基于 PageRank 和锚文本的网页排序研究

刘菁菁，林鸿飞，杨志豪

（大连理工大学计算机科学与工程系，大连 116024）

摘 要： 传统链接分析主要利用基于随机冲浪模型的 PageRank 技术，将网页入度作为评估网页重要性的一个指标。本文在利用传统链接分析成果的基础上，首先获得网页的 PageRank 值，对其进行初步排序，再利用锚文本和查询词的相似度进行二次排序。由于在某些情况下，来源于低权威性网页的锚文本更能合理描述目标网页，因此本文还对此类目标网页的排名加以修正。通过实验表明，这种方法实现了对网页较为公平合理的排序。

关键词： 链接分析；锚文本；PageRank；网页排序

A Study of Ranking Web Pages Based on PageRank and Anchor Text

Liu Jingjing, Lin Hongfei, Yang Zhihao

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

Abstract: Traditional link analysis methods mainly use PageRank technology which based on random surfing model. In degrees is considered as one factor evaluating the importance of ranking. This paper used the fruit of the link analysis algorithms. First, getting the PageRank value of every page to rank primarily and then using the similarity between anchor text and query to rank all pages again. In a way, anchor texts from low quality source pages maybe describe the target pages more properly than those from high ones. So this paper amended the ranking of the target web pages which are described by these anchor texts having large similarity. Our experiment shows that this method can rank the web pages more impartially and logically.

Keywords: link analysis; anchor text; PageRank; ranking web pages

1 引言

在传统信息检索中，计算查询词和文本内容相似度，并根据相似度大小排序，但在 WEB 检索中不再具有优势。随着技术发展，人们日益深入研究 Web 结构特征，发现纯文本和网页的一个明显区别就是网页间超链接的存在。但根据近几年 TREC 的 Web Track 测试的评测结果表明，在网页检索中过分使用超链接分析算法往往适得其反。于是网页链接的锚文本（anchor text）便引起人们关注[2]。锚文本是指当一个网页具有指向另外一个网页的链接时，与此超链接相对应的描述文字[2]。例如在大连理工大学研究生网站首页的源文件中有<a

基金资助：国家自然科学基金（60373095）

作者简介：刘菁菁（1982-），女，山东，硕士研究生. L1982@163.com. 林鸿飞（1962-），男，辽宁，教授，博士，hflin@dlut.edu.cn.

href="http://www.dlut.edu.cn ">大连理工大学这样的一条链接，则“大连理工大学”就是描述网页 http://www.dlut.edu.cn 的锚文本，大连理工大学研究生网站首页为源网页，而网页 http://www.dlut.edu.cn 则被称为目标网页。研究证明，锚文本可以高效的对网页进行排序[3]。

2 链接分析简介

2.1 基于随机冲浪模型的 PageRank

PageRank 可以被比作一个“随机冲浪”模型。该模型可以作为 PageRank 的理论基础，它描述网络用户对网页的访问行为，假设如下：

1） 用户随机的选择一个网页作为上网的起始网页；

2） 看完这个网页后，从该网页内所含的超链内随机的选择一个页面继续进行浏览；

3） 沿着超链前进了一定数目的网页后，用户对这个主题感到厌倦，重新随机选择一个网页进行浏览，如此反复[1]。

本论文试验中我们采用的公式为

$$PR(A) = (1-d) / N + d\ (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) \qquad (1)$$

离线对网页进行 PageRank 值计算，N 取的是网页集中网页总数，d 取经验值 0.85，C(Ti)代表第 Ti 个网页的出度。

2.2 HITS 模型

IBM 研究院 Clever 系统中的相应技术称为 HITS（hyperlink-indexed topic search）。Clever 描述两种类型的网页：

“权威型（Authority）网页”：对于一个特定的检索，该网页提供最好的相关信息；

“目录型（Hub）网页”：给网页提供很多指向其他高质量权威型网页的超链[1]。

进而在每个网页上定义一个“目录型权值”和“权威型权值”两个参数。当遇到一个检索时，Clever 先利用检索的关键词得到一个网页的根集合，然后根据这个集合在整个网页有向图中的位置来扩展这个根集合。在得到这个集合后，就开始计算集合中每个网页的目录型权值和权威型权值。

2.3 其他链接分析的研究

由于 PageRank 技术仅仅考虑到网页出入度对网页排名影响，很多人会利用此技术的缺陷来提高网页排名。针对上述问题，人们利用开始关注网页超链接的另一类信息——锚文本。Zhang 等人提出了 CALA 算法[2]，而文献[2]提出另一种一种新方法——LAAT。

3 相关工作和试验结果

3.1 本文主要思想

本文在考虑网页 PageRank 值高低的同时，考虑锚文本和查询词相似度大小，对于相似度较大锚文本指向的网页排序加以补偿和修正。主要分为语料预处理和 PageRank 计算，以及计算锚文本和查询词相似度，利用计算相似度进行结果调整。

