

主题爬虫的搜索策略研究

刘汉兴, 刘财兴

(华南农业大学 信息学院, 广东 广州 510642)

摘 要: 主题爬虫收集主题相关信息时, 需要评价网页的主题相关度, 并优先爬取相关度较高的网页, 在决定了搜索路径的同时也决定了主题爬虫的搜索效率。针对不同的网页评价算法, 对现有的主题爬虫的搜索策略进行分类, 指出了各类搜索策略的特点和优缺点, 总结了能够提高主题爬虫搜索效率的几方面内容。

关键词: 主题爬虫; 搜索策略; 页面评价; 搜索引擎; 优化

中图分类号: TP391 **文献标识码:** A **文章编号:** 1000-7024 (2008) 12-3160-03

Survey on searching strategies of focused crawler

LIU Han-xing, LIU Cai-xing

(College of Informatics, South China Agricultural University, Guangzhou 510642, China)

Abstract: While focused Crawler collect information, it needs to evaluate the relevance of web pages, and process firstly pages which have higher relevance, thus deciding the search path and efficiency of crawler. Web crawler's searching strategies based on the way they evaluate the web page is categorized. The character of each class of searching strategy is described and the advantage and disadvantage is discussed, several ways to improving the efficiency of web crawlers are summed up.

Key words: focused crawler; searching strategy; page evaluating; search engine; optimization

0 引言

目前的谷歌、百度等搜索引擎, 自动搜集整理互联网上的信息, 为一般用户提供检索服务, 可以称为通用搜索引擎。但对于专业用户及研究人员来说, 他们的查询往往是针对某个领域或面向特定主题, 使用通用搜索引擎进行检索效果不理想, 准确率和召回率都很低, 因此就出现了主题搜索引擎(topic-specific search engine, 又称专业搜索引擎)。

网络爬虫(Crawler, 或 Spider 程序)是一个自动下载 Web 网页的程序, 是搜索引擎的基础与核心。主题搜索引擎中的主题爬虫, 首先需要定义“主题概念”, 明确“主题”的范围和内容, 即对“主题”进行描述或定义。主题概念可以用主题词集来表示, 也可以表示为示例文档(由用户选定的种子样本), 也可来源于某一领域概念。主题爬虫在工作时, 只抓取与主题相关的网页或内容。为了保证采集到的信息主题相关性, 以何种策略来决定访问 Web 的搜索路径, 是主题爬虫研究的焦点^[1-4]。该文根据网页评价算法的不同, 对比分析了主题爬虫的几种搜索策略, 总结了提高主题爬虫搜索效率的几个方面。

1 主题爬虫的工作原理

网络爬虫在采集 Web 信息时, 通常从一个“种子集”(种子

链接)出发, 下载页面并提取其中的子链接, 然后再访问子链接对应的内容, 如此不断重复即可实现遍历 Web 信息。网络爬虫的搜索策略与搜索引擎的性质和任务密切相关^[5], 为了获得较高的 Web 覆盖率, 通用搜索引擎网络爬虫通常采用图的遍历算法搜索 Web, 如图 1(a)所示, 其中白框代表主体无关页面, 黑框代表主体相关页面, 虚线代表链接, 实箭头代表访问顺序)。

主题搜索引擎索引的内容只限于特定主题或专门领域, 因而在搜索的过程中无须对整个 Web 进行遍历, 如图 1(b)所示, 它只需选择与主题页面相关的页面进行访问。

网络爬虫对网页的抓取策略分为广度优先和最佳优先两种, 主题爬虫主要采用后者^[1-2]。广度优先能较快找到高质量的网页, 同时页面覆盖率较高, 但随着爬虫“爬行”的深入, 抓取页面的相关度也随之降低。最佳优先策略的基本思想是按照一定的网页评价算法, 计算网页与主题的相关性, 选取“价值”最高的网页中的链接进行抓取。因此, 如何评价页面价值成为研究主题爬虫搜索策略的关键。

