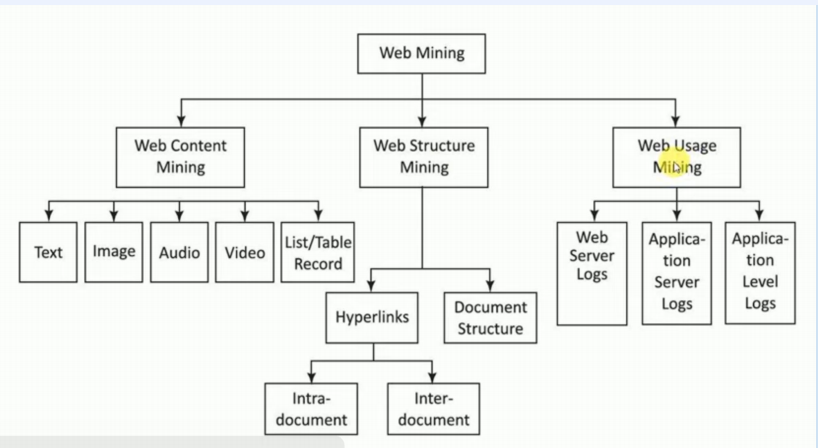
**UNIT1**

Web mining

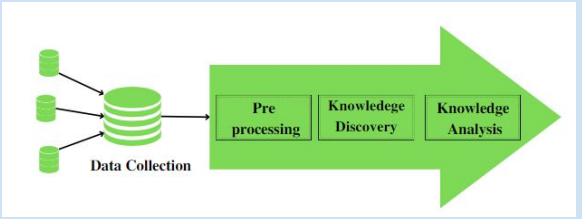
**Web mining refers to the process of extracting useful information or knowledge from web data sources. It involves various techniques and methodologies to gather data from the World Wide Web, analyze it, and derive meaningful insights.**



Web mining can be broadly categorized into three types:

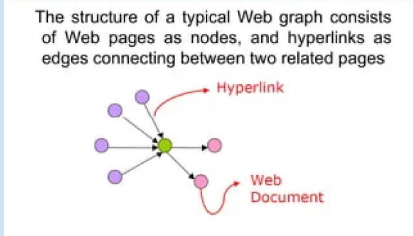
**1)Web Content Mining:**

This involves extracting useful information from web pages, including text, images, videos, and other multimedia content. Techniques such as web scraping, text mining, and image processing are often used to extract and analyze content.



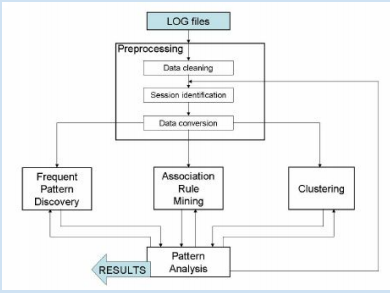
**2)Web Structure Mining:**

This focuses on analyzing the structure of the web, including links between web pages, to discover patterns, relationships, and properties of the web graph. Algorithms like PageRank and HITS (Hyperlink-Induced Topic Search) are used to analyze link structures and determine the importance or relevance of web pages.



**3)Web Usage Mining:**

This involves analyzing user interaction data collected from web servers, such as web server logs, cookies, and clickstream data, to understand user behavior, preferences, and trends. Techniques like association rule mining and sequential pattern mining are commonly used to extract useful patterns from usage data.



**Web Usage Mining Applications**

**1. Web Server logs:**

Collected by the web server and typically include IP address, page reference and access time

**2. Application Server Logs:**

Application servers typically maintain their own logging and these logs can be helpful in troubleshooting problems with services

**3. Application Level Logs:**

Recording events usually by application software in a certain scope in order to provide an audit trail that can be used to understand the activity of the system and to diagnose problems.

**Process of Web Mining**

**The process of web mining typically involves the following steps -**

**●Data collection -**

Web data is collected from various sources, including web pages, databases, and APIs.

**●Data pre-processing -**

The collected data is pre-processed to remove irrelevant information, such as advertisements and duplicate content.

**●Data integration -**

The pre-processed data is integrated and transformed into a structured format for analysis.

