第34卷 第16期 计 算 机 工 程 2008年8月
Vol.34 No.16 Computer Engineering August 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)16—0204—03 文献标识码：A 中图分类号：TP301.6

基于多 Agent 系统的定题爬虫算法

徐照财，程显毅
(江苏大学计算机科学与通信工程学院，镇江 212013)

摘 要：定题爬虫的研究是定题搜索引擎的关键技术。该文提出一种基于多 Agent 系统的爬虫算法，采用本题语义主题关键词过滤的方法来抓取与主题相关的网页，利用本体库语义网络实现本体领域中同近义词的过滤。凭借 HTML 网页标记对关键字识别的不同权重和超链接锚文本对主题相关网页进行预测，通过黑板的通信机制实现多 Agent 交互。实验结果表明算法在抓取网页的查准率、查全率方面有一定的改善。

关键词：定题爬虫；主题关键字过滤；语义

Focused Crawling Algorithm Based on Multi-agent System

XU Zhao-cai, CHENG Xian-yi
(Computer Science & Communication Engineering Institute, Jiangsu University. Zhenjiang 212013)

【Abstract】Focused crawling research is key to search engine technology. In this paper, a focused crawling algorithm based on multi-Agent system is presented, which presents a core issue of a theme key words filtering method based on ontology to collect the URL related to the themes. Semantic network based on the ontology is to achieve filtering of similar meaning. It also introduces keyword identification of different weights by HTML page tags and anchor text, which are important for the website forecast use. And system model based on the blackboard communication mechanism is explained. The experimental results show that the system has an increasing promotion in both precision and extension for crawling website.

【Key words】focused crawling; theme key words filtering; semantics

1 概述

传统的搜索引擎作为网络信息检索工具被用户广泛接受，但其依然存在诸多不足之处。由于涉及面广，因此存在大量无关的信息。为了解决上述问题，近年来，研究学者提出了新一代搜索引擎发展方向，定题检索是其中尤为突出的一种。定题搜索引擎是将信息检索限定在特定主题领域，就主题相关的信息提供检索服务。不同于通用搜索引擎，定题搜索引擎的检索范围相对小，查准率和查全率易于保证。主题相关信息的搜集是定题搜索引擎的核心。到目前为止，有2 种方法比较具有代表性：文献[1]提出的Fish-Search算法和文献[2] 提出的Shark-Search算法。后者是针对Fish-Search算法简单和精度不高的不足而提出来的。它引入了向量空间模型和链接锚文本，起到提示作用。国内学者龙宇巍等的基于反向链接与超文本分析的定题算法[3]也取得不错效果。但是执行效率不够高。由于缺乏本体知识的过滤，因此漏掉了很多与主题相关的主页。查准率和查全率都得不到保障。

本文提出了一种基于 Agent 定题爬虫模型。该模型采用了本体语义的主题关键词过滤的 URL 来收集爬行策略。通过计算相似性来对网页进行取舍，同时依据超链接锚文本分析算法，对主题相关网页进行预测，减少了对应页面的分析，以此提高效率。

2 基于多 Agent 系统定题爬虫模型及关键技术

2.1 定题爬虫模型

定题爬虫模型如图 1 所示。管理 Agent 是控制的核心，它有 2 个功能：一是根据用户需求和初始给出 URL，通过与知识库交互来获得此定题引擎的主题关键字；二是与采集 Agent 交互，将 URL 列表中的 URL 分发给采集 Agent 并将搜索的网页存入数据库。采集 Agent 的功能是：与知识库通信来实现网页的特征提取(通过特征向量空间来实现)及页面中超文本链接分析等，通过分析结果判定是否抓取页面并把它返回给管理 Agent 存入数据库。知识库在此系统中处在核心的地位，它负责判断整个网页是否和主题相关的特征向量空间提取，包括爬虫的 URL 抓取策略、本体语义网和超链接技术。它通过与管理 Agent 和各个采集 Agent 通信发挥作用。

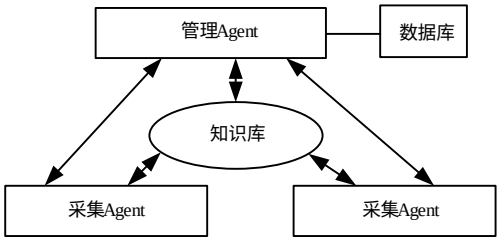

图 1 定题爬虫模型结构

2.2 本体语义

在判断主题词和网页的特征时，采取的方法是比较相同的关键字。然而，这将造成在主题特征项中有个“电脑”项等问题。而网页中有“计算机”项，但是通常简单的比较很难将上述的概念当作是相近的[4]，为此会漏掉许多相关的页面，本体(ontolgoy)技术可以应用在这种情况下，通过创建ontology知识库初步实现了特定领域的概念检索。