2 网页评价算法研究

Web 上的页面分布表面看似杂乱无章, 但主题页面的分布却有一定的规律, 可总结为 4 个特征^[3,6,10]: ①站点主题特征,

收稿日期: 2007-06-25 E-mail: liuhx666@21cn.com

基金项目: 国家 863 高技术研究发展计划基金项目 (2006AA10Z246)。

作者简介: 刘汉兴 (1971—), 男, 湖北鄂州人, 硕士, 讲师, 研究方向为智能检索、自然语言处理; 刘财兴 (1962—), 男, 副教授, 研究方向为无线传感器网络、计算机网络。

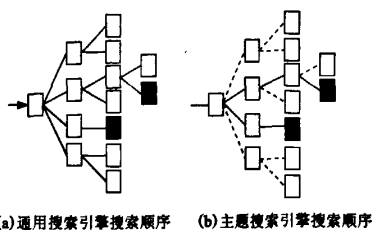


图1 两类搜索引擎爬虫搜索顺序

即一个站点倾向于说明一个或多个主题；②Hub特征，Hub页面是指该页面不但含有许多指出链接，并且这些链接趋向于同一主题；③Linkage/Sibling Locality特征，Linkage Locality是指页面倾向于拥有链接到它的页面的主题，Sibling Locality是指对于一个链接到某个主题页面的页面而言，它所链接指向的其它页面也倾向于和这个主题相关；④Tunnel特征，在不同的主题页面之间，往往是通过许多主题无链接连接在一起。由此，网页评价算法可归纳为不同类型。

2.1 基于网络拓扑结构的评价算法

Web页面是一种含有丰富链接结构的半结构化文档，其中链接结构是爬虫工作的基础。链接分析是基于这样一个前提：把超链接看作是对它所指的页面的赞许。当页面A通过超链接指向页面B时说明两点：①页面B与页面A是相关联的；②页面B是值得关注的质量较好的页面。通过网页之间的链接结构，来评价与网页有直接或间接链接关系的对象(网页或网站)的算法，本文称为基于网络拓扑结构的搜索策略。

针对网站级别的算法，主要考虑网站之间的链接关系，按照一定的模型计算链接的权重，关键之处在于站点的划分和站点等级(SiteRank)的计算^[7-9]。Wu和Aberer讨论了在分布式情况下，通过对同一个域名下不同主机、服务器的IP地址进行站点划分，构造站点图，计算出站点的SiteRank。实验表明能有效减少运算的代价，间接说明了网页的重要性，另外还可避免针对网页统计算法的欺骗行为^[7]。

针对网页级别的分析算法中，典型的有PageRank^[9]和HITS^[9]，两者都是通过对网页间链接度的递归和规范化计算，得到每个网页的重要度评价。PageRank算法的“用户冲浪”模型考虑了用户访问行为的随机性，但忽略了用户访问行为的目的性，即网页和链接与查询主题的相关性。针对这个问题，HITS算法计算页面的Authority权重和Hub权重，并以此决定页面中链接的访问顺序。

基于链接结构评价的搜索策略，考虑了链接的结构特征，对主题相关网站搜索时使用效果较好，但由于忽略页面内容与主题的相关性，容易出现搜索偏离主题的“主题漂移”问题，另外在搜索过程中需要迭代计算PageRank值或Authority及Hub权重，当页面和链接数量不断增长时计算复杂度也呈指数级增长^[9]。

2.2 基于网页内容的评价算法

基于网页内容的分析算法指的是利用网页内容(词条等)特征进行的网页评价。网页的内容由最初静态的Html页面(surface web)，发展到以动态页面(Deep Web或Hidden Web)为主的页面分布情况^[9]，相对于可以被搜索引擎直接处理的前者

不同，Deep Web主要是由结构化的数据源动态生成，搜索引擎只能覆盖大约1/3的页面。根据网页组织形式的不同，将基于网页内容的分析算法，分为两类：一类主要针对Surface Web，以分析直接可见的文本和超链接为主的网页；另一类针对Deep Web，主要分析动态生成的网页。

