

基于网页相似度的 Page Rank 算法的改进

刘金桂, 李绪蓉

(南京航空航天大学 信息科学与技术学院, 南京 210016)

摘 要:随着 Internet 上的信息量迅速增长, 用户对搜索结果的查准率提出了更高的要求。通过对 PageRank 算法进行分析, 指出 PageRank 算法不足之处, 同时提出了改进方案, 改进后的 PageRank 算法考虑了网页之间的相似度, 可提高检索结果的查准率。

关键词:信息检索; 相似度; Page Rank 算法

中图分类号: G252.7

文献标识码: A

文章编号: 1009-7961(2006)01-0008-04

A New Page Rank Algorithm Based on the Similarity of the Web Pages

LIU Jin-gui, LI Xu-rong

(College of Information Science and Technology,

Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China)

Abstract: With the increase of information on the internet, people have made more request of the precision of the search results. This paper has analyzed the Page Rank algorithm, and pointed out the disadvantages of the Page Rank algorithm. Furthermore, the improved algorithm is presented, which takes into account the similarity of the web pages and can improve the precision of the search results.

Key words: information retrieval; similarity; Page Rank algorithm

0 引 言

Internet 的诞生与发展一直是众所瞩目的焦点之一。随着信息社会的到来, 因特网作为信息交流的中心与枢纽作用也愈显重要。因特网可以称之为一个巨大的信息库, 它拥有众多但却杂乱无章的信息, 并且这些信息时刻都在以几何级数递增。但是, 网络在快捷、方便地带来大量信息的同时, 也带来了许多问题, 诸如信息超载, 信息真伪难辨, 信息安全难以保证, 信息形式不一致、难以统一处理等。用户要在如此巨大的信息海洋里查找所需信息, 就象大海捞针一样, 而搜索引擎技术恰好解决了这一难题。它可以为用户提供信息检索服务, 在一定程度上满足用户对特定信息的查询要求。目前, 搜索引擎技术正成为计算机工业界和学术界争相研究、开发的对象。如何快速、准确地从海量数据中获得有价值的信息资源, 是评价搜索引擎搜索效率的重要的指标。

在众多搜索系统所采用的算法中, 超链分析技术是很多研究者主要研究的问题, 而享有盛名的搜索引擎 Google 所采用的 PageRank 算法尤其得到认可, 从实际应用来看, 这种算法也确实解决了一些问题, 但同时由于该算法存在着一定的不足, 所以使得搜索结果的查准率不够理想。本文将对 PageRank 算法进行分析, 同时提出新的 PageRank 算法改进方案, 以改进该算法的不足之处, 从而提高检索结果的查准率。

1 PageRank 算法分析

1.1 PageRank 算法

PageRank 算法由 Standford 大学的 Brin 和 Page 提出, 它为 Web 上的每一个网页赋予一个全局的 PageRank 值, 用以评价网页的重要性。其基本思想是: 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

收稿日期: 2005-09-30; 修改日期: 2005-11-04

作者简介: 刘金桂(1975-), 女, 黑龙江宁安人, 淮阴工学院讲师, 硕士研究生, 研究方向: 系统集成。

地传递到它所引用的页。PageRank 评价标准认为每个超链接的重要性与包含这个超链接的原 Web 网页的重要性是成比例的,而不是每个链接的重要性都相同。计算 PageRank 的公式为:

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

式中 T_i ($i=1, 2, \dots, n$) 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, C 为规范化因子,是保证所有网页的 PageRank 总和为一常量。通过简单的迭代算法可以计算出 $PR(P)$ 的值。

这个定义存在一个假设前提,即认为所有的网页形成一个牢固的链接图,每个网页都能从其他网页通过超链接到达。

利用 PageRank 的公式定义可以计算网页集合中所有网页的 PageRank 值。假设 S 为所有的网页的集合, $|S|$ 为整个网页的总和,由于所有网页的 PageRank 值开始是未知的,所以在进行计算之前,首先给每个网页的 PR 值都赋予 $1/|S|$,再根据公式定义进行计算,最后对得到的值再次利用公式定义,这样循环反复,直到计算所得的 PageRank 值收敛于一个相对固定的数 L 。Page 等人通过实验,认为循环次数和链接数目是对数增长的,表示为 $\log(n)$, n 为超链接数目。

