(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 103425742 A
(43)申请公布日 2013.12.04

(21)申请号 201310298439.2

(22)申请日 2013.07.16

(71)申请人 北京中科汇联信息技术有限公司
地址 100083 北京市海淀区北四环中路 229 号海泰大厦北 527

(72)发明人 乔亚飞 田文奇 胡绍武 孟凡兴 游世学 赵丽娜

(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319
代理人 苏培华

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

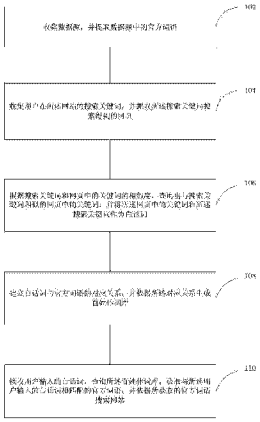
权利要求书2页 说明书10页 附图3页

(54)发明名称

一种网站的搜索方法和装置

(57)摘要

本申请提供了一种网站的搜索方法和装置，其中，所述的方法包括：收集数据源，并提取数据源中的官方词语；收集用户在所述网站的搜索关键词，并抓取所述搜索关键词搜索得到的网页；根据搜索关键词和网页中的关键词的相似度，查询出与搜索关键词相似的网页中的关键词，并将所述网页中的关键词和所述搜索关键词作为白话词；建立白话词与官方词语的对应关系，并依据所述对应关系生成百姓体词库；接收用户输入的白话词，查询所述百姓体词库，获取与所述用户输入的白话词相匹配的官方词语，并依据所获取的官方词语搜索网站。因此，本申请能够解决目前搜索结果查询不全或不准确、搜索结果不实用的问题。



CN 103425742 A

1. 一种网站的搜索方法，其特征在于，包括：

收集数据源，并提取数据源中的官方词语；

收集用户在所述网站的搜索关键词，并抓取所述搜索关键词搜索得到的网页；

根据搜索关键词和网页中的关键词的相似度，查询出与搜索关键词相似的网页中的关键词，并将所述网页中的关键词和所述搜索关键词作为白话词；

建立白话词与官方词语的对应关系，并依据所述对应关系生成百姓体词库；

接收用户输入的白话词，查询所述百姓体词库，获取与所述用户输入的白话词相匹配的官方词语，并依据所获取的官方词语搜索网站。

2. 根据权利要求 1 所述的方法，其特征在于，所述建立白话词与官方词语的对应关系，包括：

抽取出通过白话词查询网站的网页内容，对查询的网页内容进行分词处理，查询分词后词元对应的官方词语，如果查询成功，则建立所述白话词与官方词语的对应关系；

所述依据所述对应关系生成百姓体词库，包括：根据语义分析和概率统计对所述白话词与官方词语的对应关系进行筛选，筛选后生成百姓体词库。

3. 根据权利要求 1 所述的方法，其特征在于，所述提取数据源中的官方词语，包括：

从数据源中提取数据信息；

判断所述数据信息中是否含有表示官方词语的标签，若含有，则直接提取所述标签；

若不含有，则对所述数据信息进行分析得出对应的官方词语。

4. 根据权利要求 1 所述的方法，其特征在于，所述抓取所述搜索关键词搜索得到的网页之前，还包括：

依据用户在所述搜索关键词搜索得到的网页的驻留时间对所述网页排序；

所述抓取所述搜索关键词搜索得到的网页包括：抓取所述搜索关键词搜索得到的排序后的部分网页。

5. 根据权利要求 1 所述的方法，其特征在于，所述查询所述百姓体词库，获取与所述用户输入的白话词相匹配的官方词语，包括：

对用户输入的白话词进行分词，拆分成词元；

在所述百姓体词库中查询所述词元对应的官方词语；

将词元对应的官方词语合并成与所述用户输入的白话词相匹配的官方词语。

6. 一种网站的搜索装置，其特征在于，包括：

收集官方词语模块，用于收集数据源，并提取数据源中的官方词语；

收集白话词模块，包括：

收集子模块，用于收集用户在所述网站的搜索关键词；

抓取子模块，用于抓取通过所述搜索关键词搜索得到的网页；

生成白话词子模块，用于根据搜索关键词和网页中的关键词的相似度，查询出与搜索关键词相似的网页中的关键词，并将所述网页中的关键词和所述搜索关键词作为白话词；

生成百姓体词库模块，用于建立白话词与官方词语的对应关系，并依据所述对应关系生成百姓体词库；

搜索模块，用于接收用户输入的白话词，查询所述百姓体词库，获取与所述用户输入的白话词相匹配的官方词语，并依据所获取的官方词语搜索网站。

7. 根据权利要求 6 所述的装置，其特征在于，所述生成百姓体词库模块包括：

对应关系建立子模块，用于抽取出通过白话词查询网站的网页内容，对查询的网页内容进行分词处理，查询分词后词元对应的官方词语，如果查询成功，则建立所述白话词与官方词语的对应关系；

