鉴别次领域的研究水平，并帮助研究人员提高他们成果的影响力以及期刊影响力。

文献计量学方法开创了许多科学出版物、专利、引文以及另外一些指标来评估科技绩效，文献计量学可以鉴别出科学领域的基础研究结构分布，如作者、期刊、研究机构等；可以鉴别出科学领域的创新专家，以及研究进展等；可以鉴别出全球科学领域各研究组织的优势；鉴别出研究者、研究团体、研究机构、国家和地区的科研成果的影响力。

2　波音公司的基础研发知识结构与图谱

典型的科技文本挖掘研究主要是将计算机语言处理和文献计量学结合起来，既可以从宏观了解企业研发领域的基础结构分布，又可以从微观了解企业研发领域的图谱。

首先根据研究对象和目的选择合适的文献数据库。我们的目的是探讨波音公司基础研发，SCI文献数据库是比较好的选择，从Web of Science中提取波音公司发表的学术论文，时间是1995—2006年。其次，作计算机程序处理，并把结果图示出来。

2.1　文献计量学分析

80多年来，波音公司始终致力于新产品的开发和新技术的探索，从民用飞机、军用飞机到航天飞机、运载火箭、全球通信卫星网络、国际空间站。波音公司由6个主要业务集团组成：波音民用飞机集团、航天与通信集团、军用飞机与导弹集团、空中交通管理公司、波音联接公司和波音金融公司，以及联合服务集团。图1展示出波音公司1995—2006年每年的SCI文献量，波音公司在基础研发方面每年有50~200篇的论文发表，发表量是一个波动的趋势。论文主要集中在波音民用飞机集团、航天与通讯集团方面。表1列出了与波音公司科学合作的前5个国家与机构。

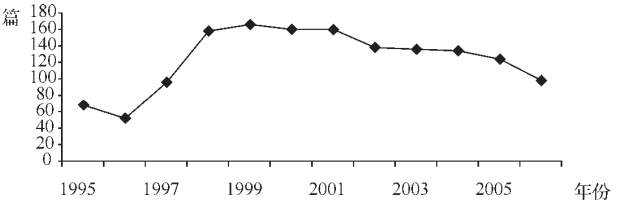

图1　1995–2006年波音公司每年的SCI文献量

表1　与波音公司科学合作的前5个国家与机构

前 5 个合作国家	英国	德国	加拿大	法国	俄罗斯
份额（%）	5.0	2.6	2.3	1.9	1.7
前 5 个合作机构	美国航空航天局	华盛顿大学	美国空军	密苏里大学	新墨西哥大学
份额（%）	6.1	5.3	4.2	2.0	1.9

波音公司的基础研发主要是在国内完成的，但也与世界30个国家有过科学合作。特别是与科学比较发达的国家合作比较密切。波音公司在国内研究的论文中有65%是公司独立完成的，在美国有700个机构参与了波音公司的基础研发，主要集中在美国的大学和研究机构，如美国国家航空航天局和美国空军研究机构，它们是波音公司的主要合作伙伴。表2列出了波音公司基础研发的前10个学科领域。表3列出了波音公司基础研发论文的前10种期刊。表4列出了波音公司发表的高被引频次论文。

表2　波音公司基础研发的前10个学科领域

波音公司基础研发的前 10 个学科领域	份额（%）
航天工程学	22.1
电学与电子工程学	17.45
材料学	11.10
机械学	8.26
机械工程	7.20
应用物理学	7.07
核科学技术	5.62
计算机科学，软件工程	4.43
计算机技术	4.10
仪器科学技术	3.83

表3　波音公司基础研发论文的前10种期刊

波音公司基础研发论文的前 10 种期刊	份额（%）
Journal of aircraft	4.82
AIAA journal	3.04
Journal of spacecraft and rockets	2.71
IEEE transactions on nuclear science	2.12
Journal of the American helicopter society	1.59
IEEE aerospace and electronic systems magazine	1.52
Aviation week & space technology	1.39
Journal of applied physics	1.32
Journal of guidance control and dynamics	1.26
Acta astronautica	1.19

