

基于中药主题的垂直搜索引擎设计与实现

彭 嫒¹, 陆安江², 董旭辉³

(1. 贵州师范大学 机械与电气工程学院, 贵州 贵阳 550014;

2. 贵州大学 计算机学院, 贵州 贵阳 550025;

3. 贵州省移动分公司, 贵州 贵阳 550001)

摘 要:以贵州中药信息化服务平台开发需求为背景,提出了一种基于 Agent 中药动态信息智能监测系统的设计方案,即主要是应用智能 Agent 技术,实时监测及整合互联网上中药行业相关的众多网站上的动态信息,建立一套庞大的经济动态数据库,实现信息的收集和发布双向互动。同时对基于统计的自然语言处理算法及基于空间向量的相似度排序算法进行了深入的研究及改进,并应用于本系统中。通过对系统的测试表明,该系统具有良好的可靠性、可移植性和应用性,达到了预期的设计效果,为中药企业带来了便利,提高了效益。

关键词:垂直搜索;信息采集;信息处理;Agent;中药主题

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2013)08-0059-03

doi:10.3969/j.issn.1673-629X.2013.08.015

Design and Implementation of Vertical Search Engine Based on Chinese Medicine Theme

PENG Man¹, LU An-jiang², DONG Xu-hui³

(1. College of Mechanical and Electrical Engineering, Guizhou Normal University, Guiyang 550014, China;

2. College of Computer, Guizhou University, Guiyang 550025, China;

3. Mobile Corporation in Guizhou Province, Guiyang 550001, China)

Abstract: On the background of Guizhou Chinese medicine information service platform, propose a design scheme based on intelligent monitoring system of Chinese medicine dynamic information of Agent, which mainly uses intelligent Agent technology to monitor and integrate the dynamic information in websites about Chinese medicine industry of Internet, establishing a huge economic dynamic database to realize the interaction between information collection and publishing. The natural language processing algorithm based on statistics and similarity ranking algorithm based on vector space are researched and improved to be used in this system. The testing result showed that the system has good reliability, portability and application, achieving the desired effect, which could bring the convenience and improve efficiency for TCM enterprises.

Key words: vertical search; information collection; information processing; Agent; Chinese medicine theme

0 引言

随着互联网的迅速发展和普及,网络信息资源呈几何指数增长,想要在海量的信息资源中快速、准确地搜集到自己所需要的信息变得越来越困难。搜索引擎作为提供搜索服务的工具已成为人们获取信息的重要途径,正在深刻地影响着人们的生活。但通用搜索引擎与用户的需求之间存在着巨大的反差,在查询结果中存在着大量重复或不相关的垃圾信息,不能满足特

定领域精确搜索的需求,这使得专业垂直搜索引擎应运而生^[1,2]。垂直搜索引擎凭借明确的检索目标定位,对网页搜索运用主题过滤技术,通过 URL 分析和实际内容过滤来进行^[3-5]。

根据贵州中药行业信息化的需求,文中设计实现了一个科学、完整的中药信息系统平台,提供中药系统的信息,包括中药的种植、中药材的生产、加工、价格等信息。

收稿日期:2012-10-22

修回日期:2013-02-26

网络出版时间:2013-04-22

基金项目:贵州省科技计划项目([黔科合社字[2009]5015])

作者简介:彭 嫒(1970-),女,河南邓州人,硕士,讲师,研究方向为通信网络与信号传输。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20130422.1722.032.html>

1 垂直搜索引擎的体系结构设计

在针对中药主题垂直搜索引擎的设计过程中,要特别注重以下几个关键问题的解决:首先是选择合适的搜索引擎过滤技术,既要满足信息抽取的广度又要针对中药信息分布的特性来进行,以上需求可以通过 URL 过滤技术和网页内容过滤技术来实现;其次是中文分词方法和基于内容的主题过滤算法的设计与实现,本系统采用的是基于统计的中文分词方法和向量空间模型相似度排序算法;最后在信息抽取过程中,采用了在 HTML 网络抽取技术的基础上将网页中非结构化的数据转换为结构化的数据,便于同时管理两种数据。

