一种基于最近搜索周期被引用频率的改进WPR算法

王旭阳 任国盛
(兰州理工大学计算机与通信学院 兰州 730000)

摘 要 针对WPR(Weighted PageRank)算法存在的在网页搜索方面的主题漂移和偏重旧网页的现象,综合网页的主题特征和最近搜索周期网页的被引用频率两个因素,提出了一种改进的算法WTFPR(Weighted Topic Frequency PageRank)。该算法通过内容分析,采用改进的TD-IDF算法来解决网页相关性,改善主题漂移现象;通过网页的最近搜索周期的被引用频率来提高那些较新而且价值较高的网页的PR值,从而改善偏重旧网页的现象。仿真结果表明,改进后的算法与WPR算法相比获得了更好的效果。

关键词 主题特征,被引用频率,偏重旧网页,搜索周期,主题漂移
中图法分类号 TP391.3 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.019

Improved WPR Algorithm Based on Referenced Frequency in Recent Search Cycle

WANG Xu-yang REN Guo-sheng
(College of Computer and Communication, Lanzhou University of Technology, Lanzhou 730000, China)

Abstract For the topic drift and bias towards the old pages of WPR(Weighted PageRank) algorithm exist in the Web search, consolidated two factors of Web pages' topic features and referenced frequency in recent search cycle, we proposed an improved algorithm WTFPR(Weighted Topic Frequency PageRank). The algorithm uses improved TD-IDF algorithm to solve relevance of page by content analysis to reduce the topic drift. The algorithm improves the PR value of new and has high quality by referenced frequency of pages in recent search cycle, reducing bias towards the old pages. Simulation results show that the improved algorithm obtaines better results compared to WPR.

Keywords Topic features, Referenced frequency, Bias towards the old pages, Search cycle, Topic drift

1 概述

当今互联网已经成为现代共享信息的主要载体,无论网站网页数量还是用户数量都特别巨大,搜索引擎在搜索信息方面占据主要地位。从用户行为上看,多数用户在使用搜索引擎的搜索结果时,只会点击搜索出来的前2页中10到20个高相关度的搜索结果。因此如何将最能满足用户需求的页面排列在搜索结果的前面变得至关重要。在网页排序算法中,最著名的是1998年由Sergey Brin和Lawrence Page提出的基于链接分析的PageRank网页排序算法[1]。而后由Wenpu Xing和Ali Ghorbani提出的WPR算法[2]是在PageRank的基础上根据网页的结构提出的一种改进算法,它根据网页重要性的不同分配不同的权值,使得重要的网页获得更大的PR值,但是它和PageRank算法一样容易出现主题漂移现象和偏重旧网页现象。Neelam Tyagi和Simple Sharma提出了基于网页的访问次数的WPR算法[3],但是它没有考虑网页的主题特征,容易出现主题漂移现象。

本文提出一种基于主题特征和最近搜索周期网页被引用频率的WPR算法,采用改进的TF-IDF算法和网页的最近搜索周期的被引用频率因子来改善主题漂移现象[4]和偏重旧网页现象[5]。

2 基本理论

2.1 WPR算法

越多的连接链向一个网页或者这个网页链向越多的网页,说明这个网页越流行。WPR算法是PageRank算法的一个扩展。它的思想是分配较大的等级值给更重要(流行)的网页,而不是一个页面中的链出网页的平均值。每个链出页面获得一个与其受欢迎程度成正比的值(链向它的和它链出的数量)。根据受欢迎程度,将一个页面的链入和链出的值分别定义为$W^{in}_{(v,u)}$和$W^{out}_{(v,u)}$。

$W^{in}_{(v,u)}$是通过计算网页u的链入数目与所有引用网页v的网页的链入数目而得出的结果:

$$W^{in}_{(v,u)}=\frac{I_u}{\sum_{p\in R(v)} I_p} \tag{1}$$

其中,I_u和I_p分别代表网页u和网页p的链入数目,$R(v)$表示所有引用网页v的网页列表。

$W^{out}_{(v,u)}$是通过计算网页u的链出数目与所有引用网页v的网页的链出数目而得出的结果:

$$W^{out}_{(v,u)}=\frac{O_u}{\sum_{p\in R(v)} O_p} \tag{2}$$

其中,O_u和O_p分别代表网页u和网页p的链出数目,$R(v)$表示所有引用网页v的网页列表。

到稿日期:2015-03-20 返修日期:2015-06-20
王旭阳(1974-),女,硕士,副教授,主要研究方向为数据挖掘、自然语言处理,E-mail:47699298@qq.com;任国盛(1987-),男,硕士生,主要研究方向为智能信息处理。

