基于最大流及页面相似度的Web结构挖掘

李 莹,吴晓军
(陕西师范大学 计算机科学学院,陕西 西安 710062)

摘 要:针对Web结构挖掘算法容易出现"主题漂移"以及主机间的多重互相加强关系的问题,提出了一种基于最大流与页面相似度值的超链接结构挖掘方法。该方法在传统的超链接结构挖掘算法HITS的基础上引入页面相似度值构造邻接矩阵,并结合基于最大流的Web社区发现技术来构建特征向量空间模型,通过迭代计算最终获得价值最高的权威结果集和中心结果集。实验结果证明该方法有较好的查准率与查全率,并有效抑制了"主题漂移"现象,具有一定的实用价值。

关键词:Web结构挖掘;主题漂移;页面相似度值
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2011)10-0112-04

Web Structure Mining Based on Maximum Flow and Page Similar Value

LI Ying , WU Xiao-jun
(School of Computer Science, Shaanxi Normal University, Xi'an 710062, China)

Abstract: Aiming to Web structure mining algorithm is easy for a "topic drift" and mutually strengthening relations among the hots of the problem, a method of hyperlink structure mining based on the maximum flow and the page similarity value is presented. On the basis of traditional HITS algorithm, this method introduced the page similarity value and adopted the Web communities identification based on the maximum flow to construct the models of feature vector space. And then the calculation eventually won the highest value of authority-set and hub-set by iterative method. Experimental results show that the method has better recall and precision, what's more it effectively inhibits the theme of Web structure mining algorithms drift, has some practical value.

Key words: Web structure mining; topic drift; page similar value

0 引 言

大量的互联网用户,在面对海量的数据信息时,虽说通过搜索引擎的使用可以大大减少无用信息的干扰,但是搜索得到的结果[1]很可能会不完整或不相关,很难完全地满足其需求。然而随着Web数据挖掘技术的不断发展,过量信息问题得到了很好地解决,这使人们在互联网络中可以更容易的得到更精确、更相关的数据信息。

HITS[2]算法是在Web结构挖掘中,最具有权威性和使用最广泛的算法,它主要是利用页面之间的引用链来挖掘隐含在其中的有用信息(如权威性),具有计算简单且效率高的特点。HITS算法认为[3]对每一个网页应该将其内容权威度和链接权威度分开来考虑,在对网页内容权威度做出评价的基础上再对页面的链接权威度进行评价,然后给出该页面的综合评价。但是在实际应用中HITS算法容易发生"主题漂移"[4~6]问题。产生这一问题[7]的根本原因就是在基集扩展的过程中引入了一些与查询主题不相关的页面以及在权威值(Authority)与中心值(Hub)的计算过程中没有考虑到Web页面内容与查询主题之间的"相关度",它忽视了Web页面超链接之间的差异性,而且将所有的超链接赋予了相同的权值,因此使得最密集链接区域中一些不相关的或者具有非支配性的主题但是链接数很多的页面影响权值过高。文中在以往改进的HITS算法基础上利用最大流社区发现方法和相似度空间向量投影方法的核心思想来解决此问题。实验结果表明该方法能大幅度提高HITS算法的精确度,很好地抑制了HITS算法的"主题漂移"现象。

收稿日期:2011-02-26;**修回日期**:2011-06-04
基金项目:中央高校基本科研业务费专项资金资助(GK201002005);陕西省工业攻关计划(2009K09-21)
作者简介:李 莹(1984-),女,硕士研究生,研究方向为嵌入式开发与模式识别;吴晓军,副教授,研究方向为系统工程、模式识别、嵌入式系统、智能系统、计算机软件。

1 基于最大流的Web社区发现

G. Flake[8,9]等人证明了采用S-T最大流算法挖掘出的Web社区页面集具有社区内页面之间的链接

要比社区外的页面链接稠密的性质。利用这一性质提出了基于最大流的 Web 社区发现算法。算法[10]具体的步骤如下:

Step1:假设 Q 为包含若干节点的种子节点,$Q=\{q_1,q_2,q_3,\cdots,q_n\}$。根据 Q 中的节点往下扩展到一定深度,生成一个 Web 网络子图 $W=<V,E>$。W 被称做为邻近图。

