InfoLab

IC612: Data Warehousing and Data Mining (Lecture 1)

Min-Soo Kim

DGIST 대구경북과학기술원
Daegu Gyeongbuk Institute of Science & Technology

Data mining

- **Data mining : study of collecting, cleaning, processing, analyzing, and gaining useful insights from data**

- **Wide variation (in real applications) in terms of**
  - problem domains
  - applications
  - formulations
  - data representations

- **Deluge of data**
  - virtually all automated systems generate some form of data either for diagnostic or analysis purposes
  - the order of petabytes or exabytes

Examples of data

- **World Wide Web**
  - # documents on the indexed Web is on the order of billions (the invisible Web is much larger)
  - User accesses to such documents create Web access logs at servers and customer behavior profiles at commercial sites
  - Linked structure of the Web is referred to as the Web graph data

- **Financial interactions**
  - Most common transactions of everyday life, such as using ATM card or a credit card, can create data in an automated way
  - Such transactions can be mined for many useful insights such as fraud or other unusual activity

User interactions

- Many forms of user interactions create large volumes of data
- Use of a telephone typically creates a record at the telecommunication company with details about the duration and destination of the call
- Many phone companies routinely analyze such data to determine relevant patterns of behavior that can be used to make decisions about network capacity, promotions, pricing, or customer targeting

Sensor technologies and the Internet of Things

- A recent trend is the development of low-cost wearable sensors, smart phones, and other smart devices that can communicate with one another
- # of such devices exceeded # of people on the planet in 2008

Goals of data mining task

- **Extract concise and possibly actionable insights from the available data for application-specific goals**
  - raw data may be arbitrary, unstructured, or even in a format that is not immediately suitable for automated processing
  - pipeline of processing collects, cleans, and transforms the raw data into a standardized format
  - data may be stored in a commercial database system and finally processed for insights with the use of analytical methods
  - pipeline of processing is conceptually similar to that of an actual mining process from a mineral ore to the refined end product

Data mining is challenging

- **Wide disparity in the problems and data types**
  - e.g., a product recommendation problem is very different from an intrusion-detection application
    - in terms of the level of the input data format or the problem definition
  - e.g., a product recommendation problem in a multidimensional database is very different from a social recommendation problem
    - due to the differences in the underlying data type

- **Nevertheless, applications are often closely connected to one of four "superproblems"**
  - association pattern mining
  - clustering
  - classification (related with machine learning)
  - outlier detection

Data formats (or types)

- **Type**
  - quantitative (e.g., age)
  - categorical (e.g., ethnicity)
  - text
  - spatial
  - temporal
  - graph-oriented

- **Most common form of data is multidimensional**

- **Precise data type may affect the behavior of a particular algorithm significantly**

