

# 基于 Lucene 的科研查新系统构建

焦 洋,王 纯,韩静茹

(中国民航科学技术研究院 北京市民航安全分析及预防工程技术中心,北京 100028)

**摘 要:**利用 Java Web 平台,以 Lucene 检索工具包为检索核心,设计实现了一套科研管理系统。调用 Lucene 分词索引功能,检索已有申报课题库,初步过滤出相关性计算得分高于 30% 的文档;通过构建领域术语库,引入查询单词的近义词或别称,一定程度上缩小了相近词文本替换的检索盲区;提出了利用 Lucene 高亮显示二次处理的方法,设置阈值为 0.7,将比例超过阈值的分散高亮单词进行平滑处理,使得对比查阅文档重复内容更加自然。在此基础上,利用高亮处理后的文本结果,重新设计文档重复率的计算算法。系统在企事业单位进行部署应用,根据系统现场运行反馈,重复率计算较准确,科研查新系统避免了项目重复申报,提高了科研经费的使用效益,为提升企事业单位科研管理水平提供了技术支撑。

**关键词:** Lucene; 查重; 领域本体; 科研

中图分类号: TP302.1

文献标识码: A

文章编号: 1673-629X(2018)05-0193-04

doi: 10.3969/j.issn.1673-629X.2018.05.043

## Construction of Scientific Research Management System Based on Lucene

JIAO Yang, WANG Chun, HAN Jing-ru

(Civil Aviation Safety Analysis and Prevention Engineering Technology Center,  
China Academy of Civil Aviation Science and Technology, Beijing 100028, China)

**Abstract:** With Java Web platform and Lucene as the retrieval kernel, we design and implement a set of scientific research management system. By calling the Lucene segmentation and index function, the existing research library is retrieved in order to filter out relevant documents that score above 30%. By constructing domain terminology libraries, synonyms or nicknames for query words are introduced. This method narrows the search blind area to some degree. We propose a secondary treatment method of Lucene highlight which sets the threshold to 0.7. If the proportion of the highlight words more than the threshold, these words must be processed for smooth, which makes it natural to read the compared document. On this basis, the algorithm of document repetition rate is redesigned by using the results of highlighted text. The system is applied in many enterprises and institutions and its repetition rate calculation is roughly accurate according to the feedback, which effectively eliminates the repeated declaration project and improves the efficiency of scientific research funds, providing technical support for improving the level of scientific research work of enterprises and institutions.

**Key words:** Lucene; duplicate checking; domain ontology; scientific research

## 0 引 言

在如今科技引领发展的时代,企业将科技创新作为工作重点。以往企事业单位申报科研课题基本通过线下纸质介质申报,管理部门人工审阅批示。但随着时间的推移,大量历史申报课题挤压,纸质文档存储的缺点逐渐暴露。第一,文档数量庞大,不利于保存,需要专门的存储空间和人员管理。第二,检索效率低。随着数量的增加,查找历史文档越来越困难。为此,利用 Web 技术开展无纸化办公成为企业信息化建设的

一项工作。目前,大部分煤矿企业拥有自己的 OA 办公系统,能够实现公文的流程。但是在科研管理工作中很少有企业开展信息化建设,往往只是利用 excel 工具进行简单的登记记录,即使有系统,功能也非常简单,仅仅实现了申报数据记录和查询的功能,每年新申请的课题无法与往年课题形成对比,以至于一些课题被重复申报,浪费科研经费。

为此,利用 Java Web 技术,以企业科研管理业务为背景,设计研发了科研管理系统。该系统嵌入 Lu-

收稿日期: 2017-05-24

修回日期: 2017-09-27

网络出版时间: 2018-02-07

基金项目: 国家自然科学基金(U1533102); 民航安全能力建设项目(DFS20150125)

作者简介: 焦 洋(1982-),男,硕士,工程师,从事航空安全方面的研究。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20180207.1525.020.html>

cene 检索工具包,实现对申报项目的检索查新,将重复率高于 30% 的已有文档进行排列显示,供科研管理部门参考,从而完善科研项目申报的鉴定方法,以提高申报项目的领域水平,促进科研工作有效开展。

