

文章编号: 1002-4026(2006)04-0011-04

# 基于 PageRank 和 HITS 的 Web 结构挖掘算法研究

刘 栋<sup>1</sup>, 刘希玉<sup>1</sup>, 郝婷婷<sup>2</sup>

(1. 山东师范大学信息科学与工程学院, 山东 济南 250014; 2. 山东大学材料科学与工程学院, 山东 济南 250061)

**摘要:**在对 Web 结构挖掘的典型算法探讨的基础上, 提出了一种 PageRank 算法和 HITS 算法相结合的改进算法, 并对该算法进行了简要分析。

**关键词:**数据挖掘; Web 结构挖掘; PageRank; HITS

**中图分类号:** TP301.6

**文献标识码:** A

随着 Internet 的快速发展, Web 正在成为一种新的数据源, 其中汇集了大量信息。但是 Web 具有无结构、动态、组织复杂的特点, 给用户搜索数据造成了很大困难。这就急需一种能自动地从 Web 资源中发现、获取信息的新技术, Web 数据挖掘技术应运而生, 并取得了一定的研究成果。本文分析了 PageRank 和 HITS 算法, 并在此基础上介绍了一种 Web 结构挖掘的改进算法。

## 1 Web 数据挖掘分类

一般来讲, Web 数据挖掘可以分为 3 类: Web 内容挖掘 (Web Content Mining), Web 结构挖掘 (Web Structure Mining) 和 Web 使用记录的挖掘 (Web Usage Mining)<sup>[1]</sup>。如图 1 所示。

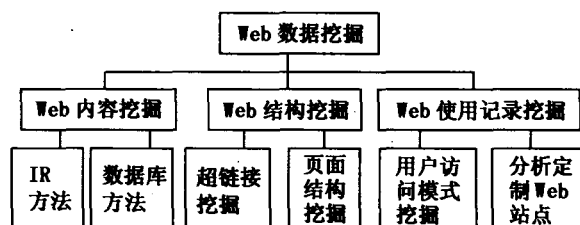


图 1 Web 数据挖掘分类

### 1.1 Web 内容挖掘

Web 内容挖掘是从大量的 Web 数据中发现并抽取有用信息的过程。这些数据既有文本和超文本数据, 也有图形、图像、语音等多媒体数据; 既有来自于数据库的结构化数据, 也有用 HTML 标记的半结构化数据和无结构的自由文本。其中, Web 内容挖掘可以分为: IR (Information retrieve) 方法和数据库方法。<sup>[2]</sup>

### 1.2 Web 使用记录挖掘

Web 使用记录挖掘即通过挖掘用户的 Web 日志记录, 发现用户访问 Web 页面的模式, 得到有价值的信息。这些数据包括: 各类服务器日志记录、浏览器日志记录、用户注册信息、用户对话或交易信息等等。目前这一方面的研究较多, 并且出现了很多种有商业价值的 Web 日志挖掘技术和工具。

收稿日期: 2006-03-01

基金项目: “泰山学者”建设工程专项经费资助; 山东省自然科学基金重大项目 (Z2004G02), 山东省教育厅计划项目 (J05G01)。

作者简介: 刘栋 (1983-), 男, 硕士研究生, 主要研究方向: 数据挖掘, 支持向量机。E-mail: ld0201@163.com

### 1.3 Web 结构挖掘

Web 结构挖掘是从 WWW 链接结构关系网络中推导知识<sup>[3]</sup>。它主要是针对 Web 页面之间的超链接结构、内部结构和 URL 中的目录路径结构进行挖掘,从中抽取知识,包括文档结构挖掘和站点结构挖掘。目前,Web 结构挖掘主要是基于超链接结构的挖掘。通过对超链接结构的研究分析,以提高搜索引擎的效率。

## 2 Web 结构挖掘典型算法

目前,Web 用户主要是使用搜索引擎在 Internet 上检索信息,但目前的搜索引擎的效率不尽如人意,往往会返回很大一部分重复的或者是与用户检索要求不相关的页面。由此,如何利用 Web 独有的结构特点,提高搜索引擎的检索效率成为当前的一个研究热点问题。

目前基于超链接结构分析的 Web 结构挖掘算法主要是将 Web 看作为有向图或无向图的形式,结合一定的启发式方法,用图论的方法进行分析研究。

### 2.1 PageRank 算法

PageRank 算法由 Stanford 大学的 Brin 和 Page 提出,是评价网页权威性的一种重要工具。著名的搜索引擎 Google 就是基于该算法实现的。

为了描述 PageRank 算法,我们作如下定义:

**定义 1**  $G(V, E)$  表示 Web 页面间的超链接结构有向图,  $V$  为 Web 页面的集合,  $E$  为页面间链接的集合。其中,有向边  $(p, q) \in E$  代表从页面  $p$  指向页面  $q$  的超链接。

