中医药主题搜索网络机器人的设计与实现

陈 勇[1],刘 勇[2]
(1.仲恺农业工程学院 计算机科学与工程学院,广东 广州 510225;
2.中国科学院 成都计算机应用研究所,四川 成都 610041)

摘 要:主题搜索网络机器人的研究对于主题搜索引擎整体性能的提高具有重要意义。鉴于国内尚缺少专门面向中医药主题的搜索引擎,针对中医药信息的特点提出了中医药主题搜索网络机器人的搜索策略和系统结构,描述了系统的基本工作流程。结合Java I/O流、套接字编程、多线程编程、中文分词和数据库JDBC连接等技术,设计和实现了中医药主题搜索网络机器人系统。面向中医药主题对如何提高主题搜索网络机器人的搜索效率和精度进行了有益的探索,对其它主题搜索网络机器人的研究和开发具有一定的借鉴作用。

关键词:搜索引擎;网络机器人;中医药
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2010)05-0162-05

Design and Implementation of Topic-Specific Robot for Traditional Chinese Medicine

CHEN Yong[1],LIU Yong[2]
(1.College of Computer Science and Engineering,Zhongkai University of Agriculture and Engineering,Guangzhou 510225,China;
2.Chengdu Institute of Computer Applications,Chinese Academy of Sciences,Chengdu 610041,China)

Abstract:Topic-specific robot is a key part of topic-specific search engine. Considering the feature of traditional Chinese medicine information,proposes the searching strategy and the architecture of the traditional Chinese medicine topic-specific robot, and then describes its working flow. With the use of technologies such as Java I/O stream, socket programming, multithread programming,Chinese words segmentation and JDBC, the design and implementation of the system's database, topic initialization module, fetching webpage module and filtering webpage module are presented. Aiming at improving the efficiency and precision of the traditional Chinese medicine topic-specific robot,this paper will be in favor of the research and development of other topic-specific robots.

Key words:search engine;web robot;traditional Chinese medicine

0 引 言

自20世纪90年代中期以来,英美等西方发达国家的医学搜索引擎层出不穷,如Medical Marix、Clini Web International、MedFinder、MedHelp等等,不胜枚举。而与此形成较大反差的是,国内虽已建立了一些中医药信息数据库检索系统和寻医问药站点,但缺少专门面向中医药主题的搜索引擎,这与中医药在医学领域和日常生活中所处的重要地位是不相称的。

由于WWW信息资源呈指数级增长且处在不断的变化之中,如果人工地去检索和分类整个WWW信息资源将是一项艰巨而几乎不可能完成的工作。由于网络机器人不需要人工干预,可以自动地在网络中穿梭,其信息采集速度、覆盖面和及时性较之人工采集大大提高,因此当前绝大多数搜索引擎都是基于Robot搜索引擎。

网络机器人作为基于Robot搜索引擎的后台部分,对于用户而言是不可见的,但是它采集的结果可以在搜索引擎的索引库覆盖范围、索引库容量和更新频率中得以体现。巧妇难为无米之炊,网络机器人的性能将直接影响到基于Robot搜索引擎的性能。所以,网络机器人在基于Robot搜索引擎中举足轻重。

主题搜索引擎作为基于Robot搜索引擎的一种新

收稿日期:2009-09-15;**修回日期**:2009-12-16
基金项目:四川省技术创新基金(2008PT013);仲恺农业工程学院博士启动基金资助项目(G2360295)
作者简介:陈 勇(1975-),男,讲师,博士,研究方向为软件工程和人工智能;刘 勇,高级工程师,博士,研究方向为智能计算和软件工程。

的发展形式,自然少不了网络机器人的参与。而且由于主题搜索引擎只提供主题领域内的信息查询,这就要求主题搜索网络机器人在进行网上信息采集时,必须按照预先规定的主题并采用主题搜索策略采集网上相关信息,过滤掉无关信息,从而减少被采集的信息数量,提高索引数据库中的信息质量。所以,相对于传统的网络机器人而言,主题搜索网络机器人具有更深的搜索深度和更短的搜索周期,从而在更大程度上直接影响到搜索引擎的整体性能。因此,主题搜索网络机器人的研究对于主题搜索引擎整体性能的提高具有重要作用。

1　中医药主题搜索网络机器人的搜索策略

文中的中医药主题搜索网络机器人在采集网页的过程中对同一站点采用宽度优先搜索策略,对采集回来的网页利用向量空间模型 VSM(Vector Space Model)进行中医药主题相关度分析后过滤掉与中医药主题相关度较小的网页,对其中的超文本链接将不予处理,而与中医药主题相关度较大的网页中的超文本链接将被提取出来并加入到 URL 等待队列中。从而达到除去多余链接、减少无用搜索、节省存储空间、提高搜索效率之目的。

