UNIT V

ADVANCED CONCEPTS

BASIC CONCEPTS IN MINING DATA STREAMS

INTRODUCTION TO STREAM CONCEPTS :

A data stream is an existing, continuous, ordered (implicitly by entrance time or explicitly by timestamp) chain of items.

It is unfeasible to control the order in which units arrive, nor it is feasible to locally capture stream in its entirety.

It is enormous volumes of data, items arrive at a high rate.

TYPES OF DATA STREAMS :

- **Data stream**

A data stream is a(possibly unchained) sequence of tuples.

Each tuple comprised of a set of attributes, similar to a row in a database table.

- **Transactional data stream**

It is a log interconnection between entities

1. Credit card – purchases by consumers from producer
2. Telecommunications – phone calls by callers to the dialed parties
3. Web – accesses by clients of information at servers

- **Measurement data streams**

1. Sensor Networks – a physical natural phenomenon, road traffic
2. IP Network – traffic at router interfaces
3. Earth climate – temperature, humidity level at weather stations

EXAMPLES OF STREAM SOURCES

SENSOR DATA

In navigation systems, sensor data is used.

Imagine a temperature sensor floating about in the ocean, sending back to the base station a reading of the surface temperature each hour.

The data generated by this sensor is a stream of real numbers.

We have 3.5 terabytes arriving every day and we for sure need to think about what we can be kept continuing and what can only be archived.

1. **Image Data**

   Satellites frequently send down-to-earth streams containing many terabytes of images per day. Surveillance cameras generate images with lower resolution than satellites, but there can be numerous of them, each producing a stream of images at a break of 1 second each.

2. **Internet and Web Traffic**

   A bobbing node in the center of the internet receives streams of IP packets from many inputs and paths them to its outputs.

   Websites receive streams of heterogeneous types. For example, Google receives a hundred million search queries per day.

CHARACTERISTICS OF DATA STREAMS:

1. Large volumes of continuous data, possibly infinite.
2. Steady changing and requires a fast, real-time response.
3. Data stream captures nicely our data processing needs of today.
4. Random access is expensive and a single scan algorithm
5. Store only the summary of the data seen so far.
6. Maximum stream data are at a pretty low level or multidimensional in creation, needs multilevel and multidimensional treatment.

APPLICATIONS OF DATA STREAMS:

1. Fraud perception
2. Real-time goods dealing
3. Consumer enterprise
4. Observing and describing on inside IT systems

ADVANTAGES OF DATA STREAMS:

- This data is helpful in upgrading sales
- Help in recognizing the fallacy
- Helps in minimizing costs

- It provides details to react swiftly to risk

DISADVANTAGES OF DATA STREAMS:

- Lack of security of data in the cloud
- Hold cloud donor subordination
- Off-premises warehouse of details introduces the probable for disconnection

DATA MINING – TIME-SERIES DATA

DATA MINING – TIME-SERIES, SYMBOLIC AND BIOLOGICAL SEQUENCES DATA

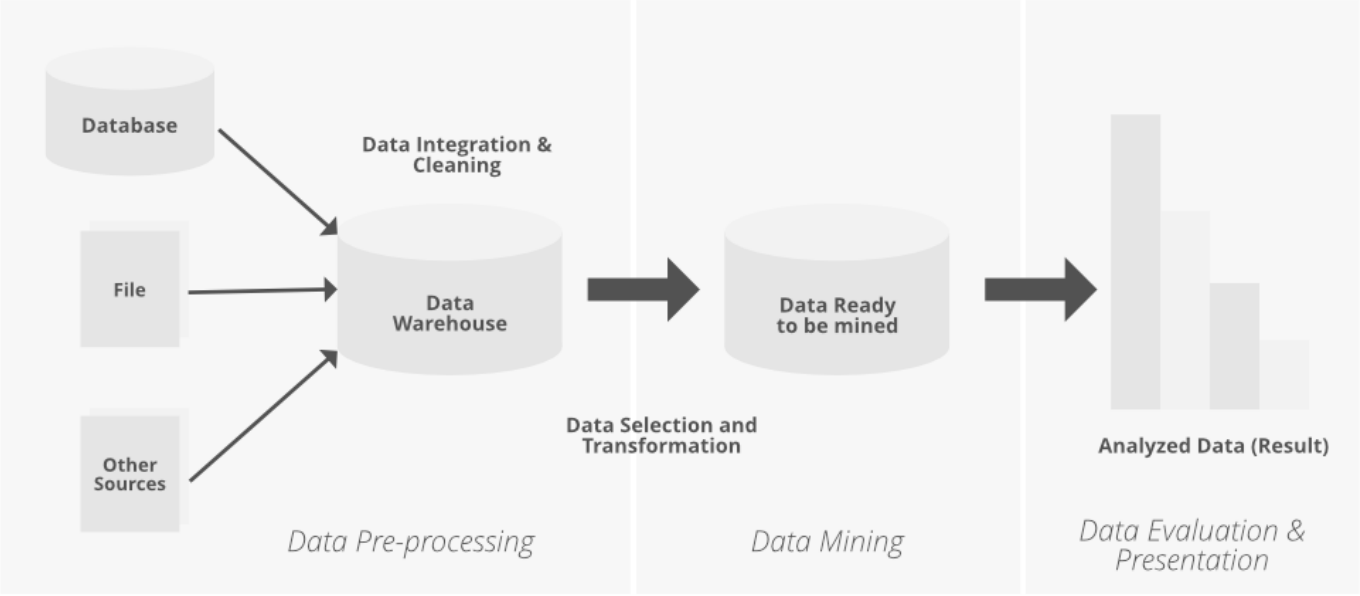
Data mining refers to extracting or mining knowledge from large amounts of data.

In other words, Data mining is the science, art, and technology of discovering large and complex bodies of data in order to discover useful patterns.

Theoreticians and practitioners are continually seeking improved techniques to make the process more efficient, cost-effective, and accurate.

Evaluation of data reached the maximum extent and may still peruse in the future.

To generalize the evaluation of data we classify them as Sequence Data, Graphs, and Network Mining, another kind of data.



A sequence is an ordered list of events.

Sequences data are classified based on characteristics as:

- Time-Series data (data with respect to time)

- Symbolic data (data with laps in an interval of time)
- Biological data (data related to DNA and protein)

TIME-SERIES DATA:

In this type of sequence, the data are of numeric data type recorded at a regular level.

They are generated by an economic process like Stock Market analysis, Medical Observations.

They are useful for studying natural phenomena.

