

一种基于结构分析的改进 HITS 算法

仲 婷, 金 浩, 冯茜芦, 潘金贵

(1. 南京大学 计算机软件新技术国家重点实验室, 江苏 南京 210093)

摘 要: Web 挖掘技术的应用之一就是 Web 搜索引擎。对于搜索引擎中的 Web 结构挖掘, 通过对经典的超链接分析算法的研究, 对 Web 超链结构进行深入分析的基础上, 针对 HITS 算法的缺陷, 通过引入权值和调整因子对其进行改进。实验表明, 改进后的算法表现更加出色。

关键词: Web 挖掘; 超链分析; HITS

中图分类号: TP393 **文献标识码:** A **文章编号:** 1001-6600(2007)02-0214-04

Web 结构挖掘, 即 Web 超链接分析是 Web 信息检索领域近来较热门的一个研究方向。目前比较流行的基本算法有 PageRank^[1], HITS^[2], SALSA^[3]等。超链接分析算法根据网页集合中网页之间的引用关系, 计算出网页在对应网页集中的“地位”, 这种地位类似于现实社会中人的声望, 因此有比较好的参考价值。

HITS 算法由 Kleinberg 在文献[2]中正式提出。但 HITS 算法最容易产生也最致命的问题是主题泛化和漂移现象。算法中所有超链接被同等对待, 缺乏语义分析, 容易产生主题漂移现象(topic drift)。

Bharat 和 Henzinger 针对这种现象提出 BHITS 算法。BHITS 加入了文本分析技术来改善 HITS^[4], 能很大程度地解决主题漂移问题。但 BHITS 算法的局限在于迭代计算时, 只针对来自于同一站点或指向同一站点的超链作了调整, 没有很好的普遍性, 更没有根本解决 HITS 算法的主题漂移问题。文献[5]提出的 WBHITS 算法对网页的超链权值进一步作了修改和限制。但仍然存在普适性不好的问题。而且算法中阈值的选取也相当困难^[6]。IBM 的 Clever 工程组通过简单的文本分析来调整超链权值。他们提出的 ARC 算法^[7]通过文本分析技术来设置权值, 取得了一定的效果, 却也增加了一定的开销。HITS 算法在扩展根集 S 到基集 T 的过程中, 加入的大量网页很可能包含很多相关性并不好的网页。文献[8]针对此问题提出了改进算法, 有效降低了迭代计算复杂度。但仍有可能在缩减过程中丢失一些相关性很不错的页面。

本文在对 Web 超链结构进行深入分析的基础上, 针对 HITS 算法的缺陷, 引入权值和调整因子。第二部分分析 HITS 并引入改进后的算法, 第三部分给出实验结果, 并进行分析比较。最后对全文进行总结。

1 基于结构分析的改进 HITS 算法

1.1 几个简单例子

HITS 算法中所有超链同等对待是主体漂移现象产生的重要原因。先来看几个简单的例子(图 1)。

例 1 设基集 T 包含图 1a 和图 1c 中所有网页, 迭代 15 步得到 B, C, D, E 的 authority 值均为 0.499 9, N 的 authority 值约为 0.026 7。随着迭代步数增加, 这两个值分别趋近于 0.5 和 0。考察这些网页时, 发现 A 若是指向大量网页的综合性网页, B, C, D 和 E 则会拥有较大 authority 值; 而 N 被较多网页指向, 说明它有一定的权威性, 理应拥有比 B, C, D, E 更大的 authority 值。因此计算 authority 值时两种超链的权值相等是不合理的。

例 2 设基集 T 包含图 1b 和图 1c 中所有网页, 可以看到 J 和 N 唯一区别是 J 比 N 多一个网页指

收稿日期: 2006-12-15

基金项目: 国家自然科学基金资助项目(60473113, 60533080)

作者简介: 仲婷(1983—), 女, 江苏江都人, 南京大学硕士研究生。

通信作者: 潘金贵(1952—), 男, 江苏南京人, 南京大学教授, 博士。

向,而 HITS 的运算结果却是 J 的 authority 值接近于 1, N 的 authority 值趋近于 0,而 J 和 N 很可能都是相关性不错的网页。这种现象称为吸收现象。它的产生会导致丢失很多相关性很好的网页。