Step2:假设对任意 $e(e\in E)$ 全部是双向的,并且 e 具有的容量 $c(e)=|Q|$。

Step3:在 V 中增加一个虚的源点 q_i,将 q_i 与 Q 中的所有节点连接,同时将这些连接在一起的边的容量设为∞。

Step4:在集合 V 中增加一个虚的汇点 p_i,将 p_i 与 $V-\{Q\cup q\cup p\}$ 中的所有节点连接,并且将这些连接的边的容量设置为 1。

Step5:运行 S-T 最大流-最小割算法,得到与源点相连接的节点集合,以成为 Web 社区的新成员。

Step6:将 Web 社区里的新添的成员增加到 Q,循环执行 Step3,直到 Web 社区的大小稳定为止。

最大流 S-T 算法并没有选用静态的 Web 数据源,它是利用集中抓取器检索页面,在抓取的过程中来得到 Web 图形,具体的过程则是从一个 Web 页面的种子集合起开始抓取,找到所有与种子节点有链接的页面,并设置一个虚拟的源点和汇点,来找到最小的割集,最后产生 Web 社区。Web 社区[11]是与主题相关的 Web 页面集合,也是 Web 网络图中节点间链接非常紧密的区域,利用社区内与社区外的链接稀疏的特点,通过求出 Web 网络图中的最小割集合,来找出最小边权值之和的划分,进而发现 Web 社区。文中则是利用最大流 Web 社区发现算法来对 HITS 进行改进。

2　基于页面相似度值的空间投影方法

2.1　页面相似度值的引入

互联网当中的 Web 页面很多都是以文本的形式存在,而 Web 内容挖掘则是通过对 Web 页面内容的分析来衡量一个页面的重要程度。现有的文本分析技术[12]比较多,应用在搜索引擎中比较广泛的则是向量空间模型(VSM),通过该模型来计算 Web 页面内容和查询主题这两者的相似度值,最后利用相似度值来对搜索的结果进行排序。

在计算文档相似度时,可以把文档 D_i 表示为 $D_i=(d_{i1},d_{i2},\cdots,d_{im})$,查询 q 可以表示为 $q=(q_1,q_2,\cdots,q_n)$,其中 q_i 代表查询主题中第 i 个关键字,利用两矢量夹角的余弦值来求得文档矢量 D_i 与查询矢量 q 之间的相似度为[13]:

$$\text{Similarity}(d,Q)=\frac{\sum_{i=1}^{m}d_iq_i}{\sqrt{\sum_{i=1}^{m}d_i^2\sum_{i=1}^{m}q_i^2}} \tag{1}$$

由公式(1)可以看出,相似度值和 Web 页面与查询主题的相关度成正比关系。因此可以利用页面相似度值作为衡量 Web 页面与查询主题的相似程度的影响因子。

2.2　构造基于相似度值的子空间

页面相似度值可以通过搜索引擎返回的结果集中得到,利用页面相似度值构造邻接矩阵 $\boldsymbol{L}$:

(1)若页面 i 与页面 j 没有链接,则矩阵元素 $\boldsymbol{L}(i,j)$ 的值为 0;

(2)若页面 i 存在与页面 j 的链接,则矩阵元素 $\boldsymbol{L}(i,j)$ 的值为搜索引擎返回的页面相似度值。

然后将 Web 网络图中所有页面的权威 Authority 值和中心 Hub 值分别用两个 n 维向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 来表示并将向量初始化为 1;再将页面相似度值用一个 n 维向量 $\boldsymbol{Z}$ 来表示。

从 $\boldsymbol{Z}$ 中选出相似度值排在前 200 的页面组成一个集合 $\boldsymbol{G}$,构造一个与 $\boldsymbol{G}$ 相对应的欧几里得子空间即基于页面相似度值的子空间。

