基于 Lucene 网页排序算法的改进①

Improvement of an Algorithm for Ranking Pages Based on Lucene

张 贤 周 娅 (桂林电子科技大学 计算机与控制学院 广西 桂林 541004)

摘 要: 在分析现有的词频位置加权排序法、Direct Hit 算法、PageRank 算法和 Lucene 的网页排序算法后，将这三种著名的算法思想运用到 Lucene 的网页排序算法中，并设计了一个基于 Lucene 的糖业专业搜索引擎，重点介绍该搜索引擎的检索功能。最后，通过在所设计好的糖业专业搜索引擎进行实验，验证改进后 Lucene 的网页排序算法，实验结果表明改进后的排序算法能够提高检索结果的质量，能够更准确地将结果信息反馈给用户。

关键词: 网页排序算法 搜索引擎 Lucene 检索功能 糖业专业搜索引擎

随着网络信息资源日益丰富，用户进行检索时，往往能搜索到成千上万的检索结果，用户希望最符合他们需求的资源能够在检索结果中前面。“据估计近 85％的用户只浏览搜索引擎返回的第一页结果[1]”如果重要的网页资源被排在检索结果的后面，被用户的点击率就很小很小。Rope Starch 调查指出，71％的用户在检索信息时受挫，绝大部分(86％)的 Internet 用户感到应当出现更有效的、准确的信息检索技术[2]。因此排序是搜索引擎的关键技术之一，排序算法决定了排序效果的优劣。

Lucene[3]是 Apache 软件基金会 Jakarta 项目组的子项目,是一个开放源代码的全文检索工具包，能够非常方便地为各种应用程序加入全文索引和搜索功能。尽管 Lucene 自问世以来就得到了开源代码社群的巨大反响，但仍存在一定的不足，Lucene 对检索结果的相关排序方面还有待完善[4]。因此，本文主要是通过将几种著名的排序算法的思想融入到 Lucene 网页排序算法，在 Lucene 原有的网页排序的基础上进行改进，使其更能满足用户的检索需求。

1 Lucene网页排序算法

Lucene 网页排序算法是搜索引擎的核心,其好坏直接影响搜索结果。Lucene 使用一个评分函数来计算某个查询 q 的文档 d 得分情况[5]，用公式（1）表示：

$$Score(d)=\sum_{t\_in\_q} tf(t\_in\_d)*idf\_t*boost(t.field\_in\_d)*lengthNorm(t.field\_in\_d)$$

其中各项参数说明如下：

①Score*(d):文档 d 的页面优先度得分。

② $\sum_{t\_in\_q}$:表示对所有的查询词 t 在查询 q 中计算合计值。

③tf（t_in_d）:tf（全称为 Term Frequency，词条频率），表示检索的词条 t 在某个文档 d 中总共出现次数。在 Lucene 中，该值是真实频率的平方根值。

④ idf_t:idf(全称为 Inversed Document Frequency，表示逆文档频率)，表示有多少个文档 d 包含该查询词 t。计算 idf_t 的常用方法： 公式(2)

$$idf\_t=\log_2(numDocs/docFreq\_t+1)+1$$

numDocs：已建索引中总共的页面文档 d 数量；

docFreq_t：包含查询词 t 的页面文档总数。

⑤oost(t.field_in_d)：boost 指在建立索引时，对每个 Field 设置的一个激励因子。默认值为 1。增加查询词 t 所在域（Field）的重要性，也就增加了文档 d 的重要性，换言之，重要性的增加就是指增加文档 d 的得分。

⑥engthNorm(t.field_in_d)：考虑文档的大小，如果文档越长，那么这个因子的值就越低，反之越高。

① 基金项目:广西青年科学基金(桂科青 0832101)
收稿时间:2008-08-03

之后将这三种分值计算出后，再相乘，就得到每个文档得分。最终 Lucene 网页排序结果是按文档得分进行自然相关度排序。

Lucene 网页排序算法主要特点为:所查询的词在一个文档中位置并不重要；若一个文档中所含该查询词的次数越多，则其得分越高；一个命中的文档中，如果除了该查询词之外，其它词越多，则其得分越少。缺点为：精确度不高；不能充分体现网页的重要性。

