資料索引 project 1

爬蟲

報告人：劉昌澈 (405410075)

1. 實驗環境
   1. 處理器：2.3 GHz Dual-Core Intel Core i5
   2. 記憶體：8 GB 2133 MHz LPDDR3
   3. 作業系統：MacOS Catalina 10.15.3
   4. 程式語言：C++ (clang++ 11.0.3)
   5. IDE: Clion 2019.3
   6. 編譯器：Apple clang version 11.0.3
2. 使用套件
   1. X DevAPI: c++ 連接mysql api

<https://dev.mysql.com/doc/x-devapi-userguide/en/>

* 1. Gumbo-parser: maintex extraction (沒有完整的教學手冊：csdn有範例教學)

<https://github.com/google/gumbo-parser>

* 1. Md5: convert url string to md5

<https://github.com/JieweiWei/md5/tree/30774d0c9ecf82fb003fc15978b1f21bc85858f7>

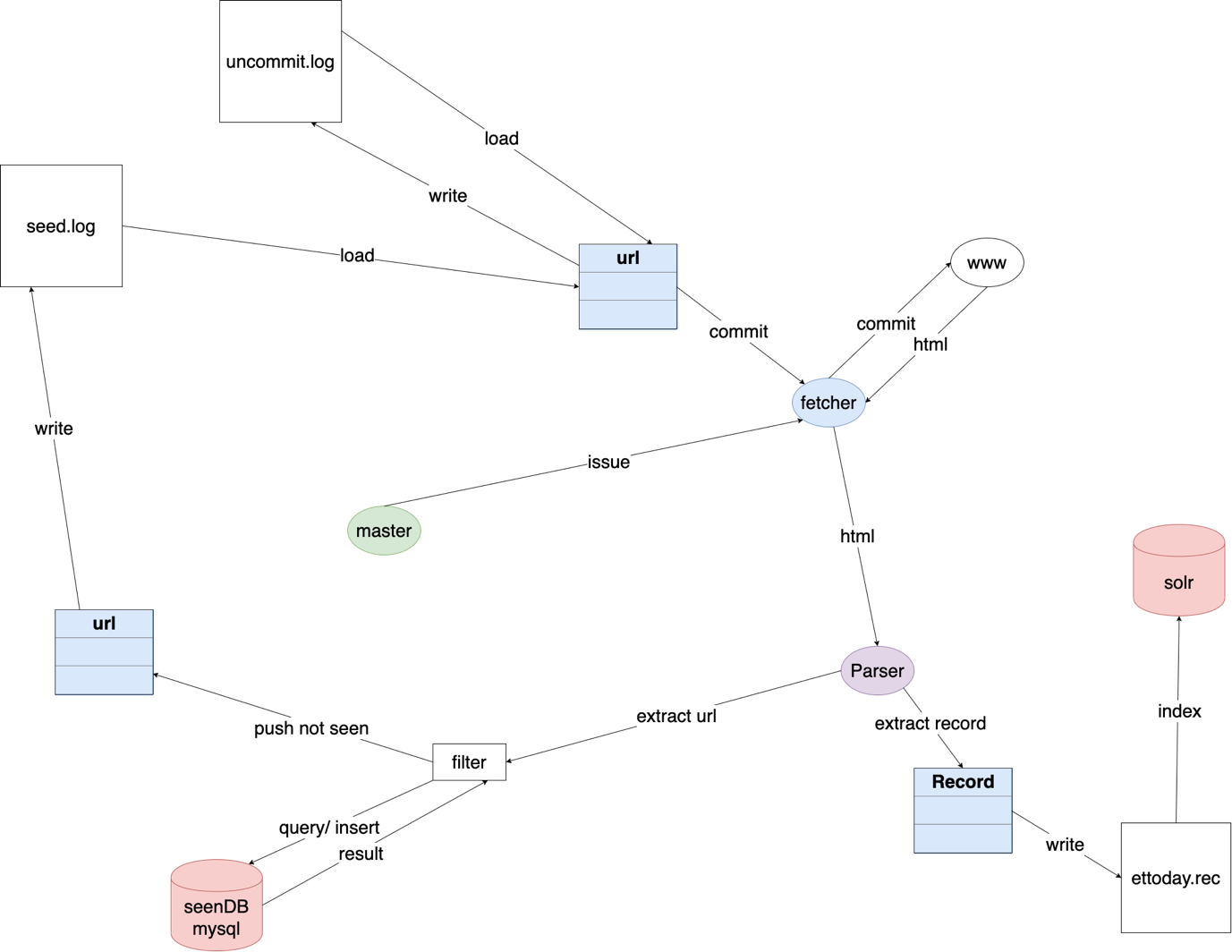
* 1. Libcurl: fetch html page

<https://curl.haxx.se/libcurl/>

* 1. Solr search engine

<https://lucene.apache.org/solr/>

1. 爬蟲架構



1. 爬蟲細節
   1. 爬蟲網站
      1. <https://www.ettoday.net/>
         1. 娛樂版 <https://www.star.ettoday.net/>
         2. 新聞雲 <https://www.ettoday.net/>
   2. 抓取名單

已regular expression來定義可接受的網址格式，目前因為maintex extraction還未完成（只能抓取固定格式的DOM tree），所以目前依底下兩種格式來抓取

* + 1. https:\/\/star.ettoday\.net[\/|\w\/.\?|=]+
    2. https:\/\/www.ettoday\.net[\/|\w\/.\?|=]+\.htm[?|=|\w]\*
  1. 不允許抓取名單

有些網站有非常好的反爬蟲裝置，如果放任爬蟲對該網站進行抓取，會一直存取到罐頭訊息，因此目前先把這些網站放入黑名單，但是黑名單也是用來防範抓取到spam，然而目前的爬蟲還未開發spam detection

* 1. User Agent

