

HITS 算法与 PageRank 算法比较分析

何晓阳 吴强* 吴治蓉

(第三军医大学图书馆) (*重庆师范大学现代信息管理系)

摘要 对 HITS 算法作了比较详细的介绍,并且与 PageRank 算法在设计思想、权重传播模型、数据处理量及适用范围等方面进行了比较分析。

关键词 搜索引擎 HITS PageRank 链接分析

随着因特网的迅猛发展,搜索引擎的应用已经非常普及。然而,人们对搜索引擎的核心技术——算法设计知之并不多。了解搜索引擎的算法设计思想及原理,有助于提高我们的信息检索能力,评价搜索引擎。更为重要的是,我国在信息技术领域内的发展情况与发达国家相比还有相当的差距。只有真正掌握了搜索引擎的核心技术,才可能开发出属于我们自己功能强大的搜索引擎,使我们在当今的信息社会中立于不败之地。

国内目前对搜索引擎排序算法的介绍较少,从已有的文献来看,多集中于对更具影响力的 PageRank 算法的介绍和分析研究,而对全球已有较大影响的 HITS 算法和 SALSA 算法介绍较少。正因为如此,才促使笔者向大家介绍 HITS 算法。

HITS 算法是由康奈尔大学(Cornell University)的 Jon Kleinberg 博士于 1998 年首先提出的,HITS 的英文全称为 Hypertext-Induced Topic Search。目前,它为 IBM 公司阿尔马登研究中心(IBM Almaden Research Center)的名为“CLEVER”的研究项目中的一部分。

1 原理

Kleinberg 认为搜索始于用户的检索提问,每个页面的重要性也依赖于用户的检索提问,他将用户检索提问分为三种:特指主题检索提问(specific queries,也称窄主题检索提问)、泛指主题检索提问(Broad-topic queries,也称宽主题检索提问)及相似网页检索提问(Similar-page queries)。而 HITS 算法则专注于改善泛指主题检索的结果。Kleinberg 将网页(或网站)分为两类,即 hubs 和 authorities。应该注意的是,每个页面也有两个级别(ranking),即 hubs(中心级别)和 authorities(权威级别),authorities 为具有较高价值的网页,依赖于指向它的页面,而 hubs 为指向较多 authorities 的网页,依赖于它所指向的页面。HITS 算法的目标就是通过一定的计算(迭代计算)方法以得到针对某个检索提问的最具价值的网页,即排名最高的 authority。

2 HITS 算法介绍

为便于理解,Kleinberg 用图来表示链接关系,可以认为超链页面的集合 V 为一个有向图 $G=(V, E)$,图中的节点对应一个网页,有向边 $(p, q) \in E$ 表示网页 p 链接指向网页 q ,节点 p 的出度(out-degree)指节点 p 链出的网页数量,而节点 p 的入度(in-degree)则指的是链接指向节点 p 的网页数量。如果集合 W 是 V 的一个子集,则用 $G[W]$ 来表示由 W 组成的有向图,它

的节点包含在 W 中,边对应于 W 中的所有链接。现在假设给定一个泛指主题检索提问 σ ,需要通过链接分析确定该提问的权威页。最先是确定 HITS 算法作用的 WWW 子集。理想地, Kleinberg 希望得到的集合 S_σ 具有以下特点: S_σ 相对较小; S_σ 中相关网页丰富; S_σ 包含多数最有价值的 authorities 页面。

2.1 针对具体的检索提问,构建关于该提问的 WWW 聚集子图 具体做法如下:

a. 用基于文本的搜索引擎如 AltaVista 或 Hotbot 来得到 σ 的查询结果集,取排名最高的前 t (t 值通常设为 200)位结果集 R_σ (称为 Root Set), Kleinberg 认为 R_σ 满足特点 a 和 b,但远不能满足 c,因此需要扩充 R_σ 。

