中国科技资源导刊 ISSN 1674－1544 CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
2009 年 3 月 第 41 卷第 2 期 37－41 ISSN1674－1544 Vol. 41 No. 2 37－41,Mar. 2009

网络多媒体主题搜索策略研究

杨仁广 孟祥增
(山东师范大学传播学院,山东济南 250014)

摘 要:针对多媒体链接在网页中分布的特点,对 PageRank、Shark－Search 两种典型的主题搜索策略进行相关参数的改进,并从网页内容和网页链接的角度计算了多媒体链接与主题的相似度。实验结果表明,改进的 Shark－Search 多媒体主题搜索策略比改进后的 PageRank 搜索策略更能有效地提高多媒体主题搜索的效率,同时也更适合网络多媒体资源的主题搜索。

关键词:多媒体;主题搜索;主题搜索策略;网络蜘蛛
中图分类号: TP391 文献标识码: A DOI: 10.3772/j.issn.1674－1544.2009.02.007

1 多媒体主题搜索原理

主题搜索是以查询和检索某一专业领域或学科领域的因特网信息资源为目的。网络蜘蛛是主题搜索的核心,对主题搜索的效率起着举足轻重的作用。它通常从一个"种子集"(如用户查询、种子链接、种子页面)出发,通过 HTTP 协议请求并下载 Web 页面,分析页面并提取链接,根据一定的搜索策略计算提取链接的权重并存入待爬行队列中,网络蜘蛛从待爬行队列中选择链接,再以循环迭代的方式访问 Web[1－2]。

主题搜索策略水平的高低决定了网络蜘蛛性能的好坏,同时也决定了主题搜索的效率和准确率的大小。目前,在该领域很多专家、学者在理论和实践上做了大量的研究工作,提出了许多主题搜索算法,包括:以Shark－Search[3]和 Best－Fish[4]为代表的基于内容评价的搜索策略;以 PageRank[5]和 HITS[6]为代表的基于链接结构评价的搜索策略。文献[7]为了增强网络蜘蛛的自适应能力,将巩固学习策略在预测远期回报的优势加入到网络蜘蛛的学习过程中,用来预测待爬行链接未来回报价值。文献[8]在巩固学习的基础上,通过构建典型的"Web 语境图"策略来估计目标页面的距离,加强了网络蜘蛛的自适应和增量反馈能力。

多媒体主题搜索和主题搜索有相似之处:都是搜索网络中特定的主题。而多媒体主题搜索还要求搜索到的网页中包含特定主题的多媒体资源。我们基于多媒体在网页中分布的特点,对已有的主题搜索策略进行分析发现:基于"巩固学习"的主题搜索策略和基于"语境图"的主题搜索策略很难应用于多媒体主题搜索中,两种搜索策略都将搜索过程分为训练和搜索两个阶段,先根据训练构建网页分类器,再用于指导搜索阶段网页主题的判断。由于多媒体在网页中分布的复杂性和多样性,通过训练很难构建具有一定普适性的分类器。因此,本文将对传统的 PageRank、Shark－Search 两种典型的主题搜索策略进行相关参数的改进,以获得更适合网络多媒体资源的主题搜索策略。

2 多媒体主题搜索策略

基于内容的搜索策略和基于链接的搜索策略

第一作者简介:杨仁广(1982－),男,硕士,研究方向是 Web 信息检索、数据挖掘。
基金项目:山东省自然科学基金资助项目(Y2005G21)。
收稿日期:2008 年 11 月 19 日。

都属于基于立即回报价值的搜索策略，通过计算在搜索过程中“在线”获得的信息，如已访问页面中的文本信息、链接周围的文本信息、页面之间的结构信息等来确定下一步网络蜘蛛爬行的方向。改进后的两种主题搜索策略，其内容相似度的计算方法是一样的，不同的是链接相似度的计算方法。

2.1　内容相似度的计算

网页一般是由超文本标记语言 HTML 编写的，其所包含的多媒体信息可以通过分析此网页的 HTML 标记获得。在网络多媒体主题搜索过程中，我们提取以下信息用来表征多媒体的主题：网页的 URL；网页< title>　标签的文本内容；网页中多媒体链接的 Anchor（锚文本）；多媒体链接的 URL。通过对大量包含多媒体的网页分析我们还发现，包含多媒体的网页链接在其父网页中通常以链接列表的形式出现。我们将这些链接列表称为“主题团”[19]，将“主题团”中包含的锚文本称为“主题团”标题，它们对这些链接的主题起着指示性作用。为了对‘主题团’的标题进行提取，提出 4 个启发性规则，每个“主题团”被限制在一对 table 标签内，并对内部嵌套的 table 标签进行合并。规则是：①该文本的字号比周围文本的大；②该文本与周围文本的颜色不同；③该文本字数很少（一般少于 10 个）；④该文本独立成段。如果满足其中任意 2 个，则将其认定为‘主题团’标题，并以此作为表征多媒体主题的重要信息。在计算多媒体内容相似度的时候，我们把这个因素加入到计算过程中，如公式（1）：

$$Content\_score(u_i) = Score(block\_title)\ [\beta * Score(anchor) + (1-\beta) * Score(url)] \quad (1)$$

