Unit 9
Mining Spatial, Multimedia, Text, and Web Data

Prepared By

Arjun Singh Saud, Asst. Prof. CDCSIT

Spatial Data Mining

- A spatial database stores a large amount of space-related data, such as maps, remote sensing data, etc.
- It is a database optimized for storing and querying data that represents objects defined in a geometric space.
- Most spatial databases allow the representation of simple geometric objects such as points, lines and polygons.
- Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. Such mining demands an integration of data mining with spatial database technologies.

Spatial Data Mining

- Spatial data mining can be used for understanding spatial data, discovering spatial relationships and relationships between spatial and non-spatial data, constructing spatial knowledge bases, reorganizing spatial databases, and optimizing spatial queries.
- It is expected to have wide applications in geographic information systems, geo-marketing, remote sensing, and many other areas where spatial data are used.
- A crucial challenge to spatial data mining is the exploration of *efficient* spatial data mining techniques due to the huge amount of spatial data and the complexity of spatial data types.

Spatial Data Cube

- We can integrate spatial data to construct a data warehouse that facilitates spatial data mining.
- A spatial data warehouse is a *subject-oriented, integrated, time-variant,* and *nonvolatile* collection of both spatial and non-spatial data in support of spatial data mining.
- The star schema model is a good choice for modeling spatial data warehouses because it provides a concise and organized warehouse structure and facilitates OLAP operations.

Spatial Data Cube

- However, in a spatial warehouse, both dimensions and measures may contain spatial components. There are three types of *dimensions* in a spatial data cube:
- A **nonspatial dimension** contains only non-spatial data. Non-spatial dimensions temperature and precipitation can be constructed for the warehouse.
- A **spatial-to-nonspatial dimension** is a dimension whose primitive-level data are spatial but whose generalization, starting at a certain high level, becomes nonspatial. For example, the spatial dimension city relays geographic data for the country map.

Spatial Data Cube

- **Spatial-to-spatial dimension** is a dimension whose primitive level and all of its high level generalized data are spatial.
- We also distinguish two types of measures in a spatial data cube:
- A **numerical measure** contains only numerical data. One measure in a spatial data warehouse could be the monthly revenue of a region, so that a roll-up may compute the total revenue by year, by county, and so on.
- A **spatial measure** contains a collection of pointers to spatial objects. In a generalization (or roll-up) with the same range of temperature and precipitation will be grouped into the same cell, and the measure so formed contains a collection of pointers to those regions.

Multimedia Data Mining

- **Multimedia database** is the collection of interrelated multimedia data that includes text, graphics (sketches, drawings), images, animations, video, audio, hypertext, etc.
- Multimedia data mining is an interdisciplinary field that integrates image processing and understanding, computer vision, data mining, and pattern recognition.
- Issues in multimedia data mining include content-based retrieval and similarity search

Multimedia Data Mining

- For similarity searching in multimedia data, we consider two main families of multimedia indexing and retrieval systems:
- **Description-based retrieval systems**: which build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation.
- **Content-based retrieval systems**: It support retrieval based on the image content, such as color histogram, texture, pattern, image topology, and the shape of objects and their layouts and locations within the image.

Mining Associations in Multimedia Data

- Association rules involving multimedia objects can be mined in image and video databases. At least three categories can be observed:
- **Associations between image content and non-image content features:** A rule like "*If at least 50% of the upper part of the picture is blue, then it is likely to represent sky*" belongs to this category since it links the image content to the keyword sky.
- **Associations among image contents that are not related to spatial relationships:** A rule like "*If a picture contains two blue squares, then it is likely to contain one red circle as well*" belongs to this category since the associations are all regarding image contents.

Mining Associations in Multimedia Data

- **Associations among image contents related to spatial relationships:** A rule like "*If a red triangle is between two yellow squares, then it is likely a big oval-shaped object is underneath*" belongs to this category since it associates objects in the image with spatial relationships.

