CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 37 卷第 2 期 2015 年 2 月
Vol. 37, No. 2, Feb. 2015

文章编号:1007-130X(2015)02-0231-07

主题网络爬虫研究综述*

于 娟,刘 强
(福州大学经济与管理学院,福建 福州 350108)

摘 要:网络信息资源呈指数级增长,面对用户越来越个性化的需求,主题网络爬虫应运而生。主题网络爬虫是一种下载特定主题网页的程序。利用在采集页面过程获得的特定信息,主题网络爬虫抓取的页面都是与主题相关的。基于主题网络爬虫的搜索引擎以及基于主题网络爬虫构建领域语料库等应用已经得到广泛运用。首先介绍了主题爬虫的定义、工作原理;然后介绍了近年来国内外关于主题爬虫的研究状况,并比较了各种爬行策略及相关算法的优缺点;最后提出了主题网络爬虫未来的研究方向。

关键词:网络爬虫;主题爬虫;搜索引擎
中图分类号:TP393 **文献标志码**:A
doi:10.3969/j.issn.1007-130X.2015.02.007

Survey on topic-focused crawlers

YU Juan, LIU Qiang
(School of Economics and Management, Fuzhou University, Fuzhou 350108, China)

Abstract: With the exponential growth of network information resources and the growing personalized demands of customers, topic-focused crawler emerges as the times require. Topic-focused crawlers are programs designed to download web pages which are relevant to specific topics. Using information gathered at running time, topic-focused crawlers explore the webs which follow promissory hyperlinks, and fetch only pages which appear to be relevant. The searching engine and corpus building based on topic-focused crawling have been widely used. We first define the goals and operating principles of focused crawling, comprehensively analyze the recent advances at home and abroad, and then compare the crawling strategies of various topic-focused crawlers as well as the advantages and disadvantages of related algorithms. Finally, we point out the future direction of topic-focused crawling.

Key words: web crawler; focused-crawler; searching engine

1 引言

网络爬虫是用户从互联网中获取信息资源的有效工具,近些年来随着网络的不断普及,网络信息资源呈爆炸式的增长,用户的需求也越来越个性化,普通网络爬虫已经难以满足用户的需要,主题网络爬虫(后简称主题爬虫)应运而生,能很好地解决这个难题。

互联网上含有超过 16 亿的网站[1],而这些被索引的网站包含至少 217 亿的页面[2]。互联网用户从网络中寻找信息资源的途径主要是通过搜索引擎,在中国互联网络信息中心 CNNIC(China Internet Network Information Center)第 31 次报告中指出[3],2012 年中国的搜索引擎用户达到了 4.51亿,可见搜索已经成为互联网用户的基本行为,而时下的一个研究热点就是基于主题爬虫的专业搜索引擎。此外,如何通过主题爬虫有效构建领域语料库也是一个研究热点。语料库指经科学取样和加工的大规模电子文本库。借助计算机分析

* 收稿日期:2013-08-27;修回日期:2013-10-18
基金项目:国家自然科学基金资助项目(71201032);福建省社会科学规划资助项目(2012C021);福建省教育厅社会科学研究资助项目(JA11040S)
通信地址:350108 福建省福州市大学城学园路 2 号福州大学经济与管理学院
Address: School of Economics and Management, Fuzhou University, Fuzhou 350108, Fujian, P. R. China

工具,研究者可开展相关的语言理论及应用研究,诸如构建本体[4]、机器翻译[5]等方面。

近几年来关于主题爬虫的研究大量涌现,因此有必要对近几年的研究做一个综述。本文对当前主题爬虫所采用的爬行策略及算法进行分析,对比各种算法及爬行策略的优缺点,并提出了以后主题爬虫的研究方向。

2 主题爬虫的定义及工作原理

网络爬虫是一种自动抓取网页并提取网页内容的程序,是搜索引擎的信息获取渠道。通常在给定的一个或多个统一资源定位符 URL(Uniform Resource Locator)种子集情况下,从种子网页开始采集,在抓取网页的过程中,不断将新的 URL 放进待爬行的 URL 队列中,直到满足一定条件(如待爬行队列为空、达到指定爬行数量)停止爬行。