基金项目：江苏省科技攻关基金资助重点项目(BE2004093)
作者简介：徐照财(1983－)，男，硕士研究生，主研方向：搜索引擎；程显毅，教授、博士生导师
收稿日期：2007-09-29 **E-mail：**xuzhaocai_2003@163.com

首先是领域本体库的构建，一般情况下需要在领域专家和本体专家的协助下，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)和词汇之间相互关系的明确定义，建立与主题相关的本体库。如图 2 所示的本体语义网的建立，即通过 5 种关系来实现概念检索。

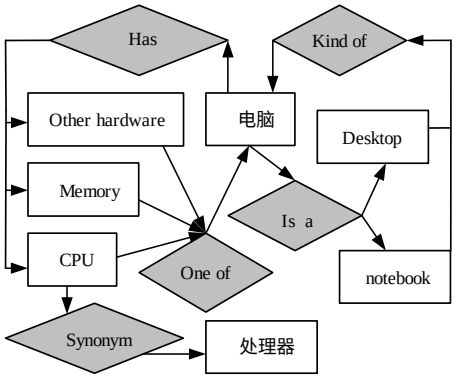

图 2　本体语义网

从图 2 可知，Synonym 表示联接的 2 个概念是同义的，如图中的 CPU 和电脑。Has 表示一个概念是由几个概念组成，如电脑由 CPU, Memory, other hardware 组成，One of 是 Has 的逆序；Is a 表示一个概念的具体化，如电脑有台式机(desktop)和笔记本(notebook)；Kind of 是 Is a 的逆序关系。

这样就可以定义 2 个类来分别实现关系和概念节点(Java 实现)：

```
public class Node        {//节点
String nodeInfo          //节点概念意义
ArrayList   synonyList //同义概念数组
ArrayList hasList        //has 概念数组
ArrayList oneOfList      //one of 概念数组
ArrayList isAList        //Is a 概念数组
ArrayList kindOfList    // kind of 概念数组
//方法省略
}
public class NodeArray{//数组中节点类
public Node node;
public degree;
}
```

通过上述的 2 个类就能很容易地实现概念和节点关系，继而实现 5 种操作：同义扩展操作(Find-synonym)，细化操作(Find-son)，范化操作(Find-father)，实例化操作(Find-subclass)，抽象化操作(Find-superclass)。通过在当前概念的同义、子概念、父概念、具体概念、抽象概念的数组中找到其概念节点的操作来实现。这样就可以提高与主题关键词合作过滤的 URL 收集方法的性能。

2.3　网页标记权重和链接分析

现在的网页一般都用基于 HTML 或 XML 这种结构化标记语言实现，除了上述基于概念语义网的方法发现意义相近的概念词语外，此类标记对发现文档中一些重要信息有着举足轻重的提示作用。通常标记<title>,<head>,<meta>,<a href>,<h1>,<h2>,中的内容比其他内容对识别文档主题来说更重要，一般给予更高的权重。大量的实践表明：为了提高被检索的几率，<head>,<meta>2 个标记有时存在对搜索引擎的欺骗，而与实际的文档内容关系不大，因而只给它普通的权重。其中<a href>之间的锚文本最重要。文献[5]提出计算特征词在文档中权重的 TF-IDF 公式：

$$W_{ik} = tf_{ik} \times IDF_k = tf_{ik} \times \lg(\frac{N}{n_k} + 0.5) \quad (1)$$

其中，tf_{ik} 表示特征词 t_k 在文档 w_i 中的词频；n_k 表示文档包含特征词 t_k 的总数；N 表示文档库中文档的总数。通过以下公式来计算网页中关键字的权重：

$$W_{i,j} = tf_{i,j} \times \lg(\frac{N'}{t_j} + 0.5) \times K \quad (2)$$

其中，N' 表示当前文档库中已存在的主题相关的文档总数，依据 HTML 的标记为 K 取值，对其取值如下：

$$K = \begin{cases} 3 & \text{在锚文本中} \\ 2 & \text{title/h1/h2/strong中} \\ 1 & \text{其他} \end{cases}$$

对文本的链接分析特别重要，抓取链接是爬虫工作的根本，对 anchor text的锚文本和附近的锚区域的文本分析可以判断该链接所对应的文本是否主题相关的，如此可以通过预测作页面的取舍，减少了抓取网页数量，从而相应地减少了对应页面的分析，减轻了系统的工作时间，提高了效率。

