(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107038172 A
(43)申请公布日 2017.08.11

(21)申请号 201610078536.4

(22)申请日 2016.02.03

(71)申请人 中国科学院沈阳自动化研究所
地址 110016 辽宁省沈阳市南塔街114号

(72)发明人 佟星 刘阳 张天石 曾鹏
于海斌 顾峰硕 许秀珍

(74)专利代理机构 沈阳科苑专利商标代理有限公司 21002
代理人 许宗富

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/27(2006.01)

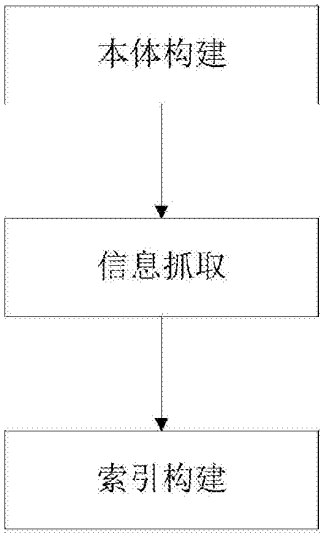
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种基于语义的油田搜索引擎构建方法

(57)摘要

本发明涉及一种基于语义的油田搜索引擎构建方法，包括以下步骤：首先构建油田本体库，然后抓取油田领域网络中的信息，最后构建油田搜索引擎。本发明实现一个面向石油行业的语义搜索引擎，提供一个专业的石油信息搜索网站，从互联网上和企业内网搜索到相关的网页信息；不仅可以根据用户的需要，对信息进行搜索，然后将通过语义检索到的信息返回给用户，从而实现为各油田网络用户和科研、生产人员方便地提供所需要的信息；还可以让企业管理层对搜索的信息进行分析，提取分析结果，进而为企业的发展提供方向，为信息化的管理等提供依据。



CN 107038172 A

1.一种基于语义的油田搜索引擎构建方法,其特征在于:包括以下步骤:

首先构建油田本体库,然后抓取油田领域网络中的信息,最后构建油田搜索引擎。

2.根据权利要求1所述的基于语义的油田搜索引擎构建方法,其特征在于:所述油田本体库的构建包括以下过程:

首先,定义类和类的层次,确保类的等级是“is-a”关系;

其次,定义类的属性和属性分面,类的属性根据内在特性、外在特性、局部关系和个体关系四种特性进行分层和定义;属性分面包括属性取值类型、允许的取值、取值个数、同义词和其它属性取值特征;

最后,实现油田本体库的构建。

3.根据权利要求1所述的基于语义的油田搜索引擎构建方法,其特征在于:所述抓取油田领域网络中的信息包括以下步骤:

步骤1:将种子网站作为抓取源头,选取目标URL,按评分由高到低依次选取若干URL;

步骤2:对蜘蛛线程进行调度,如果缓存中文件数目大于所选取URL的1/2时,蜘蛛线程休眠;否则蜘蛛线程从休眠状态唤醒,并将URL分配给该蜘蛛线程,开始爬取油田领域网络中的信息;

步骤3:如果URL分配完毕,则完成抓取过程,蜘蛛线程进入休眠状态。

4.根据权利要求3所述的基于语义的油田搜索引擎构建方法,其特征在于:所述URL与域名为一一对应关系。

5.根据权利要求1所述的基于语义的油田搜索引擎构建方法,其特征在于:所述构建油田搜索引擎包括以下步骤:

步骤1:对网页文本进行过滤,把无用广告和弹出窗口过滤掉;

步骤2:对过滤后的网页文本进行分词,并将分词结果与本体库中的油田专业词汇进行匹配,保留匹配结果;

步骤3:在本体库中找出与匹配结果同义的关键词,即为关键词的同义词;