主题爬虫是按照预先定义的爬行主题,在给定初始 URL 种子集后,根据一定的分析算法,对爬行网页进行主题相关分析,过滤与主题不相关的网页,在不断抓取相关网页的过程中,将与主题相关的链接放进待爬行队列中,重复这个过程,直到达到一定条件为止。

主题爬虫不同于普通的网络爬虫,通用网络爬虫的初始 URL 种子集可以是任意门户网站,而主题爬虫的初始 URL 种子集必须是与事先定义的主题高度相关的页面;它不必收集所有的网页,只爬取那些与主题相关的页面[6~8],关注与主题相关的网页链接,然后从已下载的页面中提取 URL 并预测该 URL 是否与给定的主题相关,并按优先级顺序对 URL 进行访问,无关的 URL 将被放弃,在爬行过程中尽可能多地发现与并下载主题相关的页面,减少无关页面的下载。

目前,主题爬虫的研究主要包括以下四个方面:

(1)主题的描述,即如何描述爬取的对象。主题爬虫与通用爬虫的不同之处在于:主题爬虫只爬取与主题相关的网页。对主题的准确定义和描述,能够帮助主题爬虫更有效地发现与主题相关的网页。

(2)主题爬行策略,即待爬行 URL 队列的访问顺序。主题爬虫在抓取网页过程中,按照一定规则将相关度递减传递给子链接,并将与主题相关的链接插入待爬行队列中。再次爬行时,并不是简单地按照广度优先或者深度优先进行爬行,而是按照链接与主题的相关度进行排序,先爬行相关度较高的 URL,不同主题爬虫之间的区别之一就是怎样计算待爬行 URL 的先后顺序。

(3)主题相关性的判断。对于已经爬取的网页,主题爬虫可以通过获取页面的文本内容,采用基于文字内容的方式来判断网页是否与主题相关。不同主题爬虫之间的区别之一在于采用什么方式来判断页面与主题是否相关。

(4)如何扩大主题爬虫的覆盖范围。在网络中,主题页面具有相邻性的规律,即主题页面中的链接很大概率是指向主题相关的页面。但是,事实情况却并非如此,文献[9]指出,与主题相关的页面并非总是聚集在一起,许多页面相关的网页是通过若干个主题不相关的页面链接起来的。Bergmark D 等人[10]的研究指出,把相邻页面的距离定义为 0,主题相关页面之间的距离从 1 到 12 不等,通常距离是 5。如果把相邻的主题相关的页面称为一个主题孤岛,那么互联网就是由一个又一个的主题孤岛构成的[11]。如何设计算法使得主题爬虫尽可能覆盖这些主题孤岛,是一个亟待解决的问题。

实际上对上述(2)、(3)问题的不同解决方法,是不同主题爬虫之间的主要区别。

3 主题爬虫研究进展

近些年来,研究者们为了使主题爬虫尽可能高效高质地获取主题相关的页面,提出多种主题定制的爬行策略和算法。本文将这些方法划分为以下几类。

3.1 传统的启发式方法

传统的启发式方法主要分为两类:基于文字内容的启发式方法和基于 Web 链接超链图的评价方法。

3.1.1 基于文字内容的启发式方法

基于文字内容的启发式方法主要利用抓取网页中的锚文本(锚文本又称锚文本链接,是链接的一种形式)、文本内容、URL 字符串信息。不同的分析算法构成了不同的策略和相应的方法。其代表方法有鱼群搜索策略(Fish Search)、鲨鱼搜索策略(Shark Search)以及最佳优先策略(Best First Search)。

鱼群搜索策略是以“鱼群搜索算法”(Fish Search)为基础的主题爬行策略。在鱼群搜索策略中,每个网页相当于一条鱼,当它们发现食物(相

关信息)时,这些鱼就继续繁殖(沿链接方向继续寻找相关页面),寻找新的食物。若没有发现食物(与主题不相关),该鱼便死亡(放弃当前链接)。关于待下载链接与主题的相关性,De Bra P M E 等人[12]提出了通过比较已下载网页内容与主题关键字是否匹配,引入二元分类方法(1 代表相关,0 代表不相关)来计量相关性。

