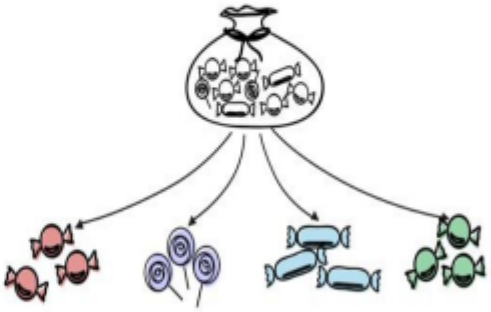
Clustering

Dr. Amit Praseed

What is clustering?

- Clustering is the task of grouping together similar data items in a dataset
- Eg: Similar users are grouped together in a recommendation system
- The class label is often absent in clustering algorithms, which differentiates it from classification



Basic Clustering Techniques

- **Partition Based Clustering**
  - divides the data into k groups such that each group must contain at least one object
  - exclusive cluster separation
- **Hierarchical Clustering**
  - creates a hierarchical decomposition of the given set of data objects
- **Density based Clustering**
  - continue growing a given cluster as long as the density in the "neighborhood" exceeds some threshold
- **Grid Based Clustering**
  - quantizes the object space into a finite number of cells that form a grid structure

k-means Algorithm

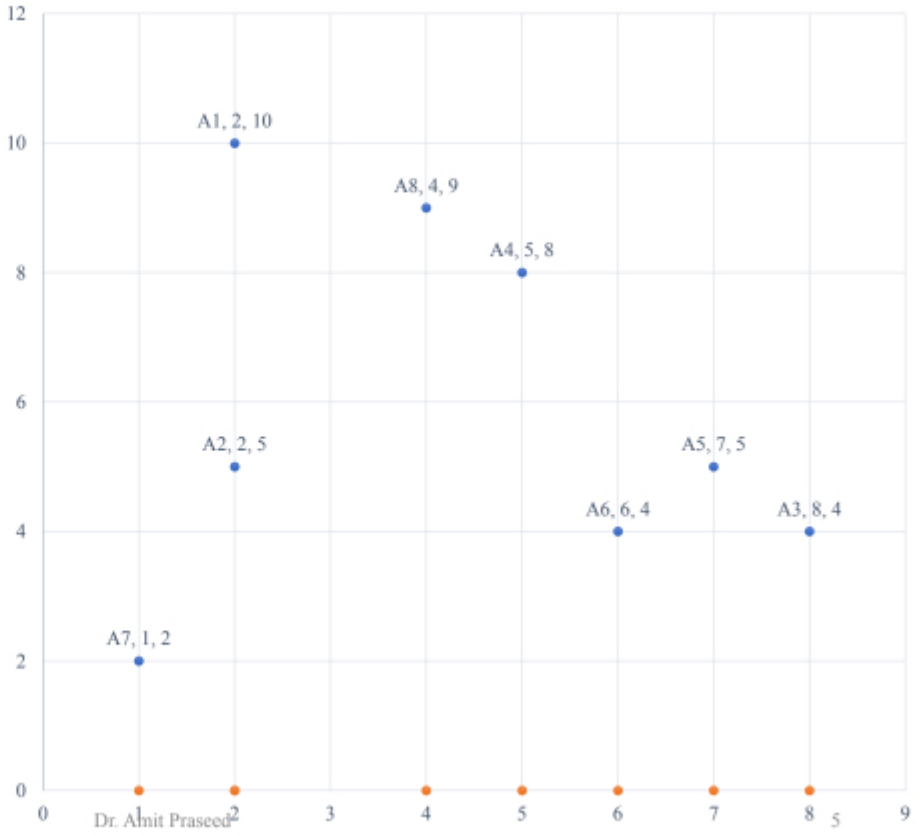
- The centroid of a cluster denotes that cluster
  - For the k-means algorithm, the mean is used to denote the centroid
- Quality of a cluster depends on how similar the items are within a cluster – minimize the within cluster variation

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, C_i)$$

- In general, the problem is NP-Hard
- k-means algorithm uses a greedy approach to approximate the process

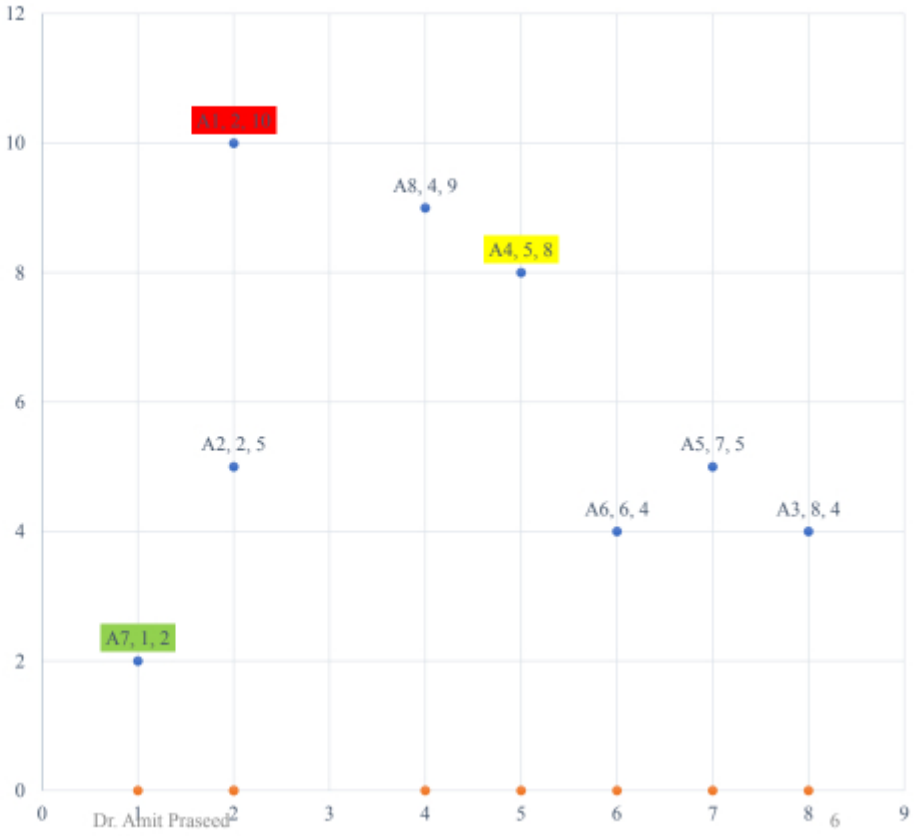
Example

Data Point	X	Y
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9



Example

Data Point	X	Y
A1 (Red)	2	10
A2	2	5
A3	8	4
A4 (Yellow)	5	8
A5	7	5
A6	6	4
A7 (Green)	1	2
A8	4	9



Calculate Distance (Eg: Manhattan Distance)

Data Point	Distance from Red Cluster (2,10)	Distance from Yellow Cluster (5,8)	Distance from Green Cluster (1,2)	Cluster
A1 (2,10)	0	5	9	Red
A2 (2,5)	5	6	4	Green
A3 (8,4)	12	7	9	Yellow
A4 (5,8)	5	0	10	Yellow
A5 (7,5)	10	5	9	Yellow
A6 (6,4)	10	5	7	Yellow
A7 (1,2)	9	10	0	Green
A8 (4,9)	3	2	10	Yellow

First Clusters

Data Point	X	Y
A1 (Red)	2	10
A2 (Green)	2	5
A3(Yellow)	8	4
A4 (Yellow)	5	8
A5(Yellow)	7	5
A6(Yellow)	6	4
A7 (Green)	1	2
A8(Yellow)	4	9

Clusters	Centroid
Red	(2,8)
Green	(1.5,3.5)
Yellow	(6,6)

Recompute Clusters

Data Point	Distance from Red Cluster (2,10)	Distance from Yellow Cluster (6,6)	Distance from Green Cluster (1.5,3.5)	Cluster
A1 (2,10)	0	8	7	Red
A2 (2,5)	5	5	2	Green
A3 (8,4)	12	4	7	Yellow
A4 (5,8)	5	3	8	Yellow
A5 (7,5)	10	2	7	Yellow
A6 (6,4)	10	2	5	Yellow
A7 (1,2)	9	9	2	Green
A8 (4,9)	3	5	8	Red

