

# 基于锚文本相似度的链接算法

刘菁菁, 董 静, 林鸿飞, 叶 正

(大连理工大学计算机科学与工程系 辽宁 大连 116024)

**摘要:** 对链接算法在搜索引擎检索结果排序中的应用进行研究, 提出基于 PageRank 和锚文本对检索结果进行二次排序, 合理评价网页重要程度. 实验结果表明, 该方法在一定程度上能提高检索效果.

**关键词:** 链接分析; PageRank; 锚文本; 相似度

**中图分类号:** TP 391

**文章编号:** 1671-6841(2007)02-0096-04

## 0 引言

当从 WWW 上查找某相关知识时, 通常只输入查询信息, 在短时间内搜索引擎就会返回不计其数的检索结果. 然而, 有时候用户阅读十几页都无法得到满意结果. 导致该问题的主要原因是文档的索引数目增加了好几个数量级, 但是用户能够看的文档数却没有增加<sup>[1]</sup>. 用户希望仅仅阅读结果的前面几个就能得到满意答案, 因此有必要研究网页排序方法, 给出合理查询结果.

按传统方式返回的检索结果在一定程度上不能很好满足用户需求, 本文利用 PageRank 和锚文本对初次查询结果进行二次排序, 并考虑三方面因素: 网页内容和查询词相似度、基于链接结构的 PageRank 分数以及锚文本与查询词相似度. 先按传统算法和 PageRank 技术初次查询, 然后利用锚文本再次排序, 进行同等与不等对待锚文本两种测试, 结果表明该方法能优化检索结果.

## 1 链接分析算法

### 1.1 引入链接分析可行性研究

超链接分析思想是利用 Web 网络的结构特征评价网页质量<sup>[2]</sup>. 任何一篇网页若是被多个网页所指向, 则该网页内容应该也是比较重要和权威的、或者有用的<sup>[1]</sup>. 该思想是引文分析法在网络应用中的体现. 引文分析法的主要依据和研究内容是科学文献之间的引证与被引证关系<sup>[3]</sup>, 认为被引用较多的文章权威性较强. 网页间链接与被链接关系类似于文献引证与被引证关系, 网页自身的“入度”可作为衡量网页重要性的一种非常有意义的指标<sup>[4]</sup>.

### 1.2 PageRank 技术

PageRank<sup>[5]</sup>技术是 Google 创始人拉里·佩奇和谢尔盖·布林提出来的. 在互联网上, 如果一个网页被很多其他网页所链接, 说明它受到普遍的承认和信赖, 排名就高. PageRank 按“被引用多的论文权威性较强”的思想, 用客观方式判定网页的相对重要性, 其计算公式为

$$PR(A) = (1-d)/N + d(PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n)) \quad (1)$$

其中,  $PR(A)$  代表网页  $A$  的 PageRank;  $d$  代表“随机冲浪”中沿链接访问网页的平均次数<sup>[2]</sup>, 经验值一般取 0.85,  $PR(T_i)$  指链接到  $A$  的网页  $T_i$  的 PageRank;  $C(T_i)$  代表网页  $T_i$  出链数.

收稿日期: 2006-11-15

基金项目: 国家自然科学基金资助项目, 编号 60373095, 60673039; 国家高科技 863 计划, 编号 2006AA01Z151.

作者简介: 刘菁菁(1982-), 女, 硕士研究生, 主要从事搜索引擎研究; 通讯作者: 林鸿飞(1962-), 男, 教授, 博士生导师, 主要从事文本挖掘和自然语言理解研究.

### 1.3 锚文本

由于网页的超链接特征,近几年提出了许多超链接分析算法来改进网页检索的性能,但是根据最近几年 TREC 的 Web Track 测试的评测结果,表明过度使用超链接分析算法往往会适得其反.原因在于:虽然网页间链接与被链接关系与科技文献间引证与被引证关系很类似,但网页在质量、用法、引用和长度上都与科技文献存在较大区别.于是,超链接的另一类信息——锚文本(anchor text)便引起了人们的关注<sup>[6]</sup>.

所谓的锚文本就是描述链接所指向网页主题信息的文字内容,对于总结网页内容和性质有着重要意义.例如:在某网页中有 `<a href="http://www.dlut.edu.cn">Dalian University of Technology</a>`,则 Dalian University of Technology 就是概括链接 `http://www.dlut.edu.cn` 所指向网页主题内容的锚文本.其优点为:①内容能比网页本身更精确地概述网页主题信息;②对链向网页的宣传有助于搜索非文本信息,如图像、程序和数据库等.

此外,锚文本还具有客观性.在某种程度上,不同作者对某特定主题的理解客观性强于某一个作者的理解.

