CLUSTERING, APPLICATIONS AND TRENDS IN DATA MINING

Cluster analysis - Types of data in Cluster Analysis - Categorization of major clustering methods - Partitioning methods - K Means - K Medoids - Hierarchical methods - Density-based methods - Grid-based methods - Model based clustering methods - Constraint Based cluster analysis - Outlier analysis - Data Mining Spatial Applications.

CLUSTER ANALYSIS

- Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.

What is Clustering?

- Clustering is the process of making a group of abstract objects into classes of similar objects.

Points to Remember

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Requirements of Clustering in Data Mining

The following points throw light on why clustering is required in data mining –

- **Scalability** – We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes** – Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape** – The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality** – The clustering algorithm should not only be able to handle low- dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** – Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** – The clustering results should be interpretable, comprehensible, and usable.

Types of Data in Cluster Analysis

- **Interval – scaled variables**
  - e.g. salary, height
- **Binary variables**
  - e.g. gender (M/F). has_canoer (T/F)
- **Nominal (Categorical) Variables**
  - e.g. religion (Christian, Muslim, Buddhist, Hindu, etc.)
- **Ordinal variables**
  - e.g., military rank (soldier, sergeant, lutenant, captain, etc.)
- **Ratio-scaled variables**
  - Population growth (1,10,100,1000….)
- **Variables of mixed types**
  - Multiple attributes with various types

CATEGORIZATION OF MAJOR CLUSTERING METHODS

- Clustering methods can be classified into the following categories –
  - Partitioning Method
  - Hierarchical Method
  - Density-based Method
  - Grid-Based Method
  - Model-Based Method
  - Constraint-based Method

PARTITIONING METHOD

- Suppose we are given a database of ‘n’ objects and the partitioning method constructs ‘k’ partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements –
  - Each group contains at least one object.
  - Each object must belong to exactly one group.

Points to remember

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

Hierarchical Methods

- This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here –
  - Agglomerative Approach
  - Divisive Approach

Agglomerative Approach

- This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach

- This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Approaches to Improve Quality of Hierarchical Clustering

- Here are the two approaches that are used to improve the quality of hierarchical clustering –
  - Perform careful analysis of object linkages at each hierarchical partitioning.
  - Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro- clusters.

Density-based Method

- This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

Grid-based Method

- In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantage

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

Model-based methods

- In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.
- This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

Constraint-based Method

- In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

PARTITIONING METHODS

K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closet centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

1: Select *K* point as the initial

2: **repeat**

3: Form *K* clusters by assigning all points to the closet centroid.

4: Recompute the centroid of each cluster.

5: **until** The centroids don't change

HIERARCHICAL METHODS

- Two main types of hierarchical cluster

- **Agglomerative:**
  - Start with the points as individual clusters
  - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left

- **Divisine:**
  - Start with one, all-inclusive cluster
  - At each step, split a cluster until each cluster contains a point (or there are k clusters)

- Traditional hierarchical algorithms use a similarity or distance matrix
  - Merge or split one cluster at a time

AGGLOMERATIVE CLUSTERING ALGORITHM

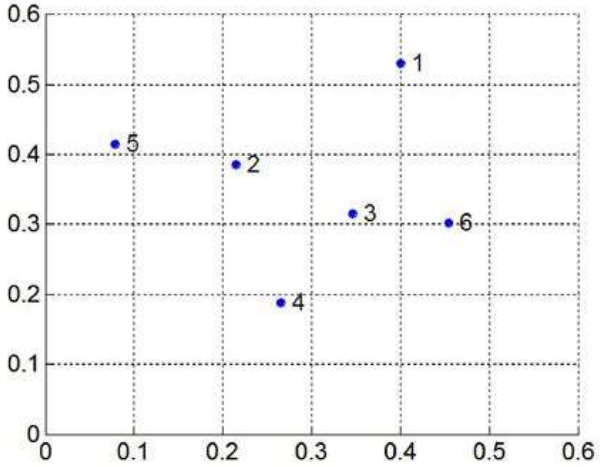
- More popular hierarchical clustering technique
- Basic algorithm is straightforward

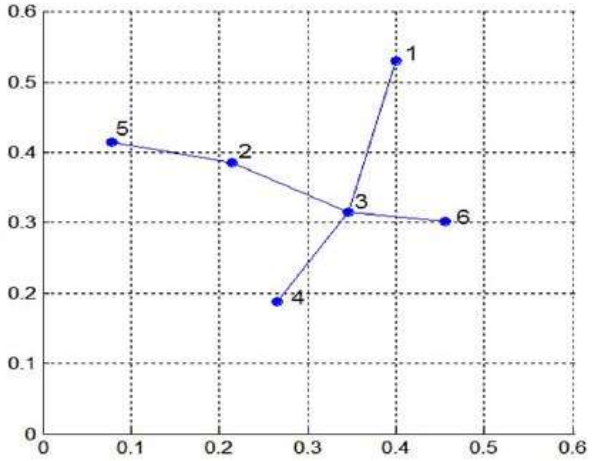
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the proximity matrix
  6. **Until** only a single cluster remains

- Key operation is the computation of the proximity of two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms.

MST: Divisive Hierarchical Clustering

- Build MST (Minimum Spanning Tree)
  - Start with a tree that consists of any point
  - In successive steps, look for the closest pair of points (p,q) such that one point (p) is in the current tree but the other (q) is not
  - Add q to the tree and put an edge between p and q





MST: Divisive Hierarchical Clustering

- Use MST for constructing hierarchy of clusters

Algorithm 7.5 MST Divisine Hierarchical Clustering Algorithm

1: Compare a minimum spanning tree for the proximity graph.
2: repeat
3: Create a new cluster by breaking the link corresponding to the largest distance (smallest similarity).
4: **until:** Only singleton cluster remain

Data Mining Applications

- Data mining is widely used in diverse areas. There are a number of commercial data mining system available today and yet there are many challenges in this field. In this tutorial, we will discuss the applications and the trend of data mining.

APPLICATIONS

- Here is the list of areas where data mining is widely used –
  - Financial Data Analysis
  - Retail Industry
  - Telecommunication Industry
  - Biological Data Analysis
  - Other Scientific Applications
  - Intrusion Detection

1. Data Mining for Financial Data Analysis

- Design and construction of data warehouses for multidimensional data analysis and data mining
- Loan payment prediction and customer credit policy analysis
- Classification and clustering of customers for targeted marketing
- Detection of money laundering and other financial crimes
- Data Mining for the Retail Industry

2. A few examples of data mining in the retail industry

- Design and construction of data warehouses based on the benefits of data mining
- Multidimensional analysis of sales, customers, products, time, and region
- Analysis of the effectiveness of sales campaigns
- Customer retention—analysis of customer loyalty
- Product recommendation and cross-referencing of items

3. Data Mining for the Telecommunication Industry

- Multidimensional analysis of telecommunication data
- Fraudulent pattern analysis and the identification of unusual patterns
- Multidimensional association and sequential pattern analysis
- Mobile telecommunication services
- Use of visualization tools in telecommunication data analysis

4. Data Mining for Biological Data Analysis

- Semantic integration of heterogeneous, distributed genomic and proteomic databases
- Alignment, indexing, similarity search, and comparative analysis of multiple nucleotide , protein sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis: identifying co-occurring gene sequences and linking genes to different stages of disease development.
- Visualization tools in genetic data analysis.

