

基于可信度模型的 HITS 算法的改进

任 平 吴 陈 雷 艳 方 李 丛

(江苏科技大学计算机科学与工程学院, 镇江 212003)

摘要 HITS 是一种经典的 Web 链接分析算法, 其主要问题是容易发生主题漂移和互相加强。针对这些问题, 提出了一种改进的算法 T-HITS。通过一种网络结构图来映射垃圾链接集与其对应的网站, 并结合链接文本将垃圾链接排除, 最后利用可信度模型来修正结果, 实验数据表明改进后的算法提高了查询结果的相关度, 减少了主题漂移的发生。

关键词 HITS 算法 可信度模型 搜索引擎

中图法分类号 TP391.3; **文献标志码** A

在不到 30 年的时间里, 因特网已经由早期连接少量站点的研究原型成长为覆盖世界上所有国家的全球通信系统, 其增长速度是非凡的。如何从这些不断增长的海量数据中寻找出有用的信息将会面临一个严峻的挑战。目前搜索引擎行业的蓬勃发展, 使得许多组织和个人试图人为地增加其排名, 以吸引更多游客到他们的网站。与此同时, 随着互联网的不断壮大, 网站之间、页面之间的导航链接也快速增加, 这给搜索引擎也带来一定影响。

搜索引擎作为一种用来获取关于某一主题信息的工具被广泛应用, 用户通过关键字指定一个主题, 搜索引擎基于各种算法对包含关键字的网页进行评分和排序, 然后按照分数递减的方式输出结果。例如, Brin 和 Page 所提出的 PageRank^[1] 算法已被谷歌采用为网页的评分算法。另一个比较著名的算法就是 Kleinberg 提出的 HITS^[2] 算法, 具有以下三个显著特点:

(1) 即使不包含关键字的网页, 只要与主题高度相关也会获得高分;

(2) 和谷歌的 PageRank 算法相比, HITS 只需要少量的网页信息, 甚至可以在个人 PC 上运行;

(3) 不需要扫描网页内容, 运行时间短。

在 HITS 刚提出的时候, 该算法运行良好, 相应的基于 HITS 的各种算法^[3-6] 也不断出现。然而, 在如今的网络环境下, 由于垃圾链接的不断增加, 原始的 HITS 算法和基于 HITS 的算法遇到了困难。虽然已经有几个人提出了寻找垃圾链接的方法^[7-9], 但它们需要太多的网页数据。例如, 由 Fetterly^[8,9] 等人提出的算法就需要提取大量的网页内容, 很显然网页内容的数据量远远超过链接的数据量, 相应的运算时间可想而知。由于链接结构的局限性, 仅凭链接结构很难区分正常链接与垃圾链接, 根据网页的结构我们可以知道, 每一个超链接都会附带一个文本, 而这个文本就是被链接网页的一个高度概括, 我们就可以采用这句文本来作为识别垃圾链接的辅助信息, 这样既可以更准确的辨别垃圾链接, 又不需要扫描整个网页的内容。垃圾链接集是一张描绘垃圾链接的有向图, 图的顶点表示每一个网页, 图的边表示两个网页之间的链接, 本文分别采用原始 HITS 算法、BHITS 算法^[3] 和我们改进后的算法 T-HITS 寻找垃圾链接集, 然后排除垃圾链接所指的网页, 并利用可信度模型对剩余的网页的评分进行修正, 最后通过实验来评价输出结果。

1 原始 HITS 算法的基本思想

2009 年 7 月 24 日收到
第一作者简介:任 平 (1985—), 男, 汉族, 山西省大同市人, 硕士研究生, 研究方向:软件开发方法与技术。

Kleinberg 最早提出的 HITS 算法中, 提出了权

威网页 (authority) 和中心网页 (hub) 的概念。权威网页 (authority) 是指那些与给定查询主题的上下文最为相关的网页, 这就是人们最需要和最关心的网页; 而中心网页 (hub) 则是那些包含了多个指向 authority 的超链接的网页。Kleinberg 认为查询结果的重要程度应该建立在用户查询条件的基础上, 而且每一个页面都分别有 authority 值和 hub 值。如果一个页面被许多相关主题的其它页面所指向, 则该页面具有较高的 authority 值; 如果一个页面指向许多相关主题页面, 则该页面具有较高的 hub 值。算法的具体过程为:

(1) 构造根集 R, 将查询主题提交给某个搜索引擎, 取前 x 个结果作为根集, Kleinberg 通常将 x 设为 200^[2]。

(2) 根据根集 R 构造链接结构图 $G = (V, E)$, 其中顶点 V 包括: R 中的所有网页、R 中的页面所指向的网页和指向 R 中页面的网页, 有向边 E 代表两个网页之间的链接和方向。

