doi:10.3772/j.issn.2095-915x.2017.05.007

一种用于科技项目查重的数据整合及描述模型

中国科学技术信息研究所 北京 100038

李善青

摘 要 整合科技项目所产出成果的信息能间接反映项目的研究内容，可以弥补项目查重过程中申报书难以获取的不足，具有重要的研究意义。本文提出一种整合科技项目相关产出信息的数据模型。该模型通过整合项目产出的科技报告、学术论文和科技成果等信息，抽取其中的关键词、标题和摘要等对项目的研究内容进行准确的描述，并强化了项目负责人和承担机构等辅助信息对项目查重的重要性，从而为解决项目查重问题提供客观的数据支撑。

关键词：数据整合，描述模型，科技项目查重，Hadoop 架构
中图分类号：G311
开放科学（资源服务）标识码（OSID）

A Data Model of Integration and Representation for Similar Scientific Projects Detection

Institute of Scientific and Technical Information of China, Beijing 100038, China

LI ShanQing

Abstract Information integration of research project outputs which are closely related to research contents can represent the research content of a project without the project proposal. This indirect description method is of important research value for the similar project detection. This paper proposed a data integration model

基金项目：本文受国家自然科学基金“大数据挖掘在科技项目查重中的应用研究”（编号：71303223）的资助。
作者简介：李善青（1981–），博士，助理研究员，研究方向：信息资源管理、大数据挖掘，Email：lishanqing@istic.ac.cn。

doi:10.3772/j.issn.2095-915x.2017.05.007

of research project outputs, which precisely represented the research content of a project with keywords, titles and abstracts extracted from its published reports, papers and achievements. The information of principle investigator and research organization was also introduced and applied to reinforce the similarity calculation. This model will provide data support and lay the foundation for similar project detection.

Keywords：Data integration, project representation model, similar scientific project detection, Hadoop architecture

1 引言

项目多头申报、重复立项是我国科研项目管理领域的突出问题之一。该问题不仅会造成国家科技资源的浪费和损失，也会导致恶性的科研竞争环境，损害鼓励创新的科研精神，对科技创新发展的危害极大。国务院 2014 年连续发布的公文《国务院关于改进加强中央财政科研项目和资金管理的若干意见》[1] 和《关于深化中央财政科技计划（专项、基金等）管理改革的方案》[2] 均强调了项目查重的重要性，并提出了建立公开、统一的国家科技管理平台的构想。

通过文献调研发现，国外的项目评审大多以同行评议的方式完成，未发现项目查重的提法。但国外在关键词抽取 [3]、自动摘要 [4]、文档检索 [5] 等方面的研究起步较早，进行了大量的研究和探索，积累了丰富的经验和成熟的技术。国内在方法研究方面起步较晚，但有针对性的开展了文本挖掘方法在科技项目管理中的应用研究 [6-13]。姜韶华 [6] 在对项目申报书进行分词的基础上提取层次特征项，并用向量空间模型对项目申报书进行建模。左川 [7] 提出了一种基于非分词技术解决科技项目查重问题的方法，通过 CHARM 算法挖掘得到的频繁闭项集作为特征向量，利用向量空间模型对项目申报书进行建模。方延风 [8] 利用向量空间模型描述项目申报书，引入特征词的位置和长度两种因素对 TF-IDF 方法进行改进。吴燕 [9] 通过对项目申报书加工处理，抽取关键词特征向量构建了项目本体，并建立了已有项目本体的层次聚类树以提升项目查重的效率。林明才等人 [10] 采用向量空间模型对项目申报书进行建模，提出了一种改进的模糊聚类算法 RM-FCM 对待判定项目集和已有项目集进行聚类分析，进而判断项目间的相似性。罗灏 [11] 利用分词技术和语义相似度网络从项目申报书中抽取关键词，结合向量空间模型和物元知识表示模型描述科技项目知识。林建海 [12] 对向量空间模型进行扩展，提出了一种基于内容项的项目知识表示模型，通过加权策略融合了语义相似度和字符串匹配相似度两种计算方法。赵士杰 [13] 通过一种基于语义理解的向量空间模型计算项目研究内容的相似度，利用编辑距离计算项目标题的相似度，并融合两种结果来计算项目之间的相似度。这些研究工作基本都是从项目申报书入手，对其进行降维处理，提取特征向量建立向量空间模型来表示项目申报书的内容，并利用特征向量