5. Data Mining in Scientific Applications

- Scientific data can be amassed at much higher speeds and lower costs.
- This has resulted in the accumulation of huge volumes of high-dimensional data, stream data, and heterogeneous data, containing rich spatial and temporal information.
- Scientific applications are shifting from the "hypothesize-and-test" paradigm toward a "collect and store data, mine for new hypotheses, confirm with data or experimentation" process.

6. Data Mining for Intrusion Detection

- Development of data mining algorithms for intrusion detection
- Association and correlation analysis, and aggregation to help select and build discriminating attributes
- Analysis of stream data
- Distributed data mining
- Visualization and querying tools

7. Trends in Data Mining

- Application exploration
- Scalable and interactive data mining methods
- Integration of data mining with database systems, data warehouse systems, and Webdatabase systems
- Standardization of data mining language
- Visual data mining
- Biological data mining
- Data mining and software engineering
- Web mining
- Distributed data mining
- Real-time or time-critical data mining
- Graph mining, link analysis, and social network analysis
- Multi relational and multi database data mining
- New methods for mining complex types of data
- Privacy protection and information security in data mining

8. Assessment of a Data mining System

- Data types
- System issues
- Data sources
- Data mining functions and methodologies
- Coupling data mining with database and/or data warehouse systems.
- Scalability
- Visualization tools
- Data mining query language and graphical user interface

9. Theoretical Foundations of Data Mining

- Data reduction
- Data compression
- Pattern discovery
- Probability theory
- Microeconomic view
- Inductive databases

10. Statistical Data Mining techniques

- Regression
- Generalized linear model
- Analysis of variance
- Mixed effect model
- Factor analysis
- Discriminate analysis
- Time series analysis
- Survival analysis
- Quality control

11. Visual and Audio Data Mining

- Visual data mining discovers implicit and useful knowledge from large data sets using data and/or knowledge visualization
- Data visualization and data mining can be integrated in the following ways:
- Data visualization
- Data mining result visualization
- Data mining process visualization
- Interactive visual data mining techniques

12. Security of Data Mining

- Data security enhancing techniques have been developed to help protect data
- Databases can employ a multilevel security model to classify and restrict data according to various security levels, with users permitted access to only their authorized level
- Privacy-sensitive data mining deals with obtaining valid data mining results without learning the underlying data values

Social Impacts of Data Mining

1. **Is Data Mining a Hype or Will It Be Persistent?**
   - ❑ Data mining is a technology
   - ❑ Technological life cycle
   - ❑ Innovators
   - ❑ Early Adopters
   - ❑ Early Adopters
   - ❑ Chasm
   - ❑ Early Majority
   - ❑ Late Majority
   - ❑ Laggards

2. Data Mining: Managers' Business or Everyone's?

- Data mining will surely be an important tool for managers' decision making
- Bill Gates: "Business @ the speed of thought"
- The amount of the available data is increasing, and data mining systems will be more affordable
- Multiple personal uses
- Mine your family's medical history to identify genetically-related medical conditions
- Mine the records of the companies you deal with
- Mine data on stocks and company performance, etc.
- Invisible data mining
- Build data mining functions into many intelligent tools

3. Social Impacts: Threat to Privacy and Data Security?

- Is data mining a threat to privacy and data security?
- "Big Brother", "Big Banker", and "Big Business" are carefully watching you
- Profiling information is collected every time
- credit card, debit card, supermarket loyalty card, or frequent flyer card, or apply for any of the above
- You surf the Web, rent a video, fill out a contest entry form,
- You pay for prescription drugs, or present you medical care number when visiting the doctor
- Collection of personal data may be beneficial for companies and consumers, there is also potential for misuse
- Medical Records, Employee Evaluations, etc.

4. Protect Privacy and Data Security

1. Fair information practices

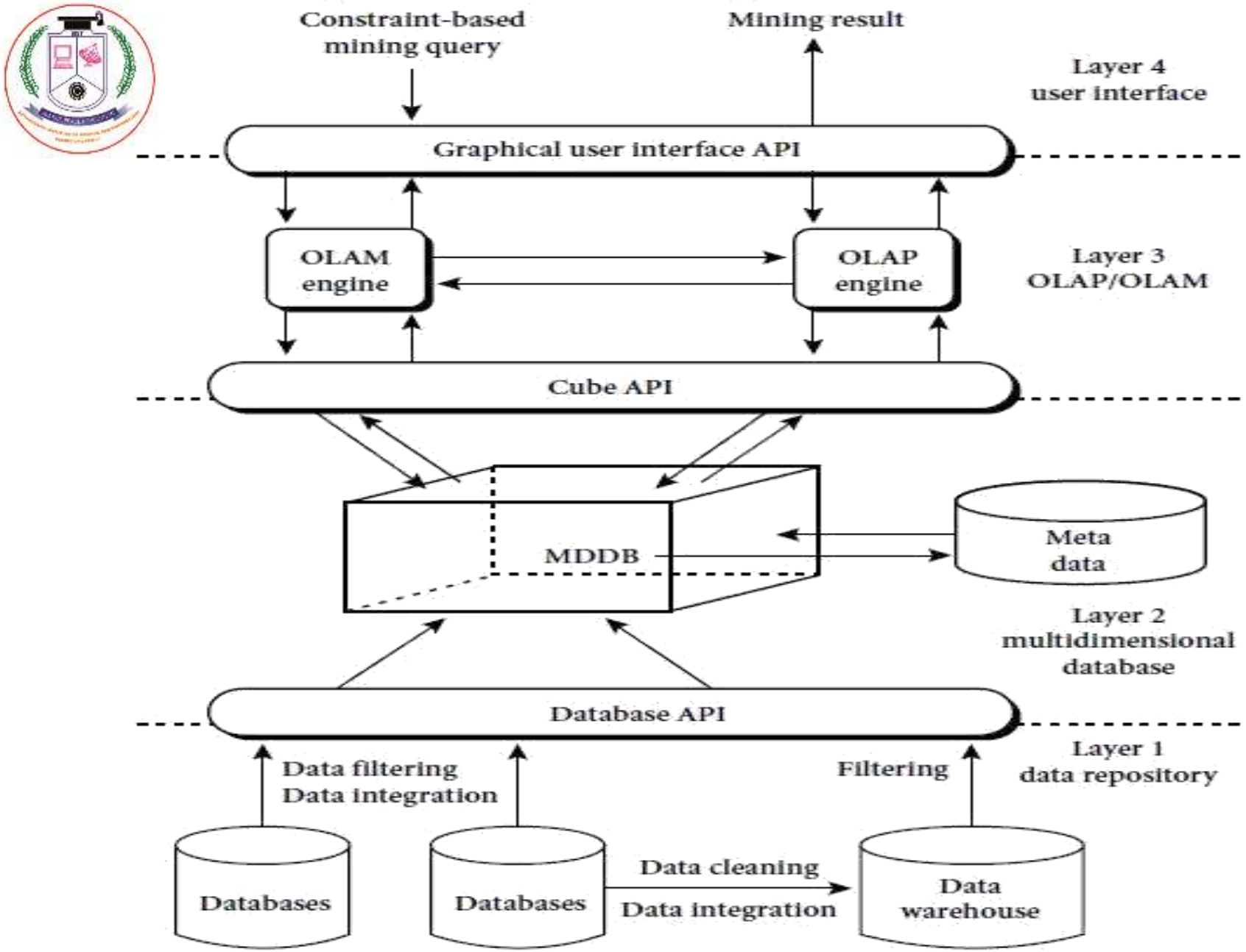
- International guidelines for data privacy protection
- Cover aspects relating to data collection, purpose, use, quality, openness, individual participation, and accountability
- Purpose specification and use limitation
- Openness : Individuals have the right to know what information is collected about them, who has access to the data, and how the data are being used