筛选子模块，用于根据语义分析和概率统计对所述白话词与官方词语的对应关系进行筛选，筛选后生成百姓体词库。

8. 根据权利要求 6 所述的装置，其特征在于，所述收集官方词语模块包括：

提取子模块，用于收集数据源，并从数据源中提取数据信息；

判断子模块，用于判断所述数据信息中是否含有表示官方词语的标签，若含有，则直接提取所述标签；若不含有，则对所述数据信息进行分析得出对应的官方词语。

9. 根据权利要求 6 所述的装置，其特征在于，所述收集白话词模块，还包括：

排序子模块，用于依据用户在所述搜索关键词搜索得到的网页的驻留时间对所述网页排序；

所述抓取子模块，用于抓取所述搜索关键词搜索得到的排序后的部分网页。

10. 根据权利要求 6 所述的装置，其特征在于，所述搜索模块包括：

分词子模块，用于接收用户输入的白话词，对用户输入的白话词进行分词，拆分成词元；

查询官方词语子模块，用于在所述百姓体词库中查询所述词元对应的官方词语；

生成官方词语子模块，用于将词元对应的官方词语合并成与所述用户输入的白话词相匹配的官方词语。

一种网站的搜索方法和装置

技术领域

[0001] 本申请涉及网站技术，特别是涉及一种网站的搜索方法和装置。

背景技术

[0002] 我国的政府网站普遍经过了“政府名片”、“新闻网站阶段”、“信息公开、在线服务、政民互动”三大定位阶段，在当前阶段每个综合性政府网站都积累了丰富的便民信息和服务，但这种“信息过载”却给网站用户查找信息带来了很大困扰。

[0003] 目前政府门户网站信息更新速度快，信息量非常大，用户在查找信息时，最近发布的很多信息查询不到，搜索结果仍停留在至少半年以前。

[0004] 有些政府门户网站检索结果非常多，但是很多检索结果和搜索关键词没有关联，或者通过标题根本看不出搜索关键词和检索结果之间的必然关系，带给用户的体验是搜索结果不准确。

[0005] 很多用户在政府门户网站查找信息时，最关注的是跟办事指南、答疑解惑等政府服务相关的内容，如教育、医疗、社保、住房、交通等与百姓生活密切相关的问题，但是搜索结果却往往大失所望。排在最前面的通常都是和新闻动态类相关的信息，而服务类的信息往往排在最后或者根本无法搜索到结果。

[0006] 因此，目前政府门户网站领域存在的主要问题有搜索结果查询不全、搜索结果查询不准确、搜索结果不实用。

发明内容

[0007] 本申请提供了一种网站的搜索方法和装置，以解决目前搜索结果查询不全或不准确、搜索结果不实用的问题。

[0008] 为了解决上述问题，本申请公开了一种网站的搜索方法，包括：

[0009] 收集数据源，并提取数据源中的官方词语；

[0010] 收集用户在所述网站的搜索关键词，并抓取所述搜索关键词搜索得到的网页；

[0011] 根据搜索关键词和网页中的关键词的相似度，查询出与搜索关键词相似的网页中的关键词，并将所述网页中的关键词和所述搜索关键词作为白话词；

[0012] 建立白话词与官方词语的对应关系，并依据所述对应关系生成百姓体词库；

[0013] 接收用户输入的白话词，查询所述百姓体词库，获取与所述用户输入的白话词相匹配的官方词语，并依据所获取的官方词语搜索网站。

[0014] 优选的，所述建立白话词与官方词语的对应关系，包括：

[0015] 抽取出通过白话词查询网站的网页内容，对查询的网页内容进行分词处理，查询分词后词元对应的官方词语，如果查询成功，则建立所述白话词与官方词语的对应关系；

[0016] 所述依据所述对应关系生成百姓体词库，包括：根据语义分析和概率统计对所述白话词与官方词语的对应关系进行筛选，筛选后生成百姓体词库。

[0017] 优选的，所述提取数据源中的官方词语，包括：

[0018] 从数据源中提取数据信息；

[0019] 判断所述数据信息中是否含有表示官方词语的标签，若含有，则直接提取所述标签；

[0020] 若不含有，则对所述数据信息进行分析得出对应的官方词语。

[0021] 优选的，所述抓取所述搜索关键词搜索得到的网页之前，还包括：

[0022] 依据用户在所述搜索关键词搜索得到的网页的驻留时间对所述网页排序；

[0023] 所述抓取所述搜索关键词搜索得到的网页包括：抓取所述搜索关键词搜索得到的排序后的部分网页。

[0024] 优选的，所述查询所述百姓体词库，获取与所述用户输入的白话词相匹配的官方词语，包括：

[0025] 对用户输入的白话词进行分词，拆分成词元；

[0026] 在所述百姓体词库中查询所述词元对应的官方词语；

[0027] 将词元对应的官方词语合并成与所述用户输入的白话词相匹配的官方词语。

[0028] 为了解决上述问题，本申请公开了一种网站的搜索装置，包括：

[0029] 收集官方词语模块，用于收集数据源，并提取数据源中的官方词语；

[0030] 收集白话词模块，包括：

[0031] 收集子模块，用于收集用户在所述网站的搜索关键词；

[0032] 抓取子模块，用于抓取通过所述搜索关键词搜索得到的网页；

[0033] 生成白话词子模块，用于根据搜索关键词和网页中的关键词的相似度，查询出与搜索关键词相似的网页中的关键词，并将所述网页中的关键词和所述搜索关键词作为白话词；

[0034] 生成百姓体词库模块，用于建立白话词与官方词语的对应关系，并依据所述对应关系生成百姓体词库；

[0035] 搜索模块，用于接收用户输入的白话词，查询所述百姓体词库，获取与所述用户输入的白话词相匹配的官方词语，并依据所获取的官方词语搜索网站。

[0036] 优选的，所述生成百姓体词库模块包括：

[0037] 对应关系建立子模块，用于抽取出通过白话词查询网站的网页内容，对查询的网页内容进行分词处理，查询分词后词元对应的官方词语，如果查询成功，则建立所述白话词与官方词语的对应关系；

