浙江理工大学学报(自然科学版),第 37 卷,第 2 期,2017 年 3 月
Journal of Zhejiang Sci-Tech University (Natural Sciences)
Vol. 36, No. 2, Mar. 2017

DOI:10.3969/j.issn.1673-3851.2017.03.015

# 基于页面相似度的 PageRank 算法

王　丰,俞成海,汪佳文,徐立波
(浙江理工大学信息学院,杭州 310018)

**摘　要**:用户通过检索平台虽然可以获得大量信息,但是搜索结果往往会出现主题漂移现象,不能满足用户的实际需求。为减少这种现象的发生,提出一种改进的 PageRank 算法。该算法基于传统的 PageRank 算法,先利用向量空间模型对页面间的相似度进行计算,然后依据相似度赋予不同的调控因子,并将它们引入到 PageRank 算法中,从而使得页面 *PR* 值的计算更加合理、科学。结果表明:改进后的 PageRank 算法在搜索应用中能够有效减少了主题漂移现象,搜索结果也更加符合用户需求。

**关键词**:PageRank 算法;主题漂移;空间向量模型;相似度;互联网
**中图分类号**:TP399　　**文献标志码**:A　　**文章编号**:1673-3851 (2017) 02-0237-05

## 0　引　言

随着互联网的爆炸式发展,网络上的信息量也越来越庞大,使得对信息检索平台性能的要求也越来越高。当用户通过检索平台查询信息时,总是希望将最权威、最相关的网页信息呈现于搜索结果的最前列。如何将用户所需要的搜索结果更快、更精准地呈现给用户是检索平台当前必须要解决的主要问题之一[1]。

目前,Google 搜索引擎无疑是众多优秀检索平台中最具代表性的一员。它采用网页排序算法[2]——PageRank 算法,该算法的成功运用使得 Google 公司取得了巨大成就。PageRank 算法是基于分析网络链接的算法,该算法的模型来源于随机冲浪者模型。PageRank 算法最早是由 Sergey Brin 和 Lawrence Page 在 1998 年提出的,并最终被 Google 搜索引擎成功运用。PageRank 算法简单描述如下:假设某个页面中有 $N$ 个正向链接,该网页的 PageRank 值为 $r$,那么 $N$ 个链接获得该网页的分配权值分别为 $r/N$,然后将每个链接获得的权值分别与其相对应的链接网页的 *PR*(PageRank) 值相加,最终得到每个正向链接网页的 *PR* 值。计算如式(1):

$$PR_{(A)} = (1-d) + d\sum_{i=1}^{n} PR_{(V_i)}/C_{(V_i)} \qquad (1)$$

其中:$PR_{(A)}$ 是 $A$ 页面的 *PR* 值;$V_i$ 是指链接到页面 $A$ 的网页;$PR_{(Vi)}$ 指页面 $V_i$ 的 *PR* 值;$C_{(Vi)}$ 指 $V_i$ 页面中的所有正向连接总数;$S_{(A)}$ 是所有链入网页 $A$ 的集合;$d$ 是处于 0 到 1 之间的衰减系数,一般取值为 0.85。

## 1　相关研究

随着 PageRank 算法的广泛应用,越来越多的学者开始对该算法进行更加深入的研究[3-8],部分专家学者对 PageRank 算法提出了不足之处,同时针对其不足之处提出了不同的改进方法。例如,Xu 等[5] 提出了一种基于时间衰退特征的 PageRank 算法模型,该模型通过分析网页或文献存在的时间,根

收稿日期:2016－04－28　　网络出版日期:2017－01－03
基金项目:国家自然科学基金项目(61502430);浙江省自然科学基金项目(LY12F02041);浙江省公益技术研究工业项目(2016C31072);浙江省高校重中之重纺织科学与工程一级学科和浙江省服装工程技术研究中心优秀青年人才培养基金(2014KF15);研究生课程建设项目(111201323315O1)
作者简介:王　丰(1988－),男,江苏连云港人,硕士研究生,主要从事移动应用方面的研究。
通信作者:俞成海,E-mail:ych@zstu.edu.cn

