文章编号：2096-1472(2018)-05-42-05
DOI:10.19644/j.cnki.issn2096-1472.2018.05.013

基于.net构建海量非结构文本与用户行为协同的搜索引擎研究

李德华[1]，巩 宇[1]，张自锋[2]，杨小龙[1]
(1.南方电网调峰调频发电有限公司，广东 广州 510630；
2.南方电网科学研究院有限责任公司，广东 广州 510623)

摘 要：从全文检索的原理出发，介绍搜索引擎和全文检索的索引原理和机制，通过实例讲述如何创建lucene索引、如何管理索引。通过索引的目录结构和内容结构的分析，讨论了基于.net平台如何使用lucene的索引功能创建搜索引擎。目前多数项目采用Java、C++等为开发语言，构建索引器。本文提出基于.net的索引器的设计实例分析，解决了在.net开发环境下，海量非结构文本搜索的问题。结合用户行为理论优化搜索引擎的用户体验，增加用户粘性，提高用户的忠诚度。

关键词：搜索引擎；索引器；.net；Lucene；用户行为
中图分类号：TP315 **文献标识码**：A

Research on the Search Engine for Building Massive Nonstructural Text and User Behavior Synergy Based on .Net

LI Dehua[1],GONG Yu[1],ZHANG Zifeng[2],YANG Xiaolong[1]
(1.*Southern Power Grid Peak Modulation FM Power Generation Co.,Ltd.,Guangzhou* 510630,*China*;
2.*Southern Power Grid Research Institute of Science Co.,Ltd.,Guangzhou* 510623,*China*)

Abstract: Starting from the principle of full-text search,this paper introduces the indexing principle and mechanism of search engine and full-text search.It describes how to create a lucene index and how to manage the index through an example. Through the analysis of the index's directory structure and content structure,we discussed how to use lucene's indexing function to create a search engine based on .net platform.At present,most projects use Java,C++,etc.as the development language and build indexers.This paper presents a design example analysis of a net-based indexer and solves the problem of massive unstructured text search under .net development environment.Combining user behavior theory to optimize the user experience of search engines,increase user stickiness and increase user loyalty.

Keywords:search engine;indexer; .net;Lucene;user behavior

1 引言(Introduction)

搜索引擎技术发展迅速，各大公司相继推出自己的产品，包括百度搜索、360搜搜、搜狗搜索、google搜索、bing搜索等，各个产品之间竞争激烈，因此如何提高用户的体验，保证产品的热度，在竞争中取得更多的市场分额，这成为了许多专家和学者研究的焦点。许多专家经过研究，提出结合用户的行为理论和用户行为属性，优化数据的索引操作(Indexing)，把用户的行为优化到搜索引擎索引器中。

2 搜索引擎与用户行为理论(Search engine and user behavior theory)

搜索引擎是一项以互联网信息为基础，进行信息资源的搜集、发现、理解、提取和处理的系统工程，为用户提供信息搜索服务，并对信息进行基于多种算法的理解和解析，为用户提供个性化的信息服务，是人们获取信息的工具，搜索引擎已经成为人们进入Internet门户。用户行为理论是指用户获取、使用相关的物品或服务，所产生的各项活动，用户

对使用对象的认知和使用，有一个完整的过程。那么用户行为理论与搜索引擎研究的结合，具体可以分为几个阶段，认知、熟悉、试用、使用、忠诚。具体内容如表1。

表1 用户使用搜索引擎各阶段分析

Tab.1 Phase analysis of user use of search engine

阶段	用户行为	参考指标
认知	访问搜索引擎主页	IP、PV、人均页面访问量、访问来源
熟悉	搜索与浏览	平均停留时长、跳出率、页面偏好
试用	用户注册	注册用户数、注册转化率
使用	用户登录	登录用户数、人均登录、访问登录比
忠诚	用户粘性	回访者比率、访问深度

搜索引擎系统一般由搜索器(Crawler)、索引器(Indexer)、检索器(Searcher)和用户接口(UI，User Interface)四个部分组成。本文主要从索引器入手，结合用户行为理论和用户行为参考指标，进行索引的优化。

2.1 基于用户属性优化索引器

索引器对获取的网页进行深度分析与挖掘，提取其中重要的信息，包括网页的标题、关键词、生成时间等。建立网页之间的相关关系，生成倒排文档等。在建设索引器过程中，为了提高用户的满意度，将结合用户的行为理论，从用户角度出发，提取更多与用户相关的属性优化索引内容。其中包括网页的地址属性、时间属性、人物属性等。

