文章编号:1671-4598(2020)02-0202-04 DOI:10.16526/j.cnki.11—4762/tp.2020.02.042 中图分类号:TP391.3 文献标识码:A

智能科技查新系统的设计与实现

黄孝伦,王 东,谭 涛,刘 芹
(重庆市卫生信息中心,重庆 401120)

摘要:为提高科技查新的效率,利用信息化技术对查新业务流程进行优化重构,在常规查新系统的基础上构建了一个智能的科技查新系统;系统设计时首先利用网络爬虫技术自动按照科技项目申请书的关键词搜索和下载相关文献资源,然后以自适应分配算法分配查新任务和遴选查新报告审核专家,最后以系统中累计的以往查新报告和文献资料为基础,利用 Lucene 检索工具对生成的查新报告进行全文检索;一是实现了文献资源检索工作的自动化,保证了检索途径、范围及检索表达式的全面性和准确性,避免了大量的人工检索,提高了文献检索效率;二是实现了任务分配的智能化,均衡分配相关任务,使查新效率最大化;三是实现与既往研究项目进行精确比对,避免了科技项目的重复申报,进一步提高查新质量。

关键词:科技查新;网络爬虫;全文检索;自适应分配模型

Construction of Intelligent Novelty Search System

Huang Xiaolun, Wang Dong, Tan Tao, Liu Qin
(Chongqing City Sanitation Information Center, Chongqing 401120, China)

Abstract: In order to improve the efficiency of sci-tech novelty retrieval, an intelligent sci-tech novelty retrieval system is constructed on the basis of conventional novelty retrieval system by using information technology to optimize and reconstruct the business process of sci-tech novelty retrieval. Firstly, the network crawler technology was used to search and download the relevant literature resources automatically according to the keywords of the application for scientific and technological projects. Secondly, the self-adaptive allocation algorithm was used to assign the task of novelty search and select the experts to audit the novelty search reports. Finally, based on the accumulated previous novelty search reports and literature in the system, the full-text search of the generated novelty search reports was carried out by using Lucene search tool. First, it realizes the automation of literature resources retrieval, guarantees the comprehensiveness and accuracy of retrieval approach, scope and expression, avoids a large number of manual searches and improves the efficiency of literature retrieval. Second, it realizes the intelligence of task allocation, balances the distribution of related tasks, and maximizes the efficiency of novelty retrieval. Third, it realizes accurate comparison with previous research projects, avoids duplicate declaration of scientific and technological projects, and further improves the quality of novelty search.

Keywords: Sci-tech novelty search; internet crawler; full-text retrieval; adaptive allocation model

0 引言

科技查新是一项复杂的脑力智慧型劳动,其主要的工作是检索相关文献并从中分析出与委托人研究点的异同,为科研立题、成果评审等科技活动的新颖性评价提供科学依据[1-2]。一份高质量查新报告的工作周期一般为 3~5 日[3]。为了提高查新人员的工作效率,减轻其工作负担,各种查新系统的构建与应用也越来越多。重庆市卫生信息中心承担了重庆市卫健委科研项目立项查新工作,每年受理科技查新 500~700 项。该中心利用网络爬虫技术、全文检索工具和智能算法对查新业务流程进行优化,构建了智能科技查询系统,进一步提高查新效率和质量。

收稿日期:2019-07-12; **修回日期**:2019-08-27。
基金项目:重庆市科委决策咨询项目(cstc2016jccx BX0067)。
作者简介:黄孝伦(1977-),男,湖北仙桃人,硕士,中级,主要从事智能计算方向的研究。
通讯作者:王 东(1979-),男,四川南充人,博士,中级,主要从事医学学术期刊市场化运营及其他相关学术方向的研究。