[0038] 筛选子模块，用于根据语义分析和概率统计对所述白话词与官方词语的对应关系进行筛选，筛选后生成百姓体词库。

[0039] 优选的，所述收集官方词语模块包括：

[0040] 提取子模块，用于收集数据源，并从数据源中提取数据信息；

[0041] 判断子模块，用于判断所述数据信息中是否含有表示官方词语的标签，若含有，则直接提取所述标签；若不含有，则对所述数据信息进行分析得出对应的官方词语。

[0042] 优选的，所述收集白话词模块，还包括：

[0043] 排序子模块，用于依据用户在所述搜索关键词搜索得到的网页的驻留时间对所述网页排序；

[0044] 所述抓取子模块，用于抓取所述搜索关键词搜索得到的排序后的部分网页。

[0045] 优选的，所述搜索模块包括：

[0046] 分词子模块，用于接收用户输入的白话词，对用户输入的白话词进行分词，拆分成词元；

[0047] 查询官方词语子模块，用于在所述百姓体词库中查询所述词元对应的官方词语；

[0048] 生成官方词语子模块，用于将词元对应的官方词语合并成与所述用户输入的白话词相匹配的官方词语。

[0049] 与现有技术相比，本申请包括以下优点：

[0050] 本申请内嵌的百姓体词库主要由官方词语和白话词构成，通过将日常工作生活中常用的白话词与政府办事服务事项中的官方词语建立对应关系，从而解决用户对政府网站业务的“理解”障碍和搜索结果不准确的问题，通过对白话词进行分词、拆分成词元，根据所述词元在百姓体词库中查询与所述词元对应的官方词语，将词元对应的官方词语进行合并生成与白话词相匹配的官方语言，从而实现了用户搜索快速、搜索结果准确。

附图说明

[0051] 图 1 是本申请实施例所述一种网站的搜索方法的流程图；

[0052] 图 2 是本申请另一实施例所述一种网站的搜索方法的流程图；

[0053] 图 3 是本申请实施例所述一种网站的搜索装置的结构图。

具体实施方式

[0054] 为使本申请的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本申请作进一步详细的说明。

[0055] 目前在政府门户网站领域的搜索引擎系统都是简单运用了将前台输入关键词与后台索引库中的词语进行匹配，然后给出对应的搜索结果。利用此原理来实现搜索功能，搜索的体验极差，基本可以概括为“冷冰冰、慢吞吞、晕乎乎”，即输入搜索关键词时没有任何引导提示，搜索响应极其慢，用户想要的信息或服务搜索不到。可以发现找不到信息的根本矛盾在于社会公众对政府业务的认知障碍，也就是说一方面政府网站的信息分类用户看不懂，另一方面各个服务事项的称谓过于专业用户无法理解，所以公众对政府业务的描述和真实网站上的对应信息或服务在字面上是完全不同的，因此在大多数情况下公众的搜索行为将是无功而返。

[0056] 本申请通过政府的相关业务收集官方词语和根据用户的搜索关键词来收集白话词，建立所述白话词与所述官方词语的对应关系，并依据所述对应关系生成百姓体词库，根据用户输入的白话词进行分词、拆分分成词元，并根据所述词语在所述生成百姓体词库中进行匹配，最终返回官方词语，从而解决用户搜索结果查询准确性、搜索结果实用性的问题。

[0057] 参照图 1，示出了本申请实施例所述一种网站的搜索方法的流程图，具体可以包括：

[0058] 步骤 102，收集数据源，并提取数据源中的官方词语；

[0059] 根据政府的政策法规、办事事项等规定从互联网、新闻等途径收集数据信息，并从中提取数据信息中的官方词语。

[0060] 所述数据源可以指办公自动化(Office Automation, 简称 OA)系统、互联网、新闻、邮件、文件、业务系统、文档、声音图像等方式。

[0061] 所述官方词语是为适应管理国家事务的需要,在国家机关、正式文件、法律裁决及国际交往等官方场合中规定为有效语言的现象。官方词语也是一个国家的公民与其政府机关通讯时使用的语言。

[0062] 步骤 104,收集用户在所述网站的搜索关键词,并抓取所述搜索关键词搜索得到的网页 ;

[0063] 用户在网站上输入要查询的搜索关键词,在网站上就会显示与用户的搜索关键词相关的网页,然后用户点击搜索到的网页,把用户点击的网页保存到网页数据库。

[0064] 步骤 106,根据搜索关键词和网页中的关键词的相似度,查询出与搜索关键词相似的网页中的关键词,并将所述网页中的关键词和所述搜索关键词作为白话词 ;

[0065] 所述网页可以指搜索关键词搜索得到的网页。

[0066] 根据用户输入的搜索关键词,在网页中查询与用户输入的搜索关键词相似的网页中的关键词,如果网页中含有与用户输入的搜索关键词相似的搜索关键词,就把网页中的关键词提取出来作为白话词。

[0067] 所述白话词主要由用户输入的搜索关键词和网页中的关键词组成。

[0068] 步骤 108,建立白话词与官方词语的对应关系,并依据所述对应关系生成百姓体词库 ;

[0069] 政府的相关业务名称都有其严格的规范,这些叫法并不为用户熟知,然而用户对于政府的业务有着自己口头叫法,针对用户的政府业务进行梳理,形成白话词,并把白话词与官方词语建立对应关系,称为百姓体词库。

[0070] 例如 :官方词语为新生婴儿出生入户,白话词为上户口、办户口、小孩办户口,上户口、办户口、小孩办户口均可以对应官方词语新生婴儿出生入户,这样对应关系就可以指白话词与官方词语的对应关系,并把所述对应关系存储到网页数据库。

[0071] 步骤 110,接收用户输入的白话词,查询所述百姓体词库,获取与所述用户输入的白话词相匹配的官方词语,并依据所获取的官方词语搜索网站。

[0072] 例如 :用户输入的白话词“旅游护照”,把“旅游护照”拆分成“旅游”和“护照”,然后在所述百姓体词库中进行查询。

[0073] 综上所述,本申请实施例所述的一种网站的搜索方法主要包括以下优点 :

