基于动态知识关联的科技项目库管理*

赵燕平[1]，张华平[2]，马宝君[3]，赵辉[4]

(1.北京理工大学管理与经济学院，北京 100081；2.北京理工大学计算机学院，北京 100081；
3.北京邮电大学经济管理学院，北京 100876；4.中国科学技术信息研究所资源共享与促进中心，北京 100039)

文　摘：随着中国经济地位和创新能力的提升，我国的科技项目投资逐年大幅提增，中国的科技成果贡献也将对国际科技和环境都产生不可估量的影响。基于此，本文尝试建立大数据背景下的中国的项目库及其与国际关联的项目知识动态分析与集成挖掘工具。通过对中国乃至全球已有项目进行关联分析和动态跟踪，评估国际发达国家的重要项目投资对中国的影响，进而为中国各级决策层提供有关中国重要项目投资动态和未来经济发展方向的辅助决策信息，以及可供全球参考的中国项目数据库服务。

关键词：大数据管理；中国科技项目库；本体与规则；多重关联

中图分类号：TP391.1

1　引言

在国际创新经济的环境下，高效、节能、环保的要求也使得对科技项目的投资变得谨慎和苛刻，各国在考虑未来投资项目时都不仅需要关注本国的实际需要，还同时要参考周边环境和未来发展的要求。然而分布于全球的项目数据是一个动态更新的庞大数据集合，且随着全球环境的变化不断调整。因此有必要建立中国的科技项目数据库，以便对国内科技项目发展情况进行跟踪，同时与国际发展进行对比，并提供国内和国际项目的检索服务。

各国科技项目数据库一般都是由一些重要的基金会或顶级研究协会建立，如美国联邦政府资助的研究与开发项目库(Federal R&D Project Summaries)[1]，美国自然基金奖励项目（摘要）数据库(Science.gov)[2]，美国能源部研发项目（总结）库，美国环保局科学发明数据库(EPA Science Inventory)，欧盟第七框架项目库服务 CORDIS (Seventh Framework Programme (FP7)) [3]，瑞典 KTH 研究项目数据库(KTH Research Project Database)[4]，德国科技研究项目数据库 FORKAT[5]，日本 JST 项目库[6]，加拿大 Alberta 生态基金项目库[7]，亚太经贸合作组织 APEC 项目数据库(The APEC Project Database)[8]，非洲撒哈拉以南地区的投资推广项目库(IPSSA investment promotion in Sub-Saharan Africa)[9]等。中国虽有国家自然科学基金(NSFC)[10]、国家社科基金项目、国家高技术研究发展计划（863计划）、国家重点基础研究发展计划（973 项目）[11]、国家科技支撑计划项目、国家发改委项目、电子发展基金、国家重点新产品计划项目、国家火炬计划项目、国家软科学研究计划项目、国家科技型中小企业技术创新基金等，但是国家自然科学基金、科技部官方网站、各省市的自科基金以及各类项目资助单位网站等其他项目数据目前却还处于分散于各种机构下，没有统一的管理和信息处理机构进行协调和提供检索服务的状态，而且还存在各数据库中混杂有公告形式的已批准项目列表文件，甚至很多还没有公开的项目数据库或公开网页可供检索等问题。

然而有价值的科技项目网络信息的集成，提供高质量的中国科技项目信息服务，不仅需要采集互联网上科技项目库方面的网站，进行信息的实时自动发现和追踪，还需要专门针对科技项目信息的特点，研制从网络上的各种资源动态搜集并整理项目库成果信息的系统。该项任务具有很大的挑战性，具体如下：

1）资源丰富，数据量巨大

多年来国际和国内对于科技项目的支持也带来了“大数据”背景下项目数据信息过载的问题。同时，网络科技项目信息正发挥越来越大的情报指引作用，据 Bright Planet 公司调查显示，早在 2003 年就存在着超过 20 万个项目数据的站点，其资源数量大约为 7500TB，是 www 的 400～550 倍，其中包括 5500 亿私人文档[12]。信息量和资源数目的巨

*基金项目：国家自然科学基金面上项目(No. 61272362), 新疆自治区高新技术计划（No.201212124），中国科技信息研究所相关项目。

作者简介：赵燕平，女，北京理工大学管理与经济学院教授；张华平，男，北京理工大学计算机学院副教授、院长助理；马宝君，男，北京邮电大学经济管理学院助理教授；赵辉，女，中国科学技术信息研究所资源共享与促进中心副研究员。

