基于 Hyperlink 和相关度发现 Web 相关文档的研究

王天江[1]，叶卫国[1]，卢正鼎[1]，李永平[2]

[1](华中科技大学 计算机学院，湖北 武汉 430074)
[2](国家药品监督管理局，北京 100810)

摘 要：分析了 Web 文档的相似度计算方法，提出了 Web 上查询相关信息发现的 SW-HITS 算法，它结合了 Web 超链接、网页知识表示的信息相关度以及 HITS 方法来搜索 Web 上相关知识．本文通过它们搜索网上有关医药的信息和知识，其效果和查准率比传统 HITS 和 IR 方法有一定提高．

关 键 词：HITS；相似度计算方法；信息检索；VSM
中图分类号：TP391；TP393　　**文献标识码**：A　　**文 章 编 号**：1000-1220(2004)08-0859-04

Finding Relevant Documents Using Hyperlink and Similarity Measure on the Web

WANG Tian-jiang[1], YE Wei-guo[1], LU Zheng-ding[1], LI Yong-ping[2]

[1](*Department of Computer Science, Huazhong University of Science and Techonology, Wuhan Hubei* 430074, *China*)
[2](*State Drug Administartion, Beijing* 100810, *China*)

Abstract: In this paper we discuss Similarity Weighted-HITS (SW-HITS) algorithms in finding relevant documents on the Web. These methods not only use the hyperlinks of web graph, but also the similarity scoring of term weights in document representations. We implement the algorithm to find Chinese medical information from the Internet. Our study showed that it has better precision than traditional IR methods and basic HITS alogorithms.

Key words: HITS; similarity scoring methods; information retrieval; VSM

1 引 言

人们已经进入信息极大丰富的时代，一方面信息来源广泛，包括 Web 文档、图书文献、数字化资料等，这些异构的信息分布在 Internet 空间中；另一方面信息量巨大．以 Web 文档为例目前已经拥有 3 亿页面，而且这个数字仍以每 4 至 6 个月翻一倍的速度增加．面对信息的海洋，人们觉得力不从心，往往花费了很多时间却所获甚少．在这种情况下如何有效地提供基于 Internet 的资源发现服务，以帮助用户从大量信息资源的集合中找到与给定的查询请求相关的、恰当的资源子集，也就成为一项重要而迫切的研究课题[1]．

传统的搜索引擎，例如 AltaVista、Google 等[2]试图解决 Internet 上的资源发现问题，但是从资源覆盖度、检索精度、检索结果可视化、可维护性等诸多方面来看其效果远不能够令人满意．假设你关注如 MEDLine 的医疗治疗信息，搜索引擎通常返回成千上万条记录，而这些记录的相关度差异很大，从包括专门信息的网站到几乎完全无用的站点．判断一个 Web 网页的相关度本身是必须由人工完成的，主要影响判断的因素有网页的组织、布局以及网页的信息质量；完全由人工进行的判断在实际中是不可行的．

目前基于 Web 超链接(Hyperlink)的算法得到了更多关注，并在解决上述问题中得到了很好的应用．Kleinberg 的 HITS[3]算法是其中应用比较普遍的、基于网页文档间的链接来排列文档的算法；它基于如下假设：一个文档如果指向很多其它文档就认为是一个 Hub，一个文档如果被很多其它文档指向就认为是一个 Authority；Hub 和 Authority 之间有互相增强关联的关系：一个好的 Hub 指向多个好的 Authority，一个好的 Authority 指向多个好的 Hub．有些研究利用 co-citation[4]和其它类型的连接关系来解决上述问题，co-citation 的含义在于如果网页 A 指向网页 B 和 C，那么 B 和 C 就有可能相关，这一思想来源于书籍检索系统．co-citation 只利用了网页的邻接图，不像 HITS 有互相增强的递归关联关系，其在 Web 上的效果比 HITS 更差．

我们关注高查准率、与给定的查询请求高相关度的 Web 资源发现问题；Web 网页有着自己的特性：动态性、局部性(或聚类性)，相关的网页间有链接的概率比较大、整个 Web 图的直径(Diameter)只有大约 19[5]．本文中我们改进了 HITS 算法，不仅利用 Hyperlink 结构，而且利用网页中的语义词汇来做权重，我们称为 SW-HITS(Similarity Weighted-HITS)算法；我们的实验表明该算法的效果有一定提高．