**●Pattern discovery -**

Web mining techniques are applied to identify patterns, trends, and relationships.

**●Evaluation -**

The discovered patterns are evaluated to determine their significance and usefulness.

**●Visualization -**

The analysis results are visualized through graphs, charts, and other visualizations.

**Applications of Web Mining**

**Web mining is the process of discovering patterns, structures, and relationships in web data. It involves using data mining techniques to analyze web data and extract valuable insights.**

**Sure, here are simplified answers to the applications of web mining:**

**1.Personalized Marketing:** Using web data to understand customer preferences and create tailored marketing campaigns.

**2.E-commerce Optimization:** Analyzing customer behavior on online stores to enhance user experience and boost sales through personalized product recommendations.

**3.Search Engine Optimization (SEO):** Using web mining to improve website visibility in search engine results and increase organic traffic.

**4. Fraud Detection:** Identifying and preventing online fraud through analysis of web data and user behavior.

**5. Sentiment Analysis:** Understanding customer opinions and emotions from social media data to make informed business decisions**.**

**6.Web Content Optimization:** Analyzing web content to improve its relevance and search engine rankings.

**7.Customer Service Improvement:** Enhancing the quality of customer service by analyzing interactions on websites and social media.

**8.Healthcare Insights:** Extracting valuable information from health-related websites to improve healthcare quality and support medical research.

**Data Mining**

**Data mining is the process of discovering patterns, trends, and insights from large datasets using various techniques and methodologies. It involves extracting knowledge and valuable information from raw data to support decision-making, improve processes, and gain competitive advantages.**

**Data Collection:**

Gather relevant data from multiple sources, including databases, data warehouses, spreadsheets, text files, and online repositories.

**Data Preprocessing:**

**Data Cleaning:**

Remove noise and irrelevant data, handle missing values, and resolve inconsistencies in the dataset.

**Data Integration:**

Combine data from different sources into a unified dataset, ensuring consistency and compatibility.

**Data Transformation:**

Convert data into a suitable format for analysis, such as normalization to scale numerical attributes or encoding categorical variables.

**Data Reduction:**

Reduce the dimensionality of the dataset by selecting relevant features or using techniques like principal component analysis (PCA) to retain essential information while reducing computational complexity.

**Data Exploration:**

Explore the dataset visually and statistically to gain insights into its distribution, relationships between variables, and potential patterns.

**Model Building:**

Select appropriate data mining algorithms and techniques based on the problem domain, data characteristics, and objectives.

**Model Evaluation:**

Assess the performance of the data mining models using evaluation metrics appropriate for the specific task, such as accuracy, precision, recall, F1-score, or area under the ROC curve (AUC).

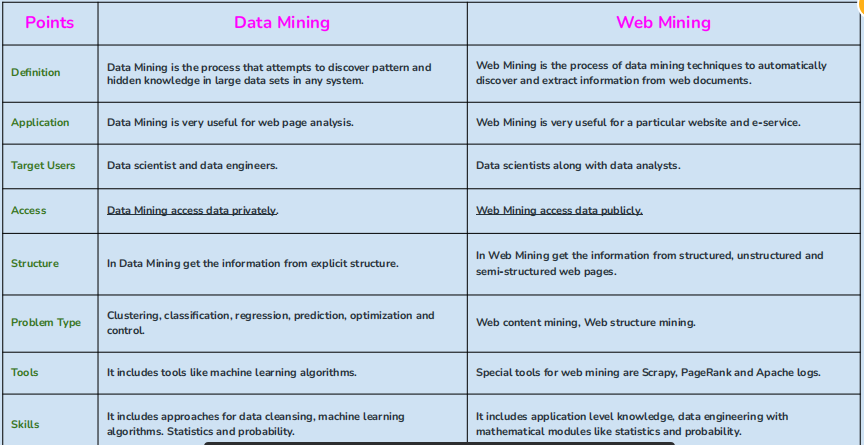
**Interpretation and Visualization:**

Interpret the patterns and insights discovered by the data mining models in the context of the problem domain.

**Deployment and Application:**

Deploy the data mining models in real-world applications to support decision-making, automate processes, or generate actionable insights.