2.2.1 基于Surface Web的网页评价算法

这类算法以分析页面可见的文本内容为主，主要利用文本相似度的计算方法评价页面文本与主题集之间相似程度，然后再根据相似度的高低决定页面中的链接的访问顺序。

Fish Search算法^[1]将用户输入的查询关键词或短语作为主题，将包含查询串的页面看作与主题相关(相关度为一个离散的值)，且仅搜索主题相关页面。优点是模式简单、动态搜索，但不能评价相关程度的高低。Shark Search算法^[1]在前者的基础上加以改进，充分利用了锚文本及其上下文，采用基于连续值的相似度函数计算链接价值，在计算出页面是否相关的同时，还可以得出相关性的大小。Best-First算法^[6]，利用向量空间模型计算页面与主题的相似度，实验说明其性能优于Page-Rank算法^[1-3]，其计算方法如公式(1)所示

$$Sim(p,q) = \frac{\sum_{k=1}^M w_{a_k} * w_{p_k}}{\sqrt{(\sum_{k=1}^M w_{a_k}^2)(\sum_{k=1}^M w_{p_k}^2)}} \quad (1)$$

式中： p, q ——主题向量和页面向量， w_{a_k}, w_{p_k} ——主题和页面的特征项的权重， M ——维数。

以上算法都考虑以文本的内容与主题的相似度来评价链接价值的高低，从而决定其搜索策略。优点是计算简单，在距离相关页面较近的地方搜索时性能较好，但由于忽略了Web页面的结构化特征，很难反映Web的整体情况，存在“近视”的缺点^[10]。

2.2.2 基于Deep Web的页面评价算法

Deep Web后台数据库多是结构化的关系数据库，因此信息的质量较高。传统的爬虫只能爬行所谓的公共可索引的页面，而Deep Web页面是为响应来自客户端的表单查询请求由服务器端后台数据库动态产生的。Deep爬虫使用启发式函数集和知识库来自动发现相关表单^[11]，同时填写表单和搜集含有匹配结果的页面集，所设计的爬虫针对特定应用领域，其工作过程可分为3步：①寻找表单；②学习填写表单；③识别和取回结果页面。其中Deep Web爬虫第一步从站点主页开始爬行表单页面，这个过程使用一组启发式规则来去除非研究表单；第二步从表单中抽取标签，配合领域本体知识库的参与，爬虫尽力学习如何正确地填写表单；最后一步提交表单，然后取回结果页面识别记录。

2.3 基于概念语义的评价算法

目前的网络爬虫主要采用机械的关键字的匹配来实现，对信息的内容缺乏知识处理和理解能力，把信息检索从基于关键词层面提升到基于知识(概念语义)层面是解决问题的关键。主题爬虫针对的“主题”是与领域相关的一组概念，有学者使用“概念空间”来表示^[12]，目前较流行的是本体ontology来表示^[13]。ontology定义为“给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外

延的规则的定义”，它包含5个基本元素：类、关系、函数、公理、实例^[14]。

用于主题爬虫中的本体，可以在领域专家的指导下构建，也可以直接使用基于字典表示的本体如 Wordnet^[15]、HowNet^[16]等。由于本体具有良好的概念层次及概念间、属性间关系定义，能较容易得到一个词语的同义词和上、下义词。Ehring^[17]在通过领域本体分析页面时，检查页面中的关键词，看能否与本体中的主题词进行匹配，如匹配则将该词转换成主题特征项，从而将关键词转换成概念，然后对网页中出现的概念进行计数和加权，根据本体的概念层次，离核心概念越近的权重越高，算出与所选领域的相关度。这种基于语义的方法比基于关键字的分类方法具有更高的准确性和效率，抗干扰能力也更强。赵丰年^[17]则是利用 HowNet，查询出主题和文本中的关键词的义原，分别计算出主题和文本的概念向量，然后通过计算两个向量之间的距离来计算相关度。实验结果表明基于概念的主题过滤功能优于关键词方法，但 HowNet 作为一个常识知识库，不适合处理某一领域，该方法使用时有所限制。