b. 扩充 R_σ 分为两个方面,一是将所有 R_σ 中页所指向的页面扩充进去,该扩充在数量上没有限制,二是将指向 R_σ 中的每一页面的链接页面取其中任意 d (d 值通常设定为 50,如果 d 不大于 50,则取其所有页面)个页面扩充到原来的 R_σ 中形成 S_σ (称为 Base Set)。通过实验表明,这样的集合 S_σ 能够较好地满足上述三个特点, S_σ 的数量范围一般在 1000 至 5000。

c. 为了排除干扰,提高计算效果, Kleinberg 还将 S_σ 作了进一步的处理,他将链接分为两种情况:一是指有链接关系的两个页面处在不同域名之间,这样的链接称为横向链接;还有一种情况是指有链接关系的两个页面处于同一域名之下,这样的链接称为内在链接。Kleinberg 认为内在链接只具有网站内部的导航功能,它几乎不能传递网页间的 authority,因此需要将这种内在链接从 S_σ 中删去,形成 G_σ 。

2.2 计算 hubs 和 authorities Kleinberg 认为 hubs 和 authorities 是相互增强的关系。一个好的 hub 页指向许多好的 authorities,同时,一个好的 authority 页也有多个好的 hubs 指向它。在许多实例中,它们之间是一种环状的关系,因此需要一定的计算方法来打破这种环状结构。

对于每一个页面 p ,用 $x^{<p>}$ 表示页面 p 的 authority weight (权威权重),用 $y^{<p>}$ 表示页面 p 的 hub weight (中心权重),满足规范化条件: $\sum_{p \in S_\sigma} (x^{<p>})^2 = 1$ 且 $\sum_{p \in S_\sigma} (y^{<p>})^2 = 1$ 。Kleinberg 将网页权重的传递分为两种方式,即 I 操作和 O 操作。I 操作为 hub 到 authority 的传递,表示为: $x^{<p>} \leftarrow \sum_{(q,p) \in E} y^{<q>}$, O 操作为 authority 到 hub 的传递,表示为: $y^{<p>} \leftarrow \sum_{(p,q) \in E} x^{<q>}$,预先设定迭代次数 k ,算法表示如下:

Iterate(G, k)

G : a collection of n linked pages

k : a natural number

Let z denote the vector $(1, 1, 1, \dots, 1) \in R_n$

Set $x_0 = z$.

Set $y_0 = z$.

For $i = 1, 2, \dots, k$

Apply the I operation to (x_{i-1}, y_{i-1}) , obtaining new x -weights x_i

Apply the O operation to (x_{i-1}, y_{i-1}) , obtaining new y -weights y_i

Normalize x_i , obtaining x_i .

Normalize y_i , obtaining y_i .

End

Return (x_k, y_k) .

根据矩阵计算知识容易证明,对于给定一个初始向量 x_0 和 y_0 ,迭代过程收敛,其最终结果 x^* 为 $A^T A$ 的主特征向量, y^* 为 AA^T 的主特征向量。

3 HITS 算法与 PageRank 算法的比较分析

显而易见,两者均是基于链接分析的搜索引擎排序算法,并且在算法中二者均利用了特征向量作为理论基础和收敛性依据。但两种算法的不同点也非常明显,下面就主要谈谈其不同点:

a. 从算法思想上看,虽然均同为链接分析算法,但二者之间还是有一定的区别。HITS 的原理如前所述,其 authority 值只是相对于某个检索主题的权重,因此 HITS 算法也常被称为 query-dependent 算法。而 PageRank 算法独立于检索主题,因此也常被称为 query-independent 算法。PageRank 的发明者 (Page&Brin) 把引文分析思想借鉴到网络文档重要性的计算中来,利用网络自身的超链接结构给所有的网页确定一个重要性的等级数。当然 PageRank 并不是引文分析的完全翻版,根据因特网自身的性质等,它不仅考虑了网页引用数量,还特别考虑了网页本身的重要性。简单地说,重要网页所指向的链接将大大增加被指向网页的重要性。

b. 从权重的传播模型来看,HITS 是首先通过基于文本的搜索引擎来获得最初的处理数据,网页重要性的传播是通过 hub 页向 authority 页传递,而且 Kleinberg 认为, hub 与 authority