Nowadays these times series are used for piecewise data approximations for further analysis.

In this time-series data, we find a subsequence that matches the query we search.

- **Time Series Forecasting:** Forecasting is a method of making predictions based on past and present data to know what happens in the future. Trend analysis is a method of forecasting Time Series. It is a function that generates historic patterns in time series that are used in short and long-term predictions. We can obtain various patterns in time series like cyclic movements, trend movements, seasonal movements as we see they are with respect to time or season. ARIMA, SARIMA, long memory time series modeling are some of the popular methods for such analysis

SYMBOLIC DATA

This type of ordered set of elements or events is recorded with or without a concrete notion of time.

Some symbolic sequences such as customer shopping sequences, web click streams are examples of symbolic data.

Sequential pattern mining is mainly used for symbolic sequence

Constraint-based pattern matching is one of the best ways to interact with user-defined data.

Apriori is an Algorithm used for this type of analysis below is an example of a symbolic date where we see customer's c1 and c2 are purchasing products at different time intervals

Tid	Time	Cid	Event(purchase products)
t1	11:45:30	c1	wheat, rice, fruit
t2	11:36:50	c2	rice, fruit

Tid	Time	Cid	Event(purchase products)
t1	12:00:01	c1	juice, rice
t2	01:00:34	c2	sugar, milk

BIOLOGICAL DATA

They are made of DNA and protein sequences.

They are very long and complicated but have some hidden meaning.

These types of data are used for the sequence of nucleotides or amino acids.

These analyses are used for aligning, indexes, analyze biological sequence and play a crucial role in bioinformatics and modern biology.

Substitution trees are used to find the probabilities of amino acids and probabilities of intersections.

BLAST-Basic Local Alignment Search Tool is the most effective tool for biological sequence.

GENERALIZED SEQUENTIAL PATTERN (GSP) MINING IN DATA MINING

GSP is a very important algorithm in data mining.

It is used in sequence mining from large databases.

Almost all sequence mining algorithms are basically based on a prior algorithm.

GSP uses a level-wise paradigm for finding all the sequence patterns in the data.

It starts with finding the frequent items of size one and then passes that as input to the next iteration of the GSP algorithm.

The database is passed multiple times to this algorithm.

In each iteration, GSP removes all the non-frequent itemsets.

This is done based on a threshold frequency which is called support.

Only those itemsets are kept whose frequency is greater than the support count.

After the first pass, GSP finds all the frequent sequences of length-1 which are called 1-sequences.

This makes the input to the next pass, it is the candidate for 2-sequences.

At the end of this pass, GSP generates all frequent 2-sequences, which makes the input for candidate 3-sequences.

The algorithm is recursively called until no more frequent itemsets are found.

Basic of Sequential Pattern (GSP) Mining:

- **Sequence:** A sequence is formally defined as the ordered set of items {s1, s2, s3, …, sn}. As the name suggests, it is the sequence of items occurring together. It can be considered as a transaction or purchased items together in a basket.
- **Subsequence:** The subset of the sequence is called a subsequence. Suppose {a, b, g, q, y, e, c} is a sequence. The subsequence of this can be {a, b, c} or {y, e}. Observe that the subsequence is not necessarily consecutive items of the sequence. From the sequences of databases, subsequences are found from which the generalized sequence patterns are found at the end.
- **Sequence pattern:** A sub-sequence is called a pattern when it is found in multiple sequences. The goal of the GSP algorithm is to mine the sequence patterns from the large database. The database consists of the sequences. When a subsequence has a frequency equal to more than the "support" value. For example: the pattern <a, b> is a sequence pattern mined from sequences {b, x, c, a}, {a, b, q}, and {a, u, b}.

SEQUENTIAL PATTERN (GSP) MINING USES

Sequential pattern mining, also known as GSP (Generalized Sequential Pattern) mining, is a technique used to identify patterns in sequential data. The goal of GSP mining is to discover patterns in data that occur over time, such as customer buying habits, website navigation patterns, or sensor data.

Some of the main uses of GSP mining include:

Market basket analysis: GSP mining can be used to analyze customer buying habits and identify products that are frequently purchased together. This can help businesses to optimize their product placement and marketing strategies.

1. **Fraud detection:** GSP mining can be used to identify patterns of behavior that are indicative of fraud, such as unusual patterns of transactions or access to sensitive data.
2. **Website navigation:** GSP mining can be used to analyze website navigation patterns, such as the sequence of pages visited by users, and identify areas of the website that are frequently accessed or ignored.
3. **Sensor data analysis**: GSP mining can be used to analyze sensor data, such as data from IoT devices, and identify patterns in the data that are indicative of certain conditions or states.
4. **Social media analysis**: GSP mining can be used to analyze social media data, such as posts and comments, and identify patterns in the data that indicate trends, sentiment, or other insights.

5. **Medical data analysis**: GSP mining can be used to analyze medical data, such as patient records, and identify patterns in the data that are indicative of certain health conditions or trends.

METHODS FOR SEQUENTIAL PATTERN MINING

- Apriori-based Approaches
    - GSP
    - SPADE
- Pattern-Growth-based Approaches
    - FreeSpan
    - PrefixSpan

Sequence Database:A database that consists of ordered elements or events is called a sequence database. Example of a sequence database:

S.No.	SID	sequences
1.	100	<a(ab)(ac)d(cef)> or <a{ab}{ac}d{cef}>
2.	200	<(ad)c(bcd)(abe)>
3.	300	<(ef)(ab)(def)cb>
4.	400	<eg(adf)CBC>

Transaction: The sequence consists of many elements which are called transactions.

<a(ab)(ac)d(cef)> is a sequence whereas (a), (ab), (ac), (d) and (cef) are the elements of the sequence.

These elements are sometimes referred as transactions.

An element may contain a set of items. Items within an element are unordered and we list them alphabetically.

For example, (cef) is the element and it consists of 3 items c, e and f.

Since, all three items belong to same element, their order does not matter.

But we prefer to put them in alphabetical order for convenience.

The order of the elements of the sequence matters unlike order of items in same transaction.

k-length Sequence:

The number of items involved in the sequence is denoted by K. A sequence of 2 items is called a 2-len sequence. While finding the 2-length candidate sequence this term comes into use. Example of 2-length sequence is: {ab}, {(ab)}, {bc} and {(bc)}.

