基于专利和期刊异种资源的集成数据库构建及其应用研究*
——面向科技发展趋势的精准预测

□ 霍翠婷 李颖 赵蕴华 桂婕 / 中国科学技术信息研究所 北京 100038
张静 邱敏清 / 北京万方数据股份有限公司 北京 100038

摘要：目前国外基于专利和期刊两类异种资源的集成数据预测科技发展趋势已成为重要的研究方向。由于集成数据库构建和分析的复杂性，国内还延续着专利、期刊数据的独立分析方法。为此，文章提出了基于专利和期刊两类异种信息资源的集成数据库构建方法，并对其应用进行了探讨。具体地说，文章分析了知名数据库提供商的专利和期刊的数据结构，从含义上选择了专利和期刊中相互匹配的字段，设计了两类异种数据信息资源的集成数据库，对该数据库的应用进行了概念分析。最后，给出了文章的结论及未来课题。

关键词：专利，期刊，异种信息资源，集成数据库构建，科技发展趋势预测
DOI：10.3772/j.issn.1673—2286.2010.09.008

1 引言

为了支持自主创新型国家的发展，基于专利、期刊等信息资源数据分析结果的科技领域研究报告，成为国家制定科技发展战略规划的重要依据，这也是中国科学技术信息研究所及所属各机构的重点工作。目前，科技领域研究报告中的数据分析及可视化，还延续着传统的专利、期刊等不同种类信息资源数据的独立分析手法，这是由于集成数据库构建和分析相对复杂而造成的。然而，在国外，几年前就出现了基于专利技术信息和科技期刊信息异种数据源的集成数据库的分析方法[1]，这一方法被认为更能准确地预测科技发展方向，是主流的研究方向。

在上述环境下，为了实现科技发展趋势的精准预测，产出更精准的科技领域研究分析报告，在中信所现有的异种数据库构建、分析及应用的基础上，本文进一步提出了如图1所示、基于专利和期刊两种不同的信息资源集成数据库的构建方法，并在此集成数据库上进行集成数据库的检索、分析及可视化等应用，以产出高质量领域研究报告的框架。本文的内容构成如下：异种信息资源集成的概念解析、集成科技数据库的构建研究、集成数据库的应用研究以及未来的规划和课题。

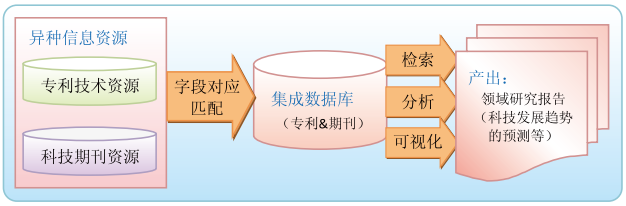

图1 基于专利和期刊异种信息源集成库的构建及应用框架

2 异种信息资源集成的概念解析

本研究选择以专利文献和期刊文献信息作为代表性的异种信息资源来构建集成科技数据库，故首先对

* 本文得到中国科学技术信息研究所“主要国家重点研发领域及主要科技计划和重大专项发展现状的调查与监测平台建设”项目（项目编号：2009FY240100）的“科技基础性工作专项”、国家科技部“十一五”科技支撑计划（项目编号：2006BAH03B03）、中国科学技术信息研究所重点工作项目（项目编号：2009KP01—7—1）、中国科学技术信息研究所2009年度预研基金项目（项目编号：YY—200906）等项目的资助。

专利文献信息、期刊文献信息以及数据集成等相关概念进行简单阐述，进而探讨基于这两类异种信息资源进行集成的可行性。

2.1 专利文献信息

专利文献是一种集工业产权情报、技术情报、商业与经济情报于一体的信息源，是自主创新的基础数据库。专利文献数量巨大且能够涵盖最新的科技信息，并且以新颖性、首创性和实用性成为越来越受到重视的信息资源，其实用价值已被广大科技人员所肯定。专利文献中含有大量的专利技术、经济、法律、战略信息，对这些专利信息进行有效的组织和开发利用可为企业技术创新、高校的科研提供重要的情报支持[2]。专利文献的二次信息（元数据）所包含的基本字段有：日期信息（申请日期、公开（公告）日期、优先权日期）、号码信息（申请号、公开（公告）号、专利号、优先权号）、发明人、申请人（专利权人）、国际专利分类（IPC）、标题、摘要、权利要求项等。基于专利进行科技发展预测都离不开这些基础数据。

