基于 Nutch 的垂直搜索引擎系统①

李耀芳[1]，张 涛[2]

[1](天津城市建设学院 电子与信息工程系，天津 300384)
[2](南开大学 信息技术科学学院，天津 300071)

摘 要：由于通用搜索引擎搜索精度不高，而国内各大物流港口搜索有效性较低，设计基于 Nutch 的港口物流垂直搜索引擎系统，实现了各个港口物流信息的快捷查询和共享。系统采用了基于向量空间模型的主题相关度判别算法并对该算法进行改进，加入元数据判别机制和重要标签所包含关键词的加权处理。加入“隧道处理”机制，以处理主题网页分离的问题，并且修改了检索结果排序的源代码，使其更适应垂直搜索引擎的要求。

关键词：Nutch 垂直搜索；向量空间模型；索引检索

Vertical Search Engine System Based on Nutch

LI Yao-Fang[1], ZHANG Tao[2]

[1](Electronic Information Engineering, Tianjin Institute of Urban Construction, Tianjin 300384, China)
[2](College of Information Technical Science, Nankai University, Tianjin 300071, China)

Abstract: Due to the low accuracy in current general search engines, and the poor search effectiveness of lower logistics in major domestic ports, the paper designs a port logistics Nutch-based vertical search engine system which achieves fast query and logistics information sharing. The system uses a theme based on vector space model identification algorithm and the relevance of the algorithm is improved by adding identification mechanism and the importance of metadata tags that contain the keywords of the weighting. Adding “tunnel handling” mechanism to deal with separation issues topic page, and modify the source code to sort search results to make it more responsive to the requirements of vertical search engines.

Key words: Nutch; vertical search; vector space model; index retrieval

随着因特网的迅猛发展，网络信息资源成几何级数增长，想要快速、准确地查找所需的信息越来越难，搜索引擎整合了互联网上众多的网页资源，能方便用户查找所需要的信息。但是目前通用搜索引擎在使用中面临着许多问题[1]，而与物流信息相关的垂直搜索引擎的检索主题相关度不高、信息更新不及时、信息量小，并且没有专门针对国内港口物流信息的搜索引擎[2]。因此，本文以天津港数字化口岸公共服务平台为研究对象，构建基于 Nutch 港口物流信息垂直搜索引擎，实现了港口物流信息的快捷查询和共享。系统对主题相关性判别、检索结果排序、隧道处理等问题在原有工作的基础上做了一些改进，提高了主题判别的准确度和效率，使信息的定位和查找更加的精确，减少了不相关信息的干扰，并提高了系统对于互联网复杂环境的处理能力。