2 背景技术

我们把 Web 网页集合 V 看作一有向 Web 图 G = (V; E)，这里 V = {v_1, v_2, ... v_n} 是等同于网页的节点集，有向

收稿日期：2003-07-23　作者简介：王天江，博士、副教授，主要研究领域为人工智能、数据挖掘．E-mail: wt_jiangmail@yahoo.com.cn；叶卫国，博士生，主要研究领域为 Web 上信息获取、计算机网络安全；卢正鼎，教授、博士生导师，主要研究领域为分布式对象平台、工程数据库、智能结构系统；李永平，博士后，主要研究领域为 Web 上信息获取、信息监管、电子商务．

边($i\rightarrow j$)表示节点 i 有一条超链接到节点 j;当从 web 网页 i 到网页 j 有边 $i\rightarrow j$ 时称 i 是 j 的父节点,j 是 i 的子节点.在 HITS 应用于 Web 上的搜索排序时,对于一个查询首先收集一个基本文档集;对于其中每个文档,递归计算其 Hub 和 Authority 值.为得到基本文档集 I,首先从搜索引擎得到一个与查询有关的根集 R;对于每个文档 $r\in R$,指向 r 和 r 所指向的文档集合作为 R 的近邻加入到集合 I 中.对于一个文档 $i\in I$,设 a_i 和 h_i 是相应的 authority 和 hub 值;刚开始时 a_i 和 h_i 被初始化为 1,当 a_i 和 h_i 没有收敛时,算法递归计算如下:

1. 对所有指向 i 的 $i'\in I(i'$ 为 i 的父节点)

$$a_i=\sum_{i'}h_{i'} \tag{1}$$

2. 对所有 i 指向的 $i'\in I(i$ 为 i' 的父节点)

$$h_i=\sum_{i'}a_{i'} \tag{2}$$

3. 标准化 a_i 和 h_i 值使得 $\sum_{i'}a_{i'}=\sum_{i'}h_{i'}=1$

Kleinberg 证明了算法最终会收敛,但递归计算的次数没有确定的界限.

文[6]中提出了改进的 HITS(BHITS)算法,在 HITS 中递归计算公式 1 和 2 中边的权重可看作 1,而 BHITS 基于站点(Host)来对 Web 图中的边赋权值:公式 1 中,如一个站点有 k 条边指向另一站点的一个文档,将这些边的 authority 权重赋为 $1/k$;公式 2 中,如某站点上的一个文档有 l 条边指向另一站点,将这些边的 hub 权重赋为 $1/l$.

尽管 BHITS 在通常情况下效果良好,但它同 HITS 一样有产生很坏结果的情况出现[3].当生成基本文档集的根集 R 有着很少的 in-links 但 out-links 数目比较大时,并且其中大多数网页与查询不相关时,HITS 以及 BHITS 的效果相当差(这称为 small-in-large-out 现象).没有分析文档的内容,HITS 和 HITS-Based 算法不可能有效、高查准率地解决问题[7].

向量空间(VSM,Vector Space Model)模型是目前 IR 系统中常采用的方法[8],它将文档和查询都用向量来表示:假设 m 为文档集合中不同词条(terms)的数目,文档 d 可以表示为范化的 m 维特征向量 $V(d)=\{(t_1,w_1(d)),\Lambda(t_m,w_m(d))\}$,其中 t_i 为词条项,$w_i(d)$ 为 t_i 在 d 中的权重值.词条权重 $w_i(d)$ 一般采用 TFIDF 函数来表示 $\varphi=tf_i(d)\times\log(N/n_i)$,其中 N 为文档集合中文档的数目,$n_i$ 为文档集合中含有词条项 t_i 的文档数目,$tf_i(d)$ 定义为 t_i 在文档 d 中出现的频率.