索引模块由信息提取模块、中文分词模块、网页提纯模块和索引模块组成。按照一定的方法对网页文档进行索引，形成索引数据库。首先是提取网页的纯文本页面，去除其中的HTML页面，然后采用分词算法，实现文本内容的切分，提取网页的关键属性，其效果如图1所示。

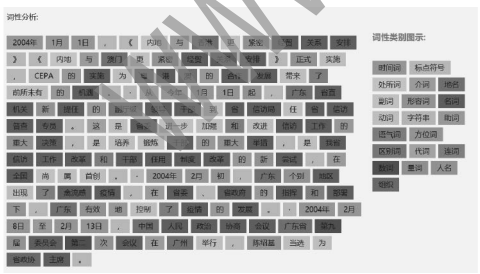

图1 文本分词及提取效果

Fig.1 Effects of text segmentation and extraction

根据文本的内容，提取其中的关键属性，结合用户的行为理论，优化索引器。当用户登录搜索界面的时候，系统获取用户的IP，进而获得用户的地址，那么用户的搜索结果，会自动根据地址的匹配，优化展示的顺序。同时根据用户的浏览结果推测用户的喜好，进而优化搜索结果的排序。

2.2 索引基本结构

以Lucene为例介绍索引的基本结构，Lucene作为一个优秀的全文检索引擎，其系统结构具有强烈的面向对象特征。Lucene是一个建立索引的框架，在这个框架下用户可以根据需要进行大量的二次开发，提高用户行为相关的各种属性的权重。Lucene定义了一个与平台无关的索引文件格式，建立用户行为相关的索引文件。

以下将讨论Lucene系统的结构组织，并给出系统结构与源码组织图，如图2所示。

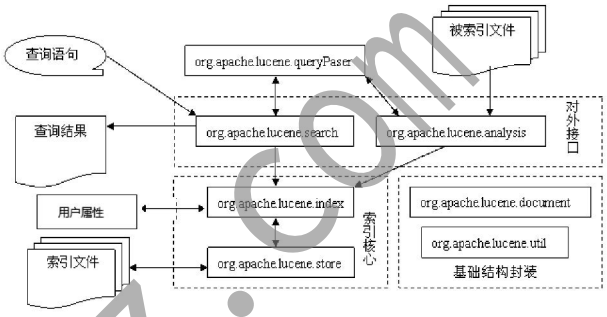

图2 系统结构与源码组织结构图

Fig.2 System structure and source code structure

说明：Lucene的最初版本是Java，经过很多技术达人的努力，开放出了的基于.net的索引器。

从图2中可以得知Lucene的系统由基础结构封装、索引核心、对外接口三大部分组成。其中直接操作索引文件的索引核心又是系统的重点。Lucene是一个基础框架，向开发人员提供完整的二次开发入口。

Lucene将所有源码分为七个模块(在.net中以文件包来表示)，各个模块所属的系统部分也如图2所示[1]。从面向对象的观点来考虑，Lucene应用了最基本的一条程序设计准则：引入额外的抽象层以降低耦合性。中国搜索引擎用户中18—24岁的用户最多，比例为35.9%；其次是25—30岁的用户，比例为18.0%40岁以上的用户比例也达到了17.0%。18岁以下用户数量最少，其比例仅为5.5%。针对以上特点索引器在构建的时候，会对内容进行特征优化，展示结果时，会根据用户的浏览和搜索行为优化搜索结果。Lucene在系统结构上的另一个特点表现为其引入了传统的客户端服务器结构以外的的应用结构。Lucene可以作为一个运行库被包含进入应用本身中去，而不是做为一个单独的索引服务器存在。根据调研结果，高频搜索内容分为三个类型娱乐型、商业型、知识型。各类型具体包含内容分别为娱乐型：电影/视频、MP3音乐、图片、博客等；商业型：商品/购物信息、财经/理财信息、企业信息；知识型：工作/学习资料、新闻资讯、软件等电脑应用、地图及本地信息、专业文档、网友回答等。通过以

上分析，索引器在构建时也必须相应的对大文本进行语义分析，抽取其中的文档属性，在Lucene开放源代码的中进行优化，这体现了Lucene在编写上的本来意图：提供一个全文索引引擎的架构，而不是实现[2]。开发人员根据需要对系统进行深度的定制优化。

