文章编号:1672 - 058X(2012)02 - 0066 - 07

基于 HMM 的主题爬虫研究

谢治军[1], 杨 武[2], 李稚楹[1], 宋静静[1]
(1. 重庆理工大学 计算机科学与工程学院,重庆 400054;2. 重庆理工大学 信息与教育技术中心,重庆 400054)

摘 要:主题爬虫是垂直搜索引擎的核心组成部分,它为面向主题的用户查询准备数据资源;提出了一种基于 HMM 的主题爬虫方法,方法不仅分析网页内容,而且还考虑网页的上下文链接结构,首先将当前网页的聚类结果作为观察状态、将当前网页到目标网页的链接距离作为隐含状态,然后通过 HMM 模型学习用户的主题浏览模式并利用它采集更多的主题网页;实验结果表明:方法能采集大量与指定主题相关的高质量网页,主题爬行效率优于 Best-First 主题爬虫。

关键词:主题爬虫;隐马尔科夫模型;向量空间模型;主题相关度;垂直搜索引擎
中图分类号:TP391 **文献标志码**: A

随着网络技术的飞速发展,万维网成为海量信息的载体,如何有效地提取并利用这些信息成为一个巨大的挑战,搜索引擎作为信息检索的工具已成为用户访问万维网的入口和指南[1]。但是,目前传统的搜索引擎正面临着索引规模、更新速度和个性化需求等多方面的挑战;对于多数用户提出的与某一主题或领域相关的查询需求,传统的搜索引擎往往不能返回令用户满意的结果,因此,适应特定主题和个性化搜索的主题爬虫便应运而生[2]。

1 主题爬虫工作原理及相关研究

网络爬虫(也称网络机器人、网络蜘蛛)是一个自动提取网页的程序,它为搜索引擎从 Web 上下载网页,是搜索引擎的重要组成部分。传统网络爬虫从一个或若干初始网页的 URL 开始,获得初始网页上的 URL 列表,在抓取网页的过程中,不断从当前页面上抽取新的 URL 放入待爬行队列,直到满足系统的停止条件[1,3]。主题爬虫(也称聚焦爬虫)是根据一定的网页分析算法过滤与主题无关的链接,保留与主题相关的链接并将其放入待抓取的 URL 队列中,然后根据一定的搜索策略从队列中选择下一步要抓取的网页 URL,并重复上述过程,直到达到系统的某一条件时停止[2,4]。主题爬虫是一种特殊的网络爬虫,其主要目标是在有限的时间与网络带宽限制下尽可能多地采集与指定主题相关的高质量网页,忽略与主题无关或低质量的网页[5]。相对于传统网络爬虫,主题爬虫加入相关准则用于采集与主题或领域有关的网页信息,它既可以提高现有查询的精度,降低爬虫对网络资源的占用,也可以缩短网页数据库更新的周期[6]。传统网络爬虫和主题爬虫的工作流程,如图 1 所示[1]。

到目前为止,研究者已经提出了多种主题爬虫搜索策略,其中主要包括:(1) 基于文字内容的启发式方

收稿日期:2011 - 05 - 25;**修回日期**:2011 - 06 - 27.
作者简介:谢治军(1984-),男,重庆市人,硕士研究生,从事信息检索研究.

