图书情报理论与实践研究丛书

用户专利文献阅读兴趣拓扑结构及其应用研究

王秀红　袁银池　金玉成 著

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书情报理论与实践研究丛书

用户专利文献阅读兴趣拓扑结构及其应用研究

王秀红　袁银池　金玉成　著

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

用户专利文献阅读兴趣拓扑结构及其应用研究/王秀红,袁银池,金玉成著. —武汉:武汉大学出版社,2016.12
图书情报理论与实践研究丛书
ISBN 978-7-307-19088-7

Ⅰ.用… Ⅱ.①王… ②袁… ③金… Ⅲ.专利文献—阅读—兴趣—研究 Ⅳ.G306.4

中国版本图书馆 CIP 数据核字(2016)第 323499 号

责任编辑:方竞男 责任校对:刘小娟 装帧设计:吴 极

出版发行:**武汉大学出版社** (430072 武昌 珞珈山)
(电子邮件:whu_publish@163.com 网址:www.stmpress.cn)
印刷:虎彩印艺股份有限公司
开本:720×1000 1/16 印张:8.25 字数:150 千字 插页:5
版次:2016 年 12 月第 1 版 2016 年 12 月第 1 次印刷
ISBN 978-7-307-19088-7 定价:52.00 元

作者简介

王秀红，1975年出生，江苏南通人，工学博士，管理科学与工程博士后，高级专利分析师，先后任系统工程和情报学硕士生导师；2011年于美国加州大学访问1年。一直从事文本相似度计算、核函数的构造与应用、专利文献检索、非结构化数据管理、信息行为等方面的研究。近4年，SSCI、SCI、EI收录第一作者期刊论文10余篇，CSSCI收录论文10余篇；以第一发明人申请发明专利6项，PCT申请1项，美国发明专利1项，已授权3项。

特 别 说 明

本专著为国家自然科学基金“专利文献的要素组合拓朴结构及向量空间语义表示与相似度计算研究”(项目编号 71403107)、中国博士后科学基金第七批特别资助“综合位置和语义的专利文献核函数构造及相似度计算研究”(项目编号 2014T70491)的研究成果之一。本专著由江苏大学专著出版基金资助出版。

本研究的相关成果已申请中国发明专利、美国发明专利。

前　言

专利文献包含丰富的技术创新信息，是重要的知识载体，具有重要的技术参考价值。阅读行为能在一定程度上反映用户的兴趣所在。将知识图谱与定性研究方法相结合来分析国内外研究现状，发现基于专利文献阅读行为的研究目前处于空白。本书主要利用专利文献自身特有的编排体例和内容特征，通过对用户专利文献阅读行为的实验分析，挖掘其对专利文献要素的兴趣拓扑结构，构建兴趣模型指标，计算兴趣度；再挖掘用户平时阅读的兴趣关键词及其所对应的国际专利分号（IPC）小类，结合兴趣拓扑结构设置关键词位置权重，从而发现待推送的专利文献，提高推荐的针对性、准确性和全面性。

本书具体包括以下几个方面的内容。

（1）分析了国内外阅读行为兴趣研究知识图谱。选择 Web of Science 数据库中阅读行为兴趣研究的相关文献，利用 Citespace Ⅲ、HistCite 软件及数据库在线分析功能进行分析和处理，揭示基于阅读行为兴趣研究的跨学科属性、重要区域、机构、研究热点、高影响力著者、代表性文献及核心刊物。

（2）构建了用户专利文献阅读行为兴趣的理论模型。利用 Tobii T60XL 眼动仪对用户阅读的眼动数据进行采集、处理与分析，从而发现用户对各兴趣区的偏好差异；通过视线扫瞄轨迹采集用户在阅读过程中的视线扫瞄模式，进一步确定了兴趣模型指标，具体包括相对访问时间、相对注视次数、瞳孔直径缩放比和回视次数；采用改进的层次分析法对兴趣模型指标进行权重设计，提供了构建用户专利文献阅读兴趣模型的一般方法。

（3）挖掘用户专利文献阅读兴趣并进行拓扑结构表示。进一步以 Tobii T60XL 眼动仪为工具，选用 26 名视力正常、从事科研工作的江苏大学教师和研究生，记录其阅读专利文献的眼动行为。实验中，将专利文献划分为 12 个要素即 12 个兴趣区，结合用户阅读实验中的阅读行为、测前问卷、测后 RTA 访谈进行分析；构建相对访问时间、相对注视次数、瞳孔直径缩放比三个指标，

