Unit-V

Advanced concepts

Mining Data streams:

streams are temporally ordered, fast changing, massive & potentially infinite.

Characteristics:

→ huge volumes of continuous data, possibly infinite
→ fast changing & requires fast, real time response
→ most stream data are at pretty low level or multi dimensional
→ Random access is expensive

Methodologies for stream data processing & Stream Data systems

1. Random sampling:

Reservoir sampling: Maintaining a set of S candidates in the reservoir which form a true random sample of the elements seen so far in the stream.

As a data stream flow, every new element has a certain probability (S/N) of replacing an old element in the reservoir

Sliding windows:

→ make decisions based only on recent data of sliding window size w.

→ An element arriving at time t expires at time t+w.

Histograms:

→ It is used to approximate the frequency distribution of element values in a stream.

→ partitions data into a set of contiguous buckets.

Multi-Resolution Models:

popular models: balanced binary trees, micro clusters, wavelets.

Sketches:

Histograms & wavelets require multi passes over the data but sketches can operate in a single pass.

Frequency Moments of Stream $A = \{a_1, a_2, \ldots a_n\}$

$$F_k = \sum_{i=1}^{v} m_i^k$$

$v \rightarrow$ the universe or domain size

$m_i \rightarrow$ frequency of i^{th} in the sequence

Given N elts & v values, sketches can approximate F_0, F_1, F_2 in $O(\log v + \log N)$ space

$F_0 \rightarrow$ no. of distinct ele in seq

$F_1 \rightarrow$ length of sequence

$F_2 \rightarrow$ self join size

Randomized Algorithms to analyze data streams

Randomized algorithms in the form of random sampling & sketching are often used to deal with massive, high dimensional data streams.

Las vegas → always return right answer but the running time varies

Monte carlo → these are time bound, may not always return the correct answer.

when random variables are used, its deviation from the expected value must be bounded.

→ **Chebyshev's Inequality**

let X be a random variable with mean M & standard deviation σ

$$P(|X-M| \geq K) \leq \frac{\sigma^2}{K^2}$$

→ **Chernoff bound**

let X be the sum of independent poisson trials $X_1, X_2, \ldots X_n$ $\delta \in [0,1]$

~~$P(|X-M|)$~~

$$P[X < (1+\delta)M] < e^{-M\delta^2/4}.$$

Data stream Management System (DSMS) & Stream Queries

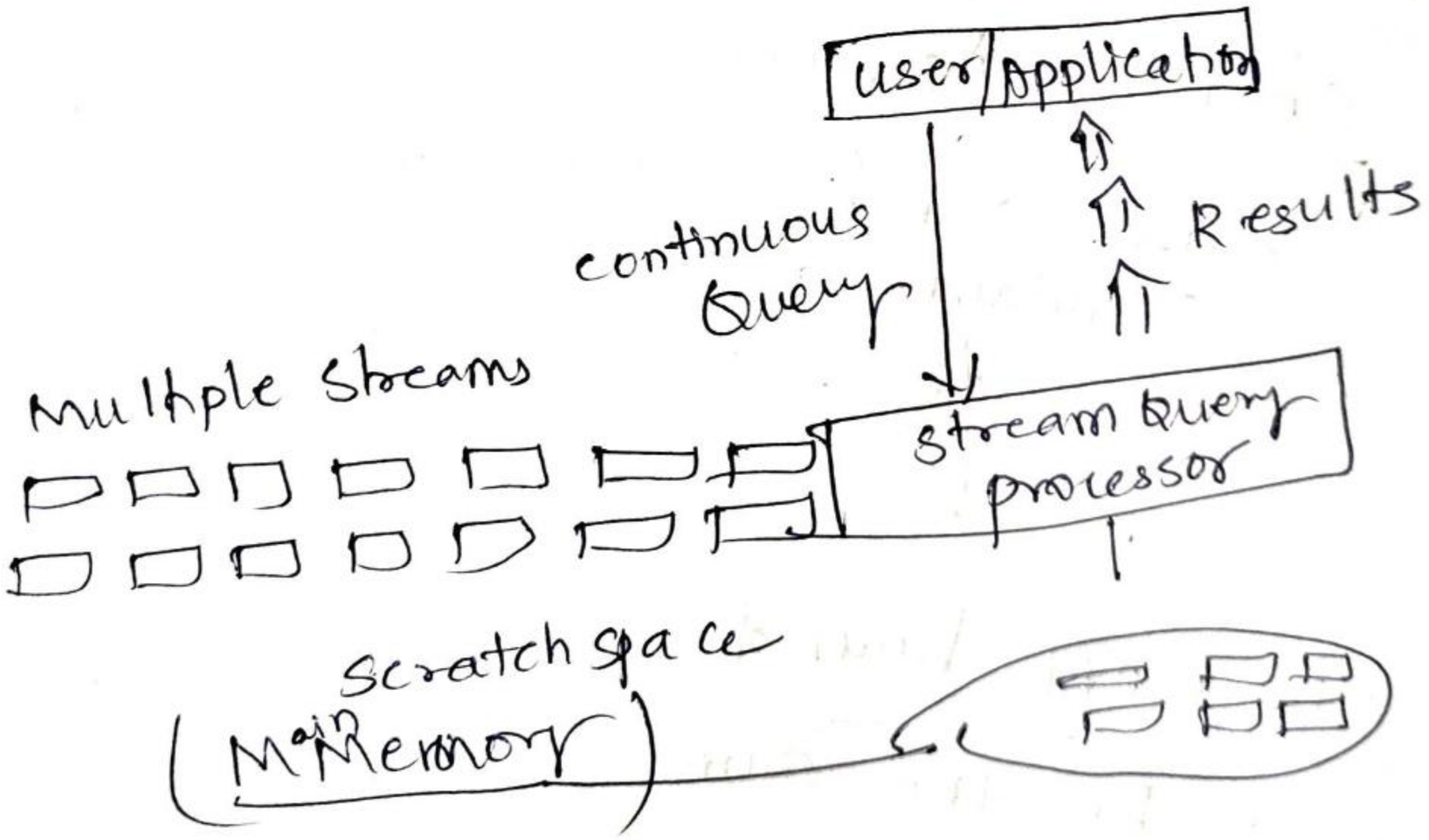
→ In traditional db systems, data are stored in the finite & persistent databases.