**定义 2** 页面  $p$  的出度是指从页面  $p$  出发的超链接 ( $\text{Line\_out}(p)$ ) 的总数,即为  $|\text{Line\_out}(p)|$ ; 其入度是指所有指向页面  $p$  的超链接 ( $\text{Line\_in}(p)$ ) 的总数,记为  $|\text{Line\_in}(p)|$ 。

**定义 3**  $|V|$  定义为 Web 有向图中 Web 页面结点的总数,其中,  $R_i$  定义为页面  $i$  指向的所有页面的集合,即为  $\text{Line\_out}(i)$ ;  $B_i$  定义为指向页面  $i$  的所有页面的集合,即为  $\text{Line\_in}(i)$ 。特殊地,对每个出度为 0 的结点  $s$ ,记  $R_s = \{\text{有向图中全部 } N \text{ 个结点}\}$ ,相应地所有其他节点的  $B_i = \{B_i \cup s\}$ 。

由前面的定义,页面  $i$  的等级 PageRank 值可  $PR(i)$  可以通过以下两步计算得出:

- (1) 以概率  $(1-m)$  随机取 Web 上任一页面  $i$ ;
- (2) 以概率  $m$  随机取指向当前页面  $i$  的页面  $j$ ,

则 PageRank 算法的具体迭代公式为:

$$PR(i) = 1 - m + m \sum_{j \in B_i} (PR(j) / |R_j|)$$

其中,参数  $m$  是取值范围在 0 到 1 之间的衰减因子,通常被置为 0.85。

PageRank 算法可以通过递归检索计算索引数据库中的超链接记录来实现,并在 Web 超链接结构分析中取得了成功,被著名的搜索引擎 Google 所采用。但是它也存在一定的缺陷,比如 PageRank 算法检索主题的无关性。目前,很多文献中提出了对 PageRank 算法的改进。如文献[4]给出了 PageRank 值的另一种加速算法——BlockRank Algorithm; 文献[5]讨论了具有时间反馈的 PageRank 算法。

### 2.2 HITS 算法<sup>[1]</sup>

1999 年 Kleinberg<sup>[6]</sup> 提出了 HITS(Hyperlink-Induced Topic Search) 算法来评定网页内容的重要性。他认为网页的重要性应该依赖于用户提出的检索主题。而且对每一个网页应该将其 Authority 权重(由网页的 outlink 决定)和 Hub 权重(由网页的 inlink 决定)分开来考虑。根据页面之间的超链接结构,将页面分为 Hub 页和 Authority 页,其中,Hub 页是一个指向权威页的超链接集合的 Web 页,而 Authority 页是被许多 Hub 页指向的权威的 Web 页。

因此,一个 Hub 页应该指向许多好的权威页,而被许多 Hub 页指向的一定是权威页。这种 Hub 与权威页面之间的相互作用,可用于权威页面的挖掘和高质量 Web 结构和资源的自动发现。具体算法描述如下:

首先, HITS 由用户的检索主题得到一初始结果集, 构成算法的根集 (root set)。比如由基于索引的搜索引擎进行查询。一般地, 根集页面取 200 个左右即可。

其次, 将根集进一步扩展为基本集 (base set), 它包含了所有由根集中的页面所指向的页, 以及所有指向根集页面的页。可以为基本集设定一个上线, 指明扩展的尺度。

最后, 按照公式 (1)、(2) 递归地计算基本集中每个页面的 Authority 权重  $a_p$  和 hub 权重  $h_p$ 。其中,  $a_p$  和  $h_p$  值初始为同一个常数。根据线性代数理论, 可以证明  $a_p$  和  $h_p$  与权重的初始设置无关。

$$a_p^{(i+1)} = \sum_{\forall q: q \rightarrow p} h_q^{(i)} \quad (1)$$

$$h_p^{(i+1)} = \sum_{\forall p: p \rightarrow q} a_q^{(i+1)} \quad (2)$$

其中,  $p \rightarrow q$  是指由页面  $p$  指向  $q$  的超链接。

### 3 基于 PageRank 和 HITS 的改进算法

传统的 PageRank 算法完全忽略掉了网页的内容, 而 HITS 算法存在“主题漂移”的现象。如用户在查询“量子物理学”时, 由于算法中需要对初次检索结果的根集扩充成基集, 最终的检索结果中会包含大量的有关“物理学”的站点<sup>[7]</sup>。基于上述算法, 本文提出了一种基于 PageRank 和 HITS 的改进算法, 可以较好地解决上述不足。

#### 3.1 算法描述

(1) 构照算法的基本集, 这一步和 HITS 中方法类似。根据用户查询请求, 首先用一个现有的商业搜索引擎进行查询, 取其部分查询结果 (约 500 个左右) 作为算法的根集, 记为  $R_0$ 。然后将  $R_0$  进行扩充, 对  $R_0$  中每一个节点, 将所有指向该节点或该节点所指向的网页补充进来, 形成基本集, 记为  $S_0$ 。