2.3 倒排索引原理

倒排索引源于实际应用中需要根据属性的值来查找记录。这种索引表中的每一项都包括一个属性值和具有该属性值的各记录的地址。由于不是由记录来确定属性值，而是由属性值来确定记录的位置，因而称为倒排索引(inverted index)。带有倒排索引的文件称为倒排索引文件，简称倒排文件(inverted file)[3-5]。

一个倒排索引由文档中所有不重复词的列表构成，对于其中每个词，有一个包含它的文档列表。例如，假设我们有两个文档，每个文档的content域包含如下内容：

(1)President Xi Jinping visits Yichang,Hubei province,on Tuesday

(2)Quick brown foxes leap over lazy dogs in summer

为了创建倒排索引，首先将每个文档的content域拆分成单独的词(称它为词条或tokens)，创建一个包含所有不重复词条的排序列表，然后列出每个词条出现在哪个文档。实现了词语文档之间的映射关系。

3 基于用户行为的索引器的设计实例分析(Analysis of design examples based on user behavior indexer)

3.1 索引器总体结构设计

该系统共有三部分功能模块组成[6]，分别是文本预处理、分词分析、索引建立与维护，具体的结构图如图3所示。

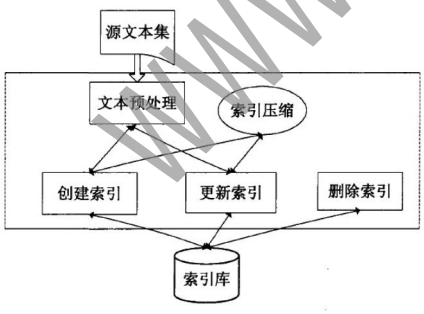

图3 系统结构图

Fig.3 System structure

由图3可知：该系统由三个功能模块、一个可调用的索引压缩类和一个索引文件库组成，其中文本预处理模块是对文本集进行预处理，索引压缩是在创建索引和更新索引时用到的一个内部类模块，所以在此特别的列出。结合用户的行为理论，本系统采用了基于单字的倒排索引文件结构。把用户的行为信息转化为索引结果的评价权重，以此优化检索的结果。

3.2 网页预处理

网页的预处理包括对搜集来的网页文档进行过滤、分词、转换等[7]。爬虫获取了各类网页之后，网页的内容是非结构化的数据，里面包含大量的不可读内容，而且没有任何保存的价值，首先是抽取其中的文本和相关文本，这将产生两方面的作用，一是提高索引的准确性，二是提取其中的链接，进而抓取更多的网页。

HTML文档与普通的文档不一样，HTML有自己的语法，通过不同的命令符来表示不同的字体、颜色、位置等版式，提取文本信息时需要把这些标示符都过滤掉。在识别这些信息的时候，需要同步记录许多版式信息，例如字体大小、是否是标题、是否加粗、是否是页面的关键词等，这些信息有利于搜索引擎判断这些单词在网页中的重要程度。

同时，对于HTML网页来说，除了标题和正文以外，会有大量的跟内容无关的标签信息、广告链接等，这些链接只能作为继续爬取信息的入口，但是和正式内容关系不是很大，在提取网页内容的时候，需要对这些链接进行识别和分类使用。网页预处理在网页搜集完成之后进行。在得到海量的原始网页集合后，网页中包含大量的HTML标记，这些标记的内容对于显示有重要价值，但是对于索引的建立则是没什么实际效果，因此必须对网页预处理，过滤掉这些HTML标记，提取出其中有用的内容。其次，由于网页上内容的多样性，有些网页的内容比较随意，或者包含大量的广告，这些信息需要在网页预处理阶段进行处理。

本设计能够对所有标记进行过滤，提取网页中的信息，对于过滤只含有广告而无文本内容的网页无法很好处理，但不影响简单的应用。网页预处理流程图如图4所示。

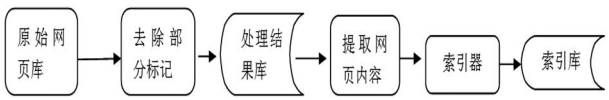

图4 网页预处理流程图

Fig.4 Flow chart of web page preprocessing

本设计中，网页预处理提取网页内容的设计过程中参考了《基于c#的HtmlParser建立文档树》该程序能够对一般的网页进行分析，然后进行节点的分类，建立相应的文档树，以HtmlParser程序来分析。但是HtmlParser该程序有一个缺陷，就是不能很好地清除<script><style>之类的节点，所以在使用HtmlParser之前，先用正则表达式去掉<script><style>之类的节点。再用HtmlParser提纯后，结果会生成很多 和空格需要再用正则表达式进一步提纯，最

