一种图像主题网络爬虫的实现方法研究

朱学芳, 韩占校

(南京大学 信息管理系, 江苏 南京 210093)

[摘要] 针对一种图像主题爬虫进行了设计研究, 采用了基于文字内容的启发式方法, 实现了借助图像文件的锚文本及其上下文进行主题相关性判定, 能更准确的抓取相关图像资源. 还对网页实现了主题相关性判定, 以便更有效地引导爬虫的爬行路经. 经实验证明, 本系统可起到一定的优化效果, 为实现定向主题的图像信息采集奠定了良好的基础.

[关键词] 链接锚文本链接上下文, 网络爬虫, JXTA, 主题爬虫

[中图分类号] TP 309 [文献标识码] A [文章编号] 1672-1292(2008)04-0115-03

Design and Implementation of a Web Crawler for Images

Zhu Xuefang Han Zhanxiao

(Department of Information Management, Nanjing University, Nanjing 210093, China)

Abstract An approach of a web crawler for images is designed and implemented in this paper. An elicitation method based on text content is adopted, and the determination of topic correlation is realized with the help of the anchor text of image files and their contexts, to snatch at resources of relevant images more accurately. The paper also carries out the determination of topic correlation of images so as to pilot more effectively the crawling path of the crawlers. Experiments prove that the system has a certain effect of optimization, and lays a good foundation of realizing the collection of image information of directional topics.

Key words anchor text; link-content Web crawler; JXTA; topical crawler

主题爬虫的功能是从 Web中收集关于一个特定主题的 Web页面, 是面向特殊化、领域化、主题化的网络资源的获取. 由于主题爬虫试图收集与预先给定主题相关的网页, 对于 Web上与主题无关的区域都不予访问, 所以能大大减少对网络信息的访问流量和文档下载量, 从而可以减少对存储资源的需求. 随着多媒体技术和计算机网络的飞速发展, 网络上的图像、视频数据呈几何级增长, 图像的传播和应用也越来越广泛, 建立高效的图像检索机制成为目前迫切需要解决的问题之一, 而建立高效的主题爬虫则成了其中的重要研究课题. 为了高效地抓取与主题相关的网络资源, 研究者提出了许多主题定制爬行策略和相关算法, 使得网络爬虫尽可能多地爬行主题相关的网页, 尽可能少地爬行无关网页, 并且确保网页的质量. 现有的主题爬虫实现方法主要有: (1)基于文字内容的启发式方法, 包括由 De Bra等人[1]提出的 Fish Search方法. 该类爬虫通常采用最相似优先的算法, 最先访问与主题相似度最高的页面, 另一种较为相似的是 Best First Search算法. (2)基于 Web超链图评价的方法. (3)基于分类器预测的方法, 它采用基于分类模型来描述用户感兴趣的主题和预测网页的主题相关度等[2]. Chakrabarti等人[3]还提出了分别基于两种不同的模型来计算网页主题相关性和 URL访问次序的方法. 计算网页主题相关性的模型可以是任意一种二值分类器, 而计算 URL访问次序的模型是通过包含父网页和子网页及其相关度的训练样本集合在线训练得到的. 实验结果表明, 爬行错误网页的数目大约减少了 30% ~ 90%. 2007年国内有研究者[4]提出了一种与上述各个爬虫都不相同的新的主题爬虫设计方法, 它利用图像本身的一些特征, 对颜色采用了基于颜色累加直方图的方法进行图像的特征提取与特征匹配, 对爬虫进行了优化, 改进爬虫的搜索策略, 提高了爬虫的搜索效率. 本文研究设计的主题爬虫是面向图像的网络爬虫, 该爬虫主要是获得与用户给出的相关主

收稿日期: 2008-06-18

通讯联系人: 朱学芳, 教授, 博士, 研究方向: 计算机图像/信号处理、模式识别、信息检索自动化理论与技术等. E-mail: xfzhu@nju.edu.cn

题的图像资源. 判定某幅图像是否与用户给定主题相关是本文所描述的爬虫系统的关键之一. 在系统中我们采用图像链接的锚文本及其上下文作为用于主题匹配的对象, 根据分析的结果以量化形式得出, 并与事先设定的量化指标比较, 如果达到量化指标, 则将该图像设置为可下载图像. 另外, 也对网页的链接上下文进行主题相关性的判定, 以便确定是否对该网页的超链进行提取, 作为新的 URL种子.

1 图像主题网络爬虫

主题爬虫的基本思路是按照预先确定的主题, 分析超链接和已经下载的网页内容, 来预测下一个要爬行的 URL, 尽可能保证多抓取与主题相关的网页, 一般要解决以下 3个关键问题: (1)一个已爬取的网页是否与主题相关的判断. (2) URL的访问次序的决定. 许多主题爬虫是根据已下载的网页的相关度, 按照一定的原则, 将相关度进行衰减, 分配给该网页中的超链接, 而后插入到优先级队列中. 不同主题爬虫之间的主要区别在于其决定 URL的爬行次序的方法. (3)主题爬虫的覆盖度的提高. 该问题要解决的是如何穿过质量不够好的(与主题不相关)网页得到真正的主题网页, 从而提高主题资源的覆盖度.

