语义相似的 PageRank 改进算法

何 明[1],周 军[1],李树友[2]
HE Ming[1],ZHOU Jun[1],LI Shu-you[2]

1.辽宁工业大学 电子与信息工程学院,辽宁 锦州 121000
2.辽宁工业大学 数理科学系,辽宁 锦州 121000
1.College of Electronic and Information Engineering,Liaoning University of Technology,Jinzhou,Liaoning 121000,China
2.Department of Mathematical Science,Liaoning University of Technology,Jinzhou,Liaoning 121000,China
E-mail:heming0405@163.com

HE Ming,ZHOU Jun,LI Shu-you.New semantic similarity PageRank algorithm.Computer Engineering and Applications, 2009,45(27):140-142.

Abstract: The PageRank algorithm is used in ranking web pages.It estimates the pages' authority by taking into account the link web structure.However,it only considers the structure of webs' link,without any topic about webs,so this algorithm leads to topic-drift.After analyzing the original PageRank algorithm several times,an improved PageRank algorithm based on semantic similarity is proposed.This new PageRank algorithm can calculate the PageRank's value according to the structure and content about web,it can avoid the topic-drift problem and improve the quality of web search effectively without adding any other extra time and space complexity degree.

Key words: PageRank algorithm;topic-drift;semantic similarity

摘 要: PageRank 算法是一种用于网页排序的算法,它利用网页间的相互引用关系评价网页的重要性。但由于它只考虑网页与网页之间的链接结构,忽略了网页与主题的相关性,容易造成主题漂移现象。在分析了原 PageRank 算法基础上,给出了一种基于语义相似度的 PageRank 改进算法。该算法能够按照网页结构和网页主要内容计算出网页的 PageRank 值,既不会增加算法的时空复杂度,又极大地减少了"主题漂移"现象,从而提高查询效率和质量。

关键词: PageRank 算法;主题漂移;语义相似度
DOI:10.3778/j.issn.1002-8331.2009.27.042 **文章编号:**1002-8331(2009)27-0140-03 **文献标识码:**A **中图分类号:**TP391.3

1 引言

Pagerank 算法是 Google 搜索引擎的核心算法,它是 1998 年由美国斯坦福大学 Larry Page 与 Sergey Brin 提出的[1]。PageRank 算法是一种与查询不相关的,对全球 Web 页面排序的算法[2]。在 PageRank 算法中,Web 页面的重要性是由互联网络超链接的拓扑结构决定的,也是通过对互联网中超链接结构的挖掘而获得的。PageRank 算法是最早将链接分析技术应用于搜索引擎的算法,也是一种能够自动判断网页客观重要性的算法。它的主要思想是:一个页面被多次引用,则这个页面很可能是重要的;一个页面尽管没有被多次引用,但被一个重要页面引用,则这个页面的重要性被均匀地传递到它所引用的页面[3]。也就是说网页被引用的次数越多,它自身的 PageRank 值可能就会越大,而一个网页引用其他网页的次数越多,它的 PageRank 值就会越小。PageRank 算法具有快速响应、成功率高等优点,被成功地应用于 Google 搜索引擎。然而,PageRank 也存在一些缺陷,例如算法过分强调链入链接而贬低链出链接、偏重于旧网页、忽视专业站点、产生"主题漂移"现象[4]等问题。

针对 Pagerank 算法产生"主题漂移"现象的问题,提出一种基于语义相似的 Pagerank 算法的改进算法。语义相似度是一种基于语言学和人工智能的词语相似的度量,网页的语义相似度是网页内容相似程度的一种刻画。由于在 PageRank 算法中加入了网页间内容相似的相关度量,改善了 PageRank 算法的忽视专业站点的缺陷,防止了"主题漂移"现象的发生。同时,也不会增加算法的时空复杂度。实验证明了算法的有效性。

2 网页的语义相似度

网页的语义相似度处理是聚类分析的一种,它用于计算网页文档之间的距离。是自然语言处理、智能检索、文档分类、自

基金项目:辽宁省教育厅科学研究基金(the Scientific Research Foundation of Liaoning Provincial Department of Education under Grant No. 20060409);辽宁省高校优秀青年骨干教师基金(the Backbone of Excellent Young Teachers in Colleges and Universities of Liaoning Province Foundation)。

作者简介:何明(1982-),男,硕士研究生,主要研究领域为 Web 结构挖掘,搜索引擎算法优化;周军(1966-),女,教授,博士,主要研究领域为数据挖掘、知识发现、近似计算等;李树友(1964-),男,教授,博士,主要研究领域为统计推断、随机信息处理。

