|  |
| --- |
| Web Mining |
| **Web Mining** is the process of [Data Mining](https://www.geeksforgeeks.org/data-mining/) techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.  **Applications of Web Mining:**   1. Web mining helps to improve the power of web search engines by classifying web documents and identifying web pages. 2. It is used for Web Searching e.g., Google, Yahoo, etc, and Vertical Searching e.g., FatLens, Become, etc. 3. Web mining is used to predict user behavior. 4. Web mining is very useful for a particular Website and e-service e.g., landing page optimization.   **Process of Web Mining:**  https://media.geeksforgeeks.org/wp-content/uploads/20221129200848/Screenshot20221129200746.png  ***Web Mining Process***  Web mining can be broadly divided into three different types of techniques of mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. These are explained as following below.  https://media.geeksforgeeks.org/wp-content/uploads/20221129202522/Screenshot20221129202430.png  ***Categories of Web Mining*** |

1. **Web Content Mining:** Web content mining is the application of extracting useful information from the content of the web documents. Web content consist of several types of data – text, image, audio, video etc. Content data is the group of facts that a web page is designed. It can provide effective and interesting patterns about user needs. Text documents are related to text mining, machine learning and natural language processing. This mining is also known as text mining. This type of mining performs scanning and mining of the text, images and groups of web pages according to the content of the input.
2. **Web Structure Mining:** Web structure mining is the application of discovering structure information from the web. The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Structure mining basically shows the structured summary of a particular website. It identifies relationship between web pages linked by information or direct link connection. To determine the connection between two commercial websites, Web structure mining can be very useful.
3. **Web Usage Mining:** Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets. And these patterns enable you to understand the user behaviors or something like that. In web usage mining, user access data on the web and collect data in form of logs. So, Web usage mining is also called log mining.

**Comparison Between Data mining and Web mining:**

|  |  |  |
| --- | --- | --- |
| Points | Data Mining | Web Mining |
| Definition | Data Mining is the process that attempts to discover pattern and hidden knowledge in large data sets in any system. | Web Mining is the process of data mining techniques to automatically discover and extract information from web documents. |
| Application | Data Mining is very useful for web page analysis. | Web Mining is very useful for a particular website and e-service. |
| Target Users | Data scientist and data engineers. | Data scientists along with data analysts. |
| Access | Data Mining access data privately. | Web Mining access data publicly. |
| Structure | In Data Mining get the information from explicit structure. | In Web Mining get the information from structured, unstructured and semi-structured web pages. |

|  |  |  |
| --- | --- | --- |
| Problem Type | Clustering, classification, regression, prediction, optimization and control. | Web content mining, Web structure mining. |
| Tools | It includes tools like machine learning algorithms. | Special tools for web mining are Scrapy, PageRank and Apache logs. |
| Skills | It includes approaches for data cleansing, machine learning algorithms. Statistics and probability. | It includes application level knowledge, data engineering with mathematical modules like statistics and probability. |

Text Mining

What is Text Mining?

**Text mining** is the process of deriving high-quality information from text. It is also referred to as text data mining in some circles and is similar in some ways to text analytics. Text mining involves the discovery of new, previously unknown information using a computer to automatically extract data from different written resources.

Text mining is widely adopted in knowledge-driven organizations. It involves examining large collections of documents, often for research purposes. Text mining is the tool that identifies patterns, uncovers relationships, and makes assertions based on patterns it discovers buried deep within layers of textual [big data](https://www.tibco.com/reference-center/what-is-big-data).

Upon extraction, the information is converted into a structured format that can either be analyzed further or sorted into clustered HTML tables, mind maps, and charts for presentation. For analysis, it can be integrated into data warehouses, databases, or [business](https://www.tibco.com/reference-center/what-is-business-intelligence) [intelligence](https://www.tibco.com/reference-center/what-is-business-intelligence) dashboards.

Types of Analytics Run on Data Extracted Through Text Mining

Data extracted through text mining can be valuable for running various types of analyses:

* [Prescriptive analytics](https://www.tibco.com/reference-center/what-is-prescriptive-analytics)
* [Predictive analytics](https://www.tibco.com/reference-center/what-is-predictive-analytics)
* Descriptive analytics
* Lexical analytics- to study frequency distribution of words
* Tagging and annotation
* Pattern recognition
* Links and associations
* [Visualization](https://www.tibco.com/reference-center/guide-to-data-visualization)

The goal is, essentially, to turn text into data for analysis, by means of the application of natural language processing (NLP), various types of algorithms, and analytical methods. The interpretation of the gathered information is an important part of this process.

