主题搜索引擎中爬虫搜索策略的研究

史宝明[1],贺元香[1],吴崇正[2]
SHI Baoming[1], HE Yuanxiang[1], WU Chongzheng[2]

1.兰州文理学院 电子信息工程学院,兰州 730000
2.兰州理工大学 计算机与通信学院,兰州 730050
1.School of Electronics and Information Engineering, Lanzhou University of Arts and Science, Lanzhou 730000, China
2.School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China

SHI Baoming, HE Yuanxiang, WU Chongzheng. Research on search strategy of web spider in topic-oriented search engines. Computer Engineering and Applications, 2014, 50(2):116-119.

Abstract: In order to solve the low efficiency problem of traditional focused crawler, web spider always selects the most valuable links to visit, so how to focus the search around a given topic is a key problem. The traditional method always only computes the relevance of the links, but ignores the relevance among the unlabeled URL, now it proposes the algorithm based on link model which combines the seed URL with unlabeled URL to compute the relevance of the other URL, and it deduces the point that initial iterative is insensitivity of the results. Compared with the methods based on traditional algorithm, experimental result proves the performance of the new algorithm is more efficient than the traditional ones.
Key words: web spider; topic-oriented search engine; search strategy; Vector Space Model(VSM)

摘 要: 为了解决传统主题爬虫效率偏低的问题,传统主题爬虫会选择最有价值的链接进行访问,仅简单地计算链接的相关性,却忽视待分析URL之间的相关性关系,致使主题爬虫爬取效率较低。提出一种基于链接模型的相关性判别算法,综合利用有标种子URL和无标的待判别URL实现对无标URL的相关性判别,并推导出迭代初值选取对结果的不敏感性。实验结果表明,与传统的网络爬虫算法相关性判别方法相比,提出的方法效率更高。
关键词: 网络爬虫;主题搜索引擎;搜索策略;向量空间模型
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.1303-0512

1 引言

随着互联网的快速发展,信息量爆炸式增长,传统的搜索引擎在信息的覆盖率和搜索结果相关性、准确性等方面呈现下降趋势。截止2011年12月底,中国网页数量为866亿个[1],比2010年同期增长44.3%,全球数字信息总量约为1.9 ZB(1 ZB=1×1 021 GB),到2020年,全球数字信息总量将达到35 ZB,面对严峻挑战,面向主题搜索引擎应运而生,基于主题爬虫的搜索引擎研究已成为搜索和Web挖掘领域研究的一个热点和难点[2]。

主题搜索引擎是以构筑专题或学科领域的Internet信息资源库为目标,智能地在互联网上搜集符合特定主题或学科需要的信息资源,包括专业信息、特定行业领域、学科信息门户、公司信息中心、行业专家等在内的用户,提供了整套的网络信息资源服务解决方案[3]。其特点是“专、精、深”,和传统搜索引擎结果的无序性相比,主题搜索引擎获取的结果更具有针对性,按照预先定义好的主题有选择地收集相关网页,大大降低了搜集信息的难度,同时提高了搜索结果的质量。

在主题搜索引擎领域的研究,一般根据返回的结果,再经人工处理从而形成面向某一领域或学科的垂直门户网站,成功的产品有Elsevier的Scirus系统、伯克利大学的Focus Project、美国国家科学数字图书馆Collection Building Program等。文献[3]提出了基于DOM树结构的过滤器方法,结合石油行业的特点,讨论了主题搜索引

基金项目: 甘肃联合大学科研能力提升计划项目(No.2012YBTS05)。
作者简介: 史宝明(1981—),男,讲师,研究领域为智能信息处理、信息安全;贺元香(1981—),女,讲师,研究领域为智能信息处理。E-mail:zlsbm@163.com
收稿日期: 2013-04-01 **修回日期:** 2013-07-05 **文章编号:** 1002-8331(2014)02-0116-04
CNKI**网络优先出版:** 2013-08-27, http://www.cnki.net/kcms/detail/11.2127.TP.20130827.1603.012.html