→ Stream data are infinite & impossible to stor fully in the database

→ In DSMS, there may be multiple streams

→ once an element from a data stream has been processed, discarded or archived, & it cannot be easily retrieved until it is explicitly stored in memory.

Stream Query processing



A SQP architecture has 3 parts, enduser, query processor & scratchspace

End user issues a query to DSMS, & the Query processor takes the query, process it using the info stored in scratch space & returns the result to the User.

Lossy counting Algorithm

→ for frequent items.

Input: min-support threshold σ & the error bound ε

- Incoming stream is divided into buckets of width $w = \lceil 1/\varepsilon \rceil$
- $N \rightarrow$ Current stream length.

Frequency list data structure is maintained

→ For each item - approximate freq count f & maximum possible error Δ are maintained

→ If the item is from the b^{th} bucket, the max. possible error Δ on the frequency count of the item is $b-1$.

→ when a bucket boundary is reached, an item entry is deleted if $f + \Delta <= b$ the current bucket number. - frequency list is kept small.

→ The frequency of an item can be under estimated by atmost εN

. $f + \Delta <= b$, $b <= N/w$,

All frequent item & some sub frequent items with frequency $\sigma N - \varepsilon N$ will be output -

Properties :

1) there are no false negatives
2) false positives have a frequency of atleast $\sigma N - \varepsilon N$.
3) Frequency of a frequent item is under-estimated by atmost εN

To find frequent itemsets

→ Load as many buckets as possible into main memory – B

→ If updated frequency $f + \Delta <= b$ where b is the current bucket number entry can be deleted

→ If an itemset has frequncy $f >= B$ & does not appear in the list, it is inserted as new entry with Δ set to $b - B$.

Strengths

→ a simple idea

→ can be extended to frequent itemsets

weakness

Space bound is not good.

Hoeffding Tree Algorithm

→ is a decision tree learning method for stream data classification.

→ It was initially used to track web click streams & construct models to predict which web hosts & web sites a user is likely to access

→ Hoeffding bound states that

$$\varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2N}}$$

$r \rightarrow$ random variable

$R \rightarrow$ Range of r

n : # independent observations

True mean of r is atleast $\bar{r} - \varepsilon$, with probability $1-\delta$ (user input)

Hoeffding Tree Input

S: sequence of examples

X: attributes

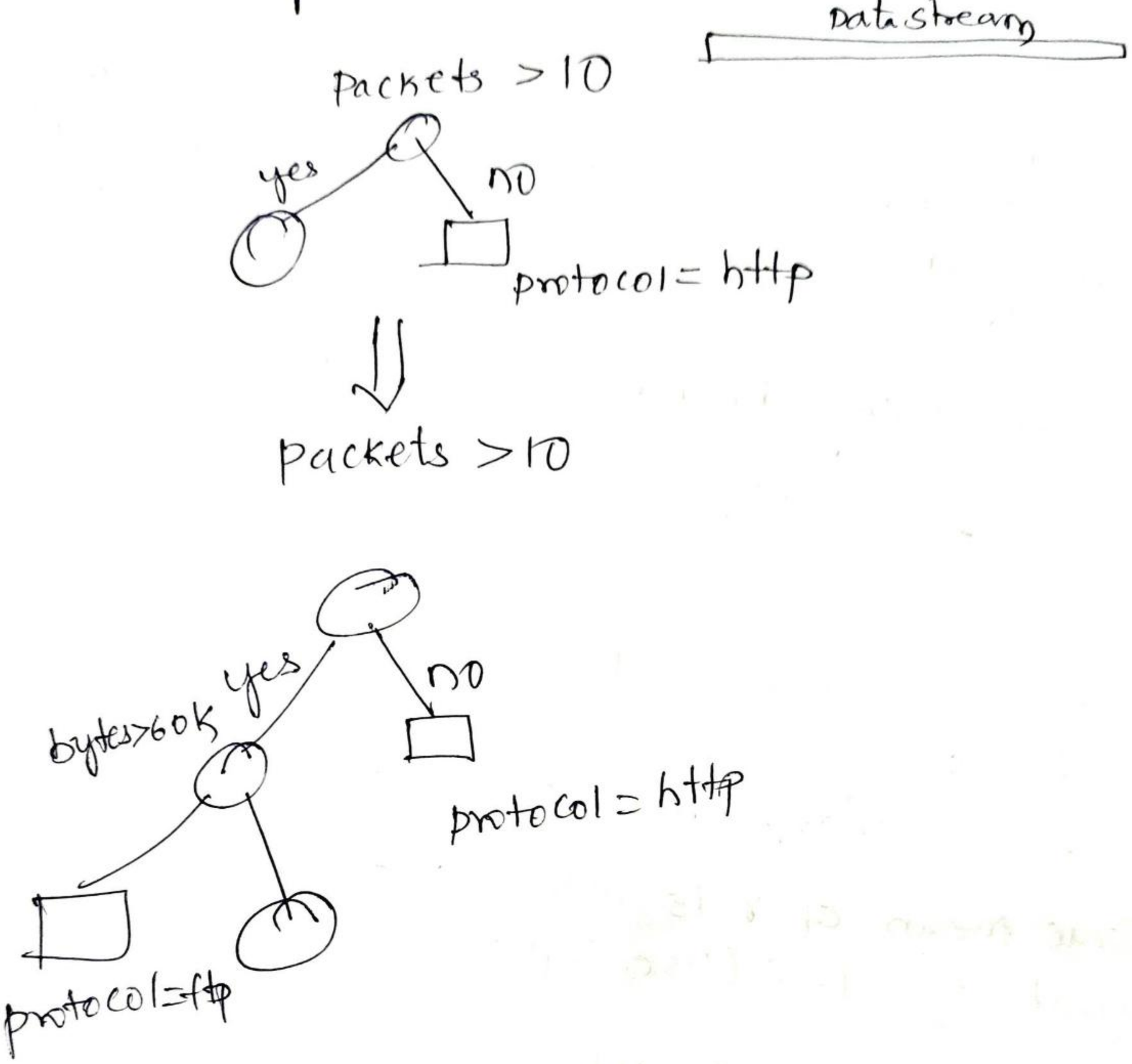
G(): evaluation function

d: desired accuracy

Alg

For each node
  retrieve $G(X_a)$ & $G(X_b)$
    if $(G(X_a) - G(X_b) > \varepsilon)$
      split on X_a
      recurse to next node
      break

