应用 Web 结构挖掘的 PageRank 算法的改进研究

范聪贤,刘秋菊,徐汀荣
FAN Cong-xian,LIU Qiu-ju,XU Ting-rong

苏州大学 计算机科学与技术学院,江苏 苏州 215006
School of Computer Science & Technology,Soochow University,Suzhou,Jiangsu 215006,China
E-mail:cxfan@163.com

FAN Cong-xian,LIU Qiu-ju,XU Ting-rong.Research and improved algorithm of PageRank based on Web structure mining.Computer Engineering and Applications,2010,46(9):127-129.

Abstract: With the development of Internet's technology,Web pages become effective approach for people to gain information. Web data mining gradually becomes hot research.This paper puts forward a kind of improved algorithm which is based on deficiency of PageRank in Web structure mining.The result of experiment proves that improved algorithm is more effective than old algorithm,which has much practical values.

Key words: Web data mining;Web structure mining;PageRank;Google

摘 要:随着 Internet 技术的发展,Web 网页成为人们获取信息的有效途径,Web 数据挖掘逐渐成为研究的热点。基于 Web 结构挖掘的 PageRank 算法存在不足的情况下,提出了一种改进的算法,实验结果证明改进的算法较原算法具有较好的效果,具有一定的实用价值。

关键词:Web 数据挖掘;Web 结构挖掘;PageRank;Google
DOI:10.3778/j.issn.1002-8331.2010.09.036 **文章编号**:1002-8331(2010)09-0127-03 **文献标识码**:A **中图分类号**:TP391

1 引言

随着 Web 信息技术的迅速发展,用户可以越来越方便快捷地获得各种信息,但同时也面临着如何从大量 Web 信息中得到相关和有用的信息问题。虽然通过使用 Google、百度、Lycos 等搜索引擎,可以大大减少无用信息的干扰,但是这些搜索引擎搜索的结果有时也不完整或不相关,很难完全满足用户的需求。而 Web 数据挖掘技术则可以解决过量信息的问题,为用户提供更精确、更相关的数据。Web 数据挖掘逐渐成为目前研究的热点。

从 Web 结构挖掘入手,针对 Web 结构挖掘中的典型算法 PageRank 忽略了 Web 页面上的文本和其他内容,只考虑页面间的超链接,容易出现主题漂移现象。基于上述情况,提出了一种改进的 PageRank 算法,改进的 PageRank 算法可以有效解决上面问题。

2 Web 数据挖掘的起源与分类

2.1 Web 数据挖掘的起源

Web 数据挖掘起源于数据挖掘,数据挖掘是指从大型数据库的数据中提取人们感兴趣的知识,而这些知识是隐含的、事先未知的、潜在的有用信息。WWW 以超文本的形式给用户提供了包含从技术资料、商业信息到新闻报道、娱乐信息等多种类别和形式的信息,可以说 Web 是当今世界上最大的电子信息仓库,蕴含着巨大潜在价值的知识。但是由于 Internet 自身的开放性、动态性和异构性等特点,导致信息、知识获取困难,在这种情况下,运用现有的数据挖掘技术对分布的、异质的 Web 信息资源进行挖掘,就成为了数据挖掘技术的挑战和未来的发展方向,由此产生了 Web 数据挖掘。

2.2 Web 数据挖掘的分类

Web 数据挖掘是一项综合技术,涉及 Web、数据挖掘、计算机语言学、信息学等多个领域。Web 数据挖掘是指从大量 Web 文档的集合 C 中发现隐含的模式 P。如果将 C 看作输入,将 P 看作输出,那么 Web 挖掘的过程就是从输入到输出的一个映射:$C\rightarrow P$。

根据对 Web 数据的感兴趣程度不同,Web 数据挖掘一般可以分为三类[1]:Web 内容挖掘(Web Content Mining)、Web 结构挖掘(Web Structure Mining)、Web 使用挖掘(Web Usage Mining)。