据存在时间的长短，从而对 *PR* 值进行更有效的分配。Zhou 等[6] 提出了基于用户点击的 *PR* 模型，该算法认为网页的重要程度和用户的点击量有极大关系，点击率越高，则网页的重要程度也就越高。曾春等[7] 提出了基于内容过滤的个性化搜索算法，通过引入用户兴趣模型来对搜索的结果进行过滤，使得查询结果更加准确。虽然这些改进的算法都通过实践证明，能够提升搜索结果的准确率，但是还是会出现主题漂移现象。

在文献[9]中提出的方法中，通过提取网页中超链接的锚文本，将锚文本的内容与当前页面进行相关性的计算，利用相关程度对正向链接的 *PR* 值进行合理分配。虽然这种方法能够在一定程度上能够解决主题漂移的现象，但是还是存在不足之处。假设一个网页中的锚文本链接虽然和网页内容具有相关性，但该链接网页的内容与当前网页毫无关联，即流氓网页。该链接网页欺骗了当前网页，从而获得了当前网页分配的部分 *PR* 值，导致在搜索结果中，该流氓网页的排名更加靠前，影响搜索结果的准确度。

本文在文献[9]的基础上，对页面和该页面的正向连接页面的关键字进行提取，通过关键字对网页和每个正向连接进行页面相似度的计算，然后通过相似度给正向连接分配合理的 *PR* 值。

## 2 基于页面相似度的 PageRank 算法改进

网页 *PR* 值的高低往往决定于链入网页的多少，以及链入网页的权威性，传统的 *PR* 算法是将当前网页的 *PR* 值平均分配给它的链出页面，忽视了链出网页与当前网页的内容是否具有相关性即相似程度。改进后的 *PR* 算法通过分析当前网页及其链出网页的内容，计算它们的相似度值，将值引入到 PageRank 的计算中，使得 *PR* 值更加合理，使得搜索结果更加准确。

### 2.1 算法思路

改进思路：存在页面 $P$ 及其它的一个正向链接页面 $Q$，若页面 $P$ 和页面 $Q$ 的主题和内容的相似度越高，那么 $P$ 分配给 $Q$ 的权重也应该越高。在页面中，关键词出现的位置以及出现的频率能够有效的体现出它与该页面的相关度，然后通过 $P$、$Q$ 两个网页的相关度来计算页面 $P$、$Q$ 之间的相似度。

计算 $P$、$Q$ 页面相似度的步骤：

a) 提取网页 $P$ 的关键字并建立关键字集合 $\{PKEY_i \mid PKEY_i$ 是网页 $P$ 的第 $i$ 个关键字$\}$。

b) 计算网页 $P$ 的关键字的方位权重 $PW_i$。根据 HTML 标签的特点，将处于不同标签中的关键字进行方位权值的分配，即方位权重。方位权重的分配情况如式(2)：

$$PW_i=\begin{cases}2.0,\text{关键字位于}<TITLE>\text{标签中}\\1.8,\text{关键字位于}<HEAD>\text{标签中}\\1.5,\text{关键字位于}<META>\text{标签中}\\1.0,\text{其他}\end{cases} \tag{2}$$

c) 统计出网页 $P$ 关键字 $PKEY_i$ 在不同标签中出现的次数 $PCK_{ij}(j=1,2,3,4)$。

d) 计算网页 $P$ 的中关键字权重。关键字权重 $PWK_i$ 的计算表示如式(3)：

$$PWK_i=\Big(\sum_{j=1}^{4}\text{方位权重}\ PW_j\times\text{关键字}\ PKEY_i\ \text{在该位置出现的次数}\Big)\Big/\Big(\sum_{i=1}^{n}\sum_{j=1}^{4}\text{方位权重}\ PW_j\times\text{关键字}\ PKEY_i\ \text{在该位置出现的次数}\Big) \tag{3}$$

其公式如式(4)所示：

$$PWK_i=\frac{\sum_{j=1}^{4}PW_j\times PCK_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{4}PW_j\times PCK_{ij}}\quad(i=1,2,\cdots,n) \tag{4}$$