### 1.4 其他链接分析算法的研究

文献[7]将超链接分析技术和网页的锚文本相关性计算结合,提出 CALA 算法,认为查询返回结果中排名高的网页除了需要质量较高之外,还需锚文本与查询词有较高相似度,如下式所示:

$$\text{CALA}(d) = \text{PR}(d) \cdot \{1 + \delta \cdot \text{sim}(d, Q)\} \quad (2)$$

其中,  $\text{PR}(d)$  为文档  $d$  的 PageRank 值;  $\text{sim}(d, Q)$  为利用锚文本计算的文档  $d$  和查询  $Q$  的相似度;  $\delta$  为调整参数.

文献[6]提出一种基于源网页质量的锚文本相似度计算方法——LAAT 算法. LAAT 算法利用网页 PageRank 值来判断各源网页上锚文本的客观性,认为质量越高即 PageRank 值越高的源网页,其锚文本的可靠程度也就越高,在计算锚文本相似度时,应侧重于这些源网页上的锚文本.

### 1.5 主要算法思想

主要算法思想如下:锚文本与用户查询的相似度是可见的,利用链接结构得到的 PageRank 值完全脱离用户查询,因此锚文本相似度优于基于结构的权威性.结构上的权威性又可为锚文本相似性提供一定参考,二者可以互相补充,综合提高网页排序的合理性.

锚文本之所以能提高网页检索的性能,是由于同一网页的锚文本往往分布在不同作者编写的网页上,因此在一定程度上比一般的文本更客观.然而,这种客观是相对的,不同作者针对同一张网页的理解不一定一致,这样编写的锚文本并不都是准确无误和高质量的.例如,让网页工程师制作关于高能物理的网页,他可能对网页内容概括得不太准确<sup>[6]</sup>.

PageRank 计算仅依赖链接数目,对于那些刚刚发布不久的网页而言,可能会因宣传不足或其他原因,得不到太多链接而被忽略.假设网页 A 的 PageRank 是 0.9,网页 B 的是 0.2,A 的锚文本与查询词相似度是 0.1,网页 B 的是 0.9,显然网页 B 与查询词相似度更高一点.因此,锚文本与查询词相似度被应用到网页排序中是合理的.锚文本具有的主观性可用 PageRank 弥补,一般认为,PageRank 值高的网页做出的理解大部分都是准确合理的.

## 2 实验与结果分析

### 2.1 评测标准及过程

实验使用的是 SEWM2006 中文 Web 信息检索评测 20G 的数据集、查询集以及结果集,并采用天网已公布的评测工具进行分析与评测.该数据集共包括 3 748 292 个网页,消重后索引网页数为 3 735 219 个,共抽取链接 73 499 436 个.使用 NPHP 查询集,包括 300 个 Topic,并提供了 34 个 Topic 答案.评测标准采用 MRR 值,即第一个正确答案出现位置的倒数平均数.

现在的网页排序基本上都采用综合排序方法,文档与查询的相关度来自于两部分贡献:一是常规相关度贡献;二是超链接分析计算出的网页质量.而常规相关度贡献又来自于两个方面:查询词语的出现次数和类型,以及查询词语之间的距离匹配,这两部分贡献之和为常规相关度<sup>[2]</sup>.按上述思想来考虑多种排序因素.

本实验利用开源 Java 实现的搜索引擎 Nutch 作为工具,对语料进行相应处理,具体过程如下:

- (1) 封装语料,将语料封装成 Nutch 可处理格式,为后期工作做准备;
- (2) 链接分析,抽取网页间链接和锚文本,计算 PageRank;
- (3) 建立索引时,采用二字串分词并索引,未做去噪处理;
- (4) 查询时,将 NPHP 查询集中的 Title 部分生成查询词,未做查询扩展处理,对每个查询词最多检索出 100 篇文档作为答案集,再进行二次检索,得到 50 篇文档作为最终提交结果;
- (5) 采用 NPHP 结果集和评测工具计算 MRR.

## 2.2 结果分析

### 2.2.1 实验 1:锚文本有无以及是否被同等对待对 MRR 值的影响

首先,输入查询词,按内容和查询词相似度,且考虑 PageRank 值对结果的影响,得到初步查询结果集,然后进行二次检索,考虑锚文本对检索效果的影响.计算其与查询词相似度时,进行两种计算:一是简单收集来源于不同网页锚文本;二是区别对待锚文本,用锚文本来源网页的 PageRank 值作评价锚文本质量的标准.实验证明,采用锚文本可提高仅考虑 PageRank 的检索效果,若再将 PageRank 对锚文本质量影响因素加以考虑,还可再次提高 MRR 值,如图 1 所示.由结果分析知,锚文本应用可在一定程度上弥补 PageRank 的不足.利用 PageRank 作为衡量锚文本质量的标准,可以减少由于锚文本错误描述网页主题对检索效果的影响,从而提高 MRR 值.