收稿日期:2008-11-03 **修回日期:**2009-01-19

动应答、词义排歧和机器翻译等很多领域的基础研究课题。网页的相似主要分为:结构相似和内容相似[5]。原 PageRank 算法已经将结构相似考虑在内,该文主要是对网页内容上的相似加以阐述。而网页在内容上是否相似是取决于网页的关键字在语义词典中词语的相似度计算的基本思想,它是一种基于语言学和人工智能的理性方法[6-7]。它利用语义词典,依据概念之间的上下位关系和同义关系,通过计算两个词语概念在树状概念层次体系中的距离来得到词语间的相似度。这种方法直观、简单有效且易于理解。

网页语义相似度函数是以知网[8]为基础的词汇相似度计算,它是当前较好的方法。为方便处理,针对关键词和非关键词的界定是以知网为标准,即知网中出现的词语是关键词,否则为非关键词。从信息论的角度来说,两个网页内容的相似度不与其个性有关,应与其共性有关。在图 1 中,如果两个网页的关键词一个是"鱼"另一个是"水果",它们之间的相似度取决于它们不同的语义距离,另一方面还与它们所包含的共同部分密切相关,即应由 D_1、D_2、D_3 这三个参数共同决定最终取值。为方便说明,做如下定义:

定义 1(义原深度) 指义原 P 在树状概念义原层次体系中所处的层数位置,记为 $Depth(P)$,其中 $Depth(P)\in Z$,Z 为整数。例如,在图 1 中,$Depth$(鱼)=7。

为方便逻辑处理,把知网中各类不同的义原体用一个虚拟根节点统一起来,构成一个互相联系的有机体,并规定根节点的义原深度为 0,它的子节点深度为 1,其他以此类推。

定义 2(重合度(Superposed Degree)) 指两个义原 P_1,P_2 在树状概念义原层次体系中所拥有的相同父节点的路径长度,记为 $Spd(P_1,P_2)$,其中 $Spd(P_1,P_2)\in Z$,Z 为整数。例如,在图 1 中,Spd(鱼,水果)=4。

定义 3(相异度(Dissimilitude Degree)) 指 2 个义原 P_1 和 P_2 在树状概念义原层次体系中沿父节点逐步上移,直到二者到达第一个共同节点,所走过的最短路径长度,记为 $Dsd(P_1,P_2)$,其中 $Dsd(P_1,P_2)\in Z$,Z 为整数。相异度与语义距离等价。例如,在图 1 中,Dsd(物质,植物)=2。

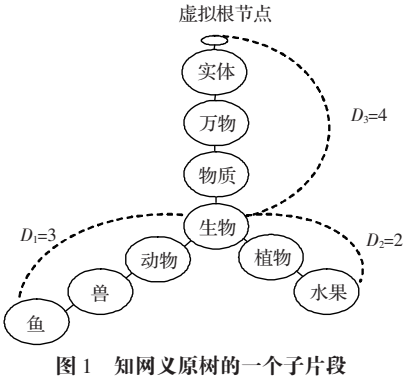

图 1　知网义原树的一个子片段

定义 4 两个网页内容为 P_1,P_2,定义网页内容相似度 $Sim(P_1,P_2)$计算公式:

$$Sim(P_1,P_2)=\frac{2\times Spd(P_1,P_2)}{Dpd(P_1,P_2)+2\times Spd(P_1,P_2)}=\frac{2\times Spd(P_1,P_2)}{Depth(P_1,P_2)+Depth(P_1,P_2)}$$

其中,$Sim(P_1,P_2)\in(0,1)$,$Spd(P_1,P_2)$为网页 P_1,P_2 的重合度,$Dsd(P_1,P_2)$为网页 P_1,P_2 的相异度,$Depth(P)$为网页的义原深度。例如,在图 1 中 Sim(鱼,水果)=(2×4)/(7+6)=8/13。

3　PageRank 算法的改进

3.1　PageRank 算法的简介

PageRank 算法是由美国斯坦福大学的工作人员开发的,用来评价 Google 搜索引擎的页面质量的算法[9]。该算法基于"被优质的网页引用的网页,必定还是优质网页"的回归关系,来判定所有网页的重要性。PageRank 算法计算出网页的 PageRank 值,来决定网页在结果集中出现的位置,PageRank 值越高的网页,在结果中出现位置就越靠前。如果网页 A 存在一条指向网页 P 的超链接,则认为 P 得到了 A 的认可,如果有许多网页指向网页 P,则可以说 P 相对比较重要。计算 PageRank 的公式为[10]:

$$PR(P)=(1-d)+d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}$$

式中 $T_i(i=1,2,\cdots,n)$为指向网页 P 的其他网页;d 为界于(0,1)区间的衰减系数;$C(T_i)$为网页 T_i 向外指出的链接数目。通过简单的迭代算法可以计算出 $PR(P)$的值。