### 1 Lucene 简介

Lucene<sup>[1]</sup>是 Apache 软件基金会 Jakarta 项目组的一个子项目,是一个纯 Java 编写的开放源代码的全文检索工具<sup>[2]</sup>。近年来,学者对基于 Lucene 全文检索的应用研究层出不穷,如行业应用、图像视频检索、全文检索技术研究以及 Web 应用等<sup>[3-6]</sup>。也出现了很多基于 Lucene 的开源软件,例如 Solr 全文搜索服务器,是一款非常优秀的全文搜索引擎<sup>[7-8]</sup>。Lucene 实际上是为检索系统提供的工具包,其主要功能是实现了全文检索。全文检索指先建立索引,后对索引进行搜索的过程,主要分为两个过程:索引创建和搜索索引<sup>[9-10]</sup>。

索引创建主要分四步:读取文本、文本分词、词元处理、创建索引。

- (1) 读取文本:获取需要创建索引的源数据;
- (2) 文本分词:将一整段文本,分为一个个 Token 及词元;
- (3) 词元处理:将词元转换为 Term 单词,例如将英语单词复数转变为单数;
- (4) 创建索引:将处理后所得的 Term 单词,出重后建立索引。

索引结构如图 1 所示。

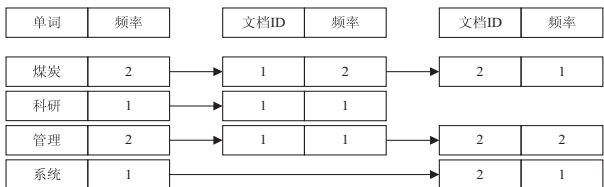


图 1 索引链表结果

搜索索引也主要分四步:查询输入、词法分析、搜索索引、相关性计算。

- (1) 查询输入:过去需要查询的文本数据;
- (2) 词法分析:类似索引创建的步骤二和三,将查询文本进行处理,处理后为单词;
- (3) 搜索索引:通过反向索引链表,找到包含处理后单词对应文档的 ID;
- (4) 相关性计算:为搜索出的每个文档进行打分,分值越高说明与查询文本相似度越大。

### 2 系统分析与设计

#### 2.1 系统设计思路

该系统基于 Java Web 技术,嵌入 Lucene 检索工具包,结合企事业单位科研管理业务,设计实现科研管

理系统。具体步骤如下:

- (1) 基于 Java Web 技术搭建 B/S 架构的系统框架,结合 SQL Server 2008 数据库,实现数据的增、删、改、查等业务操作。
- (2) 嵌入 ntoko 控件,实现电子文档的上传、下载与在线打开功能。ntoko 为能够在线打开 office、pdf 等文档的第三方控件。
- (3) 嵌入 Lucene 检索工具包,调用 Lucene 提供的 API,实现申报课题索引建立、词组划分,计算文档相关性打分,并转换为重复率。将重复率高于 30% 的文档罗列显示。
- (4) 结合 JPBM 工作流引擎,根据科研项目申报评审业务流程,实现所有申报项目的线上审批、驳回操作,基本实现无纸化科研管理。

#### 2.2 系统架构

系统采用 B/S 架构,逻辑架构主要分为 4 层:物理层、数据层、核心层、应用层,如图 2 所示。

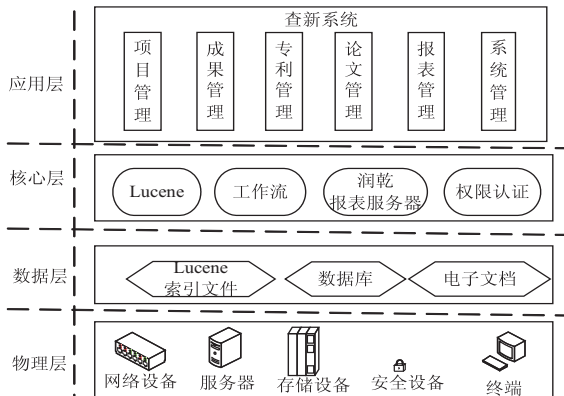


图 2 系统逻辑架构

应用层为查新系统提供的各种应用功能,包括各种页面的浏览与数据的操作,负责处理来自浏览器的业务请求。核心层作为应用层的底层支撑,一般为抽象出的成熟的工具、控件和功能模块,例如该系统调用的 Lucene 检索工具包,负责为应用层查新功能提供 API;润乾报表服务器为统计报表提供样式封装和数据分组等服务;权限认证为公用功能模块,已经应用在很多 Web 系统中。数据层表示数据的来源和存储,该系统涉及三方面的数据,系统填报数据存放放到 SQL Server 数据库中,Lucene 检索数据以文件形式存储,还有上传的电子文档也以文件形式存储在数据服务器中。物理层表示整个系统的硬件设备,该系统只使用了一些常规的网络设备和两台 IBM X3850 服务器。

