文章编号:1671-9352(2007)09-0062-05

基于网页分块的 Shark-Search 算法

陈军[1],陈竹敏[2]

(1. 山东大学 网络中心,山东 济南 250100; 2. 山东大学 计算机科学与技术学院,山东 济南 250061)

摘要:Shark-Search 算法是一个经典的主题爬取算法. 针对该算法在爬取噪音链接较多的 Web 页面时性能并不理想的问题,提出了基于网页分块的 Shark-Search 算法,该算法从页面、块、链接的多种粒度来更加有效的进行链接的选择与过滤. 实验证明,改进的 Shark-Search 算法比传统的 Shark-Search 算法在查准率和信息量总和上有了质的提高.

关键词:Shark-Search 算法; 主题爬取; 页面分块; 相关性计算

中图分类号:TP391 **文献标志码**:A

Improved Shark-Search algorithm based on page segmentation

CHEN Jun[1], CHEN Zhu-min[2]

(1. Network Center, Shandong University, Jinan 250100, Shandong, China;
2. School of Computer Science and Technology, Shandong University, Jinan 250061, Shandong, China)

Abstract: A Shark-Search algorithm is one of the classical algorithms for focused crawling. However, its performance is not ideal for crawling Web pages which contain too many noisy links. An improved Shark-Search algorithm based on page segmentation was proposed, which can accurately evaluate the relevance from three granularities: page, block and single link. Several experiments were carried out to verify that the improved Shark-Search algorithm can obtain significantly higher efficiency than traditional ones.

Key words: Shark-Search algorithm; focused crawling; page segmentation; relevance computation

0 引言

Word Wide Web 的快速发展使得 Web 上的信息快速增长. 调查显示,我国信息资源的发展势头十分强劲. 截至到 2006 年 12 月 31 日,我国网民总人数为 13 700 万人,网页总数为 44.7 亿[1]. 用户上网的最主要目的便是获取信息,强劲的需求促进了信息检索技术的发展,同时也对信息检索技术提出了更高的要求. 传统的通用搜索引擎,如百度,Google 等,成为人们检索信息不可缺少的工具. 但是,随着网络的发展,通用性搜索引擎的局限性日益突出. 如:(1)不同领域、不同背景的用户往往具有不同的检索目的和需求,通用搜索引擎返回大量的且质量不高的结果. (2) 通用搜索引擎的目标是尽可能大的网络覆盖率,有限的搜索引擎服务器资源与无限的网络数据资源之间的矛盾将进一步加深. (3) 通用搜索引擎的更新周期较长,难以及时更新某些时新性较高的网页,如新闻等. 为了解决上述问题,主题爬取[2-8]成为一种潜在的解决方案.

Shark-Search 算法(OSSA, original Shark-Search algorithm)是一种经典的主题爬取算法[4]. 但出乎意料的是,从文献[6]中进行的实验可以看出,同 Best-First, InfoSpider[7], 甚至 Bread-First 相比,OSSA 的性能是较差的. 通过对实验过程进行深入分析,认为出现这种情况的主要原因是:当前Web页面包

收稿日期:2007-06-28
基金项目:国家科技支撑计划子课题资助项目(2006BAH02A29);山东省博士基金资助项目(2006BS01016)
作者简介:陈军(1967-),男,工程师,硕士,研究方向:信息检索、文本分类、数据中心管理. Email:jchen@sdu.edu.cn

含的链接较多，而其中噪音链接占很大一部分比重，且链接的锚文本不能很好的表明该链接指向的页面与主题的相关性．基于这些问题，本文提出了基于网页分块改进的 Shark-Search 算法（SSAPS，Shark-Search algorithm based on page segmentation，），该算法从页面、块、链接的多种粒度来进行更加有效的链接选择与过滤．实验证明，SSAPS 的性能比 OSSA 有了很大的改进．