基于网页重要性的考虑，原来的 PageRank 公式修改为：

$$PR(u)=(1-d)+d\sum_{v\in B(u)}PR(v)W^{in}_{(u,v)}W^{out}_{(u,v)} \quad (3)$$

2.2 主题相关的 PageRank 算法

传统 PageRank 算法得到的权威度容易脱离用户搜索的主题范围，产生搜索结果的主题漂移，而主题相关的 PageRank 算法是将查询主题与搜索结果页面的相关性同时引入到对链接网页的 PageRank 值的迭代计算中，进而影响搜索结果的排名，以提高搜索结果的准确性。

主题相关的 PageRank 算法[6]采用相似度来度量页面与查询主题的相关性。首先从离线下载的网页中提取页面的特征项，然后根据特征项来构造索引词表，提取出现频率高的特征项并构造相应的查询主题和页面主题索引词向量空间。采用 TF-IDF(词频-逆文档频率)来计算每个特征项的权值$W_{ij}=\frac{m_j}{m}\times\log(\frac{n}{n_j}+0.01)$，$m_j$ 是特征项 j 在文档中 i 中的次数，m 是文档 i 中特征项总数，n_j 表示出现特征项 j 的次数，n 表示所有文档数。最后计算页面与查询主题的相似度：

$$sim(d_i,q)=\frac{\sum_{j=1}^{m}w_{ij}\times q_j}{\sqrt{(\sum_{j=1}^{m}w_{ij}^2)(\sum_{j=1}^{m}q_j^2)}} \quad (4)$$

其结果作为权值参与页面的 PageRank 值的计算。则 PageRank 值的计算公式为：

$$PR(A)=(1-d)+sim(A,Q)d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)} \quad (5)$$

3 改进算法

综合对上述算法的分析，根据主题网页的特征，从主题权威性、主题相关性引入了主题预测相关度、权威度并加权到 PageRank 值中。在本改进算法中新加入一个因子，即为网页的被引用频率，当一个网页的被引用次数高且被引用频率也高，那么它是非常重要的网页，权重最大；当一个新的网页的被引用总次数不高，但是被引用频率很高，说明它是比较重要的网页，权重次之；而那些被引用次数高但是被引用频率不高的网页是有一定价值的网页，权重应小于前两个权重；那些被引用次数低且被引用频率也低的网页的价值很低，权重最小。

在 WPR 算法的基础上提出改进 WPR 算法——WTFPR 算法。

$$PR(A)=(1-d)+d\times W_a\times W_c\times(\sum_{i=1}^{n}\frac{PR(T_i)}{C_i})+f' \quad (6)$$

式中，W_a 和 W_c 分别表示网页 v 的权威性权值和相关性权值，f'为被引用频率因子。

$$W_a=W^{in}_{(v,u)}=\frac{I_u}{\sum_{p\in R(v)}I_p}$$

$W^{in}_{(v,u)}$是基于入度的权重因子，I_u 是某个网页 u 的入度。根据主题网页的 Linkage/Sibling Locality 特征[7]，W_c 使用网页 U 与网页 V 的相关度来衡量，假设网页 U 与网页 V 的相关度为 $sim(U,V)$，则 $W_c=sim(U,V)$。对于 $sim(U,V)$的计算，采用了向量空间模型中计算文档相似程度的计算方法[8]：设两个文档的相似程度由表示这两个文档的向量进行内积计算。倘若网页 U 和 V 的文档向量为 $U=(U_1,U_2,\cdots,U_n)$，$V=(V_1,V_2,\cdots,V_n)$，那么它们的相关程度如下所示：

$$sim(U,V)=\frac{U\cdot V}{|U|\times|V|}=\frac{\sum_{i=1}^{n}U_iV_i}{\sqrt{\sum_{i=1}^{n}U_i^2\sum_{i=1}^{n}V_i^2}} \quad (7)$$