步骤4:将关键词以及关键词的同义词利用lucene构建倒排索引。

一种基于语义的油田搜索引擎构建方法

技术领域

[0001] 本发明涉及油田搜索领域，具体地说是一种基于语义的油田搜索引擎构建方法。

背景技术

[0002] 随着因特网应用的普及，Internet已发展成为一个巨大的分布式信息空间。由于Internet面向社会和个人，信息的产生不受约束，人类的信息世界呈现出前所未有的复杂现象。Web信息的大容量、异构性、分布性、动态性等特点造成了“信息过载”，如何有效地为用户提供Web信息检索已经成为一项重要的研究课题。20世纪60年代以来，信息检索领域取得了许多研究成果，这些成果被成功地应用在Web上，产生了搜索引擎，例如雅虎，谷歌，百度等。大量各种语言的搜索引擎的出现，使这些成熟的搜索引擎系统也成为信息的宝贵资源，成为人们获取信息的重要途径。目前，在网上常见的检索工具有几十种，按检索内容可将其大致分为3类：综合型检索工具，主题型检索工具，特殊型检索工具。综合型检索工具应用的最为广泛，这种类型检索工具可以就任何领域、任何关键词的组合进行检索，但返回结果往往成千上万，所能够检索的内容包罗万象，而检索某一特定专业网络信息的效果不太理想。主题型检索工具是针对某一特定领域的信息进行检索，由于每个学科专业都有自己专门的词汇和用语，主题型检索工具使用与之相应的标引和检索语言进行检索，其效果优于综合型检索工具。专题型检索工具为有效利用网上科技信息、共享最新科技成果的工具，但是，目前网上专题型检索工具的数量不多，无法满足社会各个行业工作者的需求。特殊型检索工具是指用来在网上检索地址、电话号码、图片、地图等特殊信息的检索工具，特殊型检索工具的开发往往受到各方面的限制。

[0003] 大型油田都具有信息来源分散、数据保密级别高和通用性较低等特点。在开展油气田数字化建设的程中，由于各个部门的要求的工作性质不一样，因此在开发过程中针对许多部门的个别要求建立了应用数据库和工作文档，虽然这些信息化的数据都能很好为各个部门服务，但也不可避免的带来了一些问题，比如信息分布散、资源共享难、建设周期长等。这些问题的存在，严重影响和制约了油田科技工作者的工作效率，也使得油气田资源并没有得到充分的开发与利用。

[0004] 在油气田开发过程中，为了进一步的为地质勘探，油田开发提供科学的方法和强有力的数据，面向油田知识的信息检索工具的研发和应用已经破在眉睫了。

发明内容

[0005] 针对现有技术的不足，本发明提供一种能够方便的实现资源共享，信息整合的基于语义的油田搜索引擎构建方法。

[0006] 本发明为实现上述目的所采用的技术方案是：

[0007] 一种基于语义的油田搜索引擎构建方法，包括以下步骤：

[0008] 首先构建油田本体库，然后抓取油田领域网络中的信息，最后构建油田搜索引擎。

[0009] 所述油田本体库的构建包括以下过程：

[0010] 首先，定义类和类的层次，确保类的等级是“is-a”关系；

[0011] 其次，定义类的属性和属性分面，类的属性根据内在特性、外在特性、局部关系和个体关系四种特性进行分层和定义；属性分面包括属性取值类型、允许的取值、取值个数、同义词和其它属性取值特征；

[0012] 最后，实现油田本体库的构建。

[0013] 所述抓取油田领域网络中的信息包括以下步骤：

[0014] 步骤1：将种子网站作为抓取源头，选取目标URL，按评分由高到低依次选取若干URL；

[0015] 步骤2：对蜘蛛线程进行调度，如果缓存中文件数目大于所选取URL的1/2时，蜘蛛线程休眠；否则蜘蛛线程从休眠状态唤醒，并将URL分配给该蜘蛛线程，开始爬取油田领域网络中的信息；

[0016] 步骤3：如果URL分配完毕，则完成抓取过程，蜘蛛线程进入休眠状态。

[0017] URL与域名为一一对应关系。

[0018] 所述构建油田搜索引擎包括以下步骤：

[0019] 步骤1：对网页文本进行过滤，把无用广告和弹出窗口过滤掉；

[0020] 步骤2：对过滤后的网页文本进行分词，并将分词结果与本体库中的油田专业词汇进行匹配，保留匹配结果；

[0021] 步骤3：在本体库中找出与匹配结果同义的关键词，即为关键词的同义词；

[0022] 步骤4：将关键词以及关键词的同义词利用lucene构建倒排索引。

[0023] 本发明具有以下有益效果及优点：

[0024] 本发明提高工作效率，为油田信息提供专业的信息检索，提高采油厂经济效益。

附图说明

[0025] 图1是本发明的方法流程图；

[0026] 图2是本发明的抓取流程图；

[0027] 图3是油田本体库构建图。

具体实施方式

[0028] 下面结合附图及实施例对本发明做进一步的详细说明。

[0029] 如图1所示为本发明的方法流程图。

[0030] 油田语义搜索引擎总体构建流程

[0031] 石油行业语义搜索引擎的处理流程分析如下：首先，构建油田语义本体；其次，由网络信息采集器依据URL专业数据库中配置的地址，抓取石油行业相关的互联网和企业内网上的网页，保存到本体文件中，网页处理后转变成纯文本格式；再次，进行分词，将有意义的关键词提取出来，经过索引后存入索引数据库，另一方面，将语义本体库中的同义词词库信息与关键词进行关联，并将这些同义词更新、补充到系统的专业数据库中；最后，当用户提出查询请求时，首先根据索引数据库找到相应数据，然后由纯文本文件生成摘要，并同时定位到网页文件位置，以便用户进一步地浏览。根据上述流程，语义搜索引擎的模块进一步细分为：本体构建、信息抓取、索引构建、信息检索。该搜索引擎为油田专业领域的专业搜索