Second Clusters

Data Point	X	Y
A1 (Red)	2	10
A2 (Green)	2	5
A3(Yellow)	8	4
A4 (Yellow)	5	8
A5(Yellow)	7	5
A6(Yellow)	6	4
A7 (Green)	1	2
A8(Red)	4	9

Clusters	Centroid
Red	(3,9.5)
Green	(1.5,3.5)
Yellow	(6.5,5.25)

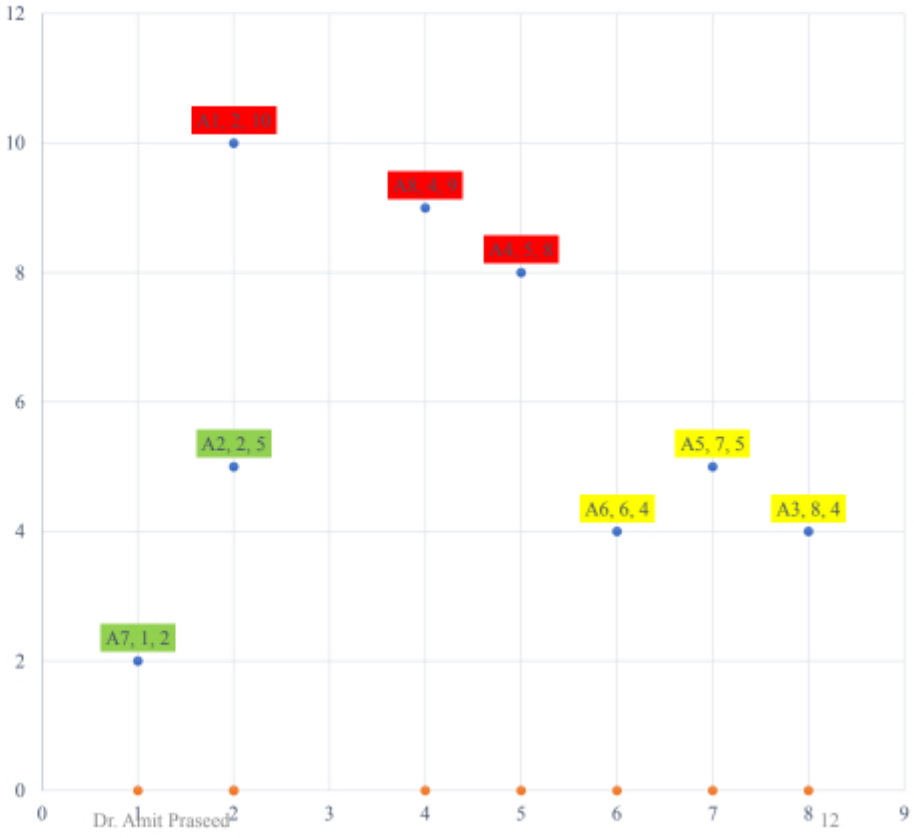
Recompute Clusters (again ☹)

Data Point	Distance from Red Cluster (3,9.5)	Distance from Yellow Cluster (6.5,5.25)	Distance from Green Cluster (1.5,3.5)	Cluster
A1 (2,10)	1.5	9.25	7	Red
A2 (2,5)	5.5	4.75	2	Green
A3 (8,4)	10.5	2.75	7	Yellow
A4 (5,8)	3.5	4.25	8	Red
A5 (7,5)	8.5	0.75	7	Yellow
A6 (6,4)	8.5	1.75	5	Yellow
A7 (1,2)	9.5	8.75	2	Green
A8 (4,9)	1.5	6.25	8	Red

Third Clusters

Data Point	X	Y
A1 (Red)	2	10
A2 (Green)	2	5
A3(Yellow)	8	4
A4 (Red)	5	8
A5(Yellow)	7	5
A6(Yellow)	6	4
A7 (Green)	1	2
A8(Red)	4	9

Clusters	Centroid
Red	(3.67,9)
Green	(1.5,3.5)
Yellow	(7,4.3)



Recompute Clusters (again ☹ ☹)

Data Point	Distance from Red Cluster (3.67,9)	Distance from Yellow Cluster (7,4.3)	Distance from Green Cluster (1.5,3.5)	Cluster
A1 (2,10)	2.67	10.7	7	Red
A2 (2,5)	5.67	5.7	2	Green
A3 (8,4)	9.33	1.3	7	Yellow
A4 (5,8)	2.33	5.7	8	Red
A5 (7,5)	7.33	0.7	7	Yellow
A6 (6,4)	7.33	1.3	5	Yellow
A7 (1,2)	9.67	8.3	2	Green
A8 (4,9)	0.33	7.7	8	Red

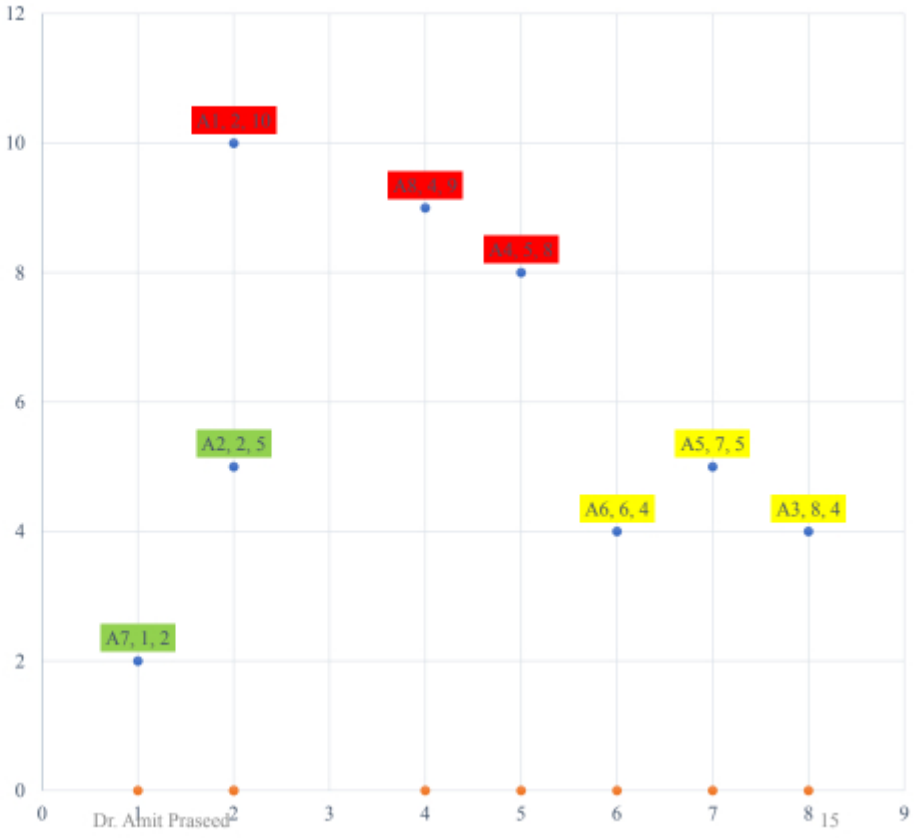
No Cluster Changes this time ☺

Data Point	Distance from Red Cluster (3.67,9)	Distance from Yellow Cluster (7,4.3)	Distance from Green Cluster (1.5,3.5)	Cluster
A1 (2,10)	2.67	10.7	7	Red
A2 (2,5)	5.67	5.7	2	Green
A3 (8,4)	9.33	1.3	7	Yellow
A4 (5,8)	2.33	5.7	8	Red
A5 (7,5)	7.33	0.7	7	Yellow
A6 (6,4)	7.33	1.3	5	Yellow
A7 (1,2)	9.67	8.3	2	Green
A8 (4,9)	0.33	7.7	8	Red

Final Clusters

Data Point	X	Y
A1 (Red)	2	10
A2 (Green)	2	5
A3(Yellow)	8	4
A4 (Red)	5	8
A5(Yellow)	7	5
A6(Yellow)	6	4
A7 (Green)	1	2
A8(Red)	4	9

Clusters	Centroid
Red	(3.67,9)
Green	(1.5,3.5)
Yellow	(7,4.3)