通过式(4)的计算得到页面 $P$ 中每个关键字的权重集合 $\boldsymbol{T}_x$，其中 $\boldsymbol{T}_x=\{(PWK_i,PKEY_i),1\leqslant i\leqslant n\}$。

e) 按照类似的步骤，提取链接网页 $Q$ 的关键字集合以 $\{QKEY_i \mid {}_QKEY_i$ 是网页 $Q$ 的第 $i$ 个关键字$\}$ 以及计算出每个关键字在网页 $Q$ 的权值集合 $\boldsymbol{T}_x^*=\{(QWK_i,QKEY_i),1\leqslant i\leqslant n\}$。

f) 网页 $P$、$Q$ 分别可由空间向量 $\boldsymbol{T}_x=\{(PWK_i,PKEY_i)\}$、$\boldsymbol{T}_x{}^*=\{(QWK_i,QKEY_i)\}$ 表示，计算 $P$ 和 $Q$ 的相似度就是计算 $\boldsymbol{T}_x$、$\boldsymbol{T}_x^*$ 的相似度。本文通过 VSM(vector space model，向量空间模型)[10] 计算网页 $P$、$Q$ 间的相似度即余弦相似度[11]，如式(5)：

$$Sim(P,Q)=Sim(\boldsymbol{T}_x,\boldsymbol{T}_x^*)=\cos\theta=\frac{\sum_{i=1}^{n}PWK_i\times QWK_i}{\sqrt{(\sum_{i=1}^{n}PWK_i^2)(\sum_{i=1}^{n}QWK_i^2)}} \tag{5}$$

其中：$Sim(P,Q)$ 是网页 $P$ 和 $Q$ 的相似度；$\cos\theta$ 是 $P$、$Q$ 在空间向量中的余弦值。

从式(5)中得出，两个网页的相似度使介于 0 和 1 之间的，若计算的结果越靠近 0，说明两者的相似度也就越小，则链出的网页被分配的权重也就相对

较低;若计算的结果靠近 1,说明两者的内容相似度很高,则链出的网页被分配的权重也就相对较高。

计算得到的两个网页的相似度,在将相似度引入 *PR* 值计算的时候,引入调控因子 $\lambda(\lambda>0)$,如式(6):

$$\lambda = \Theta_{(\cos\theta)} \tag{6}$$

式中 $\cos\theta$ 是两个页面的相似度的余弦值,$\Theta(\cos\theta)$ 的取值如式(7)。

$$\Theta_{(\cos\theta)} = \begin{cases} 0.2, 0 \leqslant \cos\theta \leqslant 0.4 \\ 0.5, 0.4 < \cos\theta \leqslant 0.7 \\ 0.8, 0.7 < \cos\theta \leqslant 1 \end{cases} \tag{7}$$

最终得到改进后的 PageRank 算法如式(8):

$$PR(A) = (1-d) + d \times \sum_{i=1}^{n} \frac{PR(V_i)}{C(V_i)} \times (1 + \lambda \times Sim(A, T_i)) \tag{8}$$

其中:$\lambda(\lambda>0)$ 是调控因子,$T_i$ 是 $V_i$ 页面中的指向 $A$ 的一个链接网页。

## 2.2　算法分析

图 1 是一个简单的页面链接关系的有向图 $\boldsymbol{G}=<V,E>$,其中 $V$ 表示网页节点集合,$E$ 表示网页节点之间的链接关系集合。假设页面的所有关键词都在 $<TITLE>$ 标签中,所有的关键词的集合为 $\boldsymbol{Key}=\{k1,k2,k3,k4,k5,k6\}$。其中页面 $A$ 的关键词 $k1,k2,k3,k5$,关键词次数分别为 4,1,1,1;页面 $B$ 的关键词 $k2,k5,k1$,关键词次数分别为 2,1,1;页面 $C$ 的关键词 $k6,k3,k4$,关键词次数分别为 1,2,2;页面 $D$ 的关键词 $k2,k1,k5$,关键词次数分别为 1,3,3;页面 $E$ 的关键词:$k1,k3,k4,k6$,关键词次数分别为 1,1,2,2。

