改进的 Shark-search 算法在网络采集中的应用

张 玲,祁玉娟,姜 华
(湖南省第一师范学院 信息科学与工程学院,湖南 长沙 410205)

摘 要:Shark-search 是一种依据链接价值的高低进行优先采集的算法,用于主题信息采集系统时由于只考虑了网页文本和链接锚文本与主题的相关性而忽略了网页的组织结构特性,在抓取有较多噪音链接的网页时效果欠佳。基于网页组织结构特性的分析研究,提出了一种基于网页主题分块的 Shark-search 算法。该算法在经典 Shark-search 算法的基础上依据网页组织结构根据网页布局标签对页面内容进行分块,从网页,块和链接三个层面与主题的相关性得到链接的综合价值,因而具有自学习功能,能统计学习与主题相关性较大的块特征,并在发生主题漂移的时候具有自调整功能,给予主题相关性较大的父页面上的链接更多被抓取的机会。采集实验结果表明,所提出的算法在经典 Shark-search 的基础上能较好地改进主题信息采集的查准率,能够更灵活地针对实际的 Web 资源状况进行自调整。

关键词:Shark-search 算法;网页分块;Web 信息搜集;链接价值;主题漂移
中图分类号:G354 **文献标识码**:A **文章编号**:1673-629X(2017)08-0192-03
doi:10.3969/j.issn.1673-629X.2017.08.040

Application of Improved Shark-search Algorithm in Web Crawler

ZHANG Ling, QI Yu-juan, JIANG Hua
(College of Information Science and Engineering, Hunan First Normal University, Changsha 410205, China)

Abstract: The Shark-search algorithm ranks Web linkages based on their topic value, which only estimates the linkage's value by pages' text content and linkages' anchor text, not taking into account the link structure of the Web and has not good enough performance in crawling web pages including many linkages irrelevant to topic. An improved Shark-search algorithm based on topical segments has been proposed, which segments the Web page into blocks on the basis of the page's structure. The linkage's integrated value is comprised of the parent page's value, the block's value and the linkage's value. Moreover, it regards the visited out links as feedback to modify the block's relevance resulting with self-learning to statistical the characteristic of blocks. It has the ability of self-adjusting in the case of topic-drift to give more chance to the linkages in the web pages more relevant to the topic. The results of experiment in Web crawler show the algorithm proposed can well improve the precision of topical information acquisition on the basis of the classical Shark-search and more flexibly adjusts according to actual Web resources status.

Key words: Shark-search algorithm; web page blocking; Web crawler; linkages' value; topic-drfit

0 引 言

近年来,互联网技术发展迅速,搜索引擎的功能要求越来越丰富,传统的搜索引擎已不能满足人们的个性化服务需求[1]。因此,“专业搜索引擎”(Topic-Specific Search Engine)成为研究热点[2]。http://search.sz.gov.cn/was40/szgonline/szgov_advsearch.jsp 就是一个专用搜索引擎的例子。该网站以 200 多个深圳市政府相关网站作为信息源,利用网络蜘蛛抓取上面的网页,最大限度地过滤信息垃圾,返回给人们权威的专门信息。与此类似的还有企业和专业网站的搜索服务。专业搜索引擎搜索的内容只限于专门领域,追求更高的搜索效率和准确率[3],在专业搜索引擎中主题网络爬虫负责采集主题页面,它的研究主要围绕四个问题展开:

(1)怎样定义待搜索的主题?

(2)怎样决定待爬行 URL 的访问次序?

(3)怎样计算网页与主题的相关性?

(4)怎样尽快穿过与主题无关的网页,找到主题相关的页面?