式中，U_i 和 V_i 为某个关键词 i 在网页 U 和 V 中的权值，该权值一般采用基于关键词频率统计的 TF-IDF 算法来计算[9]，设关键词 i 在文档 j 中的权值为 W_{ij}，则 $W_{ij}=TF_{ij}\times IDF_{ij}$，其中 TF_{ij} 为关键词 i 在文档 j 中出现的概率，IDF_{ij} 代表文档集合范围内的一种全局因子。

$$TF_{ij}=\frac{m_i}{\sum_{k=1}^{n}m_k} \quad (8)$$

$$IDF_{ij}=\log(\frac{\sum_{k=1}^{n}m_k-m_i}{m_i}+0.01) \quad (9)$$

其中，m_i 表示关键词 i 在文档 j 中出现的次数，m_k 表示关键词 k 在文档 j 中出现的次数。

在原 WPR 算法中加入被引用频率 f，主要是为了改善 WPR 算法的偏重旧网页的现象，即提高那些高质量的新网页的 PR 值，降低质量差的旧网页的 PR 值。因此 f 的值应该是网页的发布日期之类的时间函数，由于当前很多网页设计不规范，因此无法从网页中准确获取时间之类的值。考虑到一般的搜索引擎服务器的搜索周期是一个月[10]，即每隔一个月，搜索引擎就会采集一次网络上的网页，并对网页分析排序，那么可以将网页的被引用频率作为时间反馈因子，即一个网页在一个搜索周期里被引用的次数越多，则说明它质量越高，越有价值。网页的被引用频率根据搜索周期设定两种情况：①当网页第一次被搜索到时，我们认为这个网页为新生成的网页，那么它的被引用频率，即时间反馈因子设定为 $f=2\times\frac{N}{T}$，N 为新网页被引用的次数，T 为搜索周期。当 $f>1$时，$f'=1-\frac{1}{f}$，当 $f<1$ 时，$f'=(1-\frac{1}{f})/(T-1)$(进行归一化处理)。②若网页在上个周期被搜索过，那么它的被引用频率，即时间反馈因子设定为 $f=\frac{n_{new}-n_{old}}{t_{new}-t_{old}}$，$n_{new}$ 表示网页在当前搜索周期里第一次被搜索到时被引用的次数，n_{old} 表示网页在上个搜索周期里第一次被搜索到时被引用的次数，t_{new} 表示网页在当前搜索周期里第一次被搜索到的时间，t_{old} 表示网页在上个搜索周期里第一次被搜索到时被引用的时间。当 $f>1$ 时，$f'=1-\frac{1}{f}$，当 $f<1$ 时，$f'=(1-\frac{1}{f})/(t_{new}-t_{old}-1)$(进行归一化处理)。

改进算法的主要计算步骤如下：

(1)迭代计算每个网页的原始 WPR 值，即 $R(u)=\sum_{v\in B(u)}W_a\times R(v)/N(v)$；

(2)提取页面的特征项，根据特征项构造索引词表，并计算每个索引词的权值，最后根据式(4)计算页面与查询主题的相似度；

(3)根据每个网页的被引用频率计算被引用频率因子 f'；

(4)将步骤(2)和步骤(3)的结果分别与步骤(1)的结果相乘，再通过与步骤(4)的结果进行比对，按照式(6)计算出每个网页的最终 PageRank 值。

4 实验及分析

为验证改进算法的实验效果，本文从 http://www.sina.com.cn 网站获取网页数据进行仿真。

(1)网页集(M)采集。采用 Nutch 的爬虫 crawler 采集软件采集网页，总共采集 $M=754000$ 张网页，设置搜索的关键词如下：房价、高考、世界杯、反腐、大学生就业。

(2)采用 Lucene 提供的源代码对网页内容进行抽取、中文分词，以 Document(keyword1，keyword2)的形式存储在 Lucene 的存储接口中，并建立索引，其中，Document 是文档名，keyword 是关键词。采用 DOM 模型将 HTML 文档分解为 DOM 树结构，删除其中无关链接，建立 VSM 模型。根据式(6)提取文档内容和前向链接，对网页的 title 域、Meta 域、A 域和 H1 域建立索引并存储在 MySQL 数据库中。

(3)分别采用 WPR 算法和本文改进算法计算 PR 值，对网页建立索引，并按 PR 值排序保存进 MySQL 数据库。针对不同的查询，找到对应的 Lucene 索引文件，从 MySQL 数据库按 PR 值降序读出网页。

(4)对网页的 PR 值评价分析，对用户来说，越是最新的、内容越是相关的，且越具有权威的网页越容易令用户满意，也就越靠前；反之，则越靠后。

为了验证时间因子对网页 PR 值的影响，实验从一周前重复以上实验步骤。通过查询上述 5 个关键词，每个关键词搜索得到两种搜索结果，一种结果由 WPR 算法获得；另一种结果由本文改进的算法获得。为了比较两种搜索结果的效果，本文设置一个满意度 $S=\sum_{i=1}^{n}(n-i+1)s_i$，其中 n 为网页总个数，本测试选取 n 为前 20 张网页，i 为 n 个网页中的第 i 个网页，s_i 为满意系数。