擎页面预处理的方法、设计与实现；文献[4]采用向量空间模型来计算网页主题相关度，使用改进的Shark-Search网页搜索策略来决定待抓取链接的访问次序，从种子网页开始，只爬行具有较高预测相关度的链接，仅采集与主题相关的网页，多线程对网页进行下载和分析，提高了主题网页采集的精度；文献[5]提出了基于本体的主题网络爬虫解决方案；文献[6]提出基于计算链接价值度及Web页面语义主题相似度对链接分配合理权重的HITS改进算法，突出链接重要度的差异；文献[7]提出了根据改进VSM方法计算主题页面相关性，实现性能的提高；文献[8]提出了多线程网络蜘蛛的技术，使爬取效率更高；文献[9]在URL的主题相关性判别过程中引入了链接文本及相关链接属性分析，提出了URL主题相关性算法EPR算法。

还有一些研究集中在基于内容评价的主题爬虫和基于链接结构的主题爬虫上，前者是以传统信息检索模型向量空间模型为基础，利用页面中的文本信息作为领域知识指导搜索，这类算法主要有Best-First Search，Fish-Search和Shark-Search等[10]。后者主要采用基于网络链接的分析来挖掘主题信息，主要方法是利用基于PageRank算法和HITS算法来评价所要链接的主题相关性[11-12]。

以上研究有的根据网页信息评价链接，但忽视了超链接结构所包含的信息；有的通过利用超链接信息来评价链接相关性，但忽视了文本的信息。这两者都仅考虑了种子URL与待分析URL之间相关性的判断，却忽视了待分析URL之间的相关性关系[13]，本文在深入分析主题爬虫的工作原理和主题相关性判别方法的基础上，提出一种基于链接模型的相关性判别算法，综合利用了有标种子URL和无标的待判别URL实现对无标URL的相关性判别，此方法将大大提高主题页面相关性判别的效率和准确度，同时推导出迭代初值选取对结果的不敏感性，从而实现主题搜索引擎搜索性能的提高。

2　主题爬虫相关技术

2.1　主题爬虫工作原理

主题网络爬虫是基于Http协议自动提取网页的应用程序，是主题搜索引擎的重要组成部分，围绕既定的主题从网络上下载相应的网页，利用一个或多个种子URL来爬取更多的网络资源，根据主题信息，下载相关网络资源。

主题爬虫采用多线程技术，将线程分配给种子URL，对采集到的网页，根据数据库记录判断该URL是否处理过，假若没有，则下载对应的网页，要分析并尽可能多地提取其中的链接，插入到URL队列中，并分析网页文本信息，提取其中的特征项，根据相关性分析算法计算网页的主题相关度，如符合条件，则将该网页存储到网页库中。以此方式循环，直到URL队列为空[14]，如图1所示。

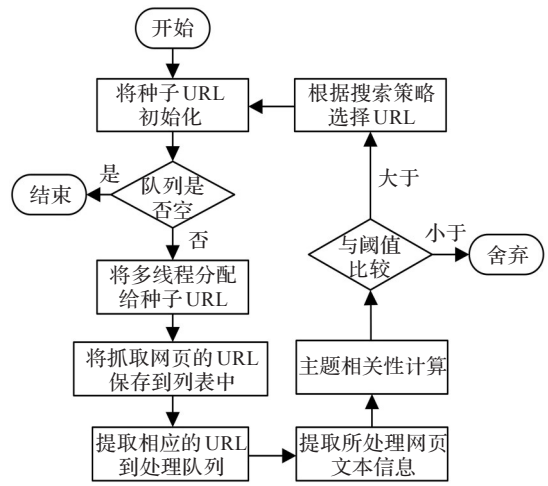

图1　主题爬虫工作流程图

2.2　页面主题相关性判断

从根本上讲，网络爬虫的搜索策略是多目标规划的问题，在有限的时间范围内，以较少的网络资源、存储资源和计算资源获得更多的主题相关页面。网络爬虫研究的核心是解决页面和URL的主题相关性判别问题，因此，如何评价链接价值是决定此类网络爬虫爬行效率的关键。

