

面向主题的垂直搜索引擎系统的研究与实现

施 隼^{1,2}, 王恒山¹, 肖仰华³, 丁卫平²

(1 上海理工大学 管理学院, 上海 200093; 2 南通大学 计算机科学与技术学院, 江苏 南通 226019;

3 复旦大学 计算机科学技术学院, 上海 200433)

摘 要: 针对通用搜索引擎的信息量大、查询不准确、深度不够等问题, 给出了面向主题的垂直搜索引擎的体系结构, 设计了垂直搜索引擎系统的爬行策略, 对系统核心信息采集模块运用了多线程技术及基于 VSM 的主题相关度判断算法进行主题网页爬行, 并通过 Lucene. Net 的索引与检索技术建立系统的检索算法, 实现了一个面向特定主题的垂直搜索引擎应用系统. 实验测试结果表明, 该系统具有较高的提取效率, 其检索的准确率、召回率均大大高于通用搜索引擎, 具有较好的实用价值和商业应用前景.

关键词: 垂直搜索; 网络爬虫; Lucene. Net; 正则表达式

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2011)07-0001-04

Research and Implementation of Subject-Oriented Vertical Search Engine System

SHI Quan^{1,2}, WANG Heng-shan¹, XIAO Yang-hua³, DING Wei-ping²

(1 College of Management, University of Shanghai for Science and Technology, Shanghai 200093, China;

2 College of Computer Science and Technology, Nantong University, Nantong 226019, China;

3 School of Computer Science, Fudan University, Shanghai 200433, China)

Abstract: A general search engine usually suffers from returning to users too much unrelated information. To avoid these problems of general search engine, a subject-oriented vertical search engine is proposed in this paper followed by the introduction of the crawling strategy and the architecture of this vertical search engine. Some distinctive characteristics of this search engine include: the system runs its core modules on multiples threads; a topic-related determination algorithm based on VSM is implemented in the system. The search engine system as well as its core algorithms is built upon lucene. Net. Experimental results show that the accuracy and recall rate of this system are higher than general search engine. Therefore, it is practically valuable and can be applied into more real applications.

Key words: vertical search; web crawler; lucene. net; regular expression

1 引言

随着 Web 信息的增长, 通用的搜索引擎在查询结果集中存在大量重复或不相关的垃圾信息, 不能满足特定领域精确搜索的需求. 垂直搜索是针对某一个行业的专业搜索引擎, 其特点就是“专、精、深”, 具有行业色彩^[1], 是搜索引擎的细分和延伸, 也是对

网页库中的某类专门信息进行一次整合. 与通用搜索引擎海量信息无序化相比, 垂直搜索引擎则显得更加专注、深入和具体, 其日益渐显的商业价值, 将成为面向特定主题的企业应用之首选.

2 垂直搜索引擎的体系结构设计

通用搜索引擎系统主要由采集器、索引器、检索

收稿日期: 2010-11-08; 修回日期: 2010-12-12

基金项目: 国家自然科学基金项目(61003001, 71071098); 江苏省自然科学基金项目(BK2010280); 南通市科技计划项目(K2008018, K2008031)

器、相关数据库及用户接口等部分组成^[2]. 而垂直搜索引擎为了实现查询结果与主题相关, 还需增加建立面向主题的特征词表以及主题识别器, 用于判断检索出的网页信息是否与此主题相关. 下面给出面向主题的垂直搜索引擎结构框架如图 1 所示.

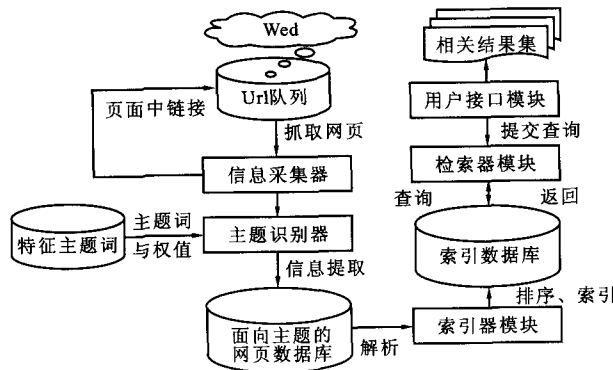


图 1 面向主题的垂直搜索引擎结构框架

(1) 信息采集器

信息采集器又叫网络爬虫(Spider)^[2], 专门用来搜集互联网上各种资源网页信息的一种智能主体. 它能自动地在互联网上搜索爬行, 负责从互联网上搜集网页. 它从初始 URL 队列出发, 遍历其相应的网页, 并把解析好的 HTML 源文档送给主题识别器来提取、过滤. 一般网络爬虫从初始的 Url 地址出发, 根据 Url 中指向其它 Url 地址的超链接而跳到其它的网页, 不断地深入和扩展, 遍历整个网络.

