計算機網路報告書

Term Project(Part 2)

資工三甲　B0929034　林芊妤

中華民國 111 年 12 月 25 日

目錄

1. 文件目的

此程式提供可遞迴下載並且離線瀏覽網站的功能，參閱此文件可以理解如何使用，以及可以達到什麼樣的需求。

1.1 程式需求

基本功能:

I. 可選擇遞迴下載的深度(深度為 0 表示不遞迴下載)

II. 顯示各物件下載狀況(資訊、進度、速度、預計剩餘完成時間)

III. 隨時更新目前的下載統計資訊(如已下載檔案總數量、已下載總容量、已下載時間等)

1.2 程式範圍

僅支援 http，以及 html 中有圖片和有外部連結(如 word, ppt, pdf)的網頁。

2. 操作說明

首先，編譯完成後有兩種模式可以輸入參數，第一種是互動模式，第二種是命令列模式。前者是在程式執行後輸出提示訊息要求輸入參數，後者是在執行時直接參數透過命令列輸入。

```
chienfish@ChienfishMacBook-Pro ~/Desktop/cn project2 chienfish ± ./B0929034
請輸入url: http://hsccl.mooo.com/index.htm
請輸入目錄: test
請輸入欲下載之遞迴深度: 0
```

```
chienfish@ChienfishMacBook-Pro ~/Desktop/cn project2 chienfish ± ./B0929034 http://hsccl.mooo.com/ind
ex.htm test
請輸入欲下載之遞迴深度: 0
```

接著，會請使用者輸入欲下載之網頁的 URL、要輸出的資料夾目錄名稱，以及想要下載的遞迴深度。

若輸入之 URL 有誤，也會跳出錯誤訊息，並直接停止執行。如下圖，我使用第二種模式輸入參數，而 URL 的網址是錯誤的，因此會跳出請確認 URL 是否有效。

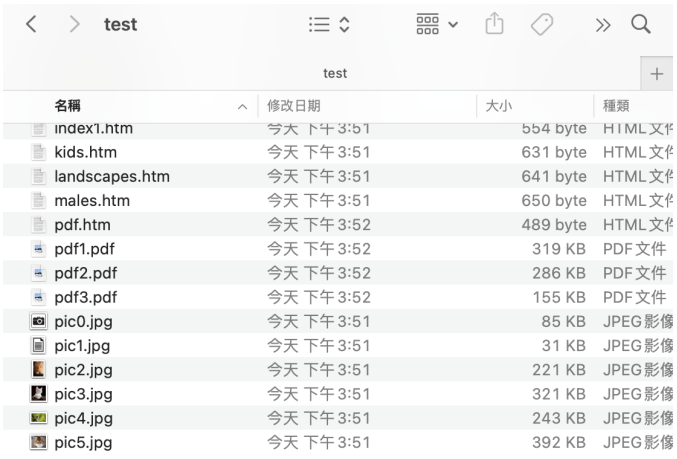
```
chienfish@ChienfishMacBook-Pro ~/Desktop/cn project2 chienfish ± ./B0929034 http://hsccl.mooo.com123/
index.htm test
請輸入欲下載之遞迴深度: 0

>>>>>>進度與步驟<<<<<<
STEP1 | 0.000042(s): Parsing url...
STEP2 | 0.001026(s): Creating/Opening test file...
請確認url是否有效以及連線狀況！
```

以下圖為例，欲下載的網址為 http://hsccl.mooo.com/index.htm，最後要輸出的資料夾目錄為 test，遞迴深度為 4。

```
chienfish@ChienfishMacBook-Pro ~/Desktop/cn project2 chienfish ± ./B0929034 http://hsccl.mooo.com
/index.htm test
請輸入欲下載之遞迴深度： 4
```

接著，在路徑下會找到建立的資料夾，打開後會看到網址上的 htm，以及所有的圖片和 htm 檔。



下載結束後會顯示狀態以及統計資訊，如檔案數量、下載總容量、下載花費時長。另外，也會顯示連線的資訊，如 domain, IP address 等。

```
>>>>>>狀態與統計資訊<<<<<<
http://hsccl.mooo.com/index.htm為有效的url
檔案數量： 52
下載總容量： 8665 KB
下載花費時間： 0.125034 (s)

>>>>>>連線資訊<<<<<<
URL: http://hsccl.mooo.com/index.htm
DOMAIN: hsccl.mooo.com
IP: 163.25.101.225
PORT: 80
```

下載的過程中，會顯示目前執行的步驟在做的事，同時列出預計剩餘完成時間。此外，在步驟和步驟間也會列出是在做哪個物件的下載。

```
>>>>>>進度與步驟<<<<<<
STEP1 | 0.000062(s): Parsing url...
STEP2 | 0.000921(s): Creating/Opening test file...
STEP3 | 0.001876(s): Connecting to server...
-- 物件1 --
STEP4 | 0.000557(s): Downloading the content of index.htm...
-- 物件2 --
STEP5 | 0.000902(s): Connecting to server...
STEP6 | 0.001939(s): Dowdloading http://hsccl.mooo.com/images/image_icon.png...
-- 物件3 --
STEP7 | 0.002515(s): Connecting to server...
STEP8 | 0.000802(s): Dowdloading http://hsccl.mooo.com/images/doc_icon.png...
```