**Comparison Between Data mining and Web mining**



SEQUENTIAL PATTERN MINING

**Data Mining:** the goal is to discover or extract useful knowledge from

data.

• Many types of data can be analyzed: graphs, relational databases,

time series, sequences, etc.

**What is a discrete sequence?**

A sequence is an ordered list of symbols.

• It is a popular data mining task, introduced in 1994 by Agrawal & Srikant.

• The goal is to find all subsequences that appear frequently in a set of discrete sequences.

• For example:

- find sequences of items purchased by many customers over time,

- find sequences of locations frequently visited by tourists in a city,

- Find sequences of words that appear frequently in a text.

**UNIT2**

**Information Retrieval**

**Information Retrieval (IR) is a software program or system designed to manage, store, retrieve, and evaluate information from document repositories**, **especially textual data.**

It is widely used in various domains, including libraries, professional research, and everyday web searches.

Here are some key points about Information Retrieval:

**Definition:** IR deals with organizing, storing, retrieving, and evaluating information, primarily textual, from large collections stored on computers.

**- Usage:** Users interact with IR systems by entering queries, which the system processes to find relevant documents.

**- Popularity:** Information Retrieval is widely used by millions of people daily, especially through web search engines.

**- Dominant Form:** It is considered the primary method for accessing information.

**- Functionality:** IR systems help users find relevant documents but do not provide explicit answers. They indicate the existence and location of potential information sources.

**- Supportive Functions:** IR systems also assist in browsing, filtering, and processing retrieved documents.

- Scale: These systems operate over vast collections of documents distributed across millions of computers.

**- Spam Filtering:** Email programs often use IR techniques, like spam filters, to classify and organize incoming emails.

**- Features**: IR systems can represent, store, organize, and access information items using keywords, which users input to search for relevant information.

**What is an IR Model?**

**An IR (Information Retrieval) model is a framework or system used to retrieve relevant information from a large collection of data, typically in the form of documents or web pages, based on a user's query.** It involves techniques like indexing, ranking, and matching to efficiently find and present the most relevant results to the user's search query.

1. Acquisition
2. Representation
3. File Organization
4. Query

**Relevance Feedback**

**The idea of relevance feedback (RF) is to involve the user in the retrieval process so as to improve the final result set.**

The basic procedure is :

1. The user issues a (short, simple) query.

2. The system returns an initial set of retrieval results.

3. The user marks some returned documents as relevant or irrelevant.

4. The system computes a better representation of the information needed based on the user feedback.

5. The system displays a revised set of retrieval result

**Inverted Index**

**An inverted index is an index data structure storing a mapping from content, such as words or numbers, to its locations in a document or a set of documents**

**Example: Consider the following documents.**

Document 1: The quick brown fox jumped over the lazy dog.

Document 2: The lazy dog slept in the sun.

Document 1: The, quick, brown, fox, jumped, over, the lazy, dog.

Document 2: The, lazy, dog, slept, in, the, sun

**There are two types of inverted indexes:**

**Record-Level Inverted Index:** Record Level Inverted Index contains a list of references to documents for each word.

**Word-Level Inverted Index:** Word Level Inverted Index additionally contains the positions of each word within a document.

**Inverted Index & Its Compression**

**An inverted index is a data structure used in information retrieval systems to quickly find which documents contain specific words or terms.** It maps each term to the list of documents that contain that term.

**Compression in the context of inverted indexes refers to reducing the size of the index to save storage space and improve search performance.** Techniques such as variable byte encoding, delta encoding, and dictionary compression can be used to compress inverted indexes.

**LATENT SEMANTIC INDEXING**

**Latent Semantic Indexing is also known as Latent Semantic Analysis. Latent Semantic Indexing is a method which we use for expanding the correctness of information retrieval.**

In simple words, to improve the understanding of the information, it finds the hidden associations between words.

This took an important step forward in text comprehension as it was responsible for the contextual nature of the language.

**Benefits of Using LSI Keywords**

1.LSI keywords help your website rank better for related searches by showing search engines that your content is relevant to a wider range of topics.

2.Using LSI keywords can boost sales by helping your business rank higher in search results with less competition, as seen with Amazon's success in using LSI versions of primary keywords.

