UNIT V PREDICTIVE MODELING OF BIG DATA AND TRENDS INDATA MINING 9

Statistics and Data Analysis – EDA – Small and Big Data –Logistic Regression Model – Ordinary Regression Model-Mining complex data objects – Spatial databases – Temporal databases – Multimedia databases – Time series and sequence data – Text mining – Web mining – Applications in Data mining

Text mining

- Rapid increment in computerized or digital information has prompted an enormous volume of information and data. A substantial portion of the available information is stored in Text databases, which consist of large collections of documents from various sources. Text databases are rapidly growing due to the increasing amount of information available in electronic form. In excess of **80%** of the present information is in the form of unstructured or semi-organized data. Traditional information retrieval techniques become inadequate for the increasingly vast amount of text data. Thus, text mining has become an increasingly popular and essential part of Data Mining. The discovery of proper patterns and analyzing the text document from the huge volume of data is a major issue in real-world

- Text mining is a process of extracting useful information and nontrivial patterns from a large volume of text databases. There exist various strategies and devices to mine the text and find important data for the prediction and decision-making process. The selection of the right and accurate text mining procedure helps to enhance the speed and the time complexity also. This article briefly discusses and analyzes text mining and its applications in diverse fields.

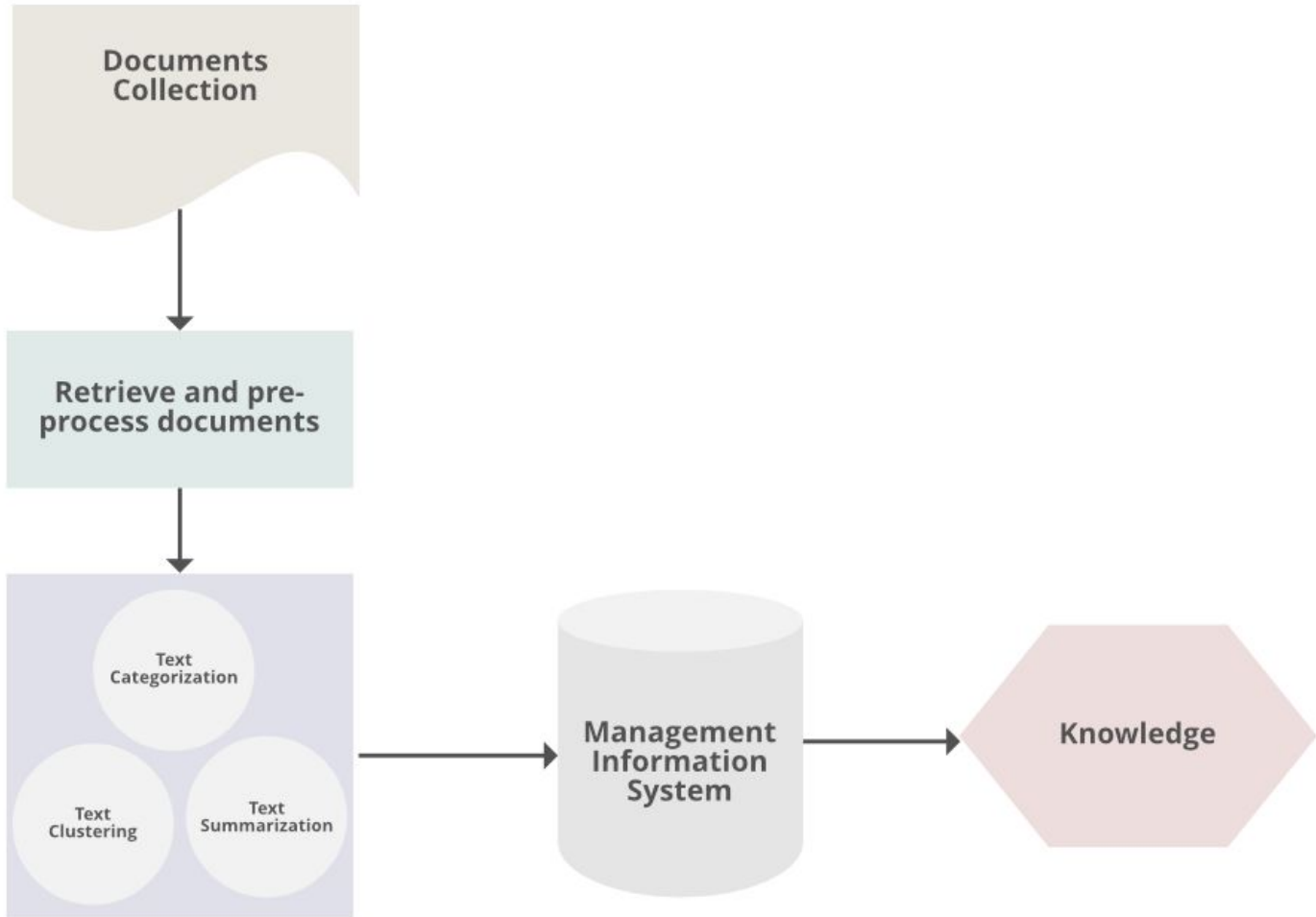
- As we discussed above, the size of information is expanding at exponential rates. Today all institutes, companies, different organizations, and business ventures are stored their information electronically. A huge collection of data available on the internet and store in digital libraries, database repositories, and other textual data like websites, blogs, social media networks, and e-mails. It is a difficult task to determine appropriate patterns and trends to extract knowledge from this large volume of data. Text mining is a part of Data mining to extract valuable text information from a text database repository. Text mining is a multi-disciplinary field based on data recovery, Data mining, AI, statistics, Machine learning, and computational linguistics.

