(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 107329969 A
(43)申请公布日 2017.11.07

(21)申请号 201710367303.0

(22)申请日 2017.05.23

(71)申请人 合肥智权信息科技有限公司
地址 230000 安徽省合肥市高新区国家大学科技园创业孵化中心C区第一层大学生梦工厂9、10号工位

(72)发明人 周钰徐

(74)专利代理机构 合肥市长远专利代理事务所（普通合伙）34119
代理人 段晓微 叶美琴

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

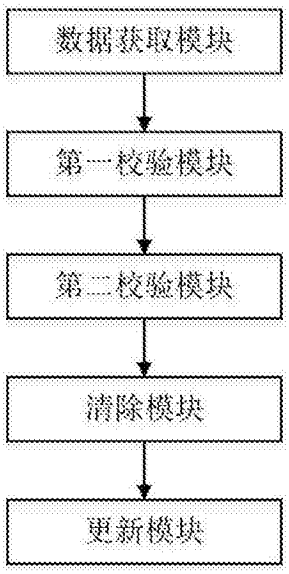
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种基于多次校验的数据信息更新系统和方法

(57)摘要

本发明公开了一种基于多次校验的数据信息更新系统和方法，其特征在于，包括：数据获取模块，用于利用网络爬虫获取数据信息；第一校验模块，用于在预设时间对网络爬虫获取的数据信息进行初步校验，得到初步数据信息集合；第二校验模块，用于将初步数据信息集合中数据信息与预设数据信息库进行重复度校验；清除模块，用于删除初步数据信息集合中重复度校验失败的数据信息；更新模块，用于将初步数据信息集合中重复度校验成功的数据信息添加到预设数据信息库中，更新预设数据信息库。如此，对网络爬虫获取大量数据信息后进行相似度校验，去除实时获取的相似度高的数据信息，避免网络爬虫获取重复数据信息，减少数据信息数量，提高重复度校验效率。



CN 107329969 A

1.一种基于多次校验的数据信息更新系统，其特征在于，包括：

数据获取模块，用于利用网络爬虫获取数据信息；

第一校验模块，用于在预设时间对网络爬虫获取的数据信息进行初步校验，得到初步数据信息集合；

第二校验模块，用于将初步数据信息集合中数据信息与预设数据信息库进行比对并判断所述数据信息是否校验成功；

清除模块，用于删除初步数据信息集合中重复度校验失败的数据信息；

更新模块，用于将初步数据信息集合中重复度校验成功的数据信息添加到预设数据信息库中，更新预设数据信息库。

2.根据权利要求1所述的基于多次校验的数据信息更新系统，其特征在于，所述第一校验模块，具体用于：

在预设时间将网络爬虫获取的数据信息进行相互比较；

删除两个数据信息中相似度大于预设相似度值的一个数据信息；

在所有数据信息均相互比较后，得到初步数据信息集合。

3.根据权利要求1所述的基于多次校验的数据信息更新系统，其特征在于，所述第二校验模块，具体用于：

获取初步数据信息集合一个数据信息；

将所述数据信息与预设数据信息库中预设数据信息进行重复度比较，得到重复度值并根据重复度值与预设阈值判断重复度校验是否成功；当所述重复度值大于预设阈值，判断所述数据信息重复度校验失败；否则，判断所述数据信息重复度校验成功；