1 基于Nutch的垂直搜索引擎的实现

1.1 系统体系结构

按照搜索引擎的一般结构[3]，系统可以分为搜索引擎内核部分和辅助部分。系统的功能框架设计如图 1 所示。

按照与搜索引擎结合的紧密程度，主题管理、资源发现、检索结果显示等内容属于辅助部分；网络爬虫、网页分析、主题过滤、网页索引、网页检索等内

① 基金项目:国家科技支撑计划(2007BAH10B01)
收稿时间:2010-12-24;收到修改稿时间:2011-02-19

容属于搜索引擎的内核部分[3]。

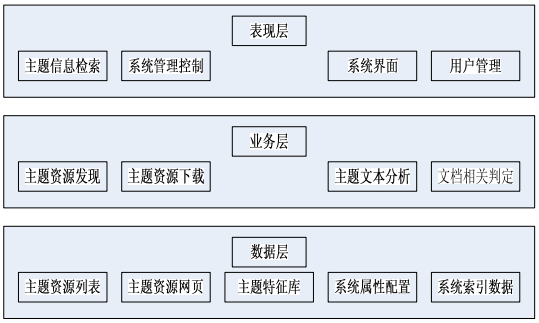

图 1　本系统功能框架图

该系统的体系结构如图 2 所示：

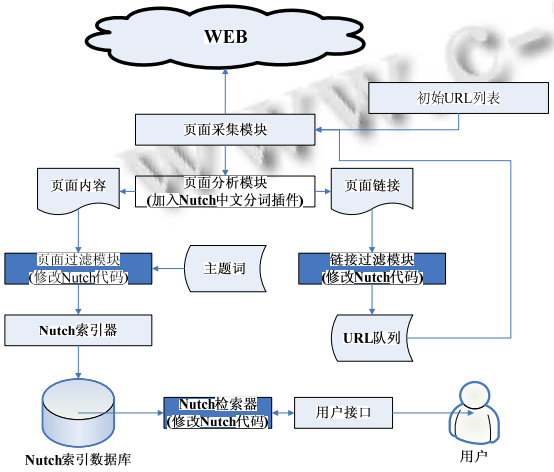

图 2　本系统体系结构图

其中蓝色背景的模块是重点要实现或改进的部分：

(1) 页面过滤模块：修改 Nutch 代码加入主题相关度判别功能，以实现对网页主题进行相关度判定和过滤。

(2) 链接过滤模块：修改 Nutch 代码加入处理“隧道现象”的功能，使爬虫可以爬取被无用页面分隔的主题页面。

(3) Nutch 检索器：修改 Nutch 代码在原有的检索结果排序的基础上加入页面主题相关性因素，使相关度高的结果优先显示给用户。

下面从以下几个重要方面介绍系统构成：

1.2 起始 URL 列表的生成

系统实现中采用了人工整理判定和元搜索相结合的 URL 列表生成策略。首先把国内各大港口的物流栏目和资讯栏目的 URL 地址加入到 URL 列表中，然后添加通过元搜索策略收集到 URL。程序实现方面，采用 HtmlParser 完成，利用主题词，生成搜索引擎的查询词列表，通过提交列表，获得搜索引擎的检索结果页面，对页面用 HtmlParser 解析提取出其中的链接，再进行人工分析[4]。

程序代码如下：

```
public class MetaSearchForURL{
public static void TravelWordTable(String filename) throws IOException
{TODO:从数据库中得到主题词表，由各个主题词构造通用搜索引擎的查询请求词列表}
public static void getBaiduURLs(String url, String pageEncoding) throws ParserException
{TODO:解析百度搜索引擎返回的页面，并提取其中的 URL}
public static void getGoogleURLs(String url, String pageEncoding) throws ParserException
{TODO:解析谷歌搜索引擎返回的页面，并提取其中的 URL}}
```

最后确定的起始 URL 列表如下：

http://www.tjportnet.com#天津港物流信息网
http://www.qingdaoport.net#青岛港物流信息
http://www.portinfo.net.cn#上海港
http://www.gzport.com#广州港
……

1.3 主题相关性判别在 Nutch 中的实现

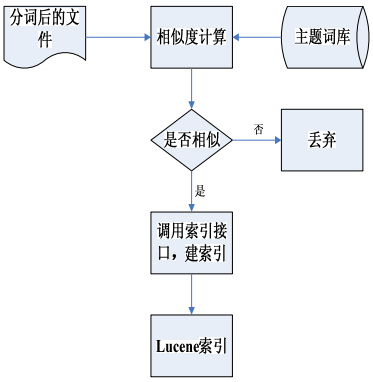

图 3　主题相关度判别功能图

Nutch 是基于整个互联网的搜索引擎，因此并没有主题相关性判别功能，要实现垂直搜索引擎的功

能需要在其基础上加以修改使其具有这项功能。在网页下载后，对网页的主题相关度进行判别，通过分析网页是否具有<meta>、<keywords>等标签，判断并计算得出该网页是否与主题相关，若相关则对其建立索引，不相关则丢弃。判别计算模型采用的是向量空间模型[5]，其基本功能流程如图 3 所示。