[0074] 本申请内嵌的百姓体词库主要由官方词语和白话词构成,通过将日常工作生活中常用的白话词与政府办事服务事项中的官方词语建立对应关系,从而解决用户对政府网站业务的“理解”障碍和搜索结果不准确的问题,通过对白话词进行分词、拆分成词元,根据所述词元在百姓体词库中查询与所述词元对应的官方词语,将词元对应的官方词语进行合并生成与白话词相匹配的官方语言,从而实现了用户搜索快速、搜索结果准确。

[0075] 基于以上内容,为使本领域技术人员更好地理解本申请,下面以一种政府门户网站的搜索方法为例对本申请进一步说明,参照图 2,其示出了本申请实施例所述一种网站搜索方法的流程图,具体如下 :

[0076] 步骤 202,收集官方词语 ;

[0077] 包括 :收集数据源,并提取数据源中的官方词语。

[0078] 所述提取数据源中的官方词语，可以包括以下过程：

[0079] 通过采集器从数据源中提取数据信息，在内容管理系统中返回提取数据信息的关键词，然后在采集层判断所述数据信息中是否含有表示官方词语的标签，若含有，则直接提取所述标签，并把标签和数据信息存储到网页数据库。若不含有，则在用户层对所述数据信息进行分析得出对应的官方词语，并把所述官方语言和数据信息存储到网页数据库。

[0080] 根据所述网页数据库建立官方词语的索引数据库。

[0081] 所述采集器的工作原理可以指按照一定的规则，自动的抓取万维网信息的程序或者脚本，最后为索引部分提供广泛的数据来源。例如：对跟踪页面的网页网址进行扩展的抓取。从一组要访问的网页地址链接开始，可以称这些网页地址为种子。爬虫访问这些链接，它辨认出这些页面的所有超链接，然后添加到这个网页地址列表，这些网页地址按照一定的策略反复访问。

[0082] 所述采集层的主要功能包括采集抓取、链接分析、采集规则、采集连接器、任务调度、采集过滤等。

[0083] 所述用户层的主要功能包括模糊检索、分类检索、组合检索、智能向导、百姓体匹配等。

[0084] 所述内容管理系统可以指集新闻管理、图库管理、视频管理、下载系统、作品管理、产品发布及留言板于一体的综合性内容管理系统。

[0085] 获取所述数据信息的标签有两种方式：

[0086] 1、直接提取所述标签，其中，标签代表的是官方词语；

[0087] 2、通过对所述数据信息分析得到标签；

[0088] 所述直接提取所述标签，例如：网页代码通常以超文本标记语言表示，如果网页代码中的标签所对应的内容是官方词语，则可以直接从所述网页代码中提取所述标签。所述超文标记语言是用于描述网页文档的一种标记语言。

[0089] 所述对所述数据信息进行分析可以指对所述数据信息中词频的分析，但不限于词频分析。例如：所述数据信息的总词数是 100，而词语“城市居民低保”在所述数据信息中出现了 5 次，那么词语“城市居民低保”在该所述数据信息中的词频是 0.05；而词语“注册企业”在所述数据信息中出现了 20 次，那么词语“注册企业”在该所述数据信息中的词频是 0.2；而词语“老年证结婚登记”在所述数据信息中出现了 2 次，那么词语“老年证结婚登记”在该所述数据信息中的词频是 0.02，挑选出词频高的词语作为所述数据信息的官方词语。基于以上分析，可以得出所述数据信息对应的官方词语为“工商注册登记”。

[0090] 步骤 204，收集白话词；

[0091] 包括：收集用户在所述网站的搜索关键词，并抓取通过所述搜索关键词搜索得到的网页；根据搜索关键词和网页中的关键词的相似度，查询出与搜索关键词相似的网页中的关键词，并将所述网页中的关键词和所述搜索关键词作为白话词。

[0092] 所述相似度属于现有技术，本领域技术人员可以采用现有技术中的任何一种相似度算法，本申请对此无需加以限制。

[0093] 优选的，所述抓取所述搜索关键词搜索得到的网页之前，还可以包括：

[0094] 依据用户所述搜索得到的网页的驻留时间对所述网页排序；

[0095] 所述抓取所述搜索关键词搜索得到的网页包括：依据所述搜索关键词搜索得到的

并且排序后的驻留时间长的网页进行抓取。

[0096] 例如：用户搜索得到的 3 个网页的驻留时间依次为 5 秒、6 秒和 300 秒，根据用户在这些网页的驻留时间长短排序，排序后的驻留时间依次为 300 秒、6 秒和 5 秒，然后抓取驻留时间为 300 秒的网页，并把网页存储到网页数据库。当然，上述驻留时间取值也仅仅用作参考。

[0097] 所述驻留时间可以指用户在会话序列中浏览某一 Web 页面的实际时间。

[0098] 本申请实施例通过对网页排序、并且依据所述搜索关键词搜索得到的并且排序后的驻留时间长的网页进行抓取，提高了用户在网站中搜索关键词的准确度。

[0099] 步骤 206，生成百姓体词库；

[0100] 包括：建立白话词与官方词语的对应关系，并依据所述对应关系生成百姓体词库；

[0101] 优选的，所述建立白话词与官方词语的对应关系，包括：

[0102] 抽取出通过白话词查询网站的网页内容，对查询的网页内容进行分词处理，查询分词后词元对应的官方词语，如果查询成功，则建立所述白话词与官方词语的对应关系。