2.3　根集子向量空间投影

(1)计算出邻接矩阵 $\boldsymbol{L}$ 的每个特征值以及相对应的特征向量,然后对特征向量取绝对值并在其子空间上进行投影 p。

(2)在投影 p 中找到特征向量 e^*(e^* 是使 $\|p\lambda_ie_i\|$ 的值达到最大时的特征向量),并归一化 e^*。

(3)将 e^* 中绝对值最大的元素来分别作为权威 Authority 的值和中心 Hub 的值返回。

其中 $\|p\lambda_i^*e_i^*\|$ 的计算实质上是将 $\|p\lambda_i^*e_i^*\|$ 与 $\boldsymbol{G}$ 向量做标积,若 e 表示归一化后的 $\boldsymbol{G}$ 向量,那么 $\|p\lambda_i^*e_i^*\|=\lambda_i^*ee_i^*$。

算法改进的核心就在于利用现有的搜索引擎所返回的页面相似度值来评估 Web 页面与查询主题之间的相关程度。

这就意味着 Web 网络社区中,一方面虽然某些社区内部的链接总数不多,但是其中一些页面与由相似度值比较高的页面组成的集合内的那些页面的链接数量非常大,那么这些页面被选中并从中提取 Authority 值的可能性就非常大。另一方面,某些社区内部虽然拥有的链接数很多,但是与高相似度值组成的页面集合的链接却并不怎么紧密,所以选中这些页面的可能性会降低。通过相似度值的引入就可以较好地抑制主题漂移现象的产生。

3 MCDHITS 算法

为了可以很好地抑制 HITS 算法存在的主题漂移现象，同时大幅度地降低算法运行时的系统开销，文中在传统的 HITS 算法基础上，提出了 MCDHITS 改进算法—基于最大流与页面相似度值的向量空间投影算法。算法的改进主要包括两方面：(1)预处理根集 R，扩展后得到高质量的基集 T；(2)引入相似度值并构造空间模型，由最相关特征向量在相似度值的子空间模型投影获取到最高的 Authority 值和 Hub 值。

算法具体描述如下：

输入：查询请求 q；根集数量设为 $k = 200$；基集数量设为 $d = 5000$；迭代次数 K；

输出：最大 Authority 值和 Hub 值。

Step1：初始化。首先将根集 R 中所包含的节点进行分组。以 $m(m > 1)$ 个节点为一组，并将其内的节点作为种子节点。如果在 R 集中有 k 个节点（$k = 200$），$R = (n_1, n_2, \cdots, n_k)$。让 R 集中的节点按照与主题的相关程度由高到低进行排列（$\mathrm{Rel}(n_i) \geqslant \mathrm{Rel}(n_i + 1)(1 \leqslant i \leqslant k - 1)$），并将相关程度比较相近的节点划分到同一组，即 $R = \{\{n_1, n_2, \cdots, n_m\}, \cdots, \{n_{k-m-1}, n_{k-m}, \cdots, n_k\}\} = \{S_1, S_2, \cdots, S_{km}\}$；

Step2：扩展 R 集。对于 S_i 中的任意一个节点进行深度为 2 的扩展，得到链接有向图 $W_i(V_i, E_i)$。加入虚拟的源点 S_i，加入虚拟的汇点 T_i。对所有在 S_i 中的节点 a，将与其连接的边 (s_i, a) 加入到 E_i 中，并设置边容量 $c(s_i, a) = \infty$。对于所有在 E_i 中的边 (a, u)（其中 $u \neq s_i$），设置 $c(a, u) = \alpha$。如果 (a, u) 不在 E_i 中，则将边 (a, u) 加入到 E_i 中，并设置边容量 $c(a, u) = \alpha$（α 的大小对社区中节点数量的发现会有影响）。对于所有在 V_i 内的节点 $a(a \neq s_i)$，将边 (a, t_i) 加入到 E_i 中，并设置边容量 $c(a, t_i) = 1$。运行最大流最小割 S-T 算法获取与源点仍然相连的节点集合 S_i'。然后合并得到的 $S_i'(1 \leqslant i \leqslant k/m)$ 为基本集 即 MCD-T 集 $= \{a \mid a \in S_i', 1 \leqslant i \leqslant \frac{k}{x}\}$。

Step3：根据基集合 MCD-T 构造子图 A；

Step4：给子图 A 中的每个页面 i，定义一个非负的权威 Authority 值 X_i 和非负的中心 Hub 值 Y_i，同时用两个 n 维向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 来分别表示所有的 X_i, Y_i，初始化两个向量为单位向量；