有多种表示文档内容的方法用来进行有效的相关度计算,主要有三种方法选择词条来组成表示 Web 网页 i 的向量[9]:

1. content-based 方法:在文档 i 中出现的词语

2. link-based 方法:由网页 j 指向 I 的网页标识符(如 URL)

3. anchor-based 方法:在网页链接点及附近出现的词语(如 Some papers can be found here),这里here为链接点

类似地查询语句 q 也可以采用向量 $Q=\{(t_1,w_1^q),\cdots,(t_m,w_m^q)\}$来表示,其权值计算通常采用 tf_i 或取为常数.向量之间的相似度函数(如内积)有着良好的数学和应用背景;IR 中有一些常用测度用于文档的相关度计算[5].文档向量 V_i 和查询向量 Q 的 *Jaccard* 测度定义为:

$$sim_J(V_i,Q)=\frac{|V_i\cap Q|}{|V_i\cup Q|}=\frac{\sum_{k=1}^{m}\min(w_k^i,w_k^q)}{\sum_{k=1}^{m}\max(w_k^i,w_k^q)} \tag{3}$$

文档向量 V_i 和查询向量 Q 的 *cosine* 相似测度定义为:

$$sim_{\cos}(V_i,Q)=Q\cdot V_i=\frac{\sum_{k=1}^{m}w_k^i\cdot w_k^q}{\sqrt{\sum_{k=1}^{m}(w_k^i)^2\cdot\sum_{k=1}^{m}(w_k^q)^2}} \tag{4}$$

Okapi 相似测度是目前传统 IR 系统中最流行的方法之一,它不仅考虑查询词条的频率,而且也考虑了整个文档集合的平均长度和文档本身的长度:

$$sim_o(V_i,Q)=Q\cdot V_i=\sum_{k=1}^{m}w_k^i\cdot w_k^q \tag{5}$$

这里 w_k^q 是查询中词条出现的频率;dl 是 web 文档的长度,$avdl$ 是集合内文档的平均长度,w_k^i 定义为:

$$w_k^i=\frac{tf_{ik}\cdot\log(\frac{N-d_k+0.5}{d_k+0.5})}{2(0.25+0.75\cdot\frac{dl}{avdl})+tf_{ik}}$$

查询和 web 文档间的相似度可定义为文档向量 V_i 和查询向量 Q 的 Jaccard、Okapi 或 Cosine 测度.本文中我们主要考虑 Jaccard 测度,因为它比较容易实现,只需要基本的 union 和 intersection 操作,不需其它复杂的计算.

3 相关信息发现的 SW-HITS 算法

在本节中我们描述全新的 Web 上查询相关信息发现的 SW-HITS(Similarity Weighted-HITS)算法;它结合运用了基本文档集 I 中的超链接、网页知识表示的信息相似度信息以及采用权重来破坏 small-in-large-out 现象.公式(1)和(2)变为:

1. 对所有是 i 的父节点的 $i'\in I$

$$a_i=\sum_{i'}w_{sim(q,i')}*h_{i'} \tag{6}$$

2. 对所有 i 的子节点,$i\in I$

$$h_i=\sum_{i'}w_{h_{i'}}*a_{i'} \tag{7}$$

在公式(6)、(7)中 $h_{i'}$ $a_{i'}$ 可以是 HITS、BHITS 或其它 HITS-based hub 和 authority 值.我们把在根集 R 中的边的 $w_{h_{i'}}$ 设为 1.1,其它设为 1.而 $w_{sim(q,i')}$ 的取值包括两部分:

1. 设所有的 $w_{sim(q,i')}=SIM_J(Q,V_{i'})$

2. 如果存在一个网页 i 其入度<3 且在基本文档集 I 组成的 Web 图中其出度为最大的三个出度,对于所有为 i 的父节点的 $i'\in I$,其 $w_{sim(q,i')}$ 乘以 4.

算法描述如下:

首先,对于一反映用户需求和兴趣的查询 q,如"癌症治疗".我们将查询 q 分解为独立的词条;并去除普通非用词汇如"在"、"和"等;只保留名词词条;对名词采用同义词替代、缩

写、扩充、及采用如 KingSoft 词典获得语义相关的关键词.我们得到一个词条集合 T,它组成了用户查询的语义集;从集合 T 得到了用户查询向量 Q;其过程见图 1.

