文章编号: 1673-1549(2011)01-0071-03

化工行业信息搜索技术的研究

刘强国[1], 高媛媛[1], 左由兵[2]

(1.四川理工学院理学院, 四川 自贡 643000; 2.四川理工学院材料与化学工程学院, 四川 自贡 643000)

摘　要: 面向行业主题的搜索在特定主题信息覆盖方面与通用搜索引擎有着截然不同的要求, 为解决行业信息搜索的问题对基于向量空间算法的化工相关度计算以及对经典的 Page-Rank页面排序算法做了研究与改进并且在 Nutch搜索引擎架构基础上, 搭建了一个面向化工行业信息资源的垂直搜索引擎。相对于通用搜索引擎来说剔除掉了不必要的搜索结果信息量, 提升了系统速度, 提高了行业信息搜索的准确度。

关键词: Nutch; 主题过滤; 页面排序

中图分类号: TP391.3　　文献标识码: A

通过运用垂直搜索引擎技术, 化工行业科研人员、技术人员可以快速有效的寻找到适合他们的个性化的化工行业知识资源和信息, 化工行业信息垂直搜索引擎不失为解决这些问题的有力工具。

Nutch[1]作为 Apache的一个开源项目, 具备搜索引擎的基本功能, 并采用网页本身的价值进行排序的算法, 另外, Nutch具有灵活而强大的插件系统, 因此可以利用 Nutch方便的搭建化工行业垂直搜索引擎[2-3]。

1　Nutch的工作原理

Nutch的工作分为两个阶段: 抓取和搜索。抓取阶段取得网页并把他们处理成倒排的索引, 后面搜索阶段的工作基于这些索引来进行, 如图 1所示。

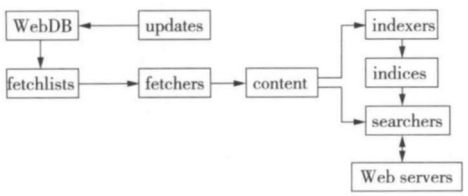

图 1　Nutch的结构

WebDB是跟踪每个已知页面及其相关联接的持久化的特定数据库。该库保存了两种类型的实体: pages和 links。page实体代表 web中的某个页面, 用 URL地址及页面内容的 MD5值来标示, 也包括页面的链接数(被称为出链)、抓取信息(页面最后一次被抓取的时间)和页面的得分(页面的重要程度)等信息。link实体代表从一个 Web页面(称为源)到另一个 Web页面(目标页面)的链接。在 WebDB的 web图中, 节点代表页面, 边代表链接。其中 page实体的成员变量如图 2所示, link实体的成员变量如图 3所示。

SF CUR_VERSION : byte
SF DEFAULT_INTERVAL : byte
□ url : UTF8
□ md5 : MD5Hash
□ nextFetch : long
□ retries : byte
□ fetchInterval : byte
□ numOutlinks : int
□ score : float
□ nextScore : float

图 2　page 实体的成员变量

SF MAX_ANCHOR_LENGTH : int
SF VERSION_1 : byte
SF VERSION_2 : byte
SF CUR_VERSION : byte
□ fromID : MD5Hash
□ url : UTF8
□ domainID : long
□ anchor : UTF8
□ targetHasOutlink : boolean

图 3　link 实体的成员变量

fetch lists是由 WebDB产生的 URL地址列表, 该表

收稿日期: 2010-09-21

基金项目: 四川理工学院人才引进科研启动项目 (07ZR41)

作者简介: 刘强国(1980-), 男, 四川成都人, 助教, 硕士, 主要从事网络计算技术方面的研究。

保存了将用于爬行的 URL 地址。

fetchers取出 fetchlists中的地址, 并从 Internet中下载相应的内容。

updates根据最后一次抓取的页面在 WebDB中出现与否, 更新 WebDB库。

content用于生成用户实际查询时所需要的索引信息。

indexers利用 content产生包含所有项及页面的倒排索引。

Web servers处理用户与搜索结果之间的交互。

整个的工作流程如下:

开始的时候向空的 WebDB 中写入 URL地址集合; 接下来生成将用于抓取的 URL地址列表的 fechlists, 然后根据 fechlists中的 URL信息进行爬行处理, 下载相应的网页内容, 这些内容在随后(或同时)被进行解析处理, 即从下载的网页内容中抽取文本及链接信息; 下一步把抽取的链接信息写入 WebDB库以便在下一个间隔时间到来时进行抓取; 同时从抽取的文本内容中建立倒排索引; 最后 searcher及 Web servers提供与索引交互的接口, 并返回用户的查询结果集。

