Week 2

From the Expert: Data

Data

A data set is a collection of data objects. Alternative names for a data object are: record, vector, pattern, sample, observation, entity, etc. The characteristics of an object are called attributes. For example, a student can be described using attributes such as student ID, year, major, and GPA. Types of data attributes can be defined as follows:

1. **Qualitative data** (or categorical data) are non-numeric can be further divided into:

   1.1 Nominal data

   As the name suggests, the values of this data are just different names with no order. Names or labels are given to various categories. For example: eye color {brown, black, green, blue}, directions {North, East, West, South}, zip codes {80221, 20500, 90210}. Even though zip codes are represented as numbers, they should be treated as symbols.

   1.2 Ordinal data

   This data type has order but does not have a numerical scale between measurements. For example, customer service ratings {poor, fair, good, best}; 'best' is better than 'good' but best is not double or triple of good.

2. **Quantitative data** or numeric data are represented by numbers and have the properties of numbers. The value can be discrete or continuous. Discrete values are represented using integer. Binary is the special case of discrete type with only two values usually represented as Boolean type such as 1/0 for Yes/No.

   2.1 Interval data

   The differences between two values are meaningful but the ratio is not meaningful, such as Intelligence Quotient (IQ) or temperature in Celsius vs. Fahrenheit. For example, 80 Fahrenheit is different from 40 Fahrenheit but is not twice (hotter) than of 40 Fahrenheit.

   2.2 Ratio data

   Both difference and ratio are meaningful such as age, weight, and height. For example, 100 pounds is different from, and twice as much as 50 pounds.

Quantitative and qualitative data are often used together to get a full picture of the population or objects in which we are interested.

Data Characteristics

Some characteristics of data sets (Tan, 2005)

1. **Dimensionality**

Dimensionality refers to the number of attributes that belong to the objects in the data set. Data with a large number of dimensionality is known as the curse of dimensionality because of the difficulty to manage and analyze.

2. **Sparsity**
   In some data sets, the attribute of an object has value of 0 most of time. Consider the items purchased from the grocery store, an item that is purchased has a value of 1 and 0 otherwise. Only a small number of items have been purchased from more than 100,000 items in the grocery stores so most of the value in the data set would be 0.

3. **Resolution**
   Generally, data have different properties at different resolutions. For example, the 93 million miles distance from the sun to earth makes the sun looks smaller than its actual size.

4. **Transaction or Market Basket Data**
   Consider the point of sale; the set of items purchased by a customer at a particular shopping trip is called transaction or market basket data.

5. **Data Matrix**
   A data set with fix number of numeric attributes. It can be displayed as m (rows) by n (columns) matrix, where m denotes each object and n signifies attributes or vice versa. Standard matrix operations can be applied to this data

6. **Sequential Data**
   It is data with associated time. This is also known as temporal data. For instance, a customer buys a house a time t1, buys a new appliance at time t2, etc. The time is important for interesting discovery such as peak home sales during summer.

7. **Sequence Data**
   It is similar to sequential data but there is no time associated to it. This data set consists of sequence of letters or words. TGGGCTGCCTGCTCGA, for instance, is an example of human genomic sequence data represented using four nucleotides A, T, C, and G.

8. **Time Series Data**
   This can be considered as a special type of sequential data. Consider, the price of different stocks that varies through the day.

9. **Spatial Data**
   This data set includes position as one of the attributes such as the temperature at different geographical locations.

Data Processing Stages:
Even though there are different types of data, the process for analyzing the data consists of these following stages, namely: data acquisition, data extraction, data preprocessing (i.e. data cleaning), data analysis and data presentation. The last two stages will be discussed in detail later.

Data Acquisition
Data can be acquired from both internal and external sources. Internal sources contain valuable data, which are obtained from within a corporates or government entities such as data warehouses, operational databases, document stores, web or email log files, etc. External data is used to supplement the solution requirements. Many public sources are available with no cost such as U.S. Census Bureau, U.K government, World Bank, Wikidata, Amazon web services, Gnip, ESRI, Standard & Poor's etc.

Database, Data Warehouse
A relational database is a collection of data items organized as a set of tables or relations. One table (entity) can reference another table using common field or 'key'. Relating one table to another table is known as 'join'. For a relational database, joins are common but it is quite an expensive operation.

A Data Warehouse is a repository of the enterprise database for modeling, analysis and strategic decision-making. It is maintained separately from the enterprise' operational database. A well-known example of a data warehouse is Facebook. Facebook collects myriad of data about users, friends, groups, etc. The relevant aggregated data are stored in a central repository database to analyze users behavior and preferences. The process of combing data from multiple sources such as relational database, flat file, on-line transaction records into a single comprehensive source is called data warehousing.

Data Extraction
Input data in different formats are obtained and converted into a format that can be further processed. The (file) format refers to structure or way to encode information in computer file storage.

- Fixed-Width and Delimited Text Files is a common format for data exported from databases and spreadsheets. Fixed width, as the name suggested, contain a specific number of characters. Whereas, delimited fields are separated by characters such as comma or tab.