1 相关工作

主题爬取就是遍历 Web 但只爬取与既定主题相关的网页．它有选择的访问万维网上的网页与相关的链接，尽可能多的爬取相关的网页，同时尽可能避免爬取不相关的网页．与通用搜索引擎不同，主题爬虫并不追求大的覆盖率，而将目标定为抓取与某一特定主题内容相关的网页，为面向主题的用户查询准备数据资源．OSSA 是一个经典的主题爬取算法，它是 Fish-Search[8] 算法的一个改进的变种．

Fish-Search 是第一个动态的主题爬取算法，将爬取过程比喻成鱼的觅食过程．它将 Web 比喻成 Pool，沿某个方向的遍历者比喻成 Fish，相关页面比喻成 Food，页面的出链比喻成 Fish 的 Children．如果 Fish 找到 Food 则产生较多的 Children；如果在一定的时间内找不到 Food，则要饿死．如果某个站点的速度较快，那么该站点的页面也会产生较多的 Children，否则产生较少的 Children．但该方法只基于页面内容与主题的相关性及链接的速度来确定待爬取 URL 的优先级，而且该方法判断页面与主题是否相关只是二元的，既相关/不相关．文献[9]给出了一个改进的 Fish-Search 算法．文中指出原先的 Fish-Search 算法的随机查找将导致重复爬取．不同的鱼应该在不同的方向上移动，移动的范围被看作不同的有向图．一个代表不同鱼的查找范围的“距离”参数用来控制鱼在不同有向子图的查找方向．“距离”被看作 2 个有向图的中心之间的距离，因此将问题转化为利用图论来求“距离”的问题．通过调整“距离”来避免了不同鱼之间的重复查找．OSSA 在 Fish-Search 的基础上主要提供了 2 种主要的改进．首先，使用一个连续的值函数来表示相关性，既取值在 0 到 1 的区间，而不是 Fish-Search 中的二进制相关性．另外，对待爬取队列中链接的潜在分数使用一个改进的方法来计算．链接分数受锚文本，锚文本上下文和祖先页面内容的影响．文献[5]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，用来计算该类与主题的相关性，用来代替 OSSA 中的链接上下文作为影响链接潜在分数的因素．

当前已有的主题爬取算法在计算相关性信息时，大都使用页面内容，锚文本及锚文本上下文．其中利用页面内容是从页面的粒度上考虑这种情况对该页面的所有链接都是平等的，显然会提高页面中许多噪音链接的优先级，而且当页面中包含大量的噪音信息时，不能准确的计算该页面与主题的相关性；锚文本及锚文本上下文是从单个链接的粒度上考虑在这种情况下，如果某个链接指向的页面是与主题相关的，但锚文本没有很好的体现出来，那么就会降低该相关链接的优先级．所以针对当前页面噪音较多且相关内容易于聚集成块的特点，本文提出了改进的基于页面，块和链接的不同粒度的 SSAPS．

2 页面分块及相关性计算

2.1 页面分块

随着 Web 的发展，当前页面上的噪音，广告链接越来越多．目前对 Web 进行分块方法的研究及基于块的信息检索方法的研究已有很多[10-12]，其中比较有效的是基于视觉的分块方法（VIPS，vision-based page segmentation）[11]．本文也是基于 VIPS 的方法进行分块，但具体分块的细节不属于本文的讨论范围．

本文对当前 Web 上的页面进行了深入分析，发现有 4 个主要的因素影响着主题爬取算法的性能：

（1）Web 页面往往有多个独立的块组成．

（2）页面包含的链接非常多，且链接一般是按块出现的．

（3）噪音链接非常多，且大部分噪音链指向的目标页面与当前页面几乎没有相关性．

（4）很大一部分出链的锚文本不能很好的表示目标页面的语义信息．

基于以上观察及主题爬取中相关性计算的需要，假设 Web 页面的内容主要有 2 种区域块组成：主题内容文本块 B_{T}，链接块 B_{L}．其中，链接块又分为相关链接块 $B_{\mathrm{L\_R}}$，导航链接块 $B_{\mathrm{L\_N}}$，噪音链接块 $B_{\mathrm{L\_A}}$．相关链接块链向的页面一般具有与当前页面相同的主题；导航链接块是为了方便读者浏览，指向的页面可能与当前页面的主题相同，也可

能不同;噪音链接块指向的页面与当前页面的主题根本没有关系,如广告链接等.