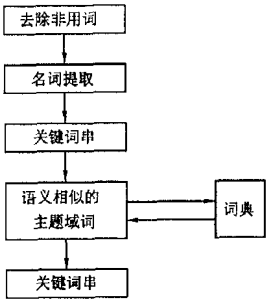

图 1　查询主题语义集生成

Fig. 1　Creation of semantic set of search toptic

根据 GVU 中心的调查报告,全世界大约 84.8%的 WWW 用户通过搜索引擎获得网页[10],且约 92%的用户访问网页后沿着网页内的 hyperlink 获取其它网页,但其访问深度不超过 2 级.因而不同于普通 HITS,我们由 R 得到基本文档集 I 及其 Web 图的方法如下:对于每个 $t\in T$ 将查询串 t 提交给一个通用的搜索引擎,搜集前 c(我们选择为 300)个排序最前面的网页,删除重复网页,得到一有向图 $G_0(V_0,E_0)$,V_0 作为根集 R,G_0 作为根图.G_0 演化生成过程如下,对于网页 $i\in R$:

Go back (B)和 *back-forward*:如果 i 有超过 B_1 个父节点,随机加入 B_1 个父节点到图中,否则添加全部父节点;如果 i 的父节点 j 有超过 B_1 个子节点,加入最先的 B_1 个子节点到图中,否则添加全部子节点;

Go forward (F)和 *forward-back*:如果 i 有超过 F_0 个子节点,随机加入 F_0 个子节点到图中,否则添加全部子节点;如果 i 的子节点 j 有超过 F_1 个父节点,加入最先的 F_1 父节点到图中,否则添加全部父节点.

我们得到演化的 Web 图 $G_2(V_2,E_2)$,V_2 作为基本文档集;这里生成基本文档集共添加了两层的上下文,而不像 HITS 只添加邻接网页.

第三步消除近似重复(near-duplicates):两个节点是近似重复的,如果它们之间有超过 10 条链接,或者它们的链接指向有 90%相同.这有可能是由于镜像网页或镜像站点产生.

最后采用公式(6)、(7)来对边或链接赋予权值,并采用 HITS 或 BHITS 算法来得到结果;算法返回 10 条(或更多)权威权值最高的网页作为查询 q 的结果.

4　实　验

国家食品与药品监督管理局负责对药品的研究、生产、流通、使用进行行政监督和技术监督,包括药品广告和信息的监管.我们用上述算法从网上搜索中文医药信息,共测试了 18 个医药查询,它们是肝炎、癌症、胃病、流感、医药电子商务、抗生素、肾炎、肺炎、化学制剂、青霉素、进口药、皮肤病、妇科病、咽炎、阿斯匹林、保健品、脑膜炎、关节炎.为了方便查找 Web 图的链接,我们爬行了四个主要中文医药站点:999.com.cn,e135.com,777.com.cn,we188.com,共有 3851 个节点和 26387 条链接.我们用 Google 得到 c 个排序最前面的网页,并取在这四个站点内的网页作为根集 R.在实验中我们取 $B_0=50$,$F_0=20$,以及 $B_1=F_1=10$.

在实验中我们选择基于 Web 文本内容(content)和链接窗方法来表示文档,并去除所有的 HTML 注释、标记和非用词语.如果边 $l=(v,u)$,对于链接窗方法文档定义为(链接正文,网页 u 的标题);链接正文定义为在 v 网页内链接窗(*anchor-window*[9])内的单词,链接窗包含链接点指向网页 u,链接窗的大小选定为 16.