主题网络爬虫一般根据链接文本内容或者 Web

收稿日期:2016-04-19 **修回日期**:2016-08-04 **网络出版时间**:2017-06-05
基金项目:湖南省教育科研基金(15C0284)
作者简介:张 玲(1979-),女(土家族),讲师,硕士,研究方向为机器学习、智能信息处理、搜索引擎。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20170605.1506.012.html
万方数据

超链图来判断链接的价值,决定待爬行链接的访问次序。判断链接与主题的相关性有多种方法,主要是根据主题信息与链接锚文本的“语义”相似度来决定链接价值的大小,相似度较大的被赋予较高的链接价值。利用 Web 超链图来判断链接价值的主要算法有 Back-Link、Hits、PageRank 等,主要是依据网页的相互引用关系来决定链接价值[4]。另外,还有利用机器学习预测链接未来价值,利用贝叶斯分类器和遗传算法来优化网络爬虫等多种方法[5]。

在经典 Shark-search 算法的基础上,将页面进行分块,由链接所在父页面、链接所在块和链接锚文本相关性共同决定待搜索链接的综合价值,以决定搜索的深度。该算法具有反馈机制,通过建立一个块标志库,将指向相关页面的块标志信息作为特征信息统计加入,用来辅助后续采集时的相关块的判断。在遇到主题漂移时,该算法会根据反馈自动调整各特征权值,重新计算队列中链接的价值,给主题相关性较大的父页面上的链接更多机会。

1　相关工作

Fish search 是经典的网络遍历算法,算法将进行网络遍历的网络采集者比做水里的鱼群。当采集者发现相关信息时,就产生更多的副本,寻找更多的相关信息;当没有相关信息时,它们就结束爬行。当一个网页被抓取后,抽取其上所有的链接,这些链接指向的网页称为孩子页面。如果抓取的网页是相关页,则孩子页面的抓取深度(depth)被设成一个预定义的值;否则孩子页面的抓取深度就被设置成一个小于其父亲网页深度的值。随着抓取的深入,当这个深度减为零时,顺着该方向的搜索就停止[6]。

Shark search 算法是 Fish search 的改进算法。在 Fish search 中,主题相关性的判断是二值的,而在 Shark 算法中相关性的判断更加量化,取值范围为 0 ~ 1[7]。Shark search 算法不但考虑了父页面与主题的相关值,还考虑了锚文字和锚文字的上下文等信息。Shark search 算法相对于 Fish search 能更好地保证搜索朝正确的方向进行,提高主题信息的发现率。

经过对很多专业主题的网页分析可知,虽然网页上有很多种类型的链接,但网页上的链接一般都是组织有序的,与页面同主题的链接被组织为一个一个的块,例如“相关文章”、“同被引文献”、“推荐阅读”等。这些块里面的有些链接即使本身的锚文本没有包含明显的主题信息,也依然有很大可能指向相关文档。网页上除了相关块以外还有其他一些特定用途的块,如导航块、广告块等,对主题搜索来说被称为“噪声”块。文中的意图是通过网页分块技术,区分出相关块和无关块,并建立一个相关块的特征库来指导以后的爬行。

目前实现 Web 页面分块比较常见的技术是 html 标签分类法和 DOM 树分类法[8],这两种分类法都是根据网页结构来进行分类。文中利用标签布局将页面中含有的链接进行分块,并将一个块中所有链接的锚文本组合起来判断块的相关性。当块中某个链接被访问后,根据它的实际主题相关性还可以反馈修改该块的相关性判断,从而修正该块中其他链接的搜索深度。此外,还建立了一个块特征库,学习具有相关性指向意义的块描述短语,例如“相关文章”、“相关文档”、“同被引文献”、“推荐阅读”等,用来辅导相关块的判断。

2　算法描述

2.1　网络蜘蛛模型

文中算法的网络蜘蛛首先对抓取页面进行页面相关性分析,再按照 Dom 树结构对网页进行分块[9],最后根据网页相关性、块相关性和链接相关性进行综合计算,得到链接的综合价值,决定链接的爬行深度[10]。如果链接价值大于某阈值,则爬行深度被设为一预定义的值,否则爬行深度被设为一小于其父亲网页爬行深度的值,直到爬行深度为零。