2.4 基于未来价值的页面评价算法

针对最佳优先算法容易过早陷入 Web 搜索空间中局部最优的缺陷，文献[3-4]提出了链接的立即回报和未来回报的概念，即衡量一个链接的价值时既要考虑可以直接计算出来的立即价值，还要考虑该链接(立即价值不高)可能指向许多主题相关页面所带来的未来价值。Diligenti^[18]提出了基于上下文图的搜索策略，分为训练和搜索两个阶段：在训练阶段构造距离种子集的不同层次集及分类器；搜索时利用该分类器判断新页面属于哪个层次集，从而估计该页面距离目标页面的远近，并优先距离目标较近的页面中的链接。林海霞^[4]提出了一种改进的 BS-BS 算法，召回率比单纯的 BestFirst 算法有所提高。

2.5 基于动态价值的评价算法

针对评价页面价值所涉及的各种因素动态变化的特点，Aggarwal^[6]提出了“智能搜索”技术，在搜索过程中“在线”学习链接的结构特征获取用户的兴趣集，用于过滤不相关页面或链接；Chakrabarti^[5]利用“在线相关反馈”技术调整搜索策略，提高搜索效率；Ester^[19]针对网页的 tunnel 特征，提出基于“隧道”技术的搜索策略，即当搜索位置距相关页面较远时，调整搜索策略扩大搜索主题范围，使网络爬虫能跨过无关页面，确保正确的搜索方向。实验表明，这类基于“动态”价值评价的搜索策略能有效提高搜索效率。

3 主题爬虫搜索策略的优化

针对主题爬虫采用不同的搜索策略而引起的性能问题，有学者研究了主题爬虫的评价方法^[1-2]，并设计了系统框架。Menczer^[2]提出了3种方法来评价主题爬虫的搜索结果：①分类器方法；②检索系统方法(SMART)；③平均相似度方法。由于缺少合适的相关集和 Web 信息动态变化的特点，直接评价主题爬虫的性能比较困难。总之，主题爬虫应该以较少的资源(网络资源、存储资源和计算资源等)代价，在较短的时间内获取更多的相关页面。因此，该文认为爬虫的搜索策略优化还应考虑以下方面：

(1)综合各类评价算法来准确评价页面以及链接价值。由于单一评价算法各自的局限性，综合各类评价方法，考虑页面的内容、结构、语义特征以及用户行为等因素，准确计算页面的价值，是主题爬虫“爬行”的重要依据和对检索结果排序的基础。在不打开链接所指向的页面的前提下判断该链接的价值则能有效提高爬虫的工作效率。

(2)主题概念的准确描述和动态适应。主题概念的提取、主题的表现形式是主题爬虫工作的前提，而主题的动态适应则是主题爬虫“智能化”的基础。

(3)采用“启发式”搜索策略。为避免采用最佳优先搜索策略容易过早陷入 Web 搜索空间中局部最优子空间的陷阱，从而导致整体回报率不高的缺点，应采用“启发式”算法，在主题爬虫搜索过程中，每次除选择价值最优的链接外，还以一定概率有限度地接收价值次优的链接。

4 结束语

随着用户对信息检索的“专业化”、“个性化”要求，主题搜索引擎将是未来搜索引擎的发展方向。在介绍了搜索引擎的工作原理的基础上，分析了主题爬虫中的页面评价算法，根据网络拓扑结构、网页内容、概念语义等方面进行分类和比较。最后归纳说明了进一步优化应考虑的方向。