Shark Search 方法[13]在 Fish Search 算法的基础上进行了改进,引入了相似度度量方法,取值为[0,1];在计算 URL 的相关性上,已下载网页中的锚文本内容、页面内容、链接内容及父页面(指向包含链接页面的 Web 页面)的相关性等都作为主要参数用来计量待下载网页与主题的相关性,通过计算确定待下载网页是否进入待爬行队列中。

在"鲨鱼搜索策略"的基础上,Menczer F 等人[14]提出了"最佳优先"搜索策略,这一策略通过计算向量空间的相关性,把相关性"最好"的页面放入最优先下载的队列。但是,由于只选择与主题相关性很大的链接,而放弃某些当前相关性不高但下级链接中可能包含很高相关性链接的网页,具有很大的贪婪性。该算法只能找到局部范围内的最优解,难以得到全局范围内的最优解。

3.1.2　基于 Web 超链图的评价方法

基于文字内容的启发式方法只考虑了页面文字内容,而忽略了通过超链接形成的网络有向图对主题爬虫的影响。基于 Web 超链图的评价算法主要以网页级别(PageRank)和 HITS(Hyperlink-Induced Topic Search)算法为代表。

PageRank[8,15,16]算法由 Brin S 与 Page L 两人提出。PageRank 背后的概念是,每个到页面的链接都是对该页面的一次投票,被链接得越多,就意味着被其他网站投票越多。PageRank 这个概念引自学术中一篇论文的被引用的频度—即被别人引用的次数越多,一般判断这篇论文的权威性就越高。

康奈尔大学(Cornell University)的 Kleinberg J 博士于 1997 年首先提出的 HITS[17]算法,该算法是建立在页面链接关系的基础上,对链接结构的改进算法。HITS 算法通过两个评价权值—Authority(内容权威度)和 Hub(链接权威度)来对网页质量进行评估。HITS 算法认为对每一个网页应该将其内容权威度和链接权威度分开来考虑,在对网页内容权威度做出评价的基础上再对页面的链接权威度进行评价,然后给出该页面的综合评价。由于 HITS 算法将网页划分为两类,计算起来比 PageRank 算法要复杂一些,在爬行表现上也没有明显优势,因此没有 PageRank 算法应用得广泛。

基于 Web 链接的评价算法偏重于旧网页,但是往往新网页中含有价值量更大的信息。另一方面这类算法适合寻找权威网页,容易出现主题漂移的情况,而且计算量通常偏大,影响了爬行速度。从某种意义上来说,PageRank 算法与 HITS 算法不能够算真正的主题爬行策略,因为其侧重的是网络链接的权威性,而忽视了目标网页的主题相关性。有实验表明[18],基于这类算法的爬虫抓下来的网页中,与主题相关的网页的比率很快就会下降趋近于零。但是另一方面,页面链接的权威性对用户的需求是有意义的,因此针对这些缺点,之后有不少人对基于 Web 超链图的算法进行了改进。

Yuan F[19]等人根据"主题随机浏览"(用户在网络中浏览某一主题信息具有随机性)的特点,在网页传递 PageRank 值的时候,同时传递各个主题的相关度,克服了 PageRank"主题漂移"的缺点;张翔等人[20]针对 PageRank 在网页时效性问题与主题漂移的问题,加入时效权重,并采用 bagging(Bootstrap aggregating)算法来解决 PageRank 的主题漂移问题,查准率均有明显提高。

3.2　基于概念语义分析的方法

上述爬行策略中对主题描述以及相关度的计算,都是以基于关键词来作为衡量标准,但是往往会有一词多义,或一义多词的情况出现,导致添加许多噪音页面或者遗漏相关页面,而采用概念语义分析能很好地解决这个问题[21]。目前基于语义分析的算法主要包括以本体和叙词表为基础的两类算法。

3.2.1　基于本体的分析方法

本体是指共享概念模型的明确的形式化规范说明,本体的目标是捕获相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇(术语)和词汇间相互关系的明确定义[22,23]。

Dong H[24]等人采用 protégé-owl(一种本体构建工具)构建运输服务本体,提取网页内容并将其中有用的信息萃取加上网络本体语言(owl)标签转化成元数据,采用扩展的基于案例的推理 ECBR(Extended Case-Based Reasoning)算法计算元数据与本体中概念的相关度,若相关度大于阈值,则将其相关度存储在元数据属性中。实验结果表明,