当初步数据信息集合中所有数据信息均进行重复度比较后，完成重复度校验。

4.根据权利要求1所述的基于多次校验的数据信息更新系统，其特征在于，所述数据获取模块包括多个数据获取子模块。

5.根据权利要求1所述的基于多次校验的数据信息更新系统，其特征在于，还包括数据信息推送模块，用于将添加到预设数据信息库中的数据信息推送给用户。

6.一种基于多次校验的数据信息更新方法，其特征在于，包括：

S1、利用网络爬虫获取数据信息；

S2、在预设时间对网络爬虫获取的数据信息进行初步校验，得到初步数据信息集合；

S3、将初步数据信息集合中数据信息与预设数据信息库进行比对并判断所述数据信息是否校验成功；

S4、删除初步数据信息集合中重复度校验失败的数据信息；

S5、将初步数据信息集合中重复度校验成功的数据信息添加到预设数据信息库中，更新预设数据信息库。

7.根据权利要求6所述的基于多次校验的数据信息更新方法，其特征在于，步骤S2，具体包括：

在预设时间将网络爬虫获取的数据信息进行相互比较；

删除两个数据信息中相似度大于预设相似度值的一个数据信息；

在所有数据信息均相互比较后，得到初步数据信息集合。

8.根据权利要求6所述的基于多次校验的数据信息更新方法，其特征在于，步骤S3，具

体包括：

S31、获取初步数据信息集合一个数据信息；

S32、将所述数据信息与预设数据信息库中预设数据信息进行重复度比较，得到重复度值并根据重复度值与预设阈值判断重复度校验是否成功；当所述重复度值大于预设阈值，判断所述数据信息重复度校验失败；否则，判断所述数据信息重复度校验成功；

S33、重复步骤S31、S32操作，直到初步数据信息集合中所有数据信息均进行重复度比较。

9.根据权利要求6所述的基于多次校验的数据信息更新方法，其特征在于，在步骤S1中，利用多个网络爬虫获取数据信息。

10.根据权利要求6所述的基于多次校验的数据信息更新方法，其特征在于，还包括步骤S6，将添加到预设数据信息库中的数据信息推送给用户。

一种基于多次校验的数据信息更新系统和方法

技术领域

[0001] 本发明涉及数据校验技术领域，尤其涉及一种基于多次校验的数据信息更新系统和方法。

背景技术

[0002] 随着互联网信息爆炸式增长，每一天互联网中的数据信息都呈现几何式的增加，用户在获取需要的数据信息时，往往会淹没于大量无用重复信息中，目前通过搜索引擎获取自己感兴趣的数据信息已经是大多数的用户推崇的便捷方式，作为搜索引擎的基础构件之一的网络爬虫，需要从互联网上获取数据信息，为用户提供数据信息的支持，但是随着互联网信息的肆意转载和多网站投放，网络爬虫获取的数据信息是否丰富、相似度和重合度是否高，均与网络爬虫的效率紧密相关。于是为了提高爬行速度，网络通常会采取并行爬行的工作方式，随之引入了新的问题：重复性，并行运行的爬虫或爬行线程同时运行时增加了重复页面、质量问题，并行运行时，每个爬虫或爬行线程只能获取部分页面，导致页面质量下降。

[0003] 目前已有的数据库查重，由于数据库数据庞大，随着抓取数据信息的增多，导致工作量较大，算法效率变低。

发明内容

[0004] 基于背景技术存在的技术问题，本发明提出了一种基于多次校验的数据信息更新系统和方法；

[0005] 本发明提出的一种基于多次校验的数据信息更新系统，包括：

[0006] 数据获取模块，用于利用网络爬虫获取数据信息；

[0007] 第一校验模块，用于在预设时间对网络爬虫获取的数据信息进行初步校验，得到初步数据信息集合；

[0008] 第二校验模块，用于将初步数据信息集合中数据信息与预设数据信息库进行比对并判断所述数据信息是否校验成功；

[0009] 清除模块，用于删除初步数据信息集合中重复度校验失败的数据信息；

[0010] 更新模块，用于将初步数据信息集合中重复度校验成功的数据信息添加到预设数据信息库中，更新预设数据信息库。

[0011] 优选地，所述第一校验模块，具体用于：

[0012] 在预设时间将网络爬虫获取的数据信息进行相互比较；

[0013] 删除两个数据信息中相似度大于预设相似度值的一个数据信息；

[0014] 在所有数据信息均相互比较后，得到初步数据信息集合。

[0015] 优选地，所述第二校验模块，具体用于：

[0016] 获取初步数据信息集合一个数据信息；

[0017] 将所述数据信息与预设数据信息库中预设数据信息进行重复度比较，得到重复度

值并根据重复度值与预设阈值判断重复度校验是否成功；当所述重复度值大于预设阈值，判断所述数据信息重复度校验失败；否则，判断所述数据信息重复度校验成功；当初步数据信息集合中所有数据信息均进行重复度比较后，完成重复度校验。