设定 2 个阈值 θ' 和 θ ($\theta' < \theta$)，分别表示与主题相关特征词个数与总的主题相关特征词的比值，算法流程如下：

Input: anchor text 和 anchor area text

Process：语义分析和权重计算，获取输入文本的相关的特征词个数与主题特征词的比值，设为 π ；

if($\pi < \theta'$)：丢弃此 Hyperlinks，不作任何处理；

else if($\pi < \theta$)：放到 URL 列表中，供后续处理；

else：直接下载对应的页面，交给管理 Agent 存入数据库中。

上述算法根据 π 的不同取值对 Hyperlinks 对应的页面进行适当的取舍操作，减少了系统的运行负担，提高了效率。

2.4　基于黑板的通信机制

Agent 系统的智能性主要体现在它的协作和协商中，这离不开多 Agent 的通信技术的实现。如果这个通信技术问题解决不好，将成为 Agent 间的“瓶颈”，严重影响系统的性能发挥。在此系统中，管理 Agent、采集 Agent 和知识库之间主要采用基于黑板的通信机制。

黑板模型结构是为了解决分布在不同物理环境下多个实体协作完成任务的并行和分布的计算模型。它能够实现异构知识源的集成、共享与交流。黑板的组成结构如图 3 所示，其由共公区和专用区组成。

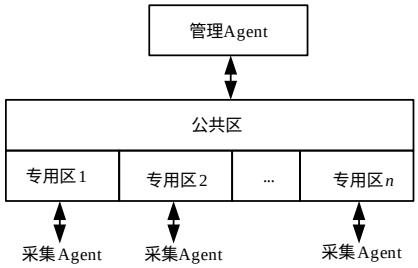

图 3　黑板结构

公共区是黑板中的一个特殊区域，任何 Agent 都可以对它进行读写操作，它就像一个通告板。公共区主要存放的是知识库，管理 Agent 和采集 Agent 都可以与之通信以对文档的主体特征进行分析。每个采集 Agent 都有自己的专用区，用于存放中间结果等属于自己私有的数据成员。协作式的通信一般按广播方式执行如下命令：

{broadcast MAgent (c1Agent(URL1), c2Agent(URL2))}

它是管理 Agent 向 $c1$ 和 $c2$ 2 个采集 Agent 发布任务处理 URL 任务的消息命令。基于黑板的通信机制简化系统协作的复杂性，可以优化系统的性能。

3 爬行策略与算法

本文提出了一种基于主题关键词过滤的 URL 收集爬行算法。该策略基于广度优先算法。首先是管理 Agent 抽取主题表示成向量空间模型(VSM)存入知识库中，从主题相关的示例网页集中提取关键词序列作为本主题的特征项。记 $w_{\text{theme}}=\{w_1,w_2,\cdots,w_n\}$ 表示主题特征项集合，n 为特征项的个数，同样对于将要采集的网页 X 的特征项记作 $w_x=\{w_{x,1},w_{x,2},\cdots,w_{x,n}\}$。然后管理 Agent 从初始的 URL 列表将 URL 传递给各采集 Agent。采集 Agent 下载获取网页内容，一方面生成网页的特征向量，按照主题关键字算法对网页进行取舍；另一方面依据锚文本分析算法(ATAA)对链接进行分析，直到管理 Agent 保存的 URL 列表为空。

本体语义主题关键字过滤算法具体实现如下：设主题关键字特征项 w_{theme} 和网页特征项 w_x 中有 m 个关键字特征项是相同的，这里的相同包括了采用了本体语义网后相近和同义的关键字，则对 m 个关键字特征项分别记为 $w_{ths}=\{w_1,w_2,\cdots,w_m\}$ 和 $w_x=\{w_{x,1},w_{x,2},\cdots,w_{x,m}\}$，则主题和节点对应的网页 y 的相似程度 $Sim(x,y)$ 可按下式计算：

$$Sim(x,y)=\frac{\sum_{i=1}^{m}(w_i-w_{\text{theme,avg}})(w_{x,i}-w_{x,\text{avg}})}{\sqrt{(\sum_{i=1}^{m}(w_i-w_{\text{theme,avg}})^2)(\sum_{i=1}^{m}(w_{x,i}-w_{\text{theme,avg}})^2)}} \quad (3)$$