2. Develop and use data security-enhancing techniques

- Blind signatures
- Biometric encryption
- Anonymous databases

An Introduction to DB Miner

An OLAM server performs analytical mining in data cubes in a similar manner as an OLAP server performs on-line analytical processing. An integrated OLAM and OLAP architecture is shown in Figure, where the OLAM and OLAP servers both accept user on-line queries (or commands) via a graphical user interface API and work with the data cube in the data analysis via a cube API. A meta data directory is used to guide the access of the data cube. The data cube can be constructed by accessing and/or integrating multiple databases via an MDDB API and/or by filtering a data warehousing via a database API that may support OLE DB or ODBC connections. Since an OLAM server may perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, time-series analysis, and so on, it usually consists of multiple integrated data mining modules and is more sophisticated than an OLAP server.



An integrated OLAM and OLAP architecture.

Mining WWW (World Wide Web)

- The World Wide Web contains huge amounts of information that provides a rich source for data mining.

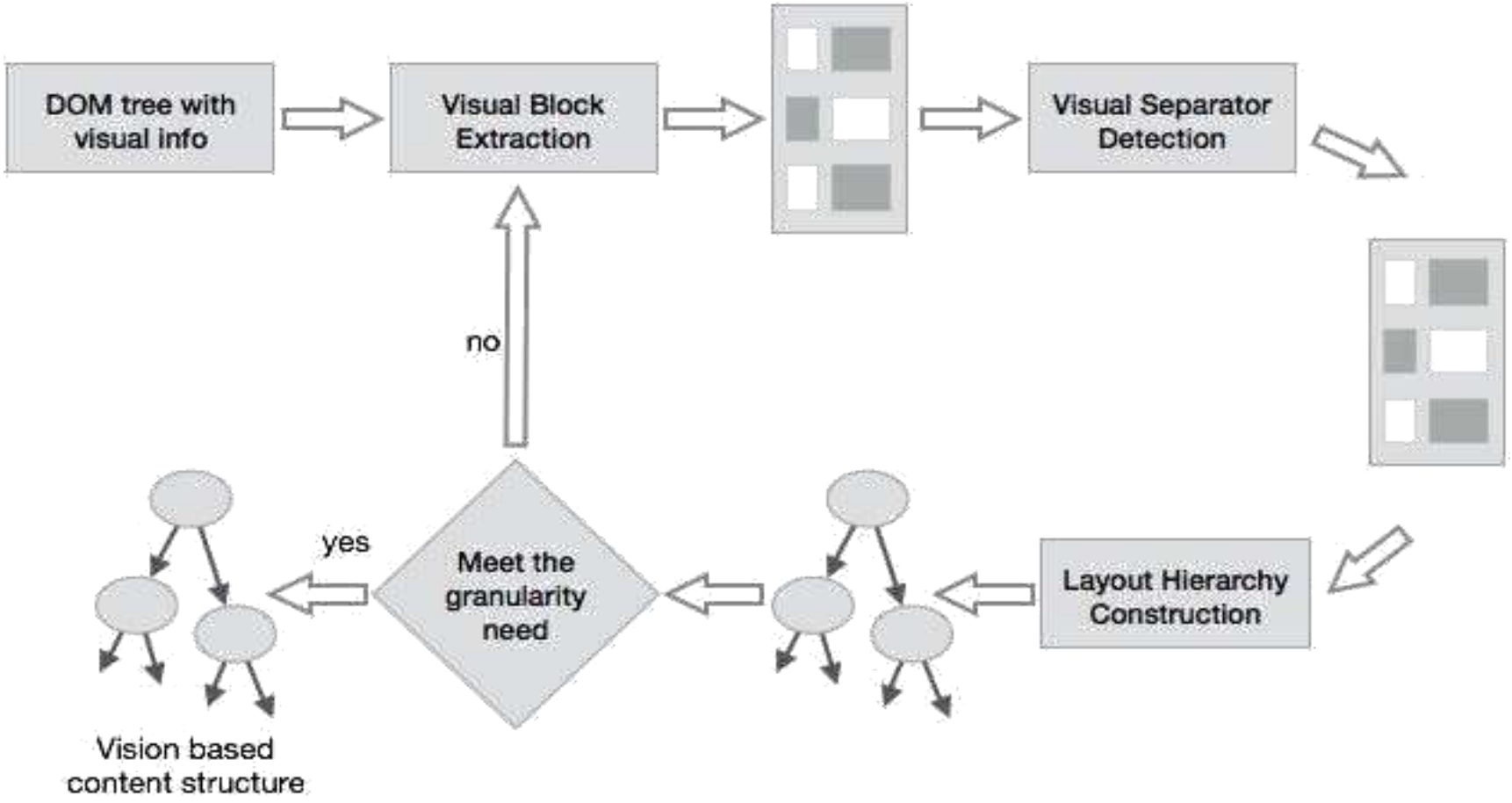
Challenges in Web Mining

- The web poses great challenges for resource and knowledge discovery based on the following observations −
- **The web is too huge** − The size of the web is very huge and rapidly increasing. This seems that the web is too huge for data warehousing and data mining.
- **Complexity of Web pages** − The web pages do not have unifying structure. They are very complex as compared to traditional text document. There are huge amount of documents in digital library of web. These libraries are not arranged according to any particular sorted order.
- **Web is dynamic information source** − The information on the web is rapidly updated. The data such as news, stock markets, weather, sports, shopping, etc., are regularly updated.