计算用户对各要素的兴趣度；构建各兴趣区回视次数的关系矩阵。根据用户在各个要素的兴趣度及不同要素间的兴趣关联程度，构建用户专利文献阅读的兴趣拓扑结构。

(4)利用兴趣拓扑结构提供主动推送微服务模型。构建专利文献树形结构模型并进行文本预处理，利用兴趣度赋以兴趣特征词微兴趣权重；结合兴趣拓扑结构赋以特征词位置权重用以计算专利文献相似度计算过程中词项的位置权值；结合特征词的长度对 TF-IDF 算法进行改进，计算特征词的综合权重，从而构建带有用户兴趣的专利特征词向量空间模型。挖掘利用兴趣特征词所对应的 IPC 小类，限定待检索的专利技术领域，在专利数据源中进行检索，得检索结果；将用户兴趣的专利特征词向量与检索结果进行相似度匹配计算和排序，得用户 TOP N 兴趣专利文献。

由于著者水平有限，书中难免有不妥之处，敬请广大读者批评指正。

著　者

2016 年 11 月

目　录

1 绪　论

1.1 研究背景

世界知识产权组织发布的《2015 世界知识产权指标：突破式创新与经济增长》指出：2014 年全球发明专利申请量增长迅速，增幅达 4.5%。全球近 270 万件的发明专利申请中，约 30%来自中国，超过美国和日本发明专利的申请总量。中国专利申请量成为推动 2014 年全球发明专利申请量增长的主要动力。其中，中国 92.8 万多件，增长率为 12.5%；美国 57.9 万多件，日本 32.6 万多件。

专利文献的特点在于集科技、法律、经济信息于一体。从它所包含的技术内容来看，它具有广博性；从它的结构编排、著录格式和采用的分类体系来看，它具有统一性和标准性；从它所阐述的权利要求及许多与法律相关的著录数据来看，它具有法律性。

专利文献充分体现了专利制度的法律保护功能和公开特点；同时在专利审查和国际交流中发挥着重要作用。专利文献是体现专利制度根本目的的媒介，是实施法律保护的依据，是进行科学审查的依据，为经济、贸易、国际交流提供参考，传播专利信息、促进科技创新。专利文献不仅全面、完整地记录了专利活动过程，反映了各领域技术活动的现状，还可用于挖掘特定领域技术活动的发展历程。从一件发明创造所拥有的同族专利的数量，以及世界各国就同一技术问题公开的专利文献数量，可分析申请人的全球市场战略以及技术的重要程度；将各国就同一技术问题所提出的不同解决方案进行对比，可以确定较为科学与经济的技术发展战略等。

专利文献在科技创新等方面具有很好的应用。专利文献为国家宏观决策提供依据，为技术评价与预测、进行科技实力对比提供参考。专利文献对技术进步有着重要的促进作用，能活跃发明创造思维、促进创新活动，为科研选题、制订科研计划提供重要的参考价值。专利文献在企业的生存与发展中也发挥重要作用。专利文献在国际技术贸易中的重要作用，可帮助确定技术进出口的最佳目标，避免技术进出口中不必要的损失。

专利文献的技术情报作用主要体现在以下几个方面：①技术创新。有关研究表明，在技术研发中阅览专利文献能缩短 60％的科研时间，节省 40％的科研经费。②新产品开发。新颖性、创造性和实用性为专利技术应具备的授权的三性，且专利文献对技术的描述往往比较具体、系统，简洁明了且完备，往往附有各类附图及公式等，各领域技术人员能够依据专利文献描述的技术方案实施专利内容，专利文献被很好地运用于开拓思路、开发研制或生产、销售新产品，避免侵权。③研发立项。为避免科研成果重复研究，在选题和立项前可以采用文献调研手段以及专利信息分析方法，对所研究的课题进行查新，明确其内容的新颖性以及创造性，并且以检索分析结果作为能否立项的依据。充分借助专利文献内容广、记载详、报道快的特点，阅读专利文献，以了解科技发展的最新动态，发现相关领域内待解决的技术问题，寻找研究着力点或创新点已成为目前科研创新的有效方法之一。

专利文献内容格式统一、形式规范，专利数据库越来越成熟，数据来源客观、有效且获取便利。用户阅读专利文献有 80％以上比较关注专利的技术内容。通过访谈得知，大部分的科研用户在需要撰写专利申请文件的时候，会集中阅读专利文献，以了解背景技术或已有的技术方案及存在的技术问题，同时规避侵犯他人专利权，且避免自己有专利而不能实施的问题，从而最终确定自己的技术方案与保护范围，使得独立权利要求的划界更准确。