1 查新系统的现状

目前,国内查新机构基本上通过信息技术对查新工作流程进行了重构,实现了在线接收委托申请、审查申请委托、分派查新任务、查新、审核、提交查新报告等功能,建立了不同的查新管理平台[4-6]。而且,许多学者还在不断地对查新系统进行设计上的优化研究,如:宋正阳等[7]提出了一种基于指标权重叠加的自适应分配模型的农业科技文献查新系统,能够一定程度地缩短任务分配时间,加速查新流程;温慧明等[8]基于 Solr 搜索应用服务器构建的科研查新系统实现了检索查新和对比查看;王华等[9]在科技查新平台模型中新增了质量控制等功能。这些系统基于不同的信息化技术在不同程度上重新构造了查新业务,规范了查新流程,提高了查新工作效率,同时也为查新业务的数据资源整合及智能化分析创造了基础。但是,目前查新系统主要以查新课题、用户管理、查新任务分配、查新课题委托等为主,如何根据课题信息自动搜索相关领域的基本知识概念,如何根据课题领域推荐相关检索资源等相关报道较少见。

2　智能科技查询系统设计

2.1　查新系统功能模块设计

2018 年，重庆市卫生信息中心研发了一套智能科技查询系统并发布于“重庆医生”平台。该系统整体包含 4 大模块：用户模块、查新员模块、审核员模块和系统模块（见图 1）。1）用户模块：用户可直接在线提交查新申请、跟踪查新处理进程、反馈对查新结果的意见；2）查新员模块：查新员负责查新申请书的受理、合同的签订、查新报告的撰写，其中文献检索功能采用网络爬虫技术实现自动检索，全文检索功能采用 Lucene 检索工具实现对查新申请书的重复性检测，而且具有对比查看功能，可展示查新申请书和数据库查新报告及文献资料的重复内容；3）审核员模块：审核员负责查新报告的审核；4）系统模块：实现对用户、查新员、审核员的管理；采用智能算法实现查新任务自动分配；实现查新进度的跟踪、提醒；用户与查新员、查新员与审核员之间可通过公共信息交流平台在线交流。

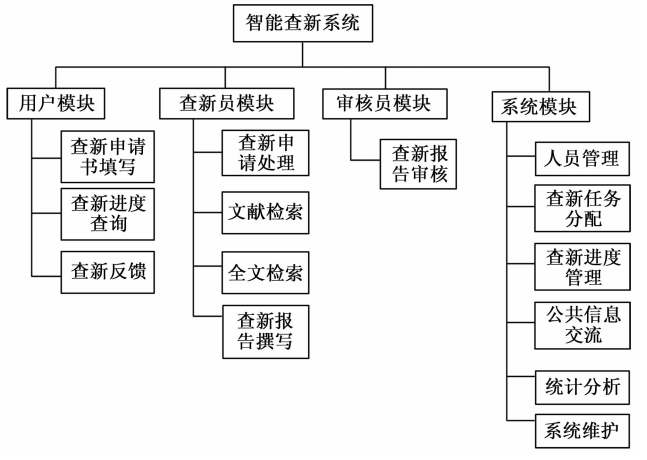

图 1　查新系统框图

2.2　查新系统架构设计

智能科技查询系统基于“重庆医生”平台在技术上实现查新各阶段的业务整合，实现文献资料的自动采集归类，实现查新服务共同体的数字化、自动化、智能化和交互性运营。整个系统的总体框架规划如图 2 所示。1）注册服务。注册服务用于各种共享服务资源的注册，通过服务资源的发布—发现—访问机制，实现服务资源共享。注册服务包括对个人、查新人员、医疗卫生术语的注册管理服务。系统针对各类实体形成各类注册库（如个人注册库、知识库等）；2）查新存储服务。查新存储服务包括一系列存储库，用于存储查询申请书、查新报告、文献资料信息，形成查新数据中心。查新存储服务除了提供查新的访问服务，也负责按查询申请书相关性采集文献资料，使其成为查新报告的基础资料；3）查新服务。查新服务用于处理系统内与数据定位和管理相关的复杂任务，是系统架构的核心组件。该服务负责实现各功能模块的互联互通，利用“重庆医生”平台内提供的组件和服务进行文献资料的采集、知识库的构建等。查新服务主要包括索引服务、业务服务、数据服务、事务处理等组件；4）信息交换层。信息交换层主要包括支持系统平台上服务与其它应用系统平台之间低级别通信的通信总线服务，还可提供可在整个平台中重复使用的通用软件功能的公共服务。