(2) 主题识别器

主题识别器采用基于主题词表匹配的向量空间模型 VSM(Vector Space Model)^[3]算法对所采集的页面进行主题相关性的判断, 以此过滤掉相关度小于预先设定阈值的页面, 并将页面信息保留至面向主题的网页数据库, 从而不仅提高了信息索引的准确率, 而且也保证了面向主题来源的相关性及专业性.

(3) 索引器模块

索引器模块的功能是对经过主题识别器所过滤的信息进行解析处理, 负责面向主题网页数据库的信息进行提取、分词等各种预处理操作, 然后将网页信息生成文档索引库.

(4) 检索器模块

检索器模块是连接用户接口界面, 在索引数据库中根据用户查询要求快速检索出用户所需的文档信息, 并对文档信息进行相关度评价, 将输出的结果进行排序反馈给用户.

(5) 用户接口以及相关数据库

用户接口是用户通过输入查询条件得到相应有效的数据信息, 其作用是为用户提供可视化的查询输入和结果输出界面. 相关数据库用于存储 URL 种子队列数据、与主题相关的特征主题词、特定主题网页信息的结果集数据以及通过索引器生成的索引数据.

3 面向主题的垂直搜索引擎关键技术

3.1 面向主题的信息采集器的设计

信息采集器主要负责从互联网上获取面向主题的网页信息并进行处理. 采集器的设计关键在于网络爬虫和网页解析器的设计. 网络爬虫是搜索引擎采集数据的主要工具, 由于垂直搜索引擎检索的信息内容仅限于特定主题或专门领域, 故网络爬虫在采集数据的过程中无需对整个 Web 进行遍历, 只需访问与主题相关的页面.

搜索引擎的信息采集过程可由管理员进行管理. 采集器首要工作先给网络爬虫设定采集规则, 即预先设定有效的正则表达式, 把采集到的信息记录入主题网页信息数据库中. 采集器由管理员设定起始 URL 集合, 起始 URL 集合是一组最初要访问的主题网站地址. 采集器启动后, 将起始 URL 集合加入待处理队列中, 然后不断地从待处理队列中取出 URL, 并对取出的 URL 进行链接访问、解析网页内容. 解析后将获得新的与主题相关的 URL 加入待处理队列. 网络爬虫不断从待处理队列取出 URL 进行处理, 直到待处理队列为空即采集结束. 网络爬虫信息采集器的流程图如图 2 所示.

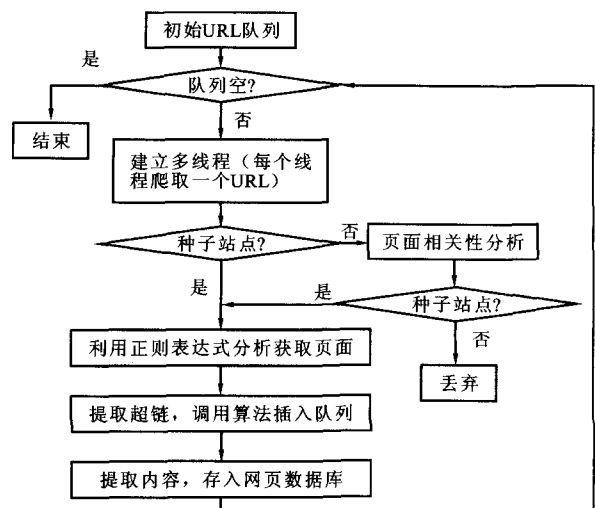


图 2 网络爬虫信息采集器的流程

根据图 2 网络爬虫信息采集器的流程图, 本系统设计针对面向特定主题的大型电子商务网站如太

平洋网等站点进行信息采集,其工作方式如下:

(1) 从初始 URL 队列(其初始值为预定的种子站点 URL)中获得一个 URL 请求页面,种子站点的建立采用了人工预选方式。

(2) 提取超链和页面内容信息,对种子站点内部链接直接插入待处理 URL 队列头部,将站点外部链接 URL 插入待处理 URL 队列最后端。

(3) 对非种子站点的 URL,提取并分析其主题相关性.若相关则按照和种子站点的页面相同的方式处理,否则直接丢弃,对整个站点都不再采集。

(4) 利用正则表达式分析获取的页面,解析获取目标网页中的标题、正文及超链接等内容信息,将结果存入到主题网页数据库。

(5) 将此 URL 加入已处理 URL 列表并获取下一个待处理的 URL,如此不断循环直到待处理 URL 列表为空。

为提高信息采集的效率,这里将网络爬虫设计成多线程的方式,用多个线程并发在网络上协同工作,一方面可以充分利用网络带宽,另一方面采用多线程开启多个连接,并发提取网页,可以有效提高网络采集效率。