2.2　相关性计算

待采集链接的预测价值由三部分组成:父页面主题相关性、所在块的主题相关性和自身锚文本的主题相关性。

页面信息与主题的相似度,采用向量间夹角余弦公式[11]:

$$s'(q,p)=\frac{\sum_{k\in q\cap p} f_{kq}f_{kp}}{\sqrt{\sum_{k\in q} f_{kq}^2 \sum_{k\in p} f_{kp}^2}} \tag{1}$$

其中,q 为主题关键字项;p 为页面块文字项;f_{kd} 为项 k 在 d 中出现的频率。

一个页面被网络爬虫抓取后,即由式(1)计算出它与主题的相关性,若与主题相关性较大,则被判为主题相关页,进入索引库,同时根据它的相关性大小反馈修改与它同属一个块的兄弟链接的价值,并根据网页布局标签进行分块[12]。位于同一分块下面的所有链接锚文本组成本块的特征文本,也按式(1)进行相关性计算,得到块的相关性价值。链接本身锚文本则直接统计主题关键词项在锚文本中出现的频率。链接最终的价值由式(2)得到:

$$v=w_1*v_p+w_2*v_b+w_3*v_l \tag{2}$$

其中,v 为待搜索链接的价值;v_p 为链接所在的父页面相关性;v_b 为链接所在块的相关性;v_l 为根据链接锚文本计算得到的链接相关性;w_1、w_2、w_3 为权值,权值越大意味着对应信息对预测链接的相关性越重

万方数据

要,文中经过反复实验比较取 $w_1=0.3$, $w_2=0.3$, $w_3=0.4$。如果块标签上具有块特征库里的信息,诸如“推荐文章”、“相关文档”、“同被引文献”等文本,则 w_1 会被乘以大于 1 的系数,以增强块同父页面的相关性,而 w_3 则相应地减少权值。

2.3 反馈机制

由式(1)计算出的块的相关性价值只是根据块中的链接锚文本与主题的文本相似性预测出来的,文中在链接被访问得到实际的网页相关性后对其所在的链接块的价值进行修正,消除预测所引起的偏差,用于指导计算块中还未采集链接的价值。如果预测的链接所在的块价值较低但是指向的网页相关度较高,则适度增大块价值,反之,适度减少块价值。块价值得到修正后,块上的链接价值也根据式(2)重新计算。对与主题相关性很大的块,还建立了一个块特征库,学习具有相关性指向意义的块描述短语,用来辅助后续采集时的相关块的判断。

2.4 主题漂移情况的处理

网络爬虫根据式(2)计算得到的链接价值进行网页采集,链接价值越大的预测网页与主题的相关性就越大,网络爬虫进行优先采集。但是网络爬虫难免遇到主题漂移的情况,即被大量无关网页包围,网页上很难再采集到链接价值较大的链接,采集队列里充斥着大量价值较低的链接。文中在遇到主题漂移时,网络爬虫会根据反馈自动调整式(2)的各权值,重新计算队列中链接的价值,适当增加式(2)中 w_1 和 w_2 的权值,减少 w_3 的权值,即给主题相关性较大的父页面上的链接更多机会,即使此链接本身的锚文本价值并不高。

3 实 验

为了比较算法性能,文中比较了两种 Web 采集者。一种是传统的 Shark-search,一种是提出的综合了页面分块和反馈调整的 Shark-search。实验目的是采集主题“计算机软件”的页面,采用搜索精度来比较算法性能,公式如下[13]:

$$\text{搜索精度}=\frac{\text{结果集中的相关文档数}}{\text{结果集中总文档数}} \tag{3}$$

分别统计这两种算法的网络采集者抓取的页面数和采集的相关计算机软件网页数,并计算两者比例。搜索精度计算如下:

$$\eta_t=\frac{t\text{ 时刻已发现的计算机网页比例}}{t\text{ 时刻已搜索的网页比例}} \tag{4}$$

图 1 为搜索精度比较图。

由图 1 可以看出,改进的 Shark-search 在采集精度上整体优于传统的 Shark-search 算法,例如,当访问了 50% 的页面时,改进的 Shark-search 比传统的 Shark-search 多采集了 15% 的主题相关页面,具有更高的搜索效率,能够更灵活地针对实际的 Web 资源状况进行自调整。