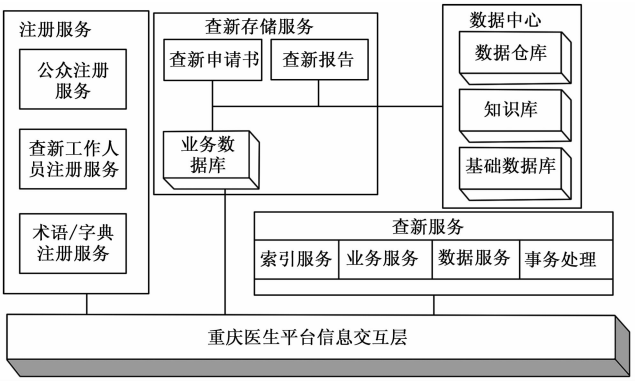

图 2　系统架构图

2.3　查新流程设计

用户在 PC 端或手机中填写、提交、查询查新申请书。系统根据查新申请书自动检索系统中数据仓库和 CNKI、万方、维普、PubMed、ScienceDirect 等中外文数据库；根据查新申请书自动更新检索知识库；每年定期更新系统中数据仓库的文献资料；根据查新要求自动分配查新任务。查新人员根据检索结果形成查新报告，并由“重庆医生”平台专家库中的专家进行审核。具体流程图见图 3。

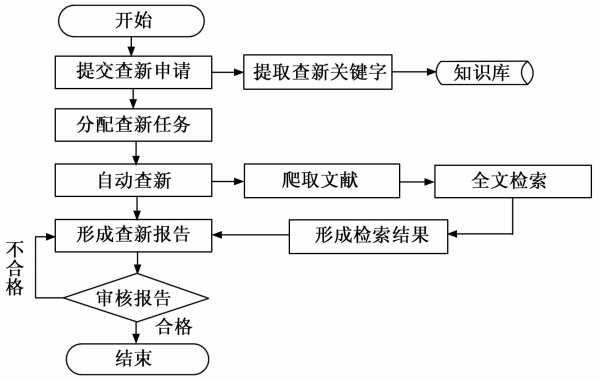

图 3　系统流程图

3　关键技术

3.1　利用网络爬虫技术实现文献资料自动抓取

查新工作是以文献检索为基础，面对海量的文献资料，文献检索工作已经成为了查新人员的一项繁重任务，以医学类课题查新为例，每次至少检索 CBM、CNKI、万方、维普等多个数据库，检索工作量较大，而且存在漏检可能。同时，为了获得所需的文献资源，查新机构每年都需支付一定费用给电子学术资源服务提供商，而每项查新工作都可能重复下载相关文献资料，因此存在重复购买相同学术资源造成的浪费。