满意系数根据用户的主观评价来确定，主题特征越明显且时间越近，则越满意。满意系数分为 4 个不同的等级，这 4 个等级分别为：

(1)非常满意，网页正文中含有关于查询主题极其重要的信息，而且是最近的，此时 $s_i=10$；

(2)满意，网页正文中含有与查询主题相关的信息，此时 $s_i=7$；

(3)稍微满意，虽然网页内容是最相关的，但网页正文中只含有少量的查询信息，此时 $s_i=3$；

(4)不满意，仅仅在网页的不重要的地方含有查询主题词，此时 $s_i=1$。

实验结果如表 1、图 1 及图 2 所示。

表 1 两种算法的结果比较

查询关键字	获得网页数量	WPR 算法网页相关度	改进算法网页相关度
房价	728	5.39	8.21
高考	516	4.95	5.78
世界杯	346	8.49	10.83
反腐	647	3.87	6.46
大学生就业	489	9.11	12.08

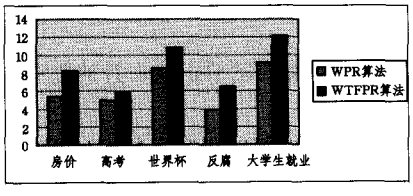

图 1 两种排序算法结果的直方图比较

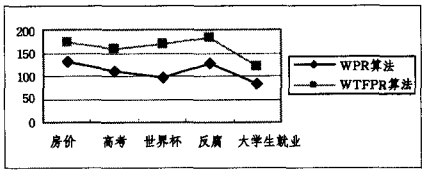

图 2 两种算法的满意度比较

由表 1 和图 1 可以看出，在相同的网页集合中，WTFPR 算法比 WPR 算法得到的网页相关度分别提高了 52%、17%、28%、67%、33%，并且对于 WTFPR 算法结果的满意度也比 WPR 算法平均提高了 47%。可以看出 WTFPR 算法的网页相关度和满意度高于 WPR 算法，改进算法即 WTFPR 算法一定程度上提高了网页排序的准确度，优化了排序的质量，具有现实可行性。

结束语 本文结合主题特征、网页权重和被引用频率 3 方面的分析，提出一种改进 WPR 的 WTPFR 算法，从网页相互链接的角度和网页内容相关的角度来解决主题网页的相关性和权威性，使质量高的网页排序靠前，质量低的网页排序下沉；通过网页的被引用频率因子的作用调整新旧网页的链接排序，使得比较权威且最新的网页排序尽量靠前。实验仿真的结果表明，改进的算法在主题特征和新旧网页的链接排序方面要比 WPR 算法更好。

WTPFR 算法在计算搜索周期里新网页的被引用频率时，使用的是搜索周期而不是新网页本身的存在时间，所以会存在一定的误差，下一步的工作可以研究如何减小新网页的被引用频率的误差，以提高网页 PR 值的准确度。

参 考 文 献

[1] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing Order to the Web[R]. Stanford Digital Libraries Working Peper, 1999

[2] Xing W, Ghorbani A. Weighted pagerank algorithm[C]// Proceedings Second Annual Conference on Communication Networks and Services Research, 2004. IEEE, 2004: 305-314

[3] Tyagi N, Sharma S. Weighted Page rank algorithm based on number of visits of Links of Web page[J]. International Journal of Soft Computing and Engineering, 2012, 2(3): 441-446

[4] Huang D, Qi H. Pagerank algorithm research[J]. Computer Engineering, 2006, 32(4): 145-146

[5] Yang J, Ling P. Improvement of PageRank Algorithm for Search Engine[J]. Computer Engineer, 2009, 35(22): 35-37

[6] Ingongngam P, Rungsawang A. Topic-centric algorithm: a novel approach to Web link analysis[C]// 18th International Conference on Advanced Information Networking and Applications, 2004(AINA 2004). IEEE, 2004, 2: 299-301

[7] Davison B D. Topical locality in the Web[C]// Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2000: 272-279

[8] Langville A N, Meyer C D. Google's PageRank and beyond: The science of search engine rankings[M]. Princeton University Press, 2011

[9] H Cheng-Hui, Y Jian, H Fang. A text similarity measurement combining word semantic information with TF-IDF method[J]. Chinese Journal of Computers, 2011, 34(5): 856-864

[10] Redlich R M, Nemzow M A. Information life cycle search engine and method: U. S. Patent 8423565[P]. 2013-4-16