(3) 对于每个顶点 v_i , 赋值 a_i 和 h_i 为 1, 分别代表 authority 值和 hub 值。

(4) 重复下面的步骤 n 次, 一般迭代 a 到 c 步骤 50 次后 a 和 h 就趋于稳定, 所以 n 通常设为 50:

$$\text{a、计算 } a_i = \sum_{(v_j, v_i)} \in E_j^h.$$

$$\text{b、计算 } h_i = \sum_{(v_j, v_i)} \in E_j^a.$$

c、将所有 a_i 和 h_i 组成的向量 $A = (a_1, a_2, \dots, a_{|V|})$ 和 $H = (h_1, h_2, \dots, h_{|V|})$ 标准化, 即, $\sum_{i=1}^{|V|} a_i = 1$, $\sum_{i=1}^{|V|} h_i = 1$ 。

(5) 输出向量 A 和 H 。

2 BHITS 算法的过程

通过上面的描述我们可以发现原始的 HITS 算法中存在两个比较明显的问题:

(1) V 中所包含的网页不一定都是和主题相关的, 这就会产生所谓的“主题漂移”^[12] 现象。

(2) “互相加强”的问题, 如果两个网站之间有

很多页面互相链接, 这两个网站就会获得很高的分数。

针对问题(2), Bharat 和 Henzinger 提出了改进的算法 BHITS^[11], BHITS 将原始算法的第 4 步中的 a 和 b 修改为:

$$\text{a'、计算 } a_i = \sum_{(v_j, v_i)} \in E \frac{h_j}{r - (\text{host}(v_j) \cdot v_i)}.$$

$$\text{b'、计算 } h_i = \sum_{(v_j, v_i)} \in E \frac{a_i}{r - (\text{host}(v_j) \cdot v_i)}.$$

其中 host 定义为每个 URL 中的域名

$\Gamma^-(v_i)$ 为 $\{v_j \in V | (v_j, v_i) \in E\}$, $\Gamma^+(V_i)$ 为 $\{v_j \in V | (v_i, v_j) \in E\}$;

$H^-(v_i)$ 为 $\Gamma^-(v_i)$ 中 host 的个数, $H^+(V_i)$ 为 $\Gamma^+(v_i)$ 中 host 的个数;

$r^-(\text{host}_k, v_i)$ 为 host_k 到页面 v_i 的链接数。

在 BHITS 算法中即使同一个网站中的多个网页同时指向另一个网页, 这个网页的分数也不会太高。虽然 BHITS 可以解决“互相加强”的问题, 但是没有考虑链接文本的内容, 还会产生“主题漂移”, 本文的改进算法 T-HITS 就是在 BHITS 的基础上继续改进, 去除影响结果的垃圾链接, 从而解决“主题漂移”的问题。

3 可信度模型的建立

可信度模型中主要有三部分组成: 黑名单集合、白名单集合、未知链接集合。通过多次实验验证, 最终将白名单中的链接对网页可信度的影响系数设为 1, 黑名单中的设为 -1, 未知链接集合中的设为 0.5。

3.1 寻找垃圾链接集, 即黑名单集合 B

Wu and Davison^[13] 将垃圾链接集定义为: 在网络图中那些高度密集在一起的垃圾链接所组成的子图。由于“互相加强”问题的存在, 垃圾链接集中的网页会得到很高的分数, 最终导致结果中有很多网页属于垃圾链接集。通过研究发现, 那些属于垃圾链接集的网页通常会有相同的 IP 地址或者使用相同的 name server, 基于这点我们将 BHITS 的算法

改进如下：

(1) 在 BHITS 算法中的第二步中增加两步预处理：

a、将具有相同 IP 地址的网页从顶点集 V 中移除，并加入黑名单集合 B 。

b、将具有相同 name server 的网页从顶点集 V 中移除，并加入黑名单集合 B 。

(2) 在第三步中增加一步后变为：

a、根据链接文本将与主题无关的链接移除，并加入黑名单集合 B 。

b、对于每个顶点 v_i , 赋值 a_i 和 h_i 为 1, 分别代表 authority 值和 hub 值。

3.2 可信链接集即白名单集合 W 的确定

这里参照了 TrustRank^[10]的思想，即“如果一个网页被另一个可以信任的网页所链接，那么这个网页也是可以信任的”。文献[11]通过实验得出在根集 R 中的网页基本上都是与主题相关的可以信任的网页，所以 Trust-Score 算法中将 R 中的网页作为信任链接基集，以此来对其他网页评分。本文中也将根集 R 作为白名单集合 W 。