Table 1　The precision of 18 queries computed by HITS, SW-HITS and co-citaion

查询	链接数	HITS	BHITS	WB-CON	WB-AN	Co-citation
1	2465	0.545	0.5	0.8	0.73	0.5
2	2321	0.5	0.505	0.615	0.62	0.465
3	2789	0.42	0.54	0.55	0.575	0.3
4	2634	0.525	0.495	0.6	0.575	0.61
5	1652	0.55	0.55	0.7	0.71	0.5
6	1995	0.625	0.615	0.64	0.665	0.645
7	1958	0.61	0.625	0.825	0.76	0.45
8	2062	0.56	0.58	0.69	0.66	0.52
9	2386	0.81	0.785	0.91	0.89	0.6
10	2461	0.795	0.78	0.96	0.895	0.65
11	2245	0.58	0.63	0.75	0.7	0.59
12	2534	0.765	0.85	0.845	0.81	0.6
13	1732	0.56	0.7	0.885	0.885	0.355
14	1625	0.62	0.62	0.865	0.82	0.625
15	1806	0.3	0.5	0.62	0.58	0.35
16	2017	0.375	0.41	0.75	0.72	0.415
17	1918	0.785	0.79	0.925	0.91	0.535
18	2023	0.81	0.885	0.875	0.88	0.5
平均查准率		0.5964	0.6311	0.767	0.7436	0.5117

可以采用其它词条权重法(如 TFIDF),但实验中我们仅选用 TF 函数来作为权重.选取算法返回的权威权值最高的 20 个网页来作评估,评估由一些专家来进行,对返回网页和查询进行相关性评价,并给出评分,评分范围 0-10,0 为不相关,10 为最相关;然后计算 20 个网页的平均相关分数.为了方便比较,我们同时采用 HITS、BHITS、Co-Citation 算法来计算;采用链接窗方法和文档内容分析的 SW-HITS 分别记为 WB-AN、WB-CON.结果见表 1.

结果的平均相关分数可以看作 IR 系统中常用的平均查准率;从表 1 中可以 SW-HITS 算法大约为 0.75,而 HITS 以及 Netscape 中 co-citation 算法的平均查准率小于 0.7,SW-HITS 算法比它们有一定的提高.WB-CON 和 WB-AN 算法在效果上的差异不大,原因可能在于二者都利用了 Web 文档的内容表示方法.

5 结束语

本文分析了 Web 文档的相似度计算方法，提出了 Web 上查询相关信息发现的 SW-HITS 算法，它结合了 Web 超链接、网页知识表示的信息相关度以及 HITS 方法来搜索 Web 上相关知识. 本文通过它们搜索网上有关医药的信息和知识，其效果和查准率比传统 HITS 和 IR 方法有一定提高. 同时结果表明基于 Web 文档内容和链接点内容的方法没有太大差异.

References:

1 Marti Hearst, Next Generation Web Search: Setting our sites [J]. IEEE Data Engineering Bulletin, 2000,23(3): 38～48.
2 Jon M. Kleinberg, Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999,46(5): 604～632.
3 S. Brin and L. Page, The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998,30(1-7): 107～117.
4 Jeffrey Dean, Monika R. Henzinger, Finding related pages in the World Wide Web[J]. Computer Networks, 1999,31(11-16): 1467～1479.
5 S. Lawrence and C. L. Giles, Searching the World Wide Web [J]. Science, 1998,280(4): 98～100.
6 Krishna Bharat, Monika R. Henzinger, Improved algorithms for topic distillation in hyperlinked environments[C]. Proc. of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 1998, 104～111.
7 Taher H. Haveliwala, Aristides Gionis, Dan Klein, Piotr Indyk, Evaluating strategies for similarity search on the web[C]. WWW 2002, May 2002,432～442.
8 Holger Billhardt, Daniel Borrajo, Victor Maojo. A context vector Model for information retrieval[J]. Journal of the American Society for Information Science and Technology, 2002,53(3): 236～249.
9 Monika R. Henzinger, Hyperlink analysis for the Web[J]. IEEE Internet Computing, 2001,5(1): 45～50.
10 Raymond Kosala, Hendrik Blockeel. Web mining research: a survey[J]. SIGKDD Explorations, July 2000,2(1): 1～15.