其中，$Score(block\_title)$ 是链接 u_i 所在‘主题团’标题与主题的相关度，计算时采用向量空间模型 VSM。在向量空间模型中，所有的检索关键词 t 形成关键词集合 $T=(t_1, t_2, \cdots, t_n)$，“主题团”标题文档 D 中的每一个文档 d 都被表示成一个范式的矢量，$V_t(d)=(t_1, w_1(d); \cdots; t_i, w_i(d); \cdots; t_n, w_n(d))$，其中 $w_i(d)$ 为 t_i 在文档 d 中的权重，权重的计算采用 TF－IDF 词频统计方法计算，如公式（2）。采用向量空间模型计算‘主题团’标题与主题的相关度，如公式（3）。$Score(anchor)$ 和 $Score(url)$ 分别表示链接 u_i 的锚文本和 URL 地址与主题的相关度，采用布尔模型进行计算；β 为相关因子，用以调节链接的锚文本和 URL 地址所占的比重。

$$w_i(d)=\frac{tf_i \lg(\frac{N}{nt_i}+0.01)}{\sqrt{\sum_{i=1}^{n}[tf_i \lg(\frac{N}{nt_i}+0.01)]}} \quad (2)$$

其中，tf_i 表示关键词 t_i 在文档 d 中出现的频率；N 表示用于特征提取的全部训练文本的文档总数；nt_i 表示出现关键词 t_i 的文档频率。

$$Score(block\_title)=sim(d,q)$$

$$=\cos(\theta)=\frac{\sum_{i=1}^{n}[w_i(d)*w_i(q)]}{\sqrt{\sum_{j=1}^{n}w_j^2(d)*\sum_{j=1}^{n}w_j^2(q)}} \quad (3)$$

其中，$w_i(q)$ 为关键词 t_i 在查询 q 中的权重，通常当查询中包含就为 1，否则就为 0。“主题团”标题与查询主题的相关度就表示为两个范化矢量之间夹角的余弦。

2.2　链接相似度的计算

对传统 PageRank 和 Shark－Search 两种主题搜索策略改进后，链接相似度的计算方法也有不同。

2.2.1　改进 PageRank 算法

PageRank 算法是一种随机漫游模型，决定一个网页重要性的主要因素是指向该网页的链接个数。PageRank 算法在迭代计算过程中，权值是按当前网页的出度平均分配，没有考虑到网页的相对重要性。但用户在实际的访问过程中，会根据链接与主题的相似度，有选择地访问网页。改进后的 PageRank 算法对待爬行队列中的页面进行排序时，把网页的链接信息相似度加入到 PageRank 计算公式中。该算法认为，用户在查资料时，点击网页内链接的机率是不相等的，和两个因素有关：链

接锚文本序列“主题团”的内容相似度和链接所指向网页的实际主题相关度（搜索过程中计算产生）。链接被点击的概率，同这两个因素成正比。这里给出改进后 PageRank 算法，如公式（4）：

$$Structure\_score(u_i) = \frac{1-d}{N} + d * \sum_{i=1}^{n} PR(T_i) * P(T_i, u_i) \quad (4)$$

其中，$Structure\_score(u_i)$ 代表网页 u_i 的 $PageRank$ 值；$PR(T_i)$ 代表网页 T_i 的 $PageRank$ 值，其中网页 T_i 指向网页 u_i；d 为阻尼系数；$P(T_i, u_i)$ 为从页面 T_i 到达页面 u_i 的概率，计算方法如公式（5）；N 为已下载到待爬行队列中与主题相关的网页数量；n 为链接到网页 u_i 的网页数量。

$$P(T_i, u_i) = \lambda * \frac{Score(block\_title)(u_i)}{\sum_{i=1}^{n} Score(block\_title)(i)} + (1-\lambda) * \frac{Sim_{link}(u_i)}{\sum_{i=1}^{n} Sim_{link}(i)} \quad (5)$$

其中，$\sum_{i=1}^{n} Score(block\_title)(i)$ 表示从网页 T_i 中链出的所有网页内容相度的集合；$\sum_{i=1}^{n} Sim_{link}$ 表示从网页 T_i 中链出的所有网页的实际主题相似度的集合；λ 是一个影响因子，取值范围为 0 ~ 1。

