农业新闻数据源增量爬虫的应用探析

杨广召　曹 叶　朱航飞　王家硕　朱家玮

(塔里木大学信息工程学院,新疆阿拉尔 843300)

摘要　随着农业新闻数据日益膨胀,以农业为主题的增量爬虫成为爬取农业信息的重要手段。增量爬虫可以依据农业新闻数据的更新爬取数据相关更新的内容,剔除已经爬取的重复内容。本文结合农业新闻数据信息的特点,提出了一种适用于农业新闻信息的基于 Redis 的布隆过滤器的增量去重方法,摆脱超大的持久化文件撑爆内存的问题。试验证明随着抓取相关农业信息的增加,该方法在保证内存不被撑爆的同时能有效提高增量爬取农业信息的效率,在增量信息爬取过程中具有很好的应用价值。

关键词　农业新闻;增量爬虫;去重

中图分类号　TP391　　**文献标识码**　A

文章编号　1007-5739(2021)02-0259-02

DOI:10.3969/j.issn.1007-5739.2021.02.103　　**开放科学(资源服务)标识码(OSID)**:

Analysis on Application of Incremental Crawler of Agricultural News Data Source

YANG Guangzhao　CAO Ye　ZHU Hangfei　WANG Jiashuo　ZHU Jiawei

(School of Information Engineering, Tarim University, Alaer Xinjiang 843300)

Abstract　With the increasing expansion of agricultural news data, incremental crawlers with the theme of agriculture have become an important means of crawling agricultural information. Incremental crawlers can crawl the updated content based on the update of the agricultural news data, and remove the duplicate content that has been crawled. Combined with the characteristics of agricultural news data, the paper proposed an incremental deduplication method based on Redis-based Bloom filter suitable for agricultural news information, to get rid of the problem of memory over flowing caused by large persistent files. Experiments proved that with the increase of related agricultural information crawled, this method could effectively improve the efficiency of incremental crawling of agricultural information while ensuring that the memory is not burst. It has good application value in the process of incremental information crawling.

Keywords　agricultural news; incremental crawler; deduplication

网络爬虫的主要抓取对象是网页,获取的数据都是用户肉眼所见的数据,所以网络爬虫的核心思想是模拟人类浏览操作,只有在模拟人类操作获取到网页内容后,才能开始解析网页、提取数据的工作。

网络爬虫具体工作流程如下:网络爬虫的爬取通常从一个起始网络地址(URL)开始抓取,通过模拟人类用户浏览行为,获取起始 URL 对应的网页 A。从网页 A 中通过网页代码解析,提取出相关的 URL 和数据,提取的 URL 会被爬虫加入待抓取 URL 列表中,而有用的数据将被保存到本地(通常会保存到数据文件或者数据库中,方便使用)。网络爬虫会从待抓取 URL 列表中依次抓取每个 URL,通过模拟人类用户浏览操作获取到相应的网页 B。从网页 B 中爬取URL 和相关数据,URL 加入待抓取列表中等候网络爬虫访问,有用的数据保存到本地。

本文提出一种基于 Redis 的布隆过滤器增量爬虫系统,该系统是根据农业新闻数据的更新来及时抓取更新的内容,该系统的核心是在爬取内容前进行重复内容过滤,可以有效避免数据库出现重复冗余数据,有效提高增量爬虫的爬取效率[1-2]。

1　去重处理

去重处理可以有效规避将重复性的农业数据保存到数据库中造成大量的无效数据存放到数据库[3]。不要在爬虫爬取数据后进行内容过滤,这样做只不过是避免后端数据库出现重复数据。去重处理对于一次性爬取是有效的,但对于增量式爬虫则恰恰相反。对于持续性的增量式爬虫,应该进行“前置过滤”,这样可以有效地减少爬虫爬取的次数[4]。

在发出请求之前检查爬虫是否爬取过该 URL,如果已爬取过,则让爬虫直接跳过该请求以避免重复出动。除了重复的 URL 指纹,还应该加上 404 与 500 错

基金项目　南疆红枣生产管理信息化系统示范与推广(19/11 17831)。

作者简介　杨广召(1992—),男,河北衡水人,在读硕士研究生。研究方向:农业信息化。

收稿日期　2020-08-02

误的 URL 过滤，因为即使目标网站上没有反爬虫机制，但绝大多数 Web 服务器程序都会有 404 与 500 错误的记录。过多的 404 与 500 很容易暴露爬虫的痕迹，因而加入异常 URL 的筛选是非常有必要的[4]。Scrapy 提供的 Request-Fingerprint 函数请求生成指纹，然后生成指纹写入 Redis 中。Redis 支持数据的持久化，可以将内存中的数据保存在磁盘中，重启时可以再次加载使用，Redis 能读的速度是 110 000 次/s，写的速度是 81 000 次/s，可以减轻数据库压力，查询内存比查询数据库效率高。

