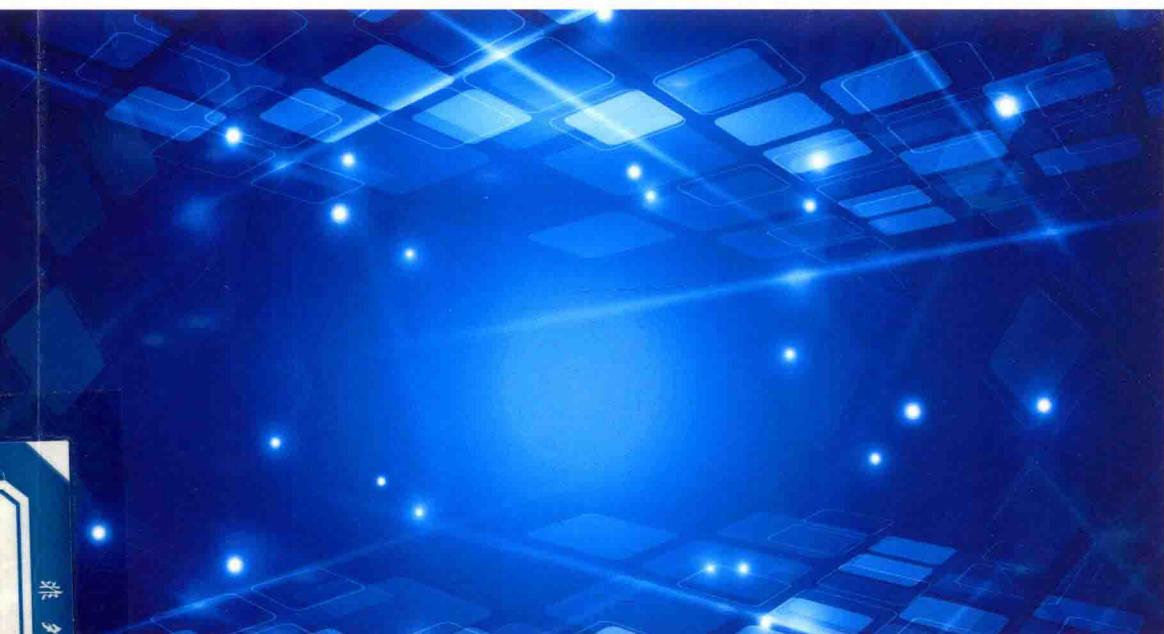


基于论文和专利资源的技术机会发现方法

徐 硕◎著



张外借



科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

国家自然科学基金项目 (ID: 71403255)

基于论文和专利资源的 技术机会发现方法

徐 硕 著



科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

· 北 京 ·

图书在版编目 (CIP) 数据

基于论文和专利资源的技术机会发现方法 / 徐硕著. —北京: 科学技术文献出版社, 2018. 1

ISBN 978-7-5189-3875-9

I. ①基… II. ①徐… III. ①论文—文献信息—信息资源—研究 ②专利文献—文献信息—信息资源—研究 IV. ① G312 ② G306.4

中国版本图书馆 CIP 数据核字 (2018) 第 011241 号

基于论文和专利资源的技术机会发现方法

策划编辑: 周国臻

责任编辑: 赵 斌

责任校对: 张叫喙

责任出版: 张志平

出 版 者 科学技术文献出版社

地 址 北京市复兴路15号 邮编 100038

编 务 部 (010) 58882938, 58882087 (传真)

发 行 部 (010) 58882868, 58882874 (传真)

邮 购 部 (010) 58882873

官 方 网 址 www.stdp.com.cn

发 行 者 科学技术文献出版社发行 全国各地新华书店经销

印 刷 者 北京教图印刷有限公司

版 次 2018 年 1 月第 1 版 2018 年 1 月第 1 次印刷

开 本 710×1000 1/16

字 数 232千

印 张 14.75

书 号 ISBN 978-7-5189-3875-9

定 价 68.00元



版权所有 违法必究

购买本社图书, 凡字迹不清、缺页、倒页、脱页者, 本社发行部负责调换

序 言

当前，全球技术发展日新月异，大量的技术创新活动为企业提供了比以往更多的技术发展机会，企业必须识别和把握可能出现的技术发展机会，才能提高企业的技术创新能力和国际竞争力。作为全球最大的技术信息源，论文信息资源通常被用于测度基础科学研究活动的水平，而专利信息资源则被用于测度产业技术的创新水平，具有生命力和潜在商业价值的技术机会在这两大信息源中通常会留下一定的足迹。因此，论文和专利信息资源的融合分析对于把握技术机会、理解科学和技术之间的联系、提高技术创新水平具有重要意义。