随着科学技术的快速发展，科技文献资源得到几何级别的增长，如何简便、高效地获取自己所需的数据资源，成为用户普遍关心的问题。用户希望能按照自身的需求订制数据资源，然后由专业的服务机构主动、及时地为其提供所需的信息资源。而采集用户阅读、浏览及访问中的兴趣行为，对用户的正常阅读不产生干扰，通过这些行为进一步确定用户感兴趣的信息，并且把这些信息主动推送给他们，有助于提高推送服务的亲和力、精准率和召回率。用户信息行为是用户兴趣研究的重要内容，能够帮助识别用户兴趣、偏好及需求的信

息行为，包括信息需求行为、信息查询行为、信息交互行为、信息浏览行为、信息选择行为和信息利用行为等。其中，信息浏览（含阅读）行为是考察用户阅读文献时，其最关注文献中的哪一部分，以及文献中各部分之间兴趣关联情况。

现有的个性化推荐模型的方法大致有以下三类：①基于内容分析、协同过滤和组合推荐，均以用户兴趣分析为基础，建立适合用户的兴趣模型。②根据用户参与兴趣模型构建过程深度的不同，其构建技术可分为用户手工订制、示例用户引导以及自动分析建模。自动分析建模是在兴趣模型构建过程中，无须用户输入其感兴趣的主题或特征词，也无须用户对阅读过的内容标注其兴趣度值，系统根据用户的阅读内容和阅读行为自动地为用户构建兴趣模型的方法，此方法有利于提高个性化服务系统的易用性。③现有的用户兴趣模型难以精确描述用户的个性特征和文献需求，从而难以提供优质的主动推送服务。

视觉是人类获取信息的主要通道。心理学家研究表明，人类获取的信息有83%来自视觉。其中，眼动技术是一种可靠、有效的方法，可通过眼动研究分析用户在阅读过程中的注意力分配情况。眼动记录方法完成对阅读过程的真实测量与还原，为阅读研究提供实时测量的技术支持。该方法不会过多干扰正常的阅读过程，能够实时、连续地记录用户的阅读行为，能对用户在阅读内容的重点区域上的行为进行分析。综合运用眼动指标不仅可以挖掘用户剖析问题的新视角，还可以展示用户词汇加工的时间进程。眼动仪被很好地应用于用户的阅读行为研究，浏览、访问或阅读中的眼动研究可充分挖掘用户偏好和潜在需求。

本研究通过分析眼动仪记录的数据，能更清楚地了解用户对文献任何一要素的访问时间、注视次数、眼跳及瞳孔直径变化情况等阅读行为。此外，眼动仪与生理多导仪、行为监测系统、用户体验软件、行为活动记录软件等通过接口能结合使用，共同采集眼动、脑电、生理方面的数据信息。本研究拟利用用户阅读时眼动数据，深入研究用户的阅读兴趣，从而为具体的用户提供基于阅读行为的主动推送微服务模式。实施时还可进一步综合脑电、肌电、眼电、心电、呼吸率及皮电等行为数据，进行更全面、深入的分析。

本研究将阅读行为记录技术和问卷访谈法相结合，对传统的主动服务模型进行改进，主要包括以下过程：计算用户阅读专利文献行为系数权重及阅读

兴趣度；利用用户基本信息、阅读文献内容及阅读行为数据等，结合问卷访谈，采用向量空间模型表示用户兴趣模型；基于文献采集与匹配系统提供主动推送文献给用户；通过专利文献的阅读行为实验，验证兴趣模型核心部分的有效性。

1.2 研究意义

1.2.1 理论意义

目前识别用户的阅读兴趣点，通常是依据检索词、字段判断用户的需求，有待全面、准确挖掘用户的需求。专利查询为专利文献的获取提供了一个有效的途径，用户阅读行为研究可充分挖掘用户潜在需求，已在网站可用性、软硬件测试等方面得到广泛的应用，但针对专利文献阅读行为的研究目前国内外属于空白。用户阅读行为是一个新兴的研究领域，对其进行研究是了解用户相关情况的一个重要途径，用以发现用户信息利用过程中存在的问题，以及用户阅读的规律与偏好等。通过对用户信息行为的研究，信息服务者以用户为本，从用户角度出发，为用户提供清晰和可理解的信息。

本研究以专利的阅读行为为基础，进一步挖掘专利阅读指标，构建专利阅读兴趣模型和主动推送模型；利用用户在不同专利文献要素的兴趣的差异，设置专利词项的位置权值，用于专利文献相似度计算，同时为提高专利推荐的准确性、客观性与可行性，提供一般理论方法。