在主题爬虫爬取相关网页时，所提取的URL需要通过主题相关性的判别，这也是主题搜索引擎与传统搜索引擎最大的区别所在。如何将与主题相关性不大的网页快速地过滤掉，是决定主题爬虫搜索策略的关键，故在提取页面文本信息后，需对页面进行主题相关性判断，以过滤掉与主题无关的网页，保留与主题相关性高的网页。

现阶段的研究大多采用向量空间模型作为页面主题相关性判别的方法，该方法将任意页面文档映射成VSM中的一个特征向量，按照VSM模型的匹配度量方法，通过计算采集页面特征向量与主题中心向量的内积来对采集页面进行评价，内积越大则说明采集页面的主题相关度越高[15]，计算公式如公式(1)所示。

$$Sim(D_i, D_j) = \frac{\sum_{k=1}^{N} w_{ik} \times w_{jk}}{\sqrt{\sum_{k=1}^{N} w_{ik}^2} \times \sqrt{\sum_{k=1}^{N} w_{jk}^2}} \tag{1}$$

其中，D_i为主题，D_j为待判别的页面，分别由N维向量$<d_{i1}, d_{i2}, \cdots, d_{in}>$和向量$<d_{j1}, d_{j2}, \cdots, d_{jn}>$表示，其中，$w_{ik}, w_{jk}$分别对应$D_i$, D_j页面中第k个词汇的权重，w_{ik}, w_{jk}通过TF-IDF方法量化处理。令$sim(D_i, D_j)$的值与阈值比较，若前者大于等于阈值，则表示页面与主题

相关,保留到数据库中,否则判为不相关,丢弃该网页。

向量空间模型的优点在于它简化了页面相关度的计算复杂度,使之转化为了向量空间运算,但其本身也有一些不足,利用向量空间模型来计算时,理想状态下,向量中的各特征项应该相互独立,既要体现页面文档内容,又要体现页面之间的差异性,而实际并非如此。另外向量的维数如果很高,会使计算的复杂度增大,那么将降低网页相关性评价效率。所以在此欲结合基于链接模型的相关性判别算法,既考虑页面之间的主题相关度,同时在计算链接相关性的同时考虑待分析URL之间的相关性关系,综合利用有标种子URL和无标的待判别URL实现对无标URL的相关性判别。

3 基于链接模型的URL相关性判别

3.1 参数定义

在主题搜索引擎中主题爬虫抓取网页URL时,根据种子URL来爬取更多的URL,将种子URL集合向量设为 $\boldsymbol{S}=\{s_1,s_2,\cdots,s_n\}$,其向量为 $\boldsymbol{X}_s=\{x_1,x_2,\cdots,x_n\}$;待分析URL向量为 $\boldsymbol{T}=\{t_{n+1},t_{n+2},\cdots,t_{n+m}\}$。当待分析URL与主题相关时,$x_i=1$,反之,$x_i=-1$。在此既要依赖种子URL提供的相关性信息对待分析URL的相关性进行判断,同时又要考虑到具有相似主题的待分析URL之间通常具有较高的向量内积,主题相关性不强的URL之间的向量内积也会较低,所以待分析URL之间的主题相似关系对于判断页面主题相关性将会有一定的作用。