后可以得到网页的具体内容。标题能够很好地反映网页的主题内容。在提取网页内容中就把标题和网页主体分开，然后建立超链接、标题和网页内容三者的联系。

处理原始网页的目的是去除网页中无关的HTML标签，只取出其中有用的信息。把HtmlParser引用到搜索引擎项目中，效果如图5所示。

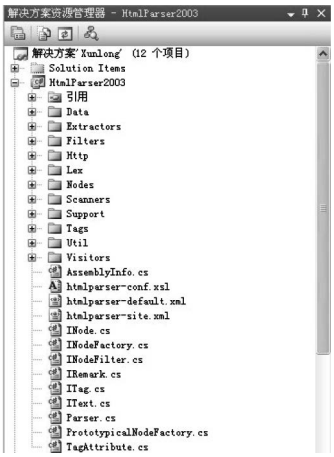

图5 加入HtmlParser

Fig.5 Interface after inserting HtmlParser

索引器的核心代码文件ClassIndex.cs，代码如下：

```
nSearch.FS.oneHtmDat newDat=nSearch.FS.ClassFSMD.GetOneDat(i);
//生成标题
nSearch.DebugShow.ClassDebugShow.WriteLineF("测试newDat："+newDat.url.ToString()+"信息");
WebRequest geturlcontent=WebRequest.Create(newDat.url);
WebResponse getresponse=geturlcontent.GetResponse();
Stream stream1=getresponse.GetResponseStream();
StreamReader read=new StreamReader(stream1,Encoding.GetEncoding("gb2312"));
string readcontent=read.ReadToEnd();
Match TitleMatch=Regex.Match(readcontent,"<title>([^<]*)</title>",RegexOptions.IgnoreCase|RegexOptions.Multiline);
string ti=TitleMatch.Groups[1].Value;
read.Dispose();
read.Close();
nSearch.DebugShow.ClassDebugShow.WriteLineF("标题"+ti.ToString());
//生成body
System.IO.Stream stream2=new
```

以上实现了网页代码中标题和网页主体的提取。

3.3 分词设计

中文分词[8](Chinese Word Segmentation)指的是根据一定的算法对文章中的句子、词语进行识别，切分成多个词语单元。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。在英文的行文中，句子中词语之间以空格进行区分，而中文只是字、句和段能通过明显的分界符来简单划界，词之间是没有明显划分的，虽然英文也同样存在短语的划分问题，在词的划分方面中文比之英文要复杂的多、困难的多，另外中文也有一词多意的现象，因此必须采用多种不停的算法，包括贝叶斯分类、神经网络识别等。

在提取了标题和网页内容后，必须利用分词系统进行词语的切分，然后才能用于建立倒排文档。提供分词工具的单位非常之多，有知名的研究机构也有互联网公司，根据项目的需要选择好的开发工具非常重要，本文根据需要，选择了中科院开发的c＃版本分词系统，而放弃了Lucene提供的中文分词工具[9](全称是Lucene.Net.Analysis.China.dll)，因为在使用Lucene的分词工具过程中，发现Lucene.Net.Analysis分词效果一般，并且已经被封装，相比中科院的分词系统缺乏灵活性，中科院的分词系统全开源，可以进行修改，特别是词典可以自由进行修改。本文采用中科院的分词系统，词典部分使用的是简化版本[10]。在分词的过程中，结合用户的行为理论，把用户相关的用户属性进行合理切分，并且作为建立倒排文档的关键词。

把分词系统导入到项目中的nSearch.Index文件中为分词系统做好准备，如图6所示。

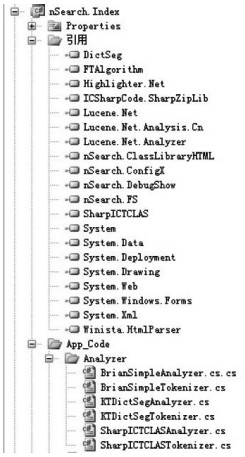

图6 分词系统及引用库文件架构图

Fig.6 Framework of segmentation system and reference library

在ClassIndex.cs建立分词的分析器，代码如下所示：

```
//定义分析器
```

```
Analyzer KTDAnalyzer=new KTDictSegAnalyzer(wordPath);
//PerFieldAnalyzerWrapper可以对不同的Field进行不同的分析
PerFieldAnalyzerWrapper wrapper=new PerFieldAnalyzerWrapper(KTDAnalyzer);
wrapper.AddAnalyzer("ID",KTDAnalyzer);
wrapper.AddAnalyzer("News_Url",KTDAnalyzer);
wrapper.AddAnalyzer("News_Date",KTDAnalyzer);
//创建IndexWriter
writer=new IndexWriter(indexDirectory,wrapper,true);
```

