改进空间向量模型主题网络爬虫系统①

徐明子[1,2], 吕 立[2], 李喜旺[2]

[1](中国科学院研究生院, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)

摘 要: 详细阐述了主题网络爬虫实现的关键技术, 将传统的空间向量模型进行改进形成自适应的空间向量模型, 结合网页内容和链接两个方面进行网页相关度计算, 设计并实现了一个面向主题的网络爬虫系统. 针对主题网络爬虫爬行中出现的页面捕捉不全问题还提出了一种改进的手动与遗传因子相结合的网页搜索策略. 最后给出实验结果, 证明该系统的可行性及优越性.

关键词: 主题爬虫; 相关度计算; 搜索策略; 遗传因子

Topic-Focused Web Crawler System

XU Ming-Zi[1,2], LV Li[2], LI Xi-Wang[2]

[1](Graduate University, Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: This paper researched key techniques of topic-focused web crawler at first, then designed and implemented a crawler system by using improved slef-adapted vector space model. It analysised documents both in text and links. As the same time, this paper also comed up with a web search stategy based on gene factor combined with manully control. This strategy can solve the problem of searching path blocked. In the end, we provide some experiment results to prove the feasibility and advantages of our system from recall ratio and precision ratio.

Key words: topic-focused web crawler; relevance calculation; search strategy; gene factor

通用的搜索引擎在处理一般用户的搜索需求时具有较好的表现, 但针对日益增长的数量巨大的网页和个性化的搜索需求, 在网页实时更新、召回率和精准率等方面存在不足. 基于这一问题, 面向特定领域的主题搜索引擎应运而生并将成为搜索引擎发展的主要趋势之一[1].

主题爬虫的设计是主题搜索引擎实现的核心, 它与通用爬虫的不同之处在于尽可能快和多的采集与主题相关的网页. 由于通用爬虫并不注重页面采集的顺序和主题相关度, 这就会造成过多的系统资源和网络带宽的消耗; 主题爬虫则通过某一相关主题将整个 Web 信息分块采集, 最后将采集结果整合到一起, 为以后的索引、检索等提供充分的相关资源库. 本文还将对当前主题爬虫中存在的一些问题进行改进, 尤其是针对目前的主题爬虫在爬取网页时会出现与主题相关网页遗漏的现象, 还将提出一种基于遗传因子的搜索策略, 实现一个在搜索准确率和查全率方面都较好的主题爬虫系统.

1 主题爬虫原理与系统框架

主题爬虫与普通爬虫不同, 它并不追求覆盖信息的广度而注重挖掘某个领域的深度. 主题爬虫是按照特定的主题从 Web 上下载网页,然后对网页内容进行相关度分析, 过滤与主题无关的链接, 保留有用的链接并放入 URL 队列, 预测下一个待抓取的 URL, 以此不断重复以上过程. 它的目标是保证尽可能多的爬行、下载与主题相关的网页. 但主题爬虫又不是绝对的独立于普通爬虫, 实际上它是通过在普通爬虫的基础上添加某些模块来实现的.

① 收稿时间:2012-12-15;收到修改稿时间:2013-01-24

本文提出的主题网络爬虫的系统框架如图 1 所示. 该系统主要包括页面采集器、页面解析器和主题过滤器三个部分. 其中主题确立模块和初始 URL 模块是附加模块, 主要用来引导爬虫抓取的方向, 使得抓取的网页始终与主题相关. 页面采集器主要负责根据初始 URL 向各个线程分配抓取任务; 待页面采集模块抓获网页以后, 页面解析器对其从网页内容和链接两个方面做相关度分析.

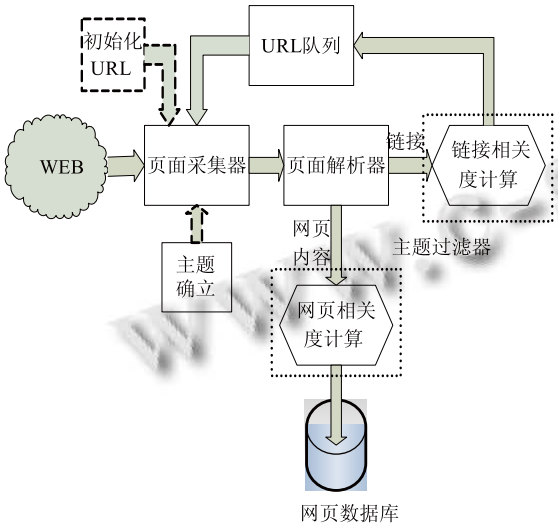

图 1 系统架构图

2 主题爬虫实现的关键技术

相对于普通爬虫, 主题爬虫主要解决以下几个问题:

