基于链接回溯的地理信息更新主题爬虫研究

吴家皋[1,2],余 浩[1,2],张雪英[3]
(1. 南京邮电大学 计算机学院,江苏 南京 210003;
2. 江苏省无线传感网高技术研究重点实验室,江苏 南京 210003;
3. 南京师范大学 虚拟地理环境教育部重点实验室,江苏 南京 210023)

摘 要:互联网的崛起为地理信息更新检索提供了一条新的途径,具有实时性强、成本低的优势。文中从实际出发,针对现有爬虫算法的缺陷,提出一种基于链接回溯的地理信息更新主题爬虫方法。首先,结合支持向量机分类技术,能够快速有效地找出一个网站中最有可能包含主题相关内容的链接方向;然后,回溯到这些链接后继续进行爬取,并通过地理信息变化要素知识库确定主题内容,从而优化爬取路径,减少低效率的爬取过程。实验结果表明,该方法可以找出最有可能包含地理信息的链接方向,大幅提高主题爬取效率,在其他主题方向也具有一定的可推广性。

关键词:主题爬虫;地理信息更新;支持向量机;回溯算法
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2014)07-0052-04
doi:10.3969/j.issn.1673-629X.2014.07.013

Study of Topic-driven Web Crawler for Geographic Information Updating Based on Link Backtracking

WU Jia-gao[1,2], YU Hao[1,2], ZHANG Xue-ying[3]
(1. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210003, China;
3. Key Laboratory of Virtual Geographic Environment of Ministry of Education, Nanjing Normal University, Nanjing 210023, China)

Abstract: The rise of Internet makes it a new way to search for information about geographic information updating, which has advantages of low cost and strong real-time. In allusion to the insufficiency of current top-driven web crawler, a new web crawler based on link backtracking algorithm is proposed in view of practice. First, it can find out the link paths in a website which most probably lead to topic information by using support vector machine classification; then, backtrack to these links and restart crawling, the theme of every links will be confirmed by using geographic information changing factor knowledge base, as a result, it will optimize crawling path and reduce low efficient crawling process. According to results from experiments, it can find out paths which lead to wanted information and enhance efficiency of crawling process, and also has a good possibility to extend to other topic areas.

Key words: topic-driven web crawler; geographic information updating; support vector machine; backtracking algorithm

0 引 言

随着中国城市的迅速发展,地理信息更新出现严重滞后现象。目前,检测地理信息变化的方法主要是通过高科技卫星遥感影像并进行人工识别,或者组织测绘小组进行实地数据采集,但无论哪种方法都需要消耗大量的人力物力。互联网作为 21 世纪新兴的信息载体,在数据现势性和丰富性方面具有显著优势,利用主题网络爬虫和信息抽取技术使得从非结构化网络文本中解析地理信息及其变化成为可能,具有周期短、成本低等特点,不仅具备较好的技术可行性,而且符合地理信息“应需适时更新”的要求,与现有更新技术手段形成优势互补。

收稿日期:2013-10-08 **修回日期**:2014-01-16 **网络出版时间**:2014-04-24
基金项目:国家测绘科技项目;江苏省自然科学基金(BK2012833);江苏省高校自然科学基金(12KJB520011)
作者简介:吴家皋(1969-),男,副教授,博士,CCF 会员,研究方向为计算机网络、移动计算、GIS 应用等。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140424.0839.087.html

主题网络爬虫[1-2]是专门进行某一特定领域或特定资源搜集的网络爬虫,相比于通用爬虫[3],主题爬虫最大的优势在于它只爬取与主题相关的链接,因此爬取策略和主题相关度计算是影响主题爬虫效率的最大因素。J. Cho 等人提出一种优先级队列的爬取策略[4],优先爬取主题相关度高的链接,由于容易陷入局部最优问题,文献[5]在此基础上进行了改进,但该方法需要锚文本完整并准确;从网络超链接结构图出发的算法有 PageRank[6-7] 和 HITS[8] 等,这些方法主要用于计算网页重要性而无法评估主题相关度,对此,文献[9]针对性地提出了一种结合 Shark-Search 算法和 HITS 算法优点的主题爬虫方法;S. Chakrabarti、皮靖等人提出了基于朴素贝叶斯分类模型的主题网络爬虫[10-11],Peng Tao 等人提出了一种基于 SVM(Support Vector Machine)分类模型的方法来引导主题爬取[12],将通过样本学习的分类方法引入了爬虫算法领域。在网页文本抽取[13]方面,文献[14]在网页树形结构基础上对标签进行映射,并采用自动训练的方法设计出一种快速有效的文本抽取方法。