(1)Web 内容挖掘。Web 内容挖掘是指对站点的页面内容进行挖掘。Web 内容挖掘的对象包括文本、图像、音频、视频、多媒体和其他各种类型的数据。

(2)Web 结构挖掘。Web 结构挖掘是对 Web 页面之间的链接结构进行挖掘。在整个 Web 空间里,有用的知识不仅包含在 Web 页面内容之中,而且也包含在页面的链接结构之中。

(3)Web 使用记录挖掘。Web 访问信息挖掘是对用户访问 Web 时在服务器方留下的访问记录进行挖掘,即对用户访问 Web 站点的存取方式进行挖掘。

作者简介:范聪贤(1979-),男,硕士研究生,主要研究领域为数据库,Web 数据挖掘;刘秋菊(1981-),女,硕士研究生,主要研究领域为图像处理模式识别;徐汀荣(1958-),男,教授,主要研究领域为算法设计,图像处理。

收稿日期:2008-09-25 **修回日期**:2008-12-19

3 PageRank 算法

该算法由 Brin 和 Page 提出[2],是最早利用超链接信息进行 Web 挖掘的算法,也是商业应用中最成功的一种算法,被Google 搜索引擎采用。该算法的基本思想如下:设页面 i 的链入集合为$\{T_1,T_2,\cdots,T_n\}$,即$\{T_1,T_2,\cdots,T_n\}$中的每一个页面都链接到页面 i,$C(i)$为页面 i 的链出页面数,则页面 i 的等级值 $PR(i)$可以通过以下两步计算得出:(1)以概率 e 随机取 Web 上任一页面。(2)以概率 1–e 随机取当前页面任一链出页面。计算公式如下:

$$PR(i)=1-e+e*(PR(\frac{T_1}{C(T_1)})+PR(\frac{T_2}{C(T_2)})+\cdots+PR(\frac{T_n}{C(T_n)}))$$

显然,$PR(i)$值越大该页面权威性越高。该算法与用户查询条件无关,只是给出每一页面的等级值,作为 Google 搜索引擎搜索结果排序的一个参考值,等级越高的页面排序越靠前。另外,从上述公式可见:(1)这个算法不以站点排序,页面的网页级别由一个个独立的页面决定;(2)页面的网页级别由链向它的页面的网页级别决定,但每个链入页面的贡献的值是不同的。如果 T_i 页面中链出越多,它对当前页面 A 的贡献就越小。A 的链入页面越多,其网页级别也越高;(3)阻尼系数的使用,它的值在 0~1 之间,减少了其他页面对当前页面 A 的排序贡献。

4 对 PageRank 算法的改进

传统的 PageRank 算法仅对互联网的超链接拓扑结构进行分析,而不考虑页面的超链接是否和该页主题相不相关,这样容易导致最后推荐的结果出现与查询主题无关但又很高的 PageRank 值的无效网页,这样就会出现主题漂移现象。另一方面,此算法是低精度和高复杂度的,它对比较流行的网页随着时间的推移赋予更高的权威值,但这些网页内容并不一定是用户想找的网页内容,导致出现主题漂移现象,所以需要改进此算法。

根据 PageRank 算法思想,提出了一种新颖的迭代方法,它是基于强化学习[3]而考虑网页之间的距离作为“处罚”因子,以此来计算 Web 网页的等级值进行排序,把它称为“距离等级”算法(DisRank)。距离等级算法是把距离定义为从一个网页到达要找的网页所经过所有中间网页的过程中平均点击网页数最小来展开的。这样做的目的是使“处罚”因子或者距离最小化的网页而具有较高的等级值,这些网页是与查询内容高度相关的。下面对改进的算法思想首先定义几个概念:

定义 1 假如由网页 i 指向网页 j,则网页 i 和网页 j 之间的链接权重值就等于 $\lg o(i)$。其中 $o(i)$表示网页 i 的出度(网页 i 链出的页面数目)。

