CS305 作業系統概論 Prog. #2 Multithreaded Programming

2022.03.29

一、 作業目的

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

二、 作業內容

【大數據中的關鍵文件】L上尉為了要在許多文件中找出關鍵文件，想要來利用電腦科技來達成目標。所謂的關鍵文件如下，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平均值最高的文件。但對於如何快速計算文件的相似度，L上尉卻毫無頭緒。

基本要求：

於是L上尉繼續來到海豚書院尋求幫助。對於這個大數據問題，海豚書院的JC王牌帶著他的高徒開發這個程式。JC王牌決定使用必殺絕技餘弦相似係數(Cosine similarity coefficient)的方法來計算，找出關鍵文件。Cosine similarity方式如下：

1. 先算出每一篇文章中的詞在文章中出現的數量，這個數量稱之為詞頻（term frequency）。
2. 按照所有文章的全部詞彙，接著將每一篇文章轉換成一個詞頻向量。
3. 接著，用出絕技，算出來源文章向量（V_s）與其他每一篇文章向量（V_x）的向量空間Cosine 值。這個公式如下：

$$Sim(V_s,V_x)=\cos(V_s,V_x)=\frac{V_s\cdot V_x}{|V_s|\times|V_x|}=\frac{\sum_{i=1}^{n}v_{s,i}\times v_{x,i}}{\sqrt{\sum_{i=1}^{n}v_{s,i}^2}\times\sqrt{\sum_{i=1}^{n}v_{x,i}^2}}$$

4. 接著就可以算出平均餘弦相似係數。

例如假設有4份文件，文件內容如下:

yuan ze university is a good university this university has many very good students

there are many students in yuan ze university many students are very good

there are good books in yuan ze university students love to read these books

there are many good teachers in yuan ze university these teachers concern these students

算出詞頻

	yuan	ze	university	is	a	good	this	has	many	students	there	are	in	very	books	love	to	read	these	teachers	concern
V_s	1	1	3	1	1	2	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0
V_1	1	1	1	0	0	1	0	0	2	2	1	2	1	1	0	0	0	0	0	0	0
V_2	1	1	1	0	0	1	0	0	0	1	1	1	1	0	2	1	1	1	1	0	0
V_3	1	1	1	0	0	1	0	0	1	1	1	1	1	0	0	0	0	0	2	2	1

得到的文章向量分別是

V_s = [1,1,3,1,1,2,1,1,1,1,0,0,0,1,0,0,0,0,0,0,0]

V_1= [1,1,1,0,0,1,0,0,2,2,1,2,1,1,0,0,0,0,0,0,0]

V_2= [1,1,1,0,0,1,0,0,0,1,1,1,1,0,2,1,1,1,1,0,0]

V_3= [1,1,1,0,0,1,0,0,1,1,1,1,1,0,0,0,0,0,2,2,1]

Cos_sim(V_s ,V_1) =0.586939185653

Cos_sim(V_s ,V_2)=0.426401432711

JC王牌同時要用multithreaded programming 的方式來設計程式。每一份文件對其他M-1份文件的平均餘弦相似係數(Avg_Cos(d_1, d_2))用一個單獨的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。餘弦相似係數如果有多位小數，需要至少精準到小數點後4位。在處理文件時，依照下面規則處理：

1. 如果文章中有標點符號，這些標點符號都先轉成空白字元。
2. 只考慮純字母組成的詞，其他如果不是純字母組成的詞則予以忽略不計。例如 “64”、 “test20”。

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. 具有最高平均餘弦相似係數的文件ID及文件內容。
   c. 主執行緒會用多少CPU時間 (以ms為單位)。
2. 子執行緒則負責計算餘弦相似係數。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，以及它的詞頻向量。
   b. 子執行緒計算餘弦相似係數時，要印出是哪兩個文件在計算，以及它們Cosine similarity。
   c. 最後的平均餘弦相似係數。
   d. 子執行緒執行會用多少CPU時間 (以ms為單位)。
3. 如果有多個文件具有相同的平均餘弦相似係數，則由它們的文件ID大小來決定。ID最小的為關鍵文件。

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 [1,1,3,1,1,2,1,1,1,1,0,0,0,1,0,0,0,0,0,0,0]

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

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

…

[TID=123] Avg_cosine: 0.511

[TID=123] CPU time: 20ms

…

[Main thread] KeyDocID:0003 Highest Average Cosine: 0.9999

[Main thread] CPU time: 2000ms

三、 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 21.10 64-bit。使用的編譯程式為gcc/g++ 編譯器：11.2。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。本項滿分10分。

ii. 能產生正確數量的 pthread。例如如果有15份文件，就要產生15個thread，分別對每個文件計算它的關鍵文件分數。一個thread只負責一份主文件。本項滿分20分。

iii. 子執行緒可以印出本身的tid。本項滿分10分。

iv. 正確計算出文件的詞頻向量。本項滿分20分。

v. 正確進行餘弦相似係數計算。程式碼中不可以使用任何套件或函式庫，必須有完整的程式碼。本項滿分20分。

vi. 每一個 thread 都印出執行過程所用的總共CPU時間，以ms為單位。本項滿分20分。注意，是CPU時間。

vii. 主執行緒找出關鍵文件並印出它的平均餘弦相似係數。本項滿分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（.c 或 . cpp）。**不可含執行檔。助教會重新編譯你們的程式**。**請注意**：也不可用 .txt檔或是 .docx檔等非正常方式繳交程式碼，如有類似情形，**助教也會扣10分**。

c. **不可以含有病毒，如果含有病毒等惡意程式，本作業0分**。

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，被判定抄襲或雷同者，一律0分。**

四、 繳交方式：

1. 最終繳交時間：

a. 程式作業檔在 2022.04.29 以前，上傳至個人portal。如有多個檔案，必須將所有檔案壓縮成一個zip（rar 亦可）檔案，然後上傳。

b. 上傳檔名格式：「學號_作業號碼.rar」或「學號_作業號碼.zip」。例如：912233_01.rar 或 912233_01.zip。

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。作業都必須以上傳至portal的方式繳交。

五、 如有未盡事宜，將在個人portal板面公告通知。

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

七、 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. Thread CPU time: https://linux.die.net/man/3/clock_gettime

5. 時間的計算：
6. https://2formosa.blogspot.com/2017/06/time-elapased-to-excute-a-command.html