该爬虫在召回率、查准率、查全率、平均查准率及错误率方面均表现出色。

Yuvarani M 等人[25]设计了一种 LSCrawler,通过利用本体来判别 URL 字符串与主题的相关性。用户每一次在搜索引擎中键入关键字时,搜索引擎返回的链接将送到 LSCrawler,下载页面并且将其中有用的信息提取出来,再与本体中的概念进行匹配来判断该链接的相关性,并对链接进行排序。

蒋国瑞等人[26]采用 owl 语言,protégé 4.0 进行领域本体建模,将领域本体引入主题爬虫的相关性分析模块中。实验表明,该方法与单纯基于关键词的主题爬虫,在查准率上有了很大的提高。

杨学明等人[27]为了方便用户在阅读时了解与文档相关的文章与信息,实现主动推荐与关联功能,设计了一种面向网络资源的本体自动构建方法,通过对网络上各领域 Web 语料文档库进行挖掘来实现本体学习,并构建基于本体的主题爬虫。其应用在数字图书馆中,作为数字图书馆和互联网间的桥梁和媒介。在每次抓取结束后得到一个文档列表,以及高频词与词项间的关系,并将其加入相应本体中。实验表明该设计有一定效果,但性能需要进一步提高。

Wang J 等人[28]尝试通过搜集全面的应急计划信息,以预防潜在的自然灾害,设计出一种基于领域本体的应急计划主题爬虫。通过构建领域本体来定义主题,采用 URL 模式库对待爬行链接的相关度进行预测,在实验中,其查准率高达 97%,效果显著。

3.2.2 基于叙词表的分析方法

叙词表是一种将文献作者、标引者和检索者使用的自然语言转换成规范化的叙词型主题检索语言的术语控制工具,亦称主题词表、检索词典。它是一种概括某一学科领域,以规范化的、受控的、动态性的叙词(主题词)为基本成分和以参照系统显示词间关系,用于标引、存储和检索文献的词典。叙词表是叙词法的具体体现。

文献[29]将叙词表同传统的信息检索技术相结合,提出用叙词表的族对爬虫的主题进行描述的方法并用该方法设计实现主题爬虫。主题爬虫的框架由两部分组成:一部分是主题域的构造,另一部分是通过主题域对网页的主题过滤和链接分析控制,实现页面主题资源的自动形成。通过实验,较 Google 与百度搜索结果相比,爬取主题网页质量明显提高,但是采集网页数量有限,结果有待进一步检验。

文献[30]提出基于叙词表来构建一种称为概念树的表示方法来描述主题的概念的分析算法框架。并且该爬虫具有一定自适应性,在分析 URL 的相关度时,首先判断其锚文本的相关度是否达到一定的阈值 σ,只有当锚文本的相关度达不到 σ 时才会去下载 URL 对应的页面进行分析,否则将锚文本的相关度作为 URL 的相关度。这样的 URL 相关度计算方法可以大大减少不必要的计算开销,又可以充分地利用锚文本的信息。

基于概念语义的分析算法能够很好地描述主题,为主题爬虫以后的工作提供一个良好的开始,并且在网页相关度计算时能极大地提高其准确性。但是,基于语义的分析算法中,重新建立本体往往显得比较复杂,难于实现,因此采用现有本体则应用的领域受到限制,采用叙词表较易实现,但很难在语义和知识层次上描述信息,应用前景不如本体。

3.3 经验爬行方法

以往的爬虫主要关注于如何寻求与主题最相关的链接,而采用各种算法来分析判别,但是绝大多数的研究并没有深刻探讨主题爬虫的初次爬行,也没有一篇文献提出关于主题爬虫再次爬行的状况。主题爬虫在爬行过程中,一般不对前一次爬行所获取的经验进行利用,再次爬行过程中,如遇到相同或相似的页面或链接需要进行重复计算,造成效率低下。