技术机会发现是指通过对特定领域内海量信息资源的深入挖掘和分析，在掌握已有技术发展趋势及其相互关系的同时，发现最新技术动向，推断该领域可能出现的技术形态或技术发展点。通过对论文和专利信息资源的融合分析，可以反映技术发展现状、挖掘研发热点、预测发展趋势、揭示竞争对手的技术实力与战略布局。

本书主要以作者带领团队所研发的大数据环境下多源信息融合的科技文献智能分析服务平台 SciTeMiner 为背景，系统阐述了该平台中技术机会分析模块所涉及的方法和技术。本书共分为四大部分：第一部分由第二章组成，主要涉及技术术语抽取方法研究；第二部分由第三章、第四章、第五章、第六章及第七章组成，主要研究技术术语间语义关系抽取方法；第三部分由第八章和第九章组成，主要探讨技术主题抽取及技术主题关联；第四部分由第十章、第十一章和第十二章组成，主要研判技术生命周期阶段及构建技术功效图。

本书得以顺利出版，要感谢国家自然科学基金的资助，同时，也要

感谢作者原单位中国科学技术信息研究所，以及新单位北京工业大学各级领导和同事的大力支持。本书作者长期从事科学前沿探测、技术预见、数据挖掘、机器学习和大数据方面的研发工作，一直酝酿出版与技术机会发现方法有关的图书，但由于种种原因，直到今天才得以实现。本书凝聚了许多人的心血和智慧，在此要特别感谢中国科学技术信息研究所的韩红旗副研究员和张兆锋博士，以及本书作者曾指导过的研究生张晗、王新和王政同学。

由于作者水平所限，书中难免存在不足，欢迎读者批评指正。

目 录

第一章 绪 论	1
1.1 背景及意义	1
1.2 国内外研究现状	2
1.3 平台及工具研发	7
1.4 章节结构安排	9
1.5 本章小节	11
参考文献	11
第二章 专利技术术语抽取方法	16
2.1 引 言	16
2.2 术语的定义、分类和特征	16
2.3 自动术语抽取的方法	19
2.4 中文术语抽取研究的概述	21
2.5 专利技术术语抽取模型	23
2.6 实验结构及讨论	31
2.7 本章小节	41
参考文献	41
第三章 共现聚类分析的新方法：最大频繁项集挖掘	44
3.1 引 言	44
3.2 共现分析法	45
3.3 最大频繁项集挖掘	50
3.4 实验结果及讨论	53
3.5 本章小结	56
参考文献	56

第四章 基于双序列比对的中文术语语义相似度计算方法	58
4.1 引 言	58
4.2 《同义词词林》简介	59
4.3 I 型问题的语义相似度计算	61
4.4 II 型问题的语义相似度计算	62
4.5 实验结果及讨论	67
4.6 本章小结	69
参考文献	70
第五章 仅根据 Proximity 数据构建向量空间模型的方法	72
5.1 引 言	72
5.2 基于 MDS 的向量空间模型构建方法	73
5.3 实验材料及数据	76
5.4 实验结果及分析	77
5.5 本章小结	82
参考文献	83
第六章 基于弱监督学习的语义关系抽取方法	85
6.1 引 言	85
6.2 国内外研究现状	86
6.3 实体关系抽取模型	91
6.4 实验结果及分析	100
6.5 本章小结	105
参考文献	106
第七章 几种叙词表复杂逻辑错误检查算法研究	109
7.1 引 言	109
7.2 预备知识	110
7.3 属/分关系中的逻辑错误检查算法	111
7.4 参考关系中的逻辑错误检查算法	113
7.5 本章小结	115

参考文献	116
第八章 融合科技文献内外部特征的主题模型发展综述	117
8.1 引 言	117
8.2 历史渊源及符号表示	119
8.3 融合科研人员特征	121
8.4 融合时间特征	126
8.5 融合参考文献特征	129
8.6 融合多个外部特征	133
8.7 本章小结	136
参考文献	137
第九章 论文和专利资源主题关联分析方法	141
9.1 引 言	141
9.2 技术路线	142
9.3 技术主题抽取	143
9.4 词项和命名实体聚类	148
9.5 主题相似度计算	150
9.6 技术主题关联	152
9.7 实验结果及讨论	154
9.8 本章小节	159
参考文献	159
第十章 基于事实型数据的技术生命周期判断方法综述	164
10.1 引 言	164
10.2 技术生命周期阶段划分	165
10.3 技术生命周期阶段判断方法	169
10.4 本章小结	177
参考文献	179
第十一章 面向情报技术领域的 Loglet 分析	182
11.1 引 言	182