1.4 隧道穿越的实现

本文提出了优先级递减和黑名单的 URL 搜集策略来解决该问题。其功能流程如图 4 所示：

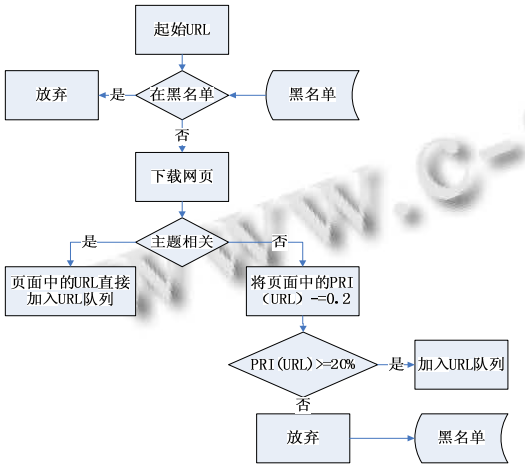

图 4 实现隧道穿越的流程图

在 Nutch 的网络爬虫(Crawl. java)实现中其 URL 队列是一个优先级队列，这样可以通过让一个与主题无关的 URL 及其子链接的优先级逐步递减，降低其优先级而不是直接删除，从而为发现另一个主题团提供了可能，且易于实现。同时将与主题彻底无关的 URL 加入黑名单中，以减少搜索范围，提高效率。

1.5 改进 Nutch 的结果排序算法

Nutch 原有的基础排序算法是 OPIC（On-line Page Importance Computation）算法[5]，OPIC 算法对于每个页面，存储两个值：cash 和 history 值。最初，对于网络图设置一个总的 cash 值，将此总的 cash 值平均的分配给每个页面。当进行计算时，页面的 cash 值存储页面从上次爬取时间开始获得的 cash 值之和，页面的 history 值存储页面从算法的开始就获得的 cash 值之和。在计算时，不断地选取页面进行抓取。当某一页面被选取，将它的 cash 值分配给它所指向的那些页面，将此 cash 值加到这个页面对应的 history 值上，最后将此 cash 值重置为 0。为了估算图中每个页面的 PageRank，用向量 Xt 表示在算法的第 t 次迭代后：

$$X_t = \frac{H_t}{\|H_t\|_1} \quad (1)$$

其中，Ht 是所有页面在第 t 次迭代后的 history 值得向量。由于本系统是在 Nutch 平台上开发的垂直搜索引擎，因此对网页的相关性要求更高，所以在对网页进行排序时，可以综合考虑主题相关度和链接分析两个关键因素，具体的算法实现中，应该对主题相关度和 PageRank 值赋予不同的权重，则网页的重要程度值可以表示为：

$$P = w1 \cdot Sim(V,D) + w2 \cdot R(u)$$

其中，Sim(V,D)是上述通过主题相关度模块计算出的主题相关度的大小，R(u)是利用 OPIC 算法计算出的可用用于页面排序的网页 PageRank 值，w1 为主题相关度的权重，w2 为 R(u)的权重，二者的取值可以根据实验需求选定，必须保证 w1 + w2 = 1。本算法拟定 w1 取 0.6，w2 取 0.4。

改进的算法将主题相关度和链接分析相结合，提高了排序结果的质量，可以对于各个因素设置权重系数，有利于灵活调整各种因素对页面优先度得分的影响程度。经过经验数据分析和人工调整，可以将搜索系统性能调整到最佳状态。