结合现有主题爬虫算法的优缺点,文中提出一种基于链接回溯的爬取方法,该方法针对指定网站的地理信息更新检索,充分利用国内主流新闻网站的链接结构特点,结合支持向量机分类技术找出最有可能包含主题相关信息的链接方向,回溯到这些链接后继续爬取,并使用地理信息变化要素知识库确定主题内容,从而避免低效率的爬取过程,提高总体资源检索效率。实验证明,该方法可以找出一个网站中最有可能包含地理信息更新内容的链接方向,大幅提高了主题爬取效率,在其他主题方向也具有一定的可推广性。

1　系统结构

文中提出基于链接回溯的地理信息更新主题爬虫方法,以广度优先通用爬虫方法为基础,针对现有主题爬虫算法在实际应用中的缺陷,引入回溯的思想,根据新闻网站的结构特点,计算出最有可能包含主题相关信息的链接方向,从而大幅提高爬取效率,获取更多与主题相关的信息。系统整体流程分为两个阶段,如图 1 所示。

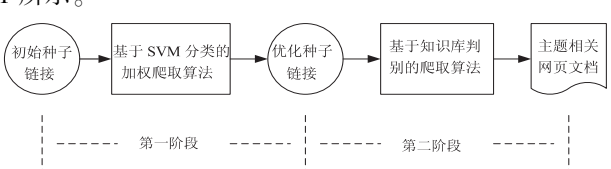

图 1　系统流程图

第一阶段以初始种子链接为基础,使用基于 SVM 分类的加权爬取算法,先指定某一层的链接为父链接组,然后以广度优先的方式进行网页爬取,在这个过程中使用 SVM 分类模型,逐条验证链接信息,找出分类准确的链接在之前指定层数的父链接,令该父链接权值增加,整个过程完成后,统计权值较高的父链接作为优化种子链接;第二阶段以优化种子链接为基础,利用基于知识库判别的爬取算法以广度优先的方式进行网页爬取,使用知识库逐条验证链接信息,判断并确定地理信息更新内容的主题,最后将优化种子链接和主题相关网页文档存入数据库,作为今后爬取的经验参考。

2　算法介绍

2.1　基于 SVM 分类的加权爬取算法

2.1.1　SVM 分类模型

支持向量机是目前分类性能最好的模型之一,文中采用支持向量机进行事件类型判断。从地理信息变化要素知识库(详见 2.2.1 节)中选择几种最常见的特征词汇作为类型关键词,以这些特征词汇及其对应的典型要素为基础收集样本文档,使用支持向量机方法进行样本机器学习,以特征词汇和典型要素作为触发词,在不影响分类精度的情况下提高分类效率,最终通过机器学习得到一个分类模型。

当判断一条链接包含的消息是否属于主题相关的范畴时,由于标题往往是一个文档包含信息最好的总结,故先判断这条标题中是否含有之前选择的特征词汇之一,如果包含,则对这条链接的正文部分使用 SVM 分类模型进行分类,如果分类得出的结果与标题中包含的特征词汇一致,则证明了这条链接中确实包含该特征词汇所代表的主题相关内容。这种基于机器学习和触发词的分类方法相比于单纯的字符匹配,可以避免字符一致但语意出现歧义的现象,并且有较高的准确率,是一种快速有效的主题相关度判别方法。

系统第一阶段是对网站所包含主题信息位置的一种试探,由于 SVM 分类性能有限,只能判断一篇文本是否属于某一特征词汇所代表地理信息范畴,无法最终确定以特征词汇和对应典型要素作为主题的内容,所以只选择几种最常见的特征词汇作为分类关键词,作为是否对父链接加权的依据。

2.1.2　算法描述

爬取过程采用广度优先的爬取策略,处理中的链接分为两个队列:待爬取队列(Uncrawled)和已爬取队列(Crawled);首先将种子链接集(Seed)加入待爬取队列,然后解析待爬取队列中链接的源代码,获得下一层超链接组,对下一层超链接组进行去重并去除已爬取过的链接,接着将待爬取队列加入已爬取队列中,下一层超链接组加入待爬取队列中,最后再次解析待爬取队列,重复上述过程,直至达到指定条件。