3.2 算法描述

主题爬虫首先初始化种子URL,然后根据这些种子URL再爬取更多的页面,根据这些页面URL之间的链接关系,它们就构成了一个张图 G,定义图 $G=<N,M>$,其中 N 包括种子URL和待分析的URL,为所有结点的集合,令 $|N|=|S|+|T|$,其中 $|S|$ 设为种子URL的总数量,$|T|$ 设为待分析URL总数量,那么 $|T|\times|N|$ 链接矩阵 $\boldsymbol{M}$ 则描述了种子结点与待分析结点和待分析结点之间的链接关系,设 M_{ij} 为结点 i 与结点 j 间的主题相似度,根据以上所设 $\boldsymbol{M}$ 可分解为两部分:子矩阵 $\boldsymbol{U}=|T|\times|S|$ 和子矩阵 $\boldsymbol{V}=|T|\times|S|$,其中 U_{ij} 表示待分析 URL_i 与种子 U_j 之间的主题相似度,V_{ij} 表示待分析 URL_i 与待分析 URL_j 之间的主题相似度,则归纳链接模型的主题相似度判别算法的迭代公式如公式(2)所示。

$$X_t^{(n)}=(1-\gamma)\boldsymbol{U}X_s+\gamma\boldsymbol{V}X_t^{(n-1)} \tag{2}$$

其中,$X_t^{(n)}$ 为第 n 次迭代后的 X_t,γ 为加权系数,$0<\gamma<1$。

在实际的情况中,可能存在一些与待分析URL链接权重较低的结点,则这些结点给予待分析URL的支持是不可靠的,同时为了保证迭代过程收敛,需对 $\boldsymbol{U}$、$\boldsymbol{V}$ 和 $\boldsymbol{X}$ 做归一化处理,经分析,给出具体算法流程描述如下。

(1)首先构建链接 $\boldsymbol{M}$。其中,链接权重利用向量空间模型方法计算出结点间的主题相似度。

(2)然后计算 $k(V_{i*}),1<i<|T|$。设 $k(V_{i*})$ 为 V 中第 i 行各元素按值由大到小排序,排名在第 k 位元素的值。

(3)在矩阵 $\boldsymbol{V}$ 中,对于任意 i 和 j,若存在 $V_{ij}\leqslant k(V_{i*})$,那么 $V_{ij}=0$;否则,V_{ij} 保持不变。

(4)同时对 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 进行联合归一化,如公式(3)所示。

$$\begin{cases}U_{ij}=\dfrac{U_{ij}}{\sum\limits_{j=1}^{|S|}U_{ij}+\sum\limits_{j=1}^{|T|}V_{ij}}\\ V_{ij}=\dfrac{V_{ij}}{\sum\limits_{j=1}^{|S|}U_{ij}+\sum\limits_{j=1}^{|T|}V_{ij}}\end{cases} \tag{3}$$

(5)对 X_s 进行归一化处理,以保证 X_s 中主题相关的种子URL分值和为1,主题不相关的分值和为 -1。

$$x_i=\begin{cases}-\dfrac{x_i}{\sum\limits_{i=1}^{|S|}x_i},x_i<0\\ \dfrac{x_i}{\sum\limits_{i=1}^{|S|}x_i},x_i>0\end{cases},i=1,2,\cdots,|S| \tag{4}$$

(6)选取迭代初值 $X_t^{(0)}$,根据公式(2)计算 X_t 不断地迭代,最后矩阵内向量收敛于某个值。

(7)同时要对 X_t 进行如下归一化:

$$x_i=\begin{cases}-\dfrac{x_i}{\sum\limits_{i=1+|S|}^{|S|+|T|}x_i},x_i<0\\ \dfrac{x_i}{\sum\limits_{i=1+|S|}^{|S|+|T|}x_i},x_i>0\end{cases} \tag{5}$$

其中 $i=|S|+1,|S|+2,\cdots,|S|+|T|$,这样就保证了 X_t 中与主题相关的URL分值和为1,与主题不相关的URL分值和为 -1,转到步骤(6),直到 X_t 的值收敛为止。

(8)依据 X_t 的值对待分析URL的主题相关性做出判断。

$$t_i=\begin{cases}\text{相关},x_i>d\\ \text{不相关},x_i<d\end{cases} \tag{6}$$

其中 $i=|S|+1,|S|+2,\cdots,|S|+|T|$,则 X_i 的绝对值大小就表示待分析URL与主题相关的强烈程度。

3.3 迭代初值选取对收敛无关性证明

根据3.2节中算法描述中(6)所述,迭代初值 $X_t^{(0)}$ 的选取对收敛结果的是否会有影响呢?本节主要讨论这个问题。下面通过数学推理证明来讨论,证明过程如下所示。