至此建立索引的准备工作基本完成。

3.4 索引创建

当所有的原始网页都处理好之后，就可以建立相应的索引，本实例为索引库建立了一个库E:\DATATEST\Index，用来保存全部的索引文件。索引建立流程[11]如图7所示。

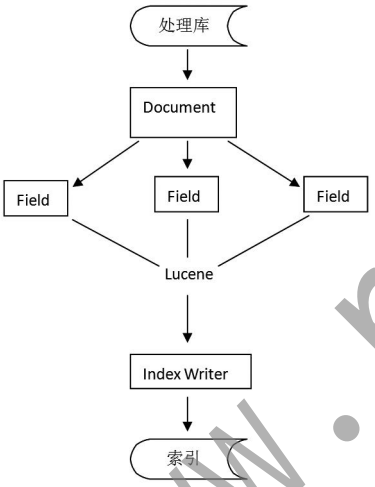

图7 索引建立流程

Fig.7 Flow chart of building indexing

首先将Lucene导入项目，效果如图8所示。

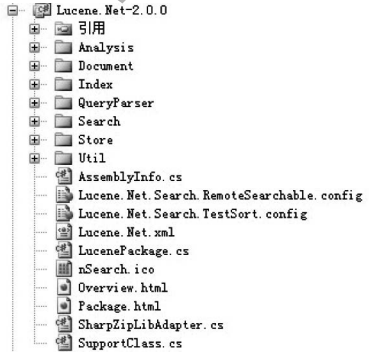

图8 Lucene文件

Fig.8 Lucene files

在引用了Lucene后在ClassIndex.cs中即可建立document，索引建立后会写到硬盘中，具体代码如下：

```
Lucene.Net.Documents.Document doc=new Lucene.Net.Documents.Document();
doc.Add(new Field("ID",i.ToString() ?? "",Field.Store.YES,Field.Index.UN_TOKENIZED));
doc.Add(new Field("News_Title",ti,Field.Store.YES,Field.Index.TOKENIZED,Field.TermVector.NO));//标题A
doc.Add(new Field("News_Body",body,Field.Store.YES,Field.Index.TOKENIZED,Field.TermVector.NO));//简要显示C
doc.Add(new Field("News_Url",newDat.url,Field.Store.YES,Field.Index.NO,Field.TermVector.NO));//url
doc.Add(new Field("News_Date",time,Field.Store.YES,Field.Index.NO,Field.TermVector.NO));//
```

以上就是一个索引建立的完整过程。

随着信息技术的发展和计算机深入广泛的应用，人们对全文检索的动态性要求变得十分的迫切，并且进一步还要求根据用户的需求进行个性化。应该说是全文索引的动态性是未来的一个必然的趋势。实现索引的动态更新的一个很大的难题在于更新时间复杂度非线性。动态全文数据库索引的创建效率非常高。由于倒排表模型在索引的创建效率上高于其他的模型，且应用也较为成熟，因此本设计主要也是使用倒排表模型的动态索引创建算法。在本设计中采用了一种准动态全文索引的更新的策略。

4 结论(Conclusion)

研究了全文检索索引器中涉及的几个关键技术，其中有中文分词技术，索引结构形式及其工作原理。建立了一个基于中文检索的索引器系统。在本文中系统的讨论了全文检索索引库所涉及的各种技术。在融合用户行为理论的要求下，必须对系统进行的各个模块进行深度优化，提升用户的满意度。在实际的应用中内存分配、分页管理、文件系统管理、虚拟内存技术、内存映射、cache管理等各种相关的技术都提高索引创建和更新的效率[12]。如何在这些方面对索引库进行进一步的优化，将是未来研究的工作和挑战。利用本文设计的索引器，可以应用于海量文本数据库中，建立资料的索引，提供快速的查找功能。

参考文献(References)

[1] 贾捷.基于MVC+Lucene.Net框架下的垂直搜索引擎研究与设计[D].吉林大学,2015.