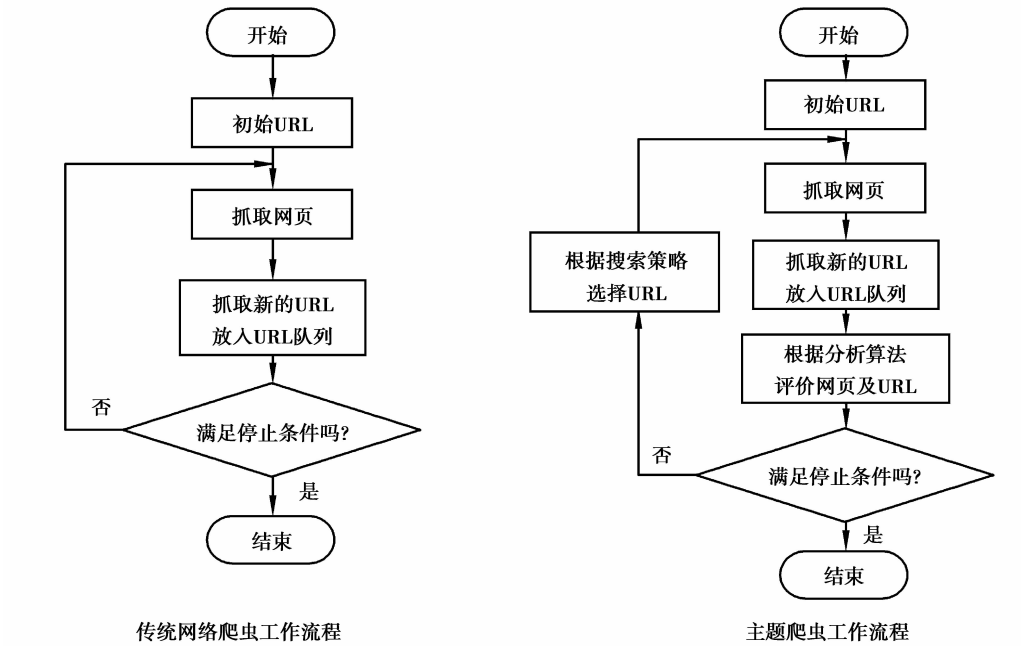

图 1　传统网络爬虫与主题爬虫工作流程

法,如 Fish Search 方法[7]、Shark Search 方法[8]、Best-First Search 方法[9]等;(2) 基于 Web 超链接图的方法,如 PageRank 方法[10]等;(3) 基于文本分类器的方法,如基于朴素贝叶斯分类器的主题爬虫[11]、基于支持向量机(SVM)分类器的主题爬虫[12]等。由此可见,主题爬虫作为垂直搜索引擎的核心组成部分,近年来得到了越来越多的研究人员的关注和研究。采用了一种基于用户浏览模式学习的 HMM 主题爬虫,首先是构建隐马尔科夫模型[13],将当前网页的聚类结果作为观察状态、将当前网页到目标网页的链接距离作为隐含状态,然后通过 Baum-Welch 学习算法学习用户的主题浏览模式,并利用模式来指导主题爬虫的爬行。

2　系统结构与算法设计

2.1　HMM 主题爬虫系统结构

一般认为主题内容相关的网页往往包含着指向相关主题网页的链接。然而,一些与主题无关的网页也同样存在着指向与主题高度相关网页的链接,正是由于这种情况,造成了一些主题爬虫丢失采集更多主题网页的机会。此外,在一些站点的网页结构布局中,为了给用户提供符合其浏览习惯的操作体验,主题相似网站的站点管理员通常将其站点结构按照主题层次式的结构对站内资源进行组织布局,用户就可以在其期望的站点位置中寻找到与主题相关的网页。因此,提出了一种基于用户浏览模式学习的 HMM 主题爬虫,HMM 模型的由以下几部分组成:

(1) 隐含状态:$S=\{T_0,T_1,\cdots,T_{n-1}\}$,其中 T_i 代表当前页面到达主题页面的距离为 i,当时 $i=0$,页面是主题页面。

(2) 观察值集合:$O=\{O_1O_2,\cdots O_m\}$,其中 O_i 表示页面所隶属的模式为类别 i。

(3) 初始状态:$\pi=\{P(T_0),P(T_1),\cdots P(T_{n-1})\}$,其中 π_i 表示初始情况下到达一个主题目标页面的距离为 i 的概率,这里的初值采用一般的均匀分布。