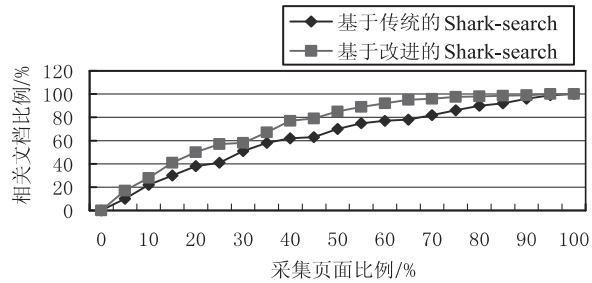

图 1 搜索精度比较图

4 结束语

为了更好地解决 Shark-search 算法在抓取网页时由于只考虑网页和链接文本与主题的相关性而忽略网页的组织结构造成的主题漂移问题,在分析了 Shark-search 算法和网页分块技术之后,提出了一种基于网页分块的 Shark-search 算法。该算法利用网页分块技术对网页进行分块,并计算链接所处块的主题相关性,链接的主题相关性由父页面价值、链接所在块价值和链接价值综合计算得到,既考虑了网页间的链接关系,也利用了网页的组织结构和主题信息,很大程度上提高了信息采集的查全率,同时也过滤了大量的噪音链接,提高了搜索效率。

在下一步的工作中,将研究采用新的网页分析技术更精确地分析网页文本信息与网页结构[14],以期进一步提高预测链接价值的精确性。

参考文献:

[1] 傅 欣. 第三代搜索引擎的智能化趋势研究[J]. 现代图书情报技术,2002(6):28-30.

[2] 张俊林. 这就是搜索引擎:核心技术详解[M]. 北京:电子工业出版社,2002.

[3] 张 博,蔡皖东. 面向主题的网络蜘蛛技术研究及系统实现[J]. 微电子学与计算机,2009,26(5):52-55.

[4] 李广丽. 基于网页内容评价和 Web 图的启发式垂直搜索策略的设计[J]. 情报理论与实践,2009,32(9):121-124.

[5] Angiulli F,Pizzuti C. Outliermining in large high-dimensional data sets[J]. IEEE Transactions on Knowledge and Data Engineering,2005,17(2):203-215.

[6] 罗方芳,陈国龙,郭文忠. 基于改进的 Fish-search 算法的信息检索研究[J]. 福州大学学报,2006,34(2):184-188.

[7] 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):139-143.

[8] 黄 歆,桑 楠. 基于 DOM 树和递归 X-Y 分割算法的

(下转第 199 页)

rad(±15)、放缩[0.5,51]倍,并对其进行全局平移,从而以较高的准确率检测到交通标志。在图(a)中可观察到,对于发生一定角度的旋转、放缩和在一定程度上被遮挡的交通标志,利用该算法可以准确检测出交通标志的位置。

CVBTM、SBTM 与 AT-SBTM 检测结果对比如表 1 和表 2 所示。

表 1　三角形交通标志检测结果

方法	三角形标志总数	检测出标志数	检测正确率/%
CVBTM	62	44	70
SBTM	62	47	75.8
AT-SBTM	62	53	85.5

表 2　圆形交通标志检测结果

方法	圆形标志总数	检测出标志数	检测准确率/%
CVBTM	568	385	67.8
SBTM	568	408	71.8
AT-SBTM	568	505	89

由表 1、2 可知,无论是检测三角形标志还是圆形标志,相比之下 AT-SBTM 的准确率都是最高的。因此,采用 AT-SBTM 可提高检测的准确率。

对于漏检和误检的圆形和三角形标志进行核查分析,主要原因如下:遮挡太严重和交通标志颜色严重褪色导致漏检;拍摄图片清晰度太差导致漏检的情况增多;交通信号灯的存在导致误检率升高。

4　结束语

在利用交通标志颜色特征的基础上,根据交通标志特有的形状特征创建模板,并对模板进行仿射变换,提出了 AT-SBTM 算法。该算法使原有的固定模板成为可变形模板,可变形模板可以更灵活准确地检测出目标。实验结果表明,该方法能有效提高交通标志检测的准确率。

参考文献:

[1] 李厚杰,邱天爽,宋海玉,等. 基于曲率尺度空间角点检测的交通标志分离算法[J]. 光学学报,2015,35(1):239-247.