3.LSI keywords improve the readability and natural flow of your content, making it more enjoyable for visitors and distinguishing it from keyword-stuffed content that may turn readers away.

**WEB SPAMMING**

Any deliberate human action that is meant to trigger an unjustifiably favorable relevance or importance for some web page, considering the page’s true value. (from Web Spam Taxonomy, Stanford)

Ex. “I received email offers to rank first in Google search results but have turned them down because the techniques used are spam.”

Types of Web Spam

There are essentially two types of spamming: boosting and hiding

**Boosting**

This is when one takes an action intended to (falsely?) increase or boost the value of a page.

**Term Spamming:**

This would be those seeking to manipulate through elements such as the page TITLE (title spam), Meta Description or Meta Keywords (meta spam)

**URL Spamming** is another area they’ve been known to also look at.

**Link Spamming** is another well-known one that also includes anchor text spamming.

**Hiding Techniques** This set of techniques is when one is using not generally noticeable methods of getting a page to rank higher. Or more appropriately, the hiding of boosting techniques. These are certainly more problematic and search engines tend to treat them as more insidious than the boosting ones.

**Content hiding:** These are techniques where terms and links are hidden when the browser renders a page. The more common approaches are using color schemes that render the elements in question effectively invisible.

**Cloaking:** This is when one identifies a search engine crawler and seeks to show a different version of the page to the spider than it would for the average user. This, one assumes, cuts down on the changes of being reported by users or competitors that might otherwise see the spammy page.

**Redirection:** The page is automatically redirected by the browser in the same manner so that the page gets indexed by the engine, but the user will never actually see it. Essentially acting as a proxy/doorway to game the engine, and misdirect the users

**UNIT3**

**OPINION MINING AND WEB USAGE MINING**

**Opinion mining, also known as sentiment analysis, is a technique that uses natural language processing to automatically extract sentiments (positive, negative, neutral) from text data.** It helps businesses understand customer opinions and emotions expressed in reviews, social media, surveys, etc., enabling them to make data-driven decisions to improve products and services. Web usage mining, on the other hand, focuses on analyzing user interactions with websites to gain insights into user behavior, preferences, and trends, aiding in website optimization and personalized marketing strategies. Both techniques leverage technology to handle large volumes of data efficiently, providing valuable insights for businesses.

**Opinion Mining Techniques & Types**

**Opinion mining, also known as sentiment analysis, involves analyzing text to determine the sentiment or opinion expressed within it.**

**There are various techniques and types of opinion mining:**

● Fine-grained sentiment analysis

● Emotion detection

● Aspect-based sentiment analysis

● Multilingual sentiment analysis

**Top Opinion Mining Applications in Business**

**Opinion mining offers a window into the thoughts and feelings of the public, allowing businesses to improve the customer experience, perform competitive research, and understand opinions.**

**Some of the most popular opinion mining sentiment analysis applications are:**

● Social media analysis

● Brand awareness

● Customer feedback

● Customer service

● Market research

● Evaluating marketing campaigns

**SENTIMENT CLASSIFICATION**

Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

**Sentiment analysis is the process of detecting positive or negative sentiment in text. It’s often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers.**

**Types of Sentiment Analysis**

**Sentiment analysis focuses on the polarity of a text (positive, negative, neutral) but it also goes beyond polarity to detect specific feelings and emotions (angry, happy, sad, etc), urgency (urgent, not urgent) and even intentions (interested v. not interested).**

Graded Sentiment Analysis If polarity precision is important to your business, you might consider expanding your polarity categories to include different levels of positive and negative:

● Very positive

● Positive

● Neutral

●Negative

● Very negative

This is usually referred to as graded or fine-grained sentiment analysis, and could be used to interpret 5-star ratings in a review, for example:

● Very Positive = 5 stars

● Very Negative = 1 star

**Emotion Detection:**

**Emotion detection in sentiment analysis goes beyond just determining positive or negative sentiment. It identifies specific emotions like happiness, frustration, anger, and sadness. Methods include lexicons (lists of words associated with emotions) or advanced machine learning algorithms.** However, lexicons can be limited as people express emotions differently, and certain words may convey multiple emotions depending on context.