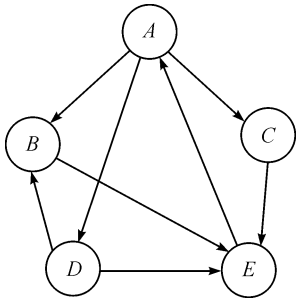


图 1　页面链接关系

运用传统的 PageRank 算法计算每个页面,页面的初始 *PR* 值 $1/M$($M$ 为页面总数),经过多次迭代得到表 1。

**表 1　传统算法的 *PR* 值**

| 迭代次数 | *PR*(*A*) | *PR*(*B*) | *PR*(*C*) | *PR*(*D*) | *PR*(*E*) |
|---|---|---|---|---|---|
| 0 | 0.200000 | 0.200000 | 0.200000 | 0.20000 | 0.200000 |
| 1 | 0.249583 | 0.159625 | 0.086667 | 0.066667 | 0.344500 |
| 5 | 0.259073 | 0.185533 | 0.111858 | 0.185533 | 0.258004 |
| 13 | 0.172668 | 0.180685 | 0.103636 | 0.18085 | 0.262327 |
| 27 | 0.272352 | 0.181031 | 0.104094 | 0.181031 | 0.261492 |
| 29 | 0.272357 | 0.181028 | 0.104089 | 0.181028 | 0.261492 |

表 2 是和表 3 分别是页面的相似度和改进后算法得到的 *PR* 值,*PR* 初始值都为 $1/M$($M$ 为页面总数),$\lambda$ 设置为 0.5。如果页面之间有链接关系,计算其相似度,否则相似度为 0。

**表 2　页面相似度**

| 相似度 | A 页面 | B 页面 | C 页面 | D 页面 | E 页面 |
|---|---|---|---|---|---|
| A 页面 | 1.0000 | 0.6556 | 0.1529 | 0.8421 | 0.3627 |
| B 页面 | 0.6556 | 1.0000 | 0 | 0.7493 | 0.1291 |
| C 页面 | 0.1529 | 0 | 1.0000 | 0 | 0.8433 |
| D 页面 | 0.8421 | 0.7493 | 0 | 1.0000 | 0.2176 |
| E 页面 | 0.3627 | 0.1291 | 0.8433 | 0.2176 | 1.0000 |

**表 3　改进算法 PR 值**

| 迭代次数 | *PR*(*A*) | *PR*(*B*) | *PR*(*C*) | *PR*(*D*) | *PR*(*E*) |
|---|---|---|---|---|---|
| 0 | 0.200000 | 0.200000 | 0.200000 | 0.200000 | 0.200000 |
| 1 | 0.200000 | 0.159623 | 0.066667 | 0.066667 | 0.445302 |
| 10 | 0.332931 | 0.337729 | 0.099542 | 0.086993 | 0.443210 |
| 40 | 0.564431 | 0.443253 | 0.139634 | 0.121426 | 0.385430 |
| 81 | 0.600157 | 0.554732 | 0.266733 | 0.201459 | 0.392130 |

通过分析表1—表3发现，引入相似度对 *PR* 值进行计算，对原有的 *PR* 值进行了修正。因为页面 *A* 与其他页面的相关性较大，所以该页面获得 *PR* 值也较高。从上述的分析中说明改进后的算法是有效可行的。

## 3 实验及结果分析

### 3.1 实验过程

为了验证改进算法的精准度，本文通过 Luncene 的网络爬虫程序[12]对 http://sina.com/进行页面的抓取，抓取页面数量为 18453，将抓取的每个网页分为 4 个域，分别为 title、head、meta 及其它，建立相关的索引并将其存入数据库中。索引建立以后对关键字进行搜索，并用传统的 PageRank 算法和改进后的 PageRank 算法计算网页的 *PR* 值，将这些 *PR* 值保存到数据库中。在进行关键字查询时，根据需求找到对应的文件，并按照 *PR* 值的降序返回结果。