2.2.2 改进 Shark – Search 算法

Shark – Search 算法是在 Fish – Search 算法基础上发展而成的。Fish – Search 算法是 De Bra 等早期提出的一个主题网页动态爬行算法。Fish 算法将主题蜘蛛在 Web 中爬取网页的过程模拟为鱼群在大海中觅食的过程，当鱼找到食物时，鱼的繁殖能力就增强，反之鱼则逐渐消亡。Shark – Search 算法在 Fish – Search 算法的基础上做了两种主要的改进。首先，用一个连续的值函数来表示相关性，取值在 0 ~ 1 之间，而不是 Fish – Search 的二值判断；另外，待爬行链接的主题相关性受锚文本、锚文本上下文和父链接相关性继承的影响。文献[10]对网页中不同区块的链接进行聚类，然后将相同类的所有链接锚文本作为该类的描述文本，用来替代 Shark – Search 算法中锚文本上下文对链接相关性的影响。文献[11]从网页页面、链接块、链接本身 3 个粒度上对网页的相似性分别进行计算，然后将三者按照不同权重结合，进而确定整个网页的相似性。在改进的 Shark – Search 算法中，我们把网页按“主题团”进行网页分块，将“主题团”标题与主题的相似性加入到算法的计算中。

在链接结构方面，我们发现包含多媒体的主题网页表现出‘资源相邻性’的特点。所谓‘资源相邻性’是指在一个网站中所包含多媒体资源往往存在于这个网站的某一部分或某几部分区域中，并且处于同一区域的多媒体资源的主题也是相同的。根据此特点我们做出如下假设：

（1）如果一个网页是与主题相关的包含多媒体的网页，那么此网页的子链接很可能是与主题相关的包含多媒体的网页。

（2）如果一个网页是与主题相关的包含多媒体的网页，那么此网页在父网页中的兄弟链接很可能是与主题相关的包含多媒体的网页。

所以，在计算网页链接相关度时，我们用父网页和兄弟网页的链接相关度来揭示链接结构对一个 URL 链接相关度的影响。为了把这种影响实时地反馈给每个子链接，引入一个动态因子。用公式（6）来表示链接结构对一个 URL 链接相关度的贡献：

$$Structure\_score(u_i) = \sum_{j=1}^{t} \lambda(d_j) P(d_j) \Big/ t \quad (6)$$

其中，u_i 是正在爬行的链接，t 是父链接的总数，$\lambda(d_j)$ 是动态因子，用公式（7）进行计算；$P(d_j)$ 表示从父链接继承来的链接相关度和已爬行过兄弟链接的平均链接相关度，来衡量通过父链接能爬行到多少主题相关页面的能力，用公式（8）进行计算。

$$\lambda(d_j) = (n' + \theta) / (n + \theta) \quad (7)$$

其中，n' 是父链接 d_j 的已爬行子链接中主题相关页面的个数，n 表示父链接 d_j 已爬行子链接的总个数，θ 是归一化因子，通常取 0.5。在爬行的过程中，$\lambda(d_j)$ 会不断地进行动态调整。

表 1　通用搜索策略实验结果

种子个数	网页总数	有效网页个数	有效网页占有率	运行时间	平均爬行速度
10	73952	3059	4.464%	30.91 小时	40 个/分种

表 2　两种搜索策略实验结果比较

算法	种子个数	网页总数	有效网页个数	运行时间	平均爬行速度	查准率	查全率
改进的 PageRank 算法	10	21691	1643	24.51 小时	14 个/分钟	7.58%	53.71%
改进 Shark－Search 算法	10	15892	2048	15.27 小时	17 个/分钟	12.89%	66.95%

$$P(d_j)=(1-\sigma)R(d_j)+\sigma\sum_{k=1,d_k\in d_j}^{N}P(d_k)/N \quad (8)$$

其中，σ 是偏置因子，$R(d_j)$ 为父链接 d_j 的主题相关度，d_k 为 d_j 的一个已爬行子链接，N 为 d_j 已爬行子链接的总数，$\sum_{k=1,d_k\in d_j}^{N}P(d_k)/N$ 是父链接 d_j 中已爬行子链接的平均链接得分。

2.3　内容相似度和链接相似度的归一化

为了提高整个网页的主题相关性和权威性，对于改进的 PageRank 算法和 Shark－Search 算法都采用内容相似度和链接相似度按不同权值相加所得结果来标识。在这里将二者归一化，计算得到的值作为“主题蜘蛛”即将爬行链接的依据。计算公式为(9)：

$$S_{(i)}=\sigma * Content\_score(u_i)+(1-\sigma) * Structure\_score(u_i) \quad (9)$$