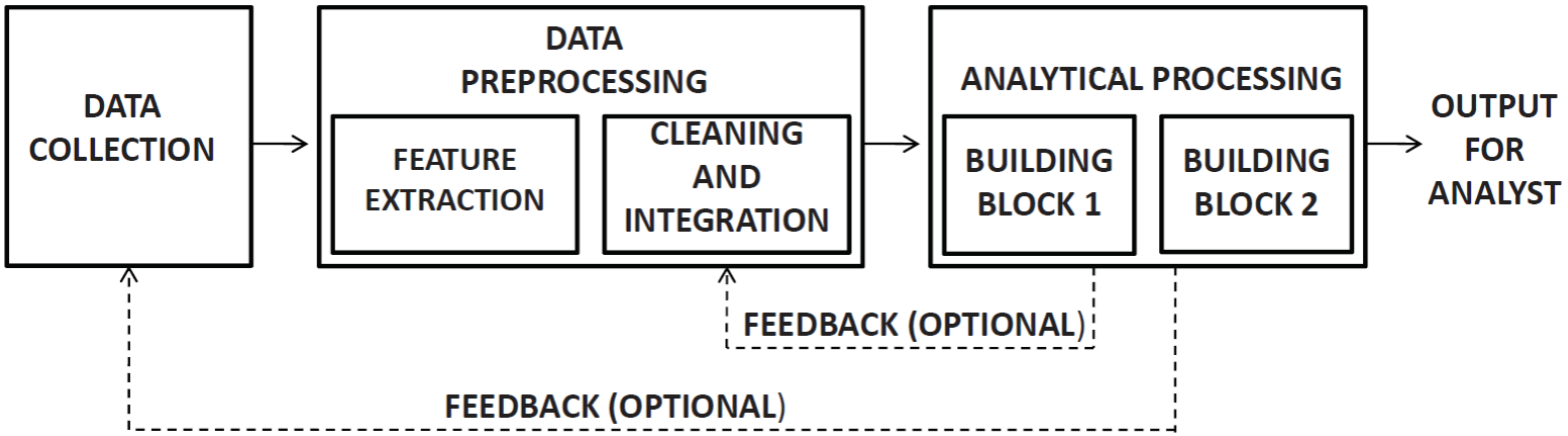
Increasing data volumes

- **Major challenge created in recent years**
- **e.g., Internet traffic generates large data streams**
  - cannot even be stored effectively
  - unless significant resources are spent on storage
- **If it is not possible to explicitly store the data, all the processing needs to be performed in real time**

Data Mining Process

1. Data collection

- sensor network, manual labor (user surveys), or software tools (e.g., Web crawler)
- collected data is stored in a database, or, a data warehouse



data processing pipeline

2. Feature extraction and data cleaning

- collected data is often not in a form that is suitable for processing
- it is essential to transform it into a format that is friendly to data mining algorithms
  - e.g, multidimensional, time-series, or semi-structured
- multidimensional format has different fields of the data for the different measured properties (features, attributes, or dimensions)
- feature extraction: extract relevant features for the mining process
- data cleaning: missing and erroneous parts of the data are either estimated or corrected
- the data is stored in a database again

3. Analytical processing and algorithms

- design effective analytical methods from the processed data
- many applications can be broken up into components that use different building blocks of four superproblems

Example: Web log mining

- **Scenario**
  - a retailer has Web logs corresponding to customer accesses to Web pages at his or her site
    - each Web page corresponds to a product
  - the retailer also stores demographic profiles for different customers
  - the retailer wants to make targeted product recommendations to customers

- **Data sources**
  - Web logs (at the site)
  - demographic profiles (collected during registration of the customer)

Unfortunately, two sources have different format

```
98.206.207.157 - - [31/Jul/2013:18:09:38 -0700] "GET /productA.htm
HTTP/1.1" 200 328177 "-" "Mozilla/5.0 (Mac OS X) AppleWebKit/536.26
(KHTML, like Gecko) Version/6.0 Mobile/10B329 Safari/8536.25"
"retailer.net"
```

- a customer at IP address 98.206.207.157 has accessed productA.htm
- customer from the IP address can be identified using the previous login information (by using cookies, or the IP address itself)
  - but, this may be a noisy process and may not always yield accurate results

Feature extraction

- raw log contains a lot of additional information that is not necessarily of any use to the retailer
- retailer creates one record for each customer (an attribute corresponds to the number of accesses to each product description)

Data cleaning

- missing entries from the demographic records need to be estimated

Data integration

- attributes are added to these records for the retailer's database containing demographic information
- this results in a single data set containing attributes for the customer demographics and customer accesses

Data mining

- determine similar groups of customers (clustering)
- make recommendations on the basis of the buying behavior of these similar groups
- the most frequent items accessed by the customers in that group are recommended (frequent pattern mining)

Basic Data Types

- **Non-dependency oriented data**
  - simple data types (e.g., multidimensional data, text data)
  - data records do not have any specified dependencies between either the data items or the attributes
  - e.g., demographic records contain age, gender, and ZIP code
- **Dependency oriented data**
  - implicit or explicit relationships may exist between data items
  - e.g., social network data contains a set of vertices (data items) that are connected together by a set of edges (relationships)
  - e.g., time series contains implicit dependencies (dependency between successive readings)
- **Dependency-oriented data is more challenging**