2　系统体系结构和基本工作流程

中医药主题搜索网络机器人的体系结构如图 1 所示。系统基本工作流程如图 2 所示。

1)主题初始化模块设置中医药主题关键词和中医药种子站点并存放在数据库中,为网页采集模块的运行作必要的准备;

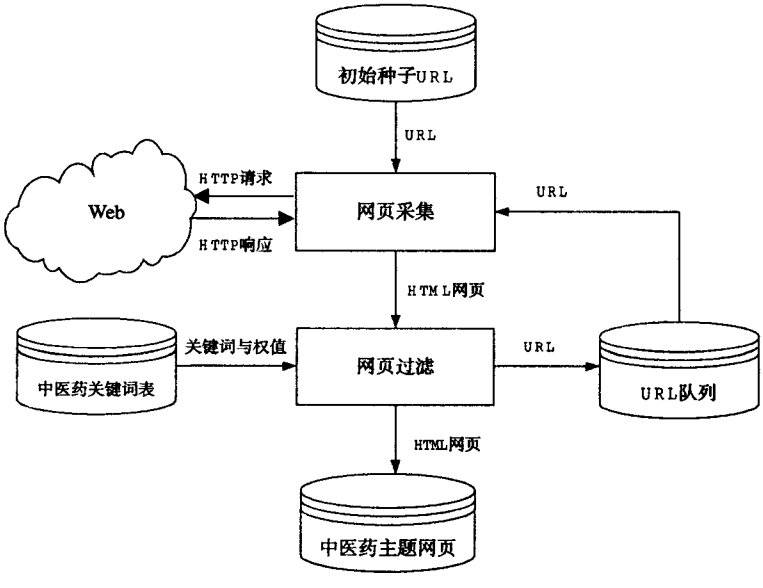

图 1　中医药主题搜索网络机器人的体系结构

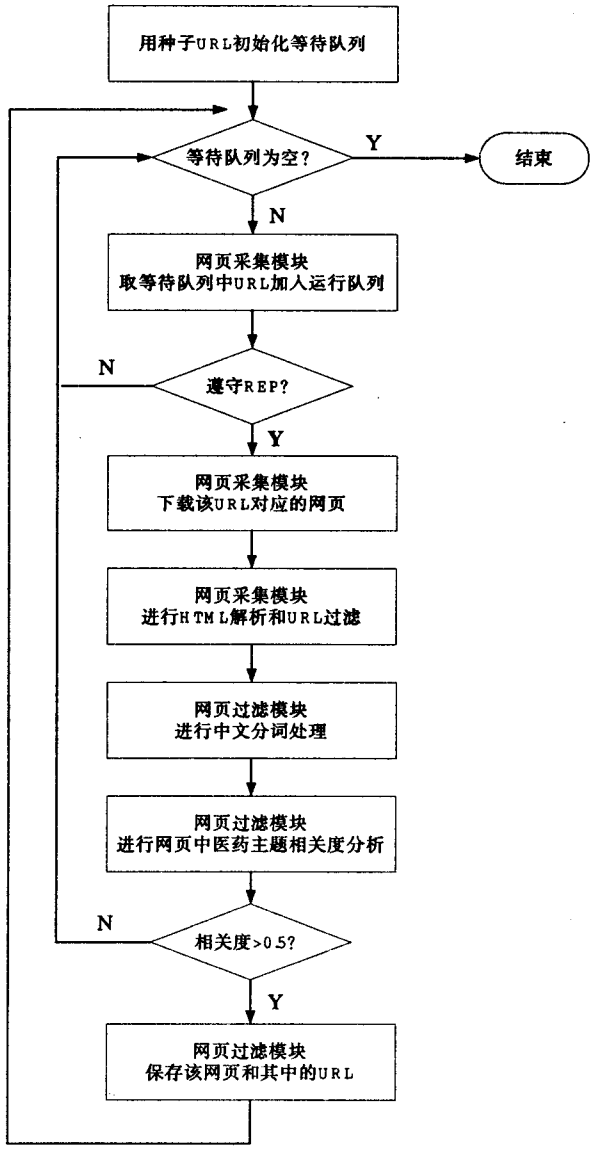

图 2　系统基本工作流程

2)网页采集模块从数据库中读取中医药种子站点 URL 加入等待队列;

3)网页采集模块当等待队列非空时取 URL 加入运行队列,否则停止;

4)网页采集模块判断 URL 是否满足 REP,满足时继续,否则转到第 3 步;