Hoeffding Tree Algorithms



Strengths

- scales better than traditional methods
  - sub linear with sampling
  - Very small memory utilization

weakness

- could spend a lot of time with ties
- Memory used with tree expansion
- Cannot handle concept drift

Very Fast Decision Tree (VFDT)

→ makes several modifications to the Hoeffding Tree Algorithm.

→ The modifications include breaking near ties during attribute selection more aggressively, computing the G after a no. of training egs, deactivating the least promising leaves whenever memory is running low, dropping poor splitting attribute & improving the initialization method.

Concept adapting VFDT

→ Time changing datastreams

→ Incorporate new & eliminate old

→ sliding window approach

CVFDT increments the counts with new example & decrement old example. when this happens it grows alternate subtrees.

Mining Time Series Data

A time series database consists of sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time intervals (eg hourly, daily, weekly).

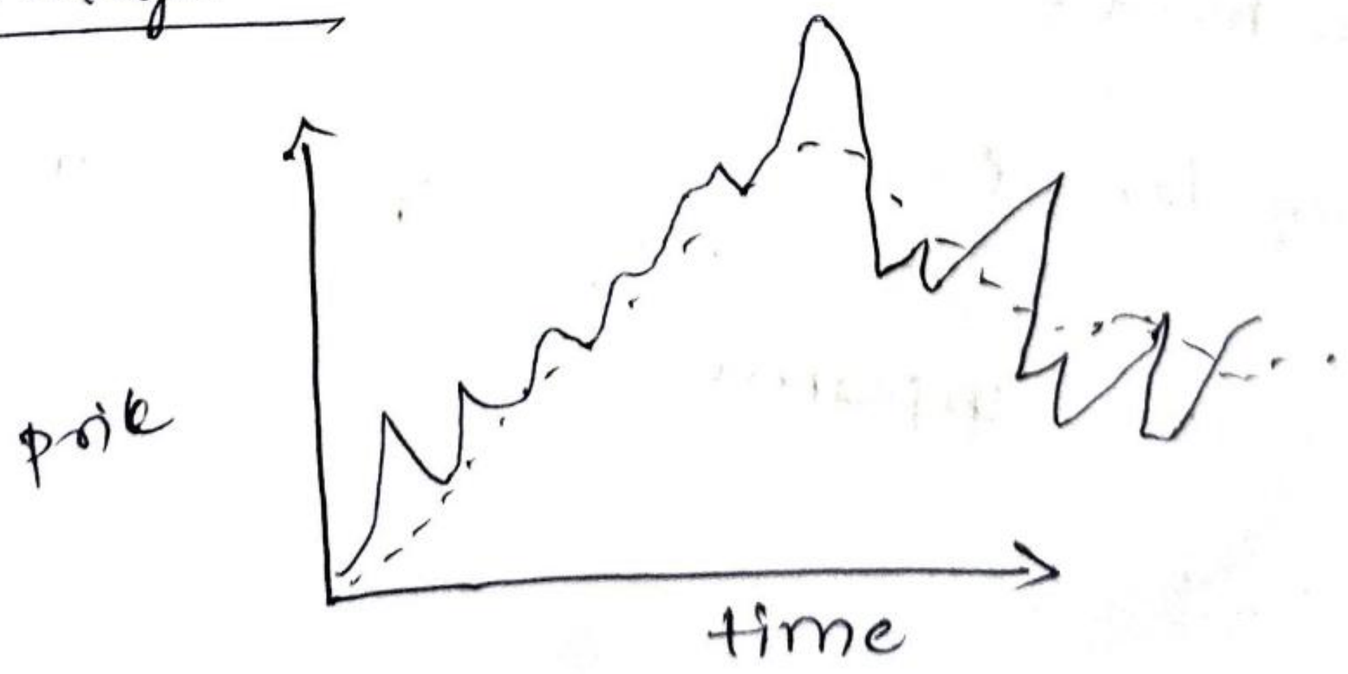
Applications:

stock Market Analysis
ecomic & sales forecasting
budgetary analysis
Medical Treatments

A time series Database is also a sequence database (Consists of a sequence of ordered events time-optional)

Time series data can be analyzed to
→ identify correlations
→ similar / Regular patterns, trends, outliers.

Trend Analysis



? sports (3),

Time series involving a variable y can be represented as a function of time t, $Y = F(t)$.

Goals of Time Series Analysis

→ Modeling time series
→ Forecasting time series.

Trend Analysis consists of following four major components.

1. Trend or long term movements:
a general direction in which a time series is moving over a long interval of time

2. cyclic movements or cyclic variation
long term oscillations about a trend line or curve
eg business cycles, may or may not be periodic.

3. Seasonal movements or Seasonal variations:
i.e almost identical patterns that a time series appears to follow during corresponding months of successive years

4. Irregular or Random movements
It changes that occur randomly due to unplanned events

Refer PPT.

Mining Object

Object Relational & object oriented Database systems deal with the efficient storage & access vast amounts of disk based complex structured data objects. These systems organize a large set of complex data objects into classes, which are in turn organized into class/subclass hierarchies.

