Web Mining :

Web Content Mining

- Is the process of extracting useful information from the contents of web documents
- Content may consist of text, images, audio, video or structured records, such as lists and tables
- Can be (i) direct mining of the contents of documents or (ii) mining through search engines, fast comparatively

Web 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

Web Content Mining 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

Web Content Mining Techniques

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

Web Content Mining Techniques - 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

Web Content Mining Techniques - Preprocessing

4. Stemming: Reduce the words to their roots; . “closed” and “closing” Root: “close”. [Porter algorithm can be used]
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 (CFt)

Web Content Mining Techniques - Preprocessing

4. Stemming: Reduce the words to their roots; . “closed” and “closing” Root: “close”. [Porter algorithm can be used]
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 (CFt)

Mining Tasks for Web Content Analytics

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

Mining Tasks for Web Content Analytics

2. **Clustering**
   - Groups the web documents with similar features
   - Uses no pre-defined perception of what the groups should be,
   - 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.
4. Categorizing the web pages into distinct topics
5. Adding a new document to a collection library
6. Finding Document relevance

Mining Tasks for Web Content Analytics

7. Concept hierarchy creation –for capturing the general relationship among web documents
8. Query-based relevance– used in information retrieval tools
9. User-based relevance – user profile based push notification services.
10. Role/task-based relevance

We learnt

- Web Content Mining Methods
- Clustering
- Classifying into categories
- Identifying topics of web documents
- Finding similar web pages
- Applications related to relevance