参考文献：

- [1] Srinivasan P, Menczer F, Pant G. A general evaluation framework for topical crawlers[J]. Information Retrieval, 2005, 8(3): 417-447.
- [2] Menczer F, Pant G, Srinivasan P. Topic web crawler: Evaluating adaptive algorithm [J]. ACM Transactions on Internet Technology, 2004, 4(4): 378-419.
- [3] 李学勇, 欧阳柳波, 李国徽, 等. 网络蜘蛛搜索策略比较研究[J]. 计算机工程与应用, 2004, 40(4): 128-131.
- [4] 林海霞, 原福永, 陈金森, 等. 一种改进的主题网络蜘蛛搜索算法 [J]. 计算机工程与应用, 2007, 43(10): 174-176.
- [5] Chakrabarti S, Punera K, Subramanyam M. Accelerated focused crawling through online relevance feedback[C]. Honolulu: Proc of the 11th International World Wide Web Conference, 2001.
- [6] Aggarwal C, AL-Garawi F, Yu S P. Intelligent crawling on the world wide web with arbitrary Predicate[C]. HongKong: Proc of the 10th International World Wide Web Conference, 2001.
- [7] Wu J, Aberer K. Using siterank for decentralized computation of web document ranking[EB/OL]. <http://lsirpeople.epfl.ch/abrerer/PAPERS/AH2004.pdf>.
- [8] Bharat K, Chang BW, Henzinger MR, et al. Who links to whom: Mining linkage between web sites [C]. California: Proc of the IEEE International Conference on Data Mining (ICDM'01), 2001.
- [9] ET O'Neill, BF Lavoie, Bennett R. Trends in the evolution of the public web [EB/OL]. <http://dlib.org/dlib/april03/lavoie/04-lavoie.html>.

(下转第 3166 页)

表2 基于二维运动场的神经网络估计仿真实验

序号	参数 μ		真实值	估计值	误差
1	0.005	R	$\begin{pmatrix} 0.69353509 & -0.33259091 & 0.63905579 \\ 0.63905579 & 0.69353509 & -0.33259091 \\ -0.33259091 & 0.63905579 & 0.69353509 \end{pmatrix}$	$\begin{pmatrix} 0.69656301 & -0.32695553 & 0.63868570 \\ 0.63602126 & 0.69345683 & -0.33859468 \\ -0.33206156 & 0.64208692 & 0.69100356 \end{pmatrix}$	0.005836640460
		T	$(13 \ -5 \ 3)^T$	$(11.79440212 \ -4.29482746 \ 2.69716239)^T$	0.100306062904
2	0.01	R	$\begin{pmatrix} 0.69353509 & -0.33259091 & 0.63905579 \\ 0.63905579 & 0.69353509 & -0.33259091 \\ -0.33259091 & 0.63905579 & 0.69353509 \end{pmatrix}$	$\begin{pmatrix} 0.69485998 & -0.32995775 & 0.63895863 \\ 0.63769352 & 0.69346100 & -0.33532450 \\ -0.33248588 & 0.64045310 & 0.69234425 \end{pmatrix}$	0.002671681838
		T	$(13 \ -5 \ 3)^T$	$(12.43967533 \ -4.70299673 \ 2.85540509)^T$	0.045652487068
3	0.01	R	$\begin{pmatrix} 0.41492507 & -0.77503961 & 0.47660345 \\ 0.48553500 & 0.63161546 & 0.60441518 \\ -0.76947576 & -0.01937935 & 0.63838190 \end{pmatrix}$	$\begin{pmatrix} 0.41519690 & -0.77474540 & 0.47683793 \\ 0.48534754 & 0.63196665 & 0.60419792 \\ -0.76944870 & -0.01943367 & 0.63841033 \end{pmatrix}$	0.000376651551
		T	$(13 \ -5 \ 3)^T$	$(12.90625477 \ -4.97099590 \ 2.99808908)^T$	0.006888649095
4	0.01	R	$\begin{pmatrix} 0.69353509 & -0.33259091 & 0.63905579 \\ 0.63905579 & 0.69353509 & -0.33259091 \\ -0.33259091 & 0.63905579 & 0.69353509 \end{pmatrix}$	$\begin{pmatrix} 0.69371420 & -0.33132705 & 0.63952029 \\ 0.63842809 & 0.69391125 & -0.33297619 \\ -0.33348405 & 0.63926381 & 0.69292045 \end{pmatrix}$	0.001117719390
		T	$(9 \ 1 \ -7)^T$	$(8.83267784 \ 1.10474038 \ -6.82978344)^T$	0.022773516445