3.2 语料预处理

语料的预处理就是提取网页的超链接和锚文本，目前对链接类型的论述主要集中在站内链接类型上[5]。本文试验主要采用SEWM2006部分语料，包括10多万个网页。在提取网页链接和锚文本过程中，如果该链接属于该网页集，则将其链接及相应的锚文本加以提取。对锚文本和查询词建立向量空间模型，计算二者相似度公式如下为

$$Similarity\ (d_i, Q) = \frac{\sum_{j=1}^{n} w_{ij} \cdot w_{qj}}{\sqrt{\sum_{j=1}^{n} w_{ij}^2 \cdot \sum_{j=1}^{n} w_{qj}^2}} \qquad (2)$$

3.3 实验部分结果

输入查询词“峨嵋”，分别利用 PageRank、查询词和锚文本相似度、综合利用前两个指标重新调整得到的部分数据结果如下，其中每个图中的 Num 代表网页序号。

若仅仅使用 PageRank 值对结果网页进行排序，结果见表 3。其中，网页 1 的入度虽然不是最多的，但指向它的源网页 PageRank 值较高，所以最后综合排名较靠前。

表 3 根据 PageRank 值排序查询结果

Tab.3 Ranking Search Result Based PageRank

Num	PageRank	URL
1	0.087205	http://cnctrip.com/tour/specialtopic/mtem/default.asp
2	0.074462	http://cnctrip.com/tour/specialtopic/mtem/ls/lsdf.asp
3	0.033773	http://cnctrip.com/service/tourlines/CNCTriplines/cnct110259.asp
4	0.014574	http://cnctrip.com/market/shoppingflat/views.asp?hw_id=151
5	0.010711	http://cnctrip.com/tour/specialtopic/mtem/emtour/index.asp

如果不考虑源网页 PageRank 值，只简单汇总所有锚文本，然后根据其与查询词相似度大小重新排序，所得结果见表 4。其中，指向网页 5 的锚文本为“峨眉山是我国的四大佛教名山之一，位于四川中南部，四川盆地西南边缘的峨眉境内，距成都约一百六十公里，在峨眉山市西南七公里处。高出五岳，秀甲天下”，其与查询词相似度较大。

表 4 根据查询词和锚文本相似度排序查询结果

Tab.4 Ranking Search Result Based Similarity between Query and Anchor Text

Num	Similarity	URL
3	0.387097	http://cnctrip.com/service/tourlines/CNCTriplines/cnct110259.asp
1	0.161290	http://cnctrip.com/tour/specialtopic/mtem/default.asp
5	0.096774	http://cnctrip.com/tour/specialtopic/mtem/emtour/index.asp
4	0.064516	http://cnctrip.com/market/shoppingflat/views.asp?hw_id=151
2	0.032258	http://cnctrip.com/tour/specialtopic/mtem/ls/lsdf.asp

利用上面相似度结果和 PageRank 值，对二者选取合适比例，综合考虑二者对排名的影响，计算得到新的排名，结果见表 5。

表 5 根据 PageRank 和查询词和锚文本相似度排序查询结果

Tab.5 Ranking Search Result Based PageRank and Similarity between Query and Anchor Text

Num	New Value	URL
3	0.281099	http://cnctrip.com/service/tourlines/CNCTriplines/cnct110259.asp
1	0.139064	http://cnctrip.com/tour/specialtopic/mtem/default.asp
5	0.070955	http://cnctrip.com/tour/specialtopic/mtem/ls/lsdf.asp
4	0.049534	http://cnctrip.com/market/shoppingflat/views.asp?hw_id=151
2	0.044912	http://cnctrip.com/service/tourlines/CNCTriplines/cnct110089.asp

4 结束语与进一步改进

本文所提出的方法，仅仅是利用链接分析提高查询结果排序的一个初步实现，在计算小规模语料的基础上利用PageRank和锚文本与查询词相似度对网页检索结果进行重新排序。该方法利用含有重要信息的锚文本，并考虑客观程度上能反映网页相对重要性的PageRank技术对网页重新排序，调整结果。目前链接分析主要考虑的是站内链接，但站外链接反映了网页对外界影响力，如何将站外链接加以综合评价还需进一步研究。此外在实验过程中锚文本对查询结果的影响力还有待进一步调整，使之更有利于网页排序。

参考文献：

[1] 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.

[2] 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 年 10 月, 第 24 卷第 5 期: 548-554

[3] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[4] N. Eiron and KS McCurley. Analysis of anchor text for web search [J]. In Proceedings of the 26th Annual International 673 Information Retrieval, ACM, 2003: 459-460.

[5] 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 年 8 月, 第 21 卷第 4 期: 401-406

[6] Kleinberg. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM, 46(5): 604–632,999

[7] 吕俊生. 网上信息资源的链接分析研究[J].情报科学,2005 年 1 月,第 23 卷第 1 期: 78-82

[8] Taher h. Haeliwala Topic-Sensitive PageRank [J]. In Proceedings of the 11th International World Wide Web Conference (WWW2003), 2003.