之间是相互增强的关系;而 PageRank 基于随机冲浪 (random surfer) 模型,可以认为它将网页的重要性从一个 authority 页传递给另一个 authority 页。

c. 从处理的数据量及用户端等待时间来分析。表面上看,HITS 算法对需排序的网页数量较小,所计算的网页数量一般为 1000 至 5000 个,但由于需要从基于内容分析的搜索引擎中提取根集并扩充基本集,这个过程需要耗费相当的时间,而 PageRank 算法表面上看,处理的数据数量上远远超过了 HITS 算法。据 Google 介绍,目前已收录的中文网页已达 33 亿以上,但由于其计算量在用户查询时已由服务器端独立完成,不需要用户端等待,基于该原因,从用户端等待时间来看,PageRank 算法应该比 HITS 要短。

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。值得关注的是,目前已有学者对这两种算法相结合的可能性作了理论上的探讨。

参考文献

- 1 Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 1999; 46(5)
- 2 Farahat, A., T. LoFaro, J. C. Miller, G. Rae and L. A. Ward. Existence and Uniqueness of Ranking Vectors for Linear Link Analysis Algorithms. <http://www.damtp.cam.ac.uk/user/jem52/hits.pdf>
- 3 R. Lempel, S. Moran. The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC effect. Computer Networks, 2000, 33
- 4 陈定权. Web 信息检索技术最新进展. 现代图书情报技术, 2002; (2)
- 5 Henzinger R. Hyperlink Analysis for the Web. IEEE Internet Computing, 2001: 45-50. <http://computer.org/internet/>
- 6 曹军. Google 的 PageRank 技术剖析. 情报杂志, 2002; (10)
- 7 S. Brin, L. Page. Anatomy of a Large-Scale Hypertextual Web Search Engine. Proc. 7th International World Wide Web Conference, 1998
- 8 <http://www.google.com/>
- 9 Liang Xu, Xueqi Cheng, and Wensi Xi. TREC10 - HITS <http://csgrad.esvt.edu/lixu1/work/CS5604/hits.htm>

(责编: 钩王京)

(上接第 84 页)

$$\begin{bmatrix} 0.3 & 0.7 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 \\ 0.295 & 0.275 & 0.355 & 0.075 \\ 0.295 & 0.475 & 0.23 & 0 \\ 0.31 & 0.47 & 0.22 & 0 \end{bmatrix} = \begin{pmatrix} 0.3175 & 0.4649 \\ 0.2008 & 0.019 \end{pmatrix}$$

根据最大隶属度原则,有 39.49% 的专家认为“良好”。假设“优秀”为 100 分,“良好”为 80 分,“合格”为 60 分,“不合格”为 40

$$\text{分,则得到矩阵 } R' = \begin{bmatrix} 100 \\ 80 \\ 60 \\ 40 \end{bmatrix}$$

该生的总得分为:

$$S = b \cdot R' = (0.3175 \quad 0.4649 \quad 0.2008 \quad 0.019) \cdot \begin{bmatrix} 100 \\ 80 \\ 60 \\ 40 \end{bmatrix} =$$

81.75 ≈ 82

因此,该生的信息能力评估结果是:“良好”。

4 结束语

目前,国内外对信息能力还没有一个统一的定义。本文通过将国内外同类指标体系进行的对比研究,草拟了高校医学本科生信息能力评估系统。通过评估实践,验证了评估结果的真实可靠,具有一定的实用价值。

参考文献

- 1 孙建军等. 面向 21 世纪的大学生信息素质教育. 中国图书馆学报, 2000; (6)
- 2 陈文勇, 杨晓光. 高等院校学生信息素养能力标准研究. 情报科学, 2000; (7)
- 3 黄晓斌. 美国高校的信息素质教育及其启示. 大学图书馆学报, 2001; (4)
- 4 <http://www.ala.org>
- 5 <http://www.cas.usf.edu/lis/il>

(责编: 王京)