**Aspect-based Sentiment Analysis:**

Aspect-based sentiment analysis focuses on identifying sentiments related to specific aspects or features in text. For example, in a product review mentioning "short battery life," an aspect-based classifier would recognize the negative sentiment about the battery life of the product.

**Multilingual Sentiment Analysis:**

Analyzing sentiment in multiple languages is complex and requires preprocessing and resources like sentiment lexicons or translated corpora. Tools for language detection and custom sentiment analysis models can be used for accurate multilingual sentiment analysis.

**Why Is Sentiment Analysis Important?**

**Since humans express their thoughts and feelings more openly than ever before, sentiment analysis is fast becoming an essential tool to monitor and understand sentiment in all types of data.** Automatically analyzing customer feedback, such as opinions in survey responses and social media conversations, allows brands to learn what makes customers happy or frustrated, so that they can tailor products and services to meet their customers’ needs

**Sentiment analysis is crucial for several reasons:**

1.Understanding Customer Sentiment

2.Improving Products and Services

3.Enhancing Customer Experience

4.Managing Brand Reputation

5.Making Data-Driven Decisions

6. Competitive Analysis

**Feature based Opinion Mining and Summarization**

**Feature-based opinion mining and summarization involve analyzing text to identify specific features or aspects of a product, service, or topic, and then summarizing opinions or sentiments related to those features. This process helps to understand what people like or dislike about certain aspects of a subject.**

For example, in a product review, feature-based opinion mining would identify features like design, performance, and usability. It would then analyze the text to determine opinions or sentiments expressed about each of these features, such as whether users found the design attractive, the performance satisfactory, or the usability intuitive.

Summarization, on the other hand, involves condensing this information into a concise summary that captures the overall sentiment or opinion about the product or topic based on its various features.

**In simple terms, feature-based opinion mining and summarization help break down opinions into specific aspects and then summarize those opinions for easier understanding.**

**CLASSIFICATION OF OPINION SUMMARIZATION TECHNIQUES**

Opinion summarization techniques can be divided into two main categories: **aspect-based and non-aspect based.**

**1. Aspect-based Opinion Summarization:**

- Focuses on specific features or aspects of a product or topic.

- Provides information about different features instead of a general summary.

- Matches what customers often look for in reviews.

- Examples of aspect-based techniques include contrastive, abstractive, multi-lingual, and entity-based summaries.

**2. Non-aspect based Opinion Summarization:**

- Creates a generalized summary without considering specific features.

- Not limited to any particular format or structure.

- Offers a broader overview of opinions without focusing on particular aspects.

**3.feature-based summarization system**

A feature-based summarization system is a type of text summarization system that focuses on extracting and summarizing information based on specific features or aspects of the text.

**COMPARATIVE SENTENCE AND RELATION MINING**

**Comparative sentence and relation mining involves identifying sentences in text that compare two sets of entities based on certain features.** For instance, in the sentence "Canon’s optics are better than those of Sony and Nikon," the comparative relation is expressed as "better" regarding the feature "optics," comparing Canon with Sony and Nikon.

**The task of comparative sentence mining is to:**

1. Identify comparative sentences in evaluative texts like reviews, forum posts, and news articles.

2. Extract comparative relations from these sentences, highlighting the comparison between entities based on specific features.

This process is valuable for various applications, such as helping product manufacturers understand customer perceptions of their products compared to competitors.

**OPINION SPAM**

**Opinion spam refers to misleading activities that promote or damage the reputation of products or services through fake reviews.** Online reviews are crucial for customers to make informed decisions, but opinion spam undermines their trust by providing biased or false information. Spammers exploit review platforms for personal gain, leading to a surge in fake reviews and deceptive practices.

**Social Network & Link Analysis**

**Link Analysis**

**Link analysis is a method used in network theory to assess connections between nodes. These nodes can be people, organizations, transactions, etc. It helps visualize data for better understanding, used in various fields like SEO, intelligence, security analysis, and research.** For example, in SEO, it's used to analyze website links and determine search engine visibility. In networking, it's used to analyze data flow, find bottlenecks, and optimize networks.

**Link analysis has three primary purposes:**

● Find matches for known patterns of interests between linked objects.