2 已有的网页排序算法

目前比较著名的几种网页排序算法有词频位置加权排序法、Direct Hit 算法、PageRank 算法等。

词频位置加权排序法[6]是通过查询关键词在页面中出现的频率，位置来给网页评级。比如说，查询关键词若出现在标题中，权重为 10；若出现在正文中，权重为 5；若为粗体字，权重为 1 等。这种算法的优点是简单、易实现。缺点是过于依赖词的重要性，不能保证页面的质量。

Direct Hit 是 Ask Jeeves 公司的一种注重信息质量和用户行为反馈的排序算法[7]。用户输入要检索的关键词后，在浏览搜索引擎提供的结果记录上的点击率和停留时间的长短。停留时间越长，说明这条记录与关键词越相关，反之则说明与这条记录与关键词的相关度很小。这种算法优点是能够帮助用户节省大量的时间，可通过阅读检索结果筛选初更符合要求的结果进行浏览。缺点是这种算法并没有进行排序，只是一种筛选和抽取，用户不可能一一查看搜索结果。

Google 是现今最受欢迎的搜索引擎,最主要的原因是优秀的排序结果[8]。PageRank 是其中最重要的因素。PageRank 算法[9]的基本思想是基于“从许多优质的网页链接过来的网页，必定还是优质的网页”的回归关系，来判定所有网页的重要性。也就是说，只要一个网页的 PageRank 值很高，它在搜索结果中的位置就会很靠前，这种方式突破以往的仅靠相关度排序的局限性，使得排序结果更加能令用户满意。PageRank 算法简要介绍如下：公式(3)

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + ... + PR(T_n)/C(T_n))$$

其中参数分别表示：

①PR(A)：表示页面 A 的网页级别。

②PR(Ti)：表示页面 Ti 的网页级别，页面 Ti 链向页面 A，其中 1≤i≤n。

③C(Ti)：表示页面 Ti 链出的链接数量。

④d：表示阻尼系数，取值在 0－1 之间，Google 通常取值 0.85。

PageRank 算法优点是注重信息的质量，只有质量高的网页才能在 PageRank 算法中获得高分，而 Google 就是利用这个原理客观地计算出 Internet 上网页的权威性相对值，也就是网页级别。

3 改进后的Lucene网页排序算法

针对 Lucene 网页排序算法中的不足，将上面的几种著名的排序算法的思想融入到 Lucene 网页排序算法中。具体改进后的 Lucene 网页排序算法应当考虑的其它因素有：

(1)关键词词频与位置关系

通常查询关键词在一般正文中出现的机会最大(即不是在标题、居中、URL 等处)，而对在其它位置出现的情况，根据所处位置决定其权值大小。若关键词在标题处出现，则它出现在正文中的频率也很高。因为通常情况下正文都是围绕标题展开，标题可认为是对正文的概括。若关键词出现在链接文字处的网页，说明该链接的网页包含了关键词的信息。若出现在文档居中的位置，那通常是文档某段正文内容的概括，相当于标题，则也说明是很重要的。根据表 1 可以对关键词出现的位置相应的赋予一定的权值。

表 1　关键词词频位置关系的权值参照表

关键词位置	权值	关键词位置	权值
外部链接	10	每句开头	1.5
标题	10	加粗或斜体	1
域名	7	文本用法	1
H1、H2号	5	Title属性	1
每段句首	5	Alt属性	0.5
路径或文件名	4	Meta描述	0.5