(2) 求  $S_0$  中页面的 PageRank 值, 作为搜索引擎结果排序的一个参考。

设  $W$  为  $S_0$  中页面的集合,  $N = |W|$ ,  $R_i$  是页面  $i$  指向的所有页面的集合,  $B_i$  是指向页面  $i$  的所有页面的集合。对每个出度为 0 或者出度页面都不在  $S_0$  中的页面  $s$ , 设  $R_s = \{S_0 \text{ 中所有页面的集合}\}$ , 则所有其他结点的  $B_i = \{B_i \cup s\}$ , 这样可以将结点  $s$  所具有的 PageRank 值均匀地传递给其他所有页面  $i$ 。由此页面  $i$  的等级 PageRank 值  $PR(i)$  可以通过以下两步计算得出:

(1) 以概率  $(1-m)$  随机取基集  $S_0$  中任一页面  $i$ ;

(2) 以概率  $m$  随机取指向当前页面  $i$  的页面  $j$ , 如果  $j \notin S_0$ , 则重新选择页面  $j$ 。

则 PageRank 算法的具体迭代公式为:

$$PR(i) = 1-m + m \sum_{j \in B_i} (PR(j) / |R_j|)$$

其中, 参数  $m$  是取值范围在 0 到 1 之间的衰减因子, 通常被置为 0.85。

#### 3.2 算法分析

PageRank 是对 WWW 的整体分析, 是独立于用户查询的, 可以对用户要求产生快速的响应。而 HITS 算法是对 WWW 的局部分析, 依赖于用户查询的, 实时性差。改进算法通过 HITS 算法得到与用户查询请求相关的 Web 页面集 (即基本集), 进而求得每一页面的  $PR(i)$  值作为搜索引擎结果排序的参考。

由于改进算法采用 HITS 构造结果的基本集, 因此它弥补了 PageRank 算法中页面内容无关性的缺点。与 HITS 算法不同的是, 改进算法采用了 PageRank 的排序机制, 通过计算基本集的 PageRank 值作为排序参考, 在一定程度上削弱了 HITS 算法中的“主题漂移”的缺点。尽管算法中搜索排序网页比 PageRank 要少的多, 但是由于需要在客户端处理一些数据, 所以响应时间没有明显提高。

### 4 结束语

Web 挖掘作为数据挖掘的一个新的主题, 是一个新兴的研究领域, 至今还未形成成熟的理论和技术。基

于链接分析的网页排序算法,目前的研究都还很不成熟,无论是 PageRank 算法,还是 HITS 算法,已有众多的国内外学者在算法的改进方面做出了努力,并且提出了一些其它的算法。对于这两种算法相结合的可能性,目前已有学者作了理论上的探讨。本文提出了两者相结合的一种改进算法,并进行了简要分析。

### 参考文献:

- [1] HAN J W, KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 等译. 北京:机械工业出版社, 2001.
- [2] KOSALA R, BLOCKEEL H. Web Mining Research: a Survey[J]. ACM SIGKDD Explorations, 2000, 2(1): 1 - 15.
- [3] 吉根林, 孙志挥. Web 挖掘技术研究[J]. 计算机工程, 2002, 28(10): 16 - 17.
- [4] SEPANDAR D, KANMVAR, TAHER H, HAVELIWALA, CHRISTOPHER D, MANNING, *et al*. Exploiting the Block Structure of the Web for Computing Pagerank[R]. Stanford: Stanford University, 2003.
- [5] 戚华春, 黄德才, 郑月峰. 具有时间反馈的 PageRank 改进算法[J]. 浙江工业大学学报, 2005, 33(3): 272 - 275.
- [6] KLEINBERG J M. Authoritative Source in s hyperlinked Environment[J]. Journal of ACM, 1999, 46(5): 604 - 632.
- [7] 杨炳儒, 李岩, 陈新中, 等. Web 结构挖掘[J]. 计算机工程, 2003, 29(20): 28 - 30.
- [8] 王艳华, 张纪. Web 结构挖掘及其算法[J]. 计算机工程, 2005, 31(增刊): 125 - 127.
- [9] 何晓阳, 吴强, 吴治蓉. HITS 算法与 PageRank 算法比较分析[J]. 情报杂志, 2004, (2): 85-86.

## Study on web structure mining algorithms based on PageRank and HITS

LIU Dong<sup>1</sup>, LIU Xi-yu<sup>1</sup>, HAO Ting-ting<sup>2</sup>

(1. School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China;

2. School of Material Science and Engineering, Shandong University, Jinan 250061, China)

**Abstract:** Based on the typical algorithms, an improved web structure mining algorithm, combining PageRank algorithm with HITS algorithm was presented.

**Key words:** data mining; web structure mining; PageRank; HITS