基于Nutch的Web网站定向采集系统

徐健　　中山大学资讯管理系
张智雄　中国科学院国家科学图书馆

1 引言

- 利用网络信息更新速度快、获取方式灵活等特点，可以实现对特定领域、学科的实时监测和有效分析。而开展这类任务的第一步，就在于如何将相关网络科技信息内容存储到本地。
- 为了实现对WEB上特定领域英文科技信息的采集，我们对目前具有代表性的开源网络抓取软件进行了分析，并最终选择在Nutch基础之上进行多种扩展和改进的专题网站定向采集方案。

2 Web抓取开源软件比较分析

- **表1 四种Web抓取开源软件的特征比较。**

比较项目	Nutch	Heritrix	WCT	Web-Harvest
操作方式	命令行	Web控制界面	Web控制界面	界面/命令行
Web抓取功能	有	有	有	有
集群扩展能力	有	无	有	无
抓取内容完整性	只对可索引内容进行抓取	完整	完整	对网页特定字段进行抓取
内容索引功能	有	无	无	无
搜索功能	有	无	无	无
内容解析	有	无	无	针对特定字段
链接解析	有	无	无	无
网页评分	有	无	无	无
采集过程管理	无	无	有	无

2 Web抓取开源软件比较分析

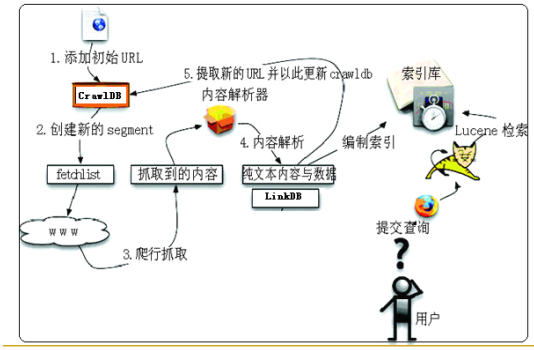

图1 Nutch的运行机制

3 Web网站定向采集系统整体设计

图2 基于Nutch的Web网站定向采集系统整体设计

3 Web网站定向采集系统整体设计

- 基于Nutch的Web网站定向采集系统整体设计具有以下几个特点：

（1）对种子站点的动态管理。
（2）抓取配置的集中管理。
（3）基于子任务的抓取管理。
（4）网页自动去噪和去重。

4 核心问题

4.1 获取种子站点
4.2 抓取任务管理
4.3 网页去噪
4.4 获取新的种子站点

4 核心问题-获取种子站点

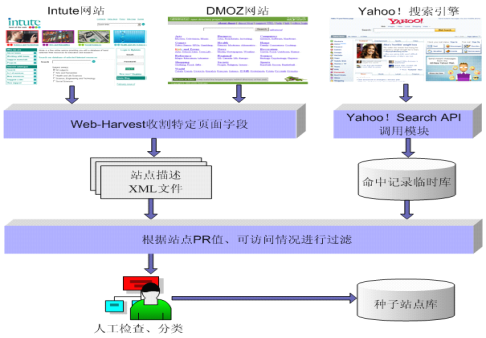

图3 获取种子站点模块设计

4 核心问题-抓取任务管理

- 实现了对种子站点和抓取过程进行管理的功能（配置统一管理和抓取过程管理）。
- 运行机制：

（1）根据用户事先设定的子任务规模，任务创建模块将自动把要抓取的若干种子站点进行分组，每一组对应一个抓取子任务。

（2）调用Nutch抓取接口，逐个运行上一步产生的子任务。运行过程中对各子任务运行状态进行记录。

（3）当各个子任务都运行完毕时，调用Nutch合并接口，将各子任务对应数据进行合并。

- 基于上述机制，特定领域大规模种子站点的个性化抓取配置和定向采集得以有效实施。

4 核心问题-网页去噪

（1）获取网页正文题名、作者/发布者以及网页发布/修改日期。
（2）使用Html Parser去除脚本、图片以及其它标签，获得只有链接和文本的字符串。
（3）根据导航栏的一般性特征（例如：“|”符号的数量，词的数量，空格数量等）去除导航栏文字。
（4）去除广告。通过两个规则来判断某一行是不是广告：
规则1，一行中的链接数不为0，且词数小于某个阈值；
规则2，一行中的链接数和词数之比大于某个阈值。
（5）去除所有以“<”和“>”标识的链接文字。
（6）去除版权声明信息。根据特征语词，且词数少于某个阈值。

4 核心问题-获取新的种子站点

图4 获取新种子站点流程

5 实验效果-收割统计数据

表2 抓取系统实验数据表（376个人工智能领域机构种子站点，三层抓取）

抓取批次	第一批	第二批	第三批
抓取时间	2008-12-05	2008-12-30	2009-01-04
抓取用时	10小时41分	15小时27分	15小时59分
单批抓取网页记录数	20005条	39616条	42039条
单批有效网页记录数	19502条	29647条	20185条
累计有效网页记录数	19502条	49149条	69334条

5 实验效果-去噪前

图5 一个原始网页示例

5 实验效果-去噪后得到的文本

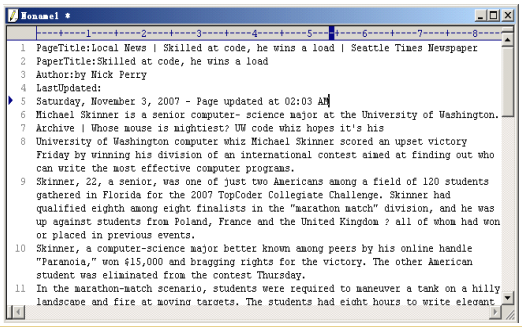

图6 去噪后的效果

5 实验效果-新种子站点的获取

- 在第三批抓取数据基础上，进行新种子站点获取实验。
- 在对3,797,988条链接数据进行共链分析、候选种子站点去重、连通情况和PageRank值过滤等步骤后，获得较高质量的候选种子站点1,065个。
- 经人工判断和分类后，共获取人工智能领域新的机构种子站点43个。被排除的站点多为新闻站点、广告站点、大学站点以及与人工智能领域相关，但不属于机构类型的站点。

6 结语-总结

- 在对Nutch开源软件进行扩展和改进的基础上，提出并实现了Web网站定向采集系统。
- 针对该系统中四个核心问题的解决方法，进行了较为深入的探讨。
- 目前，我们已经将Web网站定向采集系统应用于对科技领域站点的监测任务中，并获得了较好的效果。

6 结语-下一步的工作重点

（1）由单机抓取系统扩展为集群抓取系统。

（2）对Nutch原有接口和扩展功能接口进行标准化封装，方便任务调用和管理。

（3）新的种子站点的获取将更加智能化。

谢谢大家！

徐健
issxj@mail.sysu.edu.cn