表4 波音公司发表高被引论文

论文题目	被引次数
Experimental verification and simulation of negative index of refraction using Snell´s law	215
Directional tunneling and andreev reflection on YBa2Cu 3O7 delta single crystals: Predominance of dwave pairing symmetry verified with the generalized Blonder,Tinkham and Klapwijk theory	210
Multiphoton detection using visible light photon counter	100
Survey of numerical methods for trajectory optimization	96
Development of a high-quantum-efficiency single photon counting system	91
Long-term cholesterol-lowering effects of 4 fat-restricted diets in hypercholesterolemic and combined hyperlipidemic men The dietary alternatives study	88
Strategies for turbulence modelling and simulations	87
Airplane trailing vortices	85
An overview on the use of titanium in the aerospace industry	80
Multipole translation theory for the 3dimensional laplace and helmholtz equations	77

在SCI学科领域划分中,波音公司的基础研发涉及102个领域,但主要集中在航空工程、电子工程、材料科学、机械工程等方面。波音公司在486种期刊上发表过论文,期刊集中度较低,主要分布在工程技术类期刊,前3种期刊是美国航空航天协会主办的期刊。大部分论文发表在由美国和英国主办的期刊上。

2.2 知识图谱分析

内容结构分析是科技文本挖掘的一个重要方面,可以识别科技文献中句子的出现频率,识别出研究的主题或概念、主题之间的关系,以及它们在文献数据库中的发展和进化。其中,共词分析法和同被引分析法是研究科学知识结构的重要方法。共词分析方法属于内容分析方法的一种,它通过对一组词两两统计它们在同一篇文献中出现的次数,并以此为基础对这些词进行聚类分析,从而反映出这些词之间的亲疏关系,进而分析这些词所代表的学科和主题的结构变化。同被引分析法认为如果两个文献/著者/期刊同时被第3个文献/著者/期刊引用,则这两个文献/著者/期刊存在同被引关系。经常一起被引用的文献/著者/期刊,它们在研究主题的概念、理论或方法上是相关的。为此,共词或文献/著者/期刊共被引的次数越多,它们之间的关系就越密切,"距离"也就越近。可以利用多元统计技术如因子分析、聚类分析和多维尺度分析等,挖掘学科内的文献、关键词、著者共同体,绘制"知识地图",使之可视化。

图2给出了波音公司SCI论文高频词(前50个)共词图谱。从图2可以看出,波音公司SCI论文高频词(前50个)共词图谱密度不是很紧密,其中有9个词是孤立点,说明这些词未与其它词有任何联系。有些词构成的链条比较突出,表明这些词之间的联系比较深入,也可能预示波音公司未来需要深入研究的方向。

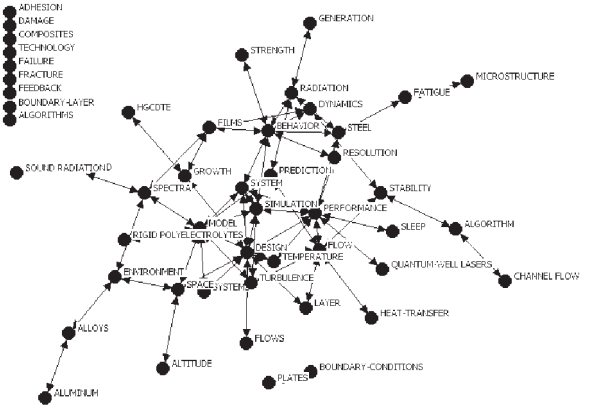

图2 波音公司SCI论文高频词(前50个)共词图谱

共词图谱k核分析。k核指的是满足一个条件的子图,即子图中的点都至少与该子图中的k个其它点邻接。通过改变k的值,就会得出不同的子图。随着k值的增加,k核的子图成员会逐渐减少,而成员之间的关系会更紧密。图3显示了当k等于3时的共词网络。