3. 程式說明

說明程式執行之流程，以及其架構和功能。

3.1 程式流程

1.判斷執行檔後是否有加入參數 -> 2.切割 URL(得到 domain, path) -> 3.創建目錄資料夾與切換路徑 -> 4.得到 request 訊息(包含 start-line 和 header) -> 5.建立 socket -> 6.建立 TCP 連線 -> 7.取得 response -> 8.把抓到的 response 解析(拿最後的 message-body) -> 9.檢查並下載 message-body 的照片 -> 10.抓所有<a>的 href -> 11.抓所有<img>的 src -> 12.建立 html 檔 -> 13.下載 src -> 14.下載 href

3.2 程式架構

根據上述流程順序說明，由於 1~9 為 part1 內容，因此從第 10 點開始。

10. 抓所有<a>的 href

allHrefHtm 用來存取所有<a href="">中雙引號裡的字串

```
// 抓所有<a>的href
vector<string> allHrefHtm;
int left = 0, right = 0;
string substr = "href=";
while ((left = content.find(substr, left)) != string::npos) {
    for (int j = left+6; j < content.length(); j++){
        if (content[j] == '"') {
            right = j;
            break;
        }
    }
    string aHREF = content.substr(left+6, right-left-6);
    allHrefHtm.push_back(aHREF);
    left += substr.length();
}
```

11. 抓所有<img>的 src

allIMG 用來存取所有<img src="">中雙引號裡的字串，left 和 right 會傳至 renameSRC 的函式，將其重新命名。

```
// 抓所有照片的src
vector<string> allIMG;
int left = 0, right = 0;
string substr = "src=";
while ((left = content.find(substr, left)) != string::npos) {
    for (int j = left+5; j < content.length(); j++){
        if (content[j] == '"') {
            right = j;
            break;
        }
    }
    string imgSRC = content.substr(left+5, right-left-5);
    renameSRC(left+5, right, allIMG.size());
    allIMG.push_back(imgSRC);
    left += substr.length();
}
```

承 11.

照片會重新被命名成 pic(數字).jpg，其中數字為現在總共存取幾張照片。之後便將 message-body 中 img tag 的 src 都改成上面這些名稱。

```
void renameSRC(int left, int right, int imgNUM) {
    string name = "pic"+ to_string(imgNUM) + ".jpg";
    content.replace(left, right-left, name);
}
```

12. 建立 html 檔

將改過 src 的 message-body 寫入檔案中，檔案已存在的話直接覆寫，若不存在則創建一個。後面 S_IWUSR, S_IRUSR 是給使用者此檔案的權限。

```
//建立html檔案
START = clock();
int fd = open(Path.substr(1, BUFFER.length()).c_str(), O_WRONLY | O_CREAT, S_IWUSR | S_IRUSR);
write(fd, content.c_str(), strlen(content.c_str()));
close(fd);
```

13. 下載 src

如果存下來的照片有 http，則可以直接去 downloadIMG 函式，如果沒有 http，則要將其補上，同時還要加上 domain，再去 download 函式下載。

```
// 下載圖片
for (int i = 0; i < allIMG.size(); i++) {
    string src = "";
    if (allIMG[i][0] == '/') allIMG[i].erase(0, 1);

    if (allIMG[i].find("http") == -1)   src = "http://"+ Domain + "/"+ allIMG[i];
    else   src = allIMG[i];
    downloadIMG(src, sock);
}
```

承 13.

一樣要先切割 URL。接著 send 會將 request 傳出去，之後用 recv 將資料讀到 buf，這邊先讀一次，因為要知道 content-length 是多少。

```
void downloadIMG(string imgSRC, int sd) {
    // 切割imgSRC -> 得到domain&path
    ParseUrl(imgSRC);

    // 發出請求並連線
    printf("-- 物件%d --\n", fileNum);
    string req = "";
    req = NonPerRequest();
    sd = ConnectHttp();
    send(sd, req.c_str(), strlen(req.c_str()), 0);

    // 建立檔案
    START = clock();
    char buf[10000]={};
    string name = "pic" + to_string(imgNUM2) + ".jpg";
    imgNUM2++;
    fileNum++;
    int pic = open(name.c_str(), O_WRONLY | O_CREAT, S_IWUSR | S_IRUSR);
    int recvSize = recv(sd, buf, sizeof(buf)-1, 0);
```

承 13.