2 化工行业信息搜索引擎研究与设计

对于特定的化工行业信息搜索来说, 爬行的对象将局限在某些固定的网站, 在此我们对基于 Nutch搜索平台的插件[4]进行二次开发, 对 Nutch的索引、评分功能进行扩展。在索引时首先对主题词进行过滤, 与主题无关或者权重较轻的关键字和网页滤过。评分过程, 则针对特定主题建立一个专业词库, 对词库中相关性较大的查询词语赋以较大的权值。因为 Nutch本身只是个搜索平台不针对任何特定行业信息, 故需要开发出主题过滤器和进行行业相关度重新排序。在本文中我们将该系统分为化工行业信息抓取、化工行业信息处理和化工行业信息检索 3个单元, 在每个单元中均须进行相应的行业信息过滤和相关度处理, 如图 4所示。

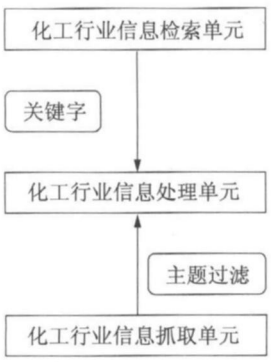

图 4 化工行业信息搜索模块

2.1 化工行业信息的抓取

化工行业信息抓取单元负责搜集与化工行业信息相关的网页, 由爬虫、WebDB、页面分析器、URL更新器和化工主题信息过滤器构成, 主要任务是化工行业信息的抓取、页面分析、URL主题过滤和页面主题过滤。爬虫通过用户初始注入的化工行业网站的 URL, 按照深度遍历的方式抓取所有网页和链接并且存储到 WebDB中; 页面分析器进行语法、词法分析, 剔除语法标记, 去掉重复 URL, 解析出网页内容和新的 URL, 并将解析出来的新 URL和页面信息送入化工行业信息过滤器进行过滤; 化工行业主题过滤器用于过滤掉与化工行业信息无关的网页和 URL; 更新器则根据过滤后的 URL更新 WebDB来进行下一轮的抓取工作。

Nutch架构本身性能良好, 但不具备专业主题过滤功能, 对此我们设计了化工行业主题方面的过滤, 主要是进行 URL和页面内容过滤两部分。在采集到页面时, 提取出链接, 根据链接和专业词典进行综合判断, 将无关链接过滤掉并将相关链接保存到 URL爬行库里; 对于化工行业主题内容的过滤, 对页面内容中的标题、摘要和总结提取出来给予高权重。

2.2 化工行业信息处理

化工行业信息抓取单元完成了对化工行业网站和 URL进行抓取和过滤; 而化工行业信息处理单元则对抓取的网页进行相关性分析, 根据抓取单元中经过分析和过滤的化工行业关键词建立倒排索引。而化工行业关键词的识别是整个搜索中十分重要的一部分, 通过关键词字典文件对网页内的词进行分析和处理, 系统所采用的化工行业关键词词典是由化工专家和化工分类专家所交融的领域所有关键词的集合。

在对网页进行解析时, 首先将其转换为纯文本信息, 然后按照字符串匹配, 进行中文分词并结合语义识别, 合理的在中文语句中提取出与系统化工行业词典匹配的关键词。

化工行业信息处理单元, 在整个化工行业信息搜索引擎中起到承上启下的关键作用。而 Nutch提供了良好的索引建立机制, 并以主流的倒排索引方式建立索引表, 而且也有 URL过滤器和中文分词, 但是需要构建化工行业词典并进行相关性分析, 使化工行业信息搜索可以建立在 Nutch之上。

2.3 化工行业信息检索

化工行业信息检索单元是用户输入查询条件, 经过中文分词, 转换为关键词, 然后按照关键词找到满足的网页, 通过一定的算法设计进行网页评分, 显示给用户相关网站信息。Nutch的评分标准是以链接入页面的链接数为依据所算出来的; 它的排序算法是借鉴 Google的 pagerank[5-6]算法, 但不进行竞价排序, 而是按照网页重要程度[7]来计算的。而本系统的排序算法按照化工行业关键词加权后进行综合排序, 将查询结果返回给用户。

3 化工行业信息搜索引擎实现的关键技术

3.1 主题过滤技术

根据化工主题, 爬虫在爬行的页面内容和链接中进行过滤。具体分为 URL化工主题过滤和页面化工主题过滤[8]两个主要技术。URL链接主题过滤方面, 将网络

爬虫爬到的链接进行分析, 去除 Http、WWW、com 和 cn 等前后缀, 从中提取出结构关键字, 通过化工词典进行 URL过滤 (例如: 认为含有 chemistry的链接既是与化工相关), 将与化工低相关、无关的 URL剔除掉进行初级化工主题过滤。仅仅对 URLL进行判断从而进行化工主题方面的简单过滤。但是大多数的与主题不相关的网页则是内容不相关。系统通过空间向量算法和化工词典相结合的算法进行化工内容方面的过滤, 具体算法如下:

(1)首先进行预处理, 将爬虫爬到的网页内容进行结构化提取, 从中提前出网页的题目、摘要和正文等。

(2)对提取出的题目、摘要再进行分词。去除助词、副词等虚词, 保留关键词, 并构造文档向量。

(3)将关键字对照化工词典, 根据关键词出现的频率进行关键词加权, 并根据词典构造词典向量。

(4)利用文档向量和词典向量进行相关度计算。

(5)计算后的相关度与阈值 Y 进行比较, 小于阈值就过滤掉, 反之保存到页面库中。

(6)同样方法用在正文中, 但关键词的权重比标题和摘要要小一些, 但是阈值保持不变。

3.2 化工主题页面排序技术

目前主流的页面排序算法[9]都是基于链接的, 主要有: pagerank算法和 HITs算法。但是 Pagerank算法有严重的"主题漂移"[10]现象; 而 HITs算法则存在"主题扩大化"[10]现象。上述两种算法都不适宜于直接用于化工网页排序, 系统采用的是综合 pagerank和 HITs算法面向化工主题的综合排序算法[11]; 利用 pagerank算法主要思想既"关键网页所链接的肯定还是关键网页", 然后通过网页中的关键词对照化工词典进行权重计算, 再按照与化工主题和用户输入词的相关度, 综合进行页面排序。

4 结束语

化工行业信息搜索技术是当前国内外研究的热点, 如何高效准确的获取网络上浩瀚的化工行业信息资源, 对于提升化工行业信息的使用效率有着深刻的意义。本文在开源项目 Nutch的二次开发基础上, 对化工行业信息主题过滤、页面排序算法做出了改进, 构建了化工行业信息垂直搜索引擎。

参考文献:

[1] Doug Cutting Internet Archive, Nutch: an Open-Source Platform for Web Search [EB/OL]. http://wiki.apache.org/nutch/NutchTutorial, 2006-08-02/2006-12-03.

[2] 欧阳柳波, 李学勇, 李国徽, 等. 专业搜索引擎搜索策略综述 [J]. 计算机工程, 2004(13): 32-33.

[3] 肖 亮. 垂直搜索引擎的研究与实现 [D]. 北京: 北京交通大学, 2008.

[4] 吴敏琦, 岳伟. 基于 Nutch的 XML网站全文搜索引擎实现 [J]. 计算机工程, 2008(15): 95-96.

[5] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the Web[C]//Google technical report Conference, Stanford Infolab, june 1998, 2: 96-98.

[6] Taher H. Haveliwala. Efficient computing of PageRank[C]//Stanford Database Group Technical Report, Stanford Infolab, Jan. 1999, 5: 101-104.

[7] Bharat K, Henzinger M R. Improved Algorithms for Topic Distillation in a Hyperlinked Environment [C]//In Proc. of {SIGIR}-98, 21st {ACM} International Conference on Research and Development in Information Retrieval, 1998, 7: 326-327.

[8] Chakrabarti S, Dom B, Gibson D, et al. Experiments in topic distillation[C]//Proc ACM SIGIR work shop on Hypertext Information Retrieval on the Web, 1998, 9: 1056-1058.

[9] 吴明礼, 施水才. 一种结合超链接分析的搜索引擎排序方法 [J]. 计算机工程, 2004(15): 144-145.

[10] 刘强国, 左志宏, 董祥千. 基于 web超链接分析算法研究综述 [J]. 计算机应用研究, 2007(7): 1324-1325.

[11] 刘强国. 主题搜索引擎设计与研究 [D]. 成都: 电子科技大学, 2007.

Research on Chemical Industry Information Search Technology

LIU Qiang-guo[1], *GAO Yuan-yuan*[1], *ZUO You-bing*[2]

(1. School of Science, Sichuan University of Science & Engineering, Zigong 643000, China;
2. School of Material and Chemical Engineering, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract The demand between the general search engine and the professional information search is mainly on the coverage of special topic information is huge different, to solving the problem which the professional information searching encountered, this paper study and give improvement on the chemical industry topic co-relation value computation based on the vector-space algorithm and the classic webpage ranking algorithm of Page-Rank, and build a vertical search engine based on the framework of Nutch. Compared to the general search engine, eliminating the unnecessary search results, improving the search system speed and the accuracy of professional information search.

Key words Nutch; topic distillation; webpage ranking