2.2 期刊文献信息

期刊文献作为信息源的一种类型，在知识经济时代，是一种有别于图书的特殊文献类型，其使用率远远高于图书和其他出版物。据统计，目前人类80%的情报信息来源于期刊，期刊文献信息正以其内容新、时效强、情报价值高、流通范围广的特点，在网络化的知识经济时代发挥着愈来愈积极的作用[3]。期刊文献的二次信息所包含的基本字段有：年份、关键词、作者、作者单位、地址、中图分类号、标题、摘要、期刊来源等。它们也是科技发展预测的有价值的参考数据。

2.3 信息集成

信息集成是一种或是针对某个既定目标，或是面向某项特定的任务，对信息进行组织和管理的概念，亦是一种使相关的多元信息有机融合并优化使用的理念。集成具有两层含义，即集合与组成。所谓集合，就是将不同分布地的信息资源通过现代技术链接在一起，运用信息技术和应用软件，形成科技信息集成服务的环境；所谓组成，就是指所集合的各种信息资源，按照用户的需求，通过各种信息技术和手段，进行规范、科学地组织，以供用户方便快捷地使用[4]。可以看出，信息的集成，更能综合、全面地挖掘信息的应用价值。

2.4 专利文献和期刊文献信息集成的可行性分析

近年来，文摘索引数据库的发展呈现出模式化、规模化的趋势，各大数据库商不断购买其他提供商的数据库或者整合一些免费资源，作为其原有内容的有效补充，旨在为用户提供一定范围内最全面的文献信息资源。比如，ISI Web of Knowledge是集成各种数据库的典型代表。而这正是本文将专利文献信息和期刊文献信息进行集成的研究前提。

从信息属性方面来看，专利文献信息和期刊文献信息都具备可共享性、记载性、创新性（前沿性）等一般属性；从数据结构方面来看，专利文献信息和期刊文献信息均为结构化的信息资源；从信息资源主要字段语义方面来看，专利文献信息的基本著录项和期刊文献信息的基本字段均属同一语义范畴，参见表1（字段说明参见表2）。比如，专利文献的发明名称和摘要与期刊文献的文章标题和文摘等对应匹配。由此可见，专利文献和期刊文献的集成数据库的构建具有可行性。

表1 专利公报与科技期刊项目的对应匹配

专利公报	科技期刊文献
发明名称	标题
专利摘要、权利要求项	文摘
国际专利分类号IPC、ECLA、USPC、FI&F-term[5]*	主题词（关键词、叙词）
公开号	序列号
发明人	作者
申请人（专利权人）	所属机构
申请号	编码（ISSN、CODEN）

* 注：ECLA为欧洲专利局的分类系统；USPC为美国专利局分类体系；FI&F-term为日本专利局分类体系。

3 集成科技数据库的构建研究

本研究基于专利文献信息和期刊文献信息两类异种信息源，提出构建集成科技数据库的主要流程（如图2所示），该流程主要包括分析多个知名数据库提供商的专利及期刊文献数据库结构，进行专利文献和期刊文献中含义相同的字段匹配研究，集成数据库结构设计等。

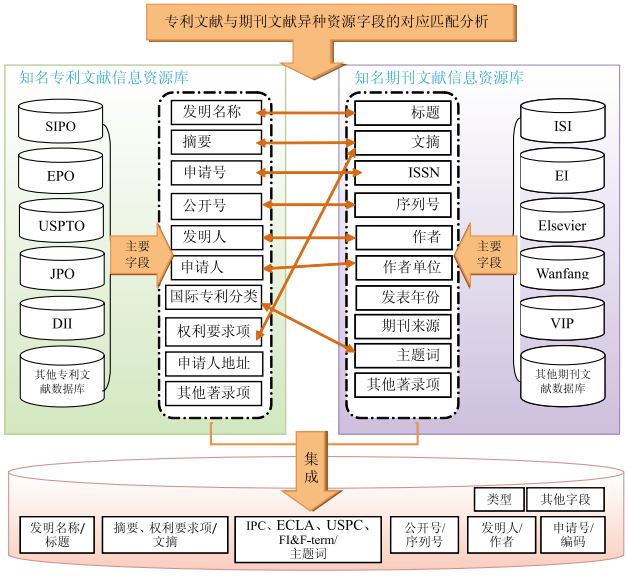

图2 基于专利文献和期刊文献信息异种数据源构建集成科技数据库的主要流程

3.1 知名数据库提供商的专利和期刊数据库的结构分析

在构建专利和期刊集成数据库之际，如何使数据库中专利技术公报的信息与科技期刊文献信息的数据项目得到含义层面上的对应？这是本研究的最大课题。为此，我们对知名数据库提供商的专利（中国知识产权局【SIPO】、欧洲专利局【EPO】、美国专利商标局【USPTO】、日本专利局【JPO】、德温特世界专利创新索引【DII】等）和期刊数据库（ISI、EI、Elsevier、Wanfang、VIP等）的结构进行了较为全面的比较分析，抽出了专利与期刊两类资源的主要字段，通过选择专利和期刊中共有的信息，进行相应的字段匹配研究（参见3.2节），从而集成了专利和期刊两类异种资源，进而构建出集成数据库，如图2所示。