从系统组成部分来讲,中药主题垂直搜索引擎系统由“资源规划和网站规则生成子系统”、“信息采集子系统”、“信息处理子系统”、“资源管理子系统”、“信息发布及应用子系统”五个部分组成,每个部分实现各自的功能,协同完成中药信息整合系统预定的目标。其体系结构图如图 1 所示。

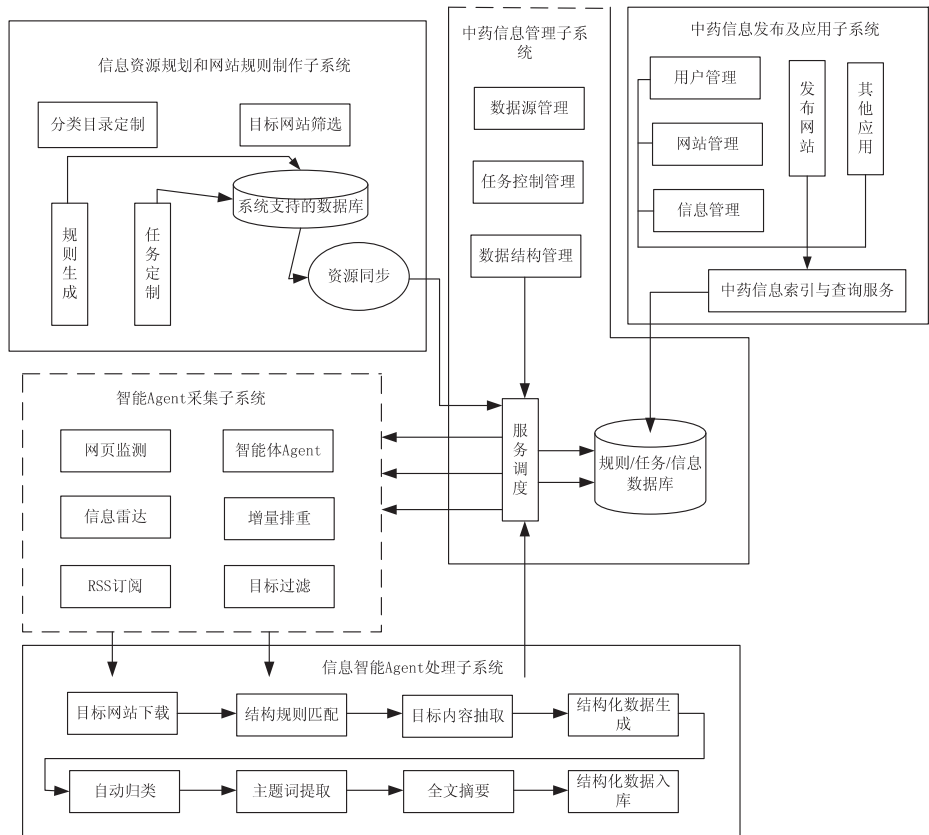


图 1 系统总体的体系结构

由图 1 可知,其各子系统详细设计如下:

1.1 中药信息资源规划和网站规则制作子系统

资源规划和网站规则制作子系统是整个系统的支持子系统,主要完成系统信息采集之前的资源规划和信息源网站规则的制作、维护等工作。信息资源规划工作需中药企业提出明确业务需求,根据业务需求,制

定一套服务于中药行业的分类目录体系,再按照分类目录体系涉及的领域,筛选目标网站、定制数据结构、采用机器学习的方式制作网站规则、分配抓取任务。

1.2 信息采集子系统

信息采集子系统是一个自觉的信息搜索和发现子系统,根据规则库中各种信息源网站规则,利用计算机人工智能 Agent 技术、互联网搜索技术、信息雷达技术及网页监测技术(包括中文分词、关键词提取、内容去重等)及时准确地发现目标网站上的最新信息,调度信息处理子系统模块对新信息进行处理^[6,7]。

该子系统采用 C/S 模式架构,可以根据系统硬件分布情况,灵活地部署一到多个监测终端,不受机器地域和终端数量的限制,符合流行的“分布式收集和处

1.3 信息处理子系统

信息处理子系统是采用基于规则的信息抽取方

法,结合互联下载技术、HTML 网页识别技术、XML 语言解析技术、自然语言处理等技术,将各种非结构化和半结构化网页信息转化成可以重复利用和检索的结构化数据^[9,10]。