### 3 关键技术

#### 3.1 Lucene 与领域本体结合

前面简单介绍了 Lucene 的搜索索引步骤,其中第一、二步都是对输入查询的操作,在实际应用中发现,

对实际输入的文本进行分词、搜索,效果比较理想,但是对采用了近义词替代的文本,则搜索效果不理想,例如运销和地销,含义相近,但搜索时不相关。因此采用了结合领域本体的办法<sup>[11-13]</sup>。在系统中建立领域术语库,将煤炭领域相近的术语以树形结构存储到数据库中,存储结构如图3所示。

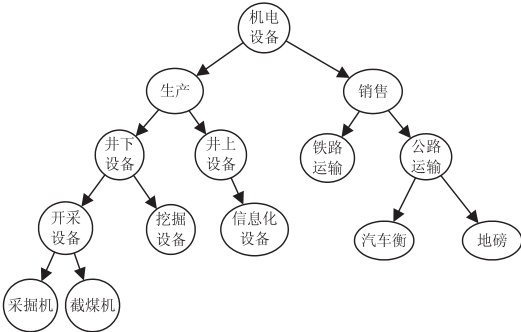


图3 领域术语存储结构

目前近似单词查询只是根据查询单词,简单地将该单词所属父节点下的叶子节点找出,取出的叶子节点表示语义近似的相关术语。例如查询单词为采掘机,根据树形结构找到父节点为开采设备,发现近义词为截煤机,同理根据汽车衡可以查询到地磅。

在查询文本处理为单词之后,进行领域本题库查询,将近似的单词一并查出后,进行合并,合并后再进行索引搜索和相关性打分操作。

3.2 高亮显示二次处理

查新结果展示中,一般会将重复的单词高亮显示,这样便于操作人员直观浏览查重结果。Lucene 提供了设置高亮的 API,对比结果可以高亮显示,但是系统应用中发现,由于高亮显示的都是词法分析后的单词,即使一句话中除了“的、地”等词,其余业务词语全部高亮,Lucene 没能形成整句或者整段的高亮,所以感觉不够连续,页面凌乱。因此,该系统对高亮显示进行了更改,在原有高亮设置的基础上,进行了二次处理。

首先,利用 Lucene 提供的 query 方法<sup>[5]</sup>,对文本进行高亮设置,代码如下:

```
StringpreHighlight = "< span style = \" background - color: yellow\">";
StringpostHighlight = "</span>";
query. setHighlightSimplePre( preHighlight );
query. setHighlightSimplePost( postHighlight );
```

设置完成后进行检索搜索,检索成功后,通过正则表达式,将查询文本和搜索文档按常用标点断句,代码如下:

```
Pattern pattern=Pattern. compile
("(^(?! [.,?!:;.,?!:;,...]))[“”]? [^.,?!:;.,?!:;,...]+
(([,.,?!:;.,?!:;,...][...]? [“”]?!$)");
```

然后循环处理每句话,公式如下:

$$P = \sum_{i=0}^n \frac{X_i}{Y}$$
 (1)

其中,  $n$  表示包含高亮字符 preHighlight 的个数;  $X_i$  表示第  $i$  个高亮区域包含的字符个数;  $Y$  表示当前断句包含的所有字符个数。计算后保留三位小数。根据实际测试效果,  $P$  取阈值为 0.7。如果  $P$  值大于 0.7,则整句话表述意思基本相同,则高亮显示当前整条语句,处理后对比显示凌乱的问题得到明显改善。

3.3 重复率计算

Lucene 是一个检索工具包,所以没有提供重复率计算的方法。系统采用 Lucene 提供的相关性打分计算公式进行重复率计算。打分方法采用的是向量空间模型(VSM)算法,将查询语句分解后的每个词和权重看作一个向量及查询向量,同理将搜索文档的每个单词和权重看作文档向量,VSM 模型算法就是计算这两个向量的夹角,夹角越小相关性越大,相关性打分公式如下:

$$\text{score}(q,d) = \text{coord}(q,d) \times \text{queryNorm}(q) \times \sum_{t \text{ in } q} (\text{tf}(t \text{ in } d) \times \text{idf}(t)^2 \times t. \text{getBoost}() \times \text{norm}(t,d))$$
 (2)