引擎;其次,该搜索引擎实现了语义关联检索。为保证这两点的完成,专利设计了专业的油田本体库,并根据本体库信息完成了语义数据的搜集,索引的构建和查询检索模块。

[0032] 如图2所示为本发明的抓取流程图。

[0033] 语义搜索引擎的搜索范围较小,搜索效率高,通过从URL数据库中获得种子网页作为搜索的起点,索引的内容只限于特定主题或专门领域,因此垂直搜索的抓取更倾向于结构化数据和元数据。垂直搜索引擎抓取是通过蜘蛛线程的工作来完成的。

[0034] 网页蜘蛛主要读取程序的配置文件,连接数据库并获取未爬取的网页URL,然后分配给各个爬取线程。爬取线程根据自己分配到的URL逐一爬取,当爬完时调用控制模块的URL分配功能为自己分配新的URL。爬取时需要及时更新对应的URL的信息;对于已经爬取过的网页,蜘蛛会根据网页是否更新(size是否改变)决定是否抓取并同时修改URL中的status,方便处理器处理。

[0035] 基本流程简要描述如下:

[0036] (1)将种子网站作为抓取源头,选取要抓取的URL,根据一个域名对应一个URL的原则,先选取2000条评分高的URL;

[0037] (2)对蜘蛛线程进行调度,当缓存中文件数目达到2500时蜘蛛线程休眠,低于2500时,即从休眠中唤醒,将URL分配给该蜘蛛线程,开始爬取;

[0038] (3)线程状态判断器判断URL是否分配完毕,如果分配完毕回到(2)重复执行,否则进入下一步(4);

[0039] (4)抓取完毕,蜘蛛线程进入休眠。

[0040] 索引构建

[0041] 垂直搜索引擎的索引构建器对网络蜘蛛抓取到的网页进行处理,任务包括文本处理、文档分析和分词、语义同义词加入、构建倒排表索引。主要任务包括以下四个方面:

[0042] (1)网页文本的过滤和预处理,自动把无用广告、弹出窗口等过滤掉,留下主体部分供后续处理;

[0043] (2)文档分析和分词,从过滤后的网页中分出主题内容经过词频统计出通用词、专业词出现的频度和次数,以及分类特征向量数据;网页的相关性的分析技术和算法,即判别该网页是否是与石油行业相关的网页;

[0044] (3)语义同义词加入,根据分词得到的油田关键字,在语义同义词词库中,找到这些关键字的同义词。

[0045] (4)将关键字以及关键字的同义词利用lucene构建倒排索引。

[0046] 如图3所示为本发明的油田本体构建图。

[0047] 本体的目标是获取、描述和表示相关领域的知识,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和词汇间相互关系的明确定义,本系统中利用OWL网络本体语言建立了油田中油井、注水井、储油罐等词汇及词汇间的关系,并利用protégé建立了油田中的本体模型。将广义数据进行层级化的划分,形成了类与子类的关系,例如日产量、含水率以及月产量为产油量的子类;油井这个类中包含了设备序号、工作状态、产油量、油压、套压等子类。

[0048] 首先,定义类和类的层次,类的定义要明确,保证无歧义,类的层次定义可根据具体情况选择自顶向下,自底向上和综合法,并确保类的等级是“is-a”关系,避免类循环和保

证一个类应有直接子类数量为2-12个;其次,定义类属性和属性分面,类的属性可根据内在特性,外在特性,局部和个体关系四种特性进行分层和定义,同时根据属性值的特征定义属性的约束及分面,一个属性可能由多个分面组成,包括属性取值类型,允许的取值,取值个数和其它属性取值特征;最后,本体实现,用形式化语言描述传感器本体。