[0103] 例如：白话词查询网站的网页内容为“山大 8 位教授获省突出贡献中青年专家称号”，把所述网页内容进行分词，可以拆分为“山 / 大 /8/ 位 / 教授 / 获 / 省 / 突出 / 贡献 / 中 / 青年 / 专家 / 称号”然后根据分词后的词元在所述百姓体词库中查找与“山”/“大”/“8”/“位”/“教授”/“获”/“省”/“突出”/“贡献”/“中”/“青年”/“专家”/”称号”对应的官方词语，查找成功，则建立了所述白话词与官方词语的对应关系。

[0104] 所述依据所述对应关系生成百姓体词库，包括：根据语义分析和概率统计对所述白话词与官方词语的对应关系进行筛选，筛选后生成百姓体词库。

[0105] 所述语义分析可以指分析所述白话词与所述官方词语是一一对应关系或者是多对一对应关系。

[0106] 例如：官方词语新生婴儿出生入户可以与白话词上户口、办户口、小孩办户口、宝宝办户口、出生入户、出生户口、小孩入户、新生儿入户、宝宝户口、新生儿户口、报户口、入户、生小孩和生孩子等对应。当用户在网站中输入搜索关键词报户口或者生孩子，在网站中就会出现官方词语新生婴儿出生入户办理手续的所有相关信息。

[0107] 所述概率统计可以指所述白话词与官方词语的对应关系在百姓体词库中出现的概率。

[0108] 例如：办户口在网站中搜索出现的概率为 0.01、小孩办户口在网站中搜索出现的概率为 0.1、宝宝办户口在网站中搜索出现的概率为 0.15、此时，就把在网站搜索中概率值偏小的删除。删除概率值偏小的原因可以指与用户搜索无关的事项或者用户误操作导致。当然，上述概率的取值也仅仅用作参考。

[0109] 为了方便本领域技术人员更好地理解本申请，通过一个示例更进一步说明本申请实施例中收集用户输入的白话词和建立所述白话词语与官方词语的对应关系，具体步骤可以包括：

[0110] 1、用户输入的白话词“办户口”；

[0111] 2、用户依次点击查询搜索结果网页中的 3 个网页，内容分别为应届毕业生办户口、外地人迁入户口、新生婴儿出生入户，其中驻留时间依次为 6 秒、5 秒和 300 秒；

[0112] 3、依据用户驻留时间对 3 个网页进行排序，并抓取用户驻留时间最长的那个网页；

[0113] 4、分析抽取网页中的关键词；

[0114] 5、依据编辑距离计算搜索关键词“办户口”和网页中的关键词的相似度，查询出与搜索关键词相似的网页中的关键词，查询出网页中的关键词中所有与“办户口”相似的词语为“婴儿入户”、“小孩办户口”；并将“办户口、“婴儿入户”和“小孩办户口”作为白话词进行查询；

[0115] 6、抽取出通过白话词查询网站的网页内容，对查询的网页内容进行分词处理，然后在官方词语的数据库中依次查找分词后的词汇，如果查找成功，则建立了所述白话词和官方词语的对应关系，并保存到网页数据库中。例如分词结果中一个官方词语为“新生婴儿出生入户”，该官方词语在数据库中查找成功，则建立“办户口”、“婴儿入户”、“小孩办户口”和“新生婴儿出生入户”的对应关系，也就是说白话词“办户口”、“婴儿入户”和“小孩办户口”对应的官方词语都是“新生婴儿出生入户”。

[0116] 所述编辑距离可以指两个字串之间，由一个转成另一个所需的最少编辑操作次数。

[0117] 所述百姓体词库还可以应用到搜素引擎、垂直检索等技术上。

[0118] 步骤 208，搜索百姓体词库；

[0119] 包括：接收用户输入的白话词，查询所述百姓体词库，获取与所述用户输入的白话词相匹配的官方词语，并依据所获取的官方词语搜索网站。

[0120] 优选的，所述查询所述百姓体词库，获取与所述用户输入的白话词相匹配的官方词语，包括：

[0121] 对用户输入的白话词进行分词，拆分成词元；

[0122] 在所述百姓体词库中查询所述词元对应的官方词语；

[0123] 将词元对应的官方词语合并成与所述用户输入的白话词相匹配的官方词语。

[0124] 例如：把用户输入的白话词进行分词，拆分词元，在所述百姓体词库中查询到所述词元对应的官方词语，根据得到的官方词语进行词语合并，对合并成的官方词语进行中文语法判断，如果符合中文语法，则形成了与白话词相匹配的官方词语；如果不符合中文语法，则对下一个合并词进行判断，重复执行此过程，直到完成所有的官方词语的合并方式。

[0125] 在具体实现时：当用户输入的白话词“旅游护照”，把“旅游护照”拆分成“旅游”和“护照”，然后在所述百姓体词库中进行查询，如果查询到与“旅游”和“护照”相匹配的官方词语为“首次”、“普通”、“护照”、“出国”和“审批”，根据得到的官方词语进行词语合并最后生成的官方词语为“居民首次申请普通护照出国审批”，然后就根据官方词语“居民首次申请普通护照出国审批”搜索网站。

[0126] 具体百姓体词库示例，请参考下表：

[0127] 表一

[0128]