3.3 未知链接集合 U

既不属于白名单也不属于黑名单的链接称为未知链接，链接结构图中的大部分属于这样的链接，我们这里借用法律上的一个概念“无罪推定原则”，假设这些链接不是垃圾链接，通过大量实验表明将这类链接的可信度影响系数设为 0.5 比较合理。

3.4 计算网页的可信度 T

(1) 对于每一个网页 v_i , 分析 v_i 中的链接，计算属于白名单的链接数 N_w , 属于黑名单的链接数和未知链接数 N_u 。

(2) 计算网页 v_i 的可信度 $T(v_i)$, 过程如下：

a、如果 $N_w + 0.5N_u > N_b$,

$$T(v_i) = \frac{N_w + 0.5N_u - N_b}{N_w + N_b + N_u};$$

b、如果 $N_w + 0.5N_u < N_b$, $T(v_i) = 0$;

c、输出向量 $T(t_1, t_2, \dots, t|V|)$ 。

3.5 最终分数的确定

最后将向量 A 与 T 相加后得到结果集 Result,

并对 Result 中的每个分量进行排序，得到网页的最终排名，按分数递减输出结果。

4 实验

实验系统设计的目的是为了验证我们对原主题精选算法 HITS 改进的有效性，我们在系统设计时遵循以下原则：

(1) 系统的实现严格按照算法设计进行，某些经验数据的选取参考相关研究中的实验，涉及信息检索中的典型操作直接采用通用的算法并遵循其要求。

(2) 由于实验中的很多具体过程(如页面下载、页面内容分析等)对时间和空间的要求比较高，实验系统不采用顺序流程，而分多个模块进行，保留多个中间结果。

因为算法的结果是按照分数递减的方式排列的，所以前 10 个结果就是与主题最相关的 10 个结果，我们就是通过评价这 10 个结果与所给主题的相关程度来评价算法的效果。本文中共采用了比较有代表性的 10 个主题作为输入条件，并对原始的 HITS 算法、HITS 的变种算法 BHITS 和我们改进后的 T-HITS 算法得到的实验结果进行比较，来验证 T-HITS 算法的有效性和优越性。

如何确定精选出的网页质量如何，是一个非常主观的概念，目前广泛采用的方法仍然是人工评价。我们也采取了类似的评价方法。取每个算法的运行结果中排在前 10 的权威网页，组成待评价的网页集，将该集合中的网页以随机的次序提交给志愿者进行评价，如果与主题相关的网页数大于 8，那么我们说算法对于该主题是“有效”的。

表 1 中 T1-T10 为比较有代表性的 10 个主题，分别为“Garden”, “Railway”, “Tennis”, “Cheese”, “Basketball”, “English”, “Email”, “Paper”, “Dictionary”, “Profession”。将每个主题分别提交给搜索引擎 Google, 取前 200 个结果作为该主题的根集 R ，然后将每个主题的根集 R 作为三种算法的输入集(原始 HITS 算法、BHITS 算法和改进后的算法 T-

HITS)。从实验的结果中我们可以清楚的看到,原始的 HITS 算法由于存在前文提到的问题 1 和 2,再加上垃圾链接的影响,找到的网页中与主题相关的网页数仅为 17,且只有 1 个主题可以得到有效的结果,BHITS 算法中与主题相关的网页数为 50,有效主题数也只有 2 个,而我们的 T-HITS 算法通过将影响结果的垃圾链接排除,并对网页赋予可信度后,找到的网页中与主题相关的网页数为 96,得到的有效值达到了 9,说明了 T-HITS 算法可以很好的解决原始 HITS 算法存在的两个问题。

表 1 实验结果

	HITS			BHITS			T - HITS		
	相关	不相关	Spam	相关	不相关	Spam	相关	不相关	Spam
T1	0	10	8	6	4	4	10	0	0
T2	2	8	7	10	0	0	9	1	0
T3	0	10	9	4	6	6	10	0	0
T4	0	10	8	5	5	4	10	0	0
T5	9	1	1	3	7	6	9	1	0
T6	3	7	6	2	8	5	10	0	0
T7	0	10	10	6	4	4	10	0	0
T8	1	9	8	3	7	5	8	2	0
T9	0	10	9	10	0	0	10	0	0
T10	2	8	8	1	9	6	10	0	0
总数	17	83	74	50	50	40	96	4	0
有效		1			2			9	