[2] 孙艺珍,季小迪,张京涛.基于.Net的全文搜索引擎设计与实现[J].西安科技大学学报,2014,34(06):701-706.

[3] 王莉.基于ASP.NET搜索引擎模型的实现[J].计算机与现代化,2011(11):199-201;205.

[4] 张文生,孙永忠.ASP.NET网站搜索引擎优化方法研究[J].信息技术,2010,34(03):146-148.

[5] 蔡建超,郭一平,王亮.基于Lucene.Net校园网搜索引擎的设计与实现[U].计算机技术与发展,2006(11):73-80.

[6] 阴爱英.基于线程并行计算的Apriori算法[J].西安科技大学学报,2014,34(1):71-74.

[7] 戚艳军,龚尚福.用户角色的XML动态加密方法研究[J].西安科技大学学报,2012,32(1):101-106.

[8] 刘磊安,符志强.基于Lucene.net网络爬虫的设计与实现[J].电脑知识与设计,2010,6(8):1870-1878.

[9] 马俊.基于正则表达式技术的信息搜集引擎应用研究[D].成都:电子科技大学,2006.

[10] Otis Gospodnetic,Erik Hatcher.Lucene in action[M]Beijing: Publishing House of Electronics Industry,2007:100-105.

[11] 武毅.基于Lucene .Net的全文检索研究与应用[D].长沙:国防科学技术大学,2011.

[12] Chih-Hao Tasi.MMSEG:a word identification system formandarin Chinese text based on two variants of the maximum matching algorithm[OL].http://technology.chtsai.org/mmseg/,2013.

作者简介：

李德华(1985-)，男，硕士，工程师.研究领域：电网调峰调频技术，企业信息化.

巩　宇(1982-)，男，硕士，工程师.研究领域：电网调峰调频技术，企业信息化.

张自锋(1988-)，男，硕士，工程师.研究领域：信息技术情报，企业信息化.

杨小龙(1982-)，男，硕士，工程师.研究领域：电网调峰调频技术，企业信息化.

(上接第56页)

带学生参与科研项目、学生创新社团的硬件条件需求。要进一步完善高端设备，能够支撑计算机类专业高年级企业项目与实践模块的教学，同时能够支持教师进行部分科研工作的需求。

(2)建立长效校企共建合作机制，共育高校应用型技术技能人才

成立本地区甚至服务“一带一路”地区的高校计算机类专业校企共建指导委员会和管理机构，制定相关制度，形成人才共育、过程共管、成果共享、责任共担的合作协同育人机制，促进校企深度融合，为高校应用型技术技能人才培养校企合作机制建设提供探索与示范。在校企合作建立创客教育实验室的基础上，结合各自学校创客实验室的特色及功能，营造实验室功能氛围、技术氛围及企业文化、学校文化；形成集教学、科研、开放和共享服务的使用规则；梳理实验项目、实践模块，开发特色的创客教育实验项目[8]。

(3)探索高校计算机类专业人才培养特色，提升学生创客实践能力

适应社会职业岗位和岗位人才需求，加强高校计算机类专业学生创客实践环节，体现计算机技术发展的时效性、创新性、先进性、科学性，使创客实践教学、社会实践、技能竞赛、管理制度等一起形成有利于培养学生创新精神和实践能力的应用型人才培养模式，实现知识、能力和素质的协调发展。

参考文献(References)

[1] 李凌云.高职院校创客教育的价值、现状及优化路径[J].教育与职业,2016(24):57-59.

[2] 苏佩尧.高校“创客”实验室开放的内涵与误区[J].实验技术与管理,2017(1):250-253.

[3] 郑燕林.美国高校实施创客教育的路径分析[J].开放教育研究,2015(3):21-29.

[4] 王芳.引入“创客教育”的高职电类人才培养模式研究[J].教育现代化,2017(11):71-73.

[5] 陈颖.地方农科院校大学生创新创业人才培养体系的创建与实践[J].云南农业大学学报,2012(6):72-76.

[6] 刘长宏.大学生创新创业训练计划项目的实践与探索[J].实验室研究与探索,2014(5):163-166.

[7] 杨威.高校开放实验室建设与管理体制探究[J].实验技术与管理,2016(3):255-257.

[8] 雒亮.开源硬件,撬动创客教育实践的杠杆[J].中国电化教育,2015(4):7-14.

作者简介：

凌　翌(1980-)，男，硕士，副教授.研究领域：计算机应用.

郎振红(1975-)，女，博士，副教授.研究领域：计算机软件开发.