Vision-based page segmentation (VIPS)

- The purpose of VIPS is to extract the semantic structure of a web page based on its visual presentation.
- Such a semantic structure corresponds to a tree structure. In this tree each node corresponds to a block.
- A value is assigned to each node. This value is called the Degree of Coherence. This value is assigned to indicate the coherent content in the block based on visual perception.
- The VIPS algorithm first extracts all the suitable blocks from the HTML DOM tree. After that it finds the separators between these blocks.
- The separators refer to the horizontal or vertical lines in a web page that visually cross with no blocks.
- The semantics of the web page is constructed on the basis of these blocks

The following figure shows the procedure of VIPS algorithm –



Text Mining

Mining Text Data

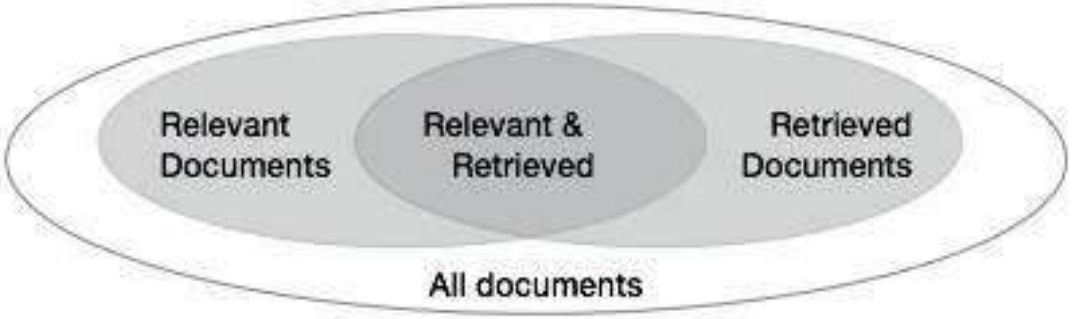
- Text databases consist of huge collection of documents. They collect these information from several sources such as news articles, books, digital libraries, e-mail messages, web pages, etc. Due to increase in the amount of information, the text databases are growing rapidly. In many of the text databases, the data is semi-structured.
- For example, a document may contain a few structured fields, such as title, author, publishing_date, etc. But along with the structure data, the document also contains unstructured text components, such as abstract and contents. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users require tools to compare the documents and rank their importance and relevance. Therefore, text mining has become popular and an essential theme in data mining.

Information Retrieval

- Information retrieval deals with the retrieval of information from a large number of text-based documents. Some of the database systems are not usually present in information retrieval systems because both handle different kinds of data. Examples of information retrieval system include −
  - Online Library catalogue system
  - Online Document Management Systems
  - Web Search Systems etc.
- **Note** − The main problem in an information retrieval system is to locate relevant documents in a document collection based on a user's query. This kind of user's query consists of some keywords describing an information need.
- In such search problems, the user takes an initiative to pull relevant information out from a collection. This is appropriate when the user has ad-hoc information need, i.e., a short-term need. But if the user has a long-term information need, then the retrieval system can also take an initiative to push any newly arrived information item to the user.
- This kind of access to information is called Information Filtering. And the corresponding systems are known as Filtering Systems or Recommender Systems.

Basic Measures for Text Retrieval

- We need to check the accuracy of a system when it retrieves a number of documents on the basis of user's input. Let the set of documents relevant to a query be denoted as {Relevant} and the set of retrieved document as {Retrieved}. The set of documents that are relevant and retrieved can be denoted as {Relevant} ∩ {Retrieved}. This can be shown in the form of a Venn diagram as follows –



- There are three fundamental measures for assessing the quality of text retrieval –
  - Precision
  - Recall
  - F-score

Precision

- Precision is the percentage of retrieved documents that are in fact relevant to the query. Precision can be defined as –

  Precision= |{Relevant} ∩ {Retrieved}| / |{Retrieved}|

Recall

- Recall is the percentage of documents that are relevant to the query and were in fact retrieved. Recall is defined as –

  Recall = |{Relevant} ∩ {Retrieved}| / |{Relevant}|

F-score

- F-score is the commonly used trade-off. The information retrieval system often needs to trade-off for precision or vice versa. F-score is defined as harmonic mean of recall or precision as follows

  F-score = recall x precision / (recall + precision) / 2

MINING SPATIAL DATABASES

- A spatial database stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, and VLSI chip layout data. Spatial databases have many features distinguishing 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 an integration of data mining with spatial database technologies. It can be used for understanding spatial data, discovering spatial relationships and relationships between spatial and non spatial data, constructing spatial knowledge bases, reorganizing spatial databases, and optimizing spatial queries.

- It is expected to have wide applications in geographic information systems, geomarketing, remote sensing, image database exploration, medical imaging, navigation, traffic control, environmental studies, and many other areas where spatial data are used. A crucial challenge to spatial data mining is the exploration of efficient spatial data mining techniques due to the huge amount of spatial data and the complexity of spatial data types and spatial access methods.
- "What about using statistical techniques for spatial data mining?" 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 handles non spatial data, one usually assumes statistical independence among different portions of data.

- However, different from traditional data sets, there is no such independence among spatially distributed data because data because in reality, spatial objects are often interrelated, ore more exactly spatially co-located, in the sense that the closer the two objects are located, the more likely they share similar properties. For example, nature resource, climate, temperature, and economic situations are likely to be similar in geographically closely located regions. People even consider this as the first law of geography: “Everything is related to everything else, but nearby things are more related than distant things”. 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 good success. Spatial data mining will further develop spatial statistical analysis methods and extend them for huge amounts of spatial data, with more emphasis on efficiency, scalability, cooperation with database and data warehouse systems, improved user interaction, and the discovery of new types of knowledge.

Spatial Data Cube Construction and Spatial OLAP

- "Can we construct a spatial data warehouse?" Yes, as with relational data, we can integrate spatial data to construct a data warehouse that facilitates spatial data mining. A spatial data warehouse is a subject-oriented, integrated, time-variant, and non volatile collection of both spatial and non spatial data in support of spatial data mining and spatial data related decision-making processes.
- Let's look at the following example.
- **Spatial data cube and spatial OLP.** There are about 3,000 weather probes distributed I British Columbia (BC), Canada, each recording daily temperature and precipitation for a designated small area and transmitting signals to a provincial weather station. With a spatial data warehouse that supports spatial OLAP, a user can view weather patterns on a map by month, by region, and by different combinations of temperature and precipitation, and can dynamically drill down or roll up along any dimension to explore desired patterns, such as "wet and hot regions in the Fraser Valley in Summer 1999".

- There are several challenging issues regarding the construction and utilization of spatial data warehouses. The first challenge is the integration of spatial data from heterogeneous sources and systems. Spatial data are usually stored in different industry firms and government agencies using various data formats. Data formats are not only structure-specific (e.g. , raster - vs. vector-based spatial data, object-oriented vs. relational models, different spatial storage and indexing structures), but also vendor-specific (e.g, ESRI, MapInfo, Intergraph). There has been a great deal of work on the integration and exchange of heterogeneous spatial data, which has paved the way for spatial data integration and spatial data warehouse construction.
- The second challenge is the realization of fast and flexible on-line analytical processing in spatial data warehouses. The star schema model is a good choice for modeling spatial data warehouses because it provides a concise and organized warehouse structure and facilitates OLAP operations. However, in a spatial warehouse, both dimensions and measures may contain spatial components.

There are three types of dimensions in a spatial data cube

- A **non dimension** contains only non spatial data. Non spatial dimensions temperature and precipitation can be constructed for the warehouse since each contains non spatial data whose generalizations are non spatial (such as “hot” for temperature and “wet” for precipitation).
- A **spatial-to-nonspatial dimension** is a dimension whose primitive-level data are spatial but whose generalization, starting at a certain high level, becomes nonspatial. For example, the spatial dimension’s spatial representation of, say, Seattle is generalized to the string “pacific_northwest.” Although “pacific_northwest” is a spatial concept, its representation is not spatial (since, in our example, it is a string). It therefore plays the role of a nonspatial dimension.
- A **spatial-to-spatial dimension** is a dimension whose primitive level and all of its high level generalized data are spatial. For example, the dimension equi_temperature_region contains spatial data, as do all of its generalizations, such as with regions covering 0-5 degrees (Celsius), 5-10 degrees, and so on.

