



(12) 发明专利申请

(10) 申请公布号 CN 105683967 A

(43) 申请公布日 2016. 06. 15

(21) 申请号 201680000295. X

(22) 申请日 2016. 01. 30

(85) PCT国际申请进入国家阶段日
2016. 04. 21

(86) PCT国际申请的申请数据
PCT/CN2016/072923 2016. 01. 30

(71) 申请人 深圳市博信诺达经贸咨询有限公司
地址 518000 广东省深圳市南山区南新路向
南海德大厦 A 座 1001

(72) 发明人 马岩

(74) 专利代理机构 深圳市科冠知识产权代理有
限公司 44355
代理人 孔丽霞

(51) Int. Cl.
G06F 17/30(2006. 01)

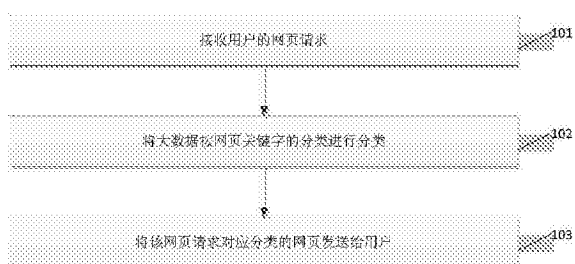
权利要求书1页 说明书3页 附图1页

(54) 发明名称

基于大数据的网页抓取方法及系统

(57) 摘要

本发明提供了一种基于大数据的网页抓取方法及系统,所述方法包括如下步骤:接收用户的网页请求;将大数据按网页关键字的分类进行分类;将该网页请求对应分类的网页发送给用户。本发明提供的技术方案具有网页抓取方便的优点。



1. 一种基于大数据的网页抓取方法,其特征在于,所述方法包括如下步骤:接收用户的网页请求;

将大数据按网页关键字的分类进行分类;

将该网页请求对应分类的网页发送给用户。

2. 根据权利要求1所述的方法,其特征在于,所述方法还包括:

如用户退回该分类的网页,将向用户发送相似分类的网页。

3. 根据权利要求1所述的方法,其特征在于,所述方法还包括:

如用户接收该分类的网页,则将该网页的分类保存。

4. 一种基于大数据的网页抓取系统,其特征在于,所述系统包括:

接收单元,用于接收用户的网页请求;

分类单元,用于将大数据按网页关键字的分类进行分类;

发送单元,用于将该网页请求对应分类的网页发送给用户。

5. 根据权利要求4所述的系统,其特征在于,所述系统还包括:

退回单元,用于如用户退回该分类的网页,将向用户发送相似分类的网页。

6. 根据权利要求4所述的系统,其特征在于,所述系统还包括:

保存单元,用于如用户接收该分类的网页,则将该网页的分类保存。

基于大数据的网页抓取方法及系统

技术领域

[0001] 本发明涉及通信及物联网领域,尤其涉及一种基于大数据的网页抓取方法及系统。

背景技术

[0002] 大数据与云计算的未来发展趋势和前景已经极其广阔,未来的互联网就是大数据和云计算的天下,大数据和云计算将成为每一个IT人员必须会的技术了。大数据中,网页数据为大数据的主要来源,而如何对网页数据进行抓取成为大数据处理中的重要技术。

发明内容

[0003] 提供一种基于大数据的网页抓取方法,其解决了现有技术的对网页数据抓取不方便的缺点。

[0004] 一方面,提供一种基于大数据的网页抓取方法,所述方法包括如下步骤:

[0005] 接收用户的网页请求;

[0006] 将大数据按网页关键字的分类进行分类;

[0007] 将该网页请求对应分类的网页发送给用户。

[0008] 可选的,所述方法还包括:

[0009] 如用户退回该分类的网页,将向用户发送相似分类的网页。

[0010] 可选的,所述方法还包括:

[0011] 如用户接收该分类的网页,则将该网页的分类保存。

[0012] 另一方面,提供一种基于大数据的网页抓取系统,所述系统包括:

[0013] 接收单元,用于接收用户的网页请求;

[0014] 分类单元,用于将大数据按网页关键字的分类进行分类;

[0015] 发送单元,用于将该网页请求对应分类的网页发送给用户。

[0016] 可选的,所述系统还包括:

[0017] 退回单元,用于如用户退回该分类的网页,将向用户发送相似分类的网页。

[0018] 可选的,所述系统还包括:

[0019] 保存单元,用于如用户接收该分类的网页,则将该网页的分类保存。

[0020] 本发明具体实施方式提供的技术方案接收用户的网页请求,将大数据按网页关键字的分类进行分类,将该网页请求对应分类的网页发送给用户,所以其具有网页数据抓取方便的优点。

附图说明

[0021] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以

根据这些附图获得其他的附图。

[0022] 图1为本发明提供的一种基于大数据的网页抓取方法的流程图；

[0023] 图2为本发明提供的一种基于大数据的网页抓取系统的结构图。

具体实施方式

[0024] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0025] 参阅图1，图1为本发明第一较佳实施方式提供的一种基于大数据的网页抓取方法的流程图，该方法由智能终端来完成，该方法如图1所示，包括如下步骤：

