CS305 作業系統概論 Homework #2 Multithreading

2018.04.18

一、 作業目的

熟悉如何使用Pthreads的API，撰寫multithreaded program。

二、 作業內容

【大數據中的關鍵文件】為了要在許多文件中找出關鍵文件，L公司想要來利用電腦科技來達成目標。對L公司而言，在一個有M個文件的文件集合$D=\{d_1,d_2,\ldots,d_M\}$中，關鍵文件d_k就是與其他文件的相似度總和最高的文件。但對於如何快速計算文件的相似度，L公司卻毫無頭緒。

於是L公司來到風之塔學院尋求幫助。對於這個大數據問題，風之塔學院的C教授帶著他的高徒開發這個程式。C教授決定先使用傑卡德相似係數(Jaccard similarity coefficient)的方法來計算，找出關鍵文件。傑卡德相似係數的公式如下，在兩個集合 A 與 B 中，兩個集合A和B的交集元素個數在A，B的聯集元素個數中所佔的比例，就是它們的傑卡德相似係數：

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}$$

因此如果對下面兩個文件

d_1= “this is a book”

d_2= “this is a pen”

$J(d_1, d_2)=3/5=0.6$

C教授同時要用multithreaded programming 的方式來設計程式。每一份文件對其他M-1份文件的平均傑卡德相似係數(Average $J(d_1, d_2)$)是由一個單獨的thread 來計算出來。所有文件會放在一個檔案中，程式須由命令列讀入檔名。檔案中最多會有50個文件。每個文件會有兩行資料，第一行是文件的ID，第二行是文件內容。文件ID會是一個字串，文件內容中的字詞會由一個或多個空白隔開。傑卡德相似係數如果有多位小數，需要至少精準到小數點後5位。在處理文件時，依照下面規則處理：

1. 只考慮純字母組成的詞。
2. 如果有重複出現的詞，只計算一次。例如： “a good book is a book” 中， “book”只算出現一次。

在程式執行時，

1. 主執行緒針對文件數量產生對應的子執行緒。例如有4份文件，就產生4個子執行緒。主執行緒並負責印出來下列事項，印出內容時，每一行需要印出 “[Main thread]”：
   a. 每一個子執行緒的 tid，以及所負責計算的主文件ID。
   b. 具有最高平均傑卡德相似係數的文件ID及文件內容。
   c. 整個程式會用多少CPU時間 (以ms為單位)。
2. 子執行緒則負責計算傑卡德相似係數。執行過程中，要列印出本身的動作，並且每一行都要印出自己的thread id 。以下是需要印出的項目：
   a. 負責計算的主文件ID編號。
   b. 子執行緒計算傑卡德相似係數時，要印出是哪兩個文件在計算，以及它們Jaccard similarity。
   c. 最後的平均傑卡德相似係數。
   d. 子執行緒執行會用多少CPU時間 (以ms為單位)。

以下是一個可能的執行過程：

> prog2 data.txt

[Main thread]: create TID:123, DocID:0001

[TID=123] DocID:0001

[TID=123] J(0001,0002)=0.6

[TID=123] J(0001,0003)=0.6

…

[TID=123] AvgJ:0.511

[TID=123] CPU time: 20ms

…

[Main thread] KeyDocID:0003 HighestJ:0.9999

[Main thread] CPU time: 2000ms

二、 作業要點

1. 請注意，本作業使用的程式語言是C/C++，測試平台的作業系統： Ubuntu 17.10 LTS 64-bit。使用的編譯程式為gcc/g++ 編譯器：7.2。其他平台或程式語言不在本次作業考慮範圍之內。如在測試平台上無法編譯與執行，都不予給分。
2. 請注意，本作業一定要用Pthread API來進行。任何不用Pthread API的程式，都不予給分。
3. 本作業的評分方式如下：
   a. 每一個項目能正確執行時，最多可得的分數如下
      i. 從命令列讀入檔名參數，20分。
      ii. 能產生 pthread，10分。
      iii. 子執行緒可以印出本身的tid，20分。
      iv. 傑卡德相似係數計算。不可以使用任何套件或函式庫，需自己完成。20分。
      v. 印出執行所用的CPU時間，20分。
      vi. 主執行緒找出關鍵文件並印出它的平均傑卡德相似係數，20分。
4. 本作業需繳交檔案：
   a. 說明報告：檔案為docx或pdf格式。
      i. 報告中必須說明程式的設計理念、程式如何編譯，以及**如何操作**。
      ii. 報告中同時必須詳細說明你完成哪些部份。如有用到特殊程式庫，請務必說明。
      iii. 請務必讓助教明白如何編譯及測試你的程式。助教如果無法編譯或測試，會寄信（最多兩次）通知你來說明，但每說明一次，**助教會少給你10分**。
   b. 完整原始程式碼檔案。程式碼檔案必須是可直接編譯的檔案。**不可含執行檔。助教會重新編譯你們的程式**。
5. 所有相關檔案，例如報告檔、程式檔、參考資料等，請壓縮成一個壓縮檔（不可超過2MB）後上傳至portal。**請注意，不可抄襲。助教不會區分何者為原始版本，被判定抄襲者，一律0分。**
6. **如果傑卡德相似係數計算有使用網路範例，務必在作業中說明。該部份將不會計分，但不會判定為抄襲。**

三、 繳交方式：

1. 最終繳交時間：
   a. 電子檔在 2018.05.11 以前，上傳至個人portal。如有多個檔案，將所有檔案壓縮成zip（rar 亦可）格式，然後上傳。
   b. 上傳檔名格式：「學號_作業號碼.docx」或「學號_作業號碼.rar」。例如：912233_01.doc 或 912233_01.rar。
2. **如有違規事項者，依照課程規定處理**。

3. 如需請假，請上portal請假，並持相關證明文件，在請假結束後的第一次上課時完成請假手續，並在一週內完成補交。補交作業將以8折計算。
4. 老師不接受「門縫」方式繳交，助教也不接受任何作業。

四、 如有未盡事宜，將在學校portal板面公告通知。

五、 If you need **any assistance in English**, please contact Prof. Yang.

六、 參考資料

1. 參考課本圖 4.9。
2. PThread: https://computing.llnl.gov/tutorials/pthreads/
3. POSIX 線程 (pthread) 入門文章分享：http://dragonspring.pixnet.net/blog/post/32963482-posix%E7%B7%9A%E7%A8%8B%28pthread%29%E5%85%A5%E9%96%80%E6%96%87%E7%AB%A0%E5%88%86%E4%BA%AB
4. Jaccard index wiki: https://en.wikipedia.org/wiki/Jaccard_index