图 1 给出了一个经过分块的典型网页的例子.图中标出了块及对应的块的名称.

图 1　网页分块示例

Fig.1　Example of Web page segmentation

2.2　基于页面分块的相关性计算

本部分给出了在 Web 页面已经分好块的前提下,基于页面,块和链接的 3 种不同粒度的相关性计算.

首先给出一些符号的定义.预先指定主题 T(一般表示为关键词或自然语言描述的形式),当前处理的 URL,表示为 U_S,其指向的已经爬取的页面 P_S.U_S 中的一个出链 URL,表示为 U_T,其指向的未爬取的页面 P_T.主题的相关性计算就是根据 P_S 中的已知信息来预测 P_T 与主题 T 的相关性.

2.2.1　基于页面的相关性计算

基于页面的相关性计算是基于以下假设:Web 上相同主题的页面一般是靠近出现(存在链接关系),既主题的局部性.因此可以通过计算 P_S 与主题的相关性来推断 P_T 与主题的相关性,既 P_T 继承了其父页面 P_S 与主题的相关性.经过分块后,$P_S=\{B_T, B_{L\_R}, B_{L\_N}, B_{L\_A}\}$,其中一个页面可能存在零个或多个 B_T, B_{L_R}, B_{L_N}, B_{L_A},且有很多噪音信息.在 OSSA 中,使用整个页面来计算与主题的相关性,因此加入了太多的噪音信息,不能正确的计算相关性.本文使用主题内容文本块和相关链接块中的文字信息来计算该页面与主题的相关性.

$$R_P=\sum_{i=0}^{n}\frac{\boldsymbol{v}_{B_T^i}\cdot\boldsymbol{v}_T}{|\boldsymbol{v}_{B_T^i}|\times|\boldsymbol{v}_T|}+\sum_{j=0}^{m}\frac{\boldsymbol{v}_{B_{L\_R}^j}\cdot\boldsymbol{v}_T}{|\boldsymbol{v}_{B_{L\_R}^j}|\times|\boldsymbol{v}_T|}. \quad (1)$$

其中 n 和 m 分别表示页面 P_S 中包含的主题内容文本块和相关链接块的个数.$\boldsymbol{v}_{B_T^i}$, $\boldsymbol{v}_{B_{L\_R}^j}$ 和 $\boldsymbol{v}_T$ 分别表示 P_S 中第 i 个主题文本内容块 B_T^i,第 j 个相关链接块 $B_{L\_R}^j$ 的文本和主题 T 的向量空间模型(VSM, vector space model).

2.2.2　基于块的相关性计算

基于块的相关性主要是基于下面的假设:Web 页面中的链接都是成块出现的,相关链接块 B_{L_R} 中的链接指向的页面与当前页面一般具有相同的主题,而不管该链接的锚文本是否与主题相关;导航链接块中当前页面的上一级链接,表示为 U_P,一般是与该页面同一主题页面的索引页相关;而噪音链接块 B_{L_A} 中的链接指向的页面与主题没有任何关系.设 $U_T\in B_L$,则 P_T 与主题的相关性可以通过链接块的文字来计算.

$$R_B=\begin{cases}0, & U_T\in B_{L\_A}\\ \dfrac{\boldsymbol{v}_{B_{L\_R}}\cdot\boldsymbol{v}_T}{|\boldsymbol{v}_{B_{L\_R}}|\times|\boldsymbol{v}_T|}, & U_T\in B_{L\_R}\\ \gamma. & U_T\in B_{L\_N}, U_T=U_P\end{cases} \quad (2)$$

其中 γ 是预定义优先级因子.R_B 不但能提高相关链接块中那些锚文本不能体现出与主题的相关性的链接的优先级,而且能过滤掉噪音链接块中的大量噪音链接.