(1) 如何描述要抓取的主题或目标;

(2) 如何判断抓取的网页或链接与主题是否相关, 这也是实现不同主题爬虫的关键点所在;

(3) 怎样确定等待 URL 队列中 URL 的优先级, 主题爬虫的爬行顺序已经不是简单的像普通爬虫那样进行广度优先或者深度优先爬行, 而是根据相关度的优先顺序进行爬行访问的;

(4) 采用何种网页搜索策略, 利用主题定制的搜索策略或算法, 使得爬虫能够更多地爬取高质量的、与主题相关的网页, 从而获得良好的网页覆盖率.

2.1 主题确立

主题的确立是主题爬虫工作的前提条件, 本文采用特征词提取法来完成抓取目标的描述, 并对不同的关键词给予不同的权值. 特征词的提取来源于两个部分, 一部分是由人工确定, 另一部分则通过初始 URL 获得. 人工确定的部分主要是通本咨询该主题领域的专家, 手动的确定一组特征词并为这些特征词确定相应的权值. 自动提取特征词是根据初始 URL 提供的训练样本利用算法得到一组能够代表主题的关键词, 并计算其权值,该权值可以采用改进的 TF-IDF 公式[2]来计算. 最后, 用人工确定的特征词指导由初始 URL 给出的训练样本得到一组更具个性化、准确率更高的特征词及权值.

因为传统的 TF-IDF 公式虽然考虑了词频和文档频率两个因素, 但是区分度在考虑主题分类方面并不占优势. 改进的公式考虑了相同的词在不同类别中的重要程度, 从而有区分的为他们分配权值. 特征词 k 在文档 i 中的权值 W_{ik} 计算公式如下:

$$W_{ik}=\frac{t_{ik}\times\log\left(\frac{N}{n_k}+0.1\right)\times W_{w_d c_j}}{\sqrt{\sum_{k\in i}\left[t_{ik}\times\log\left(\frac{N}{n_k}+0.1\right)\times W_{w_k c_j}\right]^2}}$$

其中, t_{ik} 是词 k 在文档 i 中的出现频率, N 是训练文本的总数, n_k 是训练样本中含有特征词 k 的文档数, $W_{w_k c_j}$ 是词语 w_k 关于类 c_j 的类别权重.

通过以上公式可以得到一个能够代表主题的基准文档向量: 向量的各个分量即为对应特征词的权值, 向量的维数即为特征词的个数.

2.2 相关度计算

为了确保抓取的网页以及保留的 URL 与主题具有高度的相关性, 需要对网页进行分析, 过滤与主题相关度较低(小于某一阈值)或无关的网页. 本文提出的方法从网页内容和网络拓扑两个方面完成网页的相关度计算.

2.2.1 基于内容的分析算法

基于网页内容的分析算法是指利用网页内容特征进行网页评价, 常用的方法是向量空间模型[3,4]. 传统的向量空间模型是通过初始的基准主题向量和给定的阈值来完成相关度计算, 最终完成网页的过滤. 这种方法简洁明了, 但是忽略了后来抓取内容对主题的反馈与指导作用, 因此本文在传统的向量空间模型基础上增加了自适应环节. 文本 i 和基准向量 f 的相似度计算方式如下:

$$Sim(i,f)=\frac{\sum_d W_{id} * f_d}{\sqrt{\left(\sum_d W_{id}^2\right)\left(\sum_d f_d^2\right)}}$$

其中，f_d 是主题基准向量 f 的分量.

在自适应阶段，根据后续的反馈信息，自动调整基准向量和阈值. 这种调整并不是每次完成一个网页的分析就进行的，而是间隔一定的时间.