智能科技查询系统利用网络爬虫技术实现了文献资源的搜索和下载，该技术主要采用 WebCollector＋selenium＋phantomjs 技术实现（图 4）。1）利用基于词或词组长度和频数的关键词提取算法[10]在查新申请表中自动提取基本关键词，以此为基础结合查新点确定最终关键词，然后利用数据库模拟登录及网络爬虫技术抓取需要的文献资料[11]，其中关键词按“AND”“OR”“NOT”自动组合，并能根据各数据库规则转化检索式形式；2）为了保证网络中文献资料抓取的效率、覆盖率和准确率，智能科技查询系统采用了向量空间模型的概念对网页内容和主题的相关度进行评估[12-13]，根据相关程度进行抓取，且下载的文献资料按检索关键词、检索源地址、文献资料信息、相关引用指标等保存到数据库中，为后期查新工作提供数据依据；3）为了便于查新人员进行分析，智能科技查询系统对检索得到的文献资料按内容相似度、影响因子等指标进行分类归集[2,14]，做到对比文献时有针对性、可比性和准确性，方便对“查新点”进行新颖性判断，同时也为自动生存查新报告提供了内容基础；4）通过建立自己的检索知识库（包含同义词、缩写词、同义名、学名、通用名）实现智能检索，如在“ATP6i—miRNA 抑制剂对大鼠骨质疏松的防治”查新申请表中，用户提供的“ATP6i”这一关键词为“TCIRG1”别名，通常称之为“空泡型质子泵”，系统在以“ATP6i”进行检索时，自动检出含“TCIRG1”和“空泡型质子泵”的所有文献，有效避免漏检。同时，检索知识库通过内置的词典可以识别用户自造词（如自创的药品名：开郁胶囊），避免出现检索结果为零或数量极少的现象。

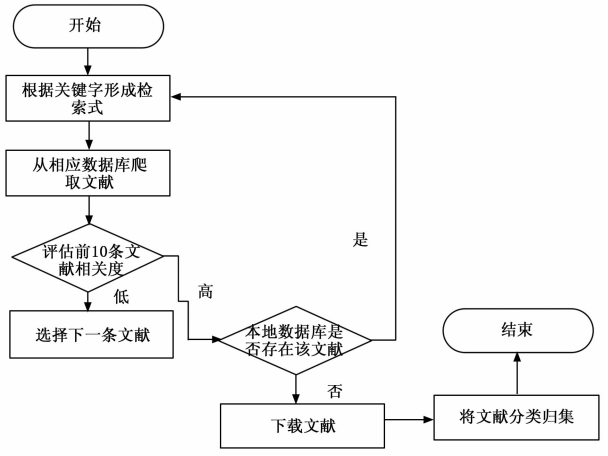

图 4 爬取流程

这些功能不仅有效提高了工作效率，而且在一定程度上弥补了查新人员、查询审核专家自身信息获取的有限性，为他们提供了充分的研究背景信息，降低了对查询申请书的原创性、新颖性、影响力等方面的评价难度。

3.2 利用智能算法优化查新任务自动分配

由于每一年的科研、项目申报、结项等在时间上比较集中，这导致查新机构经常会面临查新申请“井喷”的问题，采用按照内置规则自动分配给查新任务的任务分配系统是解决这个问题的有效办法。但是，在进行查新任务自动分配时，不同系统考虑的影响因素不同，在效率提升方面还存在着值得商榷的地方。

智能科技查询系统采用了采用自适应分配模型[14-15]，按照客户指标（项目类别、查新点数、查新时间要求、履行查新合同力度、学术水平）、查新员指标（查新熟练度、知识面广度、单位时间可接受最大任务量、已分配任务量、客户反馈情况）和查新审核专家指标（学术方向、学术水平、单位时间可接受最大任务量、已分配任务量）计算综合权重后分配查新任务和查新审核任务。模型假设查新员共有 n 位，某一时间段内系统提交了一份查新申请，则系统会根据式（1）计算出对应的结合权重 W_i，然后分配任务。

$$W_i = \sum_{i=0}^{n} A_m B_j C_k \tag{1}$$

式（1）中，$i\in\{1, 2, \cdots, n\}$；$m\in\{1, 2, 3, 4, 5\}$；$j\in\{1, 2, 3, 4, 5\}$；$k\in\{1, 2, 3, 4\}$；A，B，C 分别表示客户指标、查新员指标和查新审核专家指标的不同层级的指标的权重。其中，所有指标权重值的设定规则为：有利于加快查新项目进入查新流程并提高查新工作效率的指标权重值大，相反则小。权重取值范围为（0，1]，且属于同一级别的指标权重之和为 1。