证明 由 $X_t^{(n)}=(1-\gamma)\boldsymbol{U}X_s+\gamma\boldsymbol{V}X_t^{(n-1)}$ 可知,令 $(1-\gamma)\cdot$

$UX_s=\alpha$，则 $\Rightarrow X_t^{(n)}=(\gamma V)^n X_t^{(0)}+\sum_{i=1}^{n}(\gamma V)^{i-1}\alpha$，那么 $\lim_{n\to\infty}X_t^{(n)}=\lim_{n\to\infty}\left[(\gamma V)^n X_t^{(0)}+\sum_{i=1}^{n}(\gamma V)^{i-1}\alpha\right]$。

由3.2节的定义可知，V 中所有元素都大于0，又因为对于 U 和 V 进行联合行规范化，由此可得：

$$\sum_j(\gamma V)_{ij}^n=\sum_j\sum_k(\gamma V)_{ik}^{n-1}(\gamma V)_{kj}=\sum_k(\gamma V)_{ik}^{n-1}\sum_j(\gamma V)_{kj}<\gamma\sum_k(\gamma V)_{ik}^{n-1}$$

所以，

$$\sum_j(\gamma V)_{ij}^n<\gamma^n,\ \forall i=1,2,\cdots,|T|$$

$$\lim_{n\to\infty}(\gamma V)^n X_t^{(0)}=0$$

因此，序列 $X_t^{(n)}$ 是收敛的，即 $X_t^{(0)}$ 初值不管选哪些值，都不会影响收敛性。

4　实验与结果分析

4.1　实验设置

为验证本文所提方法的性能，以Best First Search算法及Shark Search算法作参照，从查准率、查全率两个方面，比较本文所提算法。实验平台以Heritrix 3.10作为测试的基础框架。以“计算机”为主题，选取计算机主题的网站15个，同时选取与计算机主题无关的25个网站作为实验的测试集，通过扩展Extractor组件来实现网页解析及主题相关度计算，实现对抓取的网页进行主题相关性判断；同时扩展FrontierSchedular组件来实现URL搜索策略，不断地更新URL队列直至抓取过程结束。

设定种子URL数为30个，其中包括与计算机主题相关的索引和部分不相关主题的索引，测试网页总数设为4 000个，分别统计获取页面的数量在500至4 000时获取与主题相关的页面数量，根据所得实验数据，计算查准率及查全率，比较基于Best First Search、Shark Search及本文所提方法3种爬虫实验结果。为防止搜索的规模过大造成系统崩溃，设定爬取深度为3；将 $X_t^{(0)}$ 初始化为零向量，便于计算，设置阈值为 $d=0$。根据实验知 γ 在0.6~0.7范围内选取，算法会显示其较高的性能，既考虑了种子URL与待分析之间的主题相关性判断，又考虑了待分析URL之间的主题相关性，在此 γ 选值为0.68。

4.2　实验结果

分别统计三种算法获取页面数量在500、1 000、1 500、2 000、2 500、3 000、3 500、4 000时获取与主题相关页面的个数，得到三种算法在查全率与查准率方面的实验结果对比图，如图2，图3所示。

由图2和图3可知，在查准率和查全率两方面，可以看出本文所提出的基于链接模型的URL相关性判别算法，当 γ 取0.68时，要比Best First Search和Shark Search算法有一定的优势，对比算法均只考虑种子URL与待分析之间的主题相关性判断，而没有考虑待分析URL之间的主题相关性判别，这样就可能会降低链接间的主题相关性判别，从而导致爬虫效率的降低，由此可见综合考虑有标种子URL和无标的待判别URL实现对无标URL的相关性判别，对于主题爬虫性能的提高具有一定提高。