The Abilities of Natural Language Processing Systems Today

Natural language understanding is the first step in natural language processing that helps machines read text or speech. In a way, it simulates the human ability to understand an actual language such as English or French or Mandarin.

Natural language processing combines both natural language understanding, and natural language generation. This, in turn, simulates the human ability to create natural language text. Examples of this include the ability to collate or summarize information, or participate in a conversation or dialogue.

Natural language processing has developed in leaps and bounds over the last decade, and will continue to evolve and grow. Mainstream products like Alexa, Siri and Google’s voice search use natural language processing to understand and respond to user questions and requests.

Natural language processing systems are a form of automation that has become indispensable in analyzing text-derived data today. Their abilities are manifold:

* They can run analysis on literally unlimited amounts of textual data consistently, tirelessly, and in an unbiased manner.
* They have the ability to understand sophisticated and complex concepts.
* They can detect ambiguities of language, extract relevant facts, and identify relationships.
* They can provide summaries.

The Importance of Text Mining Today

Businesses across the world today generate vast amounts of data literally every minute, simply through having an online presence and operating in the online space. This data comes in from multiple sources and is stored in data warehouses and on cloud platforms. Traditional methods and tools sometimes fall short in analyzing such gigantic data that grows exponentially by the minute, presenting a major challenge for companies.

Another major reason behind the adoption of text mining is the growing cut-throat competition in the business sphere, leading organizations to seek more value-added solutions to stay ahead of the competition.

Such is the background against which text mining applications, tools, and techniques have come into popular use; they offer a way to use all that data that has been collected, and then can help organizations use it to grow.

How Text Mining and Natural Language Processing Work Together

An example of the relevance of text mining can be seen in the context of [machine learning](https://www.tibco.com/reference-center/what-is-machine-learning). Machine learning is a widely used artificial intelligence technology which imbues systems with the ability to automatically learn from experience without having to be programmed. This technology can rival or even surpass humans in solving complex problems with great accuracy.

However, for machine learning to deliver the best outcome, it needs well-curated input to train upon. In situations where most of the available data input is in the form of unstructured text, this is difficult. An example of this is electronic health records, clinical research data sets, or full-text scientific literature.

Natural language processing is a great tool to extract structured and cleaned-up data for these advanced predictive models used in machine learning to base its training on. This reduces the need for manual annotation of such training data and saves costs.

In addition, text mining allows the analysis of large collections of literature and data to identify potential issues early on in the pipeline. This helps companies make the best use of research and development resources and avoid potentially known failures in functions like later stage drug trials.

The Multidisciplinary Nature of Text Mining

Text mining is, for all intents and purposes, a multidisciplinary field. It incorporates and integrates the tools of data mining, information retrieval, machine learning, computational linguistics and even statistics. Text mining is concerned with natural language texts stored in semi-structured or unstructured formats.

The Text Mining Process: Steps Pre-Processing Operations

* Collating unstructured text data from multiple data sources: plain text, word files, PDF files, web pages, blogs, emails, or social media.
* Hygiene and cleansing the data with the help of text mining tools and applications to detect and remove anomalies or redundancies. This part of the process is to extract

and keep only the pertinent information from the data and help identify the roots of specific words.

* Convert the above into structured formats suitable for analysis.

Analysis

* Analyze the patterns within the data via the Management Information System (MIS).
* Extract the valuable insights and move the information into a secure database to drive trend analysis.
* Use the insights for decision-making.

Text Mining Techniques

There are five commonly used and effective techniques used in text mining. Information Extraction

This technique refers to the process of extracting meaningful information from swathes of textual data, whether present in the form of unstructured or even semi-structured text formats. It focuses on identifying and extracting entities, their attributes, and their relationships. The extracted information is stored in a database for easy future access and retrieval. Precision and recall processes are used to evaluate the relevancy and efficacy of these outcomes.

Information Retrieval

The technique of information retrieval is more specific and pertains to the extraction of relevant and associated patterns based on a particular set of words or phrases. Information retrieval systems make use of algorithms to track and follow user behavior and gather relevant data. An example of this is the much-used Google search engine.

Categorization

Categorization is a form of supervised learning, in which normal language texts are sorted into a predefined bunch of topics based on their content. The system gathers text documents and analyzes them to find out the relevant topics or correct indexing for every document.