3.2 主题相关度判断算法的设计

在垂直搜索过程中除了对基于主题词的超链接进行主题相关度的判断^[4],还可进一步对目标网页内容进行主题相关度的评价,将评价结果小于设定阈值的页面过滤掉,从而提高所采集主题页面的准确率.文中采取使用较多且效果较好的基于主题词的向量空间模型 VSM 算法,以此来判断网页信息与所要搜索的主题相关与否。

设查询向量 $Q = (w_{1q}, w_{2q}, \dots, w_{nq}) \in R^t$, 网页向量 $V_j = (w_{1j}, w_{2j}, \dots, w_{nj}) \in R^t$, 则相似度由下式求解所得:

$$\text{sim}(Q, V_j) = \frac{\sum_{i=1}^t (w_{iq} \times w_{ij})}{\sqrt{\sum_{i=1}^t (w_{iq})^2 \times \sum_{i=1}^t (w_{ij})^2}} \quad (1)$$

其中, w_{iq} 和 w_{ij} 分别为检索词 i 在 Q 和 V_j 中的权重; w_{iq} 和 w_{ij} 采用传统的 TF-IDF 算法^[5]来计算。

爬取主题词与目标网页的相似程度可用向量之间的夹角来度量,若向量之间的夹角越小则说明相似度越高.页面的主题相关度判别算法具体流程如下:

(1) 预处理阶段:在采集前先对基于主题的种子 Url 页面进行主题词的提取和加权,从而得到关

于该主题的特征向量及向量的权重值。

(2) 对目标网页的正文内容进行分词,并分析出特征主题词.把主题特征向量中的主题词作为主对象,按照其在目标网页正文中出现的不同位置来计算其加权频率的值。

(3) 根据公式(1)计算出目标网页与主题词的相似度,其算法如下:

算法1 目标网页内容主题相关性判定

输入: $V = \{ (v_1, w_1), (v_2, w_2), \dots, (v_n, w_n) \}$

// 目标网页特征向量

$Q = \{ (q_1, w_1), (q_2, w_2), \dots, (q_m, w_m) \}$

// 主题描述特征的查询向量

向量: T // 预设阈值

输出: S // 相似页面集合

判定算法:

$\text{Sim} = \Phi;$

for each $v \in V$ and $q \in Q$ do

 计算 $\text{Sim}(Q, V)$ // 计算向量相似度

If $\text{Sim}(Q, V) < T$ then

 Delete (V) // 不相关并丢弃网页信息

else

 Add V to S // 相关并保存

Endif

Endfor

(4) 通过相似度计算使得每个目标网页获得一个相似度 $\text{sim}(Q, V_j)$ 值.若计算的相似度大于预先设定的阈值 T ,则该页面与主题相关,并将其采集的信息存入主题网页数据库;否则不相关,忽略此页面.此外阈值的设定可以有效地控制采集结果的规模。

3.3 基于 Lucene.net 索引的创建与检索

Lucene 是 Apache 组织用 JAVA 实现开放源代码的全文检索引擎工具包^[6]. Lucene 作为一个优秀的全文检索引擎,其系统结构运用了大量面向对象的设计思想,它定义了一个与平台无关的索引文件格式,最终达成了低耦合、高效率、容易二次开发等的检索引擎系统.本系统利用的 Lucene.net 是用 C# 语言由 Lucene 移植到 .NET 平台上的版本,系统由基础结构封装、索引核心及对外接口三大部分组成,主要包括两个主要功能:一是建立索引库,将提供的数据源写入索引或者将其从索引中删除;二是向用户提供全文搜索服务,让用户可以通过主题词定位.这里使用的 Lucene 索引结构如图 3 所示。

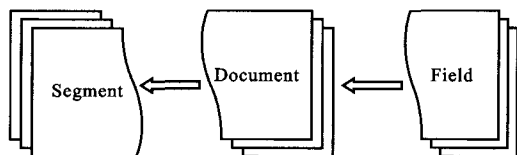


图 3 Lucene 索引结构

Lucene 在索引建立的过程中,用户首先需要按照自己的需求确定索引中 Document(文档)与 Field(字段)的组成形式,即如何确定和组织 Field。在 Field 组织好后,可以通过 Document 这个类提供的 add() 方法,将 Field 添加到 Document,然后利用 IndexWriter() 方法将构建好的 Document 添加到索引中。索引建立好以后,Document 被组织为一个一个的 segment(片段),每个 segment 代表一个完整的索引段,这样方便检索。