其中,  $t$  表示搜索的单词 Term;  $\text{coord}(q,d)$  表示当一篇文档中包含的搜索词越多,则此文档打分越高;  $\text{queryNorm}(q)$  计算每个查询条目的方差和;  $\text{tf}(t \text{ in } d)$  为  $t$  在文档  $d$  中出现的词频;  $\text{idf}(t)$  表示  $t$  在几篇文档中出现过;  $\text{norm}(t,d)$  为标准化因子;  $t. \text{getBoost}()$  指查询语句中每个词的权重。

利用上面公式计算的相关性分数,只是表示检索后的文档与查询文本的相似度,公式中很多因素,例如权重、词频、搜索词个数等,不是用来体现重复率,仅仅用相关性不能代表重复率。经过实际测试,Lucene 的打分机制并不适用于该系统,因此放弃了相关性模型计算公式,采用基于文本处理的方法进行重复率计算。

按照上述内容进行搜索高亮和高亮后二次处理,这样处理后,整片文档的相似语句已经标黄,计算重复率只要统计查询文本中标黄语句占整篇查询文本的比例,即可作为重复率使用。实现效果如图4所示,经测试重复率较准确。

立项计划:		实施计划一	
立项任务书附件:		4.22 集团科技创新基金实施办法. doc	可行性研究:
立项填写人:		测试人员一	备注:
上次审批人:		测试人员二	审批意见:
查新结果:		查重个人: 15; 查重率: 69%	
	重复率	项目（或成果）名称	类别
1	69%	批量测试项目 1	计划项目
2	69%	批量测试项目 2	计划项目
3	69%	批量测试项目 3	计划项目
4	69%	批量测试项目 4	计划项目
5	69%	批量测试项目 5	计划项目

图4 重复率显示页面

4 系统功能设计

系统功能结构如图 5 所示。

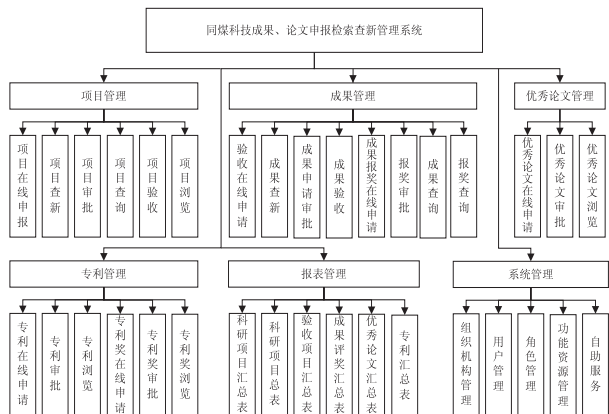


图 5 系统功能结构

项目、成果、论文、专利四个管理模块,是系统的主要业务模块,提供科技项目、科技成果、优秀论文、专利的申请、查新、审批、验收等功能,覆盖了整个系统业务流程。业务模块中涵盖了查新功能,在上面章节已经讲述。

报表管理模块负责系统统计分析数据的展示,包括历年科研经费汇总信息、申报入库信息、专利汇总信息等报表,便于企业科研管理者直观地分析科研情况。

系统管理主要包含系统的基础功能和 Lucene 领域库的配置信息。

5 系统应用

随着同煤集团技术中心的快速发展,科研项目和科技成果的数量呈现指数级增长,随之出现了申报过程繁琐、科技查新难度大和管理过程复杂等一系列问题。而科技查新针对不同类型的查新项目,通过检索文献手段,运用综合分析、对比分析等方法,为科技管理部门、评审机构和用户提供事实性情报信息的咨询服务<sup>[14-15]</sup>,是提高科技立项水平的必要手段。所以同煤集团技术中心在此背景下决定在企业内部搭建科技成果、论文申报检索查新管理系统,以实现科研项目和成果的在线申报、在线查新和统一管理。

实施编号:	PP201607280001	项目名称:	测试集团
项目开始日期:	2017-1-1	项目截止日期:	2020-12-31
项目级别:		项目类型:	推广应用
项目负责人:		项目成员:	王大伟
项目总经费:	300000.000 (双击可查看明细)	资助经费:	100000.000
主要研究内容和目标:	清洁能源研究		
实施计划:	计划 3 年完成		
立项任务附件:	立项任务书.doc	可行性研究报告附件:	可行性研究报告.doc
立项填写人:	测试员一	备注:	
审批人:	测试员二	审批意见:	