3.2　PageRank 算法的改进

搜索引擎是根据 PageRank 算法计算出来的 PR 值来决定网页在镜像中的排位顺序。网页的 PR 值不仅和该网页在网站中的链接结构有关,还和该网页的主题内容有着密不可分的联系,而原 PageRank 算法是没有考虑到这个问题的。因此它对网页在镜像中的排位是不准确、不全面的,原 PageRank 算法只是单纯地认为:一个网页只要有链入的增加,它的 PR 值就应该相应地增加,该文考虑到网页链出链接是网页作者对该网页主题的理解,这些链出的链接必定指向那些作者认为与自己主题相关的网页,所以可以把网页链出链接作为一个简单的锚页来看,因此对 PageRank 算法修正为:

$$PR(P)=(1-d)+d\sum_{i=1}^{n}Sim(P,T_i)\frac{PR(T_i)}{C(T_i)}$$

其中 $Sim(P,T_i)\in(0,1)$ 为网页 P 与网页 T_i 的相似度,$d\in(0,1)$为衰减系数,通常取 d=0.85,$T_i(i=1,2,\cdots,n)$为指向网页 P 的其他网页,$C(T_i)$为网页 T_i 向外指出的链接数目,算法的每个网页的 PR 初值为 1。通过修正的 PageRank 计算公式可以看出,当网页 P 与网页 T_i 的相似度大时,网页 T_i 对网页 P 的贡献程度也就大,计算出来的网页 P 的 PR 值相应的也就高;当网页 P 与网页 T_i 的相似度小时,网页 T_i 对网页 P 的贡献程度也就小,计算出来的网页 P 的 PR 值相应的也就低。这符合网络实际问题的情况,更大大改善搜索引擎的查全率、查准率。

4　实验分析

为了更进一步理解和验证改进后的 PageRank 算法,根据改进后的 PageRank 算法计算公式对如下网页进行了实验,同时分析了该算法所得到的排序结果,并与利用改进前的 PageRank 算法所得到的结果进行比较。假设有 5 个包含各不相同主题(鱼,兽,万物,植物,水果)的网页,它们相互之间有链接,其结构如图 2。

由网页内容相似度可知:

$$\sigma_{12}=Sim(1,2)=\frac{2\times 6}{7+6}=\frac{12}{13}$$

$$\sigma_{14}=Sim(1,4)=\frac{2\times 4}{7+5}=\frac{2}{3}$$

$$\sigma_{23}=Sim(2,3)=\frac{2\times 2}{6+2}=\frac{1}{2}$$

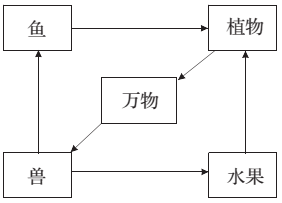

图 2 网络结构图

$$\sigma_{34}=Sim(3,4)=\frac{2\times2}{5+2}=\frac{2}{5}$$

$$\sigma_{25}=Sim(2,5)=\frac{2\times4}{6+6}=\frac{2}{3}$$

$$\sigma_{45}=Sim(4,5)=\frac{2\times5}{5+6}=\frac{10}{11}$$

其中 $\sigma_{12},\sigma_{14},\sigma_{23},\sigma_{34},\sigma_{25},\sigma_{45}$ 为网页 1 与网页 2,网页 1 与网页 4,网页 2 与网页 3,网页 3 与网页 4,网页 2 与网页 5,网页 4 与网页 5 的网页内容相似度值。为了计算方便 d 取 0.5,则各网页 PR 值计算公式如下:

$$PR(1)=0.5+0.5\times\frac{\sigma_{12}PR(2)}{2}$$

$$PR(2)=0.5+0.5\times\sigma_{23}PR(3)$$

$$PR(3)=0.5+0.5\times\sigma_{34}PR(4)$$

$$PR(4)=0.5+0.5\times(\sigma_{14}PR(1)+\sigma_{45}PR(5))$$

$$PR(5)=0.5+0.5\times\frac{\sigma_{25}PR(2)}{2}$$

各网页的 PR 值经过迭代计算如表 1。

表 1 网页 *PR* 值

次数	*PR*(1)	*PR*(2)	*PR*(3)	*PR*(4)	*PR*(5)
0	1.000 000	1.000 000	1.000 000	1.000 000	1.000 000
1	0.730 769	0.750 000	0.700 000	1.287 879	0.666 667
2	0.846 154	0.675 000	0.757 576	1.046 620	0.625 000
3	0.655 769	0.689 394	0.709 324	1.066 142	0.612 510
4	0.659 099	0.677 331	0.713 228	1.063 846	0.617 066
5	0.659 109	0.678 910	0.713 452	1.061 245	0.617 536
6	0.659 028	0.678 954	0.713 423	1.061 274	0.617 645
7	0.659 046	0.678 982	0.713 458	1.061 270	0.617 624
8	0.659 031	0.678 951	0.713 452	1.061 277	0.647 620
9	0.659 038	0.678 954	0.713 457	1.061 275	0.647 624

