**Unit 9**

**Mining 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. Due to increase in the amount of information, the text databases are growing rapidly. In many of the text databases, the data is semi-structured. For example, a document may contain a few structured fields, such as title, author, publishing-date, etc. But along with the structure data, the document also contains unstructured text components, such as abstract and contents. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users require tools to compare the documents and rank their importance and relevance. Therefore, text mining has become popular and an essential theme in data mining.

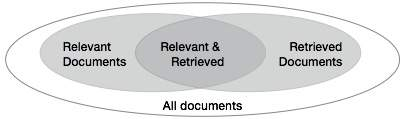
**Information Retrieval**

Information retrieval deals with the retrieval of information from a large number of text-based documents. Examples of information retrieval system include − Online Library catalogue system, Online Document Management Systems, Web Search Systems etc.

The main problem in an information retrieval system is to locate relevant documents in a document collection based on a user's query. This kind of user's query consists of some keywords describing an information need. In such search problems, the user takes an initiative to pull relevant information out from a collection. This is appropriate when the user has ad-hoc information need, i.e., a short-term need. But if the user has a long-term information need, then the retrieval system can also take an initiative to push any newly arrived information item to the user. This kind of access to information is called Information Filtering. And the corresponding systems are known as Filtering Systems or Recommender Systems.

**Basic Measures for Text Retrieval**

We need to check the accuracy of a system when it retrieves a number of documents on the basis of user's input. Let the set of documents relevant to a query be denoted as {Relevant} and the set of retrieved document as {Retrieved}. The set of documents that are relevant and retrieved can be denoted as {Relevant} ∩ {Retrieved}. This can be shown in the form of a Venn diagram as follows



There are three fundamental measures for assessing the quality of text retrieval

* Precision
* Recall
* F-score

**Precision**

Precision is the percentage of retrieved documents that are in fact relevant to the query. Precision can be defined as

Precision= |{Relevant} ∩ {Retrieved}| / |{Retrieved}|

**Recall**

Recall is the percentage of documents that are relevant to the query and were in fact retrieved. Recall is defined as

Recall = |{Relevant} ∩ {Retrieved}| / |{Relevant}|

**F-score**

F-score is the commonly used trade-off. The information retrieval system often needs to trade-off for precision or vice versa. F-score is defined as harmonic mean of recall or precision as follows

F-score = recall x precision / (recall + precision) / 2

**Mining World Wide Web**

The World Wide Web contains huge amounts of information that provides a rich source for data mining. The web poses great challenges for resource and knowledge discovery based on the following observations

* **The web is too huge** − The size of the web is very huge and rapidly increasing. This seems that the web is too huge for data warehousing and data mining.
* **Complexity of Web pages** − The web pages do not have unifying structure. They are very complex as compared to traditional text document. There are huge amount of documents in digital library of web. These libraries are not arranged according to any particular sorted order.
* **Web is dynamic information source** − The information on the web is rapidly updated. The data such as news, stock markets, weather, sports, shopping, etc., are regularly updated.
* **Diversity of user communities** − The user community on the web is rapidly expanding. These users have different backgrounds, interests, and usage purposes. There are more than 100 million workstations that are connected to the Internet and still rapidly increasing.
* **Relevancy of Information** − It is considered that a particular person is generally interested in only small portion of the web, while the rest of the portion of the web contains the information that is not relevant to the user and may swamp desired results.

**Mining Web page layout structure**

The basic structure of the web page is based on the Document Object Model (DOM). The DOM structure refers to a tree like structure where the HTML tag in the page corresponds to a node in the DOM tree. We can segment the web page by using predefined tags in HTML. Thus the DOM structure can be used to facilitate information extraction. The HTML syntax is flexible therefore, the web pages does not follow the W3C specifications. Not following the specifications of W3C may cause error in DOM tree structure.

The DOM structure was initially introduced for presentation in the browser and not for description of semantic structure of the web page. The DOM structure cannot correctly identify the semantic relationship between the different parts of a web page.

**Multimedia Data Mining**

A multimedia database system stores and manages a large collection of *multimedia data*, such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text markups, and linkages. Typical multimedia database systems include NASA’s EOS (Earth Observation System), various kinds of image and audio-video databases, and Internet databases.

Normally study of multimedia data mining focuses on image 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.

**Spatial Data Mining**

A spatial database stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, and VLSI chip layout data. Spatial databases have many features distinguishing them from relational databases. They carry topological and/or distance information. 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. It can be used for understanding spatial data, discovering spatial relationships and relationships between spatial and non-spatial data. It is expected to have wide applications in geographic information systems, geomarketing, remote sensing, image database exploration, medical imaging, navigation, traffic control, environmental studies, 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 and spatial access methods.