该算法建立了客户关系库，通过网络爬虫抓取客户以往的学术成果，评价客户学术水平，作为分配具有相应学术水平的查新审核专家的参考依据，同时采用相同方法对审核专家学术方向和水平进行跟踪维护（客户查新申请中关键词与审核专家学术方向的相关性越高，分配的可能性越大），避免专家资源的不合理使用。利用智能算法分配查新任务不但加速了查新流程，提高了客户满意度，同时避免了因数量骤升而导致的查新报告质量下降的问题。

3.3 采用全文检索工具避免重复申报

据统计，我国科研项目重复率高达 40%[17]。为了降低查新申请书的重复率，智能科技查询系统以数据库中的累计的查新报告和文献资料为基础，利用 Lucene 检索工具包对其进行全文检索，若相关性得分高于 30%则根据查重率进行排序展示[18-20]，为判断查新申请书的原创性提供依据。同时，通过检索知识库引入查询单词的近义词或别称，缩小了检索盲区。

采用全文检索工具能够较为准确地计算查新申请书的查重率，可以通过对比查看功能来实现对重复内容的详细对比，有效地杜绝了项目的重复申报，满足了项目流程管理的需求。

4 使用效果与改进策略

目前，智能科技查新系统已试运营，尝试完成 2019 年科技查新项目 700 多项。针对用户而言，很好地解决了原有手工填表时的表格填写混乱的问题，解决了查新点提炼罗列过多、内容宽泛、不具体等问题。在查新过程中，系统自动围绕“查新点”，选择数据库，拟定检索词，制定检

索式，检索并筛选文献，做文献分析对比，解决了查新报告质量高低受到查新人员学科背景、知识结构、沟通能力和理解能力的影响问题[21-22]。截至目前，该系统数据仓库爬取相关文献 10040 篇，其中学位论文 6600 篇，期刊论文 2420 篇，其他 1020；中文文献 8534 篇，外文文献 1506 篇。文献库的建立节约了检索时间，提高了检索效率和检索质量，用户从提交查新申请到收到查新报告，不超过 2 个工作日，查新部门工作在 1 周左右的时间完成超过去年同比 41.6%的查新工作量。而且查新报告的质量也受到了用户和审核专家的好评。

考虑到查新工作不仅仅是向用户提交一份查新报告，更重要的是帮助用户提炼查新点，提供相关研究热点。因此，后期计划将在在把握项目内容的基础上，启发或帮助用户对查新点进行重新提炼、修正，使其既符合查新规范的要求，又能充分体现查新项目的新颖性。同时对查新不达标的用户，把相关研究特点的文献及建议推送给用户，帮助用户提升科研水平。

5　结束语

随着信息技术的发展，查新系统会越来越智能化。实践证明，智能科技查询系统系统的应用使查新工作模式发生了根本性变化，从技术上实现了检索工作的自动化，保证了检索途径、范围及检索表达式的全面性和准确性，同时实现了任务分配的智能化，避免了重复申报，保障了查新报告质量的提升。

参考文献：

[1] 杨坤杰，亢　力，李　霞，等. 基于“中医药科技查新系统”的 2003—2012 年中国中医药文献检索中心查新项目的统计分析 [J]. 中医药导报，2015，21 (2)：4 - 8.

[2] 李　为. 科技查新工作创新发展存在的主要问题与对策 [J]. 中国高新区，2018，18 (4)：34 - 37.

[3] 马林山，郭　磊. 基于主题模型（LDA）的查新辅助分析系统设计研究 [J]. 现代情报，2018，38 (2)：80 - 83.

[4] 吴　超，赵明华，祝恋智，等. 管道科技查新平台的开发与实现 [J]. 情报探索，2017，31 (11)：65 - 69.

[5] 赵　宁. G 大学图书馆科技查新流程优化研究 [D]. 哈尔滨：学哈尔滨工业大学，2016：82 - 84.

[6] 刘迎春，杨雪萍. 基于文献计量的科技查新系统调查分析 [J]. 情报探索，2013，27 (1)：76 - 78.