The conventional process of text mining as follows:

- Gathering unstructured information from various sources accessible in various document organizations, for example, plain text, web pages, PDF records, etc.
- Pre-processing and data cleansing tasks are performed to distinguish and eliminate inconsistency from the data. The data cleansing process makes sure to capture the genuine text, and it is performed to eliminate stop words stemming (the process of identifying the root of a certain word and indexing the data.
- Processing and controlling tasks are applied to review and further clean the data set.
- Pattern analysis is implemented in Management Information System.
- Information processed in the above steps is utilized to extract important and applicable data for a powerful and convenient decision-making process and trend analysis.

Procedures of analyzing Text Mining:

- **Text Summarization:** To extract its partial content reflection its whole content automatically.
- **Text Categorization:** To assign a category to the text among categories predefined by users.
- **Text Clustering:** To segment texts into several clusters, depending on the substantial relevance.


Documents Collection
Retrieve and pre-process documents
Text Categorization
Text Clustering
Text Summarization
Management Information System
Knowledge

TextMining Techniques:

- **Information Extraction**: It is a process of extract meaningful words from documents.
- **Information Retrieval**: It is a process of extracting relevant and associated patterns according to a given set of words or text documents.
- **Natural Language Processing:** It concerns the automatic processing and analysis of unstructured text information.
- **Clustering:** It is an unsupervised learning process that grouping of text according to their similar characteristics.
- **Text Summarization:** To extract its partial content reflection it's whole content automatically.

Overview of Text Mining Techniques

	. Text Preprocessing phase	Tokenization	How can transform a text into words or text format?	Transferring strings into a single textual token.	White space separation.
	Compound word identification	How can I identify words that have a joint meaning?	Identifying words with a joint meaning that gets lost word	n-grms	
	Normalization and noise reduction	How can I cope with too many variables in my Document-Term-Matrix ?	Reducing the dimensionality of Document-Term-Matrix	Stemming, Lemmatization, Deletion of stop words. infrequent term.	
	How can I identify words with a special meaning or grammatical function?	Tagging of words	Named-entity recognition, Part-of-speech tagging		
Linguistic analysis	**2.**Content Analysis	Dictionary-based	How can I identify how latent sociological or psychological traits and states are reflected in natural language?	Measuring contextual, psychological, linguistic, or semantic concepts and constructs	pre-defined dictionaries Customized dictionaries
	Algorithmic techniques	How can I assign texts to predefined classes?	Classifying of textual entities into predefined categories	Supervised learning techniques such as binary or multi-class classifiers	
	How can I group together similar documents?	Clustering of textual entities into formerly undefined and unknown	Unsupervised learning techniques such as LDA, k-means or non-negative		