其中，$w_i\subseteq w_{ths}, w_{x,i}\subseteq w_x$；$w_{\text{theme,avg}}$ 为主题特征项的 y 的一个平均评价值；$w_{x,avg}$ 为节点网页的特征项的一个平均评价值；它们是可以被指定的 2 个常数。Sim(x,y) 越大，说明页面 x 与主题相似度越大。通过计算，当 Sim(x,y) >3.5 时，就可视为页面 x 与主题具有可以接受的相似度，然后页面被抓取下来由管理 Agent 保存到数据库中。

4 实验与性能分析

本实验以“计算机科学”为主题，构造了一个基于计算机领域简单的本体库，给出了初始的一些页面集，作为提取系统的主题特征关键字素材，然后给出爬行需要的初始 6 条 URL(从 google 搜索结果集中经挑选而获得)，为了方便测试将搜索深度限制在 4 层。系统由 Java 实现，整个系统运行在一台 Windows XP(P4 3.0)机器上，设置 20 守护线程(daemon thread)，这样可以由控制线程统一控制，通过在不同的时间长度使控制线程中断而退出，在数据库中可以统计在不同时间段 T (单位：ms)，2 种情况下的相应下载页面数和主题相关页面数 N 的数量。

从图 4 可以看出，在相同的时间内，使用基于本题和有链接预测的方法的下载页面数比没有基于本题和链接预测的要高 20%～30%，查准率要高出 8%～10%。原因是基于本题语义过滤提高了查准率以及由于采取超链接的文本预分析，减轻了系统的运行负担，提高了效率，在相同的时间内可下载更多的页面，从而间接地提高了查全率，但是约有 5%的主题相关页面也被丢弃了。系统的查准率仍然不高，下一步就如何更高地提高系统的查准率展开研究。

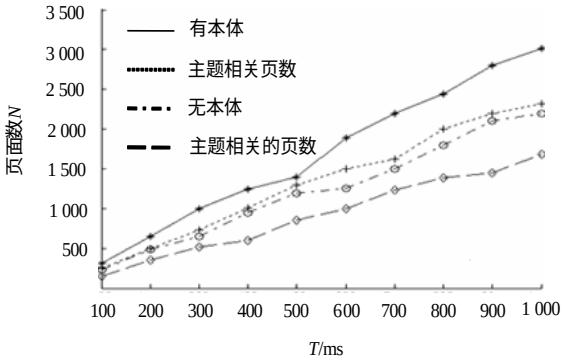

图 4 实验结果

5 结束语

定题信息搜索算法是信息过滤的关键技术。本文提出了一种基于 Agent 的定题爬虫模型。实验表明模型能在一定程度上提高爬行的查全率和查准率。机器学习在智能搜索引擎技术中已经得到深入的发展。下一步将遗传算法引入到系统中，进一步测试系统以提高查准率。

参考文献

[1] DeBra P, Post R. Information Retrieval in the World-Wide Web: Making Client-based Searching Feasible[J]. Computer Networks and ISDN Systems, 1994, 27(2): 183-192.

[2] Hersovici M, Jacov 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.

[3] 龙宇巍, 王永成, 许欢庆. 定题搜索引擎 Robot 的设计与算法[J]. 计算机仿真, 2004, 21(4): 70-76.

[4] Ehring M, Maedche A. Ontology-focused Crawling of Web Documents[C]//Proc. of the ACM Symposium on Applied Computing. [S. l.]: ACM Press, 2003: 624-626.

[5] Salton G, McGill M J. Introduction to Modern Information Retrieval[M]. New York, USA: McGraw-Hill, 1983.

(上接第 203 页)

聚类中心点选择对最终聚类结果的影响。入侵检测的实验结果说明，优化后的 K 均值算法对已知和未知攻击的检测率明显提高，同时误警率大大降低，从而提高了 K 均值算法在入侵检测领域的应用价值。

参考文献

[1] Flanagan J A. Unsupervised Clustering of Symbol Strings[C]//Proc. of Intl' Joint Conference on Neural Networks. Portland Oregon, USA: [s. n.], 2003: 3250-3255.

[2] Krishma K, Nurty M N. Genetic K-means Algorithm[J]. IEEE Trans. on System, 1999, 29(3): 433-439.

[3] Kennedy J, Eberhart R C. Particle Swarm Optimization[C]//Proc. of IEEE International Conference on Neural Networks. Perth, Australia: [s. n.], 1995: 1942-1948.

[4] Sclim S Z, Lsmailm A. K-means-type Algorithm: A Generalized Convergence Theorem and Characterization of Local Optima Reality[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1984, 6(1): 81-87.

[5] The UCI KDD Archive. KDD99 Cup Dataset[EB/OL]. (1999-07-09). http://kdd.ics.uci.edu/ databases /kddcup99/kddcup99.html.