[0026] 步骤S101、接收用户的网页请求；

[0027] 步骤S102、将大数据按网页关键字的分类进行分类；

[0028] 步骤S103、将该网页请求对应分类的网页发送给用户。

[0029] 本发明具体实施方式提供的技术方案接收用户的网页请求，将大数据按网页关键字的分类进行分类，将该网页请求对应分类的网页发送给用户，所以其具有网页数据抓取方便的优点。

[0030] 可选的，上述方法在步骤S103之后还可以包括：

[0031] 如用户退回该分类的网页，将向用户发送相似分类的网页。

[0032] 可选的，上述方法在步骤S103之后还可以包括：

[0033] 如用户接收该分类的网页，则将该网页的分类保存。

[0034] 参阅图2，图2为本发明第二较佳实施方式提供的一种基于大数据的网页抓取系统，该系统包括：

[0035] 接收单元201，用于接收用户的网页请求；

[0036] 分类单元202，用于将大数据按网页关键字的分类进行分类；

[0037] 发送单元203，用于将该网页请求对应分类的网页发送给用户。

[0038] 本发明具体实施方式提供的技术方案接收用户的网页请求，将大数据按网页关键字的分类进行分类，将该网页请求对应分类的网页发送给用户，所以其具有网页数据抓取方便的优点。

[0039] 可选的，上述系统还可以包括：

[0040] 退回单元204，用于如用户退回该分类的网页，将向用户发送相似分类的网页。

[0041] 可选的，上述系统还可以包括：

[0042] 保存单元205，用于如用户接收该分类的网页，则将该网页的分类保存。

[0043] 需要说明的是，对于前述的各方法实施方式或实施例，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，本发明并不受所描述的动作顺序的限制，因为根据本发明，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，说明书中所描述实施方式或实施例均属于优选实施例，所涉及的动作和单元并不一定是本发明所必须的。

[0044] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中未详述的部

分,可以参见其他实施例的相关描述。

[0045] 本发明实施例方法中的步骤可以根据实际需要进行顺序调整、合并和删减。

[0046] 本发明实施例装置中的单元可以根据实际需要进行合并、划分和删减。本领域的技术人员可以将本说明书中描述的不同实施例以及不同实施例的特征进行结合或组合。

[0047] 通过以上的实施方式的描述,所属领域的技术人员可以清楚地了解到本发明可以用硬件实现,或固件实现,或它们的组合方式来实现。当使用软件实现时,可以将上述功能存储在计算机可读介质中或作为计算机可读介质上的一个或多个指令或代码进行传输。计算机可读介质包括计算机存储介质和通信介质,其中通信介质包括便于从一个地方向另一个地方传送计算机程序的任何介质。存储介质可以是计算机能够存取的任何可用介质。以此为例但不限于:计算机可读介质可以包括随机存取存储器(Random Access Memory, RAM)、只读存储器(Read-Only Memory, ROM)、电可擦可编程只读存储器(Electrically Erasable Programmable Read-Only Memory, EEPROM)、只读光盘(Compact Disc Read-Only Memory, CD-ROM)或其他光盘存储、磁盘存储介质或者其他磁存储设备、或者能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质。此外,任何连接可以适当的成为计算机可读介质。例如,如果软件是使用同轴电缆、光纤光缆、双绞线、数字用户线(Digital Subscriber Line, DSL)或者诸如红外线、无线电和微波之类的无线技术从网站、服务器或者其他远程源传输的,那么同轴电缆、光纤光缆、双绞线、DSL或者诸如红外线、无线和微波之类的无线技术包括在所属介质的定义中。如本发明所使用的,盘(Disk)和碟(disc)包括压缩光碟(CD)、激光碟、光碟、数字通用光碟(DVD)、软盘和蓝光光碟,其中盘通常磁性的复制数据,而碟则用激光来光学的复制数据。上面的组合也应当包括在计算机可读介质的保护范围之内。

[0048] 总之,以上所述仅为本发明技术方案的较佳实施例而已,并非用于限定本发明的保护范围。凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

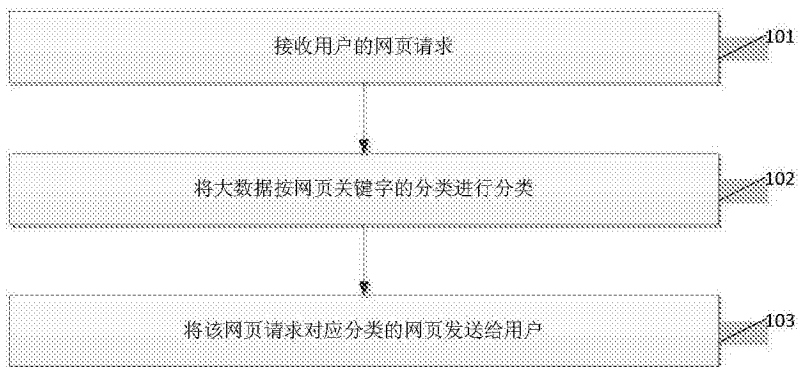


图1

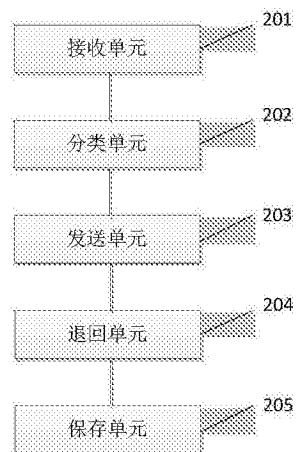


图2