CIS400/600 Fundamentals of Data and Knowledge Mining

Homework #1

Spring, 2017

Problem 1 (5 pts)

Discuss why a document-term matrix is an example of a data set that has asymmetric discrete or asymmetric continuous features.

Problem 2 (10 pts)

Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word term) in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by

$$tf'_{ij} = tf_{ij} * \log \frac{m}{df_i}$$

where df_i is the number of documents in which the i^{th} term appears and is known as the document frequency of the term. This transformation is known as the inverse document frequency transformation.

(a) What is the effect of this transformation if a term occurs in one document? In every document?

(b) What might be the purpose of this transformation?

Problem 3 (15 pts)

Download sales dataset posted in this assignment and use R to apply at least THREE most effective visualization techniques to explore different aspects of the dataset.

Please paste the visualizations with both R codes and a brief explanation of the visualizations in your answer.

Problem 4 (20 pts)

Use the same sales dataset as Question 3 and program in R to convert it into a multidimensional cube with four dimensions *product, year, month* and *state*. Each cell in the cube represents an aggregate value for a unique combination of all the dimensions.

Employ the sales data cube developed from above to answer the following questions:

1. Slice operation: compute the revenue for Laptop during January of 2013 in each state.
2. Dice operation: compute the revenue for the furniture products (Mattress and Chair) during the second quarter (April, May and June) of 2014 in each state.
3. Rollup operation: compute the annual revenue for each product and collapse the state and month dimensions.
4. Drilldown operation: compute the annual and monthly revenue for each product and collapse the state dimension.

Please include your R codes that convert sales dataset into cube and implement four OLAP operations together with various revenue computation outputs in your answer.