的相似度来度量项目申报书的相似度。这些方法存在两方面的局限性：1）项目申报书一般不对外公开，获取难度很大。因此该方法适应于单个计划内部进行项目查重，很难进行跨计划的项目查重。2）从申报书到特征向量的降维处理意味着信息的丢失，会对项目查重的准确性产生一定程度的影响。文献 [14] 总结了项目查重难以解决的三个方面的原因：1）科技项目数量迅速增长；2）项目信息公开、共享和整合程度较低；3）项目相似性判别方法单一。

针对上述问题，通过整合与项目密切相关的公开信息，建立基于大数据挖掘的查重模型，以计算机辅助的形式解决项目查重将会成为未来的趋势。随着云计算和大数据挖掘技术的不断发展和推广，尤其是国家科技报告服务系统 [15] 和国家科技成果转化项目库 [16] 等平台相继对公众提供信息服务，使得通过整合科技项目各环节的相关信息，利用大数据挖掘技术解决项目查重问题成为可能。本文将重点介绍一种以大数据挖掘为潜在应用场景，以描述科技项目的研究内容为目标，能有效整合多来源相关信息的数据模型，为解决项目查重问题奠定数据基础。

2 数据来源

本文所采用的与项目研究内容密切相关的数据主要包括项目题录、科技报告、学术论文和科技成果等。项目题录一般可从项目主管机构定期发布的立项资助公告中获取，其内容包括项目标题、负责人、承担机构、项目类型、资助金额、起始日期和结束日期等。本文将使用中国科学技术信息研究所重点项目资助建立的科技项目数据库，其涵盖了科技部支撑计划、973 计划和 863 计划，以及国家自然科学基金等项目，累计总量约为 40 万条。科技报告将主要来自国家科技报告服务系统，该系统目前已公开 6 万余篇科技报告的题录信息，包括报告题目、报告类型、报告作者、中英文摘要、中英文关键词等信息。学术论文主要来自 Web of Science、万方数据等平台收录的公开出版的论文，其题录信息是对所有用户开放的，包括论文标题、作者、机构、关键词、摘要、分类号、年卷期和基金资助等信息。科技成果是指项目所产出的新技术、新产品、新工艺、新材料、新装置及其系统等，其主要来自国家科技成果转化项目库平台。该平台目前已公开约 1.3 万条科技成果的题录信息，包括成果名称、关键词和成果简介等。注册用户可获取更详细的成果信息。本文将主要使用上述题录信息，重点利用标题、关键词和摘要等信息来间接表示项目的研究内容。

3 整合模型

针对项目申报书获取困难的现状，通过收集与项目相关的其他信息间接描述其研究内容是合理和可行的方案之一。本文给出一种整合与项目密切相关的项目题录、学术论文、科技报告和科技成果等信息的数据模型如图 1 所示。其中，科技项目表是整个模型的核心，包括项目 ID、标题、负责人 ID、机构 ID、起始日期和结束日期等字段。科技项目表通过关联表 Project-Report 建立与科技报告表之间 1 对 1

doi:10.3772/j.issn.2095-915x.2017.05.007

的对应关系；通过关联表 Project-Paper 建立与学术论文表之间的多对多的对应关系；通过关联表 Project-Achievement 建立与科技成果表之间多对多的对应关系。科技报告表、学术论文表和科技成果表分别保存了项目的不同形式产出物的题录信息，包含了反映项目研究内容的重要信息。这些表具有相似的字段结构，包括 ID、标题、关键词、摘要、作者 ID 和机构 ID 等字段。上述信息经关联整合后，通过项目 ID 可获取项目产出物的全部信息，抽取其中的标题、关键词和摘要等关键信息可实现对项目研究内容的描述和表示。