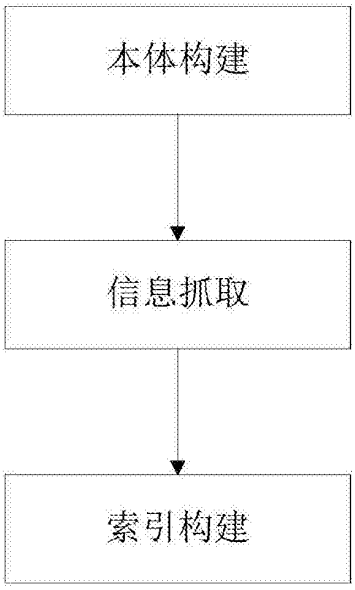

本体构建
信息抓取
索引构建

图1

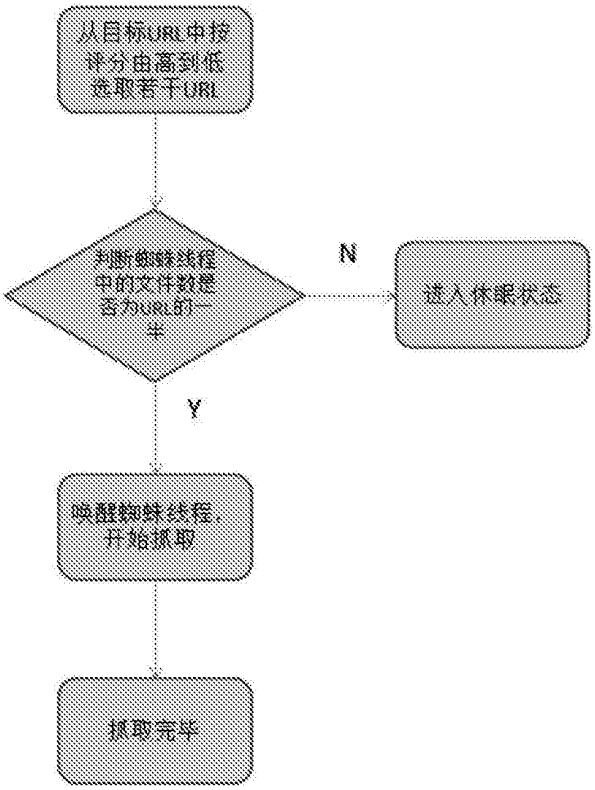

从目标URL中按评分由高到低选取若干URL
判断蜘蛛线程中的文件数是否为URL的一半
N
进入休眠状态
Y
唤醒蜘蛛线程，开始抓取
抓取完毕

图2

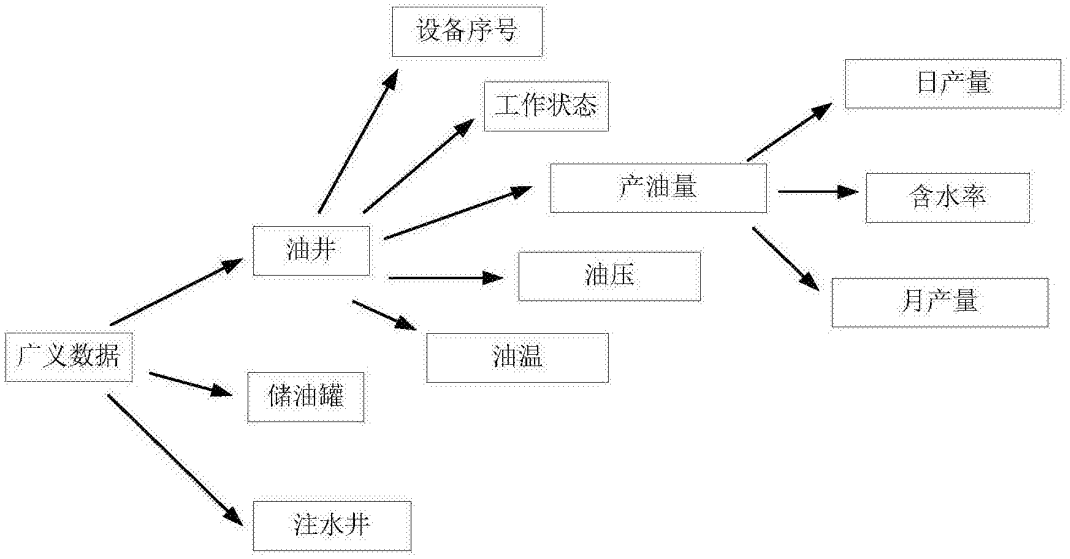

设备序号
工作状态
产油量
日产量
含水率
月产量
油井
油压
油温
广义数据
储油罐
注水井

图3