Summary of k-means Algorithm

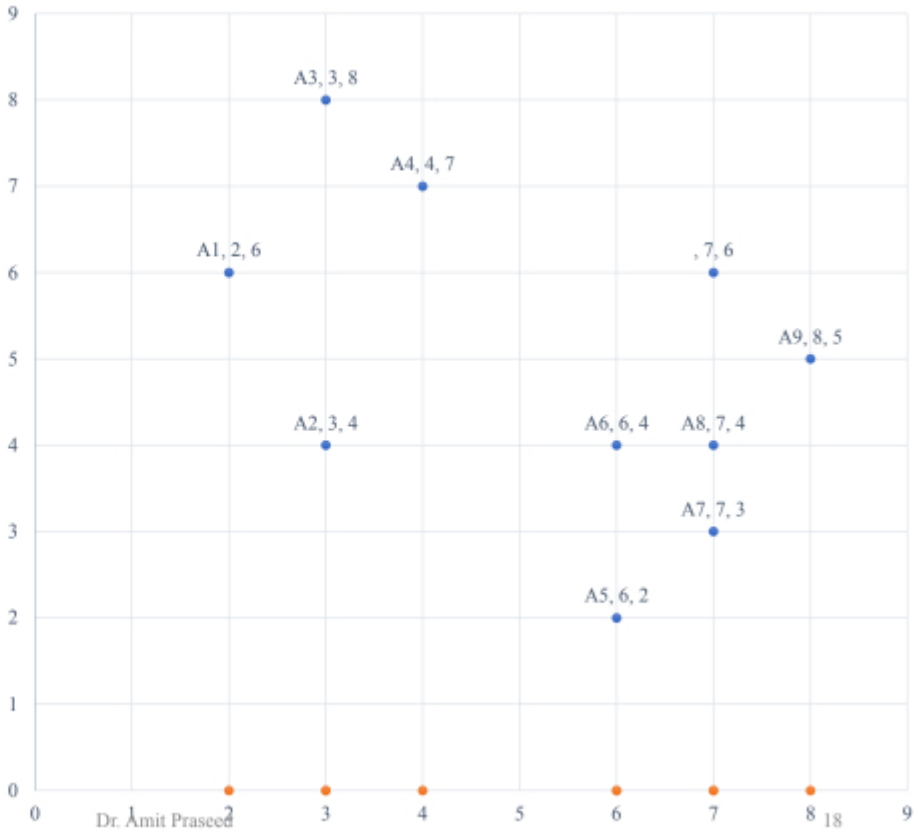
- Not guaranteed to provide a globally optimum solution
- Choosing the optimal k-value is tricky
- Only defined for data types for which mean is defined
  - K-modes is a possible modification
- Can be made more scalable using sampling

k- Medoids Algorithm

- In the k-means algorithm, the centroid is not necessarily one of the data points
  - Sensitive to outliers
- k-medoids algorithm uses a representative element within the group as the “centroid” and computes the clusters based on the medoids
- Partitioning Around Medoids (PAM) algorithm is an example

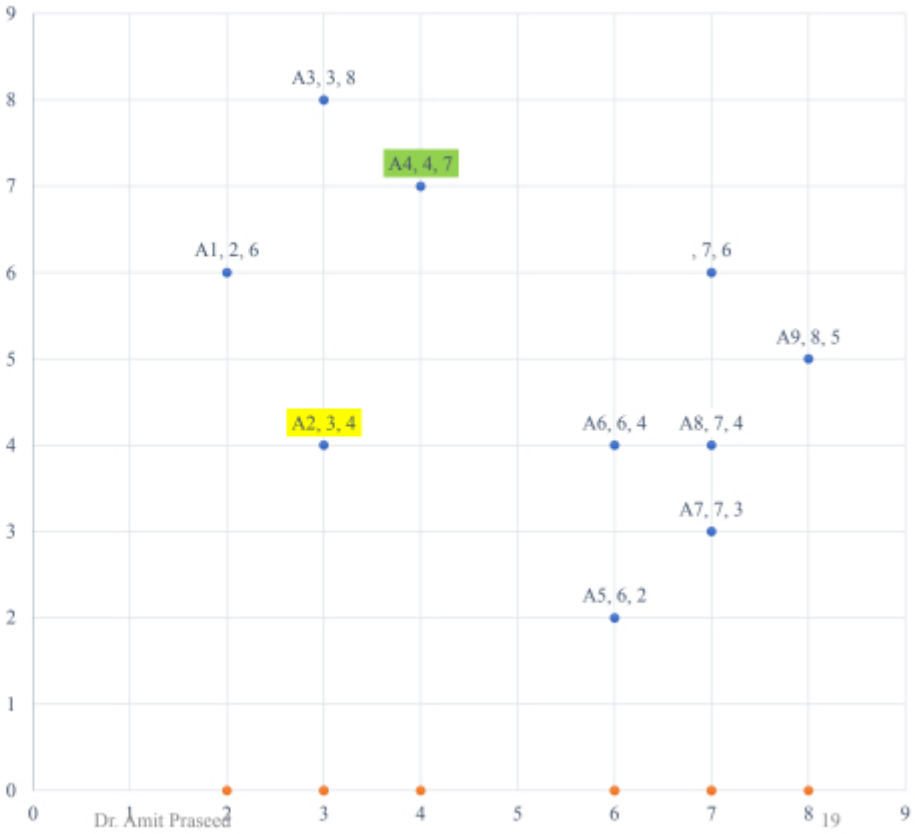
Example

Data Point	X	Y
A1	2	6
A2	3	4
A3	3	8
A4	4	7
A5	6	2
A6	6	4
A7	7	3
A8	7	4
A9	8	5
A10	7	6



Example

Data Point	X	Y
A1	2	6
A2 (Yellow)	3	4
A3	3	8
A4 (Green)	4	7
A5	6	2
A6	6	4
A7	7	3
A8	7	4
A9	8	5
A10	7	6



Calculate Distance (Eg: Manhattan Distance)

Data Point	Distance from Yellow Cluster (3,4)	Distance from Green Cluster (4,7)	Cluster
A1 (2,6)	3	3	Yellow
A2 (3,4)	0	4	Yellow
A3 (3,8)	4	2	Green
A4 (4,7)	4	0	Green
A5 (6,2)	5	5	Yellow
A6 (6,4)	3	5	Yellow
A7 (7,3)	5	7	Yellow
A8 (7,4)	4	6	Yellow
A9 (8,5)	6	6	Yellow
A10 (7,6)	6	4	Green

In case of clashes, a point is allotted to the Yellow Cluster by default

Example

Data Point	X	Y
A1 (Yellow)	2	6
A2 (Yellow)	3	4
A3 (Green)	3	8
A4 (Green)	4	7
A5 (Yellow)	6	2
A6 (Yellow)	6	4
A7 (Yellow)	7	3
A8 (Yellow)	7	4
A9 (Yellow)	8	5
A10 (Green)	7	6

Compute Absolute Error

Data Point	Distance from Yellow Cluster (3,4)	Distance from Green Cluster (4,7)	Cluster
A1 (2,6)	3	3	Yellow
A2 (3,4) Medoid	0	4	Yellow
A3 (3,8)	4	2	Green
A4 (4,7) Medoid	4	0	Green
A5 (6,2)	5	5	Yellow
A6 (6,4)	3	5	Yellow
A7 (7,3)	5	7	Yellow
A8 (7,4)	4	6	Yellow
A9 (8,5)	6	6	Yellow
A10 (7,6)	6	4	Green

E = (A1-A2) + (A5-A2)+ (A6-A2) + (A7-A2) + (A8-A2)+(A9-A2) + (A3-A4)+ (A10-A4)

= (3+4+3+5+4+6)+(2+4) =31

Example

Data Point	X	Y
A1 (Yellow)	2	6
A2 (Yellow)	3	4
A3 (Green)	3	8
A4 (Green) Medoid	4	7
A5 (Yellow)	6	2
A6 (Yellow) Medoid	6	4
A7 (Yellow)	7	3
A8 (Yellow)	7	4
A9 (Yellow)	8	5
A10 (Green)	7	6

Compute Absolute Error

Data Point	Current Cluster	Distance from Yellow Cluster (6,4)	Distance from Green Cluster (4,7)	Cluster	Error
A1 (2,6)	Yellow	6	3	Green	3
A2 (3,4)	Yellow	3	4	Yellow	3
A3 (3,8)	Green	7	2	Green	2
A4 (4,7) Medoid	Green	5	0	Green	0
A5 (6,2)	Yellow	2	5	Yellow	2
A6 (6,4) Medoid	Yellow	0	5	Yellow	0
A7 (7,3)	Yellow	2	7	Yellow	2
A8 (7,4)	Yellow	1	6	Yellow	1
A9 (8,5)	Yellow	3	6	Yellow	3
A10 (7,6)	Green	3	4	Yellow	3
					19