人员表和机构表主要用于追溯申报者和申报机构以前曾承担过的项目记录。这些信息将用于辅助计算项目的相似度，其遵循的基本假设为重复项目出现在同一人员所主持的项目中或同一机构所承担项目中的概率高于其他的情况。因此，将对满足上述假设的项目进行重点关注和排查。其中，人员表包括人员 ID、姓名、出生日期、性别、研究领域和所属机构等字段。通过人员 ID 建立的关联关系，可获取该人员所承担的项目信息，提交的科技报告信息，发表的论文信息和提交的成果信息。机构表包括机构 ID、机构名称、机构类型、机构研究领域、所属国家和联系方式等字段。通过机构 ID 建立的关联关系，可获取该机构所承担的项目信息，提交的科技报告信息，发表的论文信息和提交的成果信息。

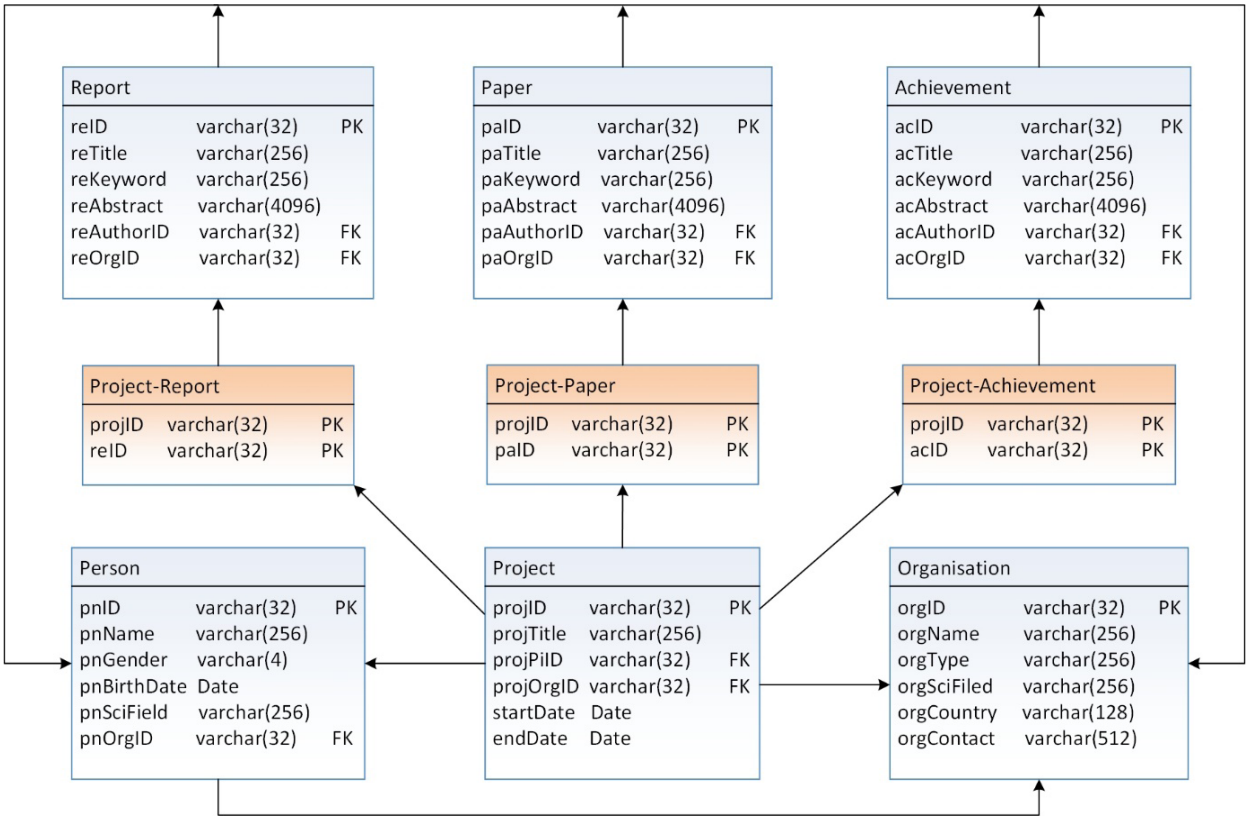

图1 项目相关信息的整合模型

4 描述模型

目前普遍采用的项目描述模型是基于项目申报书的向量空间模型，存在前面分析所指出的信息丢失和项目申报书难以获取的问题。本文尝试提出一种描述项目研究内容的数据模型如图 2 所示，不采用传统的降维处理，而是利用项目的产出物所包含的关键词、标题和摘要的内容对其研究内容进行描述。