大给及时或实时的情报搜集带来巨大困难，没有自动化的辅助工具已无法人力完成。

2）信息标准不统一，元数据缺失

互联网上的科技项目信息，很多仅仅是发布的信息，没有元数据和标引字段，需要花费大量人力来根据业务的需要，自动分析所采集到的网页格式，进行标题、摘要、学科分类、相关文章、项目金额等各种关键字段属性的提取。甚至有些是深藏在搜索引擎数据库当中的深网数据，需要核实其年代、数据提供者的身份和有效性等；

3）发展迅速，动态更新

项目数据是互联网上发展较快的动态信息资源，仅 2000 年到 2004 年期间已经增长了 3～7 倍，现已超过 30.7 万个站点，45 万个数据库和 125．8 万个界面，并且还在增长和动态更新[12]。科技项目信息往往是采用分布于各地和各部门的数据库或文件列表的形式存储，一般很难被搜索引擎发现其动态变更信息并进行采集收录；

4）需要整合加工的各种智能工具软件

目前急需对中国的项目信息建立统一的数据库，并对采集的数据进行整合加工入库的、可以智能分析文本、图形、模型等数据和信息的知识工具软件。而且普通的网络采集软件往往只能采集到普通的网页，无法针对文件型数据、Excel 表格、PPT、Model、Keywords、CAD、UML 等数据进行唯一性验证和交叉采集的质量验证。

针对上述问题和挑战，本文将提出一个对科技项目数据库进行关联动态采集的知识系统模型，该模型可以智能地整合从各种科技项目信息的网站、公告、列表、项目库等信息中提取的有关中国重要科技项目的项目信息和关联的中外文项目和资源信息，依据国际科技资源元数据发现标准和质量标准体系[13]建立初步的中国科技项目标准规范并提供唯一性核实，质量验证和动态更新等，以便对我国科技信息投资决策、重大项目关联资源配置提供客观参考依据。

本文的后续内容安排如下，第 2 节描述科技项目采集和项目资源整合的元数据模型；第 3 节设计知识本体驱动的项目采集与整合策略模板；第 4 节介绍知识库驱动的智能系统主要功能与技术方案；第 5 节展示系统检测与成果；第六节给出总结。

2　科技项目采集与资源整合元数据模型

2.1　科技项目发现与采集元数据模型

关于科技项目信息发现与采集的目标元数据模型应包括如下主要信息项：

1)　项目资源元数据信息：主要包含：项目需求分析报告、项目概要设计说明书、项目任务书、项目详细设计说明书以及项目成果，其中项目成果包括项目论文、专利、项目模型、设计图纸、项目阶段性报告、项目审核报告、项目可运行程序及源代码、项目测试报告、项目结题报告、政策、报告、PPT、公告、论著、项目验收报告及相关资料等；

2)　项目参考信息，包括参考国际 ISO 标准，国家 GB 标准，国际、国内权威组织机构的参考模型或指南等；

3)　项目关联信息，包括追加项目信息、相关联或配套项目群信息，国内各部委、办、局的项目关联项目信息、国际投资项目信息、民营或私营企业关联项目信息、后续项目信息等。

2.2　项目资源整合元数据模型

根据 2.1 中不同来源的数据，本小节将建立统一的数据获取、分立的信息整合模型以及项目本体知识库，针对关联信息分别建立不同的采集策略模板，并制定相应的采集项和采集策略和采集规则集，以便于自动处理关联信息。下文括号内为具体的信息项目的缩写与对应规则模板的采集项和质量核对项名称。

项目数据发现元数据 Discovery Metadata (参考 ISO 19115[13][14])

A）项目的标识信息

A.1)　数据资源所用的语言 Language（L）

A.2)　可用于标引的信息

A.2.1)　项目数据资源名称（DT）

A.2.2)　项目数据资源（DP），该资源的地址和启用的可获取资源

A.3）项目时间范围（TE）

A.4）项目联系人信息

与项目资源有关的负责人或组织的联系方式

A.4.1）项目承担组织名称（ON）

A.4.2）项目联系人角色 Role（R），承担者，提供者，拥有者等；　(Res. Provider [Pvd], Owner [Own])

B）项目参考信息

B.1)　参考信息的标识

B.1.1）参考资源标识码 Code（RSC），以字母数字排序的命名空间或其中的一些被引用对象的标识码

B.1.2）标识码对应的参考空间 Code Space（RSCS），可以是该空间的责任者人员名或组织名

C）项目关联信息

C.1）追加项目信息 Additional Projects 是原项目追加的款项或子项目等

C.1.1）追加项目信息标识码（APC）

C.1.2）相关联或配套项目参考空间（APCS）国内各部委、办、局的项目关联项目信息、国际投

资项目信息、民营或私营企业项目关联信息

C.2）后续相关项目信息 Subsequent related proj.