從暫存在 buf 的字串中找到 content-length，用 strstr 可以把後面的字串抓出來，while 迴圈是為了只抓出數值，不要其他的文字內容。

```
// 取得img的content-length
string contentLen = "Content-Length: ";
char *sub = strstr(buf, contentLen.c_str()) + strlen(contentLen.c_str());
string targetLen;
while (isdigit(*sub)) {
    targetLen += *sub;
    sub++;
}
```

承 13.

imgContent 是指 message-body 的內容，而 imgContentLen 是將收到的 size 減掉 imgContent-buf(此為 header 的長度)。因為讀取的長度小於緩衝區中的長度，沒辦法一次讀完，需執行 recv 函數多次，直到讀到的字串長度等於前面得到的 targetLen。

```
// 取得圖片內容並寫入檔案
char *imgContent = strstr(buf, "\r\n\r\n") + 4;
int imgContentLen = recvSize-(imgContent-buf);
write(pic, imgContent, imgContentLen);
while (true) {
    recvSize = recv(sd, buf, sizeof(buf)-1, 0);
    write(pic, buf, recvSize);
    imgContentLen += recvSize;

    if (imgContentLen >= stoi(targetLen)) break;
}
fileSize += imgContentLen;
close(pic);
shutdown(sd, SHUT_WR);
```

14. 下載 href

如果存下來的 href 有 http，則可以直接去 downloadHrefHtm 函式，如果沒有 http，則要將其補上，同時還要加上 domain，再去 download 函式下載。

```
// 下載href的htm
for (int i = 0; i < allHrefHtm.size(); i++) {
    string src = "";
    if (allHrefHtm[i][0] == '/') allHrefHtm[i].erase(0, 1);

    if (allHrefHtm[i].find("http") == -1)   src = "http://"+ Domain + "/"+ allHrefHtm[i];
    else   src = allHrefHtm[i];
    downloadHrefHtm(src, sock, depth);
```

承 14.

下載 href 分成兩種，一是下載.htm，二是下載其他(如 word, ppt, pdf)，而記錄深度就是在這邊。downloadHrefHtm 跟下載最一開始的 index.htm 一樣，都必須發出 request、建立連線；差異是在抓目前這個 htm 的<a>時要特別判斷他的 href 是外部連結(不同 Domain)，抑或是其他。

```
string aHREF = subContent.substr(left+6, right-left-6);
if (aHREF.substr(0,4) == "http") {
    if (aHREF + "/index.htm" != url) {
        string TmpDomain = Domain, TmpPath = Path, name = "index1.htm";
        ParseUrl(aHREF + "/index.htm");
        createObj();
        subContent.replace(left+6, right-left-6, name);
        Domain = TmpDomain;
        Path = TmpPath;
        left += substr.length();
    }else {
        string name = "index.htm";
        subContent.replace(left+6, right-left-6, name);
        break;
    }
}else if (aHREF.substr(aHREF.length()-4,aHREF.length()) != ".htm") { //建立連線->下載.pdf->更改名字
    int x = aHREF.find_last_of("/");
    string name = aHREF.substr(x+1, aHREF.length());
    subContent.replace(left+6, right-left-6, name);
    allHrefOther.push_back(aHREF);
    left += substr.length();
}else {
    allHrefHtm.push_back(aHREF);
    left += substr.length();
}
```

第一個 if 是判斷是否是一樣的 domain，如果不是的話，則要特別幫他取名(這邊取作 index1)，否則會跟最一開始的重複。第二個 else if 是判斷 href 是一般.htm 還是其他，若是其他則用 allHrefOther 這個 vector 來存。最後 else 是指是一般的.htm 檔，則把它放進 allHrefHtm 這個 vector。

承 14.

最後是下載，每個 vector 是不同種類，都有屬於自己的 download 函式。

```
// 下載圖片
for (int i = 0; i < allIMG.size(); i++) {…

// 下載href的htm
for (int i = 0; i < allHrefHtm.size(); i++) {…

// 下載href的其他(word, ppt, pdf)
for (int i = 0; i < allHrefOther.size(); i++) {
    string src = "";
    if (allHrefOther[i][0] == '/') allHrefOther[i].erase(0, 1);

    if (allHrefOther[i].find("http") == -1)   src = "http://"+ Domain + "/"+ allHrefOther[i];
    else    src = allHrefOther[i];
    downloadHrefOther(src, sd, depth);
}
```

downloadHrefOther 跟 downloadIMG 的讀取內容方式一樣，只差在修改.htm 名稱的地方不同，附圖為 other 的取名方式。

```
int x = href.find_last_of("/");
string name = href.substr(x+1, href.length());
```

4. 參考資料

- 建立 socket

  https://snsd0805.github.io/jekyll/update/2019/05/27/筆記-Linux 環境用 c++建立 Socket 連線.html

  http://www.tsnien.idv.tw/Internet_WebBook/chap8/8-4%20Socket%20 傳輸位址.html

  https://stackoverflow.com/questions/9400756/ip-address-from-host-name-in-windows-socket-programming

- http client request

  https://notfalse.net/47/c-socket-http-client

  https://www.796t.com/content/1550270006.html

- 設定命令列模式參數

  https://edisonx.pixnet.net/blog/post/57060736-%5Bt%5D-vs-設定命令參數列