定义 2 网页 i 和网页 j 的距离是网页 i 与网页 j 之间的最短路径权重值(这里指最小路径值)。把它命名为对数距离,并且用 d_{ij} 表示。

为了能更好地理解这个定义,举例说明,如图 1 所示,根据定义 1,网页 i,网页 t 及网页 r 的链出的权重值分别为 log(2),log(4)和 log(3)。从图 1 可以知道:从网页 i 到达网页 k 的路径有两条,一条是由 $i\to t\to k$,另一条由 $i\to r\to k$,根据定义 2,它们之间的距离分别 log(2)+log(4)及 log(2)+log(3),所以选择第二条路径作为从网页 i 到网页 k 的最短路径,这也是定义 2 中所谓的最短路径。另外从图 1 中还可以看出,从网页 i 到网页 s 的距离为 log(2)+log(4),这样可以得出:同样从网页 i 出发及经过点击两次网页到达网页 j 和网页 k,而网页 k 更接近网页i,从而具有更高的等级值,也就是说网页 i 与网页 k 具有更高的相关性。

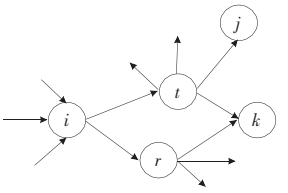

图 1 构造的 Web 有向图

定义 3 假如网页 i 和网页 j 之间的距离 d_{ij} 以定义 2 中的方法定义,那么 d_j 可以描述为网页 j 的平均距离,定义公式如下:

$$d_j=\frac{\sum_{i=1}^{v}d_{ij}}{V} \tag{1}$$

其中 V 表示网页数目。

在此定义中,通常以查找想找的网页而平均点击的网页数来定义距离,从而取代了传统对距离的定义方法。在方法中,每一个网页间链接的权重用 $\log(o(i))$来表示。自然地,假如页面 i 和页面 j 之间没有链接,则 d_{ij} 值被设定为一个较大的值(在这里设为 $\log N$,其中 N 表示互联网总的网页数)。

改进的算法像 PageRank 算法一样在构造的 Web 图中要依赖节点(网页)的出度(网页链出的页面数目)。另外,改进的算法也是使用构造的 Web 图,就像在 PageRank 算法中对于网页 i 出度的选择是根据概率 $1/o(i)$的值来确定一样,从而来构造随机冲浪模型[2]。为了说明这个事实,定义页面 i 对页面 j 的等级值可以以页面 i 到页面 j 的所有路径中对数的最短路径值来表示[4]。假如有任意一位冲浪者从网页 i 到网页 j 经过单独的三步存在一条对数的最短路径,就像 $i\to k\to l\to j$,那么从页面 i 到达页面 j 的可能性可以表示为:$(1/o(i))*(1/o(k))*(1/o(l))$。从另一个角度理解,网上随机冲浪从页面 i 到达页面 j 的概率就是$(1/o(i))*(1/o(k))*(1/o(l))$。

在考虑平均距离方面,像等级值下降的网页[5]及没有出度和入度的网页等问题不会影响改进的算法对距离的计算。为了使平均距离计算具有实际意义,提议一种新的等级值定义方法并把它命名为距离等级。直观地讲,假如网页 j 的出度为 1 及它是从页面 i 链接过来的,在这种情况下计算页面 j 的平均距离 d_j 可以根据上述的 3 个定义,由此得出下面的关系:

$$d_j=\frac{\sum_{k=1}^{v}d_{kj}}{v}=\frac{\sum_{k\neq i}^{v}(d_{ki}+d_{ij})+d_{ij}}{v}=\frac{\sum_{k\neq i}^{v}d_{ki}}{v}+d_{ij}=$$

$$\frac{\sum_{k=1}^{v}d_{ki}-d_{ii}}{v}+d_{ij}\xrightarrow{\text{公式(1)}}d_j=d_i-\frac{d_{ii}}{V}+d_{ij}\approx d_i+d_{ij}=d_i+\log(o(i)) \tag{2}$$