3　相关实验

人工选择 10 个物理教育网站作为种子（6 个包含嵌入式多媒体，4 个包含超链接式多媒体）。主题词集为物理词集(610 个词条)。先用通用搜索策略运行，结果如表 1 所示，然后分别用改进的 PageRank 搜索策略和改进的 Shark－Search 搜索策略运行，结果如表 2 所示。实验从算法的搜索效率出发对改进的两种搜索策略进行测试，结果如表 3 所示。

由表 1 和表 2 可以看出，改进的 PageRank 搜索策略和改进的 Shark－Search 搜索策略由于需要计算内容相似度和链接相似度，在平均速度上低于通用搜索策略。两种改进的策略在查准率和查全率方面比通用搜索策略都有较大的提高。表 2、表 3 显示改进的 PageRank 比改进的Shark－Search 在搜索的爬行速度、搜索精度和搜索效率上都有一定差距。根据以上分析和比较，我们认为，基于改进的 Shark－Search 的多媒体主题搜索策略，在围绕提高链接价值预测的准确性、降低计算的时空复杂性、扩大多媒体主题搜索的范围等方面进行改进后，比 PageRank 搜索策略更适合网络多媒体资源的主题搜索。

表 3　两种算法查准率随时间变化实验结果比较

查准率(%)		时间(小时)
改进 PageRank	改进 Shark－Search	
7.485%	6.527%	1
12.267%	16.463%	2
19.619%	21.493%	3
21.572%	25.691%	4

4　结　语

PageRank 和 Shark－Search 两种典型的主题搜索策略经改进，通过实验在多媒体资源中的搜索效率的比较分析表明，改进的 Shark－Search 搜索策略比改进的 PageRank 搜索策略更适合在多媒体主题搜索领域中应用。今后我们将做以下工作：①扩展基础教育主题词集，扩大搜索的范围，减少主题词集对搜索结果的影响，提高多媒体主题搜索的效率；②继续提高多媒体主题搜索策略的效率，着重优化实验算法中各个参数；③考虑对存在于多媒体网络的数据库（动态网页）中多媒体资源的获取。

参考文献

[1] Aggarwal C, AI_Garawi F, Yu S P. Intelligent Carwling on the World Wide Web with Arbitrary Predicates[C] // Proceeding of the 10th International World Wide Web Conference, 2001.

[2] Menczer F. Complementing Search Engines with Online Web mining agents[J]. Decision Support Systems, 2003, 35(2): 195 – 212.

[3] Bra D P, Houben G, Kornatzky, et al. . Information Retrieval in Distributed Hypertexts[C] // Proceeding of the 4th RIAO Conference, 1994.

[4] Cho J, Garcia – MolinaH, Page L. Efficient Crawling Through URL Ordering[J]. Computer Networks, 1998, 30(1 – 7): 161 – 172.

[5] 姜鑫维, 赵岳松. Topic PageRank——一种基于主题的搜索引擎[J]. 计算机技术与发展, 2005(5): 238 – 241.

[6] 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(7): 77 – 79.

[7] Rennie J, McCallum A. Using Reinforcement Learning to Spider the Web Efficiently[C] // Proceeding of the International Conference on Machine Learning (CML 99), 1999.

[8] Diligenti M, Coetzee F M, Lawrence S, et al. . Focused Carwling Using Context Graphs[C] // Proceeding of the International Conference on Very Large Database (VLDB '00), 2000: 527 – 534.

[9] 宋宇, 孟祥增. 基于改进 Fish – Search 算法的多媒体检索[J]. 计算机工程, 2008, 34(11): 189 – 193.

[10] 苏祺, 项锟, 孙斌. 基于链接聚类的 Shark – Search 算法[J]. 山东大学学报(理学版), 2006, 41(3): 1 – 4.

[11] 陈骏, 陈竹敏. 基于网页分块的 Shark – Search 算法[J]. 山东大学学报(理学版), 2007, 42(9): 62 – 66.

Research on Network Multimedia Topic Search Strategies

Yang Renguang, Meng Xiangzeng

(Department of Instructional Technology, Shandong Normal University, Jinan 250014)

Abstract: Based on the distribution characteristics of multimedia resources page, the two typical search strategies of PageRank and Shark – Search theme improve the relevant parameters, the improved two strategies compute the similarity of multimedia link with subjest from viewpoint of web pages content and link. The results show that the improved Shark Search theme of the multimedia search strategy than the improved PageRank search improves the efficiency of multimedia theme search, but also more suitable for the subject search of multimedia network resources.

Keywords: multimedia, topic search, topic searching strategy, web spider