在主题爬虫的每一次爬行中,都会获得相当一部分的信息,如何将这些信息反馈给爬虫,形成经验知识,指导主题爬虫的再次爬行,是近些年来研究者们的一个研究热点。

宋海洋等人[31]针对主题爬虫的“二次爬行”提出一种新的爬行算法,将爬虫“首次爬行”过程得到的一些信息形成一份经验树,采用经验树中已存储的网页的相关度来指导主题爬虫的“二次爬行”。实验结果较基于内容分析的主题爬虫、基于链接分析的主题爬虫、传统的基于内容分析与链接分析结合的主题爬虫相比,在准确率与爬行速度上都有不错的表现。

Rungsawang A 等人[32]提出一种具有自学习能力的主题爬虫。用户输入一个关键词集合,爬虫根据关键词集合进行首次爬行的经验搜集并形成知识库,指导爬虫以后每次更高效地爬行,每次爬行后对知识库进行更新。实验采取三次爬行,在结果中显示,每次爬行的查准率都有很明显的提高。

傅向华等人[33]采用快速Q学习与半监督贝叶斯分类器构建主题爬虫,将主题爬虫的爬行过程看做是一个执行序列的过程,在每一次寻找相关网页的过程中,计算该链接链路的Q值,然后根据Q值过滤无关链接;当得到主题相关页面时产生回报,将回报沿链接链路反馈,更新链路上所有链接的Q值,并选择相应的特征文本作为训练样本,增量地改善主题评估器和Q值预测器。实验中与标准聚焦爬行模型和扩展标准聚焦爬行模型进行比较,随着训练集的增加,其查准率逐渐上升,并且超过了上述两种模型。

在经验爬行方法中,由于主题爬虫在再次爬行中能够参考先前爬行的经验,能够更好地过滤不相关链接,并且通过对知识库的不断完善,能够使得对主题的定义更加准确,更易发现主题相关网页,缺点是在训练集小的情况下,爬行效果不够理想。

3.4　其他方法

针对主题爬虫在爬行过程中易陷入"局部最优的问题",贺晟等人[34]利用模拟退火算法在选择优化解方面具有"非贪婪性",结合"隧道技术"扩大主题爬虫的搜索范围,对主题爬虫进行改进。实验表明,较最佳优先策略,在查准率上有了明显的提高。利用遗传算法全局寻优概率搜索的特点,刘国靖[35]、曾广朴等人[36]先后将遗传算法引入主题爬虫的设计。刘国靖通过变异操作引入新的URL扩大搜索范围,采用交叉操作产生大量的URL,再经过选择操作选出适应度高的个体作为下一代的种子URL;曾广朴采用小生境遗传算法,将待爬行的URL作为遗传个体,采用基于主题相关度的适应度函数计算并评估个体适应度,采用概率的变迁规则和小生境淘汰运算引导它的搜索方向,实验表明,两种算法在查全率上均表现突出。Feng S等人[37]针对初始URL集与目标页面距离较远和目标页面所占总页面比例较小的问题,采用一种两步双向基于链接的相关值传递方法来计算网页的相关值,并与基于强化学习和上下文图的爬行方法进行比较,在查全率与查准率的表现上都非常好,其缺点就是不能对动态页面进行搜集。童亚拉[38]根据网络爬虫的实际搜索情况,在线调整两种价值的权重,使网络爬虫在主题相关Web社区搜索时选用基于立即价值的搜索策略,而在从无关社区过渡到相关社区的过程中提高未来价值链接的比重,既提高了网络爬虫跨越主题无关社区的能力,又提高了搜索相关主题文档的精度,具有自适应性。白玉昭等人[39]将概率模型引进到主题爬虫的构建中,并综合考虑主题相关度评价指标、历史评价指标、网页质量评价指标来计算待爬行URL优先度。通过实验表明,综合考虑三个指标的爬虫,在查准率和召回页面的主题相关度方面,要明显优于只考虑单个或两个指标的爬虫。Ahmadi-Abkenari F等人[40]针对主题爬虫爬行速度慢,而且不能很好地跟踪最新的网页,设计出一种采用点击流的方式来计算网页的相关度的并行爬虫,并且各个并行单元能够很好地相互协调,以减少各单元之间不必要的交互,从而能够提高主题爬虫的爬行速度。Maimunah S等人[44]针对主题爬虫通常会为提高查准率而牺牲召回率的问题,提出一种基于联合引用及联合参考的双向爬行策略。实验表明,即使在只有少量种子URL的情况下,在保持准确率的前提下,查全率也有了很大的提高,很好地解决了高召回率而低查准率的这一难题。