由于不同形式的产出物与项目的相关程度存在差异，如科技报告是对项目研究过程和研究内容的总结，因此应该具有最高的相关性。为体现上述差异，将为不同形式的产出物即信息来源配置不同的权重以区分其重要程度。

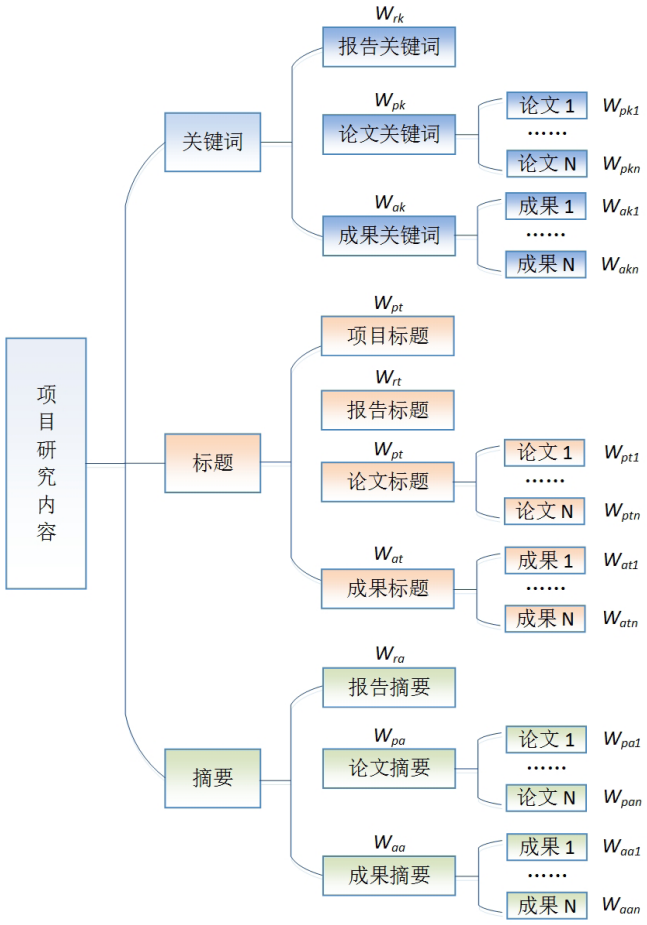

图2 项目研究内容的描述模型

项目研究内容的描述模型可表示为公式（1）：

$$C_{pj}=\{(K_r, K_p, K_a), (T_j, T_r, T_p, T_a), (A_r, A_p, A_a)\} \quad (1)$$

其中，(K_r, K_p, K_a) 是所有关键词的集合，$K_r=\{k_r, w_{rk}\}$ 表示报告的关键词及其权重的集合，$K_p=\{(k_{pki}, w_{pki})|i=1,2,\ldots N\}$ 表示论文的关键词及权重的集合，N 为该项目所产出学术论文的总数量，$K_a=\{(k_{aki}, w_{aki})|i=1,2,\ldots N\}$ 表示成果关键词及其权重的集合。