3.2 专利和期刊数据库中含义对应字段的匹配研究

由于需要分别对来自专利文献数据库和期刊文献数据库的两类异种文献信息进行集成，两者在格式和著录项等各方面存在一定程度的差异，故需对这两种文献信息可匹配的部分进行研究并建立相应的对齐关系。在调研大量专利数据库和期刊数据库结构、并参考国外案例的基础上，发现专利与期刊在含义上对应的字段主要如表1和图2所示。比如，专利摘要和权利要求项与期刊的文摘、国际专利分类号IPC/ECLA/USPC/FI&F-term与期刊的主题词（关键词、叙词）等并非简单的一对一对应关系，需要重点注意，必须保证匹配的专利和期刊字段所包含的信息内容一致。对字段匹配的研究，为集成科技数据库结构设计奠定了基础。

3.3 专利和期刊的集成数据库结构设计

根据专利文献和期刊文献两类异种资源的字段匹配的研究结果，构建出了面向专利文献和期刊文献信息的集成数据库，主要集成字段如表2所示。

4 专利和期刊集成科技数据库的应用研究

我们在本文引言部分中提到，专利和期刊集成科技数据库的构建，最终是为了实现基于异种资源统合的数据检索、分析和可视化，从而产出能精准预测科技发展趋势的领域研究报告。图3、4的概念设计代表了集成数据库重要的应用途径：集成检索、集成数据分析和集成数据的可视化。

首先，基于专利和期刊的集成数据库，将与科技发展趋势预测有关的两类异种资源有机地集成在一起，让用户利用单一界面，更加快速、有效地检索到

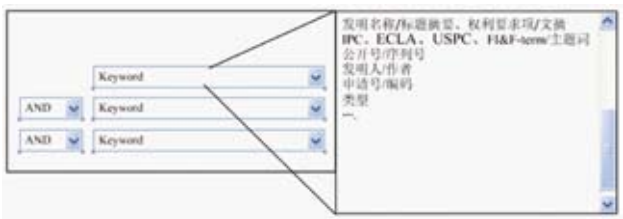

图3 专利和期刊集成数据库检索的概念图

表2 基于专利文献和期刊文献信息的集成数据库主要字段设计

字段	专利数据库对应字段	期刊数据库对应字段	字段说明
发明名称/标题	发明名称	标题	专利的发明名称或者期刊的文章题目。
摘要、权利要求项/文摘	摘要、权利要求项	文摘	专利说明书摘要和权利要求项与期刊的文摘对应。专利摘要是对发明或实用新型专利内容的简要概括，摘要是对专利的综合说明，权利要求是专利申请人请求专利保护的范围，记载了解决技术问题的必要技术特征，是保护专利权利的重要依据，也是专利的核心内容；期刊文献对应的文摘是为了使读者在阅读论文之前对文章的主题、研究方向、方法及结果有大概的了解。
IPC、ECLA、USPC、FI&F-term/主题词	专利的技术范畴	主题词	专利的范畴标引主要采用国际专利分类法IPC，它是对专利技术的分类，是中国及大多数专利局所采用的分类体系，也是ECLA、USPC及日本FI&F-term分类体系的基础；期刊的范畴分类主要是主题词，包括叙词或关键词，从内容含义上，它与专利的范畴分类对应。
公开号/序列号	公开号	序列号	专利的公开号与期刊的序列号都具有文献流水号的含义。
发明人/作者	发明人	作者	发明人是专利的创造者；而作者是文章的撰写者。
申请号/编码	申请号	编码	专利的申请号与期刊的编码，如ISSN，都具有唯一标识的作用。
类型	-	-	区分专利与期刊两类异种资源。
其他字段	-	-	可根据实际需求进行扩展。