(4) 转移概率矩阵:$A=[a_{aj}]_{n\times n}$,a_{ij}表示由 t 时刻 T_i 状态转移到 $t+1$ 时刻 T_j 状态的概率。

(5) 发射概率矩阵:$B=[b_{ij}]_{n\times m}$,b_{ij}表示 T_i 状态生成观察值 O_j 的概率。

HMM 主题爬虫系统主要由两大模块组成,即用户浏览模式学习模块和主题爬行模块,系统结构见图2。

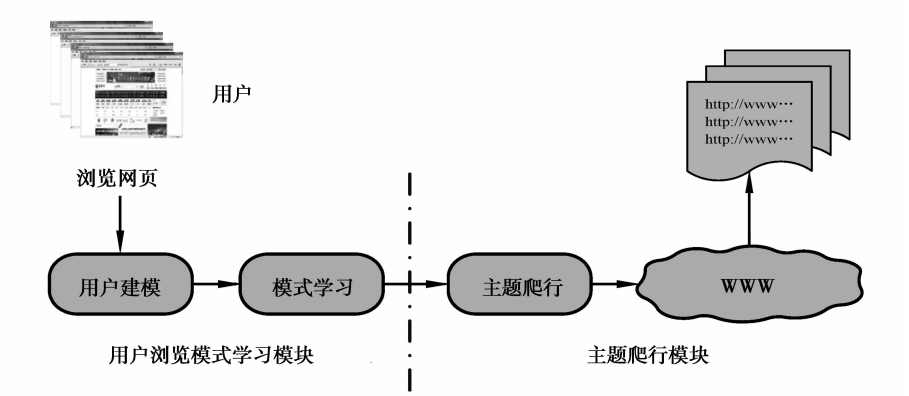

图2 系统结构

2.2 用户浏览网页的搜集

在此阶段,目的是收集用户对特定主题网页浏览访问的内容和序列。如果用户发现当前浏览网页与主题相关,那么将网页标注为目标网页。为了分析用户的浏览模式,构造出了一个既包括网页内容又包括链接结构的 Web 图,如图3所示,节点代表网页,边代表一个网页指向它引用的网页,红色节点代表目标网页。

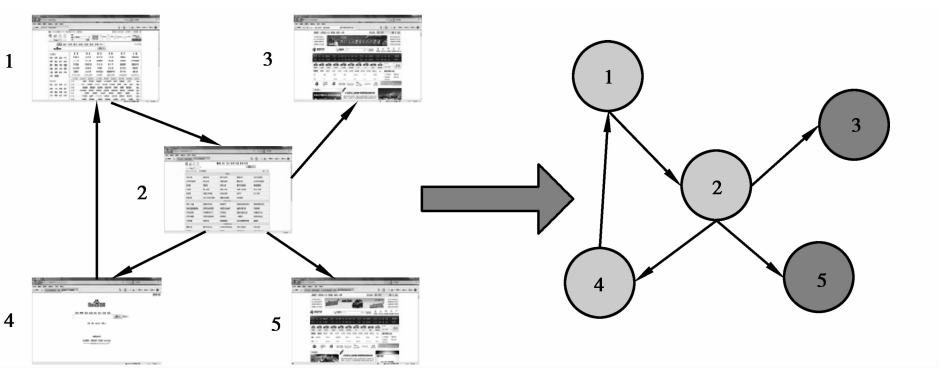

图3 用户浏览模式

例如将"汽车"作为一个主题,第1步:浏览网址 http://www.hao123.com/;第2步:浏览网址 http://www.sina.com.cn/,第3步:浏览网址 http://auto.sina.com.cn/;第4步:浏览网址 http://sports.sina.com.cn/nba/video/;第5步:由第4步后退,浏览网址 http://data.auto.sina.com.cn/carstyles/? size=a0,其中网页3和网页5,是用户感兴趣的网页,即目标网页,网页1、网页2、网页4是与主题无关的网页。然后根据用户的浏览顺序,得到用户查找主题信息的路径:网页1→网页2→网页3→网页4→网页5。