- {bc} denotes a 2-length sequence where b and c are two different transactions. This can also be written as {(b)(c)}
- {(bc)} denotes a 2-length sequence where b and c are the items belonging to the same transaction, therefore enclosed in the same parenthesis. This can also be written as {(cb)}, because the order of items in the same transaction does not matter.

Support in k-length Sequence:

Support means the frequency. The number of occurrences of a given k-length sequence in the sequence database is known as the support. While finding the support the order is taken care.

Illustration:

Suppose we have 2 sequences in the database.

s1: <a(bc)b(cd)>

s2: <b(ab)abc(de)>

We need to find the support of {ab} and {(bc)}

Finding support of {ab}:

This is present in first sequence.

s1: <a(bc)b(cd)>

Since, a and b belong to different elements, their order matters.

In second sequence {ab} is not found but {ba} is present.

s2: <b(ab)abc(de)> Thus we don't consider this.

Hence, support of {ab} is 1.

Finding support of {bc}:

Since, b and c are present in same element, their order does not matter.

s1: <a(bc)b(cd)>, first occurrence.

s2: <b(ab)abc(de)>, it seems correct, but is not. b and c are present in different elements here. So, we don't consider it.

Hence, support of {(bc)} is 1.

How to join L1 and L1 to give C2?

L1 is the final 1-length sequence after pruning. After pruning all the entries left in the set have supported greater than the threshold.

Case 1: Join {ab} and {ac}

s1: {ab}, s2: {ac}

After removing a from s1 and c from s2.

s1'={b}, s2'={a}

s1′ and s2′ are not same, so s1 and s2 can't be joined.

Case 2: Join {ab} and {be}

s1: {ab}, s2: {be}

After removing a from s1 and e from s2.

s1'={b}, s2'={b}

s1′ and s2′ are exactly same, so s1 and s2 be joined.

s1 + s2 = {abe}

Case 3: Join {(ab)} and {be}

s1: {(ab)}, s2: {be}

After removing a from s1 and e from s2.

s1'={(b)}, s2'={(b)}

s1′ and s2′ are exactly same, so s1 and s2 be joined.

s1 + s2 = {(ab)e}

s1 and s2 are joined in such a way that items belong to correct elements or transactions.

Pruning Phase: While building Ck (candidate set of k-length), we delete a candidate sequence that has a contiguous (k-1) subsequence whose support count is less than the minimum support (threshold). Also, delete a candidate sequence that has any subsequence without minimum support.

{abg} is a candidate sequence of C3.

{abg} is a candidate sequence of C3.

To check if {abg} is proper candidate or not, without checking its support, we check the support of its subsets.

Because subsets of 3-length sequence will be 1 and 2 length sequences. We build the candidate sets increment like 1-length, 2-length and so on.

Subsets of {abg} are: {ab], {bg} and {ag}

Check support of all three subsets. If any of them have support less than minimum support then delete the sequence {abg} from the set C3 otherwise keep it.

Challenges in Generalized Sequential Pattern Data Mining

The database is passed many times to the algorithm recursively. The computational efforts are more to mine the frequent pattern. When the sequence database is very large and patterns to be mined are long then GSP encounters the problem in doing so effectively.

MINING OBJECT

In an object database, data generalization and multidimensional analysis are not applied to individual objects but to classes of objects.

Since a set of objects in a class may share many attributes and methods, and the generalization of each attribute and method may apply a sequence of generalization operators, the major issue becomes how to make the generalization processes cooperate among different attributes and methods in the class(es).

“So, how can class-based generalization be performed for a large set of objects?” For classbased generalization, the attribute-oriented induction method developed for mining characteristics of relational databases can be extended to mine data characteristics in object databases.

Consider that a generalization-based data mining process can be viewed as the application of a sequence of class-based generalization operators on different attributes.

Generalization can continue until the resulting class contains a small number of generalized objects that can be summarized as a concise, generalized rule in high-level terms.

For efficient implementation, the generalization of multidimensional attributes of a complex object class can be performed by examining each attribute (or dimension), generalizing each attribute to simple-valued data, and constructing a multidimensional data cube, called an object cube.

Once an object cube is constructed, multidimensional analysis and data mining can be performed on it in a manner similar to that for relational data cubes.

Notice that from the application point of view, it is not always desirable to generalize a set of values to single-valued data. Consider the attribute keyword, which may contain a set of keywords describing a book.

It does not make much sense to generalize this set of keywords to one single value.

In this context, it is difficult to construct an object cube containing the keyword dimension.

We will address some progress in this direction in the next section when discussing spatial data cube construction.

However, it remains a challenging research issue to develop techniques for handling set-valued data effectively in object cube construction and object-based multidimensional analysis.

TYPES OF MINING OBJECTS

- Spatial
- Multimedia
- Text
- Web Data

Mining Spatiotemporal Data: The data that is related to both space and time is spatiotemporal data.

Spatiotemporal data mining retrieves interesting patterns and knowledge from spatiotemporal data.

Spatiotemporal Data mining helps us to find the value of the lands, the age of the rocks and precious stones, predict the weather patterns.

Spatiotemporal data mining has many practical applications like GPS in mobile phones, timers, Internet-based map services, weather services, satellite, RFID, sensor.

Mining Multimedia Data: Multimedia data objects include image data, video data, audio data, website hyperlinks and linkages.

Multimedia data mining tries to find out interesting patterns from multimedia databases.

This includes the processing of the digital data and performs tasks like image processing, image classification, video and audio data mining, and pattern recognition.

Multimedia Data mining is becoming the most interesting research area because most of the social media platforms like Twitter; Face book data can be analyzed through this and derives interesting trends and patterns.

Mining Web Data: Web mining is essential to discover crucial patterns and knowledge from the Web.

Web content mining analyzes data of several websites which includes the web pages and the multimedia data such as images in the web pages.

Web mining is done to understand the content of web pages, unique users of the website, unique hypertext links, web page relevance and ranking, web page content summaries, time that the users spent on the particular website, and understand user search patterns.

Web mining also finds out the best search engine and determines the search algorithm used by it.

So it helps improve search efficiency and finds the best search engine for the users.

Mining Text Data: Text mining is the subfield of data mining, machine learning, Natural Language processing, and statistics.

Most of the information in our daily life is stored as text such as news articles, technical papers, books, email messages, blogs.

Text Mining helps us to retrieve high-quality information from text such as sentiment analysis, document summarization, text categorization, text clustering.

We apply machine learning models and NLP techniques to derive useful information from the text.

This is done by finding out the hidden patterns and trends by means such as statistical pattern learning and statistical language modeling.