We distinguish two types of measures in a spatial data cube

- A **numerical measure** contains only numerical data. For example, one measure in a spatial data warehouse could be the monthly_revenue of a region, so that a roll-up may compute the total revenue by year, by country, and so on. Numerical measures can be further classified into distributive, algebraic and holistic.
- A **spatial measure** contains a collection of pointers to spatial objects. For example, in a generalization (or roll-up) in the spatial data cube, the regions with the same range of temperature and precipitation will be grouped into the same cell, and the measure so formed contains a collection of pointers to those regions.
- A non spatial data cube contains only non spatial dimensions and numerical measures. If a spatial data cube contains spatial dimensions but no spatial measures, its OLAP operations, such as drilling or pivoting, can be implemented in a manner similar to that for non spatial data cubes.

Mining Spatial Association and Co-location Patterns

- Similar to the mining of association rules in transactional and relational databases, spatial association rules can be mined in spatial databases. A spatial association rule is of the form A => B [s%, c%], where A and B are sets of spatial or non spatial predicates, s% is the support of the rule, and c% is the confidence of the rule. For example, the following is a spatial association rule:

  is_a(X,“school”)^close_to(X,“sports_center”) => close_to(X,“park”) [0.5%, 80%].

- This rule states that 80% of schools that are close to sports centers are also close to parks, and 0.5% of the data belongs to such a case.
- Various kinds of spatial predicates can constitute a spatial association rule. Examples include distance information (such as close_to and far_away), topological relations (like intersect, overlap, and disjoint), and spatial orientations (like left_of and west_of).

- Various kinds of spatial predicates can constitute a spatial association rule. Examples include distance information (such as close_to and far_away), topological relations (like intersect, overlap, and disjoint), and spatial orientations (like left_of and west_of).
- Since spatial association mining needs to evaluate multiple spatial relationships among a large number of spatial objects, the process could be quite costly. An interesting mining optimization methods called progressive refinement can be adopted in spatial association analysis. The method first mines large data sets roughly using a fast algorithm and than improves the quality of mining in a pruned data set using a more expensive algorithm.
- To ensure that the pruned data set covers the complete set of answers when applying the high-quality data mining algorithms at a later stage, an important requirement for the rough mining algorithm applied in the early stage is the superset coverage property: that is, it preserves all of the potential answers.

- In other words, it should allow a false-positive test, which might include some data sets that do not belong to the answer sets, but it should not allow a false-negative test, which might exclude some potential answers.
- For mining spatial associations related to the spatial predicate close_to, we can first collect the candidates that pass the minimum support threshold by
  - Applying certain rough spatial evaluation algorithms, for example, using an MBR structure (which registers only two spatial points rather than a set of complex polygons), and
  - Evaluating the relaxed spatial predictate, g_close_to, which is a generalized close_to covering a broader context that include close_to, touch, and intersect.

- If two spatial objects are closely located, their enclosing MBRs must be closely located, matching g_close_to. However, the reverse is not always true: if the enclosing MBRs are closely located, the two spatial objects may or may not be located so closely. Thus, the MBR pruning is a false-positive testing tool for closeness: only those that pass the rough test need to be further examined using more expensive spatial computation algorithms. With this preprocessing, only the patterns that are frequent at the approximation level will need to be examined by more detailed and finer, yet more expensive, spatial computation.
- Besides mining spatial association rules, one may like to identify groups of particular features that appear frequently close to each other in a geospatial map. Such a problem is essentially the problem of mining spatial co-locations. Finding spatial co-locations can be considered as a special case of mining spatial associations. However, based on the property of spatial autocorrelation, interesting features likely coexist in closely located regions. Thus spatial co-location can be just what one really wants to explore. Efficiency methods can be developed for mining spatial co-locations by exploring the methodologies like Aprori and progressive refinement, similar to what has been done for mining spatial association rules.

Multimedia Data Mining

- "What is a multimedia database?" A multimedia database system stores and manages a large collection of multimedia data, such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text markups, and linkages. Multimedia database systems are increasingly common owing to the popular use of audio-video equipment, digital cameras, CD-ROMs, and the Internet. Typical multimedia database systems include NASA's EOS (Earth Observation System, various kinds of image and audio-video databases, and Internet databases.
- "When searching for similarities in multimedia data, can be search on either the data description or the data content?" That is correct. For similarly searching in multimedia data, we consider two main families of multimedia indexing and retrieval systems:

1. **Description-based retrieval** systems, which build indices and perform object retrieval and
2. **Content-based retrieval** systems, which support retrieval based on the image content, such as color histogram, texture, pattern, image topology, and the shape of objects and their layouts and locations within the image.

- Description-based retrieval is labor-intensive if performed manually. If automated, the results are typically of poor quality. For example, the assignment of keywords to images can be a tricky and arbitrary task. Recent development of Web-based image clustering and classification methods has improved the quality of description-based Web image retrieval, because images surrounded text information as well as Web linkage information can be used to extract proper description and group images describing a similar theme together. Content-based retrieval uses visual features to index images and promotes object retrieval based on feature similarity, which is highly desirable in many applications.
- In a content-based image retrieval system, there are often two kinds of queries: image-sample-based queries and image features specification queries. **Image-sample-based queries** find all the images that are similar to the given image sample. This search compares the feature vector (or signature) extracted from the sample with the feature vectors of images that have already been extracted and indexed in the image database. Based on this comparison, images that are close to the sample image are returned.

- Image feature specification queries specify of sketch image features like color, texture, or shape, which are translated into a feature vector to be matched with the feature vectors of the images in the database. Content-based retrieval has wide applications, including medical diagnosis, weather prediction, TV production, Web search engines for images, and e-commerce. Some systems, such as QBIC (Query By Image Content), support both sample-based and image feature specification queries. The are also systems that support both content-based and description-based retrieval.
- Several approaches have been proposed and studied for similarity-based retrieval in image databases, based on image signature:
- **Color histogram-based signature:** In this approach, the signature of an image include color histograms based on the color composition of an image regardless of its scale or orientation. This method does not contain any information about shape, image topology, or texture. Thus, two images with similar color composition but that contain very different shapes or textures may be identified as similar, although they could be completely unrelated semantically.

- **Multifeature composed signature:** In this approach, the signature of an image includes a composition of multiple features: color histogram, shape, image topology, and texture. The extracted image features are stored as metadata, and images are indexed based on such metadata. Often, separate distance functions, can be defined for each feature and subsequent combined to derive the overall results. Multidimensional content-based search often uses one or a few probe features to search for image containing such (similar) features. It can therefore be used to search for similar images. This is the most popularly used approach in practice.
- **Wavelet-based signature:** This approach uses the dominant wavelet coefficients of an image as its signature. Wavelets capture shape, texture, and image topology information in a single unified framework. This improves efficiency and reduces the need for providing multiple search primitives (unlike the second method above). However, since this method computes a single signature for an entire image, it may fail to identify images containing similar objects where the objects differ in location or size.