2.3 用户浏览模式学习

在查找主题目标网页的路径中,浏览的许多网页都是用户根本不在意的信息,那么就应该将这些要求"告知"HMM 模型,对其进行训练,学习用户的主题浏览模式,使其具备准确预测目标网页的能力。将收集的用户浏览网页作为训练集,按照 tf-idf 的方式构建每个网页的向量空间模型(VSM),然后将目标网页作为聚类中心,其他网页按照均值方法聚类。得出一个观察序列:网页1(C2)→网页2(C3)→网页3(C0)→网页4(C1)→网页5(C0),如图4所示。

根据观察序列,使用 Baum-Welch 算法对 HMM 的参数进行优化,初始时刻先给出经验估计值,通过不断地迭代,使各个参数逐渐向更为合理的优化值。

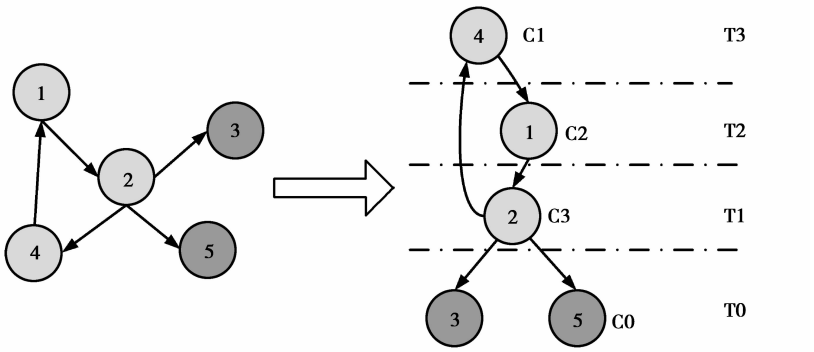

图 4　HMM 参数估计

2.4　主题相似度分析

主题爬虫需要对已抓取的页面进行分析,判断其是否与主题相关,这里采用向量空间模型的方法。网页内容的相关度计算相关度计算分为 3 步:一是确定一组特定的主题向量;二是网页向量空间模型的提取;三是计算网页与主题之间的相关度。即主题爬虫抓取到网页 P,经过预处理及中文分词后,得到 P 的向量空间模型,然后利用余弦相似度函数,计算网页 P 与主题之间的相似度。

$$\cos(p,q) = \frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2 \sum_i d_i^2}} \tag{1}$$

其中,q_i 是主题的向量空间模型权重;d_i 是网页 P 预处理后的向量空间模型权重,其计算公式如下所示:

$$d_i = \underbrace{\frac{f_i}{\max f_d}}_{tf_i} \underbrace{\log \frac{N}{N_i}}_{idf_i} \tag{2}$$

上式中 tf_i 表示词条 i 在文档 P 中的频率,idf_i 表示词条 i 在文档 P 中的逆文档频率,f_i 表示词条 i 在文档 P 中的次数,f_d 表示文档 P 的所有词的次数。N 表示是所有文档数,N_i 表示是所有文档数中出现了词条 i 的次数。

当 $\cos(p,q)$ 大于等于 δ 时,则说明网页与主题相关,当 $\cos(p,q)$ 小于 δ 时,则说明网页与主题无关,下面的公式描述了页面与主题相关性的判断。

$$R(p,q) = \begin{cases} 1 & \cos(p,q) \geqslant \delta \\ 0 & \cos(p,q) < \delta \end{cases} \tag{3}$$

