

基于相似度量矩阵 HITS 的改进算法

何 明¹, 周 军¹, 纪周鹏¹, 李树友²

(1 辽宁工业大学 电子与信息工程学院, 辽宁 锦州 121001; 2 辽宁工业大学 数理科学系, 辽宁 锦州 121001)

摘 要: HITS 算法是基于超链接结构的搜索引擎算法, 但它对超链接不加以区分, 容易产生“主题漂移”现象. 改进后的 HITS 算法在原算法的基础上, 提出了相似度量的迭代方法. 该方法是将网页超链接结构信息、文本信息、联合引用信息结合成一个相似度量权重矩阵, 通过权重矩阵来归一 HITS 算法每次产生的 Authority、Hub 值. 在查询效率和质量方面, 改进的 HITS 算法更优, 也减少了“主题漂移”现象的发生.

关键词: HITS 算法; 主题漂移; 相似度量; Authority 值; Hub 值

中图分类号: TP391.3

文献标识码: A

文章编号: 1000-7180(2010)07-0194-04

New HITS Algorithm Based on the Similarity Metric

HE Ming¹, ZHOU Jun¹, JI Zhou-peng¹, LI Shu-you²

(1 School of Electronic and Information Engineering, Liaoning University of Technology, Jinzhou 121001, China;

2 Department of Mathematical Science, Liaoning University of Technology, Jinzhou 121001, China)

Abstract: HITS based on the hyperlink structure is a search engine algorithm, but it links without any distinction, so this algorithm can lead to topic-drift. A new HITS algorithm puts forward a similarity metric iterative method is proposed after analyzed the old one. This iterative method incorporates link structure, textual information, and co-citation information into a similarity metric which gives rise to the weight matrix. This weight matrix normalizes the authority value and hub value that produced in every time of the HITS algorithm. The new HITS algorithm is more efficient and better than the old one in regard to the quality and efficiency about search. It also can avoid the problem of topic-drift.

Key words: HITS algorithm; topic-drift; similarity metric; Authority value; Hub value

1 引言

HITS(hypertext induced topic search)算法是一种有效的基于链接结构分析的网页权重提取方法, 它是 1999 年由 Kleinberg 提出的, 用于评价网页的重要性, 并以此作为核心技术解决了 IBM 搜索系统 Clever 的检索结果相关度排序问题^[1]. HITS 算法的主要思想是: 通过对网络超链接的分析, 利用页面被引用次数及其链接数目来决定不同网页的价值^[2]. 这种方法可以获得比较高的查全率和查准率. 然而 HITS 算法也有其明显的不足, 它完全将网页的文本和内容排除在外, 算法中所有超链接被同等对待, 缺乏语义分析, 容易产生“主题漂移”现象, 从而导致

查询结果与用户需求不符. 文中提出了一种基于相似度量的 HITS 改进算法. 该算法是在原来算法基础上, 加入网页内容的相似度量和网页与网页相互引用关系, 使得该算法将结构和内容结合起来. 由于在 HITS 算法中加入了相似度量矩阵, 从而避免了“主题漂移现象”的发生, 也改善了无链接增大 HITS 算法相关权值的情况.

2 相似度量

相似度量是最近发展起来的, 用来将数据对象聚类成为子群的特殊方法, 它的目的是: 使得在同子群当中的对象是相似的, 在不同子群当中的对象是不相似的. 在 Web 文档聚类问题上, 相似度量是定

收稿日期: 2009-04-30; 修回日期: 2009-08-05

基金项目: 国家自然科学基金项目(60674056); 辽宁省教育厅科学研究基金(20060409)

义两个文本相似的新方法.它是将超链接结构、文本、联合引用信息结合在一起,用矩阵的形式表现出来.其中,超链接结构信息是主要因素,文本相似度是用来调整每个超链接的强度,而联合引用信息能够更好地改善超链接服务.