Application Area of Text Mining

- **1. Digital Library**

- Various text mining strategies and tools are being used to get the pattern and trends from journal and proceedings which is stored in text database repositories. These resources of information help in the field of the research area. Libraries are a good resource of text data in digital form. It gives a novel technique for getting useful data in such a way that makes it conceivable to access millions of records online. A green-stone international digital library that supports numerous languages and multilingual interfaces give a springy method for extracting reports that handle various formats, i.e. Microsoft Word, PDF, postscript, HTML, scripting languages, and email. It additionally supports the extraction of audiovisual and image formats along with text documents. Text Mining processes perform different activities like document collection, determination, enhancement, removing data, and handling substances, and Producing summarization. There are different types of digital libraries text mining tools namely: GATE, Net Owl, and Aylien which used for text mining.

- **2. Academic and Research Field**
- In the education field, different text mining tools and strategies are utilized to examine the instructive patterns in a specific region/research field. The main purpose of text mining utilization in the research field help to discover and arrange research papers and relevant material of various fields on one platform. For this, we use k-Means clustering and different strategies help to distinguish the properties of significant data. Also, student performance in various subjects can be accessed, and how various qualities impact the selection of subjects evaluated by this mining.

- **3. Life Science**
- Life science and health care industries are producing an enormous volume of textual and mathematical data regarding patient records, sicknesses, medicines, symptoms, and treatments of diseases, etc. It is a major issue to filter data and relevant text to make decisions from a biological data repository. The clinical records contain variable data which is unpredictable, lengthy. Text mining can help to manage such kinds of data. Text mining use in biomarkers disclosure, pharmacy industry, a clinical trade analysis examination, clinical study, patent competitive intelligence also.

- **4. Social-Media**
- Text mining is accessible for dissecting analyzing web-based media applications to monitor and investigate online content like plain text from internet news, web journals, email, blogs, etc. Text mining devices help to distinguish and investigate the number of posts, likes, and followers on the web-based media network. This kind of analysis shows individuals' responses to various posts, news and how it spread around. It shows the behavior of people who belong to a specific age group and variation in like, views about the same post.

- **5. Business Intelligence**
- Text mining plays an important role in business intelligence that help different organization and enterprises to analyze their customers and competitors to make better decisions. It gives an accurate understanding of business and gives data on how to improve consumer satisfaction and gain competitive benefits. The text mining devices like IBM text analytics.
- GATE help to make the decision about the organization that produces alerts about good and bad performance, a market changeover that helps to take necessary actions. This mining can be used in the telecom sector, commerce, customer chain management system.

Issues in Text Mining

- Numerous issues happen during the text mining process:
- **1.** The efficiency and effectiveness of decision-making.
- **2.** The uncertain problem can come at an intermediate stage of text mining. In the pre-processing stage, different rules and guidelines are characterized to normalize the text that makes the text mining process efficient. Prior to applying pattern analysis on the document, there is a need to change over unstructured data into a moderate structure.
- **3.** Sometimes original message or meaning can be changed due to alteration.

Issues in Text Mining

- **4.** Another issue in text mining is many algorithms and techniques support multi-language text. It may create ambiguity in text meaning. This problem can lead to false-positive results.
- **5.** The utilization of synonym, polysemy, and antonyms in the document text makes issues for the text mining tools that take both in a similar setting. It is difficult to categorize such kinds of text/ words.