4　主题爬虫研究趋势

随着主题爬行爬虫的研究不断深入,主题爬虫的爬行策略及算法也在不断完善。未来主题爬虫研究方向主要围绕以下几个方面进行:

4.1　增加主题爬虫的自适应性

主题爬虫的自适应性主要表现在:在互联网中,不同类型网站的网页间组织形式相差很大,而目前的网络爬虫通常采用固定的搜索策略,并不能有效地搜集各类型网页,缺乏适应性,如何提高网络爬虫的自适应性有待进一步研究。

4.2　初始URL种子集的自动构建

主题爬虫的初始URL种子集的选择,对主题爬虫在之后的爬行表现有着重要的影响。一般情况下,初始URL种子集的选取方面往往需要人工与计算机相结合来选取,以保证主题爬虫的效率。但是,主题爬虫涉及的领域甚多,每次采用人工与计算机结合的方式耗时耗力,如何设计算法使得主题爬虫在面向不同领域时能够自动生成相应的初始URL种子集是未来的一个研究热点。

4.3　提高待爬行URL主题预测准确度

主题爬虫与通用爬虫的主要区别之一在于,能够有选择性地过滤与主题无关的链接,选择与主题相关的页面方向进行发掘。因此,在对待爬行URL进行主题相关度预测时,如果能够准确地判断待爬行URL与主题的相关度,过滤不相关链

接,能够大大节省时间,提高效率。

总之,如何在合理的时间限度内,以较少的网络资源、存储资源和计算资源的消耗获得更多的主题相关页面是主题爬虫追求的最终目标。

5 结束语

本文介绍了当前主题爬虫采用的爬行策略及算法,并分析了各种方法的优缺点,在此基础之上提出了主题爬虫以后的研究方向。目前关于主题爬虫的研究主要针对的是其爬行策略与页面相关度分析算法,目标是提高主题爬虫的查准率和查全率。

本文认为:主题爬虫自身的健壮性以及可扩展性,是主题爬虫提高查准率和查全率的保证,因此在设计主题爬虫时应当考虑该问题;另一方面,主题爬虫在爬取网页过程中应该注意访问网站的时间间隔,即礼貌爬行,以免频繁访问给网站带来过度负载。

参考文献:

[1] Netcraft: There are over 1.6 billion Websites in the World Wide Web (WWW). [EB/OL]. [2008-03-28]. http://it.hexun.com/2008-03-28/104849372.html.

[2] The indexed web contains at least 21.7 billion pages[EB/OL]. [2009-11-11]. http://www.worldwidewebsize.com/index.php?lang=EN.

[3] CNNIC thirty-first Internet Report[EB/OL]. [2013-01-15]. http://www.cnnic.net.cn/gywm/shzr/shzrdt/201301/t20130115_38518.htm.

[4] Yu Juan. Learning domain ontologies from Chinese text corpora [D]. Dalian: Dalian University of Technology, 2010. (in Chinese)

[5] Hu Fu-mao. The use of parallel corpora in machine translation of business correspondence [J]. China Business & Trade, 2010(14): 212-213. (in Chinese)

[6] Zareh Bidoki A M, Yazdani N, Ghodsnia P. FICA: A fast intelligent crawling algorithm[C]// Proc of the IEEE/WIC/ACM International Conference on Web Intelligence, 2007: 635-641.

[7] Xiaoqing C, Chun Y. An evolutionary relevance calculation measure in topic crawler[C]// Proc of ISECS International Colloquium on Computing, Communication, Control, and Management, 2009, 4: 267-270.

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

[9] McCallum A, Nigam K, Rennie J, et al. A machine learning approach to building domain-specific search engines[C]// Proc of IJCAI'99, 1999: 662-667.

[10] Bergmark D, Lagoze C, Sbityakov A. Focused crawls, tunneling, and digital libraries [M]// Research and Advanced Technology for Digital Libraries. Springer Berlin Heidelberg, 2002: 91-106.

[11] Ye Qin-yong. URL rule based focused crawl and its application[D]. Hangzhou: Zhejiang University, 2007. (in Chinese)

[12] De Bra P M E, Post R D J. Information retrieval in the World-Wide Web: Making client-based searching feasible[J]. Computer Networks and ISDN Systems, 1994, 27(2): 183-192.