1.2 预处理

HITS 算法在构建有向图时,去除了同一主机内部的超链。但互联网中存在大量来自相同域的网页,它们之间的超链一般也仅用于导航目的,不具备考察价值。所以在计算之前我们删除了这部分超链。

1.3 改进公式

以上几个简单的例子启发我们对原始 HITS 算法的计算过程做出适当调整。经典的 HITS 改进算法通常是引入超链的权值,此时迭代公式为:

$$a_v = \sum_{u \rightarrow v} h_u \cdot W_{auth}, h_u = \sum_{u \rightarrow v} a_v \cdot W_{hub} \quad (1)$$

其中 W_{auth} 和 W_{hub} 分别为计算文档 authority 和 hub 时的权值, u 和 v 代表网页节点, $u \rightarrow v$ 表示 u 有超链指向 v , h_u, a_u 分别代表 u 的 hub 值和 authority 值并初始化为 1。

直接利用乘法计算引入权值对计算结果影响过于激烈,反而激化了吸收现象。为减弱甚至避免吸收现象,就要减弱迭代过程中值的累积和传递,因此通过加法计算引入一个因子来调整结果,迭代公式为:

$$a_v = \sum_{u \rightarrow v} h_u + W_{auth}, h_u = \sum_{u \rightarrow v} a_v + W_{hub} \quad (2)$$

其中调整因子 W_{auth} 和 W_{hub} 由公式(3)给出, u' 和 v' 除了满足 $u' \rightarrow v$ 和 $u \rightarrow v'$ 外,均是根集 S 中的网页。

$$W_{auth} = \sqrt{\sum_{u' \rightarrow v, v' \in S} 1}, W_{hub} = \sqrt{\sum_{u \rightarrow v', v' \in S} 1} \quad (3)$$

即计算网页文档 u 的 authority 值时, W_{auth} 等于根集 S 中所有指向网页 u 的网页数目的平方根,如果有同一站点的若干网页指向 u , 只计一次;同理,计算网页文档 u 的 hub 值时, W_{hub} 等于 S 中所有被网页 u 指向的网页个数的平方根,如果 u 同时指向同一站点的若干网页, 只计一次。事实上,改进算法引入 W_{auth} 和 W_{hub} 正是为了适当降低中心度或者权威度的累积和传递,同时也有效防止吸收现象和主题漂移现象。

需要指出,通过分析和相关的实验发现,如果把根集 S 中所有指向网页 u 的网页数目直接作为 W_{auth} 引入原迭代公式效果并不理想。因为 W_{auth} 和 W_{hub} 的高低直接影响迭代的效果,取值过高会直接导致迭代过程失去意义,演变成网页的权威度只受制于 W_{auth} 和 W_{hub} 值的结论。而取值过低,则会减弱改进效果,主题漂移现象难以避免。这里采取了折衷方案,引入开方运算。后文的实验将展示这种方案的效果。

当基集的数目达到一定规模时,往往会出现大量相同站点的网页。这些网页又往往有相同或相近的链接表现(即被一定数量来自相同站点的网页指向,及指向一定数量相同站点的网页)。Bharat 的做法是在计算权威度时,若有 k 个超链来自同一主机的网页,则将这些链接的权值设为 $1/k$,且该页面的 hub 值为所有 k 个网页 hub 值的平均。对于中心度的计算类似。和 Bharat 不同,我们把处理范围从主机扩大到域名。在计算网页的权威度时,若有 k 个超链来自同一域的网页,则只考虑这其中 hub 值最大的一个而忽略其他页面(即 hub 最大的网页超链权值为 1,其他网页的超链权值为 0)。相比 BHITS,取最大值比取平均值更为合理,而且也节省了计算平均值的时间开销。改进后的公式实际上可以表示为如下形式($C_{auth}=0$ 或 1, $C_{hub}=0$ 或 1):

$$a_v = \sum_{u \rightarrow v} (C_{auth} \cdot h_u) + W_{auth}, h_u = \sum_{u \rightarrow v} (C_{hub} \cdot a_v) + W_{hub} \quad (4)$$