1.2.2 实践意义

专利文献的格式标准、规范，且具有统一的国际专利分类(IPC)体系结构，便于数据挖掘，为课题研究提供专利数据源。专利推荐为专利文献的自动个性化获取提供了一个有效的途径。用户专利需求在一定程度上能够通过阅读行为反映出来。通过专利阅读行为可以挖掘潜在、具体的。用户并没有用

语言表达出来的用户需求信息，用以揭示用户专利文献阅读的兴趣拓扑结构，用以挖掘用户对专利文献各部分的偏好，以及高度关注及需求的信息。

本研究在基础行为分析、大声思考和眼动追踪的组合模式实验中，用户做到专注于自己的任务而不用说出自己的想法。当总任务或单个任务完成后，实验者向受测者提供可视化浏览路径的眼动视频，重新播放该视频，受测者可以说出当时自己的想法。已有行为记录软硬件 Morae、眼动仪 Tobii T60XL 和 spirit 10 生理多导仪，结合“863”项目课题组成员参加实验，为本研究的模型指标和实证数据分析环节提供可靠的技术支持。通过对专利阅读兴趣进行研究，提出科学、可行的推荐服务一般理论方法，提高专利推送服务的准确性和全面性，且对专利转化实施过程中专利价值的评价、专利文献相似度计算中词项位置权系数的科学设置、相关专利的检索和利用等具有重要的实践意义。

1.3 研究方法

本研究利用文献计量方法进行现状研究，采用定性与定量相结合的方法构建兴趣模型；利用问卷调查与访谈法、实验观察与记录法相结合进行实证分析，具体包括以下几种方法。

(1)文献计量法：利用文献计量法对 Web of Science 数据库中阅读行为的兴趣研究进行检索，对有效数据进行统计分析，挖掘其学科分布、重要区域国家、机构合作、研究热点、高影响力著者、代表性文献以及期刊分布等情况。

(2)比较法：对多种阅读行为指标进行比较，提出适合专利文献的阅读行为指标，并进行详细说明，揭示指标的可行性。

(3)问卷调查与访谈法：通过用户调查研究方法来对相关的数据进行搜集，是其最重要的研究方法。对用户进行测试之前，通过制订调查问卷获得用户专利需求基本信息。对用户测试结束后，当用户对专利需求有了更直观、深刻的体会后，再进行测试后的问卷调查，得到用户相应反馈信息，用以对测试数据进行对照与结合分析。在测试和问卷调查完成后再进行用户访谈，通过访谈获得调查问卷以及现场测试以外的信息，将测试过程中遇到的问题与测试对象进行交流，以便更深入地了解用户对专利知识服务的需求与建议。

(4)大声思考法:用户在检索期间,大声将其操作行为及思维过程进行表述出来,所有的独白被录音,后期研究人员再依据录音辅助数据分析。全过程研究人员只是旁观,除必要的提示外,不与用户进行交流。大声思考法能帮助研究人员精确、详细地了解用户伴随行为变化的思维过程。大声思考法有两种主要的实施方法,一种是并发大声思考,要求用户在做任务的过程中说出自己的想法;另一种是回溯大声思考,要求用户在某一个任务或所有任务都完成后,必须描述自己在这个过程中的经历。本测试中的用户为"863"项目成员,在测试过程中,鼓励其大声思考,并做标记。而并发大声思考使用过程中可能会增加用户的认知负担,因此在分析眼动数据的可用性测试中用到回溯大声思考。

(5)实验观察与记录法:在研究用户阅读专利时,采用实验方法,观察和记录用户肢体运动和使用阅读工具的现场表现。在信息查询、阅读场景中,采用仪器设备对用户阅读专利的行为进行全程观察与记录,通过录制用户阅读、浏览行为整个过程的视频,回放视频与标记分析。利用 Morae 和 Tobii T60XL 眼动仪标记以及跟踪用户的搜寻过程和结果,通过不同维度的比较分析,结合用户特征,进行分析,找出不同因素间的因果关系或相关关系以及用户的潜在需求和兴趣点。

(6)数学方法:通过对经过处理后的指标数据进行定量分析,计算用户对专利文献各部分的兴趣度。采用改进的 TF-IDF 特征词提取方法、专利文献各要素加权法、向量相似度计算方法等,即利用数学方法对指标数据进行科学计算,量化成指标分值计算专利相似度。

最后通过示例分析法,选择合适的眼动实验案例,采用上述研究方法,对提出的基于专利文献阅读行为用户兴趣模型的方法进行计算并分析。

1.4 研究框架

本研究总体框架如图 1.1 所示。