1.2 算法优点

从 PageRank 的计算公式可以看出,一个页面会因为别的页面对自己的引用而增加自己的 PageRank 值,但并不会因为自己对别的页面进行引用而提高自己的 PageRank 值。一个页面的 PageRank 值会均匀地传给它所引用的页面,一个页面的引用越多,被引用页面所获得的 PageRank 值就越少。PageRank 技术可以有效地避免那些为了提高网站在搜索结果中的排名而故意使用链接的网页。

另外,只要提前算好了 PR 值,检索时不必重新计算,减少了在线时间,能有效地提高检索的效率。

1.3 算法缺点

PageRank 算法的计算具有全局性,要算一个网页的 PR 值就要算出文档集里所有网页的 PR 值,计算量很大;利用 PageRank 算法的计算公式计算出的 PR 值与检索主题无关;PageRank 算法不会因为一些页面引用了别的网页而提高了自己的 PR 值,这使的某些中心页也受到了同等的待遇,中心页面本身不突出,没有多少链接指向它,但它指向了某个话题最为突出的页面链接,可以说一个好

的页面由多个好的中心页面指向,一个好的中心页面指向多个好的页面。这应是一种互动的关系,但在 PageRank 算法中并没有考虑。

另外,PageRank 公式的定义的前提假设是所有网页形成一个牢固的链接图。但是,实际中的网络环境却是在不断地变化,网络超链接环境的动态多变性,使得网络超链接之间存在着两个问题:等级沉没和等级泄露,这就意味着可能对于一些网页不能很好地判断出网页的重要性。

2 网页的相似度

两个网页在内容上是否相似,可以根据向量空间模型(VSM)算法,即文献关键词的权重主要由其在文献中的出现频率决定。还有的研究者利用了超链接相似度函数来计算网页的相似度。超链接相似度函数由 Ron Weiss 等人提出。超链接相似度函数是利用网页之间的链接信息来计算网页相似度的方法。它从 3 个方面利用网页信息:两个网页间最短路径的长度;两个网页的共同祖先的数目;两个网页的共同子孙的数目。很多文献关于网页相似度的计算都进行了研究,本文在此不再赘述。

3 改进的 PageRank 算法

近些年,由于 PageRank 算法存在的一些不足,PageRank 算法发明者和一些学者对 PageRank 算法进行了改进,主要有以下几种:

修正 1:

$$PR(P) = C \sum_{i=1}^n \frac{PR(T_i)}{C(T_i)} + C \cdot E(P) \quad (2)$$

其中 C 为规范化因子, $E(P = \alpha/|S|$ ($0 < \alpha < 1$), $|S|$ 为整个网页的总和, T_i ($i=1, 2, \dots, n$) 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。

修正 2:

$$PR(P) = (1-d) + d \sum_{i=1}^n \frac{PR(T_i)}{C(T_i)} \quad (3)$$

其中 d 是 $(0, 1)$ 区间上的衰减系数, T_i ($i=1, 2, \dots, n$) 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。

修正 3:

$$PR(P) = (1-d)/m + d \sum_{i=1}^n \frac{PR(T_i)}{C(T_i)} \quad (4)$$

其中 m 表示网页节点的总数,作用在于 $1-d$ 的 PageRank 在所有网页中分配, d 是 $(0, 1)$ 区间上

的衰减系数, $T_i (i = 1, 2, \dots, n)$ 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。