5)网页采集模块利用 HTTP 协议下载相应网页;

6)网页采集模块对已下载的网页进行 HTML 解析和 URL 过滤;

7)网页过滤模块对网页文本进行中文分词处理;

8)网页过滤模块对网页进行中医药主题相关度分析;

9)网页过滤模块将与中医药主题相

关度较高的网页保存到中医药主题网页数据库,并将其中经过过滤的 URL 加入 URL 队列;与中医药主题相关度较低的网页以及其中的 URL 被直接剔除;

10)网页采集模块从 URL 等待队列中取出 URL 加入运行队列继续工作,循环到第 3 步,直至等待队列为空,并且当前已经没有网页正在处理时,系统将停止工作。

3 系统设计与实现

3.1 主题初始化模块

主题初始化模块的实现要用到两张数据库表。

(1)关键词表 tcmkeyword。

表中 KEYWORD 字段代表中医药主题的关键词;WEIGHT 字段代表着该关键词相应的权值,其取值范围从 1 到 5。

关键词表 tcmkeyword 中的记录从《医学主题词表》、《中国中医药主题词表》和《汉语主题词表》筛选而来,主要包括中医学一般概念、中医病因病机、中医治则治法、中草药名、中医药方剂、中医药古籍、经络穴位等。

(2)初始种子表 seedurl。

表中 ID 字段为种子 URL 的编号;URL 字段为初始种子站点的 URL。初始种子表 seedurl 中的最初记录可利用 Google 和百度等通用搜索引擎得到。用户根据需要可以对以上两张表中的记录进行设置和维护。

3.2 网页采集模块

3.2.1 遵守 REP 协议

网页采集模块从 URL 等待队列读入一个 URL 后,应该判断该 URL 是否在 robots. txt 禁止访问之列,以遵守 REP 协议[1,2]。

其算法描述如下:

```
Boolean robotrep(URL url)
{String strHost = url. getHost ();
//形成 robots. txt 文件的 URL 地址
String strRobot = "http://" + strHost + "/robots. txt";
URL urlRobot = new URL(strRobot);
if (robots. txt 文件不存在)
return true;
else {
If (url 包含于 robots. txt 文件的 Disallow 定义中)
return false;
else
return true;
}
}
```

3.2.2 解析 HTML 和 URL 过滤

HTML 格式的网页文件中包含有大量的格式信息,需要根据 HTML 标记提取网页中的关键信息,如标题、更新时间、正文等,过滤掉 HTML 标记及其它无用信息,保留纯文本格式文件,以便进行后续处理[3]。

特别地,为了能够连续不断的采集,还必须能够从下载的 HTML 网页中提取新的 URL。系统主要是从 HTML 文档的 A 元素的 HREF 属性的属性值中、IMG 元素的 SRC 属性的属性值中,以及 FRAME 元素的 SRC 属性的属性值中提取新的 URL。

对于提取出来的 URL,首先需要进行过滤处理,滤除含有非 HTTP 协议头的 URL(如 FTP,News,Telnet,Mailto 等)和明确含有不处理文件类型后缀名的 URL(如 rm,mp3,wav,avi,jpg,exe,zip 等)。

经过滤处理后的 URL 中,如果有相对 URL,那么还需要转化为绝对 URL,即在相对 URL 前必须加上 BASEURL。

依据 HTML4.01 和 RFC2616 规范 BASEURL 可以由下述从高到低优先次序来取得[4]:

1) 由 HTML 文档的 BASE 元素设置的 BASEURL;

2) 由在 HTTPHEADER 中发现的元数据定义的 BASEURI(涉及 ContentBase,Content Location);

3) 缺省时,BASEURL 就是当前文档的 URL。

3.2.3 URL 队列的存储

URL 队列存储在 URL 队列表 urlqueue 中,其中 URL 字段代表 URL 链接,是 urlqueue 表的主键;STATUS 代表 URL 的状态,其取值有四种情形:

STATUS='W'表示该 URL 正在等待处理;

STATUS='R'表示正在处理该 URL 对应的网页;

STATUS='C'表示该 URL 对应的网页已经完成下载;

STATUS='E'表示处理该 URL 时发生了错误。

3.2.4 网页采集实现细节

网页采集模块的具体实现主要参考 BOT 包[5],用到如下类和接口:

1)Spider 类。

Spider 类是主类,其主要功能是设置起始 URL 和 poolsize(线程池大小)等参数,为 Fetchpage 开始采集网页做准备;确定 spider 工作完成的时间;启动一个等待作业的线程,用于向作业管理器添加一个作业;处理网页等。