专业词	白话词
新生婴儿出生入户	上户口，办户口，小孩办户口，宝宝办户口，出生入户，出生户口，小孩入户，新生儿入户，宝宝户口，新生儿户口，报户口，入户，生小孩，生孩子
政策外出生婴儿申报出生入户	计划外婴儿，计划外宝宝，生二胎，偷生婴儿，偷生小孩
居民身份证办理	办身份证，二代身份证，申请身份证，临时身份证，补办身份证，小孩身份证，高考身份证，身份证损坏，领身份证，身份证
住房保障	廉租房，廉租房配房，廉租房申请，廉租房审核，廉租房配租，廉租房配租审核，公租房，经适房，限价房，廉租房，保障房，安居房，安居型商品房，廉租房补贴，申请公租房，申请经适房，申请限价房，申请廉租房，申请保障房，申请安居房，申请安居型商品房，申请廉租房补贴，经济房
深圳居民首次申请普通护照出国审批	签护照，办护照，公派留学护照，进修护照，访问护照，劳务护照，公务护照，定居护照，探亲护照，访友护照，继承遗产护照，自费留学护照，就业护照，旅游护照，护照
医师执业注册	当医师，当大夫，当医生，医生注册，注册医生，医生执业，医师执业，执业医师，医师资格证
深圳市南山区居民与国内居民结婚登记	本地人领结婚证，本地人领红本，内地人领结婚证，外地人领结婚证，结婚领证，结婚登记，结婚领证，结婚注册，领结婚证，婚姻登记，，结婚，婚姻，结婚证，红本
城市居民最低生活保障审批	城市居民低保，城市居民生活保障，低保审批，城市低保审批，低保，城镇低保，城市低保
敬老优待证办理	老人优待，老人公交免费，老人买火车票，优待卡，老人补贴，老年优待证，老年人政策，老人门票，老年证，老年人逛公园，老年人健身，老年人逛景点，南山老年人，南山老人坐公交，敬老证，老人证，优待证
工商注册登记	注册公司，注册企业，开公司，做买卖，办企业，开企业，公司登记，企业登记，工商登记，公司注册，企业注册，工商注册，开分公司，分公司，分公司注册，分公司登记，有限责任公司，有限公司，股份公司，股份合作公司，股份有限公司，外商企业，外企，外商投资，外企代表机构，外企年检，个体户，三来一补，农业合作社，农民合作社，企业集团，动产抵押，股权质押，办公司，开办有限公司，开办有限责任公司，设立有限公司，成立有限公司，开公司登记注册，注册有限公司，设立分公司，成立分公司，开设分公司

[0129] 本申请中所述对用户输入的白话词进行分词可以采用机械分词算法、基于理解的分词算法、基于统计的分词算法。现以其中的机械分词算法为例进行介绍。

[0130] 机械分词算法按照一定策略将待切分字符串与机器里预先准备的词条进行匹配，然后找出一个最长的结果。

[0131] 机械分词算法分为最大匹配法和最小匹配法。常用的机械分词算法是正向最大匹配法和逆向最大匹配法。

[0132] 正向最大匹配法是基于词典的分词系统。所述正向最大匹配法就是要求每一句的分词结果中的词汇总量要最少。

[0133] 逆向最大匹配法与正向最大匹配法相反，从句子结尾开始进行分词。

[0134] 在实际应用中，本领域技术人员根据实际应用的情况可以采用不用的分词算法也是可行的，本申请对此无需加以限制。

[0135] 综上所述，本申请实施例所述的一种网站的搜索方法主要包括以下优点：

[0136] 本申请内嵌的百姓体词库主要由官方词语和白话词构成，通过将日常工作生活中常用的白话词与政府办事服务事项中的官方词语建立对应关系，从而解决用户对政府网站业务的“理解”障碍和搜索结果不准确的问题，通过对白话词进行分词、拆分成词元，根据所述词元在百姓体词库中查询与所述词语对应的官方词语，将词元对应的官方词语进行合并生成与白话词相匹配的官方语言，从而实现了用户搜索快速、搜索结果准确。

[0137] 需要说明的是，对于前述的方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作并不一定是本申请所必需的。

[0138] 基于上述方法实施例的说明，本申请还提供了相应的一种网站的搜索装置实施例，来实现上述方法实施例所述的内容。

[0139] 参照图 3，示出了本申请实施例所述一种网站的搜索装置结构图，具体可以包括：

[0140] 收集官方词语模块 300，用于收集数据源，并提取数据源中的官方词语；

[0141] 收集白话词模块，包括：

[0142] 收集子模块，用于收集用户在所述网站的搜索关键词；

[0143] 抓取子模块，用于抓取通过所述搜索关键词搜索得到的网页；

[0144] 生成白话词子模块 302，用于根据搜索关键词和网页中的关键词的相似度，查询出与搜索关键词相似的网页中的关键词，并将所述网页中的关键词和所述搜索关键词作为白话词；

[0145] 生成百姓体词库模块 304，用于建立白话词与官方词语的对应关系，并依据所述对应关系生成百姓体词库；

[0146] 搜索模块 306，用于接收用户输入的白话词，查询所述百姓体词库，获取与所述用户输入的白话词相匹配的官方词语，并依据所获取的官方词语搜索网站。

[0147] 在本申请的一种优选实施例中，所述生成百姓体词库模块 304 具体可以包括：

[0148] 对应关系建立子模块，用于抽取出通过白话词查询网站的网页内容，对查询的网页内容进行分词处理，查询分词后词元对应的官方词语，如果查询成功，则建立所述白话词与官方词语的对应关系；