对于 PageRank 算法的计算公式及其一些修正的计算公式, 可以看到网页 P 的 PR 值与其链入的超链接的个数以及链入的超链的 PR 值有关, 也就是网页 P 的重要性只与其链入的超链接的个数和链入的超链的 PR 值有关, 而与链入的网页的内容无关。实际上, 虽然链入的网页对网页 P 投了一票, 但是它的内容与网页 P 无关, 那么链入的网页对网页 P 所投的这一票的份量也就应该大大降低, 相应的网页 P 的 PR 值也就低, 然而按照 PageRank 算法的计算公式, 只要有链入的网页, 它的 PR 值就应该相应地增加。实际上, 一个网页的重要性不仅与链入的超链的个数以及链入的超链的 PR 值有关, 而且还应该与链入的超链的内容有关, 因此本文对 PageRank 算法的计算公式修正为:

$$PR(P) = (1 - d) + d \sum_{i=1}^n (\omega_{pi} \frac{PR(T_i)}{C(T_i)}) \quad (5)$$

其中为网页 P 与网页 T_i 的相似度, d 是 $(0, 1)$ 区间上的衰减系数, $T_i (i = 1, 2, \dots, n)$ 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。通过修正的 PageRank 计算公式我们可以看出, 当网页 P 与网页 T_i 的相似度大时, 则网页 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, 则网页 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。修正后的 PageRank 计算公式的 PR 值不仅与链入的超链的个数以及链入的超链的 PR 值有关, 而且还与链入的网页的内容有关。算法如下:

$$\forall P \in S, T_i \in B(T), \omega_{pi} = \text{sim}(P, T_i),$$

/* 为所有网页的集合, 为网页与的相似度 */

/* 为所有指向的网页的集合 */

$\forall P \in S: PR(P)_0 = 1;$ /* 赋予所有网页的初始值为 1 */

{

repeat for each $P \in S$

$$PR(p)_i = (1 - d) + d \sum_{k=1}^n \frac{\omega_{pk} PR(T_k)_{i-1}}{C(T_k)}$$

/* n 为所有指向网页 P 的链接数目, d 为衰减因子 */

until $(|PR(P)_i - PR(P)_{i-1}| < \epsilon)$

/* ϵ 为事先给定的充分小的数 */

}

为了更进一步地理解改进后的 PageRank 算法, 本文根据改进后的 PageRank 算法计算公式(5)对如下网页进行了实验, 同时分析利用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 4 个包含某一相同主题的网页, 它们相互之间有链接, 其结构如图 1 所示。

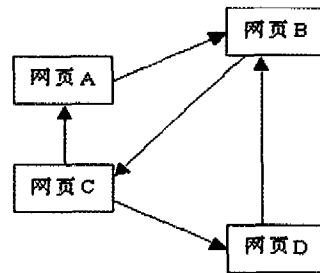


图1 网页链接结构示意图

因为网页相似度可以在获取网页的时候计算得到, 所以可以预先计算出网页相似度的值。设已知上述网页相似度的值为 $\omega_{AB} = 0.6, \omega_{AC} = 0.3, \omega_{BC} = 0.2, \omega_{BD} = 0.3, \omega_{CD} = 0.1$, 其中 $\omega_{AB}, \omega_{AC}, \omega_{BC}, \omega_{BD}, \omega_{CD}$ 分别为网页 A 与网页 B, 网页 A 与网页 C, 网页 B 与网页 C, 网页 B 与网页 D, 网页 C 与网页 D 的相似度的值, 并且为网页 A, B, C, D 赋予初始值 1, 取为 0.01 (事先给定的一个充分小的数, 用于控制迭代次数), 衰减因子 d 取值为 0.85。利用改进后的 PageRank 算法的计算公式(5)进行实验, 通过迭代计算得到实验结果为 $PR(A) = 0.1754, PR(B) = 0.2805, PR(C) = 0.1990, PR(D) = 0.1585$, 因此按照网页的 PR 值的大小, 结果排序为 B, C, A, D。此排序结果符合 PageRank 的思想, 即网页 B 被多次引用, 则网页 B 很可能是重要的; 网页 C 尽管没有被多次引用, 但被重要网页 B 引用, 则网页 B 的重要性被均匀地传递到它所引用的网页 C, 同时, 根据改进后的 PageRank 算法计算公式, 即使网页 A 和网页 D 具有相同的链入的网页 C, 但是由于网页 A 和网页 C 的相似度的值比网页 D 和网页 C 的相似度值大, 所以网页 A 比网页 D 重要。另一方面, 如果按照没有考虑网页相似度的 PageRank 公式计算 PR 值, 得到的实验结果是 $PR(B) > PR(C) > PR(A) = PR(D)$, 也就是在计算的过程中, 网页 C 对网页 A 和 D 的贡献一样大, 但是我们通过网页相似度的值知道, 网页 A 和 C 之间的相似度比网页 D 和 C 之间的相似度大, 所以网页 C 对网页 A 的贡献应该比对网页 D 的贡献大, 而不应该是网页 C 对网页 A 和网页 D

