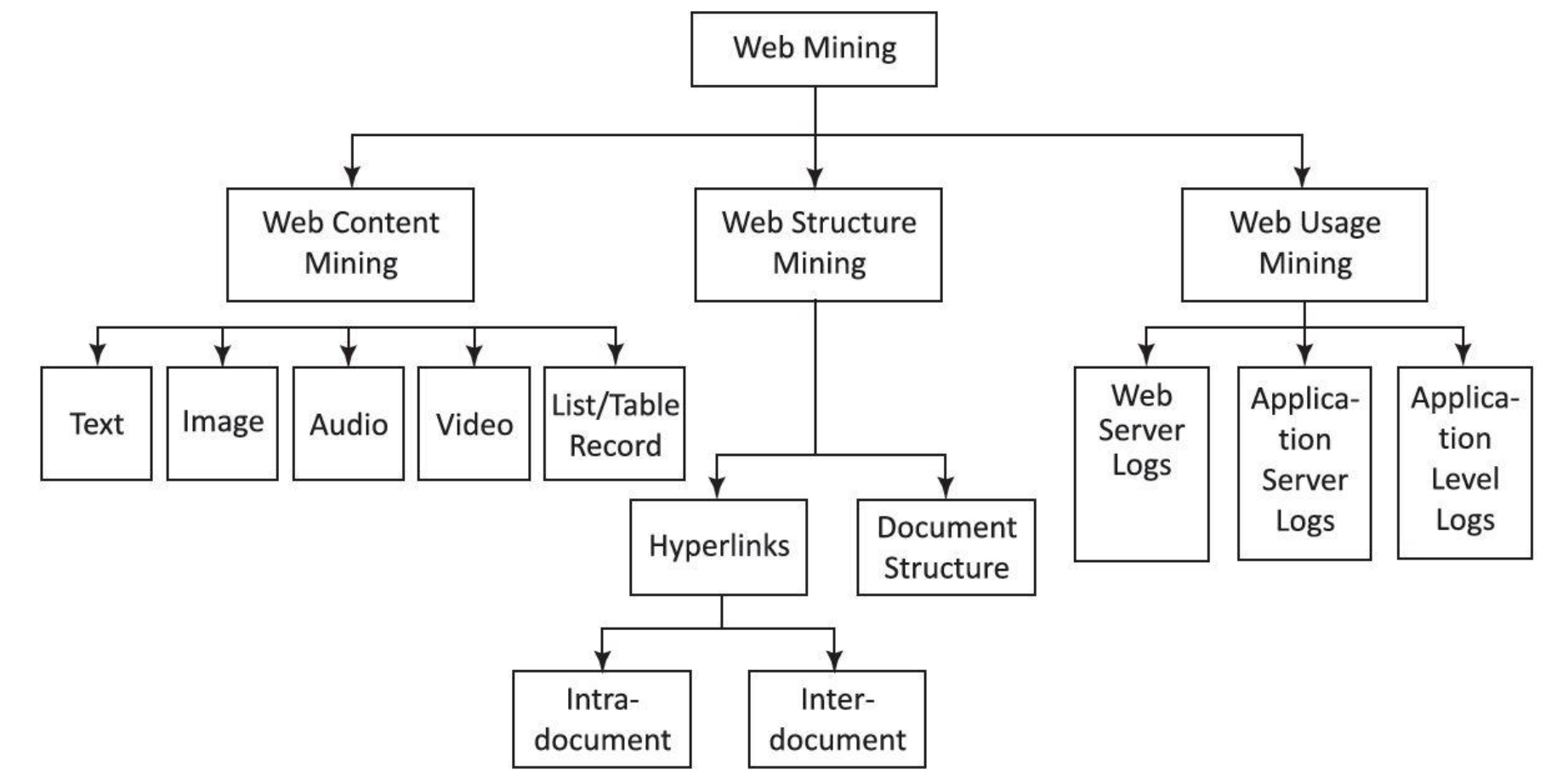
Lesson 6

Web Content Mining

Web Mining Components



Web Content Mining

- Is the process of information or resource discovery from the content of web documents
- Can be (i) direct mining of the contents of documents or (ii) mining through search engines, fast comparatively

Content mining

- Relates to text mining
- Much of the web content comprises texts.
- Web data are mainly semi-structured and/or unstructured, while data mining is structured and the text is unstructured.

Applications

1. Classifying the web documents into categories
2. Identifying topics of web documents
3. Finding similar web pages across the different web servers
4. Applications related to relevance:

Applications related to relevance Examples

(a) Recommendations – List of top "n" relevant documents

(b) Filters – Show/Hide documents based on some criterion

(c) Queries – Enhance standard query relevance with user, role, and/or task-based relevance

Web Content Mining Techniques

- Pre-processing of contents
- Clustering
- Classifying
- Identifying the associations
- Topic identification, tracking and drift analysis

Preprocessing

1. Extraction of text from HTML
2. Data cleaning by filling up the missing values and smoothing the noisy data
3. Tokenizing: Generates the tokens of words from the cleaned up text
4. Stemming: Reduce the words to their roots; . “closed” and “closing” Root: “close”. [Porter algorithm can be used]

.... Preprocessing

5. Removing the stop words: a, an, the, such as, to, in, for …
6. Calculate the multiple occurrence of a significant term (t) in a collection is called collection frequency (CF_t)
- 7. Calculate per Document Term Frequencies (DF_t).[Example 9.1]

.... Preprocessing

8. Bag of words: Represent Web document by the words it contains (and their occurrences).

- Example 9.6 for learning CFs and DFs computations

Mining Tasks for Web Content Analytics

- 1. Classification – (i) Identifies the class or category a new web documents belongs to from the set of predefined classes or categories, (ii) Categories in the form of a term vector, and
- (iii) Employs algorithms using term vector to categorize the new data

… Mining Tasks for Web Content Analytics

2. Clustering (i) Groups the web documents with similar features (ii) Uses no pre-defined perception of what the groups should be, (iii) Measures most common similarity using the dot product between two web document vectors

… Mining Tasks for Web Content Analytics

3. Identifying the association between web documents – Association rules help to identify correlation between web pages that occur mostly together

… Mining Tasks for Web Content Analytics

4. Categorizing the web pages into distinct topics
5. Adding a new document to a collection library
6. Concept hierarchy creation –for capturing the general relationship among web documents

… Mining Tasks for Web Content Analytics

7. Finding Document relevance
8. Query-based relevance─ used in information retrieval tools
9. User-based relevance ─ user profile based push notification services.
10. Role/task-based relevance

Summary

We learnt:

- Web Content Mining Methods
- Clustering
- Classifying into categories

2. Identifying topics of web documents
3. Finding similar web pages
4. Applications related to relevance

End of Lesson 6 on

Web Content Mining