提高阈值会使得抓取到的文本准确率更高，降低阈值则使得在网页主题相关度普遍较低的情况下具有更广的抓取范围. 设 S 是期望的在一个时间间隔内抓取到的文档数，T 是某一时间间隔，f_T 是在 T 时刻抓获的文档数，F_T 是到 T 时刻总共抓获的文档数，r_T 是在 T 时刻得到的主题相关的文档数，R_T 是到 T 时刻总共得到的主题相关的文档数. 本文提出的阈值(thd)调整算法是:

(1) 若 $f_T > S$ 且 $F_T > S * T$，则提高阈值 $thd = thd * \alpha$, $1.1 > \alpha > 1$;

(2) 若 $r_T < f_T * \mu$，$f_T > S$ 且 $F_T > S * T$，则需要大幅度提高阈值 $thd = thd * \beta$, $0.5 > \mu > 0.1$, $1.5 > \beta > 1.1$;

(3) 若 $f_T < S$ 且 $F_T < S * T$，则降低阈值 $thd = thd * \gamma$, $1 > \gamma > 0.9$.

基准向量的修改则是通过不间断的对抓获的文档进行分析，抽取出新的特征向量.

自适应向量空间模型的工作流程如图 2 所示.

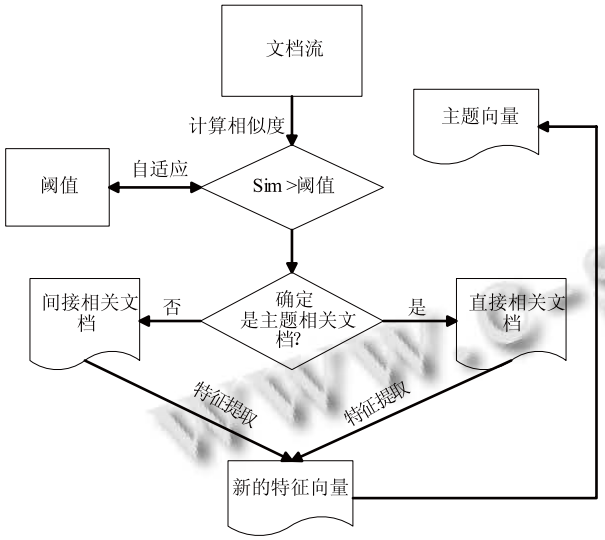

图 2　工作流程图

2.2.2 基于链接的分析算法

主题爬虫进行页面采集时，除了对页面内容进行处理以外还要提取出网页中的链出 URL 以保证爬虫可以循环往复的工作. 但是，由于链接的相关度相差很大，为了提高爬虫的抓取效率，需要对提取出的链接进行处理.

PageRank 算法是常用的计算链出网页重要程度的方法，它是一种与查询式无关[5]的算法. PageRank 算法的思想是一个网页的链入网页数量越多，说明他的重要程度越大；同时链入网页的质量对其重要程度也有影响. 网页 i 的权威度 $PR(i)$ 的计算公式如下:

$$PR(i)=(1-d)+d\times\sum_{j\in B(i)}\frac{PR(j)}{N(j)}$$

其中，d (0< d <1)是衰减因子，表示每个网页的权威度是(1- d)，因此实际每个网页用于传递的只是该网页权威度的 d 部分；$B(i)$ 是指向网页 i 的所有网页的集合；$N(i)$ 表示网页 i 中链出 URL 的数目.

因为每当爬取到新的网页都会抽取出其中包含的 URL，URL 队列中相应的优先级也需要重新计算，为了减少计算的工作量，在实际的爬虫工作时相对下载的网页进行缓冲，只有达到一定的阈值才会重新计算 URL 队列中 URL 的优先级.

2.3 爬虫抓取策略

主题信息往往只占整个 Web 中很小的一部分，所以按照传统的广度优先或深度优先进行搜索无论是在效率还是查全率上都难以达到期望要求. 通常的主题爬虫在沿着特定的方向爬行遇到道路堵塞时(即当前网页与主题不相关或小于阈值)，一般都会放弃当前的通道而另觅其他的爬行通道. 这样会导致堵塞通道上更深层的网页也会被一同丢弃，但是很多时候这些网页都是与主题相关的.

本文采用一种基于遗传因子的网页抓取策略可以有效的避免这种状况. 该算法工作过程如下:

(1) 为初始 URL 队列中的链接分配相同的主题相关程度值 Val. 因为初始 URL 队列中通过严格的筛选，他们与主题的相关度非常高，所以分配的 Val 值要比后续其他 URL 经过相关度计算得到的 Val 大. 另一方面，赋予初始 URL 较大的优先级值在以后的网页更新中依然可以被优先更新;

(2) 首先根据 Val 值大小对等待 URL 队列中的 URL 进行排序，然后再根据各个网页的相关度大小排序;

(3) 选取队列最前端的 URL 放入抓取队列，进行爬虫爬取工作;