(2)用户的行为特征

对于用户搜索的关键词的返回结果上，用户点击率高的网页，说明它与用户的相关度较高，排序时应当排在较前的位置上。

网页的平均访问时间，也是用户的一个特征表现。用户在浏览网页页面时，若用户找到自己满意的页面，会花较长的时间浏览该网页，相反则会较快地跳过该

网页。对于一些 PageRank 值较高的网页，它的平均访问时间也会较长，PageRank 值较低的网页，平均访问时间相对较短。所以可以通过平均访问页面的时间来决定网页的质量的高低，而对于网页的点击率和平均访问时间可以通过事务日志来记录。

(3)网页文档的链接关系

PageRank 算法只考虑了链入的网页，要想提高网页的 PageRank 值，就必须满足：

①网页自身的等级高，比如说一些大型的网站或者是门户网站等。

②有较多链入的网页，也就是有多个外部网页页面指向该网页。链入链接的越多说明该网页页面被引用的次数越多，也就是说它的网页质量较高，所以其 PR 值也较高。

③有 PR 值高的网页链入到该网页，也就是有 PR 值高的网页页面指向该网页。能够被等级高的 PR 值网页引用，说明该网页的质量较高，所以其 PR 值也较高。

对于那些有大量链出的网页，或者 PR 值低的网页链入到网页，它的 PR 值也仍然是很低的。

(4)网页的响应时间

网页的响应时间是评价搜索引擎的重要指标，同时，它对网页的排序得分也有影响。用户一般只会花几秒钟的时间等待一个网页的打开，如果网页打开的时间过长，用户往往不会等，而是放弃。所以对于相应时间过慢的网页，应当设置它的权值为负的。

(5)网站的刷新频率

网站的刷新频率是指某个网站最近两次变动刷新的间隔时间。

综合上述因素，改进后的 Lucene 网页排序算法的评分函数表示为：公式(4)

Score(d)=d_score*30%+d_key*30%+d_user*10%+d_PageRank*20%+d_recall*5%+d_refresh*5%

其中各项参数分别表示：

①Score(d):表示文档 d 的页面优先度得分。

②d_score:表示先前在 Lucene 网页排序算法中计算出的文档 d 的页面优先度得分,其权值设置为 30％。

③d_key:表示关键词词频与位置的因素值，其权值设置为 30％。

④d_user:表示网页的受欢迎程度,但由于用户的行为可能受随意性的影响，权值为 10％。

⑤d_PageRank:表示网页的 PageRank 值，PageRank 值根据公式（3）可计算得，其权值设置为 20％。

⑥d_recall：表示网页的响应时间。如果响应时间过长超出用户等待时间,则可认为其值为负值，也就是减去这项值，其权值为 5％。

⑦d_refresh:表示网站的刷新时间，反应网站的资源是否具有时效性，其权值为 5％。

4　糖业专业搜索引擎的设计实现

糖业专业搜索引擎不同于通用的搜索引擎，它是面向一个专业领域，目的是为用户在糖业领域内查找信息、产品分类及行业市场动态发展行情时，旨在 Lucene 基础上设计实现糖业搜索引擎,并通过改进后的 Lucene 网页排序算法为用户准确快速地提取信息。从抓取网页原理及体系结构来看，糖业专业搜索引擎所搜索是有针对性地选择一些站点进行抓取，而不像通用搜索引擎是遍布整个 Web，所以可以大大减少系统的网络开销，同时也使系统设计和实现难度降低。