的贡献相等。通过对改进前与改进后的 PageRank 算法计算公式所得到的实验结果进行分析,可以发现改进后的 PageRank 算法比改进前的算法的排序结果更准确。该实验的结果表明:利用改进后的 PageRank 算法计算公式(5)对检索结果进行排序,可以有效地提高检索结果的查准率。

此算法的优点:由于用此算法计算出来的 PR 值不仅与网页的链入的超链的个数以及链入的超链的 PR 值有关,而且还与链入的网页的内容有关,从而可以提高网页的重要性的准确度;网页的 PR 值与网页的内容有关,能使检索得到的排序结果的查准率更高。缺点:由于该算法引入了加权系数,也就是在计算网页的 PR 值之前,要计算出文档集内所有网页相似度的值,从而增加了计算量。

4 结束语

随着 Internet 上的信息的爆炸式增长,用户对检索结果的查准率的日益苛求,搜索引擎系统变得越来越复杂。PageRank 算法的提出带来了搜索引擎革命性变革,使人们将目光从注重信息量收集的传统搜索引擎转向注重搜索结果准确性的新一代搜索引擎身上。但是 PageRank 算法只注重页面之间的超链接,而忽略了链入的页面的内容,所以该算法还存在着一定的不足。本文对 PageRank 算法进行了改进,在计算公式中引入了加权系数,即网页相似度,将网页相似度引入了 PageRank 算法的计算公式,通过对实验的结果进行分析,该算法可以提高检索结果的查准率。搜索引擎的系统

是很复杂的,它所处理的网络信息是动态的,异构的,分布的,因此给搜索引擎的技术的发展提出了更严峻的挑战。而研究 PageRank 算法是为了在此基础上提出更有创意的排序算法,同时,更多新技术的应用也将使搜索引擎技术不断向人性化、智能化、多样化等方向发展。

参考文献:

- [1] DeBra P, Post R. Information retrieval in the World Wide Web: making client - based searching feasible [A]. Proc1s International World Wide Web Conference[C]. Geneva: CERN, 1994, 45 - 55.
- [2] Hersovici M, Jacovi M, Maarek Y, et al. The shark - search algorithm - an application; tailored Web site mapping [J]. Computer Networks and ISDN Sys - tem, 1998, 30: 256 - 264.
- [3] Brin S, Page L. The anatomy of a large - scale hyper - textual Web - search engine [A]. Proc 7th International World Wide Web Conference [C]. Brisbane: SIGIR, 1998. 146 - 164.
- [4] 万华,牛军钰,吴立德. 链接信息在 Web 检索中的应用[J]. 计算机工程, 2003, 28(9): 60 - 62.
- [5] 孙建军,成颖等. 信息检索技术[M]. 北京:科学出版社,2004.
- [6] 苏新宁,杨建林,邓三鸿,周军. 数据挖掘理论与技术[M]. 北京:科学文献技术出版社,2003.
- [7] 网页结构设计和大幅增加链接数量的探讨. <http://www. disksoft. net/news006/design/2004727143003. htm>.

(责任编辑:吴廷东)