[0149] 筛选子模块，用于根据语义分析和概率统计对所述白话词与官方词语的对应关系进行筛选，筛选后生成百姓体词库。

[0150] 在本申请的一种优选实施例中，所述收集官方词语模块 300 具体可以包括：

[0151] 提取子模块，用于收集数据源，并从数据源中提取数据信息；

[0152] 判断子模块，用于判断所述数据信息中是否含有表示官方词语的标签，若含有，则直接提取所述标签；若不含有，则对所述数据信息进行分析得出对应的官方词语。

[0153] 在本申请的一种优选实施例中，所述收集白话词模块 302 具体还包括：

[0154] 排序子模块，用于依据用户在所述搜索关键词搜索得到的网页的驻留时间对所述网页排序；

[0155] 所述抓取子模块，用于抓取所述搜索关键词搜索得到的排序后的部分网页。

[0156] 在本申请的一种优选实施例中，所述搜索模块 306 具体可以包括：

[0157] 分词子模块，用于接收用户输入的白话词，对用户输入的白话词进行分词，拆分成词元；

[0158] 查询官方词语子模块，用于在所述百姓体词库中查询所述词元对应的官方词语；

[0159] 生成官方词语子模块，用于将词元对应的官方词语合并成与所述用户输入的白话词相匹配的官方词语。

[0160] 本申请内嵌的百姓体词库主要由官方词语和白话词构成，通过将日常工作生活中常用的白话词与政府办事服务事项中的官方词语建立对应关系，从而解决用户对政府网站业务的“理解”障碍和搜索结果不准确的问题，通过对白话词进行分词、拆分成词元，根据所述词元在百姓体词库中查询与所述词语对应的官方词语，将词元对应的官方词语进行合并生成与白话词相匹配的官方语言，从而实现了用户搜索快速、搜索结果准确。

[0161] 对于上述一种网站的搜索装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见图 2 所示方法实施例的部分说明即可。

[0162] 本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0163] 本领域技术人员易于想到的是：上述各个实施例的任意组合应用都是可行的，故上述各个实施例之间的任意组合都是本申请的实施方案，但是由于篇幅限制，本说明书在此就不一一详述了。

[0164] 以上对本申请所提供的一种网站的搜索方法和装置，进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

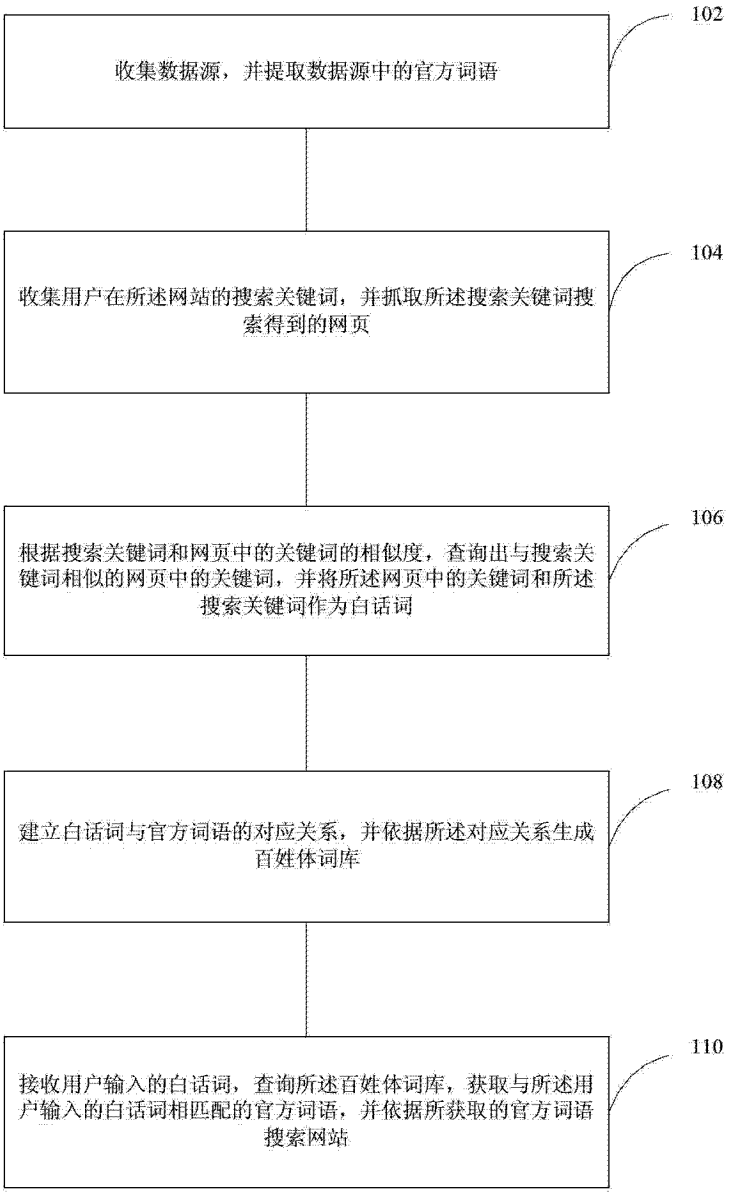

102
收集数据源，并提取数据源中的官方词语
104
收集用户在所述网站的搜索关键词，并抓取所述搜索关键词搜索得到的网页
106
根据搜索关键词和网页中的关键词的相似度，查询出与搜索关键词相似的网页中的关键词，并将所述网页中的关键词和所述搜索关键词作为白话词
108
建立白话词与官方词语的对应关系，并依据所述对应关系生成百姓体词库
110
接收用户输入的白话词，查询所述百姓体词库，获取与所述用户输入的白话词相匹配的官方词语，并依据所获取的官方词语搜索网站