11.2	增长曲线模型	183
11.3	实验数据选取	186
11.4	LogletLab 简介	190
11.5	实验结果及分析	191
11.6	本章小结	195
	参考文献	195
第十二章	专利技术功效图智能构建进展	197
12.1	引言	197
12.2	技术功效图概述	198
12.3	技术功效图构建模式	200
12.4	关键技术研究进展	203
12.5	本章小结	207
	参考文献	208
附录 1	词性说明	210
附录 2	FAO-780 数据集前 25 条高频术语组成的共现信息矩阵 C	212
附录 3	FAO-780 数据集挖掘得到的所有最大频繁项集	214
附录 4	原子术语及相应的编码	218

插图目录

图 1-1	技术机会分析工具截图	8
图 2-1	专利技术术语抽取模型	24
图 2-2	某件燃料电池专利的名称和摘要	26
图 2-3	某件燃料电池专利的名称、摘要分词和词性标注结果	27
图 2-4	专利数据采集系统模型示意	31
图 2-5	燃料电池技术术语各词长候选词语分布	32
图 2-6	C-value 和 PC-value 方法抽取的前 200 名不同 字长术语的对比	34
图 2-7	C-value 和 PC-value 方法抽取术语的准确率对比	34
图 2-8	文档频率排名前 25 位的 3 字及以上技术术语	36
图 3-1	average-linkage 层次聚类过程	49
图 3-2	正确的 average-linkage 层次聚类结果	50
图 3-3	项集 $\{A, B, C, D, E\}$ 的子集空间	52
图 3-4	项集 $\{A, B, D, E\}$ 、 $\{A, C, D, E\}$ 及 $\{B, C, D\}$ 的子集空间	53
图 3-5	25 个高频术语通过共现聚类分析得到的聚类结果	54
图 4-1	《词林 2》五级分类结构示意	60
图 4-2	与 c_1 、 c_2 有关的语义分类树片段	61
图 4-3	组成 T_1 与 T_2 的原子术语间的对应关系	63
图 4-4	计算 T_1 与 T_2 间 (a) 及 T_3 与 T_4 间 (b) 最优比对的 打分矩阵 F	66
图 4-5	采用 NW 算法得到 T_1 与 T_2 间 (a) 及 T_3 与 T_4 间 (b) 的 对应关系, 以及采用算法 1 得到的 T_1 与 T_2 间 (c) 及 T_3 与 T_4 间 (d) 的对应关系	67
图 5-1	平面内 4 点及两两间的欧氏距离	74
图 5-2	部分词汇的二维可视化表示	77

图 5-3	对比实验路线	78
图 5-4	归一化 Rand 指标及聚类个数随参数 c 的变化	81
图 6-1	弱监督学习发展历程中的关键节点	87
图 6-2	半监督学习训练过程	88
图 6-3	远程监督实体关系抽取可能遇到的各种情况	90
图 6-4	Rel-LDA 和 Rel-TNG 模型的概率图模型表示	94
图 6-5	Type-LDA 模型的概率图模型表示	97
图 6-6	Rel-LDA 和 Rel-TNG 模型的 F 值随语义关系数量 K 的 变化情况	101
图 6-7	Type-LDA 和 Type-TNG 模型的 F 值随语义关系数量 K 和 实体类型数量 $K^{(T)}$ 的变化情况	102
图 6-8	4 种模型在 GENIA 数据集上的 F 值随“种子”文档 比例变化的情况	104
图 6-9	4 种模型在 EPI 数据集上的 F 值随“种子”文档 比例变化的情况	105
图 7-1	属/分关系中的逻辑错误示意	112
图 7-2	参考关系中的逻辑错误示意	113
图 8-1	LDA 模型的概率图模型表示	120
图 8-2	主题模型发展历程	120
图 8-3	AT 模型的概率图模型表示	121
图 8-4	ART 和 RART 模型的概率图模型表示	123
图 8-5	APT 模型的概率图模型表示	124
图 8-6	AIT 和 LIT 模型的概率图模型表示	124
图 8-7	DTM 和 cDTM 模型的概率图模型表示	127
图 8-8	ToT 模型的概率图模型表示	128
图 8-9	Pairwise-Link-LDA 和 Link-pLSA-LDA 模型的概率图 模型表示	130
图 8-10	RTM 模型的概率图模型表示	131
图 8-11	LTHM 的概率图模型表示	132
图 8-12	ACT 和 AToT 模型的概率图模型表示	134
图 8-13	coAT 和 CAT 模型的概率图模型表示	135
图 9-1	标注命名实体的论文和专利样例	143