Example

Data Point	X	Y
A1 (Yellow)	2	6
A2 (Yellow)	3	4
A3 (Green)	3	8
A4 (Green) Medoid	4	7
A5 (Yellow)	6	2
A6 (Yellow) Medoid	6	4
A7 (Yellow)	7	3
A8 (Yellow)	7	4
A9 (Yellow)	8	5
A10 (Green)	7	6

The error reduces, so we use the new set of medoids

PAM Algorithm

- The algorithm starts with a randomly selected set of medoids
- Each point is allocated to a particular cluster based on how close they are to the representative elements
- Randomly select a non-representative element to replace an existing representative element
- If the cost after replacement reduces, the new set of representative elements is retained, else it is discarded
- More robust than k-means
- High complexity – $O(k(n-k)^2)$

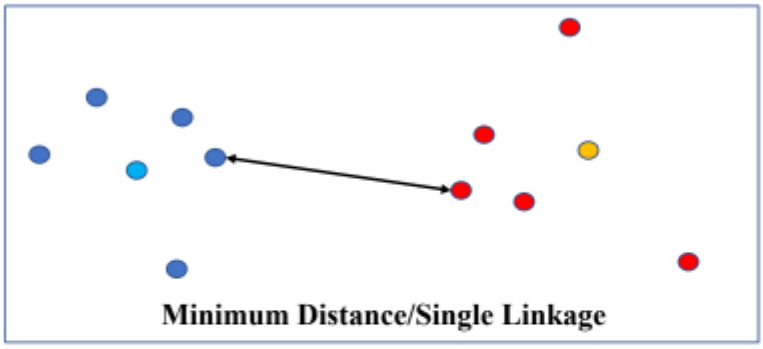
Scalable Versions of PAM

- Clustering LARge Applications (CLARA)
  - Select a random sample from the data points and perform the PAM algorithm
  - Success depends on how well the sample represents the population
- Clustering Large Applications based on RANdomised Search (CLARANS)
  - Confine the set of candidate replacement medoids to a random sample of the data

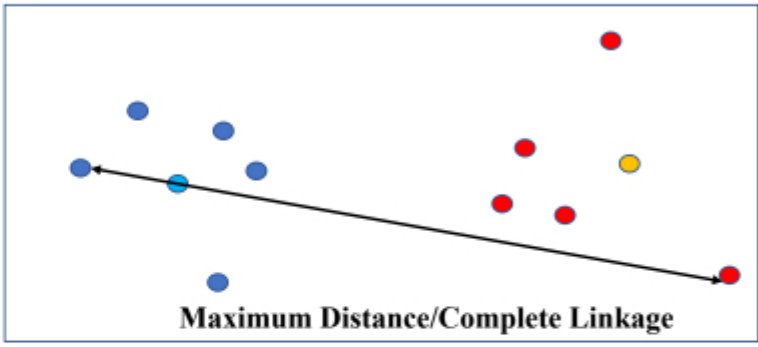
Hierarchical Clustering

- Groups data objects into a hierarchy or “tree” of clusters
- Agglomerative Hierarchical Clustering:
  - Bottom Up
  - Starts with every point in a separate cluster
  - Clusters are merged together based on how “close” they are
  - Finally, you get one “super cluster”
- Divisive Hierarchical Clustering:
  - Top Down
  - Starts with a single “super cluster”
  - Iteratively splits the clusters so that cohesion within the cluster improves
  - Finally every point becomes its own cluster