C.2.1）后续项目信息标识码（SPC）

C.2.2）相关联或配套项目参考空间（SPCS），国内各部委、办、局的项目关联项目信息、国际投资项目信息、民营或私营企业关联项目信息。

3　知识本体驱动的项目采集与整合策略模板设计

根据科技项目采集和项目资源整合元数据模型，我们设计了项目本体知识库，针对元数据模型中不同的数据源和数据集关联信息分别建立了不同的采集策略模板，制定了相应的采集项和采集策略、采集规则集等。将网站首页及信息列表页以及链接进行下载并提交策略采集模板处理，进而分析提取出下一个页面的关联信息串，集成脚本分析引擎，解析动态脚本函数，进行页面的合理并发访问，信息的补采与去重。其中项目研究进展动态跟踪策略模板示例如图 1 所示。

图 1　项目数据采集知识库策略模板示例

并将将项目数据库的属性表，与异构的全球数据结构进行匹配，建立了来自不同资源中针对同一项目数据项的多字段整合抽取模板，如图 2 所示。

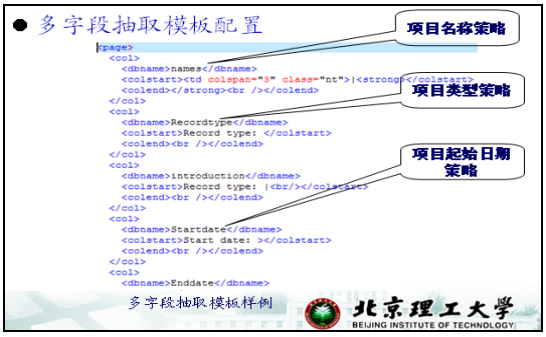

图 2　项目属性信息多字段整合抽取策略模板示例

4　知识库驱动的智能系统主要功能与技术方案

4.1　中国项目库的系统框架

中国项目库数据分类与 DOI 门户下的“中国项目库（China S&T Project Database）”的系统框架如图 3 所示:

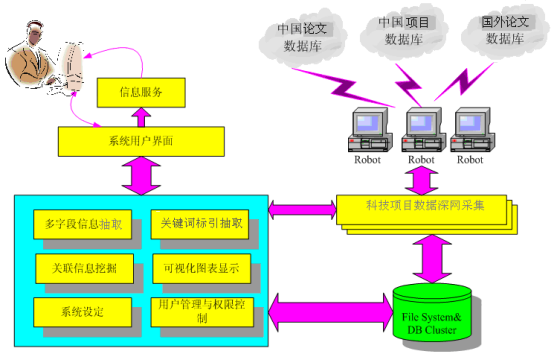

图 3　中国项目库（China S&T Proj.DB）的系统框架

4.2　中国项目库系统的主要功能

如图 3 所示，中国项目库的系统功能包括主要包括如下三大部分：

1）科技项目多源探测与采集模块（Facilitate discovery & define fitness for use，discovery metadata），主要根据元数据模型和知识驱动模块的引导。

2）科技项目采集和资源整合元数据模型和知识驱动模块，主要包括项目多字段信息抽取、项目关键字标引与抽取、项目关联信息搜索与挖掘、知识库模板，策略，规则可视化图表显示以及系统设置与权限控制等。

3）信息服务模块。

4.3　中国项目库系统的关键实现技术

1）页面大数据的并发访问

将网站首页及信息列表页以及链接进行下载并提交策略采集模板处理，进而分析提取出下一个页面的关联信息串，集成脚本分析引擎，解析动态脚本函数，进行页面大数据的合理并发访问，以提高网络信息连接效率。

2）关联动态采集规则

将采集频率可通过参数进行限定，并可通过对方提供的数据库服务接口、代理服务器接口或支持 ADSL 重新拨号进行点对点的专属或私人数据 API 接口的定期或不定期动态访问和数据采集。个别页面的需要多次关联到不同地点和不同项目组网址进行重复采集，因此为了对重复信息进行剔除和快速识别，我们设定了指纹识别策略集。