2004 年全国开放式分布与并行计算学术会议(DPCS2004)
征 文 通 知

中国计算机学会开放系统专业委员会主办、北京航空航天大学计算机学院承办、北京计算机学会协办的《2004 年全国开放式分布与并行计算学术会议》将于 2004 年 10 月 15 日在北京召开，有关信息如下：

一、征文范围(包括下列选题及相关内容)：

1. 开放式分布与并行计算模型、算法与体系结构；
2. 下一代开放式网络、数据通信、网络与信息安全、业务管理技术；
3. 开放式海量数据存储与 Internet 索引技术，分布与并行数据库及数据/Web 挖掘技术；
4. 开放式集群计算、网格计算、Web 服务、P2P 网络及中间件技术；
5. 开放式移动计算、自组网与移动代理技术；
6. 分布式人工智能、多代理与决策支持技术；
7. 开放式虚拟现实技术与分布式仿真；
8. 开放式多媒体技术(包括媒体压缩、内容分送、缓存代理、服务发现与管理技术)。

二、征文要求：

1. 论文必须是未正式发表的、或者未正式等待刊发的研究成果；
2. 论文格式仿照《计算机研究与发展》刊物的格式，应包括题目、摘要、关键词、正文和参考文献；
3. 论文中、英文均可，一般不超过 5000 字，一律用 Word2002 格式排版，提供 A4 激光打印稿一式两份，并随寄软盘或发送 Email 至 niujianwei@263.net；
4. 邮寄论文时，须在信封左下角或 Email 主题中注明《DPCS2004》；
5. 经程序委员会审查合格的论文，将收入论文集，在《北京航空航天大学学报》等核心期刊发表；
6. 论文请寄给北京航空航天大学联系人，论文自留底稿，恕不退稿。

三、重要日期与联系方式：

1. 论文须在 2004 年 7 月 15 日之前寄达(胡建平，牛建伟收)，录用通知将在 2004 年 7 月 25 日发出。
2. 联系方式：
   - 北京航空航天大学联系人：胡建平、牛建伟
     通讯地址：北京航空航天大学 601 信箱　邮编：100083
     联系电话：(010)82317601
     E-mail：niujianwei@263.net
   - 开放系统专委会联系人：陈炳从
     通讯地址：北京 619 信箱 63 号　邮编：100083
     联系电话：(010)62311951
3. 会议主页：http://ldmc.buaa.edu.cn/DPCS2004

中国计算机学会开放系统专委会

基于Hyperlink和相关度发现Web相关文档的研究

作者： 王天江， 叶卫国， 卢正鼎， 李永平
作者单位： 王天江,叶卫国,卢正鼎(华中科技大学,计算机学院,湖北,武汉,430074)， 李永平(国家药品监督管理局,北京,100810)

刊名： 小型微型计算机系统 ISTIC PKU
英文刊名： MINI-MICRO SYSTEMS
年，卷(期)： 2004,25(5)
被引用次数： 4次

参考文献(10条)

1. Raymond Kosala;Hendrik Blockeel Web mining research: a survey 2000(01)

2. Monika R Henzinger, Hyperlink analysis for the Web 2001(01)

3. Holger Billhardt;Daniel Borrajo;Victor Maojo A context vector Model for information retrieval[外文期刊] 2002(03)

4. Taher H Haveliwala;Aristides Gionis;Dan Klein;Piotr Indyk Evaluating strategies for similarity search on the web 2002

5. Krishna Bharat;Monika R Henzinger,Improved algorithms for topic distillation in hyperlinked environments 1998

6. S Lawrence;C.L. Giles Searching the World Wide Web[外文期刊] 1998(04)

7. Jeffrey Dean;Monika R Henzinger, Finding related pages in the World Wide Web 1999(11-16)

8. S Brin;L. Page The anatomy of a large-scale hypertextual Web search engine 1998(1-7)

9. Jon M Kleinberg, Authoritative sources in a hyperlinked environment 1999(05)

10. Marti Hearst Next Generation Web Search: Setting our sites 2000(03)

引证文献(4条)

1. 钱丽萍.汪立东 基于中心短语及权值的相似度计算[期刊论文]-郑州大学学报（理学版） 2007(2)

2. 王卫玲.刘培玉.刘克非 改进的Web链接主题提取算法[期刊论文]-计算机工程与设计 2007(2)

3. 梁敏 基于Web数据的距离函数研究[学位论文]硕士 2005

4. 林建方 Web页面链接文本信息抽取与分类的研究[学位论文]硕士 2005