[0018] 优选地，所述数据获取模块包括多个数据获取子模块。

[0019] 优选地，还包括数据信息推送模块，用于将添加到预设数据信息库中的数据信息推送给用户。

[0020] 一种基于多次校验的数据信息更新方法，包括：

[0021] S1、利用网络爬虫获取数据信息；

[0022] S2、在预设时间对网络爬虫获取的数据信息进行初步校验，得到初步数据信息集合；

[0023] S3、将初步数据信息集合中数据信息与预设数据信息库进行比对并判断所述数据信息是否校验成功；

[0024] S4、删除初步数据信息集合中重复度校验失败的数据信息；

[0025] S5、将初步数据信息集合中重复度校验成功的数据信息添加到预设数据信息库中，更新预设数据信息库。

[0026] 优选地，步骤S2，具体包括：

[0027] 在预设时间将网络爬虫获取的数据信息进行相互比较；

[0028] 删除两个数据信息中相似度大于预设相似度值的一个数据信息；

[0029] 在所有数据信息均相互比较后，得到初步数据信息集合。

[0030] 优选地，步骤S3，具体包括：

[0031] S31、获取初步数据信息集合一个数据信息；

[0032] S32、将所述数据信息与预设数据信息库中预设数据信息进行重复度比较，得到重复度值并根据重复度值与预设阈值判断重复度校验是否成功；当所述重复度值大于预设阈值，判断所述数据信息重复度校验失败；否则，判断所述数据信息重复度校验成功；

[0033] S33、重复步骤S31、S32操作，直到初步数据信息集合中所有数据信息均进行重复度比较。

[0034] 优选地，在步骤S1中，利用多个网络爬虫获取数据信息。

[0035] 优选地，还包括步骤S6，将添加到预设数据信息库中的数据信息推送给用户。

[0036] 本发明通过对多个网络爬虫获取的数据信息进行初步校验，过滤网络爬虫获取的数据信息中任意两个数据信息中相似度大于预设相似度值的一个数据信息，得到初步数据信息集合，再将初步数据信息集合中数据信息预设数据信息库中预设数据信息进行重复度比较，根据重复度值与预设阈值判断重复度校验是否成功，当初步数据信息集合中数据信息重复度校验失败，删除所述数据信息；当初步数据信息集合中数据信息重复度校验成功，将初步数据信息集合中数据信息添加到预设数据信息库中；如此，首先对多个网络爬虫获取大量数据信息后进行相似度校验，去除实时获取的相似度高的数据信息，避免网络爬虫获取重复数据信息，减少数据信息数量，因而提高重复度校验效率，在重复度校验中，删除重复度较高的数据信息，降低数据信息重合度；将重复度低的数据信息添加到预设数据信息库中，提高重复度校验的实时性。

附图说明

[0037] 图1为本发明提出的一种基于多次校验的数据信息更新系统的模块示意图；

[0038] 图2为本发明提出的一种基于多次校验的数据信息更新方法的流程示意图。

具体实施方式

[0039] 如图1所示，图1为本发明提出的一种基于多次校验的数据信息更新系统的模块示意图；

[0040] 参照图1，本发明提出的一种基于多次校验的数据信息更新系统，包括：

[0041] 数据获取模块，用于利用网络爬虫获取数据信息。

[0042] 在具体方案中，数据获取模块可以设置包括多个数据获取子模块，每个数据子模块可以利用多个网络爬虫用以获取数据。根据情报搜集与分析目标，利用网络爬虫，采集各类信息。

[0043] 第一校验模块，用于在预设时间对网络爬虫获取的数据信息进行初步校验，得到初步数据信息集合；

[0044] 在具体方案中，第一校验模块用于：在预设时间将网络爬虫获取的数据信息进行相互比较；删除两个数据信息中相似度大于预设相似度值的一个数据信息；在所有数据信息均相互比较后，得到初步数据信息集合。通过对多个网络爬虫获取的数据信息进行初步校验，过滤网络爬虫获取的数据信息中任意两个数据信息中相似度大于预设相似度值的一个数据信息，得到初步数据信息集合，避免网络爬虫获取重复数据信息，减少重复数据信息数量。