3）信息质量控制 （ISO-19115 子集）

项目数据页面相互链接，很多系统的标准不统一，缺少元数据等，形成采集陷阱，必须通过限制和调整采集数量和深度，防止重复和过度采集给对方造成系统负担，以及给本项目系统造成垃圾数据，因此根据指纹策略集，建立了并行策略信息的补采与去重，以保证信息及时全面。页面信息的格式多样化，通过页面分析提取技术，进行多字段的分析对比技术，在提取、标准化处理及批量入库方面进行控制。

4.4 项目资源数据动态采集模块详细实施步骤

中国项目库数据采集模块的流程图如图 4 所示。图中标号部分如下：

1)初始化[1]：读取 DPSpider.xml 知识模板文件，载入采集任务和采集策略 gather.id。

2）报告错误信息[2]：主要显示失败原因，包括：命令行参数不合法；无法找到 DPSpider.xml 文件或该文件格式不符合要求；读取任务文件失败；读取 gather.ini 失败等。

3)采集任务知识模板[3]：该模板存储在[task_dir]目录下，其值在命令行参数中指定。

4）信息动态采集[4]：支持采集过程中动态监控是否有新增的采集任务，若有则自动读取。

5）采集结果数据[5]：存储在本地[SavePath]路径下，包括原始项目资源的元信息文件、原始网页文件、网页正文文件等。用于在本地创建索引以及快照获取等。

6）复制回传结果数据[6]：将原始项目资源文件从存储采集结果数据的本地[SavePath]路径下复制到项目数据抽取和更新上传[upload_dir]目录。

7）更新 Client 端 State[7]：将分布式框架中的客户端（client）端的配置文件 client.ini 中的状态标志位 state 值设置为 1。则系统任务正常结束。

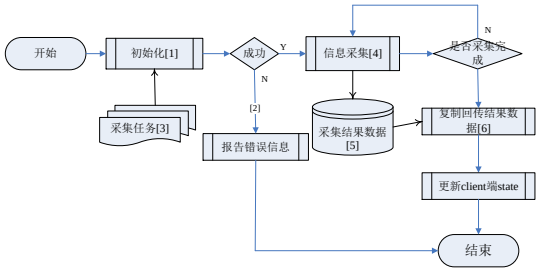

图 4 中国项目库网络数据动态采集器基本流程图

该智能采集器的具有如下特点：

- 大规模数据高效的更新

已经采集过的信息不会重复采集，新发表的网页可以在半小时以内采入本地库。

- 支持分布式部署

单个采集器能支持千万数量级的网页采集，支持分布式采集，进行无限扩展。

5 系统检测与成果

系统运行于 Windows 2003、Windows XP、Windows 7 等操作系统，Mysql5.0 数据库；Web 服务器选用 Tomcat，采用 Struts2+Hibernate+Spring 架构，普通 PC 机器服务器即可运行。一般门户建议 1G 以上内存，50G 以上硬盘空间，2M 以上带宽。目前整个科技项目库系统核心部分建设已经初步实现，并收集国内项目信息，采集的效率在 10M 网络带宽环境下，如果目标网站不限采，每小时可以采集 7 万个网页。

在北京理工大学和中信所信息共享中心相关管理专家的高度关注和指导下，本课题进展顺利，在预定的时间内我们顺利完成了系统原型的初步功能和核心知识模块的研发工作，并进行了系统测试。目前已累计采集中外相关项目信息 588,722 条，国内项目信息 350,956 条，项目成果信息 238,759 条，并具有以下特点：