In order to perform text mining, we need to preprocess the text by applying the techniques of stemming and lemmatization in order to convert the textual data into data vectors.

Mining Data Streams: Stream data is the data that can change dynamically and it is noisy, inconsistent which contain multidimensional features of different data types.

So this data is stored in NoSql database systems.

The volume of the stream data is very high and this is the challenge for the effective mining of stream data.

While mining the Data Streams we need to perform the tasks such as clustering, outlier analysis, and the online detection of rare events in data streams.

SPATIAL DATA MINING

A spatial database saves a huge amount of space-related data, including maps, preprocessed remote sensing or medical imaging records, and VLSI chip design data.

Spatial databases have several features that distinguish them from relational databases.

They carry topological and/or distance information, usually organized by sophisticated, multidimensional spatial indexing structures that are accessed by spatial data access methods and often require spatial reasoning, geometric computation, and spatial knowledge representation techniques.

Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases.

Such mining demands the unification of data mining with spatial database technologies.

It can be used for learning spatial records, discovering spatial relationships and relationships among spatial and nonspatial records, constructing spatial knowledge bases, reorganizing spatial databases, and optimizing spatial queries.

It is expected to have broad applications in geographic data systems, marketing, remote sensing, image database exploration, medical imaging, navigation, traffic control, environmental studies, and many other areas where spatial data are used.

A central challenge to spatial data mining is the exploration of efficient spatial data mining techniques because of the large amount of spatial data and the difficulty of spatial data types and spatial access methods.

Statistical spatial data analysis has been a popular approach to analyzing spatial data and exploring geographic information.

The term geostatistics is often associated with continuous geographic space, whereas the term spatial statistics is often associated with discrete space.

In a statistical model that manages non-spatial records, one generally considers statistical independence among different areas of data.

There is no such separation among spatially distributed records because, actually spatial objects are interrelated, or more exactly spatially co-located, in the sense that the closer the two objects are placed, the more likely they send the same properties.

For example, natural resources, climate, temperature, and economic situations are likely to be similar in geographically closely located regions.

Such a property of close interdependency across nearby space leads to the notion of spatial autocorrelation. Based on this notion, spatial statistical modeling methods have been developed with success.

Spatial data mining will create spatial statistical analysis methods and extend them for large amounts of spatial data, with more emphasis on effectiveness, scalability, cooperation with database and data warehouse systems, enhanced user interaction, and the discovery of new kinds of knowledge.

MULTIMEDIA DATA MINING

Multimedia mining is a subfield of data mining that is used to find interesting information of implicit knowledge from multimedia databases.

Mining in multimedia is referred to as automatic annotation or annotation mining.

Mining multimedia data requires two or more data types, such as text and video or text video and audio.

Multimedia data mining is an interdisciplinary field that integrates image processing and understanding, computer vision, data mining, and pattern recognition.

Multimedia data mining discovers interesting patterns from multimedia databases that store and manage large collections of multimedia objects, including image data, video data, audio data, sequence data and hypertext data containing text, text markups, and linkages.

Issues in multimedia data mining include content-based retrieval and similarity search, generalization and multidimensional analysis.

Multimedia data cubes contain additional dimensions and measures for multimedia information.

The framework that manages different types of multimedia data stored, delivered, and utilized in different ways is known as a multimedia database management system.

There are three classes of multimedia databases: static, dynamic, and dimensional media.

The content of the Multimedia Database management system is as follows:

- **Media data:**The actual data representing an object.
- **Media format data:** Information such as sampling rate, resolution, encoding scheme etc., about the format of the media data after it goes through the acquisition, processing and encoding phase.
- **Media keyword data:**Keywords description relating to the generation of data. It is also known as content descriptive data. Example: date, time and place of recording.
- **Media feature data:** Content dependent data such as the distribution of colours, kinds of texture and different shapes present in data.

TYPES OF MULTIMEDIA APPLICATIONS

Types of multimedia applications based on data management characteristics are:

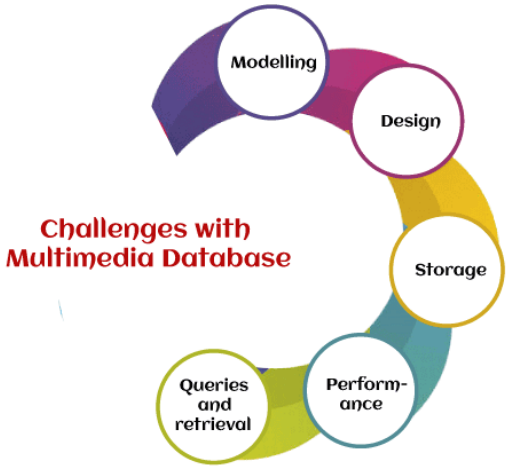
1. **Repository applications:** A Large amount of multimedia data and meta-data (Media format date, Media keyword data, Media feature data) that is stored for retrieval purposes, e.g., Repository of satellite images, engineering drawings, radiology scanned pictures.

2. **Presentation applications:** They involve delivering multimedia data subject to temporal constraints. Optimal viewing or listening requires DBMS to deliver data at a certain rate, offering the quality of service above a certain threshold. Here data is processed as it is delivered. Example: Annotating of video and audio data, real-time editing analysis.

3. **Collaborative work using multimedia information** involves executing a complex task by merging drawings and changing notifications. Example: Intelligent healthcare network.

CHALLENGES WITH MULTIMEDIA DATABASE

There are still many challenges to multimedia databases, such as:



1. **Modeling:** Working in this area can improve database versus information retrieval techniques; thus, documents constitute a specialized area and deserve special consideration.

2. **Design:** The conceptual, logical and physical design of multimedia databases has not yet been addressed fully as performance and tuning issues at each level are far more complex as they consist of a variety of formats like JPEG, GIF, PNG, MPEG, which is not easy to convert from one form to another.

3. **Storage:** Storage of multimedia database on any standard disk presents the problem of representation, compression, mapping to device hierarchies, archiving and buffering during input-output operation. In DBMS, a BLOB (Binary Large Object) facility allows untyped bitmaps to be stored and retrieved.

4. **Performance:** Physical limitations dominate an application involving video playback or audio-video synchronization. The use of parallel processing may alleviate some problems, but such techniques are not yet fully developed. Apart from this, a multimedia database consumes a lot of processing time and bandwidth.

5. **Queries and retrieval:** For multimedia data like images, video, and audio accessing data through query open up many issues like efficient query formulation, query execution and optimization, which need to be worked upon.

Where is Multimedia Database Applied?