由于互联网是巨大的,所以 v 趋向无穷大,那么上面公式 d_{ii}/v 这一项可以忽略不计。

在通常情况下,假定 $o(i)$表示网页 i 的出度(网页 i 链出的页面数目),$B(j)$表示指向网页 j 的页面集合,那么网页 j 的距离等级值 d_j 可以用下式表示:

$$d_j=\min_i(d_i+\log o(i)),i\in B(j) \tag{3}$$

借助公式(3),假设下面的公式是基于 Q-学习[6]的,Q-学习是一种强化学习算法[3],用它来计算网页 j 的距离等级值(网页 i 链向网页 j)。公式(4)如下:

$$d_{j_{i+1}}=(1-\alpha)*d_{j_i}+\alpha*\min_i(\log(o(i))+\gamma*d_{i_i}),i\in B(j),$$

$$0<\alpha\leqslant 1,0\leqslant\gamma\leqslant 1 \tag{4}$$

其中 α 表示学习的比率;γ 表示折扣因子,它通常用于在计算网页 j 的距离时控制其他链向 j 的网页对网页 j 距离的影响(例

如，假如有一条 $i \to t \to j$ 的路径，那么网页 i 对网页 j 的影响程度可以用折扣因子表示为 γ^3）；$\log(o(i))$表示从网页 i 转向网页 j 而获得的权重，即是“处罚”因子。d_{j_i}和 d_{i_t}分别表示在时间 t 内网页 j 和网页 i 的距离，而 $d_{j_{t+1}}$表示在时间 t+1 内网页 j 的距离。

公式(4)中学习比率 α 的值是根据下面公式(5)来计算的。在实验中发现假如学习比率 α 值作适当的调整，那么改进的算法在求距离等级值时将很快地收敛及达到一种稳定的状态，并具有很高的吞吐率。在算法的初始状态下，网页间的距离是未知的，所以最初把 α 设置为 1，然后指数级减少为 0。公式如下所示：

$$\alpha = e^{-\beta * t} \tag{5}$$

其中 t 表示时间；β 表示一种静态值，由它来控制学习比率的规则。

在上述公式(4)中对 DisPank 算法的 d_j 值的迭代计算与 PageRank 算法中 PR 值的迭代计算很相似。此过程反复迭代直至收敛。收敛后将得到网页 d_j 的矢量 $\boldsymbol{D}_n[j]$。网页将呈现升序排列，对于那些 d_j 值比较低的网页将具有较高的等级值。DisRank 算法将使用激励方法[7]来计算距离 d_j，具体算法描述如下：

输入：D_0

输出：$D_n[j]$

步骤 1　itr=0；//itr 表示迭代次数

步骤 2　While $\delta > \varepsilon$　//ε表示错误数

$\alpha = e^{-\beta * t}$；

$itr = itr + 1$；

for every page $j \in url\_stack$　//url_stack 是爬行算法抓取的 $D_n[j]$=$(1-\alpha) * D_{n-1}[j] + \alpha * \min_i(\log(o[i]) + \gamma * D_{n-1}[i]) i \in B(j)$，$0 < \alpha \leqslant 1$，$0 \leqslant \gamma \leqslant 1$；

$\delta = \| D_{n-1} - D_n \|$

end While

步骤 3　return $D_n[j]$。

另外，为了对改进的算法进行评估，首先要抓取一定数量的网页。具体的爬行算法如下：

输入：st_url，K=200 000，$push(url\_stack, st\_ur)$

输出：url_stack

步骤 1　while(!$empty(url\_stack)$)

　$url = pop(url\_stack)$；

　$crawl\_page(url)$；

　for each child u of url

　　if ($u \notin url\_stack$ and $u \notin crawled\_pages$)

　　$push(url\_stack)$；

　if($carwled\_pages.count()\%K == 0$)