图 9-2	论文和专利资源间主题关联分析技术路线	144
图 9-3	CorrLDA2 和 CCorrLDA2 模型的概率图模型表示	146
图 9-4	表 9-2 中论文和专利标题经过 Brown 聚类分析后 得到的二叉树	150
图 9-5	词项聚簇、词汇主题、实体类别、实体主题以及 实体聚簇组成的网络结构	151
图 9-6	论文和专利资源间的主题关联示意	154
图 9-7	论文资源中的技术主题 37	156
图 9-8	专利资源中的技术主题 45	157
图 9-9	论文与专利资源间的主题关联强度图谱	158
图 10-1	技术生命周期四阶段论	165
图 10-2	典型的四阶段论	166
图 10-3	高德纳公司的技术炒作曲线示意	168
图 10-4	韩国科学技术情报研究院研发的 InSciTe 系统截图	168
图 10-5	3 条等价曲线示意	169
图 10-6	技术就绪水平中技术成熟度的 9 个级别	170
图 10-7	专利技术生命周期图	172
图 10-8	相对增长率二维矩阵	173
图 10-9	监督判断法流程	174
图 10-10	表达式 $\ln(a)$ 和参数 b 的取值所对应的生命周期阶段	176
图 10-11	专利特性曲线	176
图 11-1	指数增长与 Logistic 增长曲线	184
图 11-2	菌落增长的 Logistic 增长曲线	185
图 11-3	Fisher-Pry 变换后菌落增长的 Logistic 增长曲线	186
图 11-4	Loglet Lab2 软件的工作界面	190
图 11-5	技术生命周期曲线分界点	192
图 11-6	论文总量、参与作者量、参与机构量、论文类型文献量的 Logistic 曲线拟合图	193
图 11-7	会议论文类型和有基金支持文献量的 Logistic 曲线拟合图	194
图 12-1	某技术主题的技术功效图	199
图 12-2	传统构建模式制作流程	201
图 12-3	智能构建模式流程	202

表格目录

表 2-1	中文术语构词规则	28
表 2-2	几个假术语例子	30
表 2-3	燃料电池技术术语候选词语的词长数量分布	32
表 2-4	燃料电池技术候选术语的词性分布	33
表 2-5	不同字长的文档频率前 5 名热点术语	35
表 2-6	燃料电池技术不同阶段的热点术语对比	37
表 2-7	增长率排名前 20 位的术语	38
表 2-8	下降率排名前 20 位的术语	39
表 2-9	只在一个阶段出现的术语	40
表 3-1	N 个高频术语形成的共词矩阵 C	46
表 3-2	用于共词分析的文档及对应的术语示例	49
表 3-3	FAO-780 数据集中标引术语数的分布情况	54
表 3-4	部分最大频繁项集挖掘结果	55
表 4-1	《词林 2》的词语编码表	61
表 4-2	实际应用中一些术语间的语义相似度	67
表 5-1	人工标注 ISTIC-NEV 分类的 1023 个词语的信息摘要	76
表 5-2	聚类结果 C 和 C' 形成的联列表	78
表 5-3	两种聚类结果的一致/不一致表	79
表 5-4	各种方法的聚类结果比较	82
表 6-1	与“Gamma Knife”和“Elekta”实体对有关的特征	91
表 6-2	语义关系抽取模型中的符号说明	92
表 6-3	GENIA 和 EPI 实验数据统计信息	101
表 6-4	GENIA 和 EPI 的准确率、召回率和 F 值	103
表 6-5	95% 置信度双尾配对样本 t 检验	103
表 7-1	逻辑错误列表	109
表 8-1	科技文献主要元数据及特征分类	118

表 8-2	符号及表示的意义	121
表 8-3	融合科研人员特征的主题模型比较	125
表 8-4	融合时间特征的主题模型的比较	128
表 8-5	融合参考文献特征的主题模型特点比较	133
表 9-1	实体主题模型中用到的符号及意义	144
表 9-2	两个论文和专利文档标题样例	149
表 9-3	论文和专利资源主题关联示例	153
表 9-4	CHEMDNER 和 CHEMDNER-patent 语料统计量信息	155
表 9-5	论文和专利资源主题关联效果比较	157
表 10-1	技术生命周期各阶段在论文和专利资源中的表现特征	170
表 10-2	专利指标的计算方法和含义	172
表 10-3	成长曲线及其特点	175
表 10-4	技术生命周期判断方法优缺点	177
表 11-1	2011 年国外图书情报类核心期刊的影响因子排名	187
表 11-2	文献统计分析归类实例	188
表 11-3	Logistic 拟合曲线参数	195
表 11-4	情报技术五阶段生命周期的分界点	195
表 12-1	构建模式对比	204