迭代到第 9 次时,PR 值基本上收敛。这时,可以认为迭代结束。$PR(1)=0.659\,038$,$PR(2)=0.678\,954$,$PR(3)=0.713\,457$,$PR(4)=1.061\,275$,$PR(5)=0.647\,624$,则有 $PR(4)>PR(3)>PR(2)>PR(1)>PR(5)$。如果按照原来的 PageRank 算法发现网页1 和网页 5 在网络中的链接结构是相同的,则计算出来的结果 $PR(4)>PR(3)>PR(2)>PR(1)=PR(5)$,搜索引擎将无法判断网页 1 和网页 5 的排名顺序。实验的结果表明:利用改进后的 PageRank 算法计算公式对检索结果进行排序,可以有效地提高检索结果的查准率。

5 结束语

搜索引擎技术是一个正在迅速发展的研究领域,用户对搜索结果的准确性和全面性也在日益苛求。因此,搜索引擎的算法也是不断增多,也综合了多种有价值的算法去决定一个网页在搜索引擎镜像中的排名。而网页的排名不仅和网页在网络中的链接结构有关,而且还和网页的主题特征密切相连。这是原 PageRank 算法无法解决的问题,将相似度引入了 PageRank 算法计算公式当中,恰恰解决了这个棘手的问题,也将信息检索和 Web 挖掘有效地结合在一起。相信不久的将来会有更多的研究成果出现,能够帮助网络用户在信息的海洋中更快速、更准确地找到需要的信息,实现搜索引擎的最终目标。

参考文献:

[1] Brin S,Page L.The anatomy of a large-scale hypertextual Web search engine[J].International Journal of Approximate Reasoning,1998,156(4):134-141.

[2] Furnkranz J.Web structure mining exploiting the graph structure of theWorld Wide Web[J/OL].Computer Networks,2005,34(1):699-711.http://www.Elsevier computer science.com.

[3] 刘金桂,李绪蓉.基于网页相似度的 PageRank 算法的改进[J].淮阴工学院学报:自然科学版,2006,15(1):8-11.

[4] Spertus E.Mining structural information on the Web[J].ComputationalStatistics&Data Analysis,2006,43(4):244-257.

[5] Runkler T A,Bezdek J C.Web mining with relational clustering[J].International Journal of Approximate Reasoning,2003,32(7):217-236.

[6] 黄德才,戚华春,钱能.基于主题相似度模型的 TS-PageRank 算法[J].小型微型计算机系统,2007,28(3):510-514.

[7] 张承立,陈剑波,齐开悦.基于语义网的语义相似度算法改进[J].计算机工程与应用,2006,32(3):17-23.

[8] 刘群,李素建.基于《知网》的词汇语义相似度计算[C]//第 3 届中文词汇语义学研讨会论文集,2002,21(3):99-104.

[9] 刘悦,程学旗,李国杰.提高 pagerank 算法效率的方法初探[J].计算机科学,2002,19(3):33-39.

[10] 宋聚平,王永成,尹中航,等.对网页 PageRank 算法的改进[J].上海交通大学学报,2003,37(3):105-110.

(上接 106 页)

[3] Wang Qin,Cao Zhenfu,Wang Shengbao.Formalized security model of multi-proxy signature schemes[C]//Proceedings of the 5th International Conference on Computer and Information Technology,2005:668-672.

[4] Elkamshoushy,AbouAlsoud D H,Madkour A K,et al.New proxy signcryption scheme with DSA verifier[C]//Radio Science Conference,2006:1-8.

[5] Ong H,Schnorr C P,Shamir A.An efficient signature scheme based on quadratic equations[C]//Proceedings of the 16th Annual ACM Symposium on Theory of Computing,1984:208-216.

[6] Shao Zuhua.Proxy signature schemes based on factoring[J].Information Processing Letters,2003,85(3):137-143.

[7] Xue Qingshui,Cao Zhenfu.Factoring based proxy signature schemes[J].Journal of Computational and Applied Mathematics,2006,195(1):229-241.

[8] 李子臣,戴一奇.二次剩余密码体制的安全性分析[J].清华大学学报:自然科学版,2001,41(7):80-82.

[9] 胡振鹏,钱海峰,李志斌.一种新的代理多重盲签名方案[J].计算机应用,2007:2718-2721.