2.2.3　基于链接的相关性计算

基于链接的相关性计算是基于以下假设:如果页面 P_S 的作者在该页面中引用 U_T,表明对页面 P_T 的一种认可,U_T 的锚文本 A 是对页面 P_T 的内容的一种描述.那么,R_A 表示锚文本与主题的相关性,其计算公式如下:

$$R_A=\begin{cases}\dfrac{\boldsymbol{v}_A\cdot\boldsymbol{v}_T}{|\boldsymbol{v}_A|\times|\boldsymbol{v}_T|}, & U_T\in B_{L\_R}\\ 0. & U_T\notin B_{L\_R}\end{cases} \quad (3)$$

其中 $\boldsymbol{v}_A$ 表示锚文本 A 的向量.

2.2.4　相关性组合

为了更准确的计算相关性,本文给出了相关性的合成公式:

$$R=w_1*R_P+w_2*R_B+w_3*R_A. \quad (4)$$

其中 w_1, w_2 和 w_3 是权重因子,用来区别 R_P, R_B 和 R_A 的不同重要度.在本文中,令 $w_1=w_2=w_3=1$,既对 R_P, R_B 和 R_A 平等对待.

可以从链接选择与过滤的角度来理解公式(4):首先从页面级别上找出那些与主题相关性高的页面中的所有链接;因为链接是按照块的单元组织的,然后识别出与主题相关性高的链接块,给予其中的链接较高的优先级,并且过滤掉噪音链接块;最后

再从锚文本的角度来识别出相关链接块中锚文本能明显体现出指向页面与主题的关系的链接，并给它们较高的优先级. 从而从多种粒度上实现了对最优链接的选择与对噪音链接的过滤.

3 基于分块的 Shark-Search 算法

本部分给出了 SSAPS 的具体描述.

输入:起始种子 U_B，深度 D，需要爬取的页面总数 N，预定义主题 T

输出:与主题相关的页面集合 S_P

1 U_B.depth = D; $U_B \rightarrow L_U$; // L_U 待爬取的 URL 优先级队列

2 while L_U 不空且爬取页面总数 < N

3 取出 L_U 中 R 最高的 URL 作为当前处理的 URL U_S;

4 爬取 U_S 指向的页面 P_S; $P_S \rightarrow S_P$; 对 P_S 进行分块;

5 根据公式(1)计算 R_P;

6 for P_S 中的每一个出链 U_T do

7 if 没有爬取过 U_T then

8 分别根据公式(2),(3)和(4)计算 R_B, R_A 和 R;

9 if U_T 在 L_U 中不存在 then 插入 U_T 且对 L_U 按照 R 排序;

10 else 更新 L_U 中存在的 U_T 的 R 值 = Max{已存在 R, 新计算 R};

11 if $R > \delta$ then U_T.depth = D else U_T.depth = U_S.depth-1, 这里 δ 是一个预定义的阈值;

12 if U_T 在 L_U 中已存在 then U_T.depth = Max{已存在 depth, 新计算 depth};

13 end while;

14 return S_P.

4 实验

用户进行信息检索时，希望尽可能多的搜索到高质量，高相关的页面. 而主题爬取的目标就是在爬取的一定数量的页面中使得与主题相关的页面的数量最多，且质量最好. 这同用户的需求是一致的.