● Find anomalies by detecting violated known patterns.

● Find new patterns of interest (for example, in social networking and marketing and business intelligence)

**Scrapy :**

**Scrapy is a Python open-source web crawling framework used for large-scale web scraping. It is a web crawler used for both web scraping and web crawling. It gives you all the tools you need to efficiently extract data from websites, process them as you want, and store them in your preferred structure and format.**

Scrapy is fast because it can make requests in parallel, unlike traditional tools, saving memory and CPU resources.

Beautiful Soup, on the other hand, is a parsing library that helps extract data from HTML and XML documents. It is not a full-fledged web scraping framework like Scrapy but is useful for parsing and navigating through the structure of web pages to extract specific information.

**In summary:**

- Scrapy: Web scraping framework, allows for parallel requests, crawling, downloading, and saving content.

- Beautiful Soup: Parsing library, helps extract data from HTML and XML documents, not a full web scraping framework.

**Social Network Analysis**

**Social network analysis (SNA) is a method of studying social structures using networks and graph theory.**

SNA tools help analyze relationships among people in groups, useful for understanding social structures and work patterns.

Data from various sources like surveys, emails, and blogs are collected, analyzed to find relationships, and used to uncover new insights.

Organizational network analysis focuses on information flow within a group, showing informal networks within organizations.

Value network analysis looks at exchanged deliverables among roles, often across different organizations.

Social influence network analysis studies social media to find influential people, associations, or trends.

**Co-Citation and Bibliographic Coupling**

**WHAT IS CITATION?**

**A citation is a way of acknowledging and giving credit to the sources you used in your work.** It helps readers find the original sources and supports your arguments by providing evidence. Proper citations also prevent plagiarism and legal issues.

**WHAT IS CO-CITATION?**

**Co-citation is when two documents are often cited together by other documents.** This shows they are related in meaning. It's like saying if many people mention two books in the same conversation, those books are probably about similar topics.

**WHAT IS BIBLIOGRAPHIC COUPLING?**

**Bibliographic coupling is a way to measure similarity between documents based on the citations they share. If two documents reference the same third document in their bibliographies, they are considered bibliographically coupled, indicating they likely cover related topics.** The more citations they share, the stronger their coupling.

Bibliographic coupling and co-citation are two notions used to describe mutual relations in a citation network.

**UNIT4**

**PageRank**

**PageRank is a method used by search engines to rank web pages based on their importance, determined by the number and quality of links pointing to them.** It takes into account the structure of the web, where pages with more backlinks from important pages are considered more valuable. This helps search engines provide more relevant results to users.

**HITS**

**The HITS (Hyperlink-Induced Topic Search) Algorithm rates web pages by looking at their link structures. It was developed by Jon Kleinberg.**

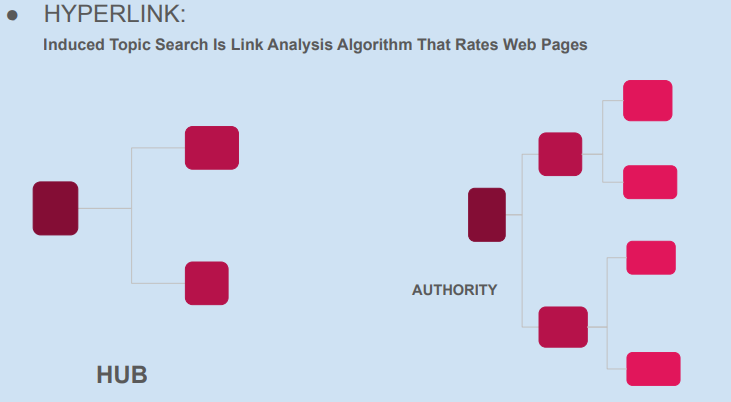
The algorithm uses two concepts:

**hubs and authorities.**

- Hubs are pages that point to many other pages.

- Authorities are pages that are linked to by many hubs.

When you search for something on a search engine, the relevant web pages are called "Roots" or potential authorities. Pages that are not directly relevant but link to these relevant pages are called hubs. So, an authority is a page that many hubs link to, while a hub is a page that links to many authorities.