[13] Hersovici M, Jacovi M, Maarek Y S, et al. The shark-search algorithm. An application: tailored Web site mapping[J]. Computer Networks and ISDN Systems, 1998, 30(1): 317-326.

[14] Menczer F, Pant G, Srinivasan P. Topical web crawlers: Evaluating adaptive algorithms[J]. ACM Transactions on Internet Technology (TOIT), 2004, 4(4): 378-419.

[15] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN systems, 1998, 30(1): 107-117.

[16] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[R]. Technical Report SIDL-WP-1999-0120, Standford: Standford Laboratory, 1999.

[17] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM (JACM), 1999, 46(5): 604-632.

[18] Novak B. A survey of focused web crawling algorithms[C]// Proc of SIKDD2004 at Multiconference IS, 2004: 55-58.

[19] Yuan F, Yin C, Liu J. Improvement of PageRank for focused crawler[C]// Proc fo the 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007: 797-802.

[20] Zhang Xiang, Zhou Min-quan, Li Zhi-jie, et al. Focused crawler based on PageRank and Bagging[J]. Computer Engineering and Design, 2010, 31(14): 3309-3312. (in Chinese)

[21] Wang Shuai, Zhou Guo-min, Wang Jian. Reviews of relevance algorithm in focused crawler[J]. Computer and Modernization, 2013(4): 27-30. (in Chinese)

[22] Studer R, Benjamins V R, Fensel D. Knowledge engineering: Principles and methods[J]. Data & Knowledge Engineering, 1998, 25(1): 161-197.

[23] Yadav P, Kalra M M, Yadav K P. Enhancing the performance of web focused crawler using ontology[J]. International Journal of Computers & Technology, 2013, 4(2): 477-482.

[24] Dong H, Hussain F K, Chang E. A transport service ontology-based focused crawler[C]// Proc of the 4th International Conference on Semantics, Knowledge and Grid, 2008: 49-56.

[25] Yuvarani M, Iyengar N C S N, Kannan A. LSCrawler: A framework for an enhanced focused web crawler based on link semantics[C]// Proc of IEEE/WIC/ACM International Conference on Web Intelligence, 2006: 794-800.

[26] Jiang Guo-rui, Wang Qiu-li. Research on focused crawler of TBT electronic information products based on ontology[J]. Journal of Intelligence, 2011, 30(7): 157-161. (in Chinese)

[27] Yang Xue-ming, Liu Bai-song. Applications of topical crawl-

er in digital libraries[J]. Library Journal, 2007(8):47-50. (in Chinese)

[28] Wang J, Dang D, Zhou P, et al. Crawling strategy based on domain ontology of emergency plans[C]// Proc of 2013 the International Conference on Education Technology and Information System (ICETIS 2013),2013:1.

[29] Xia Chong-pu, Kang Li. The focused-crawler based on thesaurus[J]. New Technology of Library and Information Service, 2007, 150(5):41-44. (in Chinese)

[30] Xie Zhi-ni. A new subject-based web crawler with concept tree[J]. Computer and Modernization,2010, 176(4):103-106. (in Chinese)

[31] Song Hai-yang, Liu Xiao-ran, Qian Hai-jun. A novel crawling strategy of focused web crawler[J]. Computer Applications and Software, 2011, 28(11):264-267. (in Chinese)

[32] Rungsawang A, Angkawattanawit N. Learnable topic-specific web crawler[J]. Journal of Network and Computer Applications , 2005(28):97-114.

[33] Fu Xiang-hua, Feng Bo-qin, Ma Zhao-feng, et al. Focused crawling method with online-incremental adaptive learning [J]. Journal of Xi'an Jiaotong University, 2004, 38(6): 599-602. (in Chinese)

[34] He Shen,Cheng Jia-xing,Cai Xin-bao. Focused crawler based on simulated anneal algorithm [J]. Computer Technology and Development, 2009, 19(12):55-58. (in Chinese)

[35] Liu Guo-jing,Kang Li,Luo Chang-shou. Strategy of focused crawler based on genetic algorithm[J]. Computer Applications, 2007, 27(12):172-174. (in Chinese)

[36] Zeng Guang-pu, Fan Hui-lian. Search strategy of focused crawler based on genetic algorithm[J]. Computer Engineering, 2010, 36(11):167-169. (in Chinese)

[37] Feng S, Zhang L, Xiong Y, et al. Focused crawling using navigational rank[C]// Proc of the 19th ACM International Conference on Information and Knowledge Management, 2010:1513-1516.