该子系统协同采集系统协作,组成了整个系统的核心处理模块,承担着系统最重要的信息处理、加工的任务,所有处理全过程无需人工干预,系统自动完成对目标信息的下载、匹配、抽取、转换、归类、入库等一系列操作。

1.4 中药信息资源管理子系统

信息资源管理子系统是布置在用户服务器端,用于控制和管理系统中所有监测的外部网上资源,分配不同任务数量,跟踪各监测和处理终端任务执行情况,调整系统数据结构,查看系统中可能出现的错误日志。在本系统设计中,互联网上的网站资源是通过规则生产程序制作成不同的规则,而规则需要通过资源管理系统生成不同的采集任务,信息检测模块按照信息采集任务自动采集互联网

上的信息资源。其流程如图 2 所示:

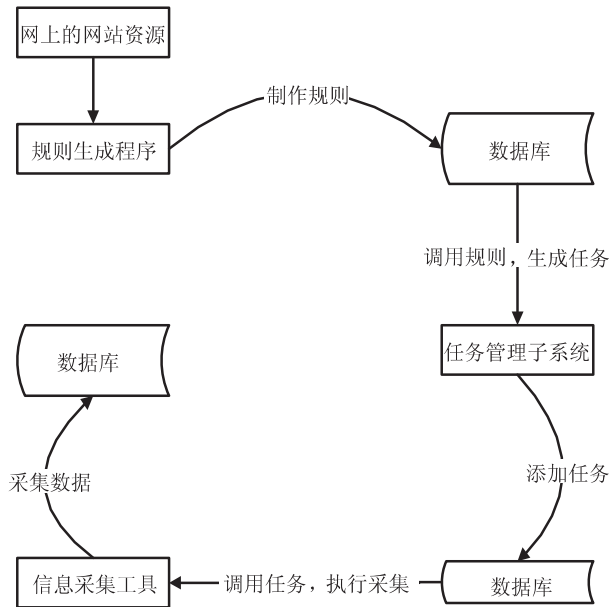


图 2 资源管理系统工作流程

资源管理子系统是整个系统的调度控制中枢,该子系统又可分为网站资源管理、任务控制管理、自定义任务管理、自定义网页监测、执行结构管理、错误日志管理等主要功能模块。

1.5 中药信息发布及应用子系统

中药信息发布及应用系统其主要功能是把从互联网上收集整理到系统数据库中的中药数据,采用不同的形式对外发布,同时还提供包括对外交流互动在内的各种工具。

该子系统又分为前台发布系统和后台控制管理系统,前台发布系统主要实现中药信息栏目展示和广告友情链接展示、分类目录信息检索、信息浏览、关键词查询、定阅、推送等服务,后台管理系统主要完成对系统网站管理、信息管理、个人用户管理等一系列综合管理控制功能。

2 系统实现

基于中药主题的垂直搜索引擎系统是在 Visual Studio2008 开发平台上使用 Sql Server2005 数据库基础上编程实现的。

信息采集系统与信息处理系统相互协作完成了对互联网上中药信息的抽取入库后,最后要通过信息发布子系统,即基于 B/S 架构的系统平台把系统数据库中收集、整理的各种分类的中药数据,采用不同方式对外公布出去。该系统平台分为前台发布应用系统和后台控制管理系统,前台主要工作是对中药信息的发布、查询、定阅、分类检索、推送等服务,后台管理系统主要完成用户管理、网站管理、信息管理等一系列综合的管理控制功能。其设计实现的部分代码如下:

```
public static DataTable GetInfoTempByTagID (int thisPage, int
onpageNum, ref int totalNum, int TagsID, string KeyWord)
{
    string spname = "sp_GetInfoTempByTagID";
    string[] spParams = SPNAME_PARAMS[spname];
    return dbo.ExecuteSP(spname, spParams, new object[] { to-
talNum, TagsID, KeyWord, (thisPage == 1? 1:(thisPage - 1)
* onpageNum+1), onpageNum },out totalNum);
}

/// <summary>
/// 获得重要新闻列表
/// 该函数调用存储过程 ( sp_GetInfoTempByImportantNews
@ NeedTotalNum @ Keyword @ BeginIndex @ Count)
/// </summary>
/// <param name = "thisPage">当前要获取第几页的数据</
param>
/// <param name = "onpageNum">每页数据条数</param>
/// <param name = "totalNum">数据总数,若不需要得到符
合条件的数据总数,请传入 0 以提高读取速度</param>
/// <param name = "@ TagsID">标签 ID</param>
/// <param name = "KeyWord">关键字</param>
/// <returns></returns>
```