**What is a web crawler?**

**A web crawler is a type of bot that scans and downloads content from the internet to create an index of web pages.** Search engines use web crawlers to gather information about different websites, which helps them provide relevant search results to users.

**Implementation Issues caused by web crawler:**

**Web crawlers could accidentally flood websites with requests to avoid this inefficiency. Web crawlers use politeness policies**. To implement politeness policy web crawler takes help of two parameters:

**1. Freshness:** As the content on webpages is constantly updated and modified web crawler needs to keep revisiting pages. For that, freshness uses HTTP protocol as HTTP has a special request type called HEAD which returns the information about the last updated date of a webpage by which crawler can decide the freshness of a webpage.

**2. Age:** An age of a webpage is T days after it has been last crawled. On average web page updates follow Poisson distribution and the older a page gets the more costs to crawl the web page so Age is more important factor for crawler than freshness.

**Applications**: This kind of web crawler is used to acquire the important parameters of the web like:

1. What are the frequently visited websites?

2. What are the websites that are important in the network as a whole?

3. Useful Information on social networks: Facebook, Twitter… etc.

4. Who is the most popular person in a group of people?

5. Who is the most important software engineer in a company

**Crawler Ethics and Conflicts**

The issue of ethics in data crawling and scraping is indeed complex and multifaceted.

**1.Subjectivity of Ethics:** Determining what constitutes ethical data crawling practices can vary depending on perspectives and contexts.

**2. Legal Requirements:**In certain situations, seeking ethical approval or adhering to legal regulations is necessary for data crawling activities.

**3.Invisibility to Users:** Many web users are unaware that their browsing data is being collected by crawlers and other automated programs.

**4.Purpose and Intent:T**he motivations behind data crawling activities play a significant role in determining their ethicality.

**5.Long-term Consequences:** Even if certain crawling activities don't cause immediate harm, they may have long-term repercussions, such as data vulnerabilities, privacy breaches, or distortions in online ecosystems.

**6. Professional Standards:** Establishing ethical guidelines and codes of conduct for data crawling practitioners can help mitigate unethical practices.

**7.Technological Controls:**While programmers can implement controls like rate limiting to manage crawler behavior, some factors like network bandwidth limitations remain beyond their control.

**Data Crawling Ethics may cause the following Issues**

**Data crawling, also known as web scraping, involves extracting information from websites using automated tools. While it can be a valuable technique for gathering data, it can also raise ethical concerns and lead to various issues.**

**1.Denial of Services:** When a crawler accesses a website too frequently or aggressively, it can overload the server and cause denial-of-service (DoS) issues. This can disrupt the website's normal functioning and affect its availability to legitimate users.

**2.Cost:** Data crawling can incur costs for both the crawler and the target website. Crawlers may consume significant bandwidth and resources, leading to increased hosting costs for the website owner.

**3.Privacy:** Crawlers may inadvertently collect and expose sensitive information, such as personally identifiable information (PII), without proper consent or authorization. This can violate individuals' privacy rights and raise legal and ethical concerns, especially if the collected data is misused or shared without consent.

**4. Copyright:** Crawlers may scrape and republish copyrighted content without permission, leading to copyright infringement issues. This can harm content creators' rights and may result in legal actions, such as takedown requests or lawsuits, against the crawler and its operators.

**The Issues and Challenges with the Web Crawlers**

**Web crawlers are vital tools for collecting data in today's big data era, particularly for businesses undergoing digital transformation or small and medium-sized enterprises aiming for growth.** However, several challenges and issues arise during the crawling process:

**1.Dynamic Content:**Web pages are frequently updated, necessitating regular crawling to ensure access to the latest information. Setting appropriate time intervals for crawling helps avoid redundant data.

**2.Anti-Crawling Measures:** Some websites implement anti-crawling mechanisms to thwart unauthorized data scraping. These measures can block crawling tools, making it difficult to extract data from such sites.

**3. Data Integrity:** Garbled data can be a common issue post-crawling, requiring additional effort to clean and preprocess the extracted information for accurate analysis.

**4.Data Analysis:** While successful crawling retrieves data, the subsequent data analysis phase can be labor-intensive and time-consuming, especially when dealing with large datasets.