Below are the following areas where a multimedia database is applied, such as:

- **Documents and record management:** Industries and businesses keep detailed records and various documents. For example, insurance claim records.
- **Knowledge dissemination:**Multimedia database is a very effective tool for knowledge dissemination in terms of providing several resources. For example, electronic books.
- **Education and training:**Computer-aided learning materials can be designed using multimedia sources which are nowadays very popular sources of learning. Example: Digital libraries.
- **Travelling:** Marketing, advertising, retailing, entertainment and travel. For example, a virtual tour of cities.
- **Real-time control and monitoring:** With active database technology, multimedia presentation of information can effectively monitor and control complex tasks. For example, manufacturing operation control.

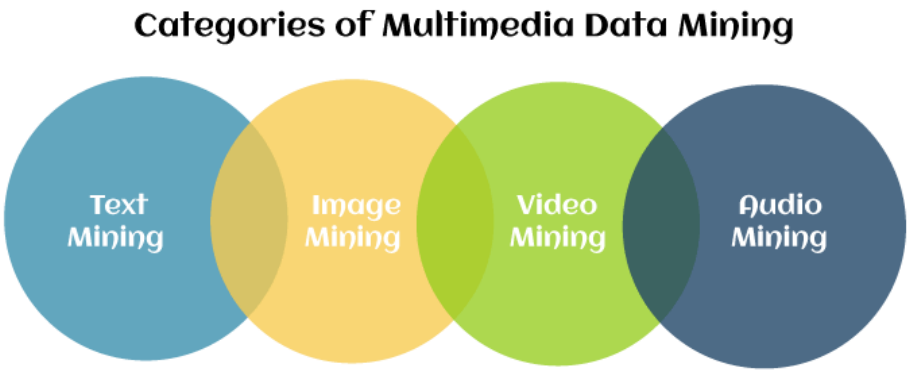
CATEGORIES OF MULTIMEDIA DATA MINING

Multimedia mining refers to analyzing a large amount of multimedia information to extract patterns based on their statistical relationships.

Multimedia data mining is classified into two broad categories:

- Static Media and
- Dynamic Media

Static media contains text (digital library, creating SMS and MMS) and images (photos and medical images). **Dynamic media** contains Audio (music and MP3 sounds) and Video (movies). The below image shows the categories of multimedia data mining.



1. Text Mining

Text is the foremost general medium for the proper exchange of information.

Text Mining evaluates a huge amount of usual language text and detects exact patterns to find useful information.

Text Mining also referred to as text data mining, is used to find meaningful information from unstructured texts from various sources.

2. Image Mining

Image mining systems can discover meaningful information or image patterns from a huge collection of images.

Image mining determines how low-level pixel representation consists of a raw image or image sequence that can be handled to recognize high-level spatial objects and relationships.

It includes digital image processing, image understanding, database, AI, etc.

3. Video Mining

Video mining is unsubstantiated to find interesting patterns from many video data; multimedia data is video data such as text, image, metadata, visuals and audio.

It is commonly used in security and surveillance, entertainment, medicine, sports and education programs.

The processing is indexing, automatic segmentation, content-based retrieval, classification and detecting triggers.

4. Audio Mining

Audio mining plays an important role in multimedia applications, is a technique by which the content of an audio signal can be automatically searched, analyzed and rotten with wavelet transformation.

It is generally used in automatic speech recognition, where the analysis efforts to find any speech within the audio.

Band energy, frequency centroid, zero-crossing rate, pitch period and bandwidth are often used for audio processing.

APPLICATION OF MULTIMEDIA MINING

There are different kinds of applications of multimedia data mining, some of which are as follows:

- **Digital Library:** The collection of digital data is stored and maintained in a digital library, which is essential to convert different digital data formats into text, images, video, audio, etc.
- **Traffic Video Sequences:** To determine important but previously unidentified knowledge from the traffic video sequences, detailed analysis and mining are to be performed based on vehicle identification, traffic flow, and queue temporal relations of the vehicle at an intersection. This provides an economic approach for regular traffic monitoring processes.
- **Medical Analysis:** Multimedia mining is primarily used in the medical field, particularly for analyzing medical images. Various data mining techniques are used for image classification. Examples, Automatic 3D delineation of highly aggressive brain tumours, Automatic localization and identification of vertebrae in 3D CT scans, MRI Scans, ECG and X-Ray.
- **Customer Perception:** It contains details about customers' opinions, products or services, customer's complaints, customers preferences, and the level of customer satisfaction with products or services, which are collected together. The audio data serve as topic detection, resource assignment and evaluation of the quality of services. Many companies have call centres that receive telephone calls from customers.
- **Media Making and Broadcasting:** Radio stations and TV channels create broadcasting companies, and multimedia mining can be applied to monitor their content to search for more efficient approaches and improve their quality.
- **Surveillance system:** It consists of collecting, analyzing, summarizing audio, video or audiovisual information about specific areas like government organizations, multi-national companies, shopping malls, banks, forests, agricultural areas and, highways etc. The main use of this technology in the field of security; hence it can be utilized by military, police and private companies since they provide security services.

PROCESS OF MULTIMEDIA DATA MINING

The below image shows the present architecture, which includes the types of the multimedia mining process.

Data Collection is the initial stage of the learning system; Pre-processing is to extract significant features from raw data.

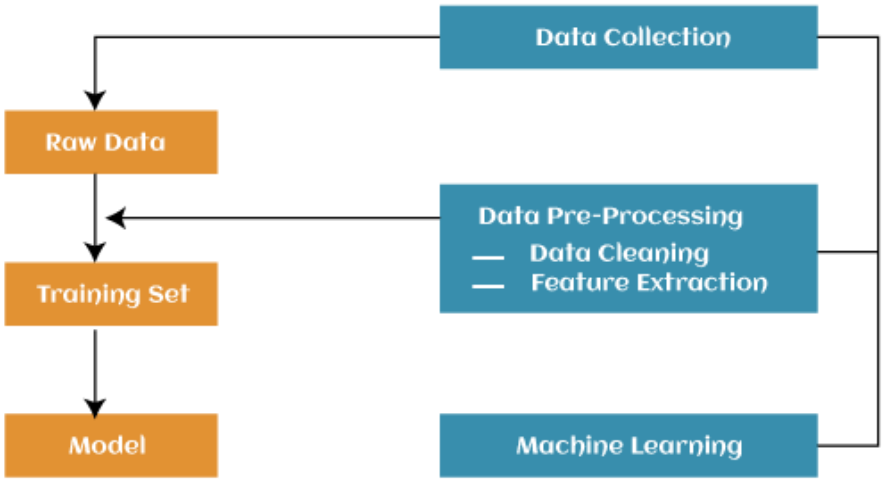
It includes data cleaning, transformation, normalization, feature extraction, etc.