对主题爬取进行评价最常用的指标主要有相关页面总数(sum_of_relevantPage)和信息量总和(sum_of_info). 相关页面总数是已爬取的页面中相关页面的总数; 信息量总和是所有已爬取的页面的相关性之和. 因为在 OSSA 中对页面与主题的相关性的计算是以整个页面的内容，而不是以主要内容文本块为单位，因此为了统一评价的标准，在评价时，采用公式(5)来计算页面与主题的相关性.

$$R'_P = \frac{\boldsymbol{v}_d \cdot \boldsymbol{v}_T}{|\boldsymbol{v}_d| \times |\boldsymbol{v}_T|}. \tag{5}$$

其中 $\boldsymbol{v}_d$ 表示整个页面内容的向量空间模型.

相关页面总数以页面为原子单位来评价爬取结果的效果. 如果一个页面与主题的相关性 R'_P 大于给定的阈值，那么该页面是一个“相关页面”.

$$\text{相关页面总数} = P_1/P_2 \tag{6}$$

这里 P_2 表示已爬取页面集合 S_P 中的页面总数; P_1 表示 S_P 中相关页面的数目. 信息量总和是把爬取的所有页面作为一个整体来评价爬取结果的效果.定义一个页面的 R'_P 为该页面的信息量. 那么，S_P 的信息量总和为:

$$\text{信息量总和} = \sum_{\text{every page in } S_P} R'_P \tag{7}$$

将 SSAPS 与 OSSA 进行了比较，试验结果如表 1. 试验中每次爬取的网页总数为 6 000.

表 1 OSSA 与 SSAPS 的爬取结果比较

Table 1 Comparison of the crawling results of OSSA and SSAPS

主题	起始种子	相关页面总数			信息量总和		
		OSSA	SSAPS	增长率	OSSA	SSAPS	增长率
中国队 足球 亚洲杯	sports.sohu.com/asiancup2007 cn.sports.yahoo.com/foot/cn/asiancup07	1 446	1 937	1.34	11.57	18.01	1.56
证券 股票 股市	business.sohu.com/stock.shtml cn.biz.yahoo.com/stock.html	781	1 190	1.52	6.51	12.11	1.86

为了更好的分析算法的动态性能,本文把6 000个页面按每 600 个页面分成一个段，观察每段的相关的页面总数和信息量总和.

从表 1，图 2,图 3,图 4,图 5 得出的结论是一致的,即不管对于相关页面总数还是信息量总和，SSAPS 都优于 OSSA.

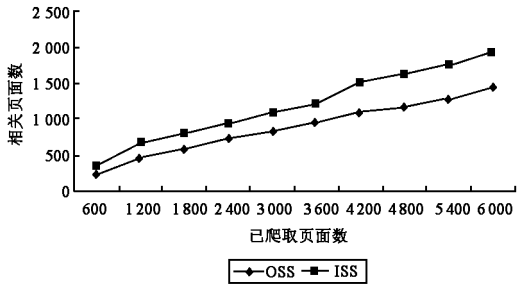

图 2　主题“中国队　足球　亚洲杯”的相关页面总数

Fig.2　The sum_of_relevantPage for the topic “Chinese Team, Football, Asian Cup”

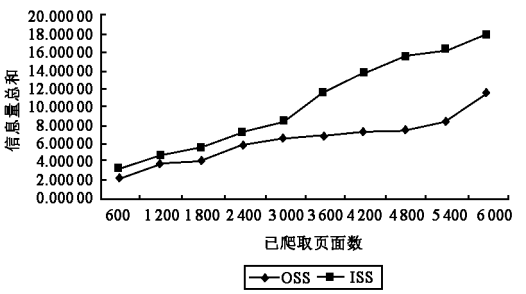

图 3　主题“中国队　足球　亚洲杯”的信息量总和

Fig.3　The sum_of_info for the topic “Chinese Team, Football, Asian Cup”

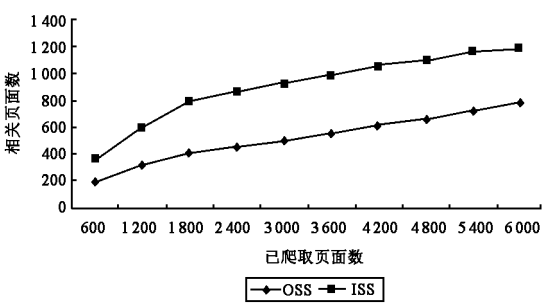

图 4　主题“证券　股票　股市”的相关页面总数

Fig.4　The sum_of_relevantPage for the topic “Securities, Stock, Stock Market”