[0045] 第二校验模块，用于将初步数据信息集合中数据信息与预设数据信息库进行比对并判断所述数据信息是否校验成功；

[0046] 在具体方案中，第二校验模块用于：获取初步数据信息集合一个数据信息；将所述数据信息与预设数据信息库中预设数据信息进行重复度比较，得到重复度值并根据重复度值与预设阈值判断重复度校验是否成功；当所述重复度值大于预设阈值，判断所述数据信息重复度校验失败；否则，判断所述数据信息重复度校验成功；当初步数据信息集合中所有数据信息均进行重复度比较后，完成重复度校验。

[0047] 清除模块，用于删除初步数据信息集合中重复度校验失败的数据信息；

[0048] 更新模块，用于将初步数据信息集合中重复度校验成功的数据信息添加到预设数据信息库中，更新预设数据信息库；

[0049] 在具体方案中，将初步数据信息集合中数据信息预设数据信息库中预设数据信息进行重复度比较，根据重复度值与预设阈值判断重复度校验是否成功，当初步数据信息集合中数据信息重复度校验失败，删除所述数据信息，降低数据信息重合度；当初步数据信息集合中数据信息重复度校验成功，将初步数据信息集合中数据信息添加到预设数据信息库中，提高重复度校验的实时性。

[0050] 数据信息推送模块，用于将添加到预设数据信息库中的数据信息推送给用户；

[0051] 在具体方案中，在两次校验后，将更新到预设数据信息库中的数据信息推送给用户。

[0052] 如图2所示，图2为本发明提出的一种基于多次校验的数据信息更新方法的流程示意图；

[0053] 参照图2，本发明提出的一种基于多次校验的数据信息更新方法，其特征在于，包括：

[0054] S1、利用网络爬虫获取数据信息；

[0055] 在具体方案中，可以设置多个网络爬虫用以获取数据，根据情报搜集与分析目标，利用网络爬虫，采集各类信息。

[0056] S2、在预设时间对网络爬虫获取的数据信息进行初步校验，得到初步数据信息集合；具体包括：在预设时间将网络爬虫获取的数据信息进行相互比较；删除两个数据信息中相似度大于预设相似度值的一个数据信息；在所有数据信息均相互比较后，得到初步数据信息集合。

[0057] 在具体方案中，通过对多个网络爬虫获取的数据信息进行初步校验，过滤网络爬虫获取的数据信息中任意两个数据信息中相似度大于预设相似度值的一个数据信息，得到初步数据信息集合，避免网络爬虫获取重复数据信息，减少重复数据信息数量。

[0058] S3、将初步数据信息集合中数据信息与预设数据信息库进行比对并判断所述数据信息是否校验成功；具体包括：S31、获取初步数据信息集合一个数据信息；S32、将所述数据信息与预设数据信息库中预设数据信息进行重复度比较，得到重复度值并根据重复度值与预设阈值判断重复度校验是否成功；当所述重复度值大于预设阈值，判断所述数据信息重复度校验失败；否则，判断所述数据信息重复度校验成功；S33、重复步骤S31、S32操作，直到初步数据信息集合中所有数据信息均进行重复度比较；

[0059] S4、删除初步数据信息集合中重复度校验失败的数据信息；

[0060] S5、将初步数据信息集合中重复度校验成功的数据信息添加到预设数据信息库中，更新预设数据信息库。

[0061] 在具体方案中，将初步数据信息集合中数据信息预设数据信息库中预设数据信息进行重复度比较，根据重复度值与预设阈值判断重复度校验是否成功，当初步数据信息集合中数据信息重复度校验失败，删除所述数据信息，降低数据信息重合度；当初步数据信息集合中数据信息重复度校验成功，将初步数据信息集合中数据信息添加到预设数据信息库中，提高重复度校验的实时性。