2.5　URL 优先级的确定

在爬行阶段,主题爬虫从待下载队列里按优先级抽取 URL,然后下载网页,提取预处理后网页的向量空间模型,利用 HMM 参数预测其链接指向目标网页的可能性。

$a(L_i,t)(j=0,1,\cdots,n)$ 代表主题爬虫在时刻 t,隐含状态为 L_j 的预测值,$a(L_j,t-1)$ 是 $a(L_j,t)$ 父网页的预测值,$a(L_j,t)$ 通过下面的迭代公式得到:

$$a(L_j,t) = b_{jc_t} \sum_{i=0}^{\text{states}} (a(L_i,t-1) * a_{ij}) \tag{4}$$

其中,a_{ij} 是状态转移矩阵从状态 L_i 到 L_j 的概率,b_{jc_t} 是发射矩阵中观察值 c_t 在 L_j 状态的概率,$a(L_j,0)$ 是初始概率 π 的值,$a(L_j,t+1)$ 的预测值可以通过下面得迭代公式得到:

$$a(L_j,t+1) = \sum_{i=0}^{\text{states}} (a(L_i,t) * a_{ij}) \tag{5}$$

最后,选取 $\max a(L_j,t+1)(j=0,1,\cdots,n)$ 作为当前网页的链接指向目标网页的预测值。

$$\text{priority}_{\text{HMM}}(p) = \max a(L_j, t+1)(j = 0,1,\cdots,n) \tag{6}$$

2.6 HMM 主题爬虫算法

HMM 主题爬虫主要负责网页抓取和 URL 过滤,它从一个候选 URL 列表中按照优先级高低来选择 URL 抓取网页,如果不是重复下载,则对该网页进行分析,如果与主题相关,则将其保存到主题网页库中并指定其子链接的优先级,它的工作流程如图 5 所示。