时,对估计值与真实设定值进行比较,根据式(25)计算两者之间的误差。

$$error = \|estimate - true\| \|true\| \quad (25)$$

式中: $estimate$ ——估计出的旋转矩阵或平移向量; $true$ ——真实设定的旋转矩阵或平移向量; $\|\cdot\|$ ——2-范数。

5 结束语

本文采用神经网络的方法分别基于三维点匹配和基于二维运动场来估计三维刚体运动参数。由于神经网络具有硬件实现的可能,并且正则化的反馈网络具有很强的运算能力。权值调整使用的 Newton-Raphson 方法也可以保证收敛。从实验结果看,基于三维点匹配估计的平移向量几乎正确;旋转矩阵的估计,基于二维运动场的结果也不如基于三维点匹配的精确。但是,基于三维点匹配的方法前提要求先建立三维点之间的匹配,这并不容易。


相比较而言,基于二维运动场的方法要求获取每一帧图像序列中刚体上感兴趣点的三维坐标以及二维运动场的前提比较容易。不过这些前提都会对估计结果产生一定误差影响,相比于仿真实验的误差比较理想,真实场景的参数估计结果会有比较大的误差。总的来说,神经网络来估计运动参数具有较满意的结果。

参考文献:

- [1] 贾云得.机器视觉[M].北京:科学出版社,2002.
- [2] Goncalves N, Araujo H. Analysis and comparison of two methods for the estimation of 3D motion parameters[J]. Robotics and Autonomous Systems, 2003, 45: 23-49.
- [3] 边肇祺,张学工.模式识别[M].北京:清华大学出版社,2000.
- [4] Chen T, Lin W C, Chen C T. Artificial neural networks for 3-D motion analysis I: Rigid motion[J]. IEEE Trans Neural Networks, 1995, 6(6): 1386-1393.
- [5] Tzovaras D, Ploskas N, Strintzis M G. Rigid 3-D motion estimation using neural networks and initially estimated 2-D motion data[J]. IEEE Trans Circuits and Systems for Video Technology, 2000, 10(1): 158-165.
- [6] 陈宝林.最优化理论与算法[M].北京:清华大学出版社,2005: 285-287.
- [7] Zhang Z Y. Camera calibration with one-dimensional objects[J]. PAMI, 2004, 26(7): 892-899.
- [8] Di Kaichang, Li Rongxing. CAHVOR camera model and its photogrammetric conversion for planetary applications[J]. Journal of Geophysical Research, 2004, 109(E4): 1-9.

(上接第 3162 页)

- [10] Ester M, Grob M, Kriegl H. Focused web crawling: A generic framework for specifying the user interest and for adaptive crawling strategies[C]. Roma: Proc of the International Conference on Very Large Database(VLDB'01), 2001.
- [11] 郑冬冬, 赵朋朋, 崔志明. Deep Web 爬虫研究与设计[J]. 清华大学学报(自然科学版), 2005, 45(9): 1896-1902.
- [12] 郑毅, 吴斌, 史忠植. 基于概念空间的文本检索系统[J]. 计算机工程与应用, 2002, 38(12): 67-69.
- [13] Ehrig M, Maedche A. Ontology-focused crawling of web documents[C]. Proc of 2003 ACM Symposium on Applied Computing. New York: ACM Press, 2003: 1174-1178.
- [14] 邓志鸿, 唐世渭, 张铭, 等. Ontology 研究综述[J]. 北京大学学报(自然科学版), 2002, 38(5): 730-738.
- [15] Cognitive Science Laboratory. Wordnet: A lexical database for the English language[EB/OL]. <http://wordnet.princeton.edu/>.
- [16] 董振东, 董强. HowNet [EB/OL]. http://www.keenage.com/zhiwang/e_zhiwang.html.
- [17] 赵丰年, 刘林, 商建云. 基于概念的文本过滤模型[J]. 计算机工程与应用, 2006, 42(4): 186-188.
- [18] Diligenti M, Coetzee F M, Law Rence S, et al. Focused crawling using context graphs [C]. Cairo: Proc of the 26th International Conference on Very Large DataBases, 2000: 527-534.