[0062] 还包括步骤S6，将添加到预设数据信息库中的数据信息推送给用户。

[0063] 在具体方案中，在两次校验后，将更新到预设数据信息库中的数据信息推送给用户。

[0064] 本实施方式通过对多个网络爬虫获取的数据信息进行初步校验，过滤网络爬虫获取的数据信息中任意两个数据信息中相似度大于预设相似度值的一个数据信息，得到初步数据信息集合，再将初步数据信息集合中数据信息预设数据信息库中预设数据信息进行重复度比较，根据重复度值与预设阈值判断重复度校验是否成功，当初步数据信息集合中数据信息重复度校验失败，删除所述数据信息；当初步数据信息集合中数据信息重复度校验成功，将初步数据信息集合中数据信息添加到预设数据信息库中；如此，首先对多个网络爬虫获取大量数据信息后进行相似度校验，去除实时获取的相似度高的数据信息，避免网络爬虫获取重复数据信息，减少数据信息数量，因而提高重复度校验效率，在重复度校验中，

删除重复度较高的数据信息，降低数据信息重合度；将重复度低的数据信息添加到预设数据信息库中，提高重复度校验的实时性。

[0065] 以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，根据本发明的技术方案及其发明构思加以等同替换或改变，都应涵盖在本发明的保护范围之内。

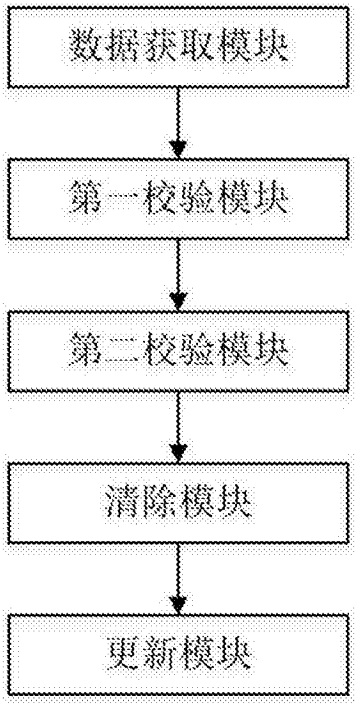


图1

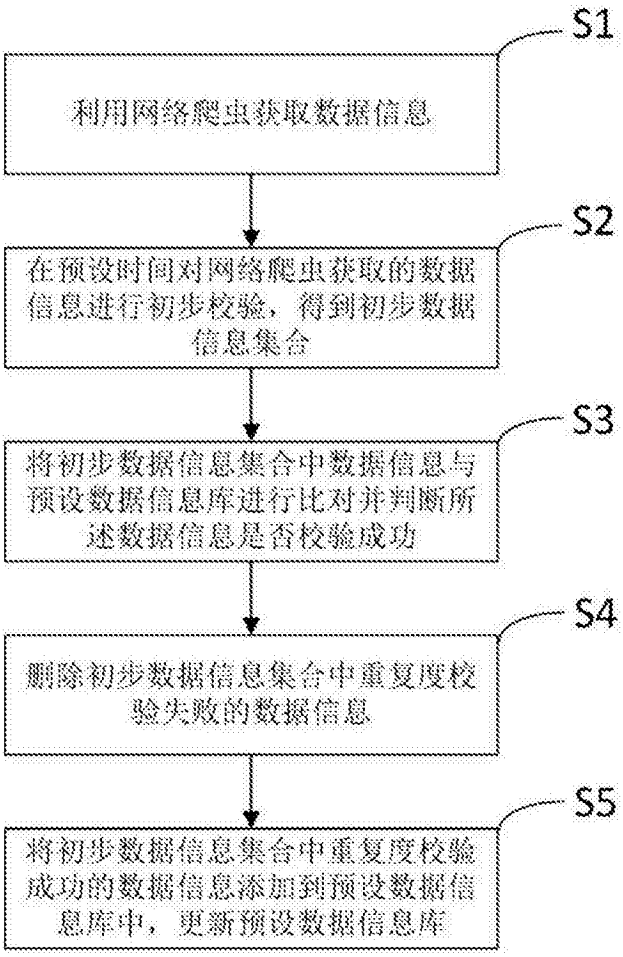

S1
利用网络爬虫获取数据信息
S2
在预设时间对网络爬虫获取的数据信息进行初步校验，得到初步数据信息集合
S3
将初步数据信息集合中数据信息与预设数据信息库进行比对并判断所述数据信息是否校验成功
S4
删除初步数据信息集合中重复度校验失败的数据信息
S5
将初步数据信息集合中重复度校验成功的数据信息添加到预设数据信息库中，更新预设数据信息库

图2