在验证过程中，取 10 个不同的关键字（旅游、大数据、足球、百慕大三角、电脑、工作、国内外趣闻、星级酒店、UFO、阅兵），请 6 名不同测试人员（程序员、大四学生、喜欢旅游的大学生、工作的大学生、硕士研究生、研究生导师）对其进行测试。测试人员分别用传统的算法和改进后的算法进行搜索测试，对每个检索词所返回的结果列表，测试人员都会打出满意度的分数（评分的范围是 0.0～0.5），其值为 SP@NSP，其中 SP 是对传统 PageRank 算法的满意值，NSP 是对改进算法的满意值。

### 3.2 结果分析

在实验过程中，以每个关键词搜索以返回结果列表前 15 项为研究基础，来分析用户对搜索结果的满意程度，然后通过用户满意度来对改进算法进行测评。表 4 和图 2 是 6 名测试人员对于返回结果的评分。

**表 4 用户满意度表 SP@NSP**

| 关键词 | 测试人员一 | 测试人员二 | 测试人员三 | 测试人员四 | 测试人员五 | 测试人员六 | 均值 |
|---|---|---|---|---|---|---|---|
| 旅游 | 0.25/0.25 | 0.32/0.32 | 0..25/0.15 | 0.15/0.15 | 0.24/0.13 | 0.2/0.2 | 0.2283/0.2067 |
| 大数据 | 0.25/0.37 | 0.1/0.2 | 0.1/0.1 | 0.1/0.23 | 0/0.12 | 0.1/0.15 | 0.1083/0.195 |
| 足球 | 0.3/0.35 | 0.1/0.23 | 0.1/0.1 | 0.2/0.25 | 0.2/0.15 | 0/0.15 | 0.15/0.205 |
| 百慕大三角 | 0.12/0.2 | 0.1/0.2 | 0.24/0.32 | 0.2/0.2 | 0.1/0.15 | 0.15/0.15 | 0.1517/0.2033 |
| 电脑 | 0.18/0.26 | 0.3/0.34 | 0.3/0.3 | 0.25/0.25 | 0.2/0.25 | 0.23/0.32 | 0.2433/0.2867 |
| 工作 | 0.2/0.4 | 0.15/0.35 | 0.1/0.3 | 0.2/0.2 | 0.26/0.34 | 0.35/0.35 | 0.3267/0.3233 |
| 国内外趣闻 | 0.25/0.3 | 0.2/0.25 | 0.2/0.2 | 0.27/0.3 | 0.3/0.3 | 0.25/0.25 | 0.245/0.2667 |
| 星级酒店 | 0.12/0.2 | 0.2/0.2 | 0.1/0.25 | 0.2/0.2 | 0.15/0.2 | 0.23/0.23 | 0.1667/0.2133 |
| UFO | 0.32/0.35 | 0.3/0.3 | 0.25/0.36 | 0.2/0.2 | 0.1/0.25 | 0.2/0.2 | 0.2283/0.2767 |
| 阅兵 | 0.2/0.35 | 0.2/0.2 | 0.3/0.35 | 0.24/0.3 | 0/0.3 | 0.25/0.25 | 0.1983/0.275 |
| 综合评价 | | | | 0.2047/0.2452 | | | |

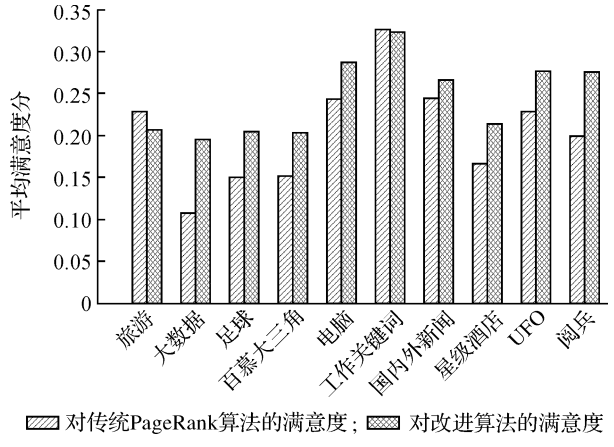


图 2 传统 PageRank 算法和改进后算法满意度对比

从表 4 和图 2 的实验数据可以看出，传统的 PageRank 算法的准确度总体来说要低于改进后的算法，改进后的算法将准确度提高了 19.785%左右。通过数据对比也说明了将页面相似度引入 PageRank 值的计算，能够有效的提高搜索结果的准确度。