- JSON (JavaScript Object Notation) is an easy to understand format in the form of plain text, which is self-describing. It is hierarchical like XML but shorter and quicker to read and write.

Syntax of JSON includes data is in name/value pairs, separated by comma, objects are contained in curly braces, whereas arrays are contained in square brackets.

JSON Example:

```
{
  “student” : [
     { “firstname”: “Annie”, “lastname”:”Johnson”,”year”:”Freshman”},
     { “firstname”: “Brook”, “lastname”:“Shield”,”year “:”Senior”},
    [
}
```

- Binary Input Formats such as PDF documents, word processor, spreadsheets, images, audio and video. This data is not in the text format. It may be necessary to convert into text-based such as CSV or text before analysis.

- SequenceFile (http://wiki.apache.org/hadoop/SequenceFile) is a flat binary file with key/value pairs. It is extensively used as input/output format in MapReduce and temporary outputs of maps.

Data Preprocessing

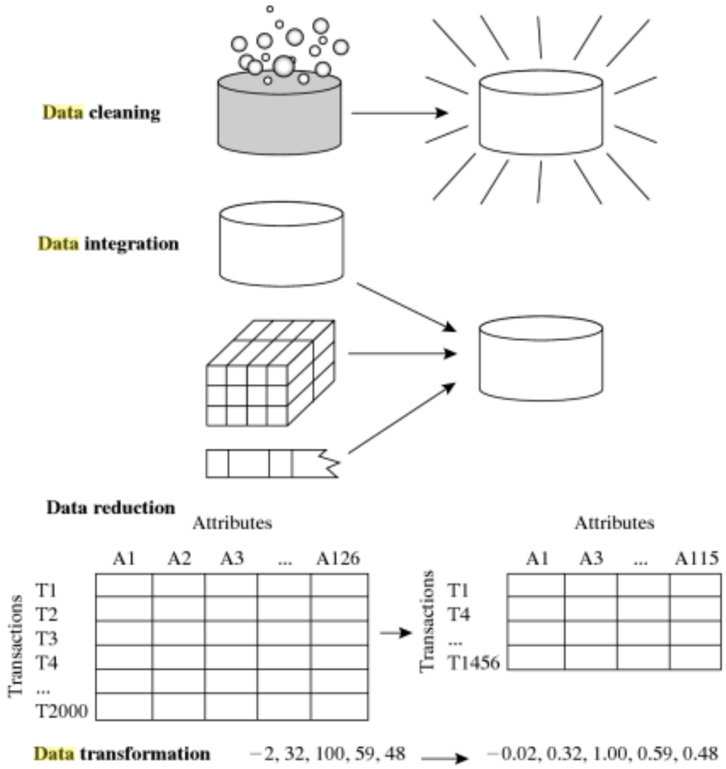
Data in the real world may be incomplete (i.e. missing attribute values), inaccurate or noisy (i.e. salary has a negative value), and inconsistent (i.e. the state of Colorado is ‘CO’ or ‘co’, different format, same person with different address). To make data more reliable and suitable for data analysis, preprocessing is mandatory.

According to (Han, J.W, 2011), major tasks in data preprocessing consists of:

- **Data cleaning**
  Dirty data can make results unreliable. Therefore, it is necessary to clean the data before analyzing it. Data cleaning processes include: filling in missing values, resolve inconsistencies, identifying outliers, etc.
- **Data integration**
  Analysis may need data from multiple sources. Therefore, an integrating process is needed to create one single database. Concerns include inconsistency (i.e., different database has different attribute names, format, different scale) and data redundancy
- **Data reduction**
  Large volumes of data can slow down the analyzing process. A reduced representation that retains the same analytical results is favorable. Examples of the strategy are dimensionality reduction and numerosity reduction (i.e. sampling, filtering). Sampling captures a subset of the available data. Filtering limits unrelated data.
- **Data transformation**
  Transformation is a process that applies to all attribute values of a variable. Common techniques are anonymization, normalization, smoothing (remove noise), aggregation (summarization), attribute construction, etc.

Anonymization is the process for removing personally identifiable information (PII) such as identification numbers (i.e. SSN), name, address, phone numbers, etc. due to laws, standard practices, and policies. Normalization or standardization is used to transform the values so that large values do not dominate the calculation of results. For instance, salary might dominate age in the results because of much wider range of salary than age. The transformation of attribute values in the range of [0,1] can be calculated from (x-min)/(max-min). When x is the attribute value, max is the maximum value of the attribute variable, and min is the minimum for that attribute variable.

- **Data discretization**
  Continuous attributes may need to be transformed into categorical data (discretization) or both continuous and discrete attribute values have to be transformed into binary (binarization) so that the analytical algorithm can be applied in the process.



Source: Data Preprocessing (Han, 2011)

References

Han, J., Kamber, M., and Pei, J. (2011). *Data mining: concepts and techniques.* (3rd edition). Morgan Kaufmann.

Tan, P., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley.

SequenceFile (n.d.) Retrieved from Hadoop Wiki,
http://wiki.apache.org/hadoop/SequenceFile