2.1 超链接结构

网页之间的链接信息来源于网络的拓扑结构.假设一个网络无向图用 $G = (V, E)$ 来表示^[3], V (vertices)表示定点, E (edges)表示边,则定义无向图图 1 的邻接矩阵 A_{ij} :

$$A_{ij} = \begin{cases} 1 & (i, j) \in E \text{ 或 } (j, i) \in E \\ 0 & \text{其他} \end{cases} \quad (1)$$

A_{ij} 为对称矩阵,通过链接结构,从 Web 中能够提取出很多这方面的信息.例如网络无向图图 1,则它的邻接矩阵为:

$$A_{ij} = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ & 0 & 1 & 0 & 0 \\ & & 0 & 0 & 1 \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix}$$

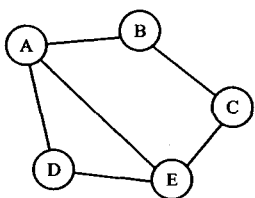


图 1 网络无向图

2.2 文本信息

搜索引擎系统需要对搜集的网页进行分析,提取出网页的文本信息,然后对提取的文本信息进行挖掘.到目前为止,还没有发现一种切实可行的能比较准确地直接从网页中抽取主要的文本信息方法.而基于相似度的文本信息提取方法,近年来受到广泛的青睐.它主要是从用户查询词条在文档中的出现情况来考虑的,包括词条频率、词条位置信息等^[4].此方法的主要思想是:(1)使用的是 Web 文本的整体,而不是通常情况下的用户查询词和锚文本;(2)文本的相似度 S_{ij} 是指两个文本 i, j 之间的相似度,而不是用户查询词和文档之间的相似度;(3) S_{ij} 是两个文档 i, j 之间超链接的强度.也就是说,如果两个文档有很小的文本相似度,即使他们被同一个超链接相连,他们也不可能属于同一个主题群^[5].

用向量空间模型将每一个 Web 文档描述成为一个向量,然后计算它们之间的相似度.每个向量的

元素 k 用如下公式表示^[6]:

$$k = tf(i, j) * idf(i) \quad (2)$$

其中, $tf(i, j)$ 是词 i 在文档 j 当中的词频(Term Frequency),表示词 i 在文档 j 中出现的次数. idf 是逆文档频率,定义如下^[4]:

$$idf(i) = \log\left(\frac{n}{n_i}\right) \quad (3)$$

其中, n 为文档的总数, n_i 为包含词 i 的文档个数.有些词在文档当中出现的次数相当频繁,如:“的”、“地”.不能用它们来识别一个文档,逆文档频率可以有效地减小这些词的影响.

如果两个文档 d_1, d_2 的向量是 x, y ,则 d_1, d_2 的相似度为^[7]:

$$S(d_1, d_2) = S(d_2, d_1) = \frac{\sum_i x_i y_i}{\|x\|_2 \|y\|_2} \quad (4)$$

其中, $\|x\|_2 = \sqrt{\sum_i x_i^2}$. 文档之间的相似度,构成相似矩阵 S .

2.3 联合引用模式

联合引用是另一个衡量两个文档之间相关性的因素.如果有很多的文档都同时指向这两个文档,则这两个文档很可能是阐述同一个话题.文档 i 和 j 的联合引用 c_{ij} 是同时指向文档 i 和文档 j 的文档个数.将它们组成为一个联合引用矩阵 C .

2.4 相似度量矩阵

将上述的链接矩阵 A 、相似矩阵 S 、联合引用矩阵 C 组合成为一个相似度量权重矩阵 W ^[8]:

$$W = \alpha \frac{A \otimes S}{\|A \otimes S\|_2} + (1 - \alpha) \frac{C}{\|C\|_2} \quad (5)$$

3 HITS 算法的改进

HITS 算法将文档视为有向图的定点,文档间的链接关系由边来表示,它通过对网络中超链接的分析,利用页面的被引用次数及其链接数目来决定不同网页的权威性. HITS 算法涉及两个重要的概念:

Authority 值:表示一个权威网页被其它网页所引用的数量,即该权威网页的入度值.若某网页被引用的数量越大,则该网页的入度值越大, Authority 值就越大; **Hub 值:**表示一个 Web 页面指向其它网页的数量,即该 Web 页的出度值,它提供了指向权威页面的链接集合.若某网页的出度值越大,则该网页的 Hub 值越大.