2)IspiderReportable 接口。

该接口定义了 spider 检索到的网页的一个消费

者,任何实现该接口的对象都可以从 spider 接收网页。

3)IworkloadStorable 接口。

该接口实现了存储 URL 队列的基本功能,以管理 spider 的作业。

4)SpiderSQLWorkload 类。

该类是 IworkloadStorable 接口在 SQL 数据库中存储作业的实现。SpiderSQLWorkload 对象的工作是预备若干条 SQL 语句,以用于创建、增加、更新或删除作业实体。它声明了一个名为 setURLStatus( )的方法用来检测指定的 URL 的状态是否存在,当状态不存在时,将为其创建一个;当状态存在时,它将被更新。

5)SpiderWorker 类。

SpiderWorker 类的基本任务是下载一个 Web 站点,并将网页内的链接加入到作业中。当 spider 启动时,它创建一个处理 spider 发现网页的 SpiderWorker 类池,以实现多线程机制。

当每个 SpiderWorker 对象 start( )方法被调用时,就启动 Spider 对象的 run( )方法。于是 run( )方法开始等待,直到它有一个作业时。在从 Spider 管理器得到一个作业后,将通知 SpiderDone 类该线程不再空闲。于是作业被传送到 processWorkload( )中进行处理。processWorkload( )方法将会下载指定的网页并解析 HTML 和提取其中的 URL。

6)SpiderDone 类。

由于有很多并发的线程,要确切地知道网页采集何时完成是很困难的。SpiderDone 类用来跟踪有多少个线程仍在运行,并等待所有线程结束。网页采集的完成应满足两个标准:

①没有活动的 Worker 线程。如果有活动的 Worker 线程,那就会有新的 URL 加入等待队列,网页采集将继续进行。

②URL 等待队列为空。URL 等待队列非空时,网页采集将继续下去。

3.3　网页过滤模块

3.3.1　中文分词的实现

该系统采用正向最大匹配法(FMM)进行中文分词。

中医药关键词表中最长的关键词含汉字个数为 7,故取被处理文本中当前字符串序列中前 7 个字为匹配字段,查找中医药关键词表,如果关键词表中有这个 7 字词,则匹配成功,匹配字段作为一个词被切分出来;如果关键词表中没有这个 7 字词,则匹配失败。于是匹配字段去掉最后一个汉字,剩下的字符作为新的匹配字段,再进行匹配,如此进行下去,直到匹配成功为止。该系统中文分词功能由 FMMSegment 类实现。其中利用正向最大匹配分词算法对单个长句进行分词的核心代码如下:

```
public int SentenceSegment(String Sentence)
{
  int senLen = Sentence.length();
  int i=0, j=0;
  int M=7;
  String word;
  boolean bFind = false;
  while(i < senLen)
  {
    int N= i+M<senLen ? i+M : senLen+1;
    bFind=false;
    for(j=N-1; j>i; j--)
    {
      word = Sentence.substring(i,j);
      if(dic.Find(word))
      {
        System.out.print(word +" ");
          bFind=true;
          i=j;
          break;
          }
        }
      if(bFind == false)
    {
    word = Sentence.substring(i, i+1);
    System.out.print(word +" ");
    i=j+1;
  }
    }
    System.out.println( );
  return 1;
}
```

3.3.2　主题相关度分析的实现

为了判断网页的中医药主题相关度,采用了向量空间模型 VSM(Vector Space Model)算法[6~8]。

其基本思想如下:把中医药主题关键词的个数 n 作为向量空间的维数,每个关键词的权值 β_i 作为每一维分量的大小,则网页的关键词权值向量对网页文本进行分析,统计其中每个关键词出现的次数,以出现次数最多的关键词作为基准,设其出现次数为 P,第 i 个关键词出现次数为 P_i,则第 i 个关键词的频率 $\alpha_i = P_i/P$。

网页的主题向量网页的中医药主题相关度用以上两个向量夹角的余弦表示,即该系统指定中医药主题相关度的阈值为 0.5,当相关度>0.5 时认为对应网页与中医药主题相关,该网页信息被存储到中医药主题

网页数据库中,其中的超文本链接 URL 将被加入到 URL 队列数据库;否则认为与中医药主题关系不大,相应网页将被简单地剔除,不再进行处理。

该系统主题相关度分析功能由 TCMRelevancy 类实现,网页关键词权值向量和网页主题向量均定义为 java. util. Vector 类。

3.3.3 中医药主题网页的存储

中医药主题网页存储在中医药信息网页表 tcmpage 中,其中 URL 字段代表 URL 链接,LASTUPDATE 代表该网页最后更新时间,TITLE 代表该网页的标题,CONTENT 代表该网页的内容。URL 字段和 LASTUPDATE 字段共同构成 tcmpage 表的主键。