Non-dependency Oriented Data

- **Typically, a set of records**
  - also referred to as a data point, instance, example, transaction, entity, tuple, object, or feature-vector
- **Each record contains a set of fields**
  - also referred to as attributes, dimensions, or features

Name	Age	Gender	Race	ZIP Code
John S.	45	M	African American	05139
Manyona L.	31	F	Native American	10598
Sayani A.	11	F	East Indian	10547
Jack M.	56	M	Caucasian	10562
Wei L.	63	M	Asian	90210

Example of a demographic multidimensional data set

Definition 1.3.1 (Multidimensional Data) *A multidimensional data set $\mathcal{D}$ is a set of n records, $\overline{X_1} \ldots \overline{X_n}$, such that each record $\overline{X_i}$ contains a set of d features denoted by $(x_i^1 \ldots x_i^d)$.*

Quantitative Multidimensional Data

- e.g., age: numeric and quantitative data (natural ordering)
- data mining textbook assume a quantitative multidimensional representation (each field's value is quantitative)
- real applications usually have a mixture of different data types

Categorical and Mixed Attribute Data

- many real applications use categorical attributes of discrete unordered values
- e.g., gender, race, and ZIP code (no natural ordering)

Binary and Set Data

- binary data: special case of either numeric or categorical data
- binary data can be also used as a representation of set-wise data
  - each attribute is treated as a set element indicator
  - value 1 indicates that the element is included in the set

Text Data

- it can be viewed either as a string, or as multidimensional data
- text documents are rarely represented as strings
- in practice, vector-space representation is used
  - precise ordering of the words is lost
  - frequencies of the words in the document are used for analysis

- document-term matrix : n×d data matrix
  - n: # of documents
  - d: # of terms(words)

- data sparsity: most attributes take on zero values (only a few attributes have non-zero values)
  - a single document may contain only a relatively small number of words out of a dictionary of size 10^5
- data sparsity significantly impacts the data mining process

Dependency Oriented Data

- **Pre-existing dependencies**
  - the knowledge about pre-existing dependencies greatly changes the data mining process
  - data mining is all about finding relationships between data items
- **Implicit dependencies**
  - the dependencies between data items are not explicitly specified but are known to “typically” exist in that domain
  - e.g., time-series data, spatial data
- **Explicit dependencies**
  - typically, graph or network data
  - edges are used to specify explicit relationships

Time-series data

- **Generated by continuous measurement over time**
  - e.g., environmental sensor measures the temperature continuously
  - e.g., electrocardiogram (ECG) measures the parameters of a subject's heart rhythm
- **Implicit dependencies**
  - the adjacent values recorded by a temperature sensor will usually vary smoothly over time
  - some forms of sensor readings may show periodic patterns of the measured attribute over time
- **Attributes are classified into two types**
  - contextual attributes: sensor data (time stamp), spatial data (two contextual attributes, x and y)
  - behavioral attributes: values measured in a particular context

Discrete sequences and strings

- **Discrete sequences**
  - categorical analog of time-series data
  - i.e., behavioral attribute is a categorical value

Definition 1.3.3 (Multivariate Discrete Sequence Data) *A discrete sequence of length n and dimensionality d contains d discrete feature values at each of n different time stamps $t_1 \ldots t_n$. Each of the n components $\overline{Y_i}$ contains d discrete behavioral attributes $(y_i^1 \ldots y_i^d)$, collected at the ith time-stamp.*

  - e.g., sequence of 100 different Web accesses (n=100, d=2)

- **String: sequence data in the univariate scenario (d=1)**

- **Contextual attribute may be a position**
  - e.g., biological sequence data

Spatial data

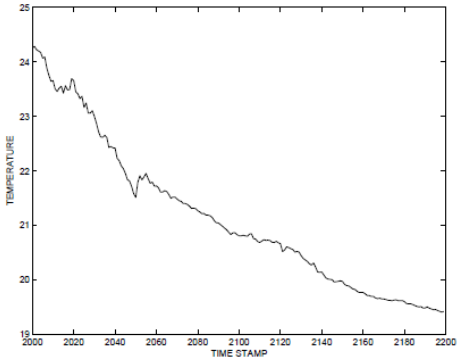
- **Many non-spatial attributes are measured at spatial locations**
  - e.g., sea-surface temperatures are often collected by meteorologists to forecast the occurrence of hurricanes
    - contextual attributes: spatial coordinates
    - behavioral attributes: temperature, pressure