1.1 锚文本及上下文抓取

链接锚文本(Anchor Text)是在 Web页面当中可以点击的高亮度显示的文本, 它含有链接属性, 即 URL的链接标识, 当用户利用鼠标点击它们时, 浏览器就会打开或跳转到该 URL链接的网页. 在网页的源文件中, 链接锚文本是由链接标签(Anchor Tag, 又称锚标签)所包围的文字来表示的, 其 HTML代码形式如下: < a href= 页面 URL> 锚文本 < /a>.

借助链接锚文本给出的目标网页的主题信息摘要, 有不少研究人员利用页面链接结构进行了有效的 web搜索. 著名的 Google搜索引擎的创始人 Brin和 Page[5]就在他们的 Google中使用了锚文本对 URL建立索引. 锚文本是对目标网页的简练概括, 因内容过于短小而制约了发展. 这是因为相对于网页中其它元素, 锚文本本身所容纳的词汇数量过于稀少, 内容涵盖范围也相对有限. 当锚文本自身对网页主题预测毫无帮助的时候, 提取锚文本之外的背景信息, 就显得至关重要了.

链接上下文(Link-content). 在一个 Web页面中, 在超链接周围出现的文本统称为链接上下文. 提取链接上下文有多种方法, 可以提取锚文本周围一定数量的单词. 在本实验系统中, 锚文本上下文最多包含围绕链接的上下各 50个字符.

1.2 链接的主题相关性判定

主题爬虫在网页爬行过程中会不断地获取新的网页链接, 并判断该链接是否同主题相关, 在满足一定要求后, 将该链接放入以后待下载的 URL队列, 否则放弃该链接. 这样可以集中抓取主题相关的网页, 提高爬虫的召回率, 并可缩小爬虫的爬行范围. 根据 1.1节可以得到该链接的链接描述文档, 利用开源软件 Apache Lucene提供的中文分词器对其进行分词, 能够得到该链接的主题特征向量表示. 链接 $link$ 对应的 LDD($link$)的特征向量可表示为:

$$WL(link) = [WT(t1, link), \cdots, WT(tN, link)]$$

链接 $link$ 在网页 p 中的链接描述文档 LDD($link(p)$)对应的特征向量 $WL(link(p)) = [WT(t_1, link(p)), WT(t_2, link(p)), \cdots WT(t_n, link(p))]$.

接下来, 将该特征向量同用户给定主题进行余弦相似度分别计算出链接的相关度:

$$TopicCorrelativity(link(p)) = d(link(p), q),$$

并用 $LocalTopicCorrelativity(link(p))$表示链接 $link$ 在网页 p 的局部主题相关度, q是用户给定的主题, 系统中的具体实现借助于 Apache Lucene提供的相关库函数.

1.3 图像主题和网页主题相关性判定

利用针对图像链接的锚文本上下文相关度进行图像主题和网页主题相关性判定, 若得到的相关度达到用户设定的阈值时, 标识该图像链接为可下载, 否则将其放弃. 为了进一步提高采集主题页面的准确性, 剔除与主题不相关或相关度不高的资源, 对网页中 URL链接上下文也进行主题相关性判定, 并根据判定结果, 将与主题不相关或相关度不高的 URL链接去除, 保留合乎用户要求的 URL链接.

2 主题相关性判定

关于某一主题的内容由于存在集中性, 我们设定了爬虫的爬行域名范围. 为了取得大量数据, 以便对结果有比较更客观的评价, 我们设置了多个主题可以下载大量的数据来进行测试, 所进行的参数设置情况见表 1.

实验中, 分别对未启用和启用主题相关性判定功能进行了测试, 以便进行比较.

2 1 未启用主题相关性判定功能时的测试

本实验持续 140m in, 共成功下载 30 502个页面, 得到相关主题图像数 2 857幅.

数据统计结果见图 1.

从图 1中可以看出: 下载网页数成近似直线增长, 表明本系统在外界网络环境稳定的情况下, 系统的数据吞吐量比较平稳. 另外, 我们还可以从图中看出相关主题图像数一开始在发现的图像数中所占比例较高. 这是因为在初始化 URL种子时, 种子站点的选择比较具有针对性. 随着爬虫的运行, 出现新的站点的加入, 不相关站点的数目随之增多, 这些都使得下载的图像数目和相关主题图像数目在总处理的 URL数目中所占比例极大减少.

2 2 启用主题相关性判定功能时的测试

本实验持续 140m in, 共成功下载 35 873个页面, 得到相关主题图像数超过 10 000万幅. 数据统计结果见图 2 从图 2我们可以知道, 下载网页数呈近似直线增长, 但比图 1动作大一些. 这主要是两者测试时间不同, 外界网络状况有所变化, 但基本相当. 图 2显示, 得到的同主题相关的图片的数量和质量都有了改善.