Each object in a class is associated with
i) an object identifier
ii) a set of attributes
iii) a set of methods

Generalization of structured Data

A set valued attribute may be homogeneous or heterogeneous type.
A set valued data can be organized by
i) generalization of each value in the set to its corresponding higher level concept
ii) Derivation of the general behaviour of the set.

eg hobby = {tennis, hockey, soccer, violin, simcity}
This set can be generalized to higher level
{sports, music, games} & count = 5
{sports(3), music(1), computer games(1)}

... ... can be.

List valued attributes / sequence

Can be generalized in a manner similar to that for set valued attributes except that the order of the elements in the list or sequence should be presered in generalization

Eg Education Record
{ BSC - 1998, MSC, 2001, PhD 2003 }

Aggregation & Approximation in Spatal & Multimedia Data Generalization

Aggregation & approximation are important techniques for Generalization. In spatial merge, it is necessary to not only merge the regions of similar types with the same general class but also to compute the total areas, average density :

A multimedia database may contain complex texts, graphics, images, voice, music.

Generalization of object Identifiers & class/subclass Hierarchies

An object identifier can be generalized as →

First the object identifier is generalized

to the identifier of lowest subclass to which the object belongs, the identifier of this subclass can then, in turn, be generalized to a higher level class/subclass by climbing up the class/subclass hierarchy.

Generalization of class composition Hierarchies

An attribute of an object may be composed of or described by another object, some of whose attributes may be in turn composed of or described by other objects thus forming a class composition hierarchy

Construction & Mining of object cubes

for efficient implementation, the generalization of multidimensional attributes of a complex object class can be performed by examining each attribute, generalizing each attribute to simple valued data, & constructing a multidimensional data cube called an object cube.

Generalization-Based Mining of plan Databases by Divide & Conquer

A plan consists of a variable sequence of actions. A plan database, a large collection of plan.

plan Mining is the task of Mining significant patterns or knowledge from a plan base.

Spatial Data Mining

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

Spatial database stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, & VLSI chip layout data.

Spatial Data cube construction & Spatial OLAP

A spatial DWH is a subject-oriented, integrated, time-variant, & non-voltaile collection of both spatial & nonspatial data in support of spatial data mining & spatial data related decision making process

There are three types of dimensions in Spatial datu cube:

A Nonspatial Dimension

contains only nonspatial data.

eg hot for temperature
wet for precipitation.

A spatial to nonspatial Dimension

is a dimension whose primitive level data are spatial whose generalization starting at a certain high level becomes nonspatial.

~~eg is_a(X, "school") ∧ close_to(X, "sports centu")~~
~~⇒ close_to(x, "park")~~
~~[0.5, 80%]~~

Spatial to spatial Dimension

is a dimension whose primitive level & all of its high level generalized data are spatial.

eg Dimension equi-temperature region contains spatial data, as do all of its generalization such as with regions covering 0.5° (Celsius), 5-10 degrees & so-on.

Two types of measures in a Spatial data cube

A Numerical measure contains only numeric data

eg monthly-revenue of a region.

A Spatial measure contains a collection of pointers to spatial objects

eg Regions of temperature & precipitation will be grouped into the same cell & the measure so formed contains a collection of pointers to those regions.

Mining Spatial Association & Co-location patterns

A spatial Association rule is of the form $A \Rightarrow B$ [s%, c%] where A & B are the sets of spatial or nonspatial predicates

eg $is\_a(x, \text{"school"}) \wedge close\_to(x, \text{"sports\_centu"}) \Rightarrow close\_to(x, \text{"park"})$ [0.5%, 80%]

Spatial clustering Methods

spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement in a large, multidimensional data set.

Spatial classification & spatial Trend Analysis

spatial classification analyzes spatial objects to derive classification schemes in relevance to certain spatial properties, such as neighborhood of a district, highway or river

Mining Raster Databases

spatial database system usually handle vector data that consist of points, lines, polygons & their compositions, such as network or partitions.

eg: Data include maps, design graphs, & 3-D representations of the arrangement of chains of proteen molecules.

UNIT V

WEB & TEXT MINING

5.1 Introduction

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what we're familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently isn't relevant to your needs in order to find the relevant information. In text mining, the goal is to discover heretofore unknown information, something that no one yet knows and so could not have yet written down.

Data mining versus Text mining

The difference between regular data mining and text mining is that in text mining
The patterns are extracted from natural language text rather than from structured databases of facts. Databases are designed for programs to process automatically; text is written for people to read. We do not have programs that can "read" text and will not have such for the foreseeable future. Many researchers think it will require a full simulation of how the mind works before we can write programs that read the way people do.

However, there is a field called **computational linguistics** (also known as natural language processing) which is making a lot of progress in doing small subtasks in text analysis. For example, it is relatively easy to write a program to extract phrases from an article or book that, when shown to a human reader, seem to summarize its contents. (The most frequent words and phrases in this article, minus the really common words like "the" are: text mining, information, programs, and example, which is not a bad five-word summary of its contents.)

Typical applications of text Mining could include Analyzing open-ended survey responses. For example, you may discover a certain set of words or terms that are commonly used by respondents to describe the pro's and con's of a product or service (under investigation), suggesting common misconceptions or confusion regarding the items in the study.

Another application include to aid in the automatic classification of texts. For example, it is possible to "filter" out automatically most undesirable "junk email" based on certain terms or words that are not likely to appear in legitimate messages, but instead identify undesirable electronic mail. In this manner, such messages can automatically be discarded. Such automatic systems for classifying electronic messages can also be useful in applications where messages need to be routed (automatically) to the most appropriate department or agency; e.g., email messages with complaints or petitions to a municipal authority are automatically routed to the appropriate departments; at the same time, the emails are screened

for inappropriate or obscene messages, which are automatically returned to the sender with a request to remove the offending words or content.

Text Mining Algorithm consist of 3 steps.