广度优先爬虫从种子链接出发,以层数为单位进

行爬取。加权算法思想是将第 S 层链接指定为父链接组,其中每条链接初始权值为 0,F 为最大爬取层数;在网页爬取的过程中,对第 $S+1$ 层至第 F 层的链接调用 2.1.1 节中提到的 SVM 分类模型进行验证,如果分类结果正确,则证明这条链接包含一定程度的主题相关信息,那么其父链接所指向的方向,就有可能包含更多与主题相关的信息,所以找到这条链接在 S 层的父链接,令其权值加 1;爬取过程结束后,统计第 S 层所有父链接的权值,选择权值最大的 K 条链接作为优化种子链接。这些链接相比于初始种子链接,指向主题相关内容的可能性更大,从而提高了整体爬取效率和准确性。S、F、K 的取值可以根据实际网站规模和结构进行调整。算法的伪代码如算法 1 所示。

算法 1:基于 SVM 分类的加权爬取算法。

```
输入:第 S 层的种子链接集 Seed;
输出:Seed 中权值最大的 K 个链接。
int[Seed.size] SWeight = 0  //第 S 层链接权重组,初值为 0
List Crawler(List Seed)
{
List Uncrawled, Crawled:={}
//Uncrawled 是待爬取队列,Crawled 是已爬取队列
Uncrawled:=seed
for i:=0 to F-S do
List temp = GetNextLayerURLs(Uncrawled)
//获取 Uncrawled 中链接的下一层链接
temp 去重并去除已爬取过的链接
Crawled:=Uncrawled∪Crawled
Uncrawled:=temp
end for
return GetTopKSeeds(SWeight,K) //返回 S 层权值最大的 K 个链接
}
List GetNextLayerURLs(List Uncrawled)
{
List temp:={}
for j:=0 to Uncrawled.size do
if SVM(Uncrawled[j])分类准确 then
int index = GetSLayerIndex(Uncrawled[j])
//返回 Uncrawled[j]在 S 层的父链接编号
SWeight[index]++
end if
temp:=temp∪GetURLs(Uncrawled[j])
//返回 Uncrawled[j]页面上所有超链接并加入 temp 中
end for
return temp
}
```

2.2 基于知识库判别的爬取算法

2.2.1 知识库判别方法

文中的应用方向是针对地理信息变化的检测,在参考《GBT13923-2006 基础地理信息要素分类与代码》中分类标准的基础上,对各类别的特征词汇和典型要素进行了总结,形成一个特征词汇对应多个典型要素的地理信息变化要素知识库;表达形式以特征词汇和典型要素两个关键词的组合来表示,例如:路+拓宽,路+通车,河+截流等,以此判断一个文本的内容是否属于地理信息范畴,并确定该文本的主题。具体步骤如下:

(1)将待处理的网络文本进行分句并编号;

(2)利用 ICTCLAS 分词软件对所有句子进行分词;

(3)从第一句开始,检查被标记为动词的词汇是否属于特征词汇集,如果匹配,则以该动词为中心,以词汇距离从近到远的顺序遍历所有被标记为名词的词汇,参照知识库中的特征词汇和典型要素匹配关系,如果配对成功,将该名词和动词作为组合抽取并记录;

(4)遍历所有句子,找出所有满足条件的组合。

其中,步骤(3)中提到的关键词匹配方法,过程举例如图 2 所示。需要处理的语句是“郑州彩虹桥隧道 5 月通车 将成北区新交通枢纽”,后缀为“/n”代表名词,“/v”代表动词,首先找到动词“通车/v”并与知识库中的典型要素进行匹配,成功后以“通车/v”为中心,以词汇距离从近到远的原则分别向左右两个方向寻找名词,向右找到名词“区/n”后,将“区+通车”与知识库进行匹配后失败,向左找到名词“隧道/n”后,将“隧道+通车”与知识库进行匹配后成功,因此停止寻找并抽取出“隧道+通车”的关键词组合作为这一句话的主题内容。