图 6 计划项目明细页面

系统通过给定的项目编号、时间段等信息可以查询各个状态的申报信息,以列表形式展示。图 6 为一

条申报信息的明细页面,点击页面中的附件信息可以下载上传的附件文档,查阅更详细的资料。系统目前已在同煤集团技术中心正式上线运行,取得了较好的应用效果。

6 结束语

以 Java Web 技术为基础,嵌入 Lucene 检索工具包,设计实现了一套科研管理系统。提出领域术语库结合 Lucene 的词法分析的想法,加入了查询单词的近义词或别称,提升了文档检索的准确性。针对 Lucene 高亮显示零散的问题,提出了高亮显示二次处理方法,该方法能够平滑处理对比文档的重复区域,使得对比查阅文档重复内容更加自然。在此基础上,依据文本高亮处理结果,重新设计了文档重复率的计算算法。系统在同煤集团技术中心进行了上线应用,根据现场运行情况反馈,满足了屏蔽项目重复申报和项目流程管理的需求,丰富了企事业单位科研管理工作的技术手段。对于查询单词的近义词搜索,目前仅仅通过简单地引用领域库资源树的叶子节点,可以进一步加入单词权重计算,更智能地搜索单词相近词语。

参考文献:

[1] 徐叶强,朱艳辉,栗春亮,等. 基于 Lucene 的海量数据库全文检索的设计与实现[J]. 湖南工业大学学报,2011,25(2):81-84.

[2] 黄江平,黄理灿,徐 玲. 基于 Lucene 的 PDF 文档的全文检索的实现[J]. 工业控制计算机,2012,25(5):103-104.

[3] 劳志佳. 基于 Lucene 3.5 搜索技术的研究与实现[J]. 现代计算机,2012(4):70-73.

[4] 夏 天,黄 文,马骏涛,等. Lucene 全文检索软件及其在学科信息服务平台中的应用[J]. 图书情报工作,2011,55(21):106-109.

[5] 李雪利,黄理灿,范晨熙. 基于 Lucene 的文档管理系统的设计与实现[J]. 工业控制计算机,2012,25(10):87-88.

[6] MILOSAVLJEVIC B,BOBERIC D,SURLA D. Retrieval of bibliographic records using Apache Lucene[J]. Electronic Library,2010,28(4):525-539.

[7] FENG Xia,TANG Xianchao. An improved dictionary-based Chinese word segmentation approach in Lucene[C]//Proceedings of 2010 international conference on services science,management and engineering. [s. l.]:IEEE,2010:363-366.

[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] 罗 刚. 解密搜索引擎技术实战: Lucene&Java 精华版[M]. 北京:电子工业出版社,2016.

[10] 吴众欣,沈家立. Lucene 分析与应用[M]. 北京:机械工业出版社,2008.



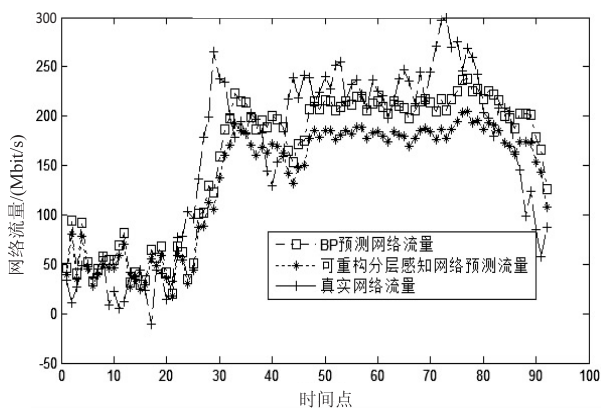


图 5 分层感知网络流量预测精度比较

评价神经网络硬件性能参数<sup>[17-18]</sup>一般采用每秒钟每核执行累加数目 (connection per second per core, CPSPC), 即神经网络每秒钟执行乘累加的数目与物理核数的比值。经过测试得出, 该算法的 CPSPC 值为 95 000, 明显高于专用神经网络处理单元。

## 4 结束语

利用多核片上网络技术实现分层感知网络进行网络流量预测, 充分发挥神经网络并行设计思想, 同时利用多核结构实现感知网络算法的可重构。仿真测试表明, 该算法具有良好的预测精度, 可扩展性好, 并行度高, 可以将其推广到其他神经网络的应用领域。