1）系统具备的功能

可以配置网站来源，并实施科技项目深网采集抽取，数据导出 XML、Excle 可定制文档集等功能的系统。

2）分类信息规模

覆盖中国自然基金、863、973 以及地方科研部门等项目信息；国内项目信息 200 万多条；

至少采集中国之外的两个国家的数据，项目信息条数至少 100 万条；

科研成果信息达到千万规模；

3）抽取信息字段质量

项目相关的字段包括项目名称、计划名称、负责人、项目类别、起始时间等关键信息完整；

科研成果相关的字段包括；成果名称、作者、出处等关键信息完整；其信息抽取与质量检测如图 5 所示。

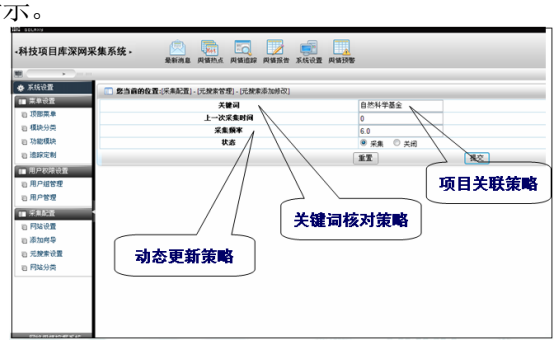

图 5 项目信息质量关联检验模板界面

经专家检测，获取的项目信息忠实反映原始网页信息，差错率不高于 1%。

6 总结

本文探讨了大数据背景下的中国科技项目数据库建设的议题，为此建立中国科技项目库及其与国际关联的项目动态分析与集成挖掘工具的意义，提出了基于科技项目知识库驱动的对中国乃至全球已有项目进行关联分析和动态跟踪的智能系统实现方案和关键技术；以此将提供全球参考的中国项目数据库服务。随着中国经济地位和创新能力的提升，本文建设的系统将为全球提供中国科技项目成果贡献的全面信息，更好地评估中国重要项目投资对中国和地区国际发达和发展中国家的影响，为中国各级决策层提供有关重要项目投资，为后期的科技成果转化和评定提供重要参考依据；为准确地把握中国和世界科技项目引领的创新需求和经济发展趋势，并将对中国和国际科技信息服务产生重要的影响。

参 考 文 献

[1] Federal R&D Project Summaries. [EB/OL]http://www.osti.gov/fedrnd/.

[2] Authoritative u.s. Government science information including research and development results. [EB/OL]Science.gov.

[3] Seventh Framework Programme (FP7). [EB/OL]. http://cordis.europa.eu/fp7/projects_en.html.

[4] KTH Research Project Database. [EB/OL]. http://researchprojects.kth.se/index.php.

[5] FORKAT. German Federal Ministry of Education and Research Project Databases [EB/OL]. http://www.stn-international.de/stndatabases/databases/forkat.html.

[6] Japan' JST Project DB.[EB/OL] http://jglobal.jst.go.jp.

[7] Canada Alberta Ecotrust projects DB. [EB/OL]. http://www.albertaecotrust.com/.

[8] The APEC Project Database. [EB/OL]. http://aimp.apec.org/PDB/default.aspx.

[9] IPSSA investment promotion in Sub-Saharan Africa. [EB/OL]. http://www.ipssa.ch/internet/osec/fr/home/ip ssa/business.html.

[10] National Natural Science Foundation of China(NSFC). [EB/OL]. http://www.nsfc.gov.cn, http://isis.nsfc.gov.cn.

[11] National Basic Research Program of China (973). [EB/OL]. http://www.973.gov.cn/English/Index.aspx.

[12] Bergman M. K. Deep Web WhitePaper [EB/OL]. 2004. http://brightplanet.com/technology/deepweb.asp.

[13] European Committee for Standardization (CEN). Metadata Standard(specification), SO-19115.

[14] Omar Boucelma. Management of Spatial Data Quality with a Service Oriented Approach[C]. COINFO'10, Beijing, Nov,23-24. 2010.

Scientific Projects Database Management Based on Dynamic Knowledge Relation

ZHAO Yanping[1]，ZHANG Huaping[2]，MA Baojun[3]，ZHAO Hui[4]

(1. School of Management and Economics, Beijing Institute of Technology, Beijing 100081；

2. School of Computer Science, Beijing Institute of Technology, Beijing 100081；

3. School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876；

4. Resource Sharing Promotion Center, Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract：With the promotion of economic status and innovation capability, China has being increased greatly in its annual investment on scientific R&D projects. China's contribution to science and technology will provide substantial influences on the world scientific achievements and environment. Based on this situation, this paper tries to establish an integrated service system for China's scientific projects and their associated international projects on big data background, with knowledge supported dynamic analysis and mining tools. This system dynamically tracks the invested projects in China and their global multiple correlated projects, which aims to assess the impact of international important investments by the developed countries on China. In the meantime, by using this system, it is possible to provide decision support information to all levels of decision-makers in China about the dynamics of China's major project investments and the future direction of economic development, as well as China's projects database services for global references.

Key words：big data management; China's S&T project database; ontology and rules; multiple relations.