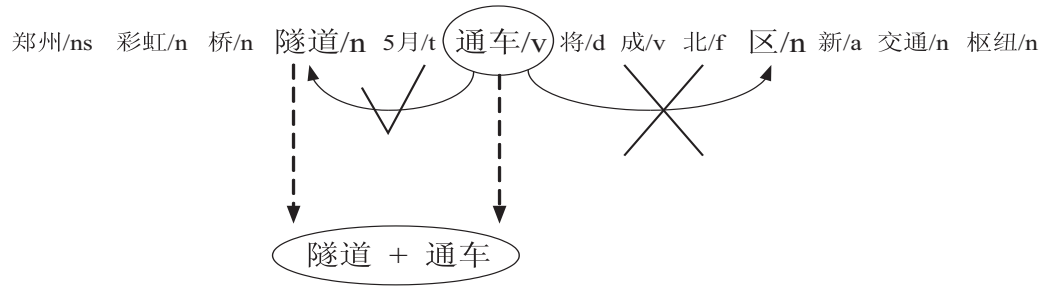

图 2 主题关键词确定示例图

2.2.2　算法描述

爬取算法依然采用广度优先的爬取策略,对爬取到的每条链接使用 2.2.1 节中的知识库判别方法,确定相关信息的主题内容,保存网页文档。

3　实验测试与性能分析

利用前文介绍的方法,使用适合网络编程的 Java 语言实现了基础地理信息更新检测原型系统,以新浪、网易和腾讯等主流新闻网站作为初始种子链接进行测试。分词工具使用中科院设计开发的 ICTCLAS 分词软件,机器学习使用目前最广泛应用的 LIBSVM 工具,由于实验设备和网络条件有限,回溯过程中设置初始层 $S=1$,爬取最大层数 $F=4$,回溯后取权值最大的 $K=2$ 条链接作为优化种子链接。

图 3 展示了两种爬虫在爬取相同数量(10 000 条)链接的情况下的整体效率,因为回溯之后再爬取是一个重复的过程,也就是为了计算出优化种子链接而付出的代价,所以比较两种爬虫的整体效率是为了检测这个重复过程对于整体效率的影响。从图 3 可以看出由于回溯过程,系统整体效率确实受到了一定的影响,但依旧好于通用的方法,并且在找出一个网站的优化种子链接后,今后对于该网站的爬取就可以直接使用优化种子链接,不再需要进行回溯的过程,所以从总体上来看,回溯方法造成的效率影响是可以接受的。

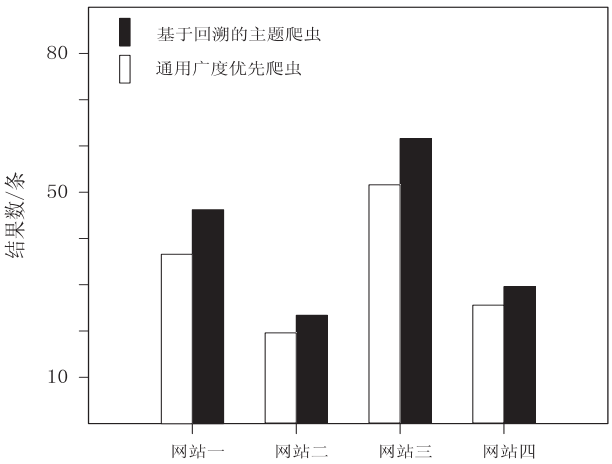

图 3　基于回溯的主题爬虫和通用广度优先爬虫性能比较图

同样在爬取相同数量(10 000 条)链接的情况下,正如之前所说的,如果不是第一次对某网站进行爬取,那么可以直接使用通过回溯后得到的优化种子链接为基础,如图 4 所示,这样找到的主题相关信息明显多于通用方法;结合图 3 和图 4 的实验结果可以证明通过文中提出的回溯方法,确实可以分析出一个网站中最有可能包含主题相关信息的链接方向,从而大大提高爬取效率,减少低效率的爬取过程。从实验结果可以看出该方法在各类新闻网站中都具有广泛的可用性,在其他主题方向也具有一定的推广性。

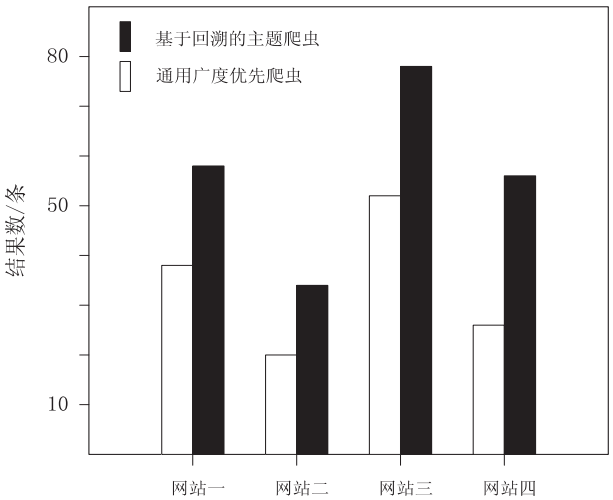

图 4　基于回溯结果的主题爬虫和通用广度优先爬虫比较图

4　结束语

主题网络爬虫能够针对用户需求,有效地进行特定主题的信息检索。文中在现有爬取策略和主题相关度算法的基础上,提出一种基于链接回溯的地理信息更新主题爬虫方法,通过实验证明了该方法确实可以提高爬取效率,设计并实现了基础地理信息更新检测原型系统,该系统适合于在新闻类网站中寻找地理信息更新的消息,后续研究可以考虑在其他主题方向的应用,以及减少回溯过程的代价来提升效率的方法。