Learning can be direct if informative types can be recognized at preprocessing stage.

The complete process depends extremely on the nature of raw data and the difficulty field.

The product of preprocessing is the training set.

A learning model must be selected for the specified training set to learn from it and make the multimedia model more constant.



Multimedia Mining Process

Converting Un-structured data to structured data:

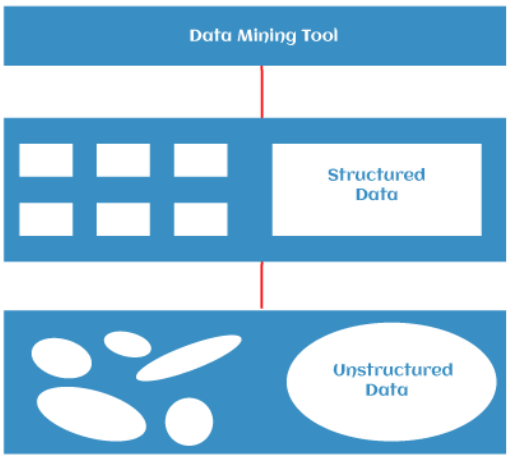
Data resides in a fixed field within a record or file is called structured data, and these data are stored in sequential form.

Structured data has been easily entered, stored, queried and analyzed.

Unstructured data is bitstream, for example, pixel representation for an image, audio, video and character representation for text.

These files may have an internal structure, but they are still considered "unstructured" because their data does not fit neatly in a database.

For example, images and videos of different objects have some similarities - each represents an interpretation of a building without a clear structure.



Unstructured Data to Structured Data Conversion

Current data mining tools operate on structured data, which resides in a huge volume of the relational database, while data in multimedia databases are semi-structured or unstructured.

Hence, the semi-structured or unstructured multimedia data is converted into structured one, and then the current data mining tools are used to extract the knowledge.

The sequence or time element is different between unstructured and structured data mining.

The architecture of converting unstructured data to structured data and which is used for extracting information from the unstructured database is shown in the above image.

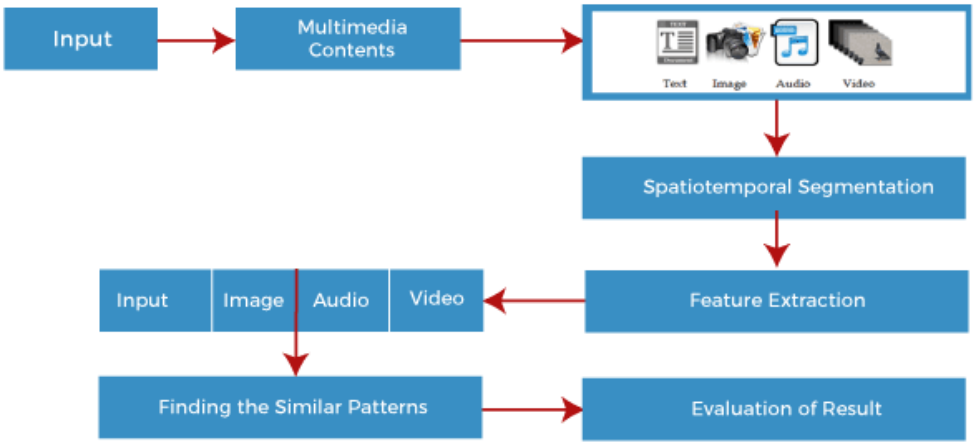
Then data mining tools are applied to the stored structured databases.

ARCHITECTURE FOR MULTIMEDIA DATA MINING

Multimedia mining architecture is given in the below image.

The architecture has several components.

Important components are Input, Multimedia Content, Spatiotemporal Segmentation, Feature Extraction, Finding similar Patterns, and Evaluation of Results.



Multimedia Data mining Architecture

1. **The input** stage comprises a multimedia database used to find the patterns and perform the data mining.

2. **Multimedia Content** is the data selection stage that requires the user to select the databases, subset of fields, or data for data mining.

3. **Spatio-temporal segmentation** is nothing but moving objects in image sequences in the videos, and it is useful for object segmentation.

4. **Feature extraction** is the preprocessing step that involves integrating data from various sources and making choices regarding characterizing or coding certain data fields to serve when inputs to the pattern-finding stage. Such representation of choices is required because certain fields could include data at various levels and are not considered for finding a similar pattern stage. In MDM, the preprocessing stage is significant since the unstructured nature of multimedia records.

5. **Finding a similar pattern** stage is the heart of the whole data mining process. The hidden patterns and trends in the data are basically uncovered in this stage. Some approaches to finding similar pattern stages contain association, classification, clustering, regression, time-series analysis and visualization.

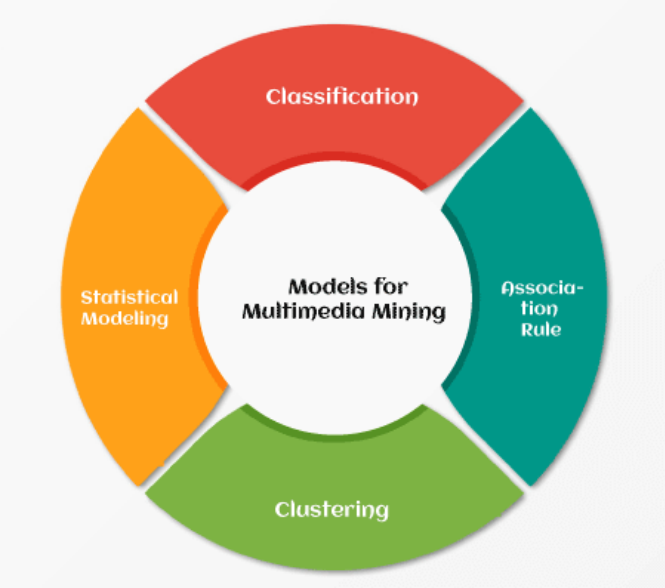
6. **Evaluation of Results** is a data mining process used to evaluate the results, and this is important to determine whether the prior stage must be revisited or not. This stage consists of reporting and using the extracted knowledge to produce new actions, products, services, or marketing strategies.