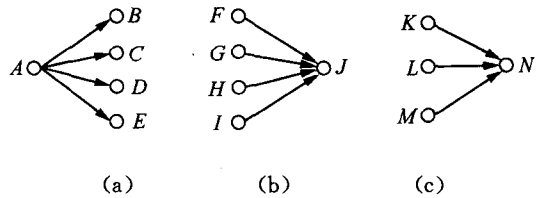


图 1 网页超链结构举例

Fig. 1 Examples of Web hyperlink structure

表 1 例 1,2 利用改进算法的运算结果(迭代 15 步)

Tab. 1 Results of example 1,2 after applying improved algorithm (iterated for 15 steps)

例	网页	权威度	中心度
例 1	A	0.000 000 034 4	0.918 917 952 4
	B,C,D,E	0.302 775 637 3	0.000 000 006 7
	K,L,M	0.000 000 034 4	0.227 735 077 6
	N	0.795 806 291 7	0.000 000 006 7
例 2	F,G,H,I	0.000 000 000 1	0.425 237 453 9
	J	0.813 774 907 3	0.000 046 699 7
	K,L,M	0.000 000 000 1	0.303 695 256 8
	N	0.581 180 178 8	0.000 046 699 7

接表现(即被一定数量来自相同站点的网页指向,及指向一定数量相同站点的网页)。Bharat 的做法是在计算权威度时,若有 k 个超链来自同一主机的网页,则将这些链接的权值设为 $1/k$,且该页面的 hub 值为所有 k 个网页 hub 值的平均。对于中心度的计算类似。和 Bharat 不同,我们把处理范围从主机扩大到域名。在计算网页的权威度时,若有 k 个超链来自同一域的网页,则只考虑这其中 hub 值最大的一个而忽略其他页面(即 hub 最大的网页超链权值为 1,其他网页的超链权值为 0)。相比 BHITS,取最大值比取平均值更为合理,而且也节省了计算平均值的时间开销。改进后的公式实际上可以表示为如下形式

2 实验

2.1 实验 1

从表 1 可以看出,1.1 节中的例子用改进算法得到了预期效果,网页 N 已经得到了稳定且合理的权威度。

2.2 实验 2: basketball

用 SHITS(modified HITS by structure analysis, with factors and weights)代表公式(4)对应的算法。关键词 basketball,根集大小 10。取返回的前 10 个网页作为根集 S ,然后扩展成基集 T 。这时 T 总共包含 346 个网页。迭代 15 步后比较结果如表 2 和表 3(只列出结果中前 10 的 URL,“—”表示相应的排名不在前 10 之内):

表 2 HITS 算法的迭代结果
Tab. 2 Iterated results of HITS

域 名	HITS	Google	Yahoo	权值
http://www.basketball.com/	1	2	8	0.270 526 112 1
http://www.nba.com/	2	1	3	0.255 485 804 2
http://www.x5t.com/	3	—	—	0.235 133 368 7
http://www.showmetickets.com/nfl/giants_tickets.htm/	4	—	—	0.232 055 191 4
http://www.nfl-football-tickets.com/ramstickets.htm/	5	—	—	0.232 055 191 4
http://www.tickets4u.com/nfl-packers-tickets.asp/	6	—	—	0.232 055 191 4
http://www.tickets4u.com/nfl-chiefs-tickets.asp/	7	—	—	0.232 055 191 4
http://www.landauction.com/	8	—	—	0.232 055 191 4
http://www.landoceanshores.com/	9	—	—	0.232 055 191 4
http://www.portestates.com/	10	—	—	0.232 055 191 4

表 3 SHITS 算法的迭代结果
Tab. 3 Iterated results of SHITS

域 名	SHITS	Google	Yahoo	权值
http://www.nba.com/	1	1	3	0.389 927 726 2
http://www.fiba.com/	2	6	1	0.352 340 483 9
http://www.wnba.com/	3	7	4	0.317 909 798 7
http://www.usabasketball.com/	4	3	2	0.287 052 621 6
http://www.basketball.com/	5	2	8	0.219 485 236 1
http://www.powerbasketball.com/	6	5	12	0.217 968 750 2
http://www.bbhighway.com/	7	4	10	0.209 330 337 9
http://www.basketball.ca/	8	10	—	0.181 214 141 5
http://dmoz.org/sports/basketball/	9	9	—	0.171 108 970 5
http://www.cnn.com/	10	—	—	0.121 219 736 0