参考文献:

[1] Zhong Shaojun, Deng Zhijuan. A web crawler system design based on distributed technology[J]. Journal of Networks, 2011,6(12):1682-1689.

[2] Batsakis S, Petrakis E G M, Milios E E. Improving the performance of focused web crawlers[J]. Data & Knowledge Engineering, 2009,68(10):1001-1013.

[3] 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.

[4] Cho J, Garcia-Molina H, Page L. Efficient crawling through URL ordering[J]. Computer Networks and ISBN Systems, 1998,30(1-7):161-172.

[5] 刘淑梅,夏　亮,许南山. 主题搜索引擎网络爬虫搜索策略的研究与实现[J]. 计算机系统应用,2010,19(3):49-52.

[6] Bourchtein A, Bourchtein L. On some analytical properties of a general PageRank algorithm[J]. Mathematical and Computer Modelling, 2013,57(9-10):2248-2256.

[7] 张　翔,周明全,李智杰,等. 基于 PageRank 与 Bagging 的主题爬虫研究[J]. 计算机工程与设计,2010,31(14):3309

(下转第 59 页)

和单步法利用下一位和下两位的信息减少了运算步骤,运算速度相对较快,但产生的中间变量太多,对存储信息量要求较高。相比之下采用三步法实现 MSD 加法。

3　结束语

三值光计算机的“进位直达”的加法进位思想,为进位设置专用通道,设法使所有的进位同步完成,称其为“进位直达”方式。依靠这一全新方式,不仅可以大大简化加法器的整体结构、缩短运算时间,而且使得运算时间定长,与数据位数无关。此设计方案为实现光学并行加法器开创了新途径,并大幅度简化了对加法器的管理难度,但目前仅有理论模型,专用进位通道的硬件尚未成型。

三值光学计算机的逻辑光学处理器是按照降值设计理论完成的,它能够完成 19 683 种二元三值逻辑运算。为了能够在该光学处理器上实现加法运算,进行了如下的研究探索:不能简单套用“先行进位”思想在光学计算机上实现加法,仿照电子计算机“分组先行进位”思想的折中方法对于三值光计算机也不行,要找出适合三值光计算机特点的数值表示和编码方式和能够发挥三值光计算机优势的算法。而 MSD 加法无需进位,只要通过多个逻辑变换就可以实现加法,正适合三值光学逻辑处理器的特点;同时利用光学处理器可以实现逻辑变换的并行处理,充分发挥处理器位数巨大的优势。

参考文献:

[1] 幸云辉,杨旭东. 计算机组成原理实用教程[M]. 北京:清华大学出版社,2001.

[2] 李育林,傅晓理. 空间光调制器及其应用[M]. 北京:国防工业出版社,1996.

[3] 金　翊,何华灿,艾丽蓉. 一种未来的计算机-三值光计算机[J]. 科技广场,2005(1):106-108.

[4] 金　翊,何华灿,艾丽蓉. 进位直达并行三值光计算机加法器原理[J]. 中国科学 E 辑 信息科学,2004,34(8):930-938.

[5] Avizienis A. Signed-digit number representations for fast parallel arithmetic [J]. IRE Trans on Electronic Computers, 1961,EC-10(3):389-400.

[6] 罗金平,陈书明,周兴铭. 基于符号替换逻辑的光学数字计算[J]. 计算机科学,1996,23(1):5-9.

[7] 李国强. 负二进制编码的光学阵列化复数运算[J]. 光学学报,1995,15(10):1419-1421.

[8] Drake B L,Bocker R P,Lasher M E,et al. Photonic computing using the modified signed digit number representation[J]. Optical Engineering,1986,25(1):38-43.

[9] Alam M S. Efficient binary signed-digit symbolic arithmetic [J]. Opt Lett,1994(19):353-355.

[10] Alam M S,Ahuja Y,Cherri A K,et al. Symmetrically recoded quaternary signed-digit arithmetic using a shared content-addressable memory[J]. Opt Eng,1996,35:1141-1149.