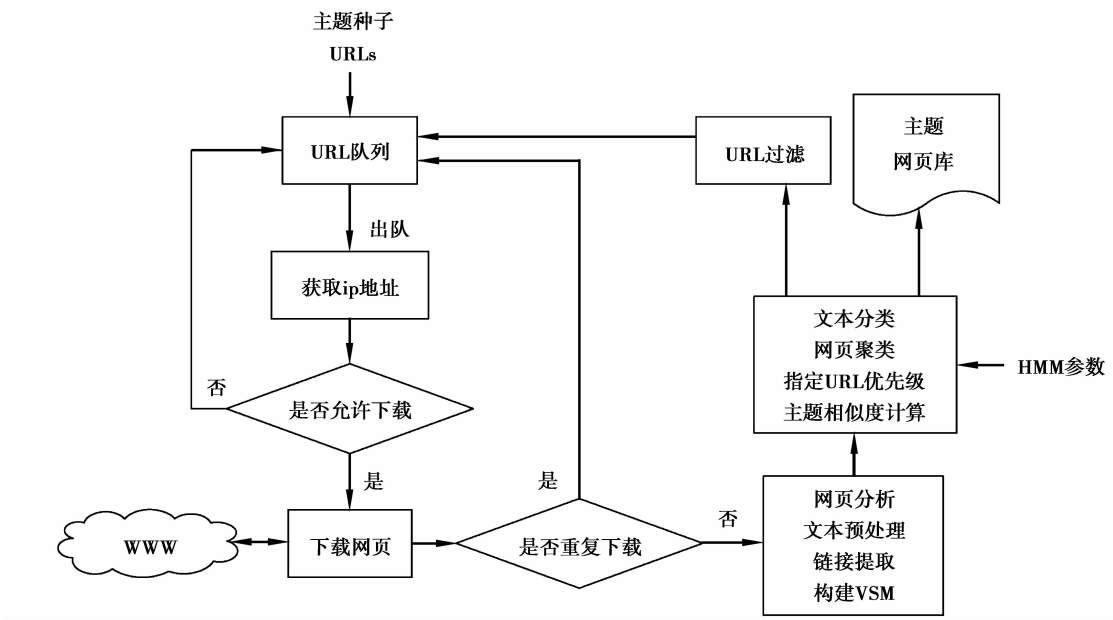

图 5 HMM 主题爬虫流程

HMM 主题爬虫的核心算法如下:

```
Focused_Crawler(HMM,δ,n)
    urlQueue: = {seed URLs};
    while (not(termination)) do
      p_t: = dequeue head of urlQueue;
      Extract from p_t its URL, parent cluster c_{t-1}, and
          a(j,t-1) for all possible states j;
      Download content of p_t;
      Parse and classify page p_t to cluster c_t;
      If cos(p_t, T arg etSet) > δ, Then store p_t as relevant;
      Calculate a(j,t) and priority for p_t's children
        If p_t is start Url,
          Then a(j,t) = π(i)b_ij;
          Else a(j,t) = ∑_i (a(i,t)) * a_ij) b_ij;
        Calculate prediction a(j,t+1) = ∑_i (a(i,t) * a_ij);
        Calculate the visit priorit y_HMM(p) = max a(j,t+1);
      For each outlink w_{t+1} of w_t with url (has the same prioprity)
        urlQueueEntry: = priority(priority, url, C_t, a(j,t));
        Enqueue(urlQueueEntry);
```

3　实验结果

系统的实现采用了 Java 语言以及集成开发环境 Eclipse,运用了开源 HMM 工具包 Jahmm 及中国科学院中文分词系统 java 版。主要通过主题爬虫采集的主题相关页面数与采集的所有页面数的比,即查准率,来评价系统的主题采集效率。

实验以“汽车”作为特定主题,收集了 100 个用户主题浏览的网页作为训练集,在主题爬行阶段,选取作为种子 URL,主题与网页的相似度阈值 δ 取 0.7。系统运行在一台装有 Windows XP 操作系统的 PC 上,内存容量为 2 GB,硬盘大小为 320 GB。为了验证基于 HMM 主题爬虫的主题爬行效率,选择了与 Best-First 主题爬虫进行了对比,实验结果如图 6 所示。

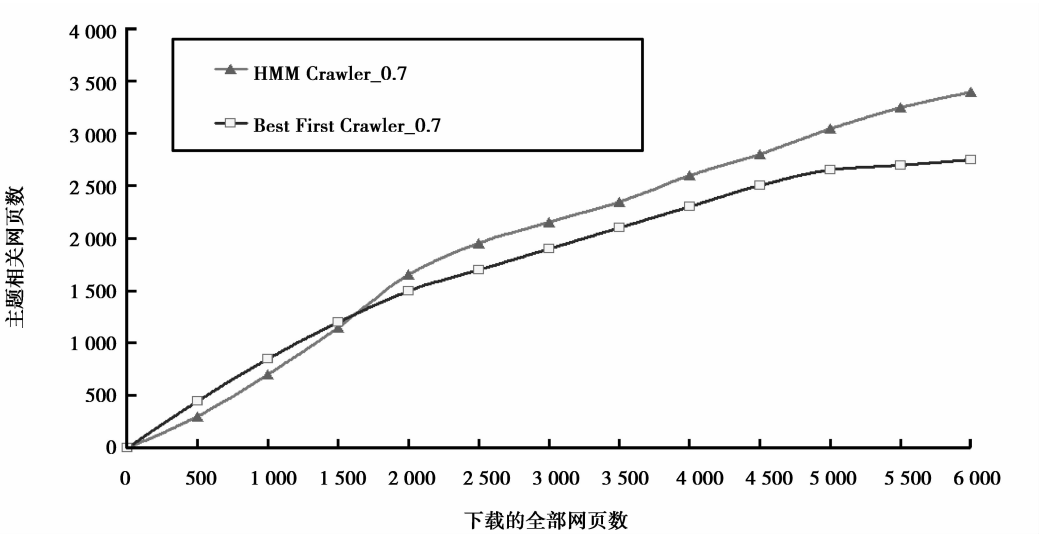

图 6　HMM 主题爬虫与 Best-First 主题爬虫对比

从图 6 可以看出,在爬行初期,基于 HMM 主题爬虫的查准率低于 Best-First 主题爬虫的查准率,但是在之后的爬行过程中,基于 HMM 主题爬虫的查准率高于 Best-First 主题爬的查准率。实验结果表明,对于采集大规模的主题网页时,相对于 Best-First 主题爬虫,基于 HMM 主题爬虫能够抓取更多的主题相关网页,主题爬行效率高于 Best-First 主题爬虫。