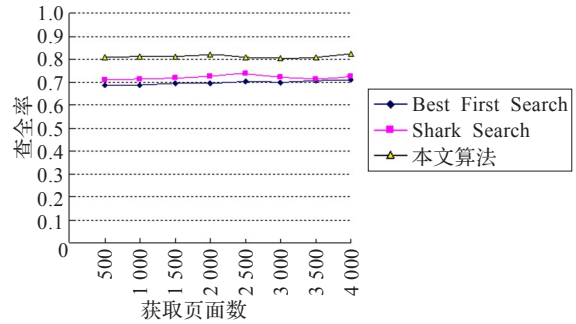

图2　3种算法查全率实验结果对比图

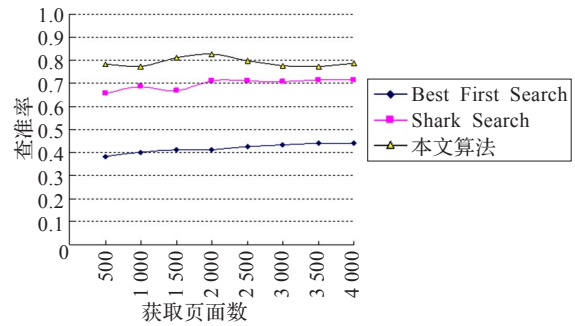

图3　3种算法查准率实验结果对比图

5　结束语

本文提出了一种基于链接模型的URL主题相关性判别方法，不仅考察了种子URL与待分析URL之间的主题相关性，而且引入了待分析URL之间的主题相关性判别，实验结果证明，在主题相关性准确率判别方面，所提方法有一定的可行性，但这组实验是在小量数据集基础上进行的，如果在实际的环境中，其算法的复杂度也要被考虑进来；基于链接和页面文本信息的主题相似度计算不能更为准确，必须采用新的方法将关键词映射到语义概念一级，从词的语义层次对页面文本信息进行主题相关性分析，同时所提算法中对图中结点的链接权重是通过VSM方法计算得到，而链接权重可以通过监督和半监督方法得到，以上所述将是进一步研究的方向。

参考文献：

[1] 中国互联网络发展状况报告[EB/OL].[2012-10-15].http://www.Cnnic.cn/research/bgxz/tjbg/201201/t20120116_23668.html.

[2] 方加沛，黄战.基于单类别文档分类的主题爬虫[J].计算工程与应用，2010，46(16)：63-66.

（下转128页）

对比来分析IMM算法的粗分召回率性能。

以句子为切分单元,定义粗分召回率γ=正确切分句子数/切分句子总数×100%,由于IMM算法保留了句子中的歧义字段,因此称句子的切分结果为正确切分结果,若该结果满足以下任一条件:

(1)切分结果与语料库标准切分结果一致。

(2)在不满足条件(1)时,通过切分结果中若干词的组合可还原正确结果,如“尉健行”被切分为“尉/健/行”,与正确切分结果不同,但该切分结果的组合可以得到正确结果,故为正确切分,而“尉健行李岚清”被切分为“尉/健/行李/岚/清”,无法通过若干词组合得到正确结果,故为错误切分[10]。

(3)在不满足条件(1)时,通过切分结果中单个词的拆分可以得到正确结果。如“奏凯歌颂伟业”切分结果为“奏/凯歌颂/伟业”,虽然“颂伟业”与正确结果不一致,但通过该单个词的拆分可得到正确结果“凯歌/颂”,故为正确切分。

通过对人民日报语料(含125 498个句子)进行测试,得到125 369个正确切分结果,即IMM算法的粗分召回率为99.90%,相当于文献[10]中8-最短路径的召回率,可见IMM算法的粗分召回率也是令人十分满意的。