2　Bloomfilter

在网络爬虫中使用 Bloomfilter 可以实现高效去重。Bloomfilter 是一个很长的二进制向量和一系列随机映射 Hash 函数。通常辨别某个元素是否在集合中的常用方法是用已知元素和集合中的元素进行对比。Bloomfilter 能够在较短时间内检查某一元素是否在集合内。

通常会把一些数据放在 Redis 内缓存，例如商品信息。因此，有查找请求可以依据商品 ID 直接去缓存中读取数据，不需要通过读取数据库，这样可以极大地提升性能。查询的相关请求流程：首先查找缓存，如果缓存内存在直接返回，如果在缓存内找不到相关数据再去数据库查找，最后把将数据库内获取的相关数据存放到缓存内。一旦有大量数据请求涌进来，并且在请求某个不存在的商品 ID，如果商品 ID 不存在，那么缓存内肯定没有相关数据，所以相关的数据请求都涌现在数据库，数据库处理的数据会很多，一旦处理数据的压力超过数据库的极限，极有可能将数据库压垮。因此，可以使用 Bloomfilter，Bloomfilter 可以解决缓存被穿透的相关问题。有大量的相关查询数据，并且相关数据的大小范围极大地超过了服务器的内存大小，如果再给一个相关数据，如何判断此数据是否存在其中。如果服务器的内存空间足够大，可以应用 HashMap 来解决相关问题，从原理上来说，其时间复杂度可以达到 O(1)，但是已有的数据量大小已经极大超过了服务器的内存存储范围，因而 HashMap 不能使用了，所以可以使用 Bloomfilter 来解决相关问题，查询的时间复杂度是常数 O(k)。Bloomfilter 可以不用存储元素本身，Bloomfilter 可以用来表示数据全集，其他相关数据结构都不能表示。

3　农业新闻数据源增量爬虫描述

3.1　增量爬虫整体设计

与用户的交互由用户接口模块负责，链接生成模块负责提取 URL 地址，任务控制模块来整体控制爬取的开始和结束，处理器模块是对爬取页面的处理，数据模块负责处理过程中数据的存储和抓取结果数据的存储。与其他网络爬虫不同之处在于，链接生成模块需要对 URL 进行分类和过滤，将 URL 分为不同的类别，同时过滤其他站点的 URL，只保留指定站点的 URL。在处理器模块，通过对页面变化情况的比较分析，来实现对产品详细信息页面中相关农业名称信息、简介、图片等信息的提取，这一过程称为信息抽取。通过处理器模块中时间控制子模块来实现页面定时更新计划。要摆脱超大的持久化文件撑爆内存的问题，核心去重技术是将布隆过滤器持久化数据保存到数据库中。本文以 Redis 作为布隆过滤器的数据载体，Redis 支持数据的持久化，可以将内存中的数据保存在磁盘中，布隆过滤器可以实现高效去重，Redis 原生就有 BitSet 类型，非常容易操控，增量爬虫的流程如图 1 所示。

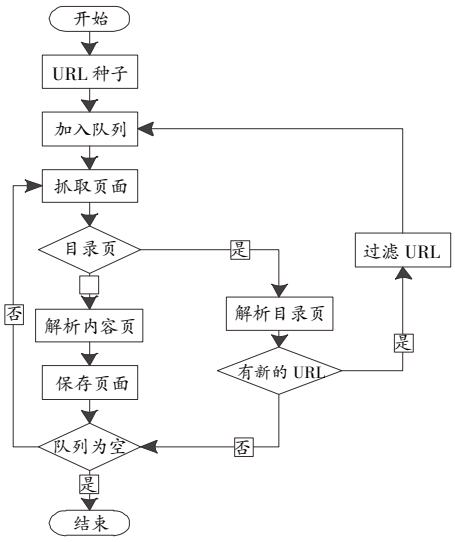

图 1　增量爬虫流程

3.2　主要算法描述

URL 分配算法：读取 URL 的更新历史记录，包括抓取次数、检测到的更新次数、上次抓取时间。

更新判断：①从缓存中读取 URL 及页面信息；②调用信息抽取模块，抽取页面包含的关键信息；③从数据库中读出 URL 对应的结构化信息；④将②中得到的数据与③中得到的数据进行对比；⑤调用更新处理模块。

更新处理：①接收更新判断模块的判断结果；②根据网页是否发生变化，更新数据库中网页对应的抓取次数、更新次数等值。如果网页发生变化，执行③；③将结构化后的网页关键信息存入数据库。