4　结束语

通过对主题爬虫工作原理的深入研究,针对一些主题爬虫存在的缺陷,实现了一个基于 HMM 的主题爬虫,方法通过训练集学习用户的主题网页浏览模式,然后利用模式指导主题爬虫的爬行。实验结果表明:方法能采集大量与指定主题相关的高质量网页,主题爬行效率优于 Best-First 主题爬虫。

参考文献:

[1] 周立柱,林玲. 聚焦爬虫技术研究综述[J]. 计算机应用,2005,25(9):1965-1969

[2] 刘金红,陆余良. 主题网络爬虫研究综述[J]. 计算机应用研究,2007,27(10):26-29

[3] CHO J,GARCIA-MOLINA H,PAGE L. Efficient Crawling Through URL Ordering[A]. Proceedings of the seventh international conference on World Wide Web 7 [C]. 1998

[4] CHAKRABARTI S,MARTIN D,DOM B. Focused crawling:a new approach for topic specific resource discovery[A]. Proceedings of the 8th International World Wide Web Conference (WWW8)[C]. 1999

[5] 邹永斌,陈兴蜀,王文贤. 基于贝叶斯分类器的主题爬虫研究[J]. 计算机应用研究,2009,26(9):3418-3420

[6] SOTIRS B,EURIPIDES G M,PETRAKIS, et al. Improving the performance of focused web crawlers[J]. Data & Knowledge Engineering,2009,68 :1001-1013

[7] BRA P,HOUBEN G,KORNATZKY Y,et al. Information retrieval in distributed hypertexts,Proceedings of RIAO'94,Intelligent Multimedia[A]. Information Retrieval Systems and Management[C]. New York,1994:481-491

[8] 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(1-7):317-326

[9] MENCZER F,PANT G,SRINIVASAN. Topical web crawlers:evaluating adaptive algorithms[J]. ACM Transactions on Internet Technology (TOIT),2004,4(4):378-419

[10] PAGE L,BRIN S,MOTWANI R, et al. The PageRank Citation Ranking:Bringing Order to the Web[A]. Stanford Digital Library Technologies Project[C]. 1998

[11] MCCALLUM A,NIGAM K,RENNIE J,et al. Building domain-specific search engines with machine learning technique[J]. In Procs. of AAAI Spring Symposium on Intelligents Engine in Cyberspace,1999

[12] PANT G,SRINIVASAN P. Link contexts in classifier-guided topical crawlers[J]. IEEE Transactions on Knowledge and Data Engineering,2006,18(1):107-122

[13] RABINER L,JUANG B. An introduction to hidden Markov models[J]. ASSP Magazine IEEE,1986,3(1):4-16

Research on Focused Crawler Based on HMM

XIE Zhi-jun[1] ,YANG Wu[2] ,LI Zhi-ying[1] ,SONG Jing-jing[1]

(1. School of Computer Science and Engineering,Chongqing University of Technology,Chongqing 400054,China;
2. Information and Education Technology Center,Chongqing University of Technology,Chongqing 400054,China)

Abstract: Focused crawler is a core component of the vertical search engine,it collected data resources for the subject-oriented user's query. This paper proposes an approach for focused crawler based on HMM,it not only considers the web content,but also analyzes the context of web link structure. Firstly,the observation state represents the clustering of the current web page,the hidden state represents the link distance from current web page to target web page,then through the HMM model learning user browsing patterns, more topic webpages are downloaded by using the model. Experiments show that the focused crawler based on HMM can capture a large number of high quality web pages related to target topics,and its crawling performs better than Best-First crawler.

Key words: focused crawler; Hidden Markov Model; Vector Space Model; topic correlativity; Vertical Search Engine

责任编辑:代小红
校　　对:田　静