Step5：定义一个 n 维向量 $\boldsymbol{W}$，其中每一个分量都取值为对应页面的相似度值（相似度值由搜索引擎返回的结果获得）。权值向量 $\boldsymbol{W}_i$ 由相似度计算得到，即 $\boldsymbol{W}_i = \mathrm{Similarity}(d_i, Q)$；

Step6：利用根集子空间向量投影算法，获取特征向量 e^* 绝对值最大的元素分别作为 Authority 值和 Hub 值返回；

Step7：返回 Authority 值和 Hub 值最大的前 10 个网页。

该算法第一步可以求出与 R 集相邻的密集区域，这就相当于一个社区发现的过程。首先，分别对每一组种子节点进行 2 层扩展来得到 Web 网络子图，然后对子图进行合理改造，使其符合 S-T 算法运行的条件，条件满足后利用最大流最小割算法划分出密集区域的边界，最后合并各个小组的结果得到精简后的基集。而算法第五步引入页面相似度值则可以衡量 Web 页面与查询主题的相关度，这大大降低了算法的系统消耗。

4 实验结果与分析

MCD-T 集可以在根集扩展的过程中，通过对边赋予不同的权值来控制其网页规模，权值越大，网页的规模就越大。在该实验中，权值 $\alpha = 5$。查询主题为“gulf war”，则两种算法的运行结果如表 1 所示（取排名前 6 的页面结果集）。

表 1 HITS 与 MCD-HITS 对于主题 gulf war 的运行结果比较

传统 HITS 算法	权威值
http://smtp.ngwrc.org/exchange	0.456331
http://www.flash8.net/flash/6128.shtml	0.447066
http://www.pbs.org	0.440955
http://www.chinadaily.com.cn	0.326199
http://english.cri.cn	0.324076
http://www.yxdown.com/SoftView/	0.321684
MCDHITS 算法	**权威值**
http://www.gulfwarvets.com	0.439776
http://en.wikipedia.org/wiki/Gulf_War	0.281571
http://en.wikilib.com/wiki/Gulf_War	0.250083
http://www.csmonitor.com	0.248829
http://www.ngwrc.org	0.1793 73
http://www.va.gov/gufwar	0.167373

从表 1 中的结果数据可以看出，在传统的 HITS 算法查询得到的结果中出现了一些与慈善、商业机构相关的页面，很明显主题发生了漂移。而改进后的 MCDHITS 算法所返回结果大都是与主题相关的。

HITS 算法和 MCDHITS 算法针对五个实验主题（见表 2）分别从基集 Web 页面数目、权威页面主题相关性和中心页面主题相关性三个方面对实验结果进行分析。如图 1 ~ 图 3 所示。