页面下载：①从 URL 队列中读取相应的 URL 信息；②将 URL 分配给不同的抓取进程；③下载页面；④

（下转第 264 页）

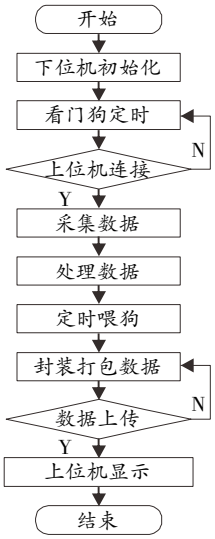

图 7　工作逻辑流程

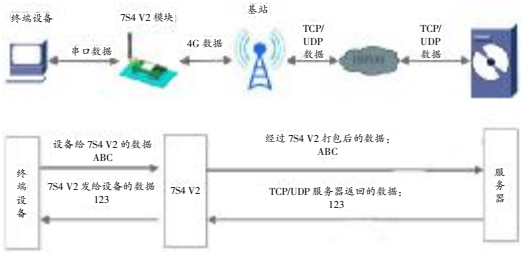

图 8　透传模式工作

所示。

在该网络透传模式下，还使用了心跳包监测工作，用户可以选择让模块发送心跳包以实现特定的需求。心跳包可以向网络端发送，也可以向串口设备端发送，心跳包的作用是为了保证连接可靠，同时通过心跳包了解模块的连接情况，用户在不确定设备是否连接的情况下，可以向服务器发送心跳包来确认各模块是否在线而不需要发送特定指令来确认，从而节省流量，反应更快[9-10]。

4　结论

农业冷链物流总体来说是一个高风险、高收益的行业，但是对其主要载体，也就是冷藏车内部的监测与探究并不多。针对现有冷链运输设备存在的问题，我们对其进行了进一步的改进。以往的冷链运输车监测系统只监测了温度与湿度，并没有对其内部气体监测进行过多探究。我们在腐败气体以及功能优化方面进行了深度探究，将腐败气体的监测融入冷链监测仪中，并在整合后优化了冷链监测仪的稳定性，为农业冷链运输车内部的监测提供了一种精准、便携、多功能的冷链物流监测仪，提高了农业冷链运输效率，可有效降低农业运输中的损失。

5　参考文献

[1] 陈志新，董瑞雪，卢成林，等.基于双模定位的冷链物流实时监测系统[J].保鲜与加工，2019，19(5)：178-184.
[2] 刘富奇.我国冷链物流发展的大数据模式探究[J].全国流通经济，2019(10)：29-30.
[3] 李锦晶.浅谈生鲜食品电商背景下冷链物流的发展趋势和要求[J].时代经贸，2019(9)：9-12.
[4] 雷佳雨，贾家鑫.以农产品为例探究冷链物流问题及对策[J].现代经济信息，2019(5)：387.
[5] 张凯，陈令芳，张恒，等.基于 STM32 的冷链物流监测系统的设计[J].现代电子技术，2018，41(4)：23-26.
[6] 刘影，王智霖，王宛，等.基于 RFID 冷链物流监测系统[J].物联网技术，2018，8(5)：97-99.
[7] 王倩.唐山市农产品物流问题研究[D].秦皇岛：河北科技师范学院，2018.
[8] 胡冠山.嵌入式城市冷链物流智能车载终端的研究与设计[J].科技创新导报，2018，15(9)：1-2.
[9] 张仕臻.超市电商化生鲜肉安全信息实时监测及配送方法研究[D].武汉：湖北工业大学，2015.
[10] 沈敏燕.果蔬类农产品冷链物流信息溯源研究[D].苏州：苏州科技大学，2017.

(上接第 260 页)

将页面信息 URL 信息存入缓存；⑤触发更新判断模块。

4　运行效果

通过对改进的增量爬虫与普通增量爬虫对农业新闻网站的爬取数据对比分析得出，随着新闻数据量的增加，普通增量爬虫处理速度有明显降低，过期页面也相应增加。将改进的增量爬虫运行一段时间发现，过期页面数量增加，改进的增量爬虫较普通增量爬虫爬取速度上有较大提高，该系统网络流量的增速明显降低。

5　结语

随着大数据和数据挖掘的发展，增量爬虫再一次受到人们的关注，基于 Redis 的 Bloomfilter 去重，既发挥了 Bloomfilter 的海量去重能力，又发挥了 Redis 的可持久化能力，基于 Redis 也方便分布式机器的去重。在使用过程中，要估算好待去重的数据量，适当地调整 seed 的数量和 blockNum 数量。

6　参考文献

[1] 杨颂.面向电子商务网站的增量爬虫设计与实现[D].长沙：湖南大学，2010.
[2] 刘芳云，张志勇，李玉祥.基于 Hadoop 的分布式并行增量爬虫技术研究[J].计算机测量与控制，2018，26(10)：269-275.
[3] 韩逸.基于增量式爬虫的搜索引擎系统的设计与实现[D].沈阳：东北大学，2015.
[4] 张皓，周学广.基于网页去噪 Hash 的增量式网络爬虫研究[J].舰船电子工程，2014，34(2)：86-90.