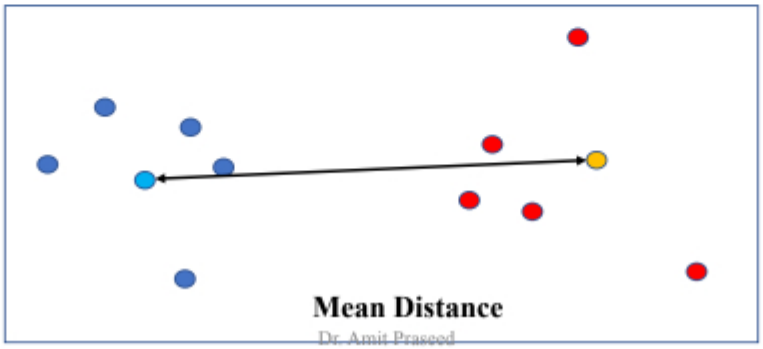
Linkage Measures between Clusters



Minimum Distance/Single Linkage

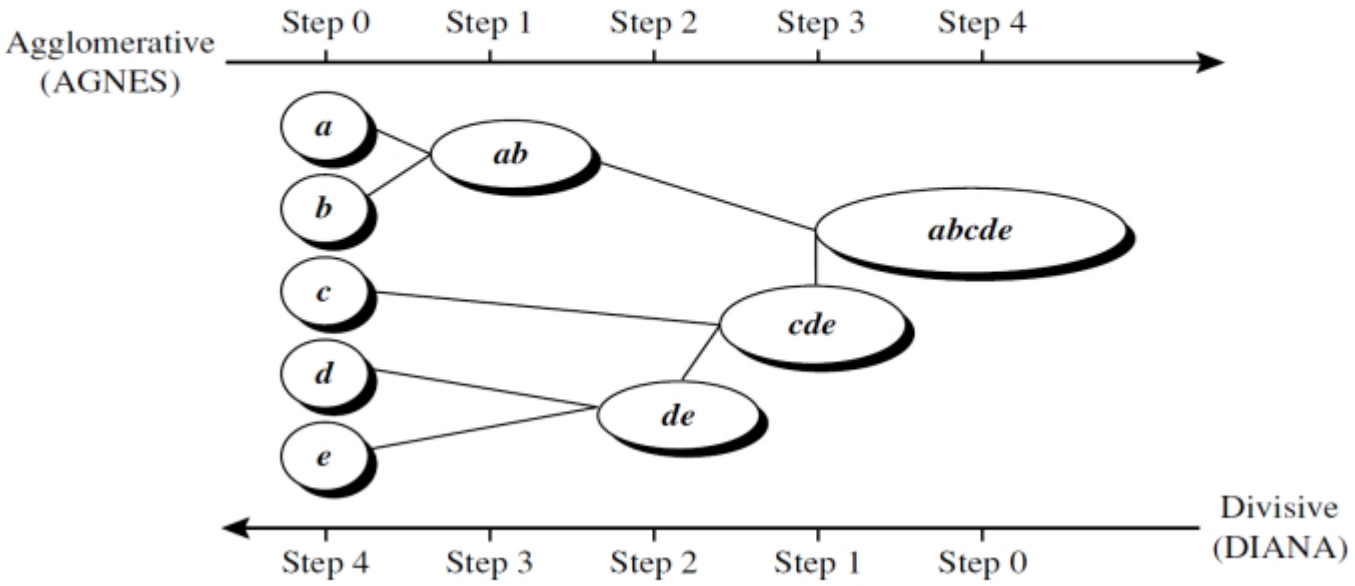


Maximum Distance/Complete Linkage



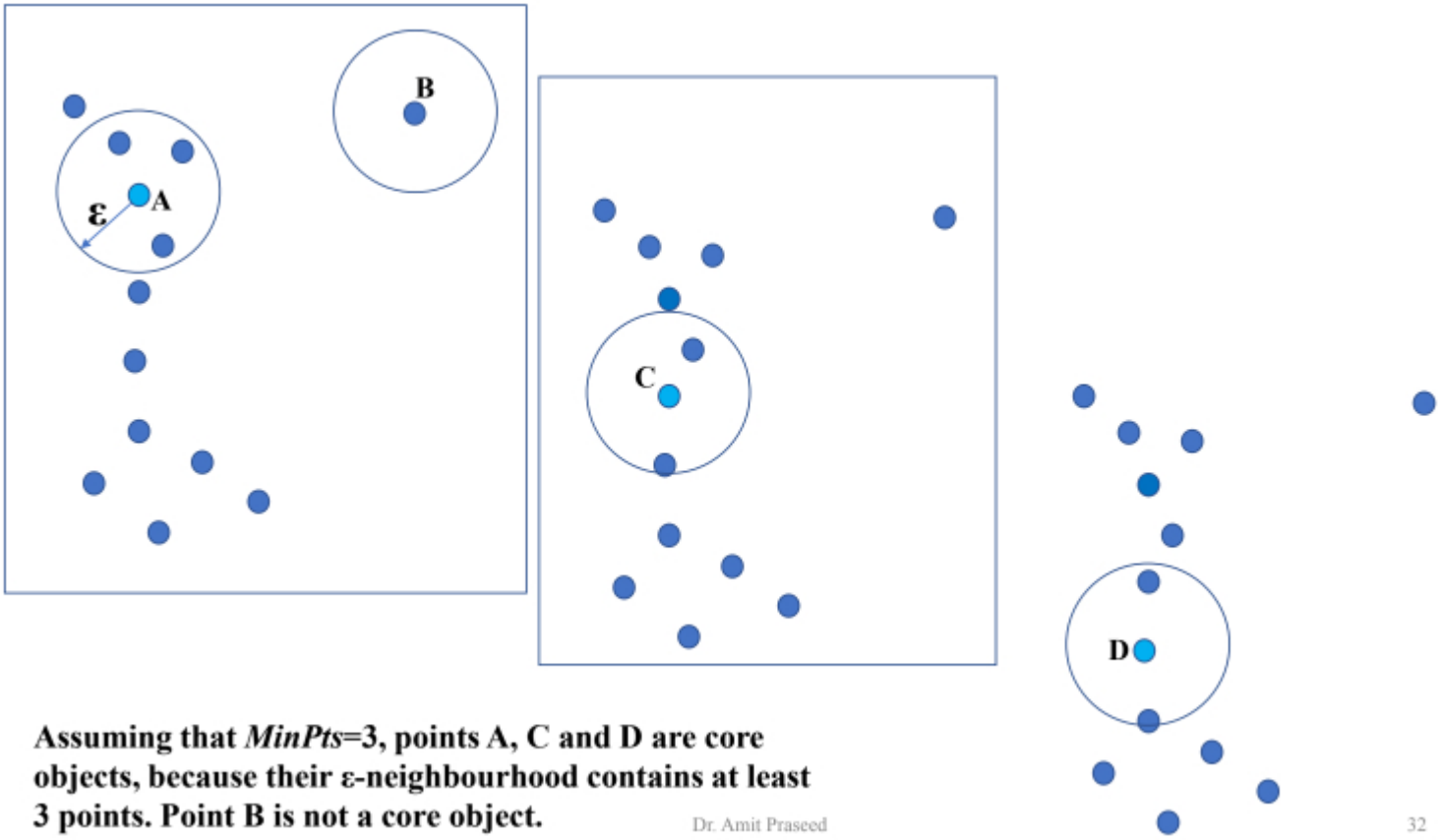
Mean Distance

Dendrogram Representation



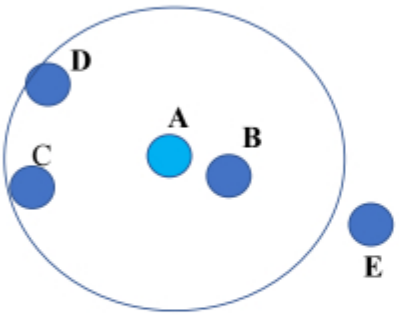
Density Based Clustering

- The density of an object *o* can be measured by the number of objects close to *o*.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) finds core objects, that is, objects that have dense neighborhoods.
  - It connects core objects and their neighborhoods to form dense regions as clusters.
- A user-specified parameter ε is used to specify the radius of a neighborhood we consider for every object.
- An object is a core object if the ε - neighborhood of the object contains at least *MinPts* objects.

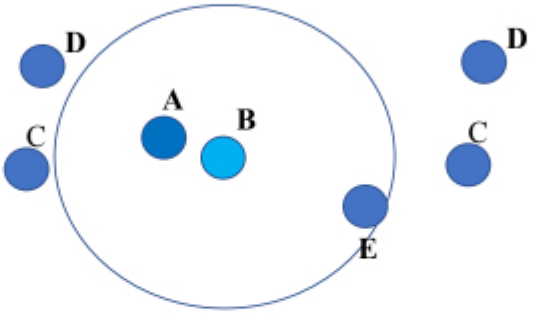


Assuming that *MinPts*=3, points A, C and D are core objects, because their ε-neighbourhood contains at least 3 points. Point B is not a core object.

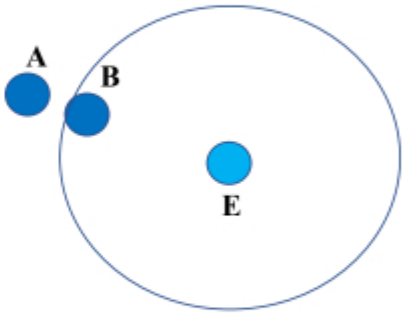
Core, Border and Noise Objects



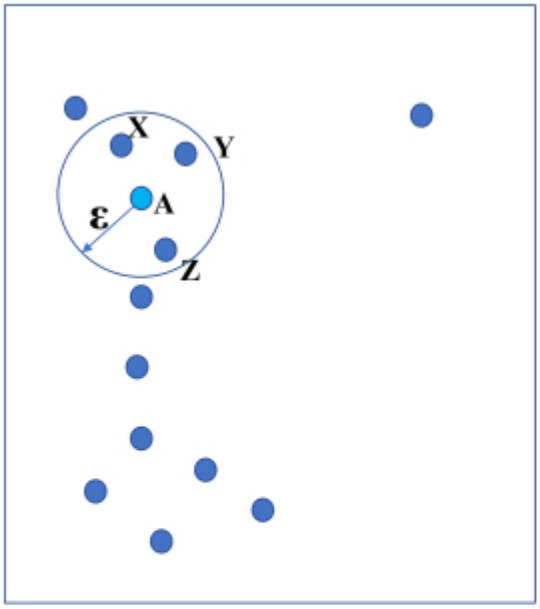
Assuming that *MinPts*=4, object A is a core object because its ε-neighbourhood contains 5 objects



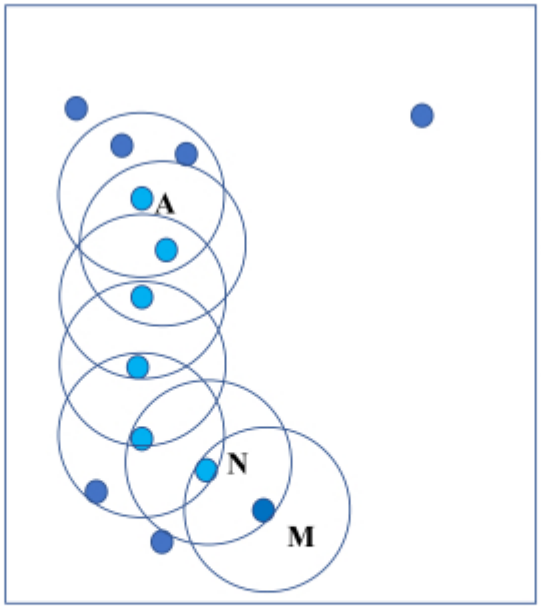
Assuming that *MinPts*=4, object B is called a Border object, because it lies in the neighbourhood of a core object (A), but is itself not a core object



Assuming that *MinPts*=4, object E is called a Noise Object, because it is neither a core object, nor a border object



X, Y and Z are said to be direct density reachable from A

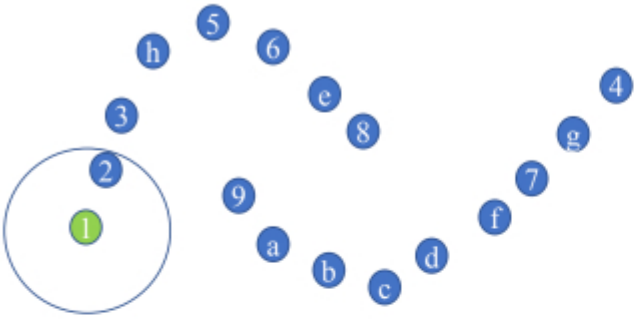


M and N are said to be density reachable from A

Here N is density reachable from A and A is density reachable from N.
Hence, we say that A and N are density connected.