本系统建立索引的过程:将采集的记录从数据库读取出来;遍历记录,针对每条记录,建立一个 Lucene Document 对象。其实现的算法如下:

```
IndexWriter writer=new IndexWriter(INDEX_STORE
_PATH, new StandardAnalyzer(), true);
for (int i = 0; i < dt.Rows.Count; i++)
{Document document = new Document();
document.Add(new Field("ID", dt.Rows[i]["ID"].
ToString()) ?? "", Field.Store.YES, Field.Index. TO-
KENIZED));
// 对应 Lucene Document 中的 Field 进行添加字段,
// 存储并索引
..... // 存储并索引其它如 Title, Content, Url, Price,
// imgUrl 等字段同上
writer.AddDocument(document); //然后将该条 docu-
ment 加入到索引中
writer.Optimize(); //进行优化
write.close(); //关闭
IndexWrite; //写入索引文件
}
```

经过上面几步,即可将建立的索引信息写入到了索引文件中。

接下来设计检索器功能模块,检索器功能是检索系统中的核心问题,它首先接收用户输入的关键字,对其进行分析,然后在索引中进行查找,并将相应的检索结果返回给用户。检索器主要包含三个部分:

(1) 初始化 Lucene 的检索索引工具 IndexSearcher;

(2) 构建查询 Query,这里使用 QueryParser 解析用户的输入,通过扫描用户输入的字符串,生成

Query 对象;

(3) 搜索并处理返回结果,在构建完 Query 对象后,即可以使用前面已经初始化好的 IndexSearcher 工具来进行检索。IndexSearcher 提供了良好的检索接口,用户只需简单地将 Query 对象传入,即可以得到一个返回结果。最后利用返回结果构建了一个数据表 DataTable,将结果以列表的形式返回给用户。

通过上述关键技术实现了图 4 面向电子商务特定主题的垂直搜索引擎系统。

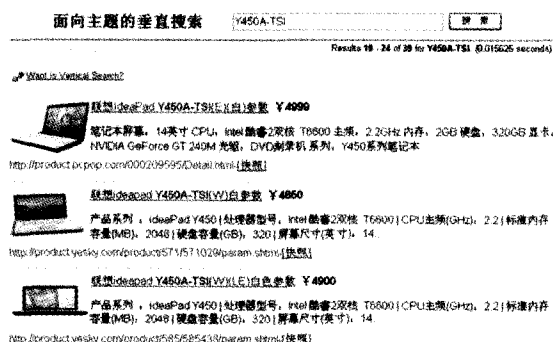


图 4 垂直搜索引擎系统搜索界面

4 垂直搜索引擎系统的性能分析

多线程爬行性能的好坏是度量网络搜索速度的重要指标^[7]。在多线程环境下,选取面向电子商务的特定主题 URL 种子如 <http://www.pconline.com.cn/> (太平洋网) 等站点进行信息采集,搜索的一级主题词如“电脑”、“手机”、“数码产品”、“办公设备”等,二级主题词“台式机”、“服务器”、“笔记本”、“一体机”等。实验采用多个线程进行同时爬行,测试结果如图 5 所示。

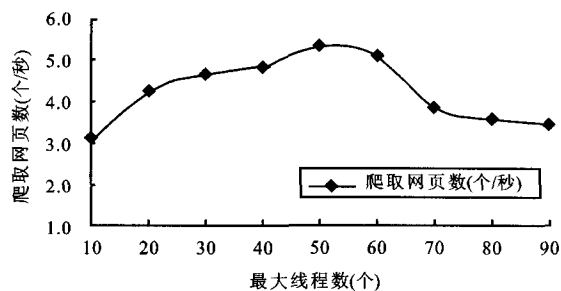


图 5 多线程爬行性能测试

从图 5 可以看出,随着线程数的增多网页爬行速度得到不断提升,但线程数目并非越多越好,当线程数到达一定数量时,网络爬行速度不升反而下降,这是因为线程数量增加后会导致大量的并发操作,

(下转第 8 页)

5 结束语

本方案对缩短 SoC 测试时间具有明显效果,可在峰值测试功耗及 TAM 宽度限制条件下帮助 SoC 测试人员寻找优化测试方案。

参考文献:

- [1] Huang Y, Reddy S M, Cheng Wutang, et al. Optimal corewrapper width selection and SOC test scheduling based on 3-D bin packing algorithm[C]// Proceedings IEEE of International Test Conference (ITC). Washington, DC, USA: IEEE Computer Society, 2002: 74-82.
- [2] Zhao D, Upadhyaya S. Dynamically partitioned test scheduling with adaptive TAM configuration for power-constrained SoC testing[J]. IEEE Trans Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2005, 24(6): 956-965.
- [3] Pouget J, Larsson E, Peng Z. Multiple-constraint driven system-on-chip test time optimization[J]. Journal of Electronic Testing, 2005, 21(6): 599-611.
- [4] Colirni A, Dorigo M, Mainieszo V. An investigation of some properties of ant algorithm[C]// Proceeding of Parallel Problem Solving form Nature Conference (PPSN). brussels, belgium: elsevier publishing, 1992: 509-520.

- [5] 崔小乐,程伟. SoC 测试访问机制和测试壳的蚁群联合优化[J]. 计算机辅助设计与图形学学报, 2009, 21(4): 461-466.
- [6] Ahn JinHo, Kang Sungho. SoC test scheduling algorithm using ACO-based rectangle packing[C]// Proceedings of the International Conference on Intelligent Computing. Berlin, Heidelberg: Springer-Verlag, 2006 (4114): 655-660.
- [7] Cui X L, Cheng W. Hierarchical SoC testing scheduling based on the ant colony algorithm[C]// Proc The IEEE 8th International Conference on ASIC. Changsha, Hunan: IEEE, 2009: 1221-1224.
- [8] Chakrabarty K, Marinissen E J. Integrated wrapper TAM co-optimization, constraint-driven test scheduling, and tester data volume reduction for SOC's[C]// Proc Design Automation Conference (DAC). New York, NY, USA: ACM, 2002: 685-690.

作者简介:

崔小乐 男, (1975-), 博士, 副教授. 研究方向为可信计算、电子系统测试验证技术、VLSI 测试与 DFT、芯片设计、软件可靠性与安全性。

程伟 男, (1984-), 硕士, 工程师. 研究方向为 VLSI 测试与 DFT。

李崇仁 男, (1946-), 博士, 教授. 研究方向为 VLSI 测试与 DFT、芯片设计、半导体器件与工艺。

(上接第 4 页)

而导致额外的资源耗费. 因此根据不同的搜索环境设置合理的线程数, 才能更好的发挥出多线程技术的网络爬行的作用。

5 结束语

本系统运用了多线程技术及基于 VSM 的主题相关度判断算法进行主题网页爬行, 并通过 Lucene. Net 提供的索引与检索技术建立系统的检索核心算法, 实现了一个面向特定主题的垂直搜索引擎应用系统. 系统通过对多个面向特定主题的大型电子商务网站进行数据采集测试. 实验结果表明, 采集入库的数据较为完整, 采集效率较高, 其检索的准确率、召回率均大大高于通用搜索引擎, 本系统的应用具有较好的实用价值和商业应用前景。

参考文献:

- [1] 刘运强. 垂直搜索引擎的研究与设计[J]. 计算机应用与软件, 2010, 127(7): 130-132.
- [2] 赵洋, 滕桂法, 张玉新. 基于 Internet 的农业信息垂直搜索

引擎的设计[J]. 河北农业大学学报, 2009, 32(6): 125-128.

- [3] 张昌年. 一种基于 VSM 的检测相似重复记录的方法[J]. 微电子学与计算机, 2008, 25(8): 184-187.
- [4] 吴金红, 张玉峰, 王翠波. 面向主题的网络竞争情报采集系统[J]. 现代图书情报技术, 2006, (12): 55-57.
- [5] 刘海峰, 姚泽清, 汪泽焱, 等. 基于位置的文本特征加权方法研究[J]. 微电子学与计算机, 2009, 26(2): 188-192.
- [6] 李刚, 宋伟, 邱哲. 征服 Ajax + Lucene 构建搜索引擎[M]. 北京: 人民邮电出版社, 2006: 156-228.
- [7] 李广丽, 刘觉夫. 垂直搜索引擎系统的研究与实现[J]. 情报杂志, 2008, 28(10): 144-147.

作者简介:

施佳 男, (1973-), 博士研究生, 副教授. 研究方向为网络数据库、复杂网络。

王恒山 男, (1948-), 博士生导师, 教授. 研究方向为复杂网络、管理信息系统。

肖仰华 男, (1980-), 博士, 讲师. 研究方向为图数据库、复杂网络。

丁卫平 男, (1979-), 博士研究生, 讲师. 研究方向为粗糙集、数据挖掘。