7 结束语

本文在最大匹配算法的基础上,提出了一种新的中文分词粗分方法,通过引入广义词的概念,将词的范围从词表中已收录词扩展至交叉型歧义词条,并引入汉字序列诱导词集的概念用于最长广义词的匹配和判定。本算法不仅能够正确切分无歧义的语句,也可以准确检测并标记所有的交叉型歧义,最大程度上简化后续歧义消解过程,而且算法执行效率比较高,算法复杂度为$O(N)$。IMM算法已用于某公司的舆情监控系统,能够满足大规模网络文本语料处理的速度需求,且处理效果也非常令人满意。但为快速高效地完成整个汉语分词过程,IMM算法仍需要歧义消解算法的配合,因此,高准确率的歧义消解算法将是下一步工作的研究重点。

参考文献:

[1] 付年钧,彭昌水,王慰.中文分词技术及其实现[J].软件导刊,2011,10(1):18-20.

[2] 赵伟,戴新宇,尹存燕,等.一种规则与统计相结合的汉语分词方法[J].计算机应用研究,2004,21(3):23-25.

[3] 张劲松,袁健.回溯正向匹配中文分词算法[J].计算机工程与应用,2009,45(22):132-134.

[4] 金在全,赵照,杜秀全,等.一种改进的增字最大匹配算法[J].科学技术与工程,2007,7(18):4761-4764.

[5] 孙茂松,黄昌宁,邹嘉彦,等.利用汉语二元语法关系解决汉语自动分词中的交集型歧义[J].计算机研究与发展,1997,34(5):332-339.

[6] 何国斌,赵晶璐.基于最大匹配的中文分词概率算法研究[J].计算机工程,2010,36(5):173-175.

[7] 张玉茹.中文分词算法之最大匹配算法的研究[J].现代计算机:专业版,2011(19):24-26.

[8] 闻玉彪,贾时银,邓世昆,等.一种改进的最大匹配中文分词算法[J].计算机技术与发展,2011,21(10):91-94.

[9] 王瑞雷,栾静,潘晓花,等.一种改进的中文分词正向最大匹配算法[J].计算机应用与软件,2011,28(3):195-197.

[10] 张华平,刘群.基于N-最短路径方法的中文词语粗分模型[J].中文信息学报,2002,16(5):1-7.

(上接119页)

[3] 梁党卫,彭文滔.垂直搜索引擎中过滤器的设计与实现[J].计算机应用与软件,2009,26(12):148-151.

[4] 张博,蔡皖东.面向主题的网络蜘蛛技术研究及系统实现[J].微电子学与计算机,2009,26(5):52-55.

[5] 戚欣.基于本体的主题网络爬虫设计[J].武汉理工大学学报,2009,31(3):138-141.

[6] 吕林涛,陈丽萍.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.

[7] 施佺,王恒山.面向主题的垂直搜索引擎系统的研究与实现[J].微电子学与计算机,2011,28(7):1-5.

[8] 王明国,胡敬仓.主题搜索引擎中网络蜘蛛搜索策略的研究[J].微处理机,2011(4):34-37.

[9] 李勇,韩亮.主题搜索引擎中网络爬虫的搜索策略研究[J].计算机工程与科学,2008,30(3):4-6.

[10] 汲业,陈燕.生活服务领域垂直搜索引擎的设计与实现[J].计算机工程,2010,36(24):24-26.

[11] 叶育鑫,欧阳丹彤.基于语义的主题爬行策略[J].软件学报,2011,22(9):2075-2088.

[12] Bussche F.Not so creepy crawler:easy crawler generation with standard XML queries[C]//Proceedings of the 19th International Conference on WWW,Raleigh,North Carolina,USA,2010:1305-1308.

[13] 丁军平,蔡皖东.面向P2P特定信息的爬虫改进技术[J].计算机工程与应用,2011,47(29):23-26.

[14] Patel A.An adaptive updating topic specific web search system using T-Graph[J].Journal of Computer Science,2010,6(4):450-456.

[15] Barbosa L,Freire J,Taylor R C.An adaptive crawler for locating hidden web entry points[C]//Proceedings of the 18th International Conference on World Wide Web,Madrid,Spain,2009:681-697.