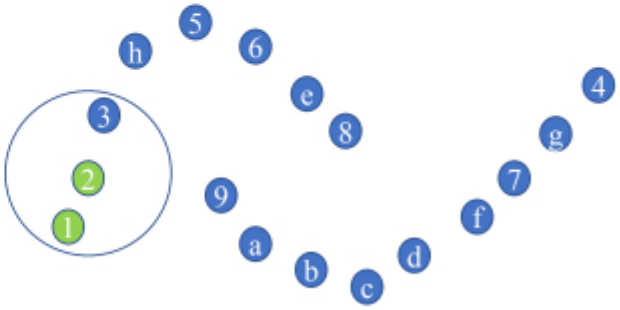
We cannot say the same for A and M.

Let MinPts=2

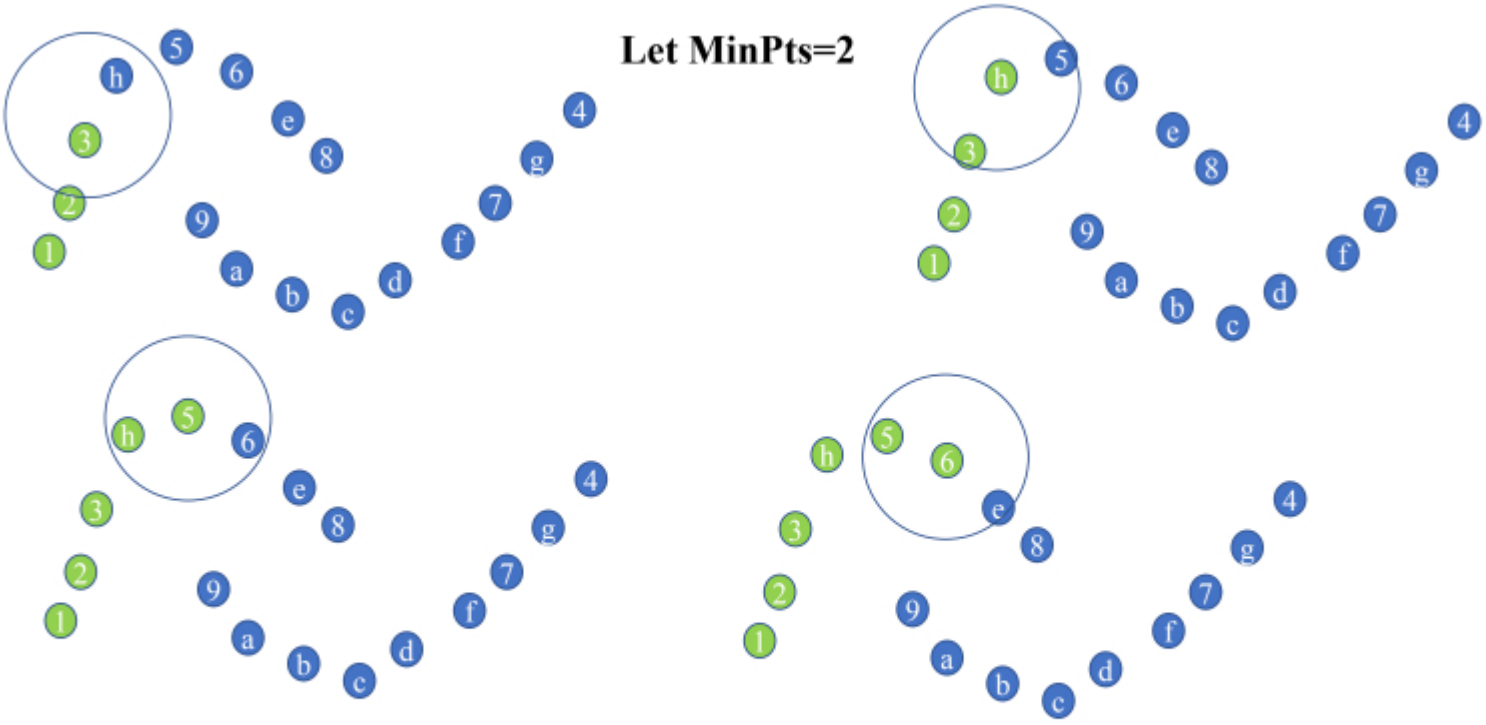


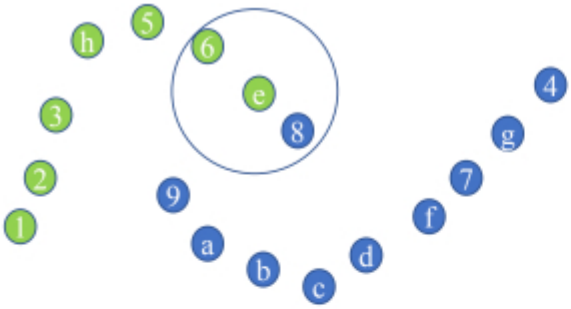
For every unvisited node in N, assign it to the current cluster (green) and add its neighbours to N if it is a core node

Select a random, unvisited object. If it is a core object assign it to a cluster, say green and add all of its neighbours into a candidate set N. Otherwise mark it as a noise node.

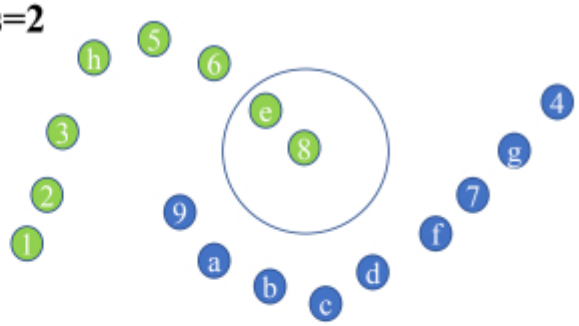


Let MinPts=2

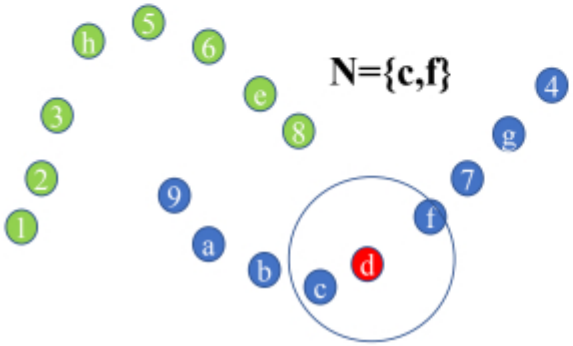


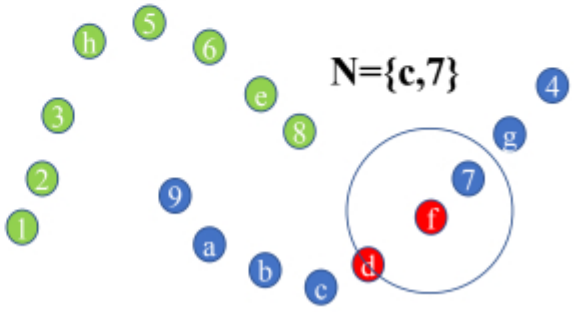


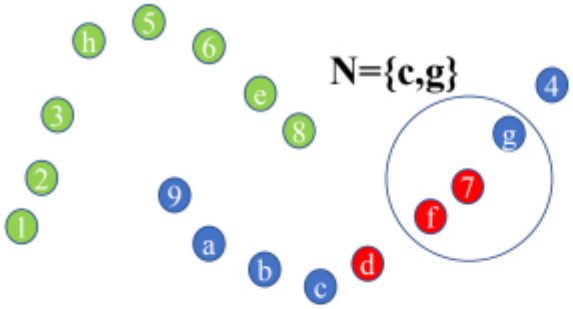
Let MinPts=2

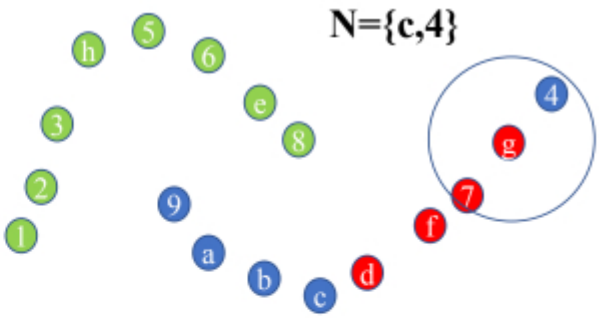


At this point, the set N becomes empty. So DBSCAN picks another unvisited node and adds it to a new cluster.