[7] 宋正阳，胡玉清. 基于自适应分配模型的查新系统 [J]. 计算机应用，2013，33 (S1)：104 - 106.

[8] 温慧明，宫晓辉. 基于 Solr 的科技成果查新系统的构建研究 [J]. 计算机技术与发展，2014，24 (6)：67 - 70.

[9] 王　华，梅江林. 论科技查新平台模型的构建 [J]. 图书情报导刊，2017，2 (2)：53 - 57.

[10] 陈伟鹤，刘　云. 基于词或词组长度和频数的短中文文本关键词提取算法 [J]. 计算机科学，2016，43 (12)：61 - 63.

[11] 杨　洋，李晓风，赵　赫，等. 基于网络爬虫的文献检索系统的研究和实现 [J]. 计算机技术与发展，2014，24 (11)：35 - 38.

[12] 李凤侠，战玉华，赵军平. 清华大学科技查新系统的开发与实践 [J]. 大学图书馆学报，2014，32 (2)：33 - 38.

[13] 代　宽，赵　辉，韩　冬，等. 基于向量空间模型的中文网页主题特征项抽取 [J]. 吉林大学学报（信息科学版），2014，32 (1)：88 - 94.

[14] 江哲雅. 聚类挖掘在电信客户分类中的研究与应用 [D]. 上海：上海交通大学，2013：40 - 45.

[15] 许川佩，陈征南，任智新. 基于云自适应遗传算法的 NoC 路径分配研究 [J]. 计算机测量与控制，2012，20 (9)：2516 - 2519.

[16] 王　解，郭晓松. 基于简化模型与自适应滤波的车载 SINS 基座快速对准 [J]. 计算机测量与控制，2017，25 (7)：261 - 263.

[17] 党彦龙. 科技查新工作效率及服务质量的提高 [J]. 河南科技，2010，35 (3)：11 - 13.

[18] 焦　洋，王　纯，韩静茹. 基于 Lucene 的科研查新系统构建 [J]. 计算机技术与发展，2018，28 (5)：193 - 196.

[19] 张　俊，李鲁群，周　熔. 基于 Lucene 的搜索引擎的研究与应用 [J]. 计算机技术与发展，2013，25 (6)：230 - 232.

[20] 邹敏昊. 基于 Lucene 的 HBase 全文检索功能的设计与实现 [D]. 南京：南京大学，2013：37 - 40.

[21] 牟韶彬，马　磊，魏　晓. 科技查新中查新点提炼的实例分析 [J]. 江苏科技信息，2017，34 (27)：11 - 13.

[22] 朱玉奴，缪家鼎，张冬梅，等. 论科技查新中的科学技术要点和查新点 [J]. 图书馆理论与实践，2014，36 (6)：16 - 18.

（上接第 201 页）

[14] Wu Z，Huang N E. Ensemble EmpiricalMode Decomposition：A Noise—assisted Data Analysis Method [J]. Advances in Adaptive Data Analysis，2009，1 (1).

[15] 王科俊，王克成. 神经网络建模预报与控制 [M]. 黑龙江：哈尔滨工程大学出版社，1996.

[16] Mirjalili S，Lewis A. The whale optimization algorithm [J]. Advances in Engineering Software，2016，95：51 - 67.

[17] Tizhoosh H R. Oppositionbased learning：a new scheme for machine intelligence [A]. Computational Intelligence for Modelling，Control and Automation，2005 International Conference on Intelligent Agents，Web Technologies and Internet Commerce [C]. IEEE，2005，1：695 - 701.

[18] Rahnamayan S，Tizhoosh H R，Salama M M A. Opposition—based differential evolution [J]. IEEE Transactions on Evolutionary Computation，2008，12 (1)：64 - 79.

[19] Xiao C F F. Development of prediction models for next day building energy consumption and peak power demand using data mining techniques [J]. Applied Energy，2014，127 (6)：1 - 10.