图4 集成数据分析和集成数据的可视化的概念图

所需数据；其次，比起单一信息资源，即专利或期刊数据，基于专利和期刊集成数据库而检索到的结果数据，更加综合地反映了科技发展趋势。在此基础上，运用单一种类数据资源的分析和可视化手段，进行集成数据库数据的分析和可视化应用，产出的领域研究报告价值更高。同时，由于集成库中有文献类型的区分，也可以进行单一种类数据的分析等功能。

5 结论与未来课题

本研究提出了基于专利文献和期刊文献两类异种信息资源来构建集成科技数据库的方法，并根据集成科技数据库进行了应用研究的探讨。该方法主要针对各类较为规范化（比如，主题词及分类的规范化、发明人及作者的归一化【清洗】）的专利与期刊数据库，内容适用于各类科技资源，同时也不受限于语种，主要解决国内长期基于单一资源进行科技发展趋势预测的局限性。

在国外研究趋势的驱动下，我们进行了集成库的构建，这仅仅是一个开端，作为未来的课题，我们将利用该集成库，选择国家重大领域，比如太阳能电池，进行数据的导入、分析和可视化研究，对单一种类数据的分析结果与集成库的分析结果进行实际的对比研究，从而产出更优的技术领域研究报告，推进国家科技创新战略的实施，进一步带动地方信息研究所的发展。

参考文献

[1] KOUDA A, MORITA U. Development of a database system integrating technological part of patent and science & technology information [J]. 情报知识学会,2008(2):200-203.
[2] 陈燕,等.专利信息采集与分析[M]. 北京:清华大学出版社,2006.
[3] 王蔚,展群霞,赵肖峰.知识经济与期刊文献资源[J]. 兵团教育学院学报,2001(2).
[4] 毕强,史海燕. 网络信息集成服务研究综述[J]. 情报理论与实践,2004(1):20-24.
[5] 李颖,等. 日本专利检索体系中主题分类“FI/F-term”的理论与应用研究[J]. 数字图书馆论坛,2008(11):11-17.

作者简介

霍翠婷（1984-），加拿大温莎大学硕士，研究方向：专利分析、专利数据挖掘、信息资源服务研究等。通讯地址：北京市复兴路15号 中国科学技术信息研究所 100038。E-mail: huoct@wanfangdata.com.cn
李颖，日本筑波大学信息学博士，信息系统专业。主要研究方向是：Web信息知识系统、基于XML的跨媒体数字出版。最近研究课题：基于XML的数字出版、基于DOI的文献链接系统、跨语言检索、专利分析、数字版权保护等。通讯地址：北京市海淀区 复兴路15号 中国科学技术信息研究所 信息技术支持中心 100038。E-mail: liying@istic.ac.cn
赵蕴华 (1967-），硕士，副研究馆员，研究方向：信息咨询和信息资源服务研究。 通讯地址：北京市海淀区复兴路15号，中国科学技术信息研究所 100038。
桂婕 (1976-)，博士，助理研究员。研究方向：专利分析和科技创新管理。通讯地址：北京市海淀区 复兴路15号 中国科学技术信息研究所 信息技术支持中心 100038。E-mail: guij@istic.ac.cn
张静（1975-），博士，研究方向为数据挖掘、商业智能、信息分析；发表相关文章十余篇。目前研究方向为专利信息挖掘、开放获取、知识组织。通讯地址：北京市海淀区复兴路15号，中国科学技术信息研究所219室 100038。E-mail: jane.zh.t@gmail.com
邱敏清（1985-），中国科学技术信息研究所硕士，研究方向：专利分析、信息服务等。通讯地址：北京市复兴路15号万方数据股份有限公司技术研究院 100038。

Construction and Application of Integrated Database Based on Patent and Journal Information Resources: For Trend Prediction in Science and Technology

Huo Cuiting, Li Ying, Zhao Yunhua, Gui Jie / Institute of Scientific and Technical Information of China, Beijing, 100038
Zhang Jing, Qiu Minqing / WanFang Data, Beijing, 100038

Abstract: Currently, research on trend prediction in science and technology based on integrated database of patents and journal resources has become a hot topic. As database integration and analysis using it are complex, independent analysis of patent and journal data has been continued in China. This paper proposes a construction method of integration database based on patent and journal, and discusses its applications. Specifically, the paper analyzes patent and journal DBs schemas provided by leading providers, chooses matched fields according to the meaning of data, designs interrogation database of the two types of heterogeneous information resources, and conceptually analyzes the applications of the integration database. Finally, it gives the conclusions and the future works.

Keywords: Patent, Journal, Heterogeneous information resources, Construction of integration database, Trend prediction in science and technology

（收稿日期：2010-08-15）