(4) 下载网页到网页数据库并对其建立索引，将

URL 放入完成队列;

(5) 通过页面解析器对文本和文本中的链接进行分析, 并计算链接的权威度 val;

(6) 将 val 值与相关度阈值 f 进行比较: a.若链接相关度 val 大于相关度阈值 f, 即 val>f, 则将此链接放入等待队列, 上一个网页的 Val 值直接传递给此链接; b.若链接相关度 val 小于相关度阈值 f, 即 val<f, 则将上一个网页的 Val 值乘以遗传因子 r 传递给改链接, 因此该链接的 Val 值为 Val 父网页*r;

(7) 将解析出的 URL 及其相关度值 Val, val 放入等待队列, 重复(2);

(8) 算法结束.

算法利用初始相关度值 Val 和网页分析计算所得的相关度值 val 来保证爬虫始终在规定的主通道上进行爬行. 当通道阻塞时, 从主通道开辟一个次通道并在此通道上继续爬行, 从而避免了为了获得局部的最优忽略其他很多相关网页的问题.

3 实验结果和分析

为了验证该爬虫系统的可用性及合理性, 本文对该系统进行了实验检验并进一步对实验结果进行分析. 实验环境具体参数为: 酷睿 2 双核处理器, 2G 内存, 线程数为 200, 初始种子数为 20, 相关度阈值 f=0.15, 初始相关度值 Val 设为 100, 语言环境为 Java.

以电网为主题, 表 1 给出的是本文的爬虫系统与普通爬虫 Heritrix 的爬行结果对比. 可以看出普通爬虫比主题爬虫爬取的范围广, 但是他们的抓取时间相差并不大. 虽然主题爬虫还要对文档内容进行相关性分析计算, 但它会淘汰很多小于相关度阈值的网页, 而且随着搜索深度的增加, 主题爬虫和普通爬虫的抓取时间差距会越来越小.

表 1 本文爬虫与 Heritrix 爬取数据对比

爬虫名称	发现文档	下载文档		下载失败文档	总搜索时间（秒）	实际爬行时间（秒）
		主题相关文档	丢弃文档			
Heritrix	5618	1849	0	11	651	235
本文爬虫	3745	2106	1632	7	752	401

然后, 从查准率和查全率两个方面, 将本文的爬虫系统与基于文本内容的 best first search[6]算法及基于链接的 PageRank 算法的爬虫效率进行比较. 其中, 查准率=抓取主题相关文档数/发现文档数, 查全率=抓取主题相关文档数/主题相关文档总数, 实验结果图 3、图 4 所示.

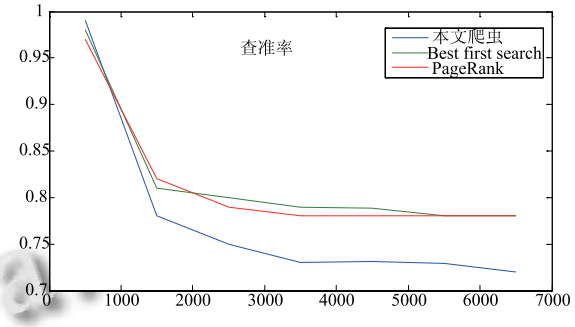

图 3 查准率对比图

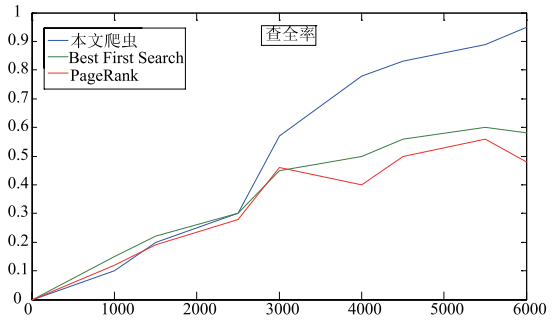

图 4 查全率对比图

从图 3 可以看出, 本文的爬虫系统相对于另外两种算法在查准率上接近平行相等, 另外两种爬虫算法查准率稍高; 但图 4 则显示, 在查全率上其有突出的优越性. 在文档数量较少时由于初始 URL 所开辟的主通道相关度都很高, 因此各个算法的查全率相差不大, 但在抓取后期本文的爬虫系统效率明显更高. 这是因为一方面采用自适应空间向量模型可以更准确的定位相关网页, 另一方面改进的搜索算法可以发现大量的被遗弃网页. 由此, 本文所设计实现的爬虫系统优势得以体现.