作者: 刘汉兴, 刘财兴, [LIU Han-xing](#), [LIU Cai-xing](#)
作者单位: 华南农业大学, 信息学院, 广东, 广州, 510642
刊名: [计算机工程与设计](#) 
英文刊名: [COMPUTER ENGINEERING AND DESIGN](#)
年, 卷(期): 2008, 29(12)
被引用次数: 12次

参考文献(18条)

1. [Srinivasan P;Mencezer F;Pant G A general evalution framework for topical crawlers](#)[外文期刊] 2005(03)
2. [Menczer F;Pant G;Srinivasan P Topic web crawler:Evaluating adaptive algorithm](#)[外文期刊] 2004(04)
3. 李学勇;欧阳柳波;李国徽 [网络蜘蛛搜索策略比较研究](#)[期刊论文]-[计算机工程与应用](#) 2004(04)
4. 林海霞;原福永;陈金森 [一种改进的主题网络蜘蛛搜索算法](#)[期刊论文]-[计算机工程与应用](#) 2007(10)
5. [Chakrabarti S;Punera K;Subramanyam M Accelerated focused crawling through online relevance feedback](#) 2001
6. [Aggarwai C;AL-Garawi F;Yu S P Intelligent crawling on the world wide web with arbitrary Predicate](#) 2001
7. [Wu J;Aberer K Using siterank for decentralized computation of web document ranking](#)
8. [Bharat K;Chang BW;Henzinger MR Who links to whom:Mining linkage between web sites](#) 2001
9. [ET O'Neill;BF Lavoie;Bennett R Trends in the evolution of the public web](#)
10. [Ester M;Grob M;Kriegel H Focused web crawling:A generic framework for specifying the user interest and for adaptive crawling stratrgies](#) 2001
11. 郑冬冬;赵朋朋;崔志明 [Deep Web爬虫研究与设计](#)[期刊论文]-[清华大学学报\(自然科学版\)](#) 2005(09)
12. 郑毅;吴斌;史忠植 [基于概念空间的文本检索系统](#)[期刊论文]-[计算机工程与应用](#) 2002(12)
13. [Ehrig M;Maodche A Ontology-focused crawling of web documents](#) 2003
14. 邓志鸿;唐世渭;张铭 [Ontology研究综述](#)[期刊论文]-[北京大学学报\(自然科学版\)](#) 2002(05)
15. [Cognitive Science Laboratory Wordnet:A lexical database for the English language](#)
16. 董振东;董强 [HowNet](#)
17. 赵丰年;刘林;商建云 [基于概念的文本过滤模型](#)[期刊论文]-[计算机工程与应用](#) 2006(04)
18. [Diligenti M;Coetzee F M;LawRence S Focused crawling using context graphs](#) 2000

本文读者也读过(10条)