抓取網頁時如果沒有提供適當的User Agent Header，而發出request的話會照成網頁無法擷取的狀況，因此我使用我本機google瀏覽器的Header，之後會將此處修改成網頁要求的User Agent Header

* 1. 爬蟲元素
     1. Master
        1. Master 負責發送尚未抓取網頁的網址給Parser
     2. Parser
        1. Fetcher抓取網站的功能
        2. maintex extraction
        3. extract link
  2. 爬蟲流程

目前該版的爬蟲為單線程爬蟲，使用者在開始進行爬抓取前須先將預抓取的網域放入seed.log在開始爬蟲。爬蟲開始時，master會從seed.log當中將要抓取的網址放入CommitQueue中，並一個一個傳入由Parser，Parser會向該網站發出request，如果獲取到網頁後，對該網頁進行解析，並提煉出外部連結與正文，正文會先放入RecordQueue中等到RecordQueue達到最大5MB(使用者可自行定義)時，RecordQueue在寫入ettoday.rec，而連結則是會經過filter來判斷該連結是否已看過，如果沒看過就存入Queue中（uncommit.log和Filter間的Queue）等到5MB(使用者可自行定義)時，在寫入uncommit.log。這邊的filter是直接和mysql DB進行query詢問的，上述的部份會不斷重複直到達到使用者規定的紀錄數量

1. 上述關於從queue寫入到file中只要其中一個queue要寫入，其他所有的queue都會跟著寫入
2. commitQueue沒資料時會要求所有的queue寫出資料
3. 重新load資料時commitQueue會先去uncommit.log讀資料，再去seed.log讀去資料（原先filter uncommit間的queue：會直接將資料複寫到seed.log）
4. Uncommit.log的資料是原先在CommitQueue中還未送入Parser的網址
5. 實驗結果
   1. 目前抓取資料量：12677 筆
   2. 抓取時間（網速 15 ~ 20M）
      1. 1159 筆 946 秒
      2. 1191 筆 973 秒
      3. 1068筆 956秒
      4. 1095筆 2398秒
      5. 1115筆 1355秒
      6. 1220筆 959秒
6. 待完成功能
   1. Memory leak
   2. 適當的User Agent Header
   3. 多線程爬蟲
   4. 平行化加速
   5. 廣用正文抽取
   6. 將mysql db改成hash: 增加效能
   7. Server-client 架構
   8. 爬蟲dashboard
   9. 將mysql db改成hash: 增加效能
   10. 漂亮的UI介面
   11. Spam detection