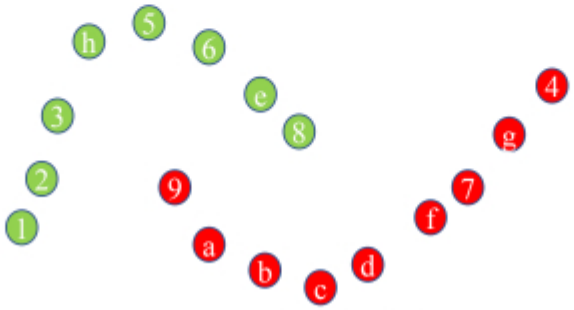








Finally, when no more unvisited nodes are left…



Summary of DBSCAN Algorithm

- Capable of detecting non-spherical clusters
- Drawbacks:
  - Algorithm is sensitive to the value of ε and MinPts, which are difficult to estimate
  - In real world scenarios, use of a global density value may not yield good results

Clustering in the Presence of Query Conditions

- One of the common applications of clustering is spatial data mining
- It is common to encounter problems where the required clusters have conditions attached to them
  - Eg: Select the maximal regions that have at least 100 houses per unit area and at least 70% of the house prices are above $400K and with total area at least100 units with 90% confidence
- Every query requires a separate clustering operation
  - Each clustering operation depends on the number of objects *n* in the dataset
  - Extensive computation required *per query*

Grid Based Clustering

- Grid-based clustering method takes a space-driven approach
  - Partition the embedding space into cells
  - Independent of the distribution of the input objects
- Quantizes the object space into a finite number of cells that form a grid structure
- Fast processing time, typically independent of the number of data
  - dependent on only the number of cells in each dimension in the quantized space.
- Eg: STING, CLIQUE

STING Algorithm

- STING (STatistical INformation Grid) is a grid based clustering algorithm for answering queries
  - Eg: Select the maximal regions that have at least 100 houses per unit area and at least 70% of the house prices are above $400K and with total area at least100 units with 90% confidence
- Divides the entire search space hierarchically into cells
- At the bottom layer, metrics such as number of points, mean, standard deviation, min, max etc. are maintained
- The information of higher layers can be computed from the information at lower layers.

STING Algorithm

- Query answering starts at a particular layer.
  - Cells which satisfy the constraints are selected based on confidence interval
  - Processing at the next layer only requires selected cells
  - Proceed till the last layer
- Query independent
- Cost depends on the granularity at the lowest level
- All cluster shapes are isothetic

CLIQUE Algorithm

- CLIQUE (CLustering In QUEst) is a grid based method for finding density based clusters in subspaces
- Uses the Apriori property:
  - A *k*-dimensional cell *c* can have at least *l* points only if every *(k-1)*-dimensional projection of *c* has at least *l* points
- Dense clusters are identified in (k-1) dimension and the candidate clusters for the kth dimension are found similar to apriori algorithm

Advanced Topics in Data Mining

Dr. Amit Praseed

Mining Data Streams

Techniques for Dealing with Streaming Data

- Sampling
  - Reservoir Sampling: Maintain a set of s candidates in the reservoir, which form a true random sample of the elements seen so far in the stream. Every new element has a certain probability of replacing an old element in the reservoir
- Sliding Windows
- Histograms
- Multiresolution Methods
- Sketching

Frequent-Pattern Mining in Data Streams

- Frequent pattern mining algorithms require at least 2 scans over the input data
  - Not feasible for streaming data
- Two approaches
  - Keep track of only a predefined, limited set of items and itemsets.
    - Very limited usage and expressive power
  - Approximate itemset counting
    - A router keeps track of items whose frequency is at least 1% of the entire traffic stream seen so far. It is felt that 1/10 of min support is an acceptable margin of error ($\varepsilon = 0.1\%$). All frequent items with a support of at least min support will be output, but some items with a support of at least (min support$-\varepsilon$) will also be output.
    - Eg: Lossy Counting Algorithm

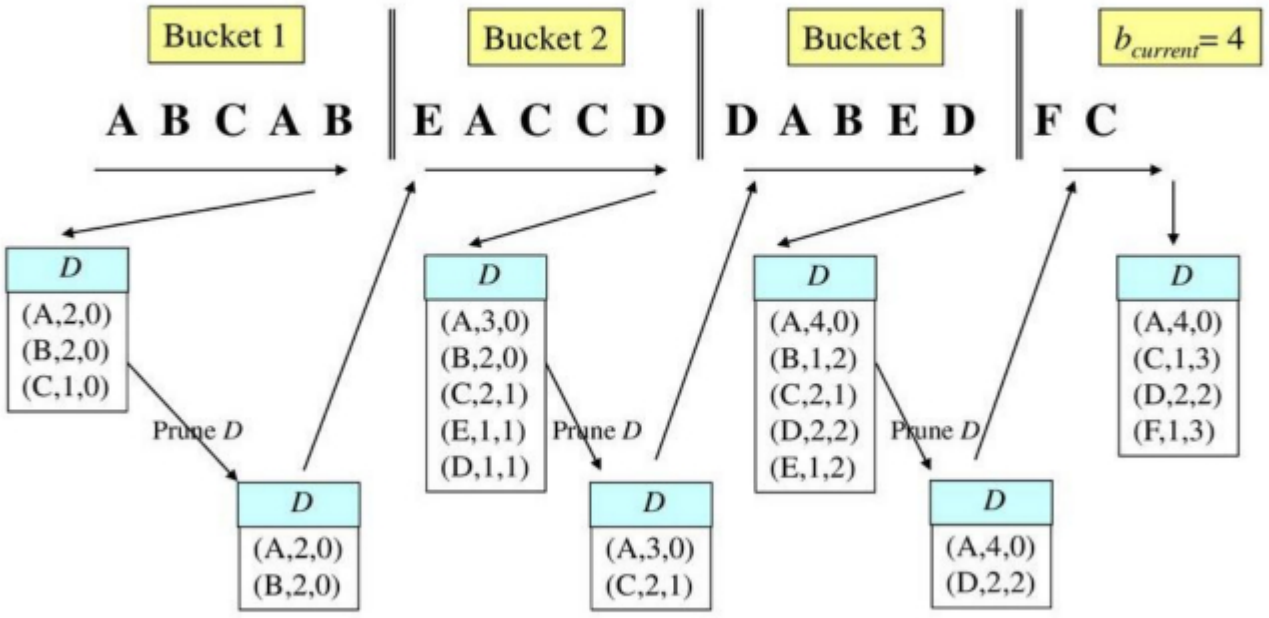
Lossy Counting Algorithm

- Two input parameters:
  - min support threshold, σ
  - error bound, ε.
- The incoming stream is conceptually divided into buckets of width $w = \lceil \frac{1}{\varepsilon} \rceil$.
- For each item, the algorithm maintains f , the approximate frequency count, and Δ, the maximum possible error of f .
- If the new item is from the bth bucket, we set Δ, the maximum possible error on the frequency count of the item, to be b-1

Lossy Counting Algorithm

- Whenever a bucket "boundary" is reached
  - frequency list is examined. Let b be the current bucket number. An item entry is deleted if $f + \Delta <= b$.
  - The frequency count stored for each item will either be the true frequency of the item or an underestimate of it
- The Lossy Counting algorithm has the following properties
  - No false negatives
  - Few false positives, but all of the false positives have a support value of at least $\sigma-\varepsilon$

Example: ε=0.2, w=5, N=17, $b_{current}$=4



14

Classification in Data Streams

- Classification in data streams also faces a similar problem
  - Testing can be done
  - Training is difficult as the data needs to be resident on disk/memory
- Need for approximations – for example within a decision tree
- Normal DT algorithms require the data to be resident on system for training
  - Multiple iterations may be required on the training data to find the splitting attribute
  - Possible to get an approximate value of the splitting criteria using mathematical approximations

Hoeffding's Inequality

- Suppose we make N independent observations of a random variable r with range R, where r is an attribute selection measure. If we compute the mean, r', of this sample, the Hoeffding bound states that the true mean of r is at least r'- ε, with probability 1- δ , where δ is user-specified and