4 结语

通过对主题爬虫关键技术的讨论, 本文设计并实现了一个面向主题的爬虫系统. 该系统的创新之处在于对传统空间向量模型进行了改进, 添加自适应模块使爬虫始终不偏离主题方向; 另一方面从网页和链接两个方面对网页进行相关度分析使得爬行结果更加准

(下转第 52 页)

务平台，远程客户端可以利用网页浏览器通过服务器主机查看各个摄像头的实时及历史视频流. 服务器主机采用 LAMP(Linux+Apache+MySQL+PHP)构建，每一个组件都是免费或者开源软件，不需要为软件的发布支付任何许可证费就可以开发和应用基于 LAMP 的工程，放大了项目的衍生性. 为该系统的开发节约了成本.

网页服务器的设计中，使用 MySQL 数据库记录登录数据浏览系统的用户信息及视频文件位置信息，不同的用户提供不同的访问权限，如此提高系统的安全性，系统结合 PHP 向用户提供动态的网页浏览页面.

3 结语

本系统基于网络的监控系统实现了视频的分布式采集和集中处理. 利用网络摄像头采集数据，对图像进行人脸识别和步态识别，检测出异常人体，进而进行监视跟踪.

系统利用网络传输视频数据，结合 C/S 和 B/S 架构，使得系统具有很大的可扩展性，要监视某个区域只需在该区域安装网络摄像头，系统自动检测摄像头的加入，进而利用已经保存的人体数据模型进行匹配、识别. 用户查看各个摄像头只需连接到网页服务器，使得数据的查看和管理更为方便.

在人脸的检测方面，利用 Haar 分类器，具有较高的检测率. 在人脸图像识别中，运用 PCA 算法，提高了识别速度. 在步态识别方面，采用基于特征的的人体运动模型，利用一维距离降低了数据复杂程度.

参考文献

1 乔丽娟,许文力.浅析动态网站的交互性.科技信息(科学教研),2008,19:367,419.

2 Richardson Iain. h.264 和 mpeg-4 视频压缩:新一代多媒体的视频编码技术.长沙:国防科技大学出版社.

3 Lienhart R, Maydt J. An Extened Set of Haar-like Features for Rapid Object Dection, IEEE ICIP, 2002: 900–903.

4 Turk M, Pentland A. Eigenfaces for Recognition. Cognitive Neuroscience, 1991,3(1):71–86.

5 黄红梅,李广林,等.远距离人体步态识别算法.计算机工程,2007,21:100–105.

6 Hoffmann H. Tracking Faces in Grayscale Video Sequences with “Binary Direction Vectors”-Introduction and Evaluation of Binary Direction Vectors as a New Local Structural Feature for Tracking Faces with the OpenCV CAMSHIFT Tracker VDM Verlag Dr.Mueller e.K.2008.

7 刘瑞祯,于仕琪.Opencv 教程:基础篇.北京:北京航空航天大学出版社.

8 Stevens WR, Rago SA. UNIX 高级编程.第 2 版.北京:人民邮电出版社.

9 Stevens WR, Fenner B, Rudoff AM.UNIX 网络编程(第一卷):套接口 API.北京:清华大学出版社.

10 Prinz PG, Grawford T. C 语言核心技术.北京:机械工业出版社.

(上接第 39 页)

确；最后，针对目前主题爬虫中存在的网页捕捉不全问题提出了基于遗传因子网页搜索策略，大大提高了相关网页的覆盖率.

参考文献

1 刘金红,陆余良.主题网络爬虫研究综述.计算机应用研究,2009,24.

2 罗欣,夏德麟,晏蒲柳.基于词频差异的特征选取及改进的 TF-IDF 公式.计算机应用,2005,25(9).

3 Salton G. Developments in automatic text retrieval.Science, 1991,253(5023):974–979.

4 吴立德,黄萱青.大规模文本处理.上海:复旦大学出版社.1997,102–118.

5 Karch S, Heilig L. SAP NetWeaver Roadmap. America: SAP Press, 2005.

6 Rin S, PAGE L.The anatomy of a large-scale hypertexual Web search engine. Proc. of the 7th World Wide Web Conference.

7 Menczer F, Pant C, Srinivasan P. Topic-driven crawlers: machine learning issues. [2002-05-15]. http:// dollar.biz.uiowa.edu/～fil/papers.html