　$reorder\_stack(url\_stack)$；

　end while；

步骤 2　return url_stack。

5　实验结果与分析

5.1　实验结果

以上述爬行算法对 http://www.sohu.com 网站进行网页的抓取，从而抓取 20 万张有效的网页，分别利用传统的 PageRank 算法和改进的 DisRank 算法进行网页等级值的计算。并模拟查询 HIV、Olympic Games、Earthquake、Stamp、Education、Economy、University、Environment、Military Affairs、Stock 10 个不同的主题，每次获取结果集的前 100 项。同时，为取得标准的结果集，利用 Google(http://www.google.com)搜索引擎的高级搜索功能，单独在 http://www.sohu.com 网站查询刚才的 10 个主题，取每次查询结果的前 100 项为标准结果集。为了对改进的算法从整体查看效果，引用文献[8]的一个度量来衡量算法的返回结果，此文中提出查全率，就是返回的相关网页与全部相关网页的百分比。如表 1，其中系列 1 是改进算法获取结果集的查全率，系列 2 是传统算法获得的结果的查全率。改进算法(DisRank)及传统算法(PageRank)实验数据如表 1 所示。

表 1　两种算法的查全率数据

	1	2	3	4	5	6	7	8	9	10
DisRank	0.89	0.93	0.96	0.85	0.79	0.86	0.98	0.87	0.86	0.85
PageRank	0.60	0.58	0.65	0.72	0.66	0.73	0.75	0.50	0.54	0.61

5.2　实验分析

为了更好地对改进的算法进行分析。把上述实验数据用 Excel 绘成柱形图来显示，如图 2 所示。

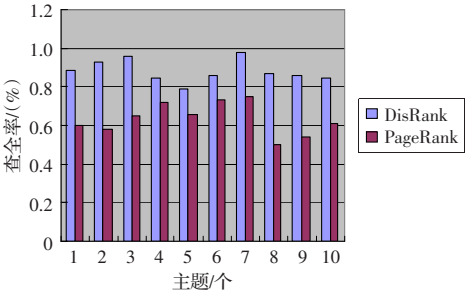

图 2　两种算法的比较图

从图 2 中可以清晰地看出，改进算法较传统算法表现出较好的查全率，有效地遏制了主题漂移现象。

6　小结

首先对 Web 数据挖掘的来源及分类进行了概述，接着对 Web 结构挖掘的 PageRank 算法进行了详细的描述，然后基于 PageRank 算法存在不足的情况下提出一种改进的算法，最后给出了实验结果。创新点在于对改进的算法以计算网页之间距离的大小来确定网页的等级高低，这样以来可以降低算法的复杂度；另外在进行实验时采用真实的网络体系结构，而不是像传统算法采用冲浪模型。下一步的研究重点是寻找更有效的 Web 结构挖掘算法，进一步提高 Web 挖掘的效率。

参考文献：

[1] 毛国君，王实.数据挖掘原理与算法[M].2 版.北京：清华大学出版社，2007.

[2] Brin S，Page L.The anatomy of a large-scale hypertextual web search engine[C]//Proceedings of 7th World Wide Web Conference，1998：107-117.

[3] Sutton R S，Barto A G.Reinforcement learning：An Introduction[M].Cambridge，MA：MIT Press，1998.

[4] Baeza-Yates D，Boldi P，Castillo C.Generalizing PageRank：Damping functions for link-based ranking algorithms[C]//SIGIR，2006：308-315.

[5] Arasu A，Cho J，Garcia-Molina H，et al.Searching the web[J].ACM Transactions on Internet Technology，2001，1：2-43.

[6] Watkins C J C H.Learning from delayed reward[D].University of Cambridge，England，1989.

[7] Watkins D S.Fundamentals of matrix computations[M].[S.l.]：John Wiley & Sons，2002.

[8] Dean J，Henzinger M R.Finding related pages in the World Wide Web[J].Computer Networks，1999，31(11/16)：1467-1469.