[2] Sindha P D,Shah D M,Patel A. A color and shape based real time traffic sign detection and recognition system[J]. International Journal in IT & Engineering,2015,3(1):36-42.

[3] 汤智超,苏　琳,何　超,等. 导盲机器人的交通标志视觉识别技术研究[J]. 计算机技术与发展,2014,24(9):23-27.

[4] 陈兴华,万幼川,王晓华. 基于街景影像的交通标志识别[J]. 地理空间信息,2014,12(5):75-77.

[5] Wang G Y,Ren G H,Jiang L H,et al. Hole-based traffic sign detection method for traffic signs with red rim[J]. The Visual Computer,2014,30(5):539-551.

[6] 贾永红,胡志雄,周明婷,等. 自然场景下三角形交通标志的检测与识别[J]. 应用科学学报,2014,32(4):423-426.

[7] 金旭晖. 基于区域颜色分割的交通标志检测和识别[J]. 电气自动化,2016,38(3):14-16.

[8] 陈亦欣,叶　锋,肖　锋,等. 基于 HSV 空间和形状特征的交通标志检测识别研究[J]. 江汉大学学报:自然科学版,2016,44(2):119-125.

[9] Garcíagarrido M A,Ocaña M,Llorca D F,et al. Complete vision-based traffic sign recognition supported by an I2V communication system[J]. Sensors,2012,12(2):1148-1169.

[10] 江治国,陈小林. 基于特征匹配的交通标志识别算法[J]. 吉首大学学报:自然科学版,2013,34(1):28-32.

[11] 房泽平,段建民,郑榜贵. 基于特征颜色和 SNCC 的交通标志识别与跟踪[J]. 交通运输系统工程与信息,2014,14(1):47-52.

[12] 汤　凯,李实英,刘　娟,等. 基于多特征协同的交通标志检测[J]. 计算机工程,2015,41(3):211-217.

[13] Chen Y X,Xie Y,Wang Y L. Detection and recognition of traffic signs based on HSV vision model and shape features[J]. Journal of Computers,2013,8(5):1366-1370.

[14] 傅建安,万　文,熊震宇. 双能图像的亚像素匹配方法[J]. 南昌航空大学学报:自然科学版,2015,29(4):40-44.

[15] 潘铭星,孙　涵. 自然场景中道路交通标志形状的检测与校正[J]. 计算机与现代化,2016(2):5-10.

(上接第 194 页)

Zone 树模型[J]. 计算机工程,2009,35(5):53-55.

[9] 张瑞雪,宋明秋,公衍磊. 逆序解析 DOM 树及网页正文信息提取[J]. 计算机科学,2011,38(4):213-215.

[10] Brin S,Page L. The anatomy of a large-scale hypertextual Web search engine[C]//International conference on WWW. [s.l.]:[s.n.],1998:107-117.

[11] Srinivasan P,Pant G,Menczer F. Target seeking crawlers and their topical performance[C]//Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval. Tampere,Finland:ACM Press,2002:113-117.

[12] 黄文蓓,杨　静,顾君忠. 基于分块的网页正文信息提取算法研究[J]. 计算机应用,2007,27:24-26.

[13] Diligenti M,Coetzee F M,Lawrence S,et al. Focused crawling using context graphs[C]//Proceedings of the 26th international conference on very large databases. Cairo,Egypt:[s.n.],2000:527-534.

[14] 张　岭,马范援. 加速评估算法:一种提高 Web 结构挖掘质量的新方法[J]. 计算机研究与发展,2004,41(1):98-103.

万方数据