糖业专业搜索引擎一般包括 Web 文档提取、文档预处理、中文分词、信息索引、信息检索以及用户界面等几个重要模块,图 1 给出了整个系统实现框架图。

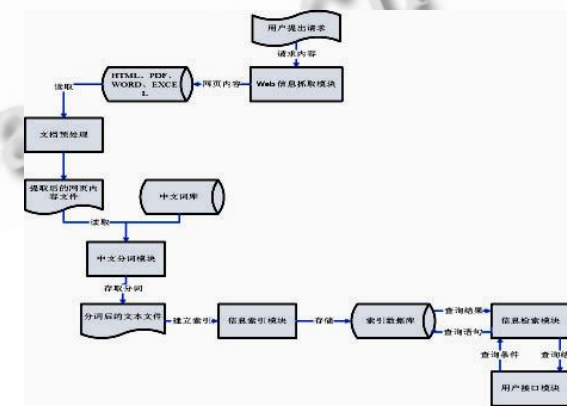

图 1　糖业专业搜索引擎系统实现框架图

在 Lucene 基础上设计实现的糖业专业搜索引擎，Lucene 提供的信息检索接口是 IndexSearcher 类，由 Query 类封装查询词、字段和语言分析器，交由 IndexSearcher 查询。由于影响排序结果的只有 Lucene 索引中存储的 2 个参数：DocID 和 score，所以查询结果使用改进后的 Lucene 网页排序算法，

需要修改 Lucene 中的 org apache lucene index 模块的 IndexSearcher.java 中的 HitCollector 过程，然后用户再通过 Hit 对象访问 Document→Field 的内容即可。

5 实验分析

为了测试改进后的基于 Lucene 网页排序算法的效果，本文在所设计的糖业专业搜索引擎的基础上进行 Lucene 网页排序算法改进前后性能测试。首先用设计好的糖业专业搜索引擎在中国化工网上进行搜索关键词为“甘蔗”的网页，选取前面的 30 个网页。用改进前和改进后的算法进行测试，将上述影响网页排序结果的因素考虑进来后，并经过统计分析算法改进前的第一页中的 10 条记录的页面情况，分析如下，改进后的排序结果如表 2 所示：

表 2　改进前后网页排序结果对照表

改进前排名	改进后排名	相关链接数	查询词是否出现在特殊位置（T/F）	网页是否及时相应（T/F）	网页是否及时刷新（T/F）	用户点击率
1	1	45	T	T	T	1520
2	3	20	T	T	T	890
3	5	42	T	T	T	1210
4	7	13	F	F	F	630
5	2	35	T	T	T	1114
6	9	4	F	F	F	140
7	4	30	T	F	T	1052
8	11	2	F	F	F	230
9	14	20	T	T	T	753
10	17	4	F	F	F	102

改进后的基于 Lucene 网页排序算法可以使一些 PR 值高的页面、用户点击率高的页面、网页响应时间快、关键词出现的位置占重要比例的一些网页尽可能排在第一页，便于用户在第一时间就能找到需要的信息。排序结果相比之前有了很大地改善，达到设计的预期目的。

6 结束语

Lucene 作为开源全文信息检索包，越来越多的应用程序通过它来提供搜索功能。本文在 Lucene 基础上设计实现糖业专业搜索引擎，并对 Lucene 的网页排序算法进行改进，除先前自身所考虑的因素外，还考虑网页的 PageRank 值、关键词词频与位置关系、网页响应时间快、用户的行为特征等。实验表明改进后的基于 Lucene 网页排序算法确实能够提高搜索引擎的准确性和重要性。

参考文献

1 Lan Huang. A Survery on Web information Retrieval Technologies. State University of New York, Department of Computer Science ESCL Technical Report TR-120. 2000－02.

2 周晋等.搜索引擎输入方式的研究.计算机科学.2002,29(8):9－12.

3 Hatcher E, Gospodnetic O. Lucene in Action. [2005]. http:// lucenebook.com/.

4 http://www.0240.cn/html/2006/0104/30042.html.

5 邱哲,符滔滔.开发自己的搜索引擎 Lucene2.0+ Heritrix.北京:人民邮电出版社,2007:105－108.

6 Atsushi Matsumura. Effect of relationships between words on Japanese information retrieval. ACM Transactions on Asian Language Information Proeessing. 2006:264－289.

7 M.M. Sufyan Beg. A subjective measure of web search quality. Information Science—— Informatics and Computer Science:An International Journal. 2005: 365－38l.

8 曹军.Google 的 PageRank 技术剖析.情报杂志,2002,(10):15－18.

9 Matthew Richardson. Beyond PageRank: machine learning for static ranking. International World Wide Web Conference. 2006:707－715.