第一章 绪 论

1.1 背景及意义

当前,全球技术发展日新月异,大量的技术创新活动为企业提供了比以往更多的技术发展机会,企业必须识别和把握可能出现的技术发展机会,才能提高企业的技术创新能力和国际竞争力。论文和专利信息资源是全球最大的两个技术信息源,具有生命力和潜在商业价值的技术机会在这两大信息源中通常会留下一定的足迹,只是这些足迹不太明显,很难被觉察到,而且不同的信息资源揭示技术发展阶段的不同信息。

论文信息资源通常被用于测度基础科学研究活动的水平,而专利信息资源则被用于测度产业技术的创新水平,因此,论文和专利信息资源的融合分析对于把握技术机会、理解科学和技术之间的联系、提高技术创新水平具有重要意义,而且大量的实证研究^[1-4]也表明,这两种资源的集成揭示分析有助于理解技术发展趋势、产学研关系、度量创新水平等。

技术机会发现是指通过对某领域内海量信息资源的深入挖掘和分析,在掌握已有技术发展趋势及其相互关系的同时,发现最新技术动向,推断该领域可能出现的技术形态或技术发展点。通过对论文和专利信息资源的融合分析,可以反映技术发展现状、挖掘研发热点、预测发展趋势、揭示竞争对手的技术实力与战略布局。

实际上,技术机会发现的有效性主要取决于两个因素:①多源异构信息资源的序化和关联程度。但在当前的文献服务体系中,论文与专利文献资源表现出明显的局部有序但整体无序的孤岛特征^[5],两者之间没有进行有效的整合,造成了数据冗余、相互关联程度低,大量的信息孤岛出现。②数据分析与挖掘方法的效用。目前的研究及实证多是针对单一或同类信息资源,对多源异构信息资源尚缺乏有效的处理手段,导致分析结果存在较大偏差,从而经常错过真正有价值的技术机会。因此,有必要在增强论文和专利信息

资源间的序化和关联程度的同时,突破多源异构信息资源深度融合分析的关键技术。

随着大数据时代的到来,机器学习和数据挖掘技术的飞速发展,以及新的技术机会发现理论和方法的涌现,为融合论文和专利信息资源的技术机会发现带来了新的机遇和挑战。相对于企业应用和产业政策制定的实际需求而言,对该问题的研究出现了较大的滞后,出现了大量如上所述的前沿问题有待人们加以解决,使得在多源异构信息的基础上发现技术机会成为一项极具研究意义的课题。把这些问题研究清楚,有助于洞察未来的技术机会,发现科学和技术之间的发展规律,为国家制定产业政策,为创新主体把握特定技术研发投资方向、寻找合适的战略合作伙伴等提供科技情报支撑。

1.2 国内外研究现状

从多源异构资源融合分析、技术机会发现及技术生命周期模型 3 个方面对国内外研究现状进行简单介绍。

1.2.1 在多源异构资源融合分析方面

20 世纪 70 年代,CHI Research 公司与美国国家科学基金委员会合作利用论文与专利指标评价公司的价值^[6]。1980 年,Narin 团队^[7]分析了专利对论文的引用在数量、时间和种类等方面的分布,有三大发现:当时美国专利对论文的引用很普遍,而且 90% 的引用都是基础和应用科学的论文,而非工程技术文献;专利所引用的论文平均发表于专利申请前的 3~5 年,而这个时滞与论文之间相互引用的时滞是相当的;专利引用的论文往往集中在 SCI 收录期刊的核心部分,这与论文引用很相似。Narin 于 1994 年正式提出了“专利文献计量学”的概念(Patentometrics)^[8],并以专利分析与 GDP 之间的关系为例证,展示专利分析在研究国家生产方面的应用,说明包括专利分析在内的整个文献分析方法,都是技术创新研究的重要工具。Narin 通过专利对论文的引用分析了技术与科学的联动。Leydesdorff 等人^[3,4]基于论文、专利及政策指标提出“产学研三螺旋管理”理论,探讨科学和技术之间存在的内在联系。

2005 年,上海科技情报研究所整合发明专利申请和 SCI 论文收录情况,初步考察了中国高校的创新能力。2006 年,清华大学涂俊和吴贵生运用三