Data Preprocessing

- Data Preprocessing : An Overview

Data Quality

Major Tasks in Data Preprocessing

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization

- **Data Cleaning**

Fill in missing values, smooth noisy data , identify or remove outliers and resolve inconsistencies.

- **Data Integration**

Integration of multiple databases , data cubes, or files

- **Data reduction**

Dimensionality reduction

Numerosity reduction

Data Compression

- **Data transformation and data discretization**

Normalization

Concept hierarchy generation

Data Cleaning

Data in the Real World is dirty , lots of potentially incorrect data .

Eg occupation =""(missing data) , Age = 32, DOB = 7/12/76, rating = 'A,B,C'(inconsistent) , Salary = "-10"(error)

Missing data may due to

- Equipment malfunction
- Inconsistent with other recorded data and thus deleted
- Data not entered due to misunderstanding
- Certain data may not be considered important at the time of entry
- Not register history or changes of the data

Data Integration

- Data Integration

Combines data from multiple sources into a coherent store

- Scheme integration eg A cust –id , A.cust-#

Integrate megadata from different sources

- Entity identification problem:

Identify real world entities from multiple data sources , eg Bill Clinton = Willaim

Clinton

- Detecting and resolving data value conflicts
- For the same real world entity , attribute values from different sources are different
- Possible reasons: different representations , different scales

Data Redundancy

- An attribute (such as annual revenue, for instance) may be redundant if it can be derived from another attribute or set attributes
- Some redundancies can be detected by correlation analysis .Given two attributes , such analysis can measure how strongly attribute implies the other , based on the available data
- For nominal data we use X^2 (Chi-Square test)
- Chi - squared test (X^2) = Σ(observed – expected)^2 / expected

The larger the X^2 value the more likely variables are related

For numeric we can use correlation coefficient or covarince

Panorama Stitching Technique

- Image Stitching or photo stitching is the process of combining multiple photographic images with overlapping fields of view to produce a segmented panorama or high resolution image.

- Although some stitching algorithms actually benefit from differently exposed images by doing high dynamic range imaging in regions of overlap.

- **The fundaments of the typical image stitching algorithm require four key steps:**

1) Detecting keypoints(DoG, Harris, etc.) and extracting local invariant descriptors(SIFT,SURF,etc) from two input images

2)Matching the descriptors between the images

3) Using the RANSAC algorithm to estimate a homography matrix using our matched feature vectors.
4) Applying the warping transformation using the homography matrix obtained from step.

SIFT and SURF are recent key-point or interest point detector algorithms but a point to note is that these are patented and their commercial usage restricted. Once a feature has been detected , a descriptor method like SIFT descriptor can be applied to later match them.

To estimate a robust model from the data , a common method used is known as RANSAC ('Random SAmple Consensus').If the ratio of number of outliers to data points is very low, the RANSAC outputs a decent model fitting the data.