Definition 1.3.4 (Spatial Data) *A d-dimensional spatial data record contains d behavioral attributes and one or more contextual attributes containing the spatial location. Therefore, a d-dimensional spatial data set is a set of d dimensional records $\overline{X_1} \ldots \overline{X_n}$, together with a set of n locations $L_1 \ldots L_n$, such that the record $\overline{X_i}$ is associated with the location L_i.*

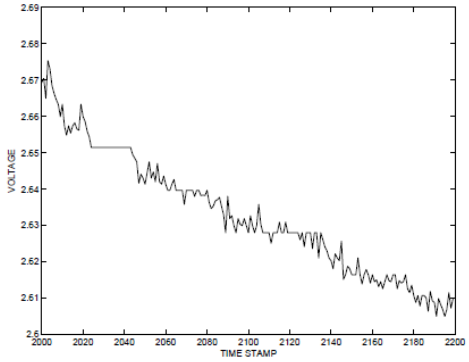
- **Spatial data mining is closely related to time-series data mining**
  - behavioral attributes in spatial applications are usually continuous
  - value continuity is observed across contiguous spatial locations

Spatiotemporal data

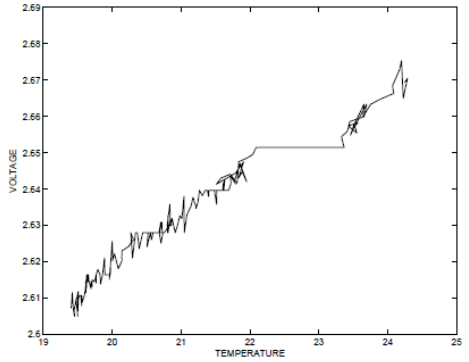
- **Containing both spatial and temporal attributes**
- **Two kinds of spatiotemporal data**
  - ➢ both spatial and temporal attributes are contextual
    - spatial and temporal dynamics of particular behavioral attributes are measured simultaneously
    - e.g., sea-surface temperature need to be measured over time
  - ➢ temporal attribute is contextual, but spatial attributes are behavioral
    - .e.g, trajectory data



(a) Temperature



(b) Voltage



(c) Temperature-voltage trajectory

Network and Graph Data

- **Data values correspond to nodes**
- **Relationships among data values correspond to edges**
  - directed : Web graph
  - undirected : friendships in the Facebook social network
- **Attributes may be associated with nodes**

Definition 1.3.5 (Network Data) *A network $G = (N, A)$ contains a set of nodes N and a set of edges A, where the edges in A represent the relationships between the nodes. In some cases, an attribute set $\overline{X_i}$ may be associated with node i, or an attribute set $\overline{Y_{ij}}$ may be associated with edge (i, j).*

- **e.g., specialized forms of social networks**
  - email or chat-messenger networks (edges may have content)
- **e.g., chemical compound databases**
  - nodes are the elements; the edges are the chemical bonds

Major Building Blocks: A Bird's Eye View

- **Data matrix D: n x d (multidimensional database)**
  - *n* records, *d* attributes
- **Data mining : all about finding summary relationships between the entries in the data matrix that are**
  - either unusually frequent
  - or unusually infrequent
- **Relationships**
  - between columns: the frequent or infrequent relationships between the values in a particular row are determined
    - e.g., data classification
  - between rows: determine subsets of rows, in which the values in the corresponding columns are related
    - e.g., data clustering, outlier analysis

Association pattern mining

- **Sparse binary databases**
  - data matrix contains only 0/1 entries
  - most entries take on the value of 0

Definition 1.4.1 (Frequent Pattern Mining) *Given a binary $n \times d$ data matrix D, determine all subsets of columns such that all the values in these columns take on the value of 1 for at least a fraction s of the rows in the matrix. The relative frequency of a pattern is referred to as its support. The fraction s is referred to as the minimum support.*