(T_j, T_r, T_p, T_a) 是所有标题的集合，包括项目

doi:10.3772/j.issn.2095-915x.2017.05.007

的标题及其权重的集合 $T_j=\{t_j, w_{jt}\}$，报告的标题及其权重的集合 $T_r=\{t_r, w_{rt}\}$，论文的标题及其权重的集合 $T_p=\{(t_{pti}, w_{pti})|i=1,2,\ldots N\}$，成果的标题及其权重的集合 $T_a=\{(t_{ati}, w_{ati})|i=1,2,\ldots N\}$。

(A_r, A_p, A_a) 是所有摘要的集合，其中，$A_r=\{a_r, w_{ra}\}$ 表示报告的摘要及其权重的集合，$A_p=\{(a_{pai}, w_{pai})|i=1,2,\ldots N\}$ 表示论文的摘要及其权重的集合，$A_a=\{(a_{aai}, w_{aai})|i=1,2,\ldots N\}$ 表示成果的摘要及其权重的集合。

描述模型不同要素的权重反映了该要素与项目研究内容的相关性，其配置方式将会影响模型描述的准确程度。本文将采用层次分析法确定不同要素的权重，由于论文数量较多，且难以区分不同论文对项目研究内容的重要程度，因此在算法实现时将其设置为相同的权重以简化其配置过程。科技成果存在相同问题，也将采用上述权重配置方法。

项目的相似度将考虑四个方面的因素计算得到，即分别由关键词集合、标题集合和摘要集合计算得到的相似度，以及由相关辅助信息确定的影响因子，其计算公式表示为公式（2）：

$$S(F, P)=S(F, K)+S(F, T)+S(F, A)+\Delta_{p+o} \quad (2)$$

其中 $S(\cdot)$ 表示相似度函数，F 为输入的用于描述项目特征的检索词的集合，P 为待判定的项目，K、T、A 分别表示关键词、标题和摘要的集合，Δ_{p+o} 表示申报者和申报机构等辅助信息对项目查重的影响因子。

根据关键词集合计算得到的相似度可表示为 $S(F, K)=S(F, K_r)*w_{rk}+\sum_{i=1}^{N}S(F, K_{pi})*w_{pki}+\sum_{i=1}^{N}S(F, K_{ai})*w_{aki}$。根据标题集合计算的相似度可表示为 $S(F, T)=S(F, T_j)*w_{jt}+S(F, T_r)*w_{rt}+\sum_{i=1}^{N}S(F, T_{pi})*w_{pti}+\sum_{i=1}^{N}S(F, T_{ai})*w_{ati}$。由摘要集合计算的相似度表示为 $S(F, A)=S(F, A_r)*w_{ra}+\sum_{i=1}^{N}S(F, A_{pi})*w_{pai}+\sum_{i=1}^{N}S(F, A_{ai})*w_{aai}$。

基于重复项目出现的一般规律，我们假设重复项目出现在同一人员所主持的项目中或同一机构所承担项目中的概率高于其他的情况。为体现上述因素的影响，将引入因子 $\Delta_{p+o}=\Delta_{per}+\Delta_{org}$ 提升满足假设条件项目的相似度，其中 Δ_{per} 为同一人员所产生的影响，Δ_{org} 表示同一机构所产生的影响。

经上述计算过程后，可得到检索词集合与全部已有项目的相似度，选取超出阈值的项目作为重复项目的候选，最后由专家小组审核后判定该项目是否为重复。需要指出的是，重复项目的判定是一个复杂的过程，需要综合的背景知识和较高的判断力，因此计算产出的候选集合仅提供可疑项目的清单和客观的事实依据，最终是否为重复项目的判定将由专家小组作出。