MODELS FOR MULTIMEDIA MINING

The models which are used to perform multimedia data are very important in mining.

Commonly four different multimedia mining models have been used.

These are classification, association rule, clustering and statistical modeling.



1. **Classification:** Classification is a technique for multimedia data analysis that can learn from every property of a specified set of multimedia.

   It is divided into a predefined class label to achieve the purpose of classification.

   Classification is the process of constructing data into categories for its better effective and efficient use; it creates a function that well-planned data item into one of many predefined classes by inputting a training data set and building a model of the class attribute based on the rest of the attributes.

   Decision tree classification has a perceptive nature that the users conceptual model without loss of exactness.

   Hidden Markov Model is used to classify multimedia data such as images and videos as indoor-outdoor games.

2. **Association Rule:** Association Rule is one of the most important data mining techniques that help find relations between data items in huge databases.

   There are two types of associations in multimedia mining: image content and non-image content features.

   Mining the frequently occurring patterns between different images becomes mining the repeated patterns in a set of transactions.

   Multi-relational association rule mining displays multiple reports for the same image.

   In image classification also, multiple-level association rule techniques are used.

3. **Clustering:** Cluster analysis divides the data objects into multiple groups or clusters.

   Cluster analysis combines all objects based on their groups.

   In multimedia mining, the clustering technique can be applied to group similar images, objects, sounds, videos and texts.

   Clustering algorithms can be divided into several methods: hierarchical methods, density-based methods, grid-based methods, model-based methods, k-means algorithms, and graph-based models.

4. **Statistical Modeling:** Statistical mining models regulate the statistical validity of test parameters and have been used to test hypotheses, undertake correlation studies, and transform and make data for further analysis.

   This is used to establish links between words and partitioned image regions to form a simple co-occurrence model.

ISSUES IN MULTIMEDIA MINING

Major Issues in multimedia data mining contains content-based retrieval, similarity search, dimensional analysis, classification, prediction analysis and mining associations in multimedia data.

1. Content-based retrieval and Similarity search

Content-based retrieval in multimedia is a stimulating problem since multimedia data is required for detailed analysis from pixel values.

We considered two main families of multimedia retrieval systems, i.e. similarity search in multimedia data.

- **Description-based retrieval system** creates indices and object retrieval based on image descriptions, such as keywords, captions, size, and creation time.
- **Content-based retrieval system** supports image content retrieval, for example, colour histogram, texture, shape, objects, and wavelet transform.
- **Use of content-based retrieval system:** Visual features index images and promote object retrieval based on feature similarity; it is very desirable in various applications. These applications include diagnosis, weather prediction, TV production and internet search engines for pictures and e-commerce.

2. Multidimensional Analysis

To perform multidimensional analysis of large multimedia databases, multimedia data cubes may be designed and constructed similarly to traditional data cubes from relational data.

A multimedia data cube has several dimensions.

For example, the size of the image or video in bytes; the width and height of the frames, creating two dimensions, the date on which image or video was created or last modified, the format type of the image or video, frame sequence duration in seconds, Internet domain of pages referencing the image or video, the keywords like a color dimension and edge orientation dimension.

A multimedia data cube can have additional dimensions and measures for multimedia data, such as color, texture, and shape.

The Multimedia data mining system prototype is Multimedia Miner, the extension of the DBMiner system that handles multimedia data.

The Image Excavator component of Multimedia Miner uses image contextual information, like HTML tags on Web pages, to derive keywords.

By navigating online directory structures, like Yahoo! directory, it is possible to build hierarchies of keywords mapped on the directories in which the image was found.

3. Classification and Prediction Analysis

Classification and predictive analysis has been used for mining multimedia data, particularly in scientific analysis like astronomy, seismology, and geoscientific analysis.

Decision tree classification is an important method for reported image data mining applications.

For example, consider the sky images, which astronomers have carefully classified as the training set.

It can create models for recognizing galaxies, stars and further stellar objects based on properties like magnitudes, areas, intensity, image moments and orientation.

Image data mining classification and clustering are carefully connected to image analysis and scientific data mining.

The image data are frequently in large volumes and need substantial processing power, such as parallel and distributed processing.

Hence, many image analysis techniques and scientific data analysis methods could be applied to image data mining.

4. Mining Associations in Multimedia

Data Association rules involving multimedia objects have been mined in image and video databases. Three categories can be observed:

- Associations between image content and non-image content features
- Associations among image contents that are not related to spatial relationships
- Associations among image contents related to spatial relationships

First, an image contains multiple objects, each with various features such as colour, shape, texture, keyword, and spatial locations, so that many possible associations can be made.

Second, a picture containing multiple repeated objects is essential in image analysis.

The recurrence of similar objects should not be ignored in association analysis.

Third, to find the associations between the spatial relationships and multimedia images can be used to discover object associations and correlations.

With the associations between multimedia objects, we can treat every image as a transaction and find commonly occurring patterns among different images.

TEXT MINING

Text mining is also known as text analysis.

It is the procedure of transforming unstructured text into structured data for simple analysis.

Text mining applies natural language processing (NLP), enabling machines to know the human language and process it automatically.

It is defined as the procedure of deriving significant information from standard language text.

Some data that it can generate via text messages, records, emails, files are written in common language text.

It is generally used to draw beneficial insights or patterns from such data.

Text mining is an automatic method that uses natural language processing to derive valuable insights from unstructured text.

It can be converting data into information that devices can learn, text mining automates the method of classifying texts by sentiment, subject, and intent.

In text data mining, it is used on textual data.

It can read and analyze textual information.

In text mining, the pattern are extracted from the unstructured data or natural language text.

In text mining, the input is unstructured text and then the output is structured text.

Text Mining includes a set of text documents are in the form of pdf, doc, Docx, txt, etc.

After receiving the document, using Pre-processing (compare to NLT – Natural Language Text) of text and then Text Mining approaches.

Thus, analyzing the text document finally find the knowledge.

There are two methods are involved as Filtering and Streaming.

Filtering can remove unwanted words or relevant information.

Streaming words provide the root for the associated words.

After using the streaming method every word is designed by its root node.

Text Mining is an area that is an unexpected explosion in adoptions for business applications.

The explosion in adoption is triggered by heightened information about TM and the lowered price points at which TM tools are available today.

Manual analysis of unstructured textual data is more impractical, and accordingly, text mining methods are being developed to automate the process of analyzing the data.

The primary objective of text mining is to allow users to extract records from textbased assets and handles the services like Retrieval, Extraction, Summarization, Categorization (supervised), and Clustering (unsupervised), Segmentation, and Association.