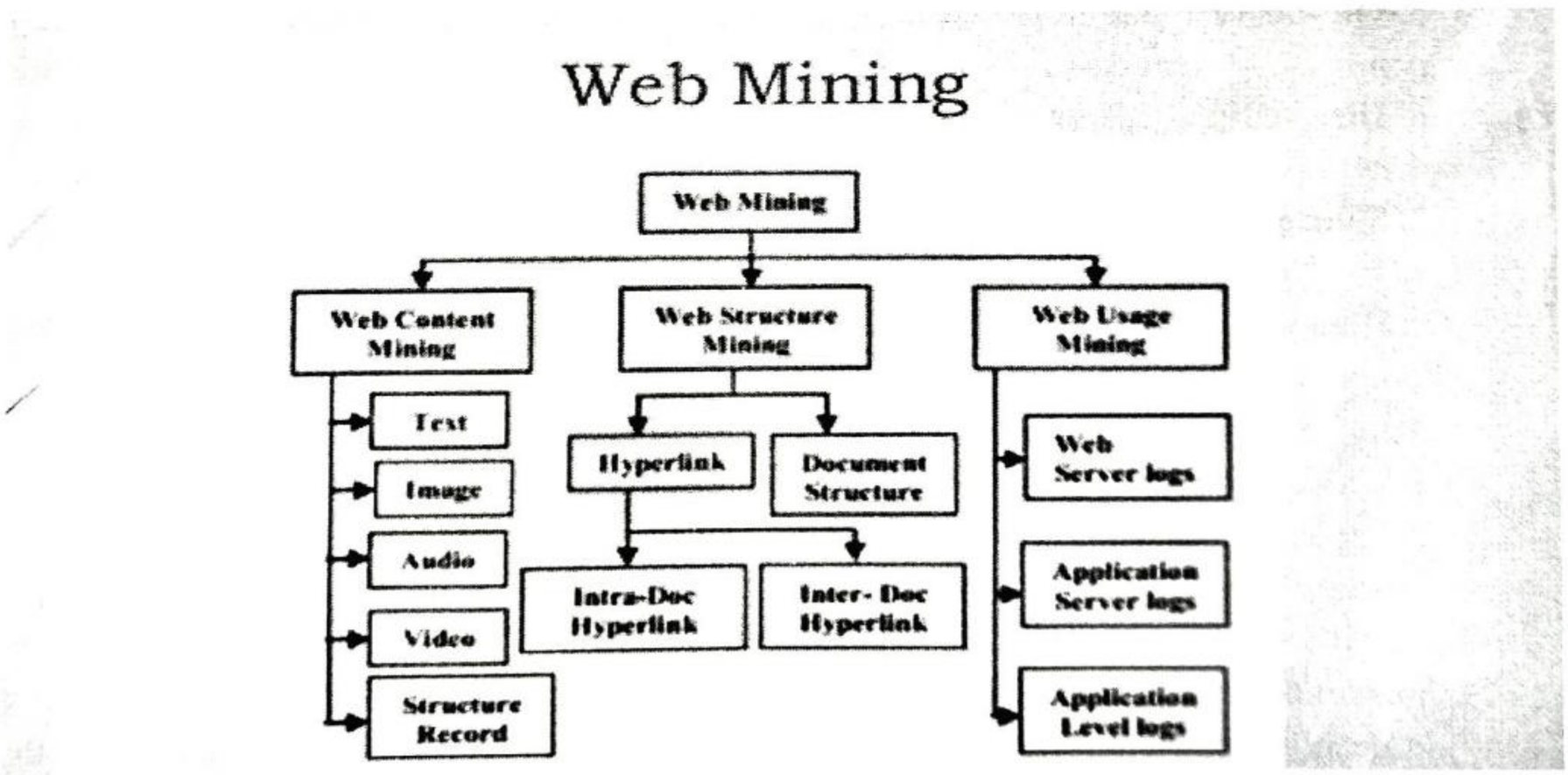
1.Train.create attribute dictioary where the attribute represents words from articles related to a particular topic. Choose only words that occur a minimum number of times.
2. **Filter.** Remove the common words known to be useless in the differentiating articles. Eg. The,As, We etc.
3. **Classify.** Check each document to be classified for the presence and frequency of the chosen attributes. Classify the document under a particular topic if it contains a predetermined minimum number of references to the chosen attributes for the topic.

5.1 Web Mining

Web mining is the process which includes various data mining techniques to extract knowledge from web data categorized as web content, web structure and data usage. It includes a process of discovering the useful and unknown information from the web data.

Web mining can be classified based on the following categories:

1. Web Content
2. Web Structure
3. Web Usage



Let's understand concepts of various categories included in web mining.

5.1.1 Web Content Minings

Web content mining is defined as the process of converting raw data to useful information using the content of web page of a specified web site. The process starts with the extraction of structured data or information from web pages and then identifying similar data with integration. Various types of web content include text, audio, video etc. This process is called as text mining. Text Mining uses Natural Language processing and retrieving information techniques for a specific mining process.

5.1.2 Web Structure Mining

Web graphs include a typical structure which consists of web pages such as nodes and hyperlinks which will be treated as edges connected between web pages. It includes a process of discovering a specified structure with information from the web.
This category of mining can be performed either at document level or hyperlink level. The researsch activity which involves hyperlink level is called hyperlink analysis.

Terminologies associated with Web structure:

1.Webgraph: It is a directed graph which represents the web.
2.Node: Each web page includes a node of the web graph.
3.Link: Hyperlink is a type of directed edge of the web graph.
4. In-degree: In-degree specifies the number of distinct links that point to a specified node.
5. Out-degree: Out-degree specifies the number of distinct lakes originating at a node that points to othernodes.
6. Directed path: Directed path includes a sequence of links starting from a specified node that can befollowedtoreachanothernode.
7. Shortest Path: The shortest path will be the shortest length out of all the paths between p and q.
8. Diameter: The maximum of the shortest path between a pair of nodes p and q for all pairs of nodes p and q in the web graph.