Text Mining

- Text mining is the process of deriving high-quality information from text databases.
- Text databases consist of huge collection of documents. They collect these information from several sources such as news articles, books, digital libraries, e-mail messages, web pages, etc.
- Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output.

Text Mining

- Typical text mining tasks include text classification, text clustering, sentiment analysis, document summarization, information extraction, etc.
- The automatic extraction of structured data such as entities, entities relationships, and attributes describing entities from an unstructured source is called **information extraction (IE)**.
- In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP).

Text Mining

- Information extraction can be applied to a wide range of textual sources: from emails and Web pages, reports, legal documents and scientific papers.
- Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written.
- NLP involves Natural Language Understanding (NLU) and Natural Language Generation(NLG).
- NLU is the NLP task of understanding meaning of the natural language sentences whereas NLG is the task of generating natural language sentences from machine representation.

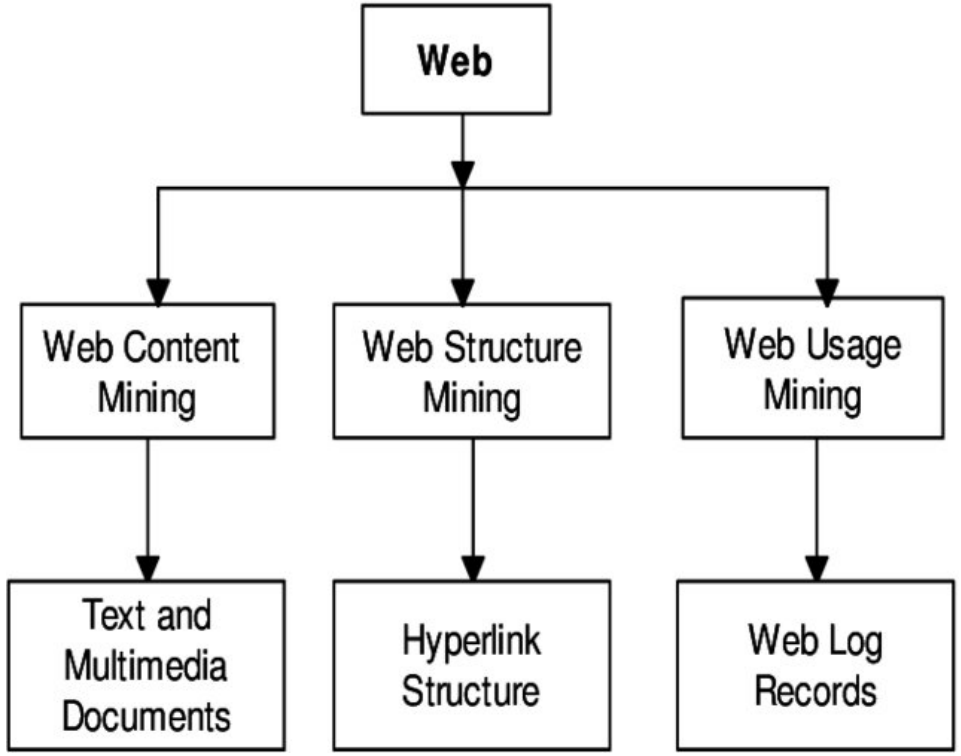
Web Mining

- Web mining is the application of data mining techniques to discover patterns from the World Wide Web.
- It uses automated methods to extract both structured and unstructured data from web pages, server logs and link structures.
- There are three main sub-categories of web mining: *Web Content Mining, Web Structure Mining, and Web Usage Mining*.

Web Mining

- Web content mining extracts information from within a page. It mines text and multimedia data present in web pages.
- Web structure mining discovers the structure of the hyperlinks between documents, categorizing sets of web pages and measuring the similarity and relationship between different sites.
- Web usage mining finds patterns of usage of web pages. It does so by mining web log records.

Web Mining


Web
Web Content Mining
Web Structure Mining
Web Usage Mining
Text and Multimedia Documents
Hyperlink Structure
Web Log Records