1. 汪涛, 樊孝忠, 顾益军, 刘林 [基于概念分析的主题爬虫设计](#)[期刊论文]-[北京理工大学学报](#)2004, 24(10)
2. 张翔, 周明全, 李智杰, 董丽丽, ZHANG Xiang, ZHOU Ming-quan, LI Zhi-jie, DONG Li-li [基于PageRank与Bagging的主题爬虫研究](#)[期刊论文]-[计算机工程与设计](#)2010, 31(14)
3. 刘林, 汪涛, 樊孝忠 [主题爬虫的解决方案](#)[期刊论文]-[华南理工大学学报\(自然科学版\)](#) 2003, 32(z1)
4. 罗林波, 陈绮, 吴清秀, LUO Lin-bo, CHEN Qi, WU Qing-xiu [基于Shark-Search和Hits算法的主题爬虫研究](#)[期刊论文]-[计算机技术与发展](#)2010, 20(11)
5. 陈从丛 [主题爬虫搜索策略研究](#)[学位论文]2009
6. 汪涛, 樊孝忠 [主题爬虫的设计与实现](#)[期刊论文]-[计算机应用](#)2004, 24(z1)

7. [吕赛辉](#) [主题爬虫关键技术研究及应用](#)[学位论文]2009
8. [张航](#) [主题爬虫的实现及其关键技术研究](#)[学位论文]2010
9. [刘国靖](#), [康丽](#), [罗长寿](#) [基于遗传算法的主题爬虫策略](#)[期刊论文]-[计算机应用](#)2007, 27 (z2)
10. [金明珠](#), [丁岳伟](#), [JIN Ming-zhu](#), [DING Yue-wei](#) [基于统计模型的主题爬虫的研究与实现](#)[期刊论文]-[计算机工程与设计](#)2010, 31 (16)

引证文献(14条)

1. [姜鹏](#), [宋继华](#) [一种主题爬虫文本分类器的构建](#)[期刊论文]-[中文信息学报](#) 2010 (6)
2. [张丽敏](#) [垂直搜索引擎的主题爬虫策略](#)[期刊论文]-[电脑知识与技术](#) 2010 (15)
3. [王绮卉](#), [田泽](#), [赵彬](#) [基于 HKS1553BCRT 芯片的1553B总线通信软件设计](#)[期刊论文]-[计算机技术与发展](#) 2012 (8)
4. [徐敏](#), [杨应全](#), [陈祖琴](#) [学科发展热点推荐平台的实施模式研究——以信息采集模块为例](#)[期刊论文]-[现代情报](#) 2011 (1)
5. [张福泉](#) [人工智能在主题搜索策略中的应用](#)[期刊论文]-[重庆科技学院学报（自然科学版）](#) 2009 (4)
6. [郑志高](#), [刘庆圣](#), [陈立彬](#) [基于主题网络爬虫的网络学习资源收集平台的设计](#)[期刊论文]-[中国教育信息化·高教职教](#) 2010 (1)
7. [夏小云](#), [吴为波](#) [AJAX技术的搜索引擎优化问题研究](#)[期刊论文]-[江西理工大学学报](#) 2008 (5)
8. [王芳](#), [陈海建](#) [深入解析Web主题爬虫的关键性原理](#)[期刊论文]-[微型电脑应用](#) 2011 (7)
9. [谢志妮](#) [一种新的基于概念树的主题网络爬虫方法](#)[期刊论文]-[计算机与现代化](#) 2010 (4)
10. [徐敏](#), [杨应全](#), [陈祖琴](#) [学科联盟热点推荐平台的实施模式研究](#)[期刊论文]-[情报杂志](#) 2011 (3)
11. [韩宇](#), [黄青松](#) [基于改进PageRank的情报主题相关度预测策略](#)[期刊论文]-[微型电脑应用](#) 2010 (3)
12. [赵思佳](#), [尹婷](#) [基于规则引擎的个性化主题网页爬虫的研究](#)[期刊论文]-[计算机技术与发展](#) 2011 (3)
13. [杨定中](#), [赵刚](#), [王泰](#) [网络爬虫在Web信息搜索与数据挖掘中应用](#)[期刊论文]-[计算机工程与设计](#) 2009 (24)
14. [冯硕](#), [李书琴](#), [杨会君](#) [基于Web挖掘的化学物质信息提取应用研究](#)[期刊论文]-[计算机工程与设计](#) 2012 (8)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_jsjgcysj200812053.aspx