### 参考文献:

- [1] 田中大, 李树江, 王艳红, 等. 基于混沌理论与改进回声状态网络的网络流量多步预测[J]. 通信学报, 2016, 37(3): 55-70.
- [2] 邵 忻. 一种新的基于 ARIMA-SVM 网络流量预测研究[J]. 计算机应用研究, 2012, 29(5): 1901-1903.
- [3] PENG Y, CHEN K, WANG G, et al. Towards comprehensive traffic forecasting in cloud computing: design and application[J]. IEEE/ACM Transactions on Networking, 2016, 24(4): 2210-2222.
- [4] 田中大, 高宪文, 李树江, 等. 遗传算法优化回声状态网络的网络流量预测[J]. 计算机研究与发展, 2015, 52(5): 1137-1145.

- [5] CHAN K Y, DILLON T S, SINGH J, et al. Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg-Marquardt algorithm[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(2): 644-654.
- [6] 高述涛. CS 算法优化 BP 神经网络的短时交通流量预测[J]. 计算机工程与应用, 2013, 49(9): 106-109.
- [7] 赖锦辉, 梁 松. 基于 ACS 优化 BP 神经网络的交通流量短时预测方法[J]. 计算机工程与应用, 2014, 50(10): 244-248.
- [8] 冯华丽, 刘 渊, 陈 冬. QPSO 算法优化 BP 网络的网络流量预测[J]. 计算机工程与应用, 2012, 48(3): 102-104.
- [9] 万 勇, 王 沁, 李占才, 等. 一种神经网络硬件实现的可重构设计[J]. 计算机应用, 2006, 26(1): 202-203.
- [10] 李 昂, 王 沁, 李占才, 等. 基于 FPGA 的神经网络硬件实现方法[J]. 北京科技大学学报, 2007, 29(1): 90-94.
- [11] JERRAYA A, TENBUNEN H, WOLF W. Multiprocessor systems-on-chips[J]. IEEE Computer, 2005, 38(7): 36-40.
- [12] YANG H, KIM S, HA S. An MILP-based performance analysis technique for non-preemptive multitasking MPSoc[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2010, 29(10): 1600-1613.
- [13] 刘 澍, 王宏远. 基于混合遗传算法优化的 MLP 神经网络的调制方式识别[J]. 武汉大学学报: 理学版, 2008, 54(1): 104-108.
- [14] 朱新召, 胡哲琨, 周 莉, 等. 基于多核处理器的多层感知神经网络设计和实现[J]. 微电子学与计算机, 2014, 31(11): 27-31.
- [15] 张 帅, 宋凤龙, 王 栋, 等. 多核结构片上网络性能-能耗分析及优化方法[J]. 计算机学报, 2013, 36(5): 988-1003.
- [16] 付斌章, 韩银和, 李华伟, 等. 面向高可靠片上网络通信的可重构路由算法[J]. 计算机辅助设计与图形学学报, 2011, 23(3): 448-455.
- [17] 李 洋. 基于 QoS 保证的 2D-mesh 片上网络延时评价与性能优化研究[D]. 长春: 吉林大学, 2015.
- [18] 王 磊, 陆 超, 章隆兵, 等. 基于神经网络预测模型的异构多核处理器调度[J]. 高技术通讯, 2015, 25(6): 567-574.

(上接第 196 页)

- [11] LIU Tianyuan, SONG Meina, ZHANG Xiaoqi. Research of massive heterogeneous data integration based on Lucene and Xquery[C]//Proceedings of 2010 IEEE 2nd symposium on Web society. Beijing: IEEE, 2010: 648-652.
- [12] GANTER B. Two basic algorithms in concept analysis[M]. Berlin, Germany: Springer-Verlag, 2010.
- [13] RESNIK P. Semantic similarity in a taxonomy: an informa-

tion based measure and its application to problems of ambiguity and natural language[J]. Journal of Artificial Intelligence Research, 1999, 11: 95-130.

- [14] 马小雨. 基于 AHP 的煤炭科研项目评价系统的设计与实现[D]. 北京: 北京邮电大学, 2007.
- [15] 李广利, 李书宁. 科技查新报告自动生成软件的设计与实现[J]. 现代图书情报技术, 2013(2): 82-87.