表 2 搜索关键字列表

关键字	Gulf war	HIV	Table	Bicycle	Climbe
编号	1	2	3	4	5

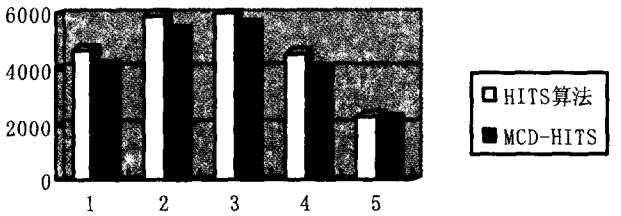

图 1　基集 web 页面数目比较

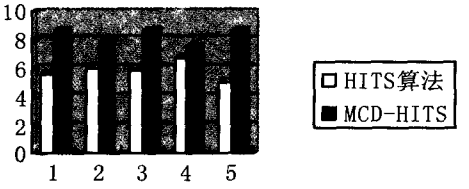

图 2　权威页面主题相关性比较

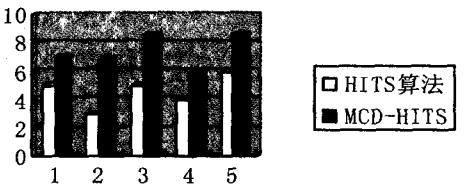

图 3　中心页面主题相关性比较

从上图 1 可以看出,MCDHITS 算法在优化了基集之后获得的 Web 页面数目明显少于传统 HITS 算法,这很大程度上节约了算法运行的系统开销。

图 2、图 3 则可以看出,改进后的 MCDHITS 算法大大提高了查询结果的精确度,同时较大程度上降低了传统 HITS 算法发生主题漂移问题的可能性。

5　结束语

文中在深入研究了网页超链接结构之后,在传统的 HITS 算法基础上,针对原算法出现的问题提出了一种改进算法 MCDHITS。该算法一定程度上解决了 HITS 算法的主题漂移问题,并提高了 HITS 的精确度。算法通过对 R 集进行的两层扩展而得到一个比原来范围更广的 Web 网络子图,之后采用最大流最小割 S-T 算法思想来发现以 R 集为中心的社区。社区很好地优化了基集,在高质量的基集上利用页面相似度的空间向量投影算法思想,最终获取高的 Authority 页面和 Hub 页面。实验结果表明 MCDHITS 算法提高了查询结果的查全率和查准率,并降低了算法运行的系统开销。

参考文献:

[1] 欧阳柳波,李学勇,李国徽,等. 专业搜索引擎搜索策略综述[J]. 计算机工程, 2004(13):32-33.

[2] 杜光芹,张化祥,赵瑞东. 主题 Web 挖掘研究 [J]. 计算机技术与发展, 2008,18(2): 94-97.

[3] 卢虹宇. Web 结构挖掘中 HITS 算法的研究 [D]. 成都:西南交通大学, 2008.

[4] 王宇新,刘海峰,郭　禾,等. 一种有效的专题信息集中和检索策略[J]. 计算机应用研究, 2010(6): 2106-2108.

[5] 常　庆,周明全,耿国华. 基于 PageRank 和 HITS 的 Web 搜索[J]. 计算机技术与发展, 2008,18(7): 77-79.

[6] 罗林波,陈　绮,吴清秀. 基于 Shark-Search 和 Hits 算法的主题爬虫研究[J]. 计算机技术与发展,2010,20(11): 76-79.

[7] 邱东洋,汤小春. 一种基于超链和锚文本分析的主题发现算法[J]. 微电子学与计算机, 2009(6):125-128.

[8] Flake G W,Lawrence S, Giles C L. Efficient identification of web communities[C]//Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. USA: ACM Press, 2000: 150-160.

[9] Flake G W,Lawrence S, Giles C L,et al. Self-organization of the web and identification of communities[J]. Communities, 2002, 35(3): 66-71.

[10] 杨　楠,弓丹志,李　忱,等. Web 社区发现技术综述[J]. 计算机研究与发展, 2005(3):439-447.

[11] 罗彩君. Web 社区结构挖掘的研究与应用[D]. 西安: 西北大学, 2008.

[12] Deng R M A G. A mixed weight HITS in link analysis[C]// Proceedings of the 2010 International Conference on Circuit and Signal Processing (ICCSP 2010) & 2010 Second IITA International Joint Conference on Artificial Intelligence (IITA-JCAI 2010). Shanghai:[s. n.],2010

[13] 王艳华,张　纪. Web 结构挖掘及其算法 [J]. 计算机工程, 2005(S1):125-127.

(上接第 111 页)

images/2010_BPM_Handbook_Free_Chapters. pdf.

[9] Baeyens T. Process Virtual Machine [EB/OL]. [2009-11-21]. http://docs. jboss. com/jBPM/pvm/article.

[10] Activiti Team. Activti User Guide[EB/OL]. [2010-09-11]. http://www. activiti. org/userguide/index. html.

[11] jBPM Team. jBPM User Guide[EB/OL]. [2010-08-11]. http: //www. jboss. org/jBPM.

[12] 胡长城. 开源工作流平台 jBPM:过程组件模型与 PVM[J]. 程序员,2008(5):116-118.

[13] Workflow Management Coalition. The Workflow Reference Model[EB/OL]. [2010-06-20]. http://www. wfmc. org/Published-Research/View-category. html.

[14] 叶　娜,李　健. 工作流引擎推进过程中 m 选 n 问题的研究[J]. 计算机应用研究,2009,26(11):4098-4100.

[15] 余　阳,汤　庸,潘茂林,等. 时态工作流过程模型及其合理性验证[J]. 软件学报,2010,21(6):1233-1251.