The co-referencing process is used as a part of natural language processing to extract not just meanings but actual synonyms and abbreviations from text data sets. At present, this process is an automated one with widespread applications, from personalized commercials to spam filtering. It is extensively used in categorizing web pages under hierarchical definitions. Its uses are many.

Clustering

As the name suggests, this text mining technique seeks to identify and locate intrinsic structures within a text database and organize them into subgroups (or, ‘clusters’) for further analysis. This is a vital and standard text mining technique.

The biggest challenge in the cluster-forming process is to create meaningful clusters from unclassified, unlabeled textual data with no prior lead information. Cluster analysis is used in data distribution. It also acts as a pre-processing step for other text mining algorithms and techniques that can be applied downstream on detected clusters.

Summarization

Text summarization is the process of auto-generating a compressed version of a specific text, that contains information that may be useful to the end user. The goal of the summarization technique is to look through multiple sources of textual data to put together summaries of texts containing a sizable amount of information in a concise format. The overall meaning and intent of original documents is kept essentially unchanged. Text summarization integrates the various methods that use text categorization, such as decision trees, neural networks, swarm intelligence or regression models.

Applications and Benefits of Text Mining

Text mining tools and techniques are being deployed in a variety of industries and areas today; academia, healthcare, organizations, social media platforms, to name a few.

Text Mining for Risk Analysis, Assessment and Risk Management

Often organizations launch new products and services without conducting a sufficient amount of risk analysis. Improper risk analysis puts the organization behind on key information and trends, contributing to them missing out on opportunities for growth or for connecting better with their target audience.

Text mining technologies are the drivers for risk management software that can be integrated into a business’s operations. Such text mining technologies can collate information from a multitude of text data sources and create links between relevant insights.

Adoption of text mining technologies enables organizations to remain up-to-date on current market trends, get the right information at the right moment, and identify potential risks in a timely fashion. This means that organizations can mitigate risks and be agile in making business decisions.

Fraud Detection with Text Mining and Text Analytics

This application of text analytics and the text mining tools within remain a mainstay of insurance and finance companies. Such organizations gather a majority of their data in the text format. Structuring this data and subjecting it to text analyses using text mining tools

and techniques helps such companies detect and prevent fraud. Text mining also helps companies process warranty or insurance claims faster.

Text Mining for Superior Business Intelligence

A lot of organizations across various industries are increasingly leveraging text mining techniques for superior business intelligence insights. Text mining techniques yield deep insights into customer/ buyer behavior and market trends.

Text mining also helps organizations complete a strength, weakness, opportunity and threat analysis of their own business as well as their competition and gain an edge in the market.

Text mining tools and techniques also yield insights on how marketing strategies and campaigns are performing, what customers are looking for, their buying preferences and trends and the shifting market.

Improving Customer Care Services Using Text Mining Techniques

Text mining techniques are increasingly being adopted in the field of customer care services to enhance their overall customer experience. Natural language processing is a frontrunner in this area. Companies are investing in [text analytics](https://www.tibco.com/reference-center/what-is-text-analytics) software that patrols text data from customer surveys, feedback forms, voice calls, emails and chats.

The goal of text mining and analytics is to reduce the response time to a call or query and deliver faster, more efficient turnaround in addressing customer complaints. This has the benefit of customer longevity, less churn, and faster resolution of complaints.

Social Media Analysis Using Text Mining Tools

With the text-heavy nature of social media, text mining tools shine in terms of analyzing the number of posts, likes, comments, referrals and follower trends of your brand. In fact, there are several text mining tools designed just for analyzing how your brand performs on various social media platforms.

Text mining on social media is also an invaluable tool to understand reactions and behavior patterns of a large number of people interacting with your brand and online content, often in real-time.

This enables text mining and text analysis to help organizations capitalize on the hot trends of the moment that are captivating their target audience. What is going viral? What content is engaging users? How can a business use this information to increase their market share and grow sales?

Disadvantages of Text Mining

While the text mining, or web mining technology itself does not create problems, its application on datasets of a private nature can lead to ethical concerns. This includes using text mining on personal medical records, or to create group profiles. Privacy issues are a highly criticized ethical issue linked with the unscrupulous use of text mining.

Also, companies may conduct text mining for a purpose, but could use the data for another, unstated or undisclosed purpose. In a world where personal data is a big commodity, such misuse presents a major threat to an individual’s data privacy.

Despite this, text mining remains a highly powerful tool that many organizations can use to their advantage for everything from streamlining day-to-day operations to making strategic business decisions.