3.1 HITS 算法描述

HITS 算法如下:对于一个给定的查询,得到一

个基于搜索引擎的 Web 集合,称为“根集” R ,然后将 R 所有指向的网页集合以及其它指向 R 的网页集合包含进来形成“基集” S . 基集 S 当中的页面分别是: $1, 2, \dots, n$, 每个页面 i 有一个 Authority 值 a_i 和 Hub 值 h_i , 且将所有页面的 a_i 和 h_i 初始值设为 1, 页面 i 的入链页面集表示为 $B(i)$, 出链页面集表示为 $F(i)$. 则 Hub 权重 h_i 和 a_i 的迭代计算公式为:

$$a_i = \sum_{j \in B(i)} h_j, h_i = \sum_{j \in F(i)} a_j \quad (6)$$

每次计算后,都要对 a_i 和 h_i 做规范化处理. 上面两个公式说明一个页面的权威权重为所有指向它的页面的现有 Hub 权重之和, 一个页面的 Hub 权重为该页面链接的所有页面的现有权重之和. 一个页面如果有许多好的 Hub 页面所指, 则其权威权重会相应地增加; 一个页面若指向许多好的权威页面, 则其中心权重也会相应的增加.

对页面的 Authority 权重和 Hub 权重进行计算并且输出一组具有较大 Authority 权重的页面和具有较大 Hub 权重的页面.

3.2 HITS 算法的改进

HITS 算法将超链接创造性地应用到搜索引擎系统当中, 并产生了很好的效果, 但是其不足之处也是显而易见的, 首先, HITS 算法完全不考虑页面的文本内容, 在实际应用中会产生所谓的“主题漂移”, 其主要原因是算法在实现过程中将所有超链接设为相同权重. 其次, 当一个 Hub 页面链接到许多无关 Web 页时, 它的 Hub 权重一般较高, 导致输出的 Hub 页面偏离主题. 这是由于 HITS 算法在计算 Hub 权重时, 采用 Hub 所指向的所有 Authority 权重之和所至.

基于上述两点缺陷, 文中将 HITS 算法的迭代邻接矩阵 A , 改为相似度量矩阵 W , 改进的算法如下:

(1) 设定一个允许误差 T , 作为收敛的最终标准;

(2) 设定页面初值 a_i, h_i , 将它们带入公式:

$$a_i = \sum_{j \in B(i)} h_j, h_i = \sum_{j \in F(i)} a_j$$

(3) 将每一次计算出来的 a_i, h_i 归一化:

$$\sum_i a_i^2 = 1, \sum_i h_i^2 = 1;$$

(4) 如果 a_i, h_i 不收敛, 则转(3);

(5) 将所有 a_i, h_i 大于 T 的值选出, 按大小排序, 输出查询结果.

步骤(3) 细化为: 将网页构成的有向图看作 G

$= (V, E)$, V 为定点集合, E 为边集合, 则存在着 $n \times n$ 的相似度量矩阵 W , 将所有页面对应的 a_i 值和 h_i 值分别定义为向量, 可以根据公式:

$$x \leftarrow W^T y \leftarrow W^T W x \leftarrow (W W^T) x$$

$$y \leftarrow W^T x \leftarrow W^T W y \leftarrow (W W^T) y$$

进行多次迭代, 得到最终的 a_i, h_i . 根据线性代数的理论, 迭代序列经过标准化最终将收敛于矩阵的特征向量.

4 实验

实验用 4 个网页组成了一个网络结构图. 这里只是做了一个实验, 来证明算法的有效性, 网页的链接情况如图 2. 则邻接矩阵为 A , 联合引用矩阵 C 为:

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}, C = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

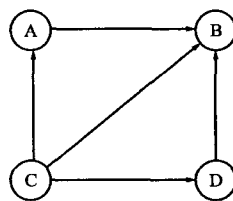


图 2 网络有向图

由于相似矩阵涉及到网页的具体内容, 而此实验中的网页没有具体的文本, 所以假设相似矩阵为 S :

$$S = \begin{bmatrix} 1 & 0.8 & 0.5 & 0.2 \\ 0.8 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.2 \\ 0.2 & 0.5 & 0.2 & 1 \end{bmatrix}$$

则根据表达式(4) α 取 0.8, 有相似度量权重矩阵 W :

$$W = \begin{bmatrix} 0 & 128424 & 0.214 & 125 \\ 128424 & 0 & 0.214 & 12714 \\ 0.3424 & 0.214 & 0 & 0.214 \\ 125 & 128424 & 0.214 & 0 \end{bmatrix}$$

则经过多次迭代如表 1、表 2 所示.

最终收敛于第 14 次, 即 $a_B > a_D > a_A > a_C; h_C > h_D > h_A > h_B$, 网页的权重值被排列出来.