5 结束语

基于链接分析的算法,提供了一种衡量网页质量的客观方法,独立于语言,独立于内容,不需人工干预就能自动发现 WEB 上重要的资源,自动实现文档分类。但是也有一些共同的问题影响着算法的精度,比如根集的质量,垃圾链接的与日俱增,人为恶意修改等。当然这些问题带有很大的主观性如根集质量不能精确的定义,链接是否包含重要的信息也没有有效的方法来准确的判定,分析链接文本又涉及到语义问题,查询的分类也没有明确界限。如果算法要取得更好的效果,在这几个方面需要继续做深入的研究。本文在 HITS 算法的基础上

提出一种基于可信模型的 T-HITS 算法,实验表明改进后的算法较大提高了返回结果的相关度,减少了发生主题漂移的可能性。但 Web 是一个高速发展的结构体,相信还有很多未知的关于 Web 结构的知识有待认识开发。未来的工作要紧密结合 Web 链接结构的分析进行研究,以及对 HITS 算法主题漂移的程度度量和执行时间上作进一步研究。

参 考 文 献

- 1 Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web, <http://www-db.Stanford.edu/~backrub/pageranksub.ps>, 1998
- 2 Kleinberg J. Authoritative sources in a hyperlinked environment. Proc 9th ACM-SIAM Symposium on Discrete Algorithms, 1998; 668—677
- 3 Bharat K, Henzinger M R. Improved algorithms for topic distillation in a hyperlinked environment. Proc 21st ACM SIGIR Conference, 1998; 104—111
- 4 Lempel R, Moran S. The stochastic approach for link-structure analysis (SALSA) and the tlc effect. Proc 9th WWW Conference, 2000; 387—401
- 5 Li L , Shang Y. Zhang W. Improvement of HITS-based algorithms on Web documents. Proc 11th WWW Conference, 2002; 527—535
- 6 Wang X , Lu Z Zhou A. Topic exploration and distillation for Web search by a similarity-based analysis. Proc 3rd WAIM Conference, 2002; 316—327
- 7 Costa Carvalho A, Chirita P, Moura E. Site level noise removal for search engines. Proc 15th WWW Conference, 2006; 73—82
- 8 Fetterly D, Manasse M, Najork M. Spam, damn spam, and statistics: Using statistical analysis to locate spam Web pages. Proc 7th International Workshop on the Web and Databases, 2004; 1—6
- 9 Fetterly D ,Manasse M ,Najork M , et al. Detecting spam Web pages through content analysis. Proc 15th WWW Conference, 2006; 83—92
- 10 Gyongyi Z , Garcia-Molina H, Pedersen J. Combating Web spam with TrustRank. Proc 30th VLDB Conference, 2004; 576—587
- 11 Asano Y, Yu Tezuka, Nishizeki T. Improvement of HITS algorithms for spam links. APWeb/WAIM 2007, LNCS 4505, 2007; 479—490
- 12 Chakrabarti S, Dom B, Gibson D, et al. Automatic resource list compilation by analyzing hyperlink structure and associated text. Proc 7th International World Wide Web Conference, 1998
- 13 Wu B, Davison B D. Identifying link farm spam pages. Proc 14th WWW Conference, 2005; 820—829

Improvements of the HITS Algorithm Based on the Model of Credibility

REN Ping, WU Chen, LEI Yan-yun, LI Cong

(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, P. R. China)

[Abstract] HITS is one of the classical Web link analysis algorithms. The main problem is the topic drift and mutual reinforcement. According to these problems, an improved algorithm T-HITS is proposed, using a network structure to map spam collection with corresponding website and exclude the spam links with link text. At last the model of credibility is used to fix the results, and the experimental data shows that the improved algorithm has a big improvement about the degree of correlation of the results, and can decrease the probability of the topic drift.

[Key words] HITS algorithm the model of credibility search engine

(上接第 6389 页)

To Design Remoting Examining System of Speed-down on GA and NN

SU Jun, CHEN Shu-xia¹, WU Zhao-yang

(School of Mechanical, Nantong University, Nantong 226019, P. R. China Nanfong Profession Universing¹, Nanfong 221097, P. R. China)

[Abstract] Through to the speed reducer failure mode's analysis, uses the BP neural network establishment failure diagnosis model, the use genetic algorithm optimization neural network weight, the threshold value, the network architecture, will inherit the neural network model to apply in the long-distance speed reducer's failure diagnosis design. Compared with the sole neural network and the heredity neural network's training curve of error, obtains the heredity neural network is higher than the neural network training model by far in the training speed and the accuracy.

[Key words] NN GA weights threshold error curve