4 结束语

文中面向中医药主题,对如何提高主题搜索网络机器人的搜索效率和精度进行了有益的探索。但是文中采用的中文分词方法还比较粗糙,在较大程度上影响到网页中医药主题相关度分析的准确性,在以后的研究过程中要下大力气加以改进。

其次,文中采用向量空间模型 VSM 算法初步实现了网页过滤功能,但是要想得到更好的效果,还必须综合应用数据挖掘、人工智能、神经网络、粗集理论等各方面的方法和技术。

再次,网络机器人还无法实现信息的准确分类,这在一定程度上将影响到搜索引擎的检索效果。这是目前该领域的研究热点之一,也将是以后努力的重要方向之一。

参考文献:

[1] 谭淑英,刘丽华. Web Robot 技术及其 Java 实现[J]. 中南工业大学学报,2001,32(3):325-327.

[2] 洪光宗,王 皓. 搜索引擎 Robot 技术实现的原理分析[J]. 现代图书情报技术,2002(1):48-50.

[3] Reilly D,Reilly M. java 网络编程与分布式计算[M]. 沈凤,译. 北京:机械工业出版社,2003.

[4] 潘春华,常 敏,武港山. 面向 Web 的信息收集工具的设计与开发[J]. 计算机应用研究,2002,19(6):144-147.

[5] Heaton J. Programming Spiders,Bots,and Aggregators in Java [M]. Sybex,Alameda,USA:[s. n.],2002.

[6] 丁国良,王嘉祯. 专题式 Web 信息检索系统的设计与实现[J]. 军械工程学院学报,2000,12(1):58-61.

[7] 汪 涛,樊孝忠. 主题爬虫的设计与实现[J]. 计算机应用,2004,24(6):270-272.

[8] 戴先宇,王明文,吴水秀,等. 带参数的搜索引擎[J]. 江西师范大学学报:自然科学版,2002,26(4):344-348.

(上接第 141 页)

个相同操作连续执行效率将大大提升。因此在计算中将累计值放置在多个变量中,使处理器不会因为数据相关而暂停。将代码改进之后测试结果如表 2 所示。

表 2 代码优化后运算时间(秒)

实现方式	数据量			
	8192	16384	32768	65536
串行实现	0.005607	0.013340	0.027731	0.060521
Windows 多线程实现	0.004529	0.006921	0.014531	0.037868
OpenMP 多任务实现	0.003110	0.007707	0.016510	0.034331

3 结束语

通过实验对比,发现几种并行实现的方式中 OpenMP 多任务实现的效果最好,并行粒度越细,效果越差,数据量越大,并行化的优势越明显。原因主要在于 OpenMP 的 Fork - Join 执行模型,OpenMP 在被调用时从线程池唤醒线程,当线程执行结束时又重新放到线程池中。并行粒度越细 fork 和 Join 的次数就越多,开销就越大,当粒度细到一定程度时,创建线程的开销将超过并行运算所带来的效益。因此,在利用 OpenMP 进行多线程程序开发时,并行的粒度不应过小,每个并行模块的运算量要尽量平均,总的运算量要达到一定规模,这样才能充分发挥 OpenMP 的优势。

参考文献:

[1] 潘晓杰,刘涤尘. 谐波分析高效算法的研究[J]. 阜阳师范学院学报:自然科学版,2005,22(3):13-16.

[2] 柯建东,刘文江,祝叶华. 多载波中的实数 FFT 及其离散 Hartley 变换实现[J]. 信息技术,2005(10):15-17.

[3] 冷建华. 傅里叶变换[M]. 北京:清华大学出版社,2004.

[4] 铁满霞,董玉红. 快速傅里叶变换的多机并行计算[J]. 航空计算技术,2000,30(3):5-7.

[5] Olejniczak F J, Ribeiro P. Time varying harmonics: Part I: Characterizing measured data[J]. IEEE Trans on Power Delivery,1998,13(3):938-944.

[6] 史旭光,裴海龙. 一种改进的 FFT 方法在谐波测量中的应用[J]. 计算技术与自动化,2005,24(2):24-26.

[7] Morl H,Itou K. An artificial neural net based method for predicting power system voltage harmonic[J]. IEEE Trans on Power Delivery,1992,7(1):402-409.

[8] Monteiro M E,Moura E S, Drago A B,et al. An Internet-Based Power Quality Monitoring System[C]// IEEE International Symposium on Industrial Electronics. [s. l.]:[s. n.],2003:333-336.