2 系统运行和测试结果

2.1 系统运行步骤

初次运行时需要首先确定与本领域相关的主题词和起始 URL 列表，然后将其加入系统的配置文件中，运行流程如图 5 所示：

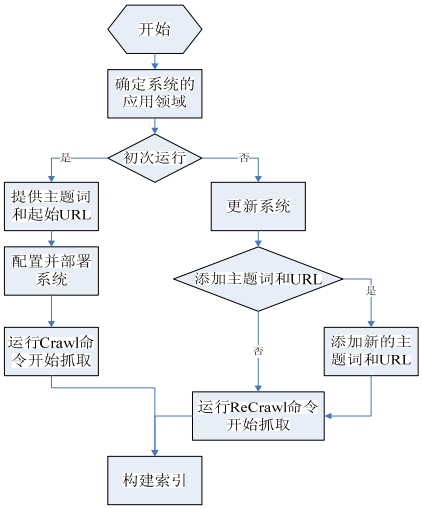

图 5 本系统运行步骤

以下是实际运行时的界面：

(1) 主题词和起始 URL 配置界面如图 6 所示：

图 6　本系统配置工具

在“主题词管理”栏可以添加、修改主题词和权重，也可以删除主题词；在“起始 URL 管理”栏可以打开起始 URL 文件，在其中添加或者删除起始 URL；在“运行管理”栏可以点击“初次运行”按钮或“更新”按钮，已开启 Cygwin 的命令行界面。

(2) 爬虫运行界面

Nutch 是为在 Linux 系统下运行而开发的，因此在 Windows 下需要安装 Cygwin 工具来模拟 Linux 环境才能使用 Nutch，启动 Cygwin 以后，会在 Windows 下得到一个 Bash Shell，在该命令行下输入：

bin/nutch crawl urls –dir tjportinfo –depth 5 –topN 100 –threads 10

如图 7 所示：

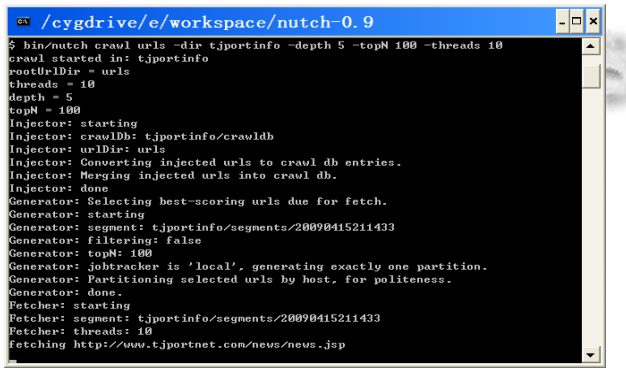

图 7　实际抓取

(3) 检索页面

运行 Tomcat，将 nutch-0.9.war 文件解压后放到 tomcat\webapps 文件夹下并替代原来的 ROOT 文件夹，打开 ROOT\WEB-INF\classes 下的 nutch-site.xml 文件，在<configuration></configuration>标签中添加索引库路径：

<property>

<name>searcher.dir</name>

<value>E:\workspace\nutch-0.9\tjport</value><description></description>

</property>

其中的<value></value>标签指定索引库的原始路径。

Tomcat 运行后启动的检索页面如图 8 所示：

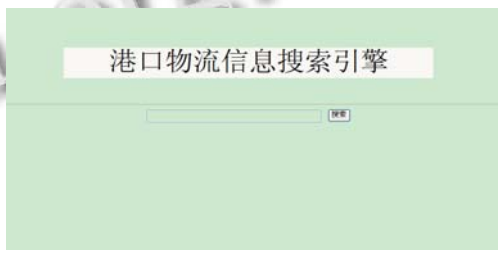

图 8　检索页面

(4) 检索结果页面

输入“天津港”之后的检索结果页面如图 9 所示：

图 9　结果显示页面

2.2 测试结果与分析

本文所实现的系统在收集页面数达到 5000 个时停止收集，从中随机选择 1000 个页面，这 1000 个页面中包含主题相关页面和主题无关页面，在系统评价的结果上进行人工评价，得出系统评价精度。

(下转第 47 页)