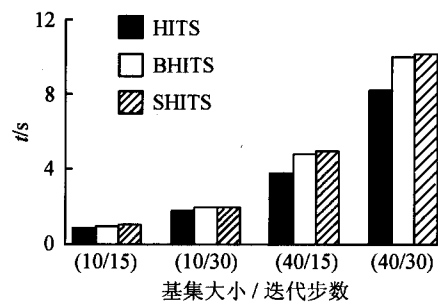


图 2 HITS, BHITS 和 SHITS 的时间性能对比

Fig. 2 Performance comparison of HITS, BHITS and SHITS

从表中看出, HITS 算法出现了反常,前 10 个网页中有 8 个相关性很差,大部分相关性好的网页落到了后面。而 SHITS 表现稳定,排名较前的网页相关性都不错。研究 HITS 的结果时发现,排名 4 到 16 的网页 authority 值相等,进一步分析其链接关系时发现这 13 个网页均同时被且仅被另外 9 个网页引用。这两组网页相互促进和增强,导致前者有很大的 authority 值而后者有很大的 hub 值。排名前三的网页也被这 9 个网页引用。

2.3 实验 3:效率

把 HITS, BHITS 和 SHITS 算法在迭代计算过程中的时间开销作横向对比。实验仍是基于 basketball 查询,见图 2,数据为多次试验后的平均值。由于本实验目的是横向对比,算法在具体实现时没有采取过多的优化手段,加上硬件条件有限,导致时间消耗偏大,

但从图中的数据我们可以看出相比 HITS 和 BHITS,SHITS 只是在时间开销上略有增加,考虑到性能的显著提高,这是完全可以接受的结果。

3 总结

本文以 HITS 算法为基础,在深入研究了网页超链接结构之后提出了改进算法。可以看到只是在增加极有限的计算开销的代价下使算法效果获得了明显改善,稳定性得到了很大程度提升。但由于 Web 规模庞大,结构复杂,当基集中包含大量相互链接紧密,但主题相关度不是很好的网页文档时,算法仍有可能产生偏差。所以在实际应用中,我们还将改进算法的基础上结合其他挖掘方法做进一步的研究。

参 考 文 献:

- [1] PAGE L,BRIN S,MOTWANI R,et al. The PageRank citation ranking,Bringing order to the Web[R]. Stanford,CA: Stanford Digital Libraries Working Paper,1998.
- [2] KLEINBERG J. Authoritative sources in a hyperlinked environment[C]//Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms. New Orleans;ACM Press,1997:668-677.
- [3] LEMPEL R,MORAN S. SALSA:The stochastic approach for link-structure analysis[J]. ACM Transactions on Information Systems,2001,19(2):131-160.
- [4] BHARAT K,HENZINGER M R. Improved algorithms for topic distillation in a hyperlinked environment[C]//21st International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne;ACM Press,1998:104-111.
- [5] LI Long-zhuang,SHANG Yi,ZHANG Wei. Improvement of HITS-based algorithms on Web documents[C]//Proceedings of the eleventh international conference on World Wide Web. New York;ACM Press,2002:527-535.
- [6] CHAKRABARTI S,DOM B E,GIBSON D,et al. Mining the link structure of the world wide Web[J]. IEEE Computer,1999,32(8):60-67.
- [7] CHAKRABARTI S,DOM B E,GIBSON D,et a. Automatic resource compilation by analyzing hyperlink structure and associated text[C]//Proceedings of the 7th International WWW Conference. Amsterdam;Elsevier Science Publisher,1998:65-74.
- [8] NOMURA S,OYAMA S,HAYAMIZU T,et al. Analysis and improvement of HITS algorithm for detecting web communities[J]. Systems and Computers in Japan,2004,35(13):32-42.

An Improved HITS Algorithm Based on Structure Analysis

ZHONG Ting,JIN Hao,FENG Xi-lu,PAN Jin-gui

(State Key Lab for Novel Software,Nanjing University,Nanjing 210093,China)

Abstract: Web search engine is one of the applications of web mining technologies. Web structure mining used in web search engines is reviewed. After deeply analyzing the web hyperlink structure,the HITS algorithm is improved by importing weights and regulative factors. Experiments show that the improved algorithm works much better.

Key words: Web mining;hyperlink analysis;HITS

(责任编辑 王龙杰)