表 1 单结点下针对图像的主题测试参数设置表

Table 1 Parameters with single node for image themes

参数名	参数值
下载线程数	200
同一服务器访问间隔	30 s
初始化 URL种子	根据百度网址选取 30个知名新闻网站以及一些政府网站
设置主题	地震、救助、四川
要求	有关这 3个主题之一的即可

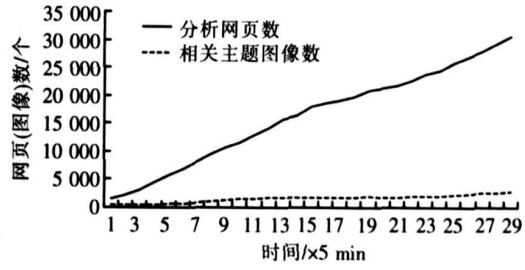

图 1 未启用主题相关判断

Fig.1 Relativity determination without using themes

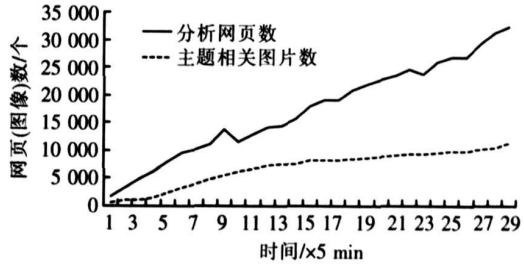

图 2 启用主题相关性判断

Fig.2 Relativity determination using themes

3 结语

本文详细介绍了设计和实现用于图像检索系统中的分布式网络爬虫, 探讨了其中的关键技术. 首先, 系统采用基于图像链接锚文本上下文的主题判定策略, 经实验数据证明使用该策略能获得较好抓取效果. 同时, 为了防止爬虫访问大量主题无关的网页, 采用类似的策略对网页的 URL链接上下文也进行判定, 减少了爬虫的无效负载. 文中所涉及到爬虫的爬取对象是图像文件, 其中主题可由用户根据需求进行设定. 经实验证明, 取得比较好的效果.

[参考文献] (References)

[1] De Bra P, Houben G, Kornatzky Y, et al Information retrieval in distributed hypertexts[C] //Proc of the 4th RIAO Conference New York, 1994: 481-491.

[2] 刘金红, 陆余良. 主题网络爬虫研究综述[J]. 计算机应用研究, 2007, 24(10): 26-29, 47.
Liu Jinhong Lu Yuliang. Survey on topic-focused Web crawler[J]. Application Research of Computers, 2007, 24(10): 26-29, 47. (in Chinese)

[3] Chakrabarti S, Punera K, Subramanyam M. Accelerated focused crawling through online relevance feedback[C]. Proc of the 11th International World Wide Web Conference Hawaii [s n.], 2002.

(下转第 166页)

[参考文献](References)

[1] 罗军辉, 罗勇江, 白义臣, 等. Matlab7.0在数字信号处理中的应用[M]. 北京: 机械工业出版社, 2005: 121-129.
Luo Junhui, Luo Yongjiang, Bai Yicheng et al. The Application of Matlab7.0 in the Digital Signal Processing [M]. Beijing: China Machine Press, 2005: 121-129. (in Chinese)
[2] 张贤达. 现代信号处理[M]. 2版. 北京: 清华大学出版社, 2002.
Zhang Xianda. Modern Signal Processing[M]. 2nd ed. Beijing: Tsinghua University Press, 2002. (in Chinese)
[3] Xilinx Co. The Programmable Logic Data Book[Z]. Xilinx Co, 2006.
[4] 郭成彬, 蒋危平. 认识数字超声探伤仪[J]. 无损检测, 2004, 26(3): 149-154.
Guo Chengbing, Jiang Weiping. Acquaintanceship of digital ultrasonic testing instrument[J]. Non-destructive Testing, 2004, 26(3): 149-154. (in Chinese)
[5] 常晓明. Verilog-HDL实践与应用系统设计[M]. 北京: 北京航空航天大学出版社, 2003.
Chang Xiaoming. Practice and Application System Design of Verilog-HDL[M]. Beijing: BUAA Press, 2003. (in Chinese)

[责任编辑: 顾晓天]

(上接第 117页)

[4] 张磊, 林坤辉, 周昌乐, 等. 基于图像内容检索的主题爬虫设计方法[J]. 广西师范大学学报: 自然科学版, 2007, 25(2): 182-185.
Zhang Lei, Lin Kunhui, Zhou Changle, et al. Design method of theme crawler of content based image retrieval[J]. Journal of Guangxi Normal University: Natural Science Edition, 2007, 25(2): 182-185. (in Chinese)
[5] Brin S, Page L. The anatomy of a large-scale hypertextual Web search Engine[C]. Proc the 7th World Wide Web Conference, [s.n.], 1998: 146-164.
[6] Lucene[EB/OL]. http://lucene.apache.org/, 2008.7.21.

[责任编辑: 孙德泉]