[38] Tong Ya-la. Application in topic information extraction from web based on adaptive dynamical evolutional particle swarm optimization[J]. Geomatics and Information Science of Wuhan University, 2008, 33(12):1296-1299. (in Chinese)

[39] Bai Yu-zhao, Liang Jiu-zhen. Research and implementation for focused crawler based on probabilistic model [J]. Computer Engineering and Science, 2013, 35(1):160-165. (in Chinese)

[40] Ahmadi-Abkenari F, Selamat A. Application of clickstream analysis in a tailored focused web crawler[J]. Journal of Communications of SIWN, The Systemic and Informatics World Network, 2010,10:137-144.

[41] Maimunah S, Widyantoro D H, Sastramihardja H S. Co-citation & co-reference concepts to control focused crawler exploration[C]// Proc of 2011 International Conference on Electrical Engineering and Informatics, 2011:1-7.

附中文参考文献:

[4] 于娟. 基于文本的领域本体学习方法及其应用研究[D]. 大连:大连理工大学, 2010.

[5] 胡富茂. 平行语料库在商务信函机器翻译中的应用[J]. 中国商贸, 2010(14):212-213.

[11] 叶勤勇. 基于URL规则的聚焦爬虫及其应用 [D]. 杭州:浙江大学, 2007.

[20] 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):3309-3312.

[21] 王帅, 周国民, 王健. 主题爬虫相关度算法研究综述[J]. 计算机与现代化, 2013 (4):27-30.

[26] 蒋国瑞, 王秋利. 基于本体的TBT电子信息产品领域主题爬虫研究[J]. 情报杂志, 2011, 30(7):157-161.

[27] 杨学明, 刘柏嵩. 主题爬虫在数字图书馆中的应用[J]. 图书馆杂志,2007(8):47-50.

[29] 夏崇锴, 康丽. 基于叙词表的主题爬虫技术研究[J]. 现代图书情报技术, 2007, 150(5):41-44.

[30] 谢志妮. 一种新的基于概念树的主题网络爬虫方法[J]. 计算机与现代化,2010, 176(4):103-106.

[31] 宋海洋, 刘晓然, 钱海俊. 一种新的主题网络爬虫爬行策略[J]. 计算机应用与软件, 2011, 28(11):264-267.

[33] 傅向华, 冯博琴, 马兆丰, 等. 可在线增量自学习的聚焦爬行方法[J]. 西安交通大学学报, 2004, 38(6):599-602.

[34] 贺晟, 程家兴, 蔡欣宝. 基于模拟退火算法的主题爬虫[J]. 计算机技术与发展, 2009, 19(12):55-58.

[35] 刘国靖, 康丽, 罗长寿. 基于遗传算法的主题爬虫策略[J]. 计算机应用, 2007, 27(12):172-174.

[36] 曾广朴, 范会联. 基于遗传算法的聚焦爬虫搜索策略[J]. 计算机工程, 2010, 36(11):167-169.

[38] 童亚拉. 自适应动态演化粒子群算法在Web主题信息搜索中的应用[J]. 武汉大学学报(信息科学版), 2008, 33(12):1296-1299.

[39] 白玉昭, 梁久祯. 基于概率模型的主题爬虫的研究和实现[J]. 计算机工程与科学, 2013, 35(1):160-165.

作者简介:

于娟(1981-),女,山东青岛人,博士,副教授,研究方向为知识管理和领域本体。**E-mail**:infoyujuan1@163.com

YU Juan, born in 1981, PhD, associate professor, her research interests include knowledge management, and domain ontology.

刘强(1987-),男,湖南攸县人,硕士生,研究方向为数据挖掘与决策支持系统。**E-mail**:haha13743@126.com

LIU Qiang, born in 1987, MS candidate, his research interests include data mining, and decision support system.