表 1 Authority 值 &Hub 值表(A 和 B)

time	a_A	h_A	a_B	h_B
1	1	1	1	1
2	1	1	3	0
3	0.04579	0.0214	0.0835	0.0148
4	0.03925	0.0263	0.0955	0.0121
5	0.03862	0.0279	0.0976	0.0107
6	0.03858	0.0286	0.0983	0.0101
7	0.03845	0.0291	0.0992	0.0091
8	0.03839	0.0304	0.0999	0.0086
9	0.03828	0.0319	0.1065	0.0092
10	0.03816	0.0326	0.1098	0.0098
11	0.03795	0.0330	0.1167	0.0094
12	0.03764	0.0331	0.1279	0.0094
13	0.03721	0.0338	0.1384	0.0096
14	0.03724	0.0339	0.1383	0.0097

表 2 Authority 值 &Hub 值表(C 和 D)

time	a_C	h_C	a_D	h_D
1	1	1	1	1
2	0	3	1	1
3	0.00184	0.089	0.040	0.035
4	0.00153	0.092	0.042	0.043
5	0.00144	0.097	0.042	0.051
6	0.00136	0.106	0.047	0.059
7	0.00118	0.114	0.049	0.061
8	0.00110	0.119	0.056	0.057
9	0.00091	0.128	0.047	0.062
10	0.00078	0.137	0.043	0.069
11	0.00064	0.145	0.059	0.071
12	0.00049	0.159	0.054	0.073
13	0.00043	0.165	0.054	0.074
14	0.00049	0.164	0.057	0.073

5 结束语

HITS 算法是搜索引擎核心算法之一. 文中在分析网络结构的基础上,介绍了一种基于 Web 文本相似度量的 HITS 算法,该算法能够有效地改进原 HITS 算法的“主题漂移”现象. 而且也提高了算法的效率和正确性,搜索引擎使用它时会更进一步提

高精度度. 该算法挖掘了 Web 超链接之间潜在的语义关系,进一步分析了网络中超链接之间的联系,把超链接分析与内容分析有机地结合起来,并使用户满意的有效网页能更多的排在搜索结果集的靠前位置,从而提高搜索引擎的查全率和查准率.

参考文献:

[1] He Xiaofeng, Zha Hongyuan, Chris H Q Ding. Web document clustering using hyperlink structures[J]. Computational Statistics & DataAnalysis,2002,41(8):19-45.

[2] Soumen Chakrabart. Mining the web discovering knowledge from hypertext date[M]. Indian Institute of Technology: Bombay, Library of Congress Control, 2002:136-140.

[3] Ricardo Baeza-Yates. Information retrieval in the web beyond current search engines[J]. International Journal of Approximate Reasoning,2003,34(4):97-104.

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

[5] 汤亚玲,崔志明. 遗传算法在 Web 关联挖掘中的应用研究[J]. 微电子学与计算机,2006,23(6):126-129.

[6] 崔志明,谢春丽. 基于 Web 的文本挖掘研究[J]. 微电子学与计算机,2002,10(4):51-53.

[7] 夏天. 汉语词语语义相似度计算研究[J]. 计算机工程, 2007,33(6):191-194.

[8] 张敏,高剑峰,马少平. 基于链接描述文本及其上下文的 Web 信息检索[J]. 计算机研究与发展,2004,41(1):36-44.

作者简介:

何 明 男,(1982-),硕士研究生. 研究方向为 Web 结构挖掘、搜索引擎算法优化.

周 军 女,(1966-),博士,教授. 研究方向为数据挖掘、知识发现、近似计算等.

纪周鹏 男,(1981-),硕士研究生. 研究方向为 Web 使用挖掘、模糊集理论.

李树友 男,(1964-),博士,教授. 研究方向为统计推断、随机信息处理.

(上接第 193 页)

[7] Byeong-Tae Lee, Sunwoo Kim. Scalable DV-Hop localization for wireless sensor networks[C]// International Conference on Computer Communications and Networks. Korea:Hanyang University, 2008:1-4.

[8] 王刚,王长山. Leach 协议簇头节点选取策略的改进[J]. 微电子学与计算机,2009,26(6):254-256.

作者简介:

刘丽君 女,(1983-),硕士研究生. 研究方向为无线传感器网络.

陈志奎 男,(1968-),教授,博士生导师. 研究方向为无线传感器网络、图像处理、无线通信等.