- **Wavelet-based signature with region-based granularity:** In this approach, the computation and comparison of signatures are at the granularity of regions, not the entire image. This is based on the observation that similar images may contain similar regions, but a region in one image could be a translation or scaling of matching region in the other. Therefore, a similarity measure between the query image Q and a target image T can be defined in terms of the area of the two images covered by matching pairs of regions from Q and T. Such a region-based similarity search can find images containing similar objects, where these objects may be translated or scaled.

AUDIO AND VIDEO DATA MINING

- Besides still images, an incommensurable amount of audiovisual information is becoming available in digital form, in digital achieves, on the World Wide Web, in broadcast data streams, and in personal and professional databases. This amount is rapidly growing. There are great demands for effect content-based retrieval and data mining methods for audio and video data.

- Typical examples include searching for and multimedia editing of particular video clips in a TV studio, detecting suspicious persons or scenes in surveillance videos, searching for particular events in a personal multimedia repository such as MyLifeBits, discovering patterns and outliers in weather radar recordings, and finding a particular melody or tune in your MP3 audio album.
- To facilitate the recording, search, and analysis of audio and video information from multimedia data, industry and standardization committees have made great strides toward developing a set of standards for multimedia information description and compression. For example, MPEG-k (developed by MPEG: Moving Picture Experts Group) and JPEG are typical video compression schemes. The most recently released MPEG-7, formally named “Multimedia Content Description Interface”, is a standard for describing the multimedia content data. It supports some degree of interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer.

- MPEG-7 is not aimed at any one application in particular, rather, the elements that MPEG-7 standardizes support as broad a range of applications as possible. The audiovisual data description in MPEG-7 includes still pictures, video, graphics, audio, speech, three-dimensional models, and information about how these data elements are combined in the multimedia presentation.
- The MPEG committee standardizes the following elements in MPEG-7: (1) a set of descriptors, where each descriptor defines the syntax and semantics of a feature, such as color, shape, texture, image topology, motion, or title; (2) a set of descriptor schemes, where each scheme specifies the structure and semantics of the relationships between its components (descriptors or description schemes); (3) a set of coding schemes for the descriptors, and (4) a description definition language (DDL) to specify schemes and descriptions. Such standardization greatly facilitates content-based video retrieval and video data mining.
- It is unrealistic to treat a video clip as a long sequence of individual still pictures and analyze each picture since there are too many pictures, and most adjacent images could be rather similar.

- In order to capture the story or event structure of a video, it is better to treat each video clip as a collection of actions and events in time and first temporarily segment them into video shots. A shot is a group of frames or pictures where the video content from one frame to the adjacent ones does not change abruptly. Moreover, the more representative frame in a video shot is considered the key frame of the shot. Each key frame can be analyzed using the image feature extraction and analysis methods studied above in the content-based image retrieval. The sequence of key frames will then be used to define the sequence of the events happening in the video clip. Thus the detection of shots and the extraction of key frames from video clips become the essential tasks in video processing and mining.
- Video data mining is still in its infancy. There are still a lot of research issues to be solved before it becomes general practice. Similarity-based preprocessing, compression, indexing and retrieval, information extraction, redundancy removal, frequent pattern discovery, classification, clustering, and trend and outlier detection are important data mining tasks in this domain.

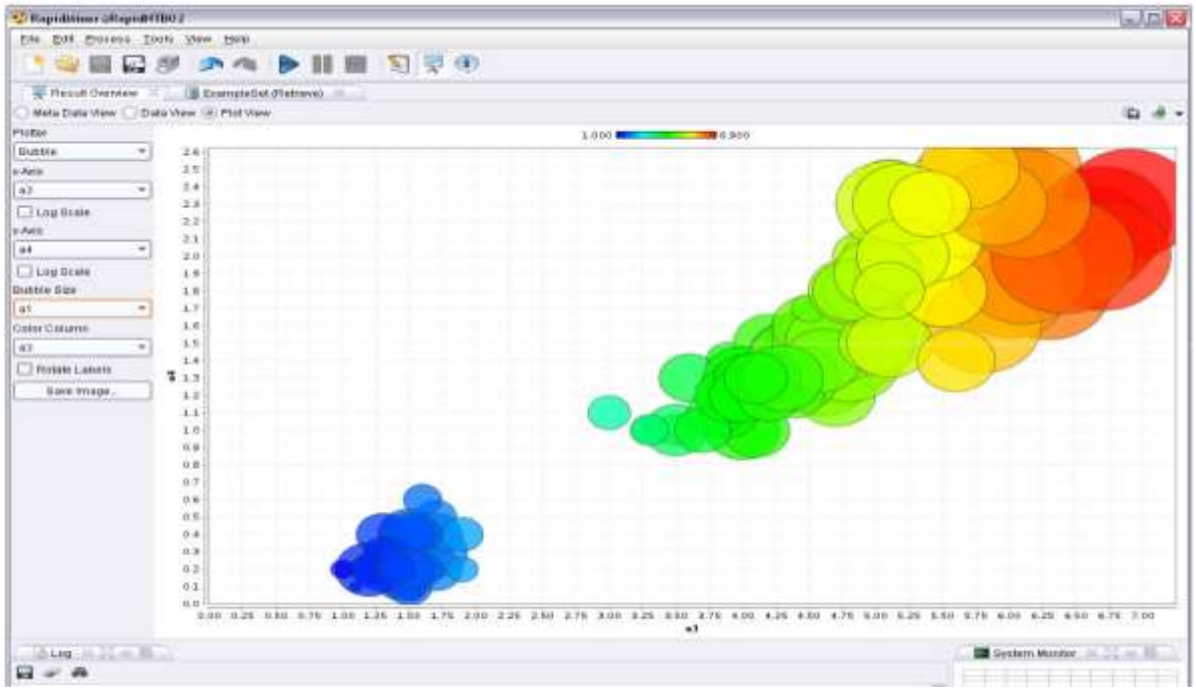
TOOLS

- Data mining or “Knowledge Discovery in Databases” is the process of discovering patterns in large data sets with artificial intelligence, machine learning, statistics, and database systems. The overall goal of a data mining process is to extract information from a data set and transform it into an understandable structure for further use. Here is a simple but fascinating example of how data mining helped dissipate wrong assumptions and conclusions about girls, and take action with tremendous social impact. For long time, the high rate of dropout of girls in schools in developing countries were explained with sociological and cultural hypothesis: girls are not encouraged by indigenous societies, parents treat girls differently, girls are pushed to get married earlier or loaded with much more work than boys. Some others using economic theories, speculated that girls education is not seen by those societies as a good investment.
- Then, in the years 90s, came a group of young data miners who plugged into several schools records on absenteeism, and slowly discovered that girls were missing schools for few days every month, with stunning regularity and predictability.

- A little bit more analysis reveals that girls were missing schools mostly during their menstruation period, and because there were no safe way for them to feel clean and comfortable to come to school during that period.
- Consequence, “millions of girls living in developing countries like Uganda skip up to 20% of the school year simply because they cannot afford to buy mainstream sanitary products when they menstruate. This deliberate absenteeism has enormous consequences on girls’ education and academic potential.”
- In western countries and in Asia, companies and governments are using data mining to make great discoveries. We can do the same in Africa. There are numerous free tools to do so. I have collected the best of them here for you. Try it, start slowly but persist with patience. It could yield amazing and transformational results like Afripads is now helping African girls stay at school.