当用户想了解中药相关的具体信息,点击进入自己所关心的列表页,从信息监测平台列表页中可以看到从网络上采集并且入库的信息详情,其中包括标题、信息内容详情、所发布的日期、信息来源网址以及浏览次数等,该系统可以完成对网络上有关中药信息的实时监测,达到了预定的效果,可以为中药企业以及政府部门带来便利,同时也可以给企业带来效益。

3 结束语

文中提出了一种基于中药信息化的垂直搜索引擎的解决方案,并实现了该方案。通过试运行,文中提出的实现方案是切实可行的,系统基本上达到了预期的设计目标,但仍有需要改进的地方:如在去重的问题上需要进一步完善,现在的系统针对互联网上的中药信息列表页仅能达到相同 URL 地址及相同标题去重;应该尽量减少编程,因为编程实现的权限不仅复杂而且难以维护,难免出现漏洞。需进一步研究如何将实际需求中的权限用适当的安全控制来实现。

参考文献:

[1] Darie C, Tosa F C, Brinzarea B. Ajax 与 PHP Web 开发[M]. 北京:人民邮电出版社,2007:211-215.
[2] 彭 洁,赵 辉,齐 娜. 信息资源整合技术[M]. 北京:科学技术文献出版社,2008.

面关于差分进化算法的收敛性分析是在 $\Psi \subset \text{SP}$ 的前提下进行的。

定理4 当 $\Psi \subset \text{SP}$ 时,即最优种群状态集合为种群状态空间的真子集,差分进化算法无法保证全局收敛。

证明:因为 $\Psi \subset \text{SP}$,因此在可行解空间内至少存在一点 x ,使得 $f(x) > f(g^*)$ 。构造一个种群状态 $S_i = (x_{i1}, x_{i2}, \dots, x_{iN_p})$,并且其所有的 N_p 个体的状态都为 x 。并假定 $\text{SP}_i = \{S_i\}$ 为状态空间的一个子集。因为 $\exists S_i \notin \text{SP}_i$ 和 $\forall S_i \in \text{SP}_i$,则种群状态转移 $S_i \rightarrow S_i$ 的概率为 $P(S_i \rightarrow S_i) = \prod_{k=1}^{N_p} P(x_{ik} \rightarrow x_{ik}) > 0$ 。但对于 $\forall S_j \notin \text{SP}_i$ 和 $\forall S_i \in \text{SP}_i$,若要实现种群的状态转移 $S_i \rightarrow S_j$,则种群状态 S_i 中至少需要一个个体 I_i 发生如下状态转移 $x_{ii} \rightarrow x_j$ 并且 $f(x_j) \neq f(x_{ii})$ 。但是由差分进化算法的迭代公式和定理1与2可知, $P(x_{ii} \rightarrow x_{ii}) = 1$,并且可以得到 $P(S_i \rightarrow S_i) = 1$,因此 $P(S_i \rightarrow S_j) = 0$ 。由于 $f(x) > f(g^*)$,所以 SP_i 是完全不同于最优状态集合 Ψ ,即 $\text{SP}_i \cap \Psi = \emptyset$ 。由定理3可知,种群状态序列 $\{S(t) \mid t \geq 0\}$ 为有限齐次马尔可夫链。又因为种群状态序列 $\{S(t) \mid t \geq 0\}$ 这一状态空间中存在真子集 SP_i ,且 $\text{SP}_i \cap \Psi = \emptyset$ 。 SP_i 为这一有限齐次马尔可夫链的吸收态。系统一旦进入该吸收态,将无法跳出。综上所述,可知当 $\Psi \subset \text{SP}$ 时,差分进化算法无法保证全局收敛。至此定理证毕。

3 结束语

文中对差分进化算法作了深入的理论分析,这将为在具体工程实践中改进差分进化算法提供指导思想。首先,对差分进化算法的基本概念作了严格的数学描述和定义;然后建立差分进化算法的 Markov 链模型,并证明差分进化算法优化过程中群体状态的转移过程是有限齐次 Markov 链;最后,分析差分进化算法的收敛性,并证明了差分进化算法无法保证全局收敛。虽然差分进化算法无法保证全局收敛,但是仿真实验

和实际应用显示该优化算法具有很强的全局搜索能力。粒子群算法同样无法保证全局收敛,但是也被广泛地应用于解决各种各样的全局优化问题,并显示出搜索能力强、稳定性好等优势。

参考文献:

[1] Storn R, Price K. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces [J]. Global Optimization, 1997, 11(4): 341-359.

