COMP SCI 4094/4194/7094 - Distributed Databases and Data Mining
Assignment 1

DUE: 23:30 Friday, 15th Sept, 2017

Important Notes

- Handins:
  - The deadline for submission of your assignment is **23:30 Friday, 15th Sept, 2017**.
  - You must do this assignment individually and make individual submissions.
  - Your program should be coded in **C++** and pass test runs on the two test files. The sample input and output files are downloadable from the track of Assignment 1 at MyUni (`https://myuni.adelaide.edu.au/courses/25366/assignments/40064`).
  - You need to use `svn` to upload and run your source code in the web submission system following Web-submission instructions stated at the end of this sheet. You should attach your name and student number in your submission.
  - Late submissions will attract a penalty: the maximum mark you can obtain will be reduced by 25% per day (or part thereof) past the due date.
- Marking scheme:
  - 12 marks for testing on 2 standard tests: 6 marks per test.
  - 3 marks for the code and comments.
  - **Note:** If it is found your code did not implement the required computation tasks in this assignment, you will receive zero mark regardless of the correctness of test output.

If you have any questions, please send them to the student discussion forum. This way you can all help each other and everyone gets to see the answers.

The assignment

In the past decade, network management has been benefited from traffic classification to classify the raw internet traffic packets to different applications, such as http, smtp, dns, msn and ftp. For a given traffic pack, the attributes we may obtain include source address, source port, destination address, destination port and packet length. Assume that the raw traffic data are distributed over a set of servers at which users submit their applications (queries) as needed. It is observed that some attributes are often accessed together by user applications and hence show a high *affinity*. The affinity $aff(A_i, A_j)$ between attributes A_i and A_j is defined by Cosine-based similarity (`https://en.wikipedia.org/wiki/Cosine_similarity`) and computed by the following equation, where n is the number of attributes, m is the number of sites, m_k is the number of times query q_k is issued from all sites, and a_{ik} is the total number of times attribute A_i is accessed by q_k.

$$aff(A_i, A_j) = \frac{\sum_{k=1}^{m_k} a_{ik} \times a_{jk}}{\sqrt{\sum_{k=1}^{m_k} (a_{ik})^2} \times \sqrt{\sum_{k=1}^{m_k} (a_{jk})^2}} \tag{1}$$

$$a_{ik} = use(q_k, A_i) \times \sum_{j=1}^{m} acc_j(q_k) \tag{2}$$

In this assignment, to localise data accesses of users applications, you are required to code a C++ program that partitions the given traffic flow attributes according to user applications such that attributes of a high affinity are placed within the same group. Your code should round off the decimals in the results of all divisions to four digits only.

Hint: You should apply the appropriate algorithms learned from the first part of this course – distributed databases.

Example

You are given the following inputs:

- Attributes (A_i):

Label	Name
A_1	SrcAddr
A_2	SrcPort
A_3	DstAddr
A_4	DstPort

- Queries (q_i) and their access frequencies at different sites (S_i):

	Queries	S_1	S_2	S_3
q_1	SELECT DstAddr FROM PROJ WHERE SrcAddr=Value	15	20	10
q_2	SELECT SrcPort, DstAddr FROM PROJ	5	0	0
q_3	SELECT SrcPort FROM PROJ WHERE DstPort=Value	25	25	25
q_4	SELECT DstAddr FROM PROJ WHERE DstPort=Value	3	0	0

Your partitioning program will generate the the following em clustered affinity matrix as output:

	A1	A3	A2	A4
A1	1.0000	0.9917	0.0000	0.0000
A3	0.9917	1.0000	0.0073	0.0026
A2	0.0000	0.0073	1.0000	0.9970
A4	0.0000	0.0026	0.9970	1.0000

Web-submission instructions

- First, type the following command, all on one line (replacing xxxxxxx with your student ID):
```
svn mkdir --parents -m “DDDM”
https://version-control.adelaide.edu.au/svn/axxxxxxx/2017/s2/dddm/assignment1
```

- Then, check out this directory and add your files:
```
svn co https://version-control.adelaide.edu.au/svn/axxxxxxx/2017/s2/dddm/assignment1
cd assignment1
svn add PartitionAttributes.cpp
```

. . .
svn commit -m “assignment1 solution”

- Next, go to the web submission system at:
  `https://cs.adelaide.edu.au/services/websubmission/`
  Navigate to 2017, Semester 2, Distributed Databases and Data Mining, Assignment 1. Then, click Tab “Make Submission” for this assignment and indicate that you agree to the declaration. The automark script will then check whether your code compiles. You can make as many resubmissions as you like. If your final solution does not compile you will not get any marks for this solution.

- **Note:**

  1. Your PartitionAttributes.cpp should accept two input files in the order of Attributes, Queries and Access frequencies then print the CA matrix as the output.
  2. Please follow the forms in sample output files.
  3. Your local file path will not work with our web-submission system.