- **e.g., basket data**
  - if the columns of the data matrix D corresponding to Bread, Butter, and Milk take on the value of 1 together frequently in a customer transaction database
  - it implies that these items are often bought together

Data clustering

Definition 1.4.3 (Data Clustering) *Given a data matrix D (database $\mathcal{D}$), partition its rows (records) into sets $\mathcal{C}_1 \ldots \mathcal{C}_k$, such that the rows (records) in each cluster are "similar" to one another.*

- **Often defined as an optimization problem**
  - variables: cluster memberships of data points
  - objective function: maximizes intra-group similarity (mathematical quantification) in terms of these variables
- **Appropriate similarity function is very important**
  - computation of similarity depends on the underlying data type
- **e.g., customer segmentation, data summarization**

Outlier detection

- **Outlier: a data point that is significantly different from the remaining data**
  - referred to as abnormalities, discordants, deviants, or anomalies
- **Recognition of unusual characteristics of systems provides useful application-specific insights**

> **Definition 1.4.4 (Outlier Detection)** *Given a data matrix D, determine the rows of the data matrix that are very different from the remaining rows in the matrix.*

- **Related to the clustering problem by complementarity**
  - outliers: dissimilar data points from the main groups in the data
- **e.g., intrusion-detection, credit card fraud, interesting sensor events, medical diagnosis, earth science**

Data classification

- **Learning the relationships of between**
  - the special feature (called class label) and
  - the remaining features
- **Training data: data used to learn these relationships**
  - learned model may be used to determine the estimated class labels for test records
  - test record: record whose class label is unknown

Definition 1.4.5 (Data Classification) *Given an $n \times d$ training data matrix D (database $\mathcal{D}$), and a class label value in $\{1 \ldots k\}$ associated with each of the n rows in D (records in $\mathcal{D}$), create a training model $\mathcal{M}$, which can be used to predict the class label of a d-dimensional record $\overline{Y} \notin \mathcal{D}$.*

- **Supervised learning vs. unsupervised learning**

- **Related to association pattern mining**
  - association pattern mining is used to solve classification problem (rule-based classifiers)
  - entire training data (including class label) : n×(d+1) data matrix
  - frequent patterns containing the class label provide useful hints about the correlations of other features to the class label
- **Related to outlier detection**
  - outlier detection problem is unsupervised by default
  - many variations of the problem are either partially or fully supervised
    - some examples of outliers are available
    - rare class vs. normal class
- **e.g., target marketing, intrusion detection, supervised anomaly detection**

Impact of complex data types on problem definitions

- **Data type has an impact on the kinds of problems**

Problem	Time Series	Spatial	Sequence	Networks
Patterns	Motif Mining Periodic Pattern	Co-location Patterns	Sequential Patterns Periodic Sequence	Structural Patterns
	Trajectory Patterns			
Clustering	Shape Clusters	Spatial Clusters	Sequence Clusters	Community Detection
	Trajectory Clusters			
Outliers	Position Outlier Shape Outlier	Position Outlier Shape Outlier	Position Outlier Combination Outlier	Node Outlier Linkage Outlier Community Outliers
	Trajectory Outliers			
Classification	Position Classification Shape Classification	Position Classification Shape Classification	Position Classification Sequence Classification	Collective Classification Graph Classification
	Trajectory Classification			

Some examples of variation in problem definition with data type

Scalability Issues & Streaming Scenario

- **Two important scenarios for scalability**
  - data is stored on one or more machines, but it is too large to process efficiently (main memory or disk)
  - data is generated continuously over time in high volume, and it is not practical to store it entirely (data streams, online approach)

- **Major challenges in data stream processing**
  - one-pass constraint: algorithm needs to process the entire data set in one pass
    - raw item is discarded and is no longer available for processing
  - **concept drift**: in most applications, the data distribution changes over time
    - e.g., the pattern of sales in a given hour of a day may not be similar to that at another hour of the day

InfoLab

Thank you!

DGIST 대구경북과학기술원
Daegu Gyeongbuk Institute of Science & Technology