The main reason after the adoption of text mining is more powerful competition in the business industry, several organizations seeking value-added solutions to play with other organizations.

With raising completion in business and changing user perspectives, organizations are getting huge investments to get a solution that is able of analyzing user and adversary data to improve competitiveness.

MINING THE WORLD WIDE WEB

Over the last few years, the **World Wide Web** has become a significant source of information and simultaneously a popular platform for business.

Web mining can define as the method of utilizing data mining techniques and algorithms to extract useful information directly from the web, such as Web documents and services, hyperlinks, Web content, and server logs.

The World Wide Web contains a large amount of data that provides a rich source to data mining.

The objective of Web mining is to look for patterns in Web data by collecting and examining data in order to gain insights.

What is Web Mining?

Web mining can widely be seen as the application of adapted data mining techniques to the web, whereas data mining is defined as the application of the algorithm to discover patterns on mostly structured data embedded into a **knowledge discovery process**.

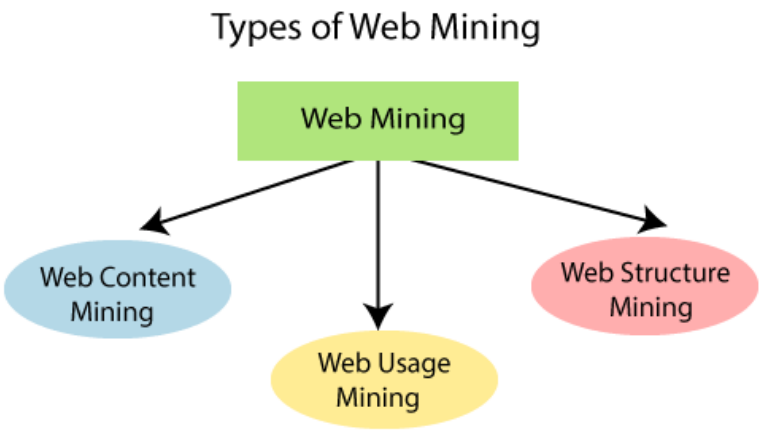
Web mining has a distinctive property to provide a set of various data types.

The web has multiple aspects that yield different approaches for the mining process, such as web pages consist of text, web pages are linked via hyperlinks, and user activity can be monitored via web server logs.

These three features lead to the differentiation between the three areas are web content mining, web structure mining, web usage mining.

There are three types of data mining:

- Web Content Mining
- Web Usage Mining
- Web Structure Mining



1. Web Content Mining:

Web content mining can be used to extract useful data, information, and knowledge from the web page content.

In web content mining, each web page is considered as an individual document.

The individual can take advantage of the semi-structured nature of web pages, as HTML provides information that concerns not only the layout but also logical structure.

The primary task of content mining is data extraction, where structured data is extracted from unstructured websites.

The objective is to facilitate data aggregation over various web sites by using the extracted structured data. Web content mining can be utilized to distinguish topics on the web.

For Example, if any user searches for a specific task on the search engine, then the user will get a list of suggestions.

2. Web Structured Mining:

The web structure mining can be used to find the link structure of hyperlink.

It is used to identify that data either link the web pages or direct link network.

In Web Structure Mining, an individual considers the web as a directed graph, with the web pages being the vertices that are associated with hyperlinks.

The most important application in this regard is the Google search engine, which estimates the ranking of its outcomes primarily with the Page Rank algorithm.

It characterizes a page to be exceptionally relevant when frequently connected by other highly related pages. Structure and content mining methodologies are usually combined.

For example, web structured mining can be beneficial to organizations to regulate the network between two commercial sites.

3. Web Usage Mining:

Web usage mining is used to extract useful data, information, knowledge from the weblog records, and assists in recognizing the user access patterns for web pages.

In Mining, the usage of web resources, the individual is thinking about records of requests of visitors of a website, that are often collected as web server logs.

While the content and structure of the collection of web pages follow the intentions of the authors of the pages, the individual requests demonstrate how the consumers see these pages.

Web usage mining may disclose relationships that were not proposed by the creator of the pages.

Some of the methods to identify and analyze the web usage patterns are given below:

I. Session and visitor analysis:

The analysis of preprocessed data can be accomplished in session analysis, which incorporates the guest records, days, time, sessions, etc.

This data can be utilized to analyze the visitor's behavior.

The document is created after this analysis, which contains the details of repeatedly visited web pages, common entry, and exit.

II. OLAP (Online Analytical Processing):

OLAP accomplishes a multidimensional analysis of advanced data.

OLAP can be accomplished on various parts of log related data in a specific period.

OLAP tools can be used to infer important business intelligence metrics

CHALLENGES IN WEB MINING:

The web pretends incredible challenges for resources, and knowledge discovery based on the following observations:

- **The complexity of web pages:**

The site pages don't have a unifying structure. They are extremely complicated as compared to traditional text documents. There are enormous amounts of documents in the digital library of the web. These libraries are not organized according to a specific order.

- **The web is a dynamic data source:**

The data on the internet is quickly updated. For example, news, climate, shopping, financial news, sports, and so on.

- **Diversity of client networks:**

The client network on the web is quickly expanding.

These clients have different interests, backgrounds, and usage purposes.

There are over a hundred million workstations that are associated with the internet and still increasing tremendously.

- **Relevancy of data:**

It is considered that a specific person is generally concerned about a small portion of the web, while the rest of the segment of the web contains the data that is not familiar to the user and may lead to unwanted results.

- **The web is too broad:**

The size of the web is tremendous and rapidly increasing.

It appears that the web is too huge for data warehousing and data mining.

MINING THE WEB'S LINK STRUCTURES TO RECOGNIZE AUTHORITATIVE WEB PAGES:

The web comprises of pages as well as hyperlinks indicating from one to another page.

When a creator of a Web page creates a hyperlink showing another Web page, this can be considered as the creator's authorization of the other page.

The unified authorization of a given page by various creators on the web may indicate the significance of the page and may naturally prompt the discovery of authoritative web pages.

The web linkage data provide rich data about the relevance, the quality, and structure of the web's content, and thus is a rich source of web mining.

APPLICATION OF WEB MINING:

Web mining has an extensive application because of various uses of the web.

The list of some applications of web mining is given below.

- Marketing and conversion tool
- Data analysis on website and application accomplishment.
- Audience behavior analysis
- Advertising and campaign accomplishment analysis.
- Testing and analysis of a site.