[11] Li Guoqiang,Liu Liren,Cheng Huiquan,et al. Simplified quaternary signed-digit arithmetic and its optical implementation [J]. Opt Commun,1998,137:389-396.

[12] 左开中,金　翊,严军勇. 三值光计算机的数值表示及其基本算法[J]. 计算机技术与发展,2007,17(9):8-10.

[13] Huang Hongxin,Itoh M,Yatagai T. Modified signed-digit arithmetic based on redundant bit representation[J]. Applied Optics,1994,33(26):6146-6156.

[14] Qian Feng,Li Guoqiang,Ruan Hao,et al. Two-step digit-set-restricted modified signed-digit addition-subtraction algorithm and its optoelectronic implementation [J]. Applied Optics, 1999,38(26):5621-5630.

[15] Barua S. Carry-free optical binary adders[C]//Proc of SPIE. [s. l.]:[s. n.],1990.

(上接第 55 页)

-3312.

[8] 吕林涛,陈丽萍,周红芳. 面向垂直搜索引擎的主题提取算法[J]. 计算机工程,2009,35(15):44-46.

[9] 罗林波,陈　绮,吴清秀. 基于 Shark-Search 和 Hits 算法的主题爬虫研究[J]. 计算机技术与发展,2010,20(11):76-79.

[10] Charkrabarti S,Dom B,Indyk P. Enhanced hypertext categorization using hyperlink[C]//Proc of the ACM SIGMOD international conference on management of data. [s. l.]:[s. n.], 1998:307-318.

[11] 皮　靖,邵雄凯,肖雅夫. 基于朴素贝叶斯算法的主题爬虫的研究[J]. 计算机与数字工程,2012,40(6):76-78.

[12] Peng Tao,Zou Wanli,He Fengling. SVM based adaptive learning method for text classification from positive and unlabeled documents[J]. Knowledge and Information Systems,2008,16(3):281-301.

[13] 张俊英,胡　侠,卜佳俊. 网页文本信息自动提取技术综述[J]. 计算机应用研究,2009,26(8):2827-2831.

[14] 陈治昂,周知予,李大学. 一种基于模板的快速网页文本自动抽取算法[J]. 计算机应用研究,2009,26(7):2646-2649.

基于链接回溯的地理信息更新主题爬虫研究

作者： 吴家皋， 余浩， 张雪英， WU Jia-gao， YU Hao， ZHANG Xue-ying

作者单位： 吴家皋,余浩,WU Jia-gao,YU Hao(南京邮电大学 计算机学院，江苏 南京 210003; 江苏省无线传感网高技术研究重点实验室，江苏 南京 210003)， 张雪英,ZHANG Xue-ying(南京师范大学 虚拟地理环境教育部重点实验室,江苏 南京,210023)

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(7)

参考文献(14条)

1. Zhong Shaojun;Deng Zhijuan A web crawler system design based on distributed technology 2011(12)

2. Batsakis S;Petrakis E G M;Milios E E Improving the per-formance of focused web crawlers 2009(10)

3. 刘金红;陆余良 主题网络爬虫研究综述 2007(10)

4. Cho J;Garcia-Molina H;Page L Efficient crawling through URL ordering 1998(1-7)

5. 刘淑梅;夏亮;许南山 主题搜索引擎网络爬虫搜索策略的研究与实现 2010(03)

6. Bourchtein A;Bourchtein L On some analytical properties of a general PageRank algorithm 2013(9-10)

7. 张翔;周明全;李智杰 基于PageRank与Bagging的主题爬虫研究 2010(14)

8. 吕林涛;陈丽萍;周红芳 面向垂直搜索引擎的主题提取算法 2009(15)

9. 罗林波;陈绮;吴清秀 基于Shark-Search和Hits算法的主题爬虫研究 2010(11)

10. Charkrabarti S;Dom B;Indyk P Enhanced hypertext categori-zation using hyperlink 1998

11. 皮靖;邵雄凯;肖雅夫 基于朴素贝叶斯算法的主题爬虫的研究 2012(06)

12. Peng Tao;Zou Wanli;He Fengling SVM based adaptive learn-ing method for text classification from positive and unlabeled documents 2008(03)

13. 张俊英;胡侠;卜佳俊 网页文本信息自动提取技术综述 2009(08)

14. 陈治昂;周知予;李大学 一种基于模板的快速网页文本自动抽取算法 2009(07)