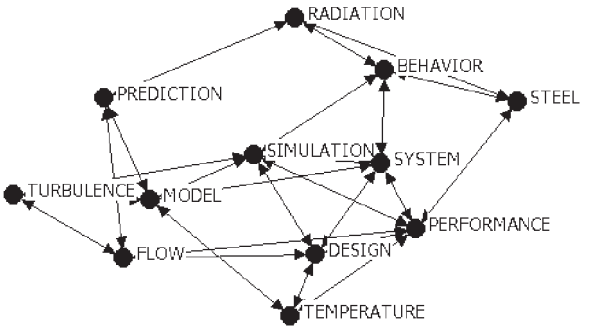

图3 k等于3时的共词网络

在波音公司SCI论文高频词(前50个)共词网络中k核最高的级数为3。也就是说在k等于3的核中,每一个词都最少和同一核中的其它词同现次数等于或大于3,是整个共词网络中连接最强,也是关系最为紧密的一个部分。每个顶点的大小与其在共词网络中出现的次数成正比。从图2可以直观地解读波音公司基础研发的重点领域。

3 结论与讨论

科技文本挖掘是一种对文献进行探测的有效方法,而可视化知识图谱提供了看得见的科学研究领域的结构和动态状况。通过对航空巨头—波音公司的基础研发的知识挖掘,揭示了波音公司基础研发的各种结构分布模式,以及研发产出知识结构。这种方法可以推广到探测国际大公司企业基础研发的结构动向。对于国内公司做大做强有一定的现实意义。

同时我们也可以看到,国际制造业大公司不断加强研究与发展。在愈来愈激烈的竞争面前,大公司把加强研究和发展放在了首位,力争走在同行的前面。在整个研发结构中,基础研究处在研发体系的前端,也是产业创新的重

要支撑体系。SAPPHO项目研究结果也显示，基础研究确实有利于企业从事产业创新。正是由于基础研究在产业创新中具有越来越重要的作用,所以许多大公司越来越重视建立自己的基础研究机构和培养基础研究能力，如贝尔(Bell)公司、通用(GE)公司、道化学(Dow)公司等都有强大的基础研究能力，它们基本上以实施进攻型产业创新为主。基础研究是技术创新,特别是核心技术产生的主要源泉,是增强自主创新能力所必备的基础。上个世纪90年代以来,随着某些高技术领域的基础研究成果可迅速转化为商品,基础研究逐渐从专注于创造新知识的“生产导向”,向“生产导向”与“扩散导向”并重转化。

在进行基础研发的过程中,大公司充分利用各种资源进行国内外合作,包括政府研究机构、大学、公司等等,这也是国际大公司,特别是R&D密集型的工业采取的一种普遍策略。当前工业发达国家各企业之间的共同研究主要属于前竞争活动,多公司集团集中于创造前竞争的技术信息和制造工艺技术。这两种技术都是很多公司感兴趣的,因为它们渴望得到技术提高的机会,也希望战略工业部门的基础设施日渐完善以利于有效地进行竞争。

参考文献:

[1] 毛蕴诗,孙景武,杜慕群,等.世界500强的特征及其对中国企业的启示[N].中山大学学报,2002,42(5):76-83.

[2] FERN NDEZ CANO A,TORRALBO M,VALLEJO M.Reconsidering price´s model of scientific growth:an overview[J].Scientometrics,2004,61(3):301-321.

[3] KOSTOFF R N,GREEN K A,TOOTHMAN D R,HUMENIK,J.A. Database tomography applied to an aircraft science and technology investment strategy[J].Journal of Aircraft,2000,37(4):727-730.

[4] RONALD N.KOSTOFF,ROBERT R.SCHALLER.Science and technology roadmaps [J].IEEE Transactions on Engineering Management,2001,48(2):132-143.

[5] Ronald N.Kostoff.Role of technical literature in science and technology development and exploitation [J].Journal of Information Science,2003,29(3):223-228.

(责任编辑:万贤贤)

A study of Knowledge Mining and the Visualization of Enterprises' Basic R&D

Yue Hongjiang[1],Liu Sifeng[2]

(1.Institute of Public Management and Performance Evaluation,Nanjing Audit University,Nanjing 210029,China;
2.Economic and Management School,Nanjing University of Aeronautics and Astronautics,Nanjing 210016,China)

Abstract: Technology-oriented R&D is the basis of international big company R&D.Science & Technology (S&T)knowledge mining is used to extract technical intelligence from the open source global anthrax research literature.R&D literature infrastructure is obtained using bibliometrics and literature of the co-keyword network is visualized.

Key Words: Technology-oriented R&D;Science and Technology Text Mining;Boeing