1. Rapid Miner

- Rapid Miner is unquestionably the world-leading open-source system for data mining. It is available as a stand-alone application for data analysis and as a data mining engine for the integration into own products. Thousands of applications of Rapid Miner in more than 40 countries give their users a competitive edge.

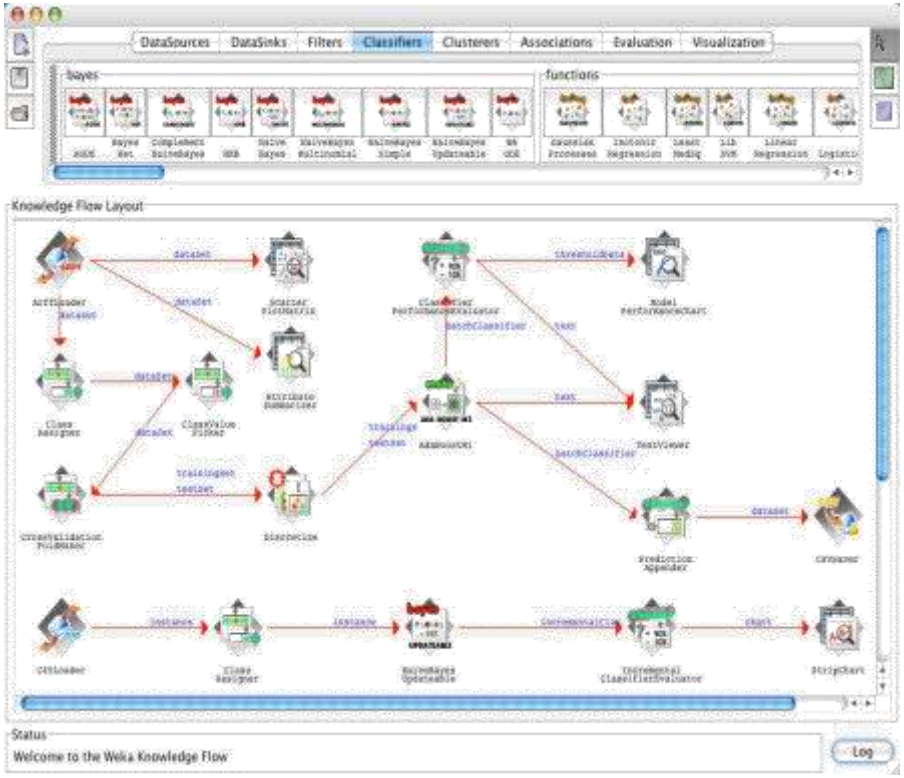


2. Rapid Analytics

- Built around Rapid Miner as a powerful engine for analytical ETL, data analysis, and predictive reporting, the new business analytics server Rapid Analytics is the key product for all business critical data analysis tasks and a milestone for business analytics.

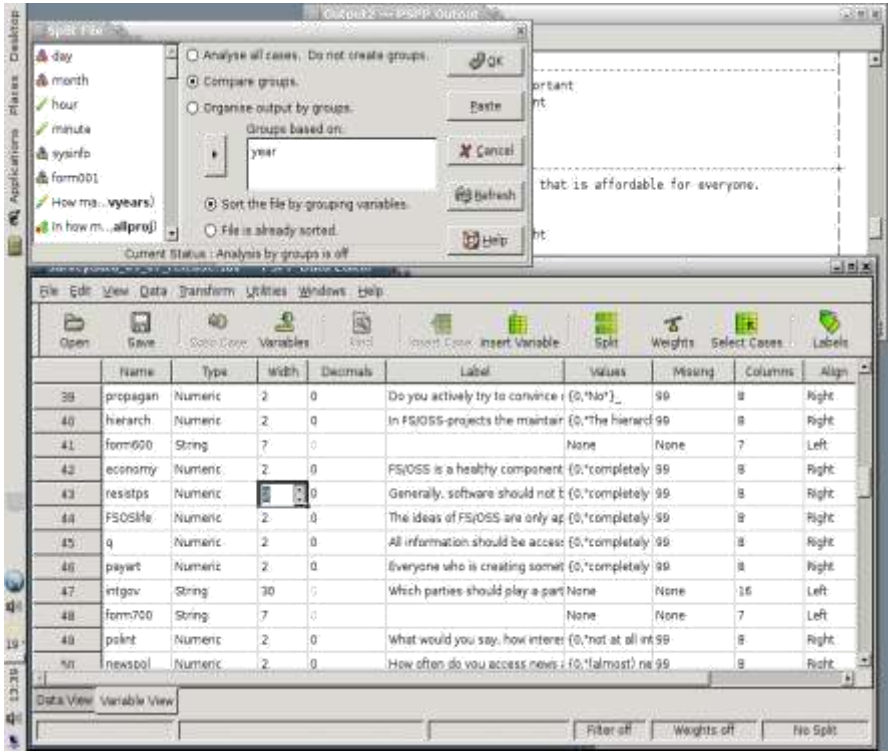
3. Weka

- Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.



4. PSPP

- PSPP is a program for statistical analysis of sampled data. It has a graphical user interface and conventional command-line interface. It is written in C, uses GNU Scientific Library for its mathematical routines, and plotutils for generating graphs. It is a Free replacement for the proprietary program SPSS (from IBM) predict with confidence what will happen next so that you can make smarter decisions, solve problems and improve outcomes.

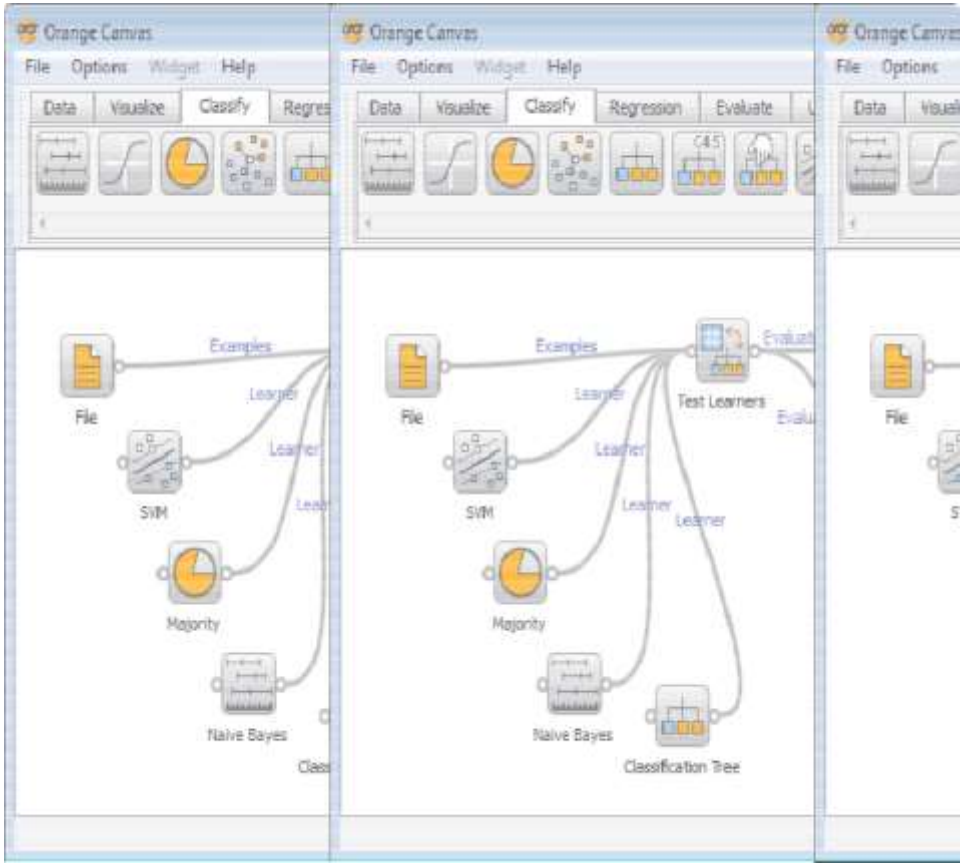


5. KNIME

- KNIME is a user-friendly graphical workbench for the entire analysis process: data access, data transformation, initial investigation, powerful predictive analytics, visualisation and reporting. The open integration platform provides over 1000 modules (nodes)

