

COMP9318 Assignment 1

Due Date: 23:59 4 June, 2009 (Thu)

DESCRIPTION

Q1

(30%) Despite the economy downturn, you are hired by the *National Aussie Bank* (NAB) to design a data warehouse to analyze the **loans**. The relational schema of the operational database is given below.

Branch(branchNo, bStreetAddress, bCity)

LoanManager(empNo, empName, phone, branchNo)

Customer(custNo, custName, profession, streetAddress, city, state)

Account(accNo, accType, balance, accDate, custNo)

LoanContract(contractNo, loanType, amount, loanDate, empNo, custNo)

where primary key attributes have been underlined. Most attribute names are self-explanatory. *loanDate* records the date the loan is established. There could be more than one loan manager in a branch.

Design a star-schema for analyzing the loans. You can refer to the following frequently issued queries as the statement of requirements.

QueryID	Description
Q1	The total amount of loans in 2008.
Q2	For the type of loans with more than 10 loan contracts, the type of loan and the number of contracts.
Q3	For each loan type and each loan city and state where customer resides, the total loan amount in 2007.
Q4	For each type of loan and each customer profession, list the total loan amount.
Q5	The performance (aka. total loan amount) of each loan manager in 2008.
Q6	The total loan amount of each branch.
Q7	The total loan amount for each branch and city.

You may make any reasonable assumptions (state them).

Q2

(40 + 5%) Consider the following database of strings.

ID	String
1	ababa
2	abcaba
3	aacabc
4	abcdefgh

1. Show all the 3-grams of the strings.
2. Find all pairs of strings within edit distance 1 using the prefix-based similarity join algorithm. You need to illustrate your steps.
3. A problem with the edit distance is that it favors short strings (if we assume humans make typographical errors with a constant probability per key stroke, then we can expect finding more typos in long strings than in short strings). One proposal to remedy this problem is to use the *normalized edit distance*, defined as

$$ned(s, t) = \frac{ed(s, t)}{\max(|s|, |t|)}$$

where $ed(\cdot, \cdot)$ is the edit distance, and $|s|$ is the length of the string s . Proof the following constraint on the lengths of two strings that satisfy a certain *ned* threshold of α .

For two strings s and t such that (1) $|s| \leq |t|$, and (2) $ned(s, t) \leq \alpha$, the following must be true:

$$|s| \leq |t| \leq \frac{1}{1 - \alpha} |s|$$

4. (**Bonus**) How to adapt the prefix-based similarity join algorithm for normalized edit distance with threshold of α ?

Q3

(30%) Consider the given **similarity** matrix. You are asked to perform average-link hierarchical clustering on this dataset. You need to show the steps and final result of the clustering algorithm. You will show the final results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

	p_1	p_2	p_3	p_4	p_5
p_1	1.00	0.10	0.41	0.55	0.35
p_2	0.10	1.00	0.64	0.47	0.98
p_3	0.41	0.64	1.00	0.44	0.85
p_4	0.55	0.47	0.44	1.00	0.76
p_5	0.35	0.98	0.85	0.76	1.00

SUBMISSION DETAILS

1. This is an individual assignment.
2. You should submit a PDF document (in electronic form) named `ass1.pdf`.
3. Please submit the document via the **GIVE** system before the deadline.
The command is:

```
give cs9318 ass1 ass1.pdf
```

4. Please also make sure all the documents can be opened and **PRINTED** correctly.
5. Late submission policy: we use soft penalty at the rate of -10% if late by one day, and -20% per day onwards.
6. The size of the `ass1.pdf` file should not exceed 1MB.