## 4 结 语

本文通过页面之间的相似度对传统的 PageRank 算法进行改进，在一定的程度上解决了主题漂移问题。通过实验证明改进后的算法是可行的，其搜索结果准确度也高于传统算法搜索的准确度。接下来在此基础上，将继续对 PageRank 算法进行更深一步的研究，通过引入一些其他相关因素，如时间因子、用户的收藏情况、用户的兴趣度等等，以便充分完善 PageRank 算法的计算，最终让搜索排序的结果更加符合用户的需求。

**参考文献:**

[1] PRAKASHA S, SHASHIDHAR H, RAJU G T. Structured intelligent search engine for effective information retrieval using query clustering technique and semantic Web[C]// Contemporary Computing and Informatics (IC3I), 2014 International Conference on . IEEE, 2014:688-695.

[2] JAIN A, SHARMA R, DIXIT G, et al. Page ranking algorithms in Web mining, limitations of existing methods and a new method for indexing Web pages [C]// International Conference on Communication Systems and Network Technologies. IEEE, 2013:640-645.

[3] HE X, LI Y, FAN C. Web-based links and authoritative content pagerank improvement [C]// E-Business and E-Government (ICEE), 2010 International Conference on. IEEE, 2010:5016-5019.

[4] LEI J, CHEN H F. Distributed randomized PageRank algorithm based on stochastic approximation [J]. IEEE Transactions on Automatic Control, 2015, 60(6):1641-1646.

[5] XU H, MARTIN E, MAHIDADIA A. Contents and time sensitive document ranking of scientific literature [J]. Journal of Informetrics, 2014, 8(3):546-561.

[6] ZHOU C L, CHEN K, LI S S. Improved PageRank algorithm based on feedback of user clicks [C]// Computer Science and Service System (CSSS), 2011 International Conference on. IEEE, 2011:3949-3952.

[7] 曾春,邢春晓,周立柱. 基于内容过滤的个性化搜索算法[J]. 软件学报, 2003, 14(5):999-1004.

[8] GONG C, FU K, LOZA A, et al. PageRank tracker: from ranking to tracking. [J]. IEEE Transactions on Cybernetics, 2013, 44(6):882-893.

[9] 王钟斐, 王彪. 基于锚文本相似度的 PageRank 改进算法[J]. 计算机工程, 2010, 36(24):258-260.

[10] NAKANISHI T. Semantic context-dependent weighting for vector space model [C]// IEEE International Conference on Semantic Computing. IEEE, 2014:262-266.

[11] 黄承慧,印鉴,侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011, 34(5):856-864.

[12] 周德懋,李舟军. 高性能网络爬虫:研究综述[J]. 计算机科学, 2009, 36(8):26-29.

# PageRank Algorithm Based on Page Similarity

*WANG Feng, YU Chenghai, WANG Jiawen, XU Libo*

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

**Abstract**: Users can get a lot of information through the search platform,, but the theme drift phenomenon often appears to search results. Thus, users' actual needs cannot be met. In order to reduce the occurrence of this phenomenon, an improved PageRank algorithm is proposed. The algorithm based on traditional PageRank algorithm, first applies the vector space model (VSM) to calculate the similarity between pages, then gives different regulatory factors according to the similarity, introduces them to the PageRank algorithm, and finally makes *PR* value calculation more reasonable and scientific. The result shows that the improved PageRank algorithm can effectively reduce the theme drift phenomenon in the search application, and the search results are more in line with users' needs.

**Key words**: PageRank algorithm; theme drift; VSM; similarity; internet

**(责任编辑: 陈和榜)**