5 讨论

本文从信息整合的角度提出了一种用于整合项目产出物信息和描述项目研究内容的数据模型，解决了项目查重所需的数据标识、描述和整合机制问题。该模型的应用场景将是面向大数据背景下的项目查重，将对海量的信息进行采集和加工，因此需要制定一系列的标准和规范来保证数据加工的准确性，并建立严格的工作流程实现数据的处理。

该数据模型采用关键词、标题和摘要等原始信息建模和描述项目的研究内容，未采用信息降维等处理方法。其优点是在最大程度上保证了信息的完整性，提升了项目描述的准确性，

但缺点是大幅增加了计算的复杂度。为解决上述问题，我们后续拟研究和利用大数据挖掘的思想来构建分布式的项目查重系统，以提高其查重计算的速度。其中，Hadoop 是一种开源且相对成熟的技术，其基本原理为“分而治之”的思想，分别利用 Map 和 Reduce 操作对业务逻辑进行拆分和对结果进行归纳，从而实现快速的分布式计算。如何针对 Hadoop 技术框架的特点建立相应的项目查重的业务逻辑将是未来需解决的关键问题。

6 结束语

本文提出了一种用于项目查重的数据模型，通过整合与项目相关的科技报告、学术论文和科技成果等信息，抽取其中的关键词、标题和摘要信息对项目的研究内容进行间接的描述。该模型提供了一种描述项目研究内容的新思路和新方法，在一定程度上解决了项目申报书难以获取的问题，可在更大的范围内实施项目查重，因此具有重要的研究和应用价值。

参考文献

[1] 《国务院关于改进加强中央财政科研项目和资金管理的若干意见》 (国发[2014]11号)[EB/OL]. [2016-10-11]. http://www.gov.cn/zhengce/content/2014-03/12/content_8711.htm.

[2] 《关于深化中央财政科技计划（专项、基金等）管理改革的方案》 (国发[2014]64号)[EB/OL]. [2016-10-17]. http://www.most.gov.cn/ztzl/shzyczkjjhglgg/wjfb/201501/t20150107_117294.htm.

[3] Siddiqi S, Sharan A. Keyword and Keyphrase Extraction Techniques: A Literature Review[J]. International Journal of Computer Applications, 2015, 109(2): 18-23.

[4] Gambhir M, Gupta V. Recent Automatic Text Summarization Techniques: A Survey[J]. Artificial Intelligence Review, 2016: 1-66.

[5] Jin Y K, Croft W B. A Field Relevance Model for Structured Document Retrieval[C]// European Conference on Advances in Information Retrieval. 2012: 97-108.

[6] 姜韶华. 科研项目管理中文本挖掘方法研究及应用[D]. 大连: 大连理工大学, 2006.

[7] 左川. 基于非分词技术的科技项目查重研究与实现[D]. 重庆: 重庆大学, 2010.

[8] 方延风. 科技项目查重中特征词TF-IDF值计算方法的改进[J]. 情报探索, 2012(1): 1-3.

[9] 吴燕. 基于层次聚类的科技项目分类与查重研究[D]. 天津: 天津财经大学, 2008.

[10] 林明才, 康耀红, 张诚一. 基于科研立项管理应用的模糊C均值算法研究[J]. 计算机工程与设计, 2010, 31(7): 1570-1572.

[11] 罗灏. 基于语义的科技项目相似度计算研究[D]. 杭州: 杭州电子科技大学, 2012.

[12] 林建海. 相似度计算在科技项目管理系统中的研究及应用[D]. 杭州: 杭州电子科技大学, 2013.

[13] 赵士杰. 面向科技项目的相似度计算和聚类算法研究[D]. 杭州: 杭州电子科技大学, 2015.

[14] 李善青, 赵辉, 宋立荣. 基于大数据挖掘的科技项目查重模型研究[J]. 图书馆论坛, 2014, 34(2): 78-83.

[15] 国家科技报告服务系统 [DB/OL]. [2017-03-15]. http://www.nstrs.cn/.

[16] 国家科技成果转化项目库 [DB/OL]. [2017-04-15]. http://www.nstad.cn/.