### 2.2.2 实验 2:锚文本与查询词相似度所占权重对 MRR 值的影响

在进行二次检索过程中,调整锚文本与查询词相似度所占权重,从而获得 MRR 值随着锚文本与查询词相似度所占权重的变化曲线,如图 2 所示.结果表明,当锚文本所占权重 boost 为 0.6 时,MRR 值最高,而且无限增大 boost 并不能提高 MRR 值.在排序过程中,锚文本所占比例一定要合理,否则会降低检索效果.

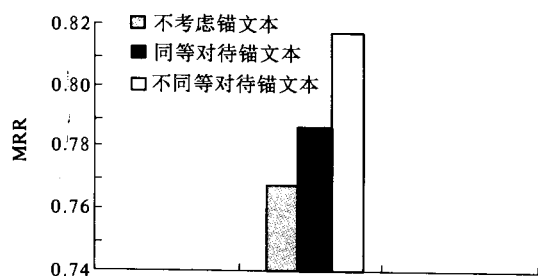


图 1 锚文本有无以及是否被同等对待对 MRR 值的影响

Fig. 1 Effect on MRR with anchor texts or not and be treated equally or not

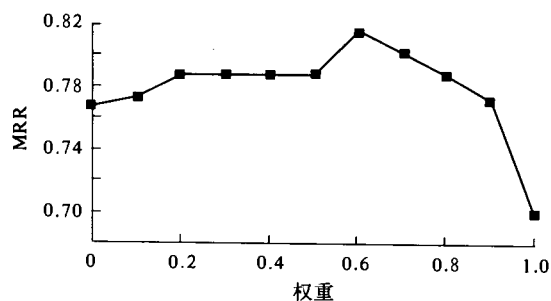


图 2 锚文本相似度所占权重对 MRR 值的影响

Fig. 2 Effect on MRR with different proportion of anchor texts similarity

## 3 小结

应用锚文本与查询词相似度再次检索,可弥补 PageRank 的不足,而利用 PageRank 作为衡量锚文本质量的标准,又可减少锚文本错误描述网页主题信息对检索精确性的影响,从而优化检索效果.但这仅是一个较小规模的尝试,如何有效利用锚文本,链接分析在检索过程中所占比例及将链接分析作用发挥到最大,还需要进一步的改进与提高.

## 参考文献:

- [1] Brin S, Page L. The anatomy of a large scale hypertextual Web search engine[C]//Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, 1998;107-117.
- [2] 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,2004,30(15):143-145.
- [3] 刘雁书,方平.利用链接关系评价网络信息的可行性研究[J].情报学报,2002,21(4):401-406.
- [4] 李晓明,闫宏飞,王继民.搜索引擎:原理、技术与系统[M].北京:科学出版社,2005:165-167.

- [5] Page L., Brin S., Motwani R., et al. The PageRank citation ranking: bringing order to the Web[R]. Stanford: Stanford Digital Library Technologies Project, 1998.
- [6] 陆一鸣, 胡健, 马范援. 一种基于源网页质量的锚文本相似度计算方法——LAAT[J]. 情报学报, 2005, 24(5): 548-554.
- [7] Zhang Ling, Ma Fanyuan, Ye Yunming. CAIA: a new Web page ranking algorithm for search engines[C]// Third International Conference on Intelligent Data Engineering and Automated Learning (IDEAL2002). Springer Verlag, 2002.

## Hyperlink Algorithm Based on Anchor Texts Similarity

LIU Jing-jing, DONG Jing, LIN Hong-fei, YE Zheng

(Department of Computer Science and Engineering, Dalian University of Technology,  
Dalian 116024, China)

**Abstract:** The application of link analysis to search engine for ranking search results is studied, and anchor text is used to rank again and give proper evaluation of importance of each page. The results show that it can improve searching quality to a certain extent.

**Key words:** link analysis; PageRank; anchor text; similarity

(上接第 91 页)

- [5] Park H S, Lee S M. Off-line recognition of large-set handwritten character with multiple Hidden Markov Models[J]. Pattern Recognition, 1996, 29(2): 231-244.

## On-line Chinese Character Recognition Based on 2-classifier Combination

ZHENG Zhen, KANG Yao-hong

(College of Information Science & Technology, Hainan University, Haikou 570228, China)

**Abstract:** The method proposed is based on 2-classifier combination. Firstly the whole features and the strokes features are extracted from the input Chinese characters. During the recognition of single character, GA-BP neural network based on stroke classifier is used as a primary tool to recognize the Chinese characters. But if the words written by consecutive and metamorphic strokes cause the exclusion to appear unexpectedly, HMM based on whole classifier is introduced to improve the whole recognition rate.

**Key words:** 2-classifier combination; whole character; stroke; GA-BP neural network; HMM