图 1

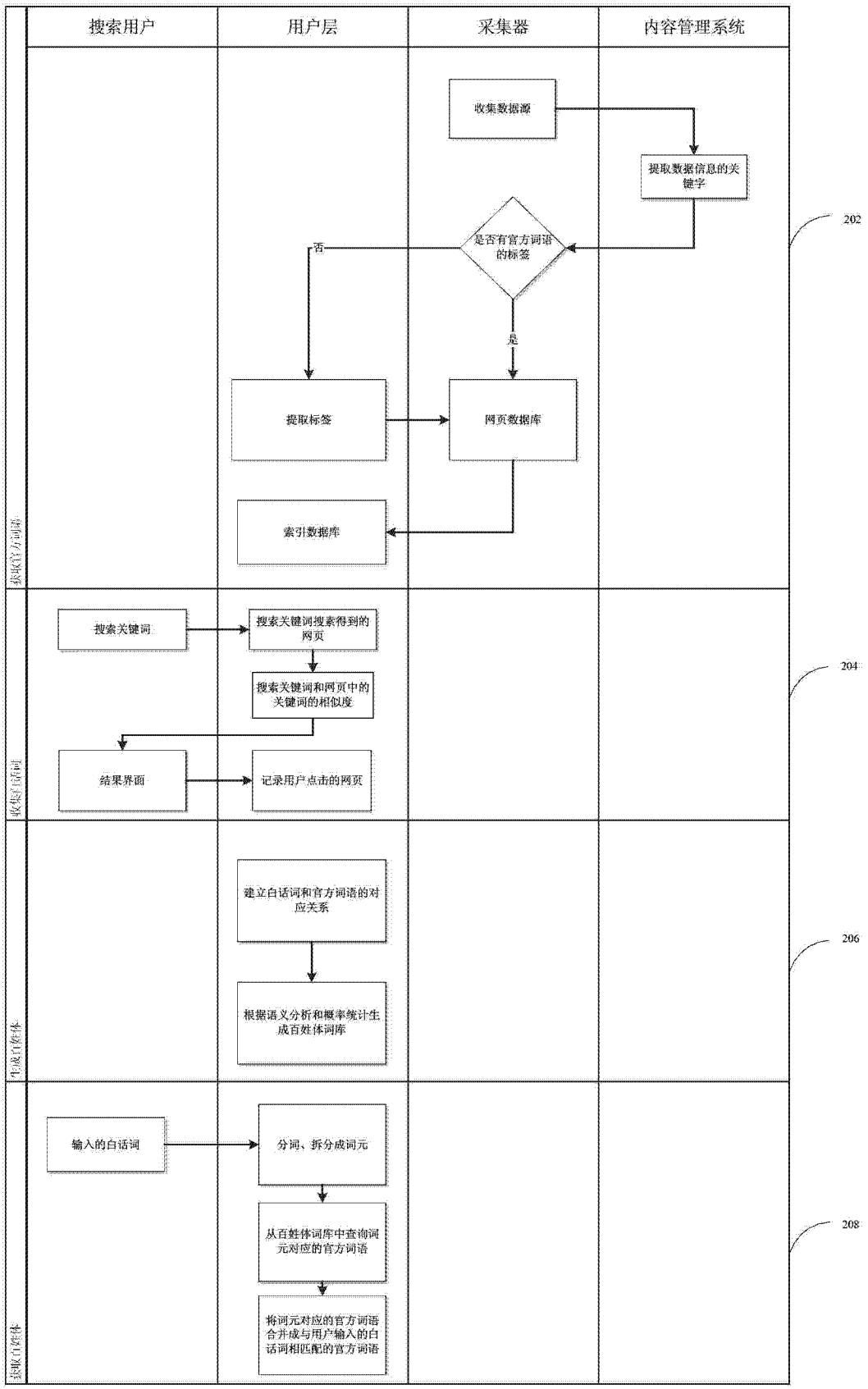

搜索用户
用户层
采集器
内容管理系统
获取官方词语
收集白话词
生成白姓体
获取白姓体
收集数据源
提取数据信息的关键字
是否有官方词语的标签
否
是
提取标签
网页数据库
索引数据库
搜索关键词
搜索关键词搜索得到的网页
搜索关键词和网页中的关键词的相似度
结果界面
记录用户点击的网页
建立白话词和官方词语的对应关系
根据语义分析和概率统计生成百姓体词库
输入的白话词
分词、拆分成词元
从百姓体词库中查询词元对应的官方词语
将词元对应的官方词语合并成与用户输入的白话词相匹配的官方词语
202
204
206
208

图 2

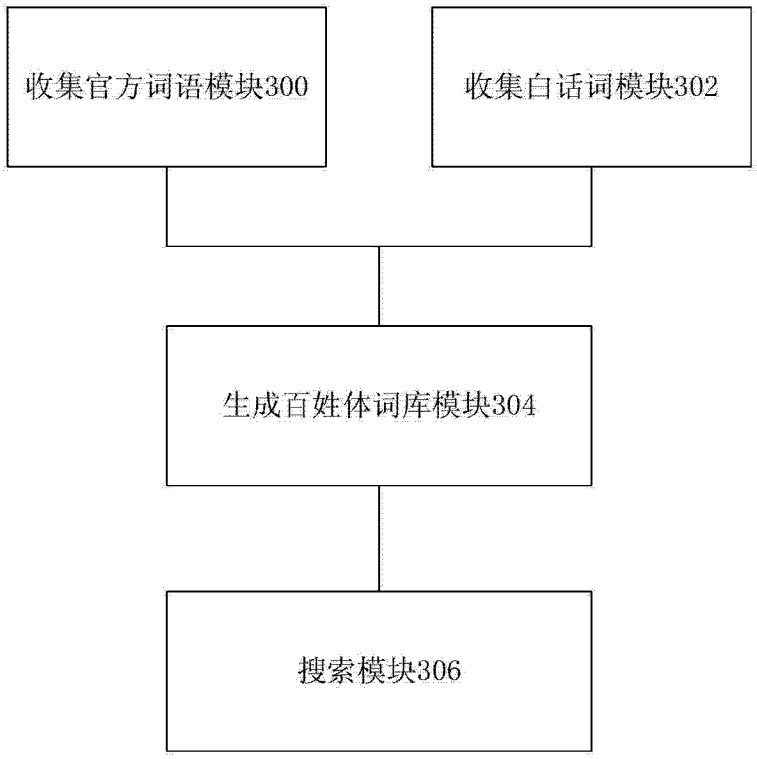


图 3