[2] Paterlinia S, Krinkb T. Differential evolution and particle swarm optimization in partitional clustering[J]. Computational Statistics & Data Analysis, 2006, 50(5): 1220-1247.

[3] Mandal K K, Chakraborty N. Short-term combined economic emission scheduling of hydrothermal power systems with cascaded reservoirs using differential evolution[J]. Energy Conversion and Management, 2009, 50(1): 97-104.

[4] Yuan X H, Wang L, Zhang Y C, et al. A hybrid differential evolution method for dynamic economic dispatch with valve-point effects [J]. Expert System with Application, 2009, 36(2): 4042-4048.

[5] Pan Q K, Wang L, Qian B. A novel differential evolution algorithm for bi-criteria no-wait flow shop scheduling problems [J]. Computer & Operations Research, 2009, 36(8): 2498-2511.

[6] 包融, 王伟业, 顾汉杰, 等. 订单可分的协作计划模型及其进化算法[J]. 计算机技术与发展, 2010, 20(10): 58-61.

[7] Fan H Y, Lampinen J. A trigonometric mutation operation to differential evolution[J]. Journal of global optimization, 2003, 27(1): 105-129.

[8] Lakshminarasimman L, Subramanian S. A modified hybrid differential evolution for short-term scheduling of hydrothermal power systems with cascaded reservoirs [J]. Energy Conversion and Management, 2009, 49(10): 2513-2521.

[9] Hu C, Yan X F. An immune self-adaptive differential evolution algorithm with application to estimate kinetic parameters for homogeneous mercury oxidation [J]. Chinese Journal of Chemical Engineering, 2009, 17(2): 232-240.

.....

(上接第 61 页)

[3] 刘刚, 于力超. 搜索引擎中网络蜘蛛的设计与实现[J]. 电脑与信息技术, 2007, 15(4): 39-42.

[4] Roslak J. Active lighting systems for improved road safety [C]//Proc. of IEEE Intelligent Vehicles Symposium. [s. l.]: [s. n.], 2004: 682-685.

[5] 陈勇, 刘勇. 中医药主题搜索网络机器人的设计与实现[J]. 计算机技术与发展, 2010, 20(5): 162-166.

[6] 尹西杰. 基于智能 Agent 的 Web 个性化信息检索系统[D]. 济南: 山东大学, 2006.

[7] 何顺志, 徐文芬, 黄敏, 等. 贵州中药资源种类与分布的研究[J]. 世界科学技术: 中医药现代化, 2005, 7(2): 95-102.

[8] 胡元军. 基于 Agent 的分布式专业信息采集系统[D]. 北京: 北京化工大学, 2007.

[9] 王汝传, 徐小龙, 黄海平. 智能 AGENT 及其在信息网络中的应用[M]. 北京: 北京邮电大学出版社, 2007.

[10] 刘迁, 贾惠波. 中文信息处理中自动分词技术的研究与展望[J]. 计算机工程与应用, 2006, 42(3): 175-177.

基于中药主题的垂直搜索引擎设计与实现

作者:

[彭嫚](#), [陆安江](#), [董旭辉](#), [PENG Man](#), [LU An-jiang](#), [DONG Xu-hui](#)

作者单位:

[彭嫚, PENG Man \(贵州师范大学 机械与电气工程学院, 贵州 贵阳, 550014\)](#), [陆安江, LU An-jiang \(贵州大学 计算机学院, 贵州 贵阳, 550025\)](#), [董旭辉, DONG Xu-hui \(贵州省移动分公司, 贵州 贵阳, 550001\)](#)

刊名:

[计算机技术与发展](#)

ISTIC

英文刊名:

[Computer Technology and Development](#)

年, 卷(期):

2013 (8)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjtz201308015.aspx