而左车灯保持中间初始位置不变，所以表 1 中的左车灯的理论和实际转角值为 0°。同理，当向左打方向盘时，右车灯的理论和实际转角值为 0°。

2）当车辆行驶速度低于 5km/h 时或者方向盘转角不大于 5°时，AFS 功能不启用，此时左、右车灯的理论和实际转角值为 0°。

3）左、右车灯的旋转角度不同，右车灯的水平最大转动角度可达到 18°，左车灯的水平最大转动角度可达到 15°。

4）车灯的理论转动角度和实际转动角度的最大误差不超过 0.2°，满足实际使用要求。

5 结语

本文根据自适应前照灯系统的功能要求，开发了基于汽车驾驶模拟器的 AFS 系统半实物硬件仿真平台，该半实物仿真平台引入驾驶模拟器作为 AFS 控制模型中所需的车辆各种驾驶信息的信息源，并在保持 AFS 控制器硬件不变条件下，将其原先所具有的传感器采集、数据解算、LIN 总线传输功能转换为 CAN/LIN 网关功能。实验证明，AFS 半实物仿真平台能够成为 AFS 硬件平台设计和控制器设计阶段和实车验证阶段之间一个有效的缓冲，为后续为 AFS 系统控制策略研究和算法设计验证奠定必要的基础。

参考文献

1 Roslak J. Active lighting systems for improved road safety. IEEE Intelligent Vehicles Symposium, 2004,6: 682–685.

2 Hacidekir T, Karaman S, Aksun G. Adaptive head-light system design using hardware-in-the-loop simulation. International Conference on Control Applications, 2006, 5: 915–920.

3 张新江.Ford 发表全新 AFS 智能型主动转向头灯技术.轻型汽车技术,2008,2:11.

4 左国章.现代轿车车灯的流行趋势.汽车与配件,2003,43: 27–29.

5 古强.智能照明系统.世界汽车,2006,3:74–75.

6 雷玲.智能车灯控制器的硬件在环测试系统研制[硕士学位论文].武汉:武汉理工大学,2009.

7 雷玲,吴青,陈建林,初秀民.基于 xPC 的智能车灯硬件在环仿真系统开发.武汉理工大学学报,2009,31(23):126–129.

(上接第 196 页)

表 1 主题相关度判断结果分析

系统评价结果		人工实际评价结果		主题评价精度
评价结果	页面数	主题相关	主题无关	
主题相关	867	732	135	84.4%
主题无关	133	37	96	72.2%

从表 1 中可以发现，本文系统在当前的阀值下，已经达到了一定的主题评价精度。

为了说明本系统的主题搜索性能，选择通用搜索引擎 Google 对关键词“港口 物流”进行搜索，同时使用本系统进行同样的搜索，对两个系统结果集的前 200 个页面进行了主题相关度评价，对比数据结果如表 2 所示：

表 2 与通用搜索引擎比较

搜索引擎	相关时间	主题相关网页数	查准率
本系统	1.54s	173	86.5%
Google	0.21s	89	44.5%

实验结果表明，港口物流信息垂直搜索引擎具有明显的主题倾向性，结果的查准率优于通用搜索引擎。

3 小结

本文就垂直搜索引擎的关键技术进行了研究，并提出了一种基于 Nutch 平台的垂直搜索引擎解决方案，研究并实现了港口物流信息垂直搜索引擎。实验证明系统设计和实现方案是切实可行的，基本达到了预期设计目标。该系统的研究促进了港口物流信息化的发展并对该领域的从业人员起到了很好的帮助作用。

参考文献

1 Gulli A, Signorini A. The Indexable Web is More than 11.5 billion pages. Proc. of the 14th International World Wide Web Conference. Chiba, Japan, 2005.

2 Sullivan D. Fifth Annual Search Engine Meeting Report. Boston, MA, 2000.

3 印鉴,陈忆群,张钢.搜索引擎技术研究与发展.计算机工程,2005,l4.

4 李世明.专业搜索引擎中信息过滤的研究与实现.北京:北京化工大学,2005.

5 Menczer F, Pant G, Srinivasan P. Topical Web Crawlers: Evaluating Adaptive Algorithms. ACM Trans. on Internet Technology, 2004.