Application of Web Content and Web Structure Mining

Structure mining can aid to this goal, by identifying popular sites (so-called 'authorities'), for example, by analysing the number of links that refer to a particular site. Web content and structure mining are not only used to improve the quality of public search engines. Special search services can also be offered. Content and structure mining tools can for instance track down online misuse of brands , or analyse the content and structure of competitive web sites in detail to gain some strategic advantage . With content and structure mining tools, things like online curriculum vitae or personal home pages can be collected. After interpreting the personal data found on personal pages this information could be used for marketing purposes. Profiles on potential customers can be produced and more detailed information is added to profiles of current customers. So mining the web not only contributes to acquiring new customers, it can also aid in retaining existing ones.

5.1.3 Web Usage Mining:

Web includes a collection of interrelated files with one or more web servers. It includes a pattern of discovery of meaningful patterns of data generated by the client-server transaction.

The typical sources of data are mentioned below:

1. Data which is generated automatically is stored in server access logs, referrer logs, agent logs and client-sidecookies.
2. Informationofuserprofiles.
3. Metadata which includes page attributes and content attributes.

Application of web usage mining

Using web usage mining, it can extract useful information from the clickstream analysis of web server log containing details of webpage visits, transactions. Web server log analyzer may include software such as NetTracker, AwStats to view how often is the website visited, which kind of product is the best and worst sellers in a e-commerce website. The ability to track web users' browsing behaviour down to individual mouse clicks makes it possible to personalise services for individual customers on a massive scale. This 'mass customisation' of services not only helps customers by satisfying their needs, but also results in customer loyalty. Due to a more personalised and customer-centred approach, the content and structure of a web site can be evaluated and adapted to the customer's preferences and the right offers can be made to the right customer.

Web server log:

Server logs created by the server record all activities. The page forwarded to the web server includes every single piece of basic information about URL.

5.2 Text Mining

Text and data mining are considered as complementary techniques required for efficient business management. Data mining and text mining tools have gathered its primary location in the marketplace. Natural Language processing is a subset of text mining tools which is used to define accurate and complete domain specific taxonomies. This helps in effective metadata association. Text mining is more mature and efficient in comparison with data mining process. 80 percent of the information is made of text.

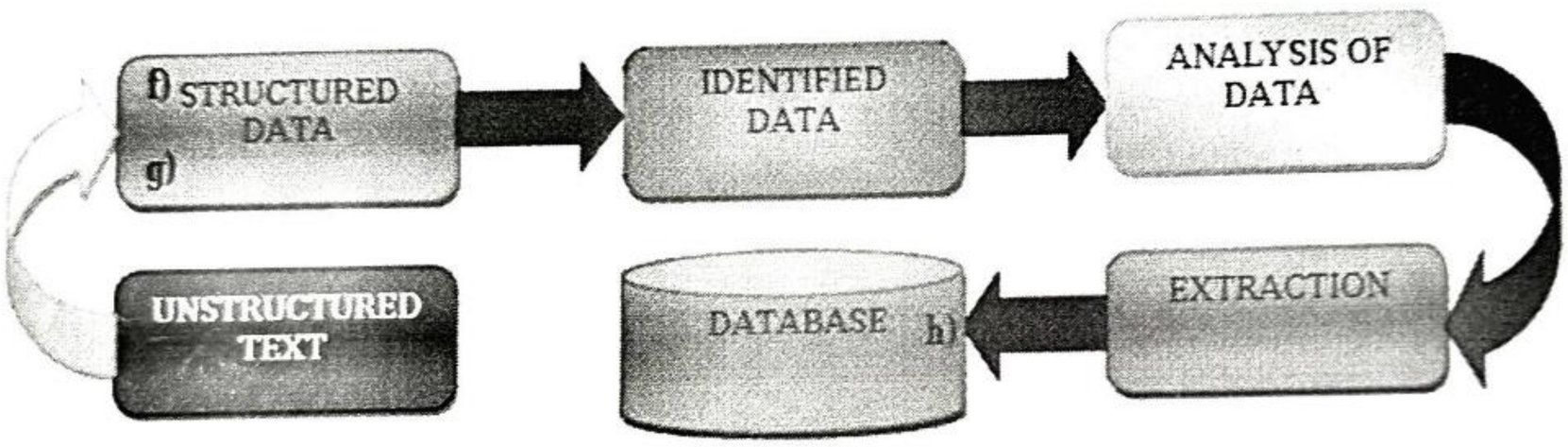
The objective of text mining is to exploit information which is included in textual documents in various patterns and trends in association with entities and predictive rules. The results are manipulated and used for:

1. The analysis of a collection

2. Providing information about intelligent navigation and browsing method.

The five fundamental steps involved in text mining are:

- Gathering unstructured data from multiple data sources like plain text, web pages, pdf files, emails, and blogs, to name a few.
- Detect and remove anomalies from data by conducting pre-processing and cleansing operations. Data cleansing allows you to extract and retain the valuable information hidden within the data and to help identify the roots of specific words.
- Convert all the relevant information extracted from unstructured data into structured formats.
- Analyse the patterns within the data via Management Information System (MIS).
- Store all the valuable information into a secure database to drive trend analysis and enhance the decision-making process of the organisation.



Text Mining Techniques

various text mining techniques:

Information Extraction

Information Extraction (IE) refers to the process of extracting meaningful information from vast chunks of textual data. This method focuses on identifying the extraction of entities, attributes, and their relationships from semi-structured or unstructured texts. Whatever information is extracted is then stored in a database for future access and retrieval. The efficacy and relevancy of the outcomes are checked and evaluated using precision and recall processes.

Information Retrieval

Information Retrieval (IR) refers to the process of extracting relevant and associated patterns based on a specific set of words or phrases. IR systems make use of different algorithms to track and monitor user behaviours and discover relevant data accordingly. Google and Yahoo search engines are the two most renowned IR systems.

Categorization

Text categorization is a form of "supervised" learning wherein normal language texts are assigned to a predefined set of topics depending upon their content. Thus, categorization or

rather Natural Language Processing (NLP) is a process of gathering text documents and processing and analysing them to uncover the right topics or indexes for each document. The co-referencing method is commonly used as a part of NLP to extract relevant synonyms and abbreviations from textual data. Today, NLP has become an automated process used in a host of contexts ranging from personalized commercials delivery to spam filtering and categorizing web pages under hierarchical definitions, and much more.