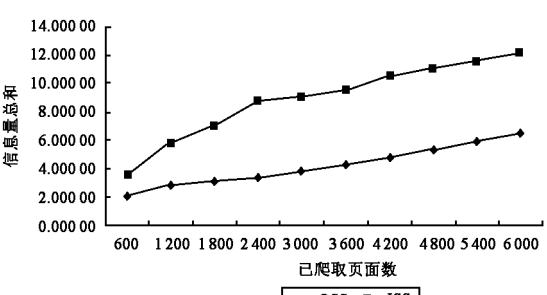

图 5　主题“证券　股票　股市”的信息量总和

Fig.5　The sum_of_info for the topic “Securities, Stock, Stock Market”

5　结语

本文分析了当前 Web 页面的特点，既页面内容及链接一般是成块出现，噪音链接较多，且大部分链接的锚文本不能很好的体现出它指向的页面与主题的关系，因此导致 OSSA 的性能不是很理想．针对这种情况，本文提出了基于分块改进的 Shark-Search 算法，该算法从页面，块和链接的 3 个粒度上来更加准确的实现对最优链接的选择与过滤．实验证明了该方法的有效性．在下一步工作中，我们将对主题的描述方法进行研究，使得主题描述更加满足用户的需求，更加个性化．

参考文献：

[1] 中国互联网信息中心．第 19 次中国互联网络发展状况统计报告[EB/OL].(2007-01)[2007-06-20]. http://www.cnnic.net.cn/index/0E/00/11/index.htm.

[2] 周立柱，林玲．聚焦爬虫技术研究综述[J]．计算机应用，2005，25(9):1965-1989.

[3] NOVAK B. A survey of focused web crawling algorithms[C]// Proceedings of SIKDD 2004 at Muticonference IS. Slovennia: ACM, 2004: 55-58.

[4] HERSOVICI M, JACOVI M, MAAREK Y, et al. The Shark-Search algorithm-an application: Tailored web site mapping [C]// Proceedings of the Seventh International World Wide Web Conference. Brisbane, Australia: Elsevier Science Publishers B V, 1998: 317-326.

[5] 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.

[6] MENCZER F, PANT G, SRINIVASAN P. Topical web crawlers: Evaluating adaptive algorithms[J]. ACM Transactions on Internet Technology, 2004, 4(4): 378-419.

[7] MENCZER F, PANT G, RUIZ M, et al. Evaluating topic-driven Web crawlers[C]// Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, USA: [s.n.], 2001: 241-249.

[8] BRA P De, HOUBEN G, KORNATZKY Y, et al. Information retrieval in distributed hypertexts[C]// Proceedings of the 4th RIAO Conference. New York:[s.n.], 1994: 481-491.

[9] LUO Fang-fang, CHEN Guolong, GUO Wenzhong. An improved “Fish-Search” algorithm for information retrieval[C]// Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering. [S.l.]: [s.n.], 2005: 523-528.

[10] 宋睿华，马少平，陈刚，等．一种提高中文搜索引擎检索质量的 HTML 解析方法[J]．中文信息学报，2003，17(4):19-26.

[11] CAI Deng, YU Shipeng, WEN Jirong, et al. VIPS: A vision-based page segmentation algorithm [EB/OL]. (2003-11-01) [2007-06-20]. http://www.ews.uiuc.edu/dengcai2/VIPS/VIPS_July-2004.pdf.

[12] LIN S H, HO J M. Discovering informative content blocks from web documents[C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (SIGKDD 2002). Alberta, Canada: ACM Press, 2002: 588-593.

(编辑:孙培芹)