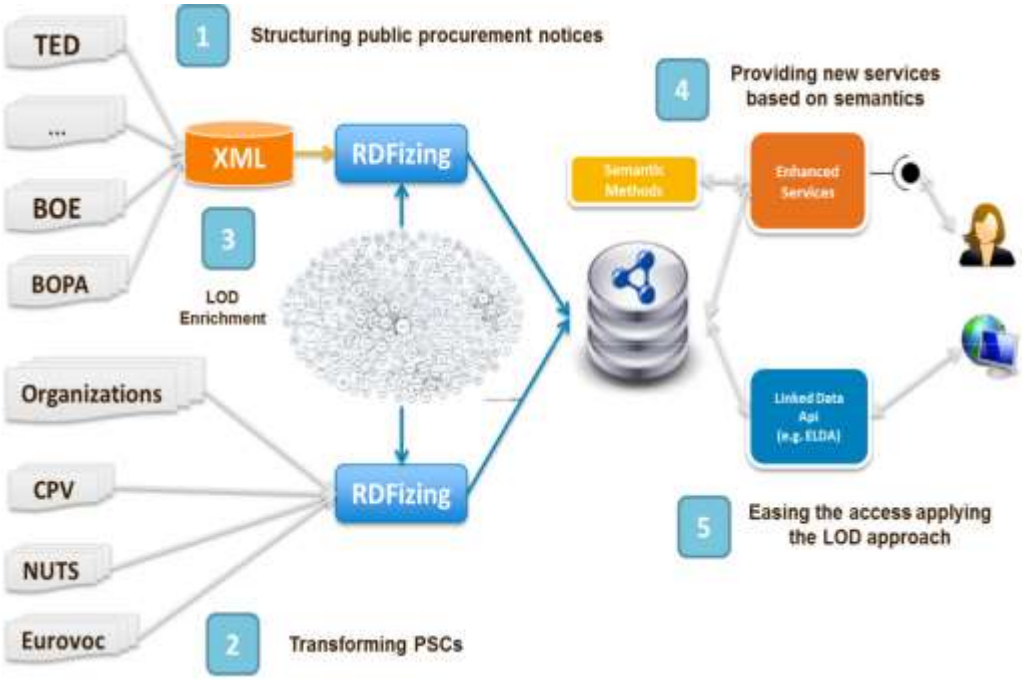
6. Orange

- Orange is an Open source data visualization and analysis for novice and experts. Data mining through visual programming or Python scripting. Components for machine learning. Add-ons for bioinformatics and text mining. Packed with features for data analytics.



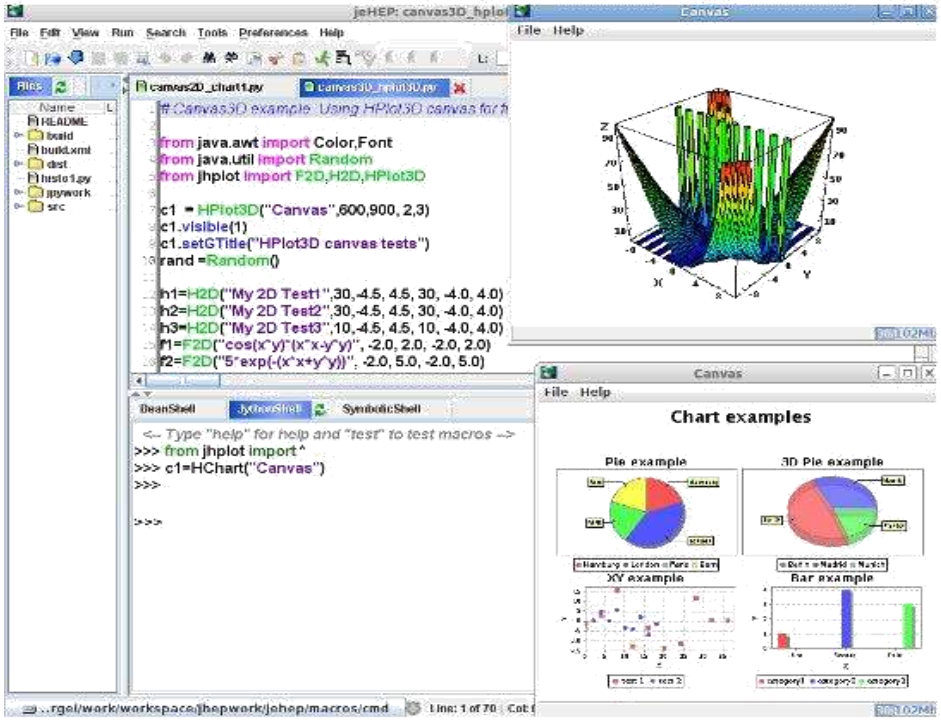
7. Apache Mahout

- Apache Mahout is an Apache project to produce free implementations of distributed or otherwise scalable machine learning algorithms on the Hadoop platform.
- Currently Mahout supports mainly four use cases: Recommendation mining takes users' behavior and from that tries to find items users might like. Clustering takes e.g . text documents and groups them into groups of topically related documents. Classification learns from existing categorized documents what documents of a specific category look like and is able to assign unlabelled documents to the (hopefully) correct category. Frequent itemset mining takes a set of item groups (terms in a query session, shopping cart content) and identifies, which individual items usually appear together.



8. jHepWork

- jHepWork (or “jWork”) is an environment for scientific computation, data analysis and data visualization designed for scientists, engineers and students. The program incorporates many open-source software packages into a coherent interface using the concept of scripting, rather than only-GUI or macro-based concept.
- jHepWork can be used everywhere where an analysis of large numerical data volumes, data mining, statistical analysis and mathematics are essential (natural sciences, engineering, modeling and analysis of financial markets).



9. Rattle

- Rattle (the R Analytical Tool To Learn Easily) presents statistical and visual summaries of data, transforms data into forms that can be readily modelled, builds both unsupervised and supervised models from the data, presents the performance of models graphically, and scores new datasets.
- It is a free and open source data mining toolkit written in the statistical language R using the Gnome graphical interface. It runs under GNU/Linux, Macintosh OS X, and MS/Windows. Rattle is being used in business, government, research and for teaching data mining in Australia and internationally.

Thank you