Clustering

Clustering is one of the most crucial techniques of text mining. It seeks to identify intrinsic structures in textual information and organize them into relevant subgroups or 'clusters' for further analysis. A significant challenge in the clustering process is to form meaningful clusters from the unlabeled textual data without having any prior information on them. Cluster analysis is a standard text mining tool that assists in data distribution or acts as a per-processing step for other text mining algorithms running on detected clusters.

Summarisation

Text summarisation refers to the process of automatically generating a compressed version of a specific text that holds valuable information for the end user. The aim here is to browse through multiple text sources to craft summaries of texts containing a considerable proportion of information in a concise format, keeping the overall meaning and intent of the original documents essentially the same. Text summarisation integrates and combines the various methods that employ text categorization like decision trees, neural networks, regression models, and swarm intelligence.

Applications Of Text Mining

Text mining techniques are rapidly penetrating the industry, right from academia and healthcare to businesses and social media platforms. Here are a few applications of text mining being used across the globe today.

1.Risk Management

One of the primary causes of failure in the business sector is the lack of proper or insufficient risk analysis. Adopting and integrating risk management software powered by text mining technologies such as SAS Text Miner can help businesses to stay updated with all the current trends in the business market and boost their abilities to mitigate potential risks. Since text mining technologies can gather relevant information from across thousands of text data sources and create links between the extracted insights, it allows companies to access the right information at the right moment, thereby enhancing the entire risk management process.

2.Customer care service

Text mining techniques, particularly NLP, are finding increasing importance in the field of customer care. Companies are investing in text analytics software to enhance their overall customer experience by accessing the textual data from varied sources such as surveys, customer feedback, and customer calls, etc. Text analysis aims to reduce the response time of the company and help address the grievances of the customers speedily and efficiently.

3.Fraud Detection

Text analytics backed by text mining technologies provides a tremendous opportunity for domains that gather a majority of data in the text format. Insurance and finance companies are harnessing this opportunity. By combining the outcomes of text analyses with relevant structured data these companies are now able to process claims swiftly as well as detect and prevent frauds.

4.Business Intelligence

Organisations and business firms have started to leverage text mining techniques as a part of their business intelligence. Apart from providing profound insights into customer behaviour and trends, text mining techniques also help companies to analyse the strengths and weaknesses of their rivals, thus, giving them a competitive advantage in the market. Text mining tools such as Cogito Intelligence Platform and IBM text analytics provide insights on the performance of marketing strategies, latest customer and market trends, and so on.

5.Social Media Analysis

There are many text mining software packages designed exclusively for analysing the performance of social media platforms. These help to track and interpret the texts generated online from the news, blogs, emails, etc. Furthermore, text mining tools can efficiently analyse the number of posts, likes, and followers of your brand on social media, thereby allowing you to understand the reaction of people who are interacting with your brand and online content. The analysis will enable you to understand 'what's hot and what's not' for your target audience.

5.3. Unstructured text

Unstructured data is information, in many different forms, that doesn't hew to conventional data models and thus typically isn't a good fit for a mainstream relational database. Thanks to the emergence of alternative platforms for storing and managing such data, it is increasingly prevalent in IT system and is used by organizations in a variety of business intelligence and analyticsapplications.

Traditional structured data, such as the transaction data in financial systems and other business applications, conforms to a rigid format to ensure consistency in processing and analyzing it. Sets of unstructured data, on the other hand, can be maintained in formats that aren't uniform, freeing analytics teams to work with all of the available datawithout necessarily having to consolidate and standardize it first. That enables more comprehensive analyses than would otherwise be possible.

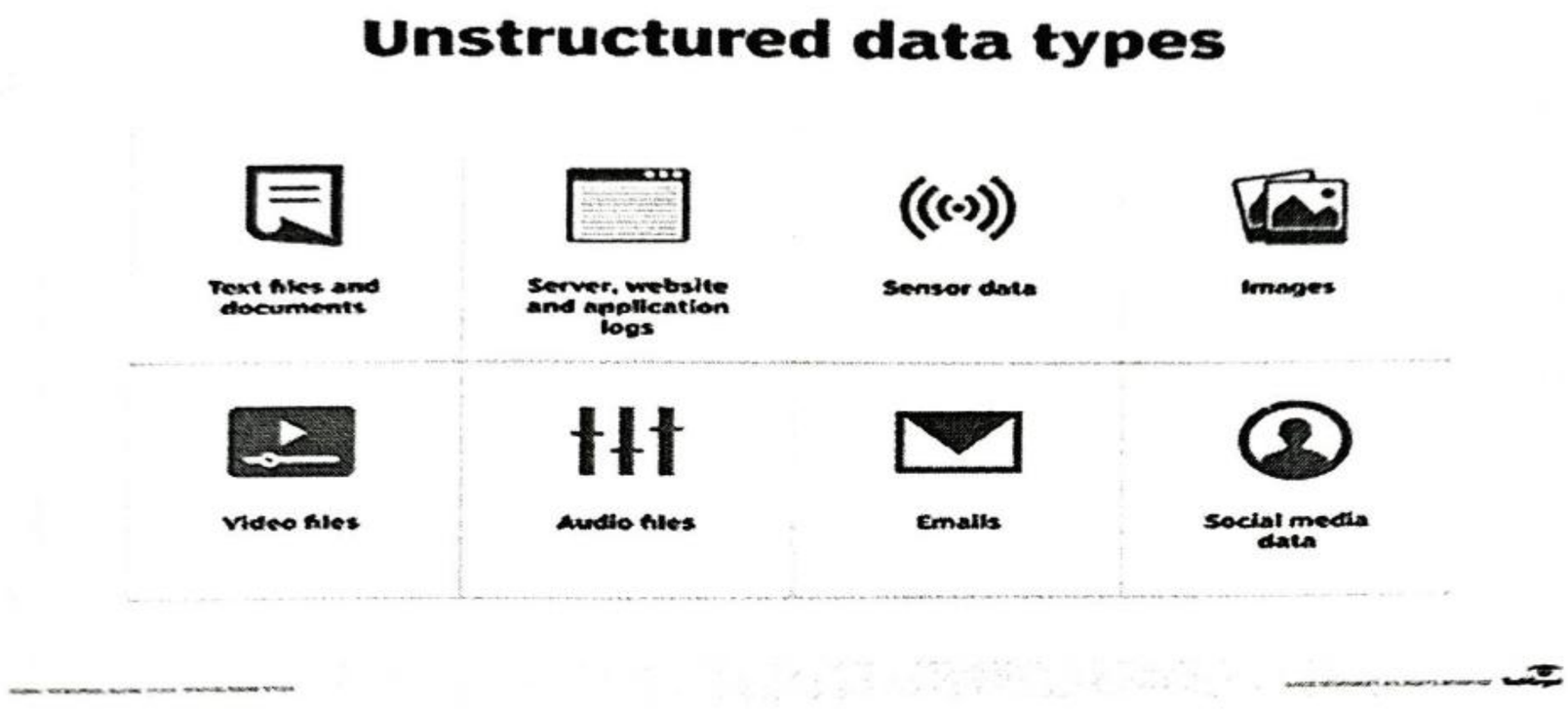
Types of unstructured data

One of the most common types of unstructured data is text. Unstructured text is generated and collected in a wide range of forms, including Word documents, email messages, PowerPoint presentations, survey responses, transcripts of call center interactions, and posts from blogs and social media sites.

Other types of unstructured data include images, audio and video files. Machine data is another category, one that's growing quickly in many organizations. For example, log files from websites, servers, networks and applications -- particularly mobile ones -- yield a trove of activity and performance data. In addition, companies increasingly capture and

analyze data from sensors on manufacturing equipment and other internet of things (IoT) connected devices.

In some cases, such data may be considered to be semi-structured -- for example, if metadatatags are added to provide information and context about the content of the data. The line between unstructured and semi-structured data isn't absolute, though; some data management consultants contend that all data, even the unstructured kind, has some level of structure.

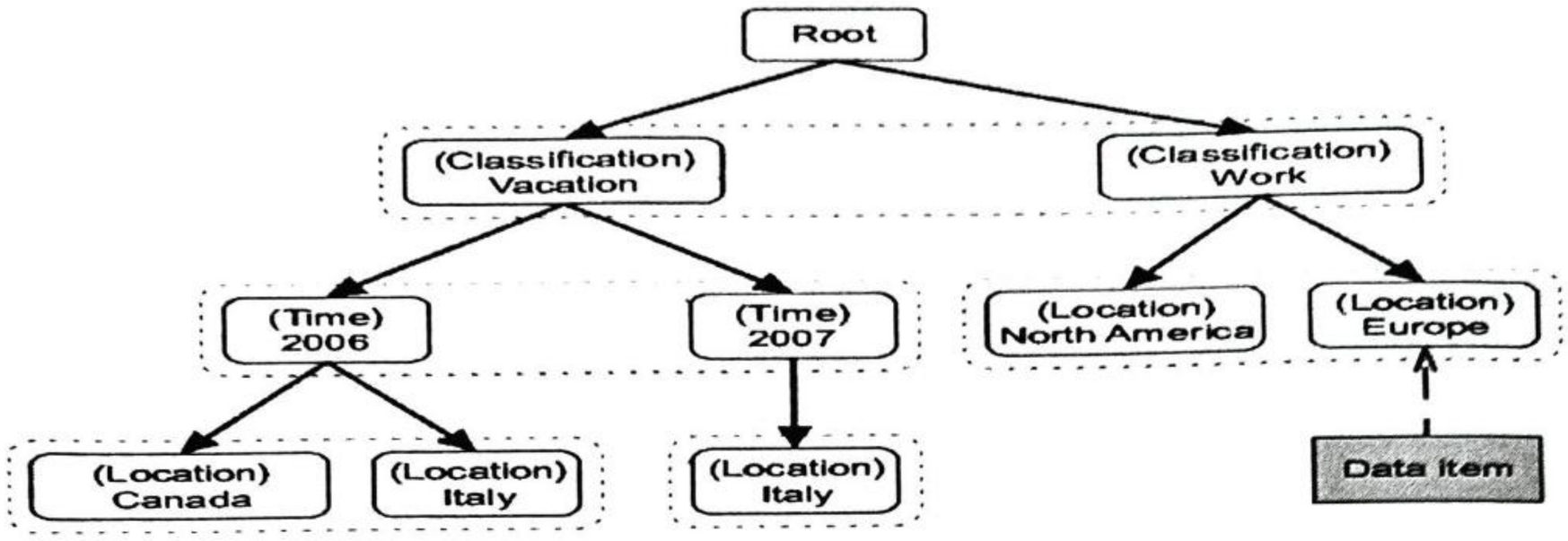


5.3.1. Episode rule discovery for texts

Episode rules are event patterns mined from a single event sequence. They are mainly used to predict the occurrence of events (the consequent of the rule), once the antecedent has occurred. The occurrence of the consequent of a rule may however be disturbed by the occurrence of another event in the sequence (that does not belong to the antecedent). We refer such an event to as an influencer event. To the best of our knowledge, the identification of such events in the context of episode rules has never been studied. However, identifying influencer events is of the highest importance as these events can be viewed as a way to act to impact the occurrence of events, here the consequent of rules. We propose to identify three types of influencer events: distance influencer events, confidence influencer events and disappearance events.

5.3.2. Hierarchy of categories

Concept hierarchies may also be defined by discretizing or grouping values for a given dimension or attribute, resulting in a **set-grouping hierarchy**. A total or partial order can be defined among groups of values. An example of a set-grouping hierarchy is shown in figure for the dimension Root, where an interval (vacation...Work) denotes the range from vacation (exclusive) to Work (inclusive).



5.4 Text clustering

Text clustering is the application of cluster analysis to text-based documents. It uses machine learning and natural language processing (NLP) to understand and categorize unstructured, textual data.