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

Hoeffding Trees and Variations

- The algorithm calculates the information gain of every attribute as the data comes in
  - Only need to maintain the count of tuples belonging to a particular class
- If the difference between Gain(Aa) and Gain(Ab) exceeds ε, we can decide upon Aa as the splitting attribute with confidence of 1- δ
- VFDT (Very Fast Decision Tree) algorithm makes several modifications to the Hoeffding tree algorithm
  - breaking near-ties during attribute selection more aggressively
  - computing the G function after a number of training examples
  - deactivating the least promising leaves whenever memory is running low
  - dropping poor splitting attributes
  - and improving the initialization method
- Concept Adapting VFDT (CVFDT) is used to manage concept drift by giving more weight to newer samples and growing parallel subtrees

Clustering Data Streams

- STREAM Algorithm
  - Input m input data points
  - Perform clustering and obtain k cluster centres
  - Final cluster centres can be computed by performing clustering on the cluster centres themselves
  - If it is difficult to maintain cluster centres for buckets in memory, perform a clustering on the centres and store a single unified value for the cluster centres

Mining Time Series Data

Time Series Data

- A time-series database consists of sequences of values or events obtained over repeated measurements of time.
- Two goals in time-series analysis:
  - modeling time series
  - forecasting time series
- Trend analysis consists of 4 major components
  - Trend or long-term movements: Indicate the general direction in which a timeseries graph is moving over a long interval of time. Typical methods for determining a trend curve or trend line include the weighted moving average method and the least squares method
  - Cyclic movements or cyclic variations: These refer to the cycles, that is, the long-term oscillations about a trend line or curve, which may or may not be periodic.
  - Seasonal movements or seasonal variations: These are systematic or calendar related.
  - Irregular or random movements: These characterize the sporadic motion of time series due to random or chance events

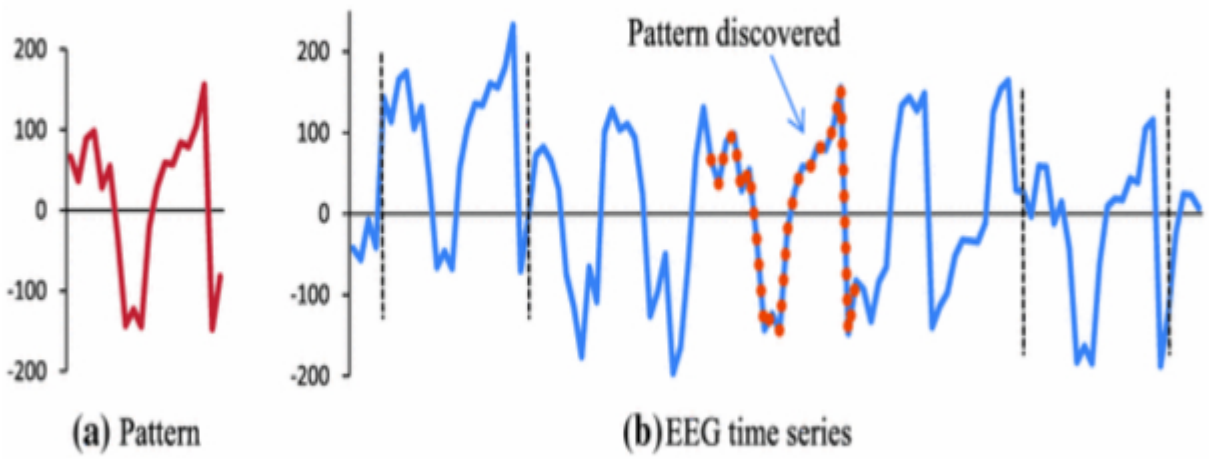
Trend Analysis in Time Series Data

- Correcting for seasonal fluctuations
  - Seasonal fluctuations can be identified by autocorrelation between the ith and (i-k)th elements
  - Deseasonalization can be done using seasonal index
- A common method for determining trend is to calculate a moving average of order n as the following sequence of arithmetic means

Similarity Search in Time Series Analysis

- Two types
  - Subsequence matching finds the sequences in S that contain subsequences that are similar to a given query sequence x
  - whole sequence matching finds a set of sequences in S that are similar to each other (as a whole)
- Data Reduction
  - DFT
  - DWT
  - SVD

Similarity Search in Time Series Data



Text Mining

Text Mining and Information Retrieval

- Typical problem is to locate relevant documents from a document collection based on a user query
- Methods of text retrieval:
  - Document Selection:
    - A document is represented by a set of keywords
    - User issues a query, possibly a Boolean expression
    - Documents satisfying the Boolean expression are returned
  - Document ranking
    - Match the keywords in the query with the words in the document
    - Score each document
    - Present a ranked list

Vector Space Model

- Documents (and queries) are represented as vectors using a vector space model
- Eg: S1: It is very hot today

  S2: It was raining heavily yesterday
- Using vector space model, these texts can be represented as follows:

S = [It	is	very	hot	today	was	raining	heavily	yesterday]	
S1=[1	1	1	1	1	0	0	0	0	]
S2=[1	0	0	0	0	1	1	1	1	]

- The distance between two documents (or between a document and a query) can be represented using the cosine similarity between them

Example

- T1: The movie was not good, but I don’t like scary movies in general
- T2: I found the movie to be very creepy
- T3: The movie was not as scary as other movies in the movie franchise

- W=[movie, not, good, don’t, like, scary, general, found, very, creepy, other, franchise]
- T1=[2,1,1,1,1,1,1,0,0,0,0,0]
- T2=[1,0,0,0,0,0,0,1,1,1,0,0]
- T3=[3,1,0,0,0,1,0,0,0,0,1,1]
- Q={creepy movie}
  =[1,0,0,0,0,0,0,0,0,0,1,0,0]

Example

- T1=[2,1,1,1,1,1,1,0,0,0,0,0]
- T2=[1,0,0,0,0,0,0,1,1,1,0,0]
- T3=[3,1,0,0,0,1,0,0,0,0,1,1]
- Q= [1,0,0,0,0,0,0,0,0,0,1,0,0]
- Sim(T1,Q)=?
- Sim(T2,Q)=?
- Sim(T3,Q)=?

Example

- T1=[2,1,1,1,1,1,1,0,0,0,0,0], $||T1||=\sqrt{10}$
- T2=[1,0,0,0,0,0,0,1,1,1,0,0], $||T2||=2$
- T3=[3,1,0,0,0,1,0,0,0,0,1,1] $||T3||=\sqrt{13}$
- Q= [1,0,0,0,0,0,0,0,0,0,1,0,0] $||Q||=\sqrt{2}$
- Sim(T1,Q)=2/ $\sqrt{20}$=0.447
- Sim(T2,Q)=1/ $\sqrt{2}$=0.707
- Sim(T3,Q)=3/ $\sqrt{13}$=0.832

Vector Space Model

- Stop words: Words such as “a”, “an” etc. which do not contribute to the meaning of text
- Stop words are removed in the pre-processing stage to reduce the dimensionality of text data and improve the efficiency
- All words are not equally important!!
  - Words that occur frequently in all documents contribute less to document relevance than words that occur rarely
  - Instead of storing word counts, it is better to score the TF-IDF values

$$TF - IDF = TF * log\left(\frac{N}{DF}\right)$$

  - Words that occur frequently in a document, but occur rarely in the corpus are given a higher TF-IDF value

Text 1	i love natural language processing but i hate python
Text 2	i like image processing
Text 3	i like signal processing and image processing

	and	but	hate	i	image	language	like	love	natural	processing	python	signal
Text 1	0	1	1	2	0	1	0	1	1	1	1	0
Text 2	0	0	0	1	1	0	1	0	0	1	0	0
Text 3	1	0	0	1	1	0	1	0	0	2	0	1

Term	and	but	hate	i	image	language	like	love	natural	processing	python	signal
IDF	0.47712	0.47712	0.4771	0	0.1760913	0.477121	0.1760913	0.477121	0.47712125	0	0.477121	0.477121

	and	but	hate	i	image	language	like	love	natural	processing	python	signal
Text 1	0	0.47712	0.4771	0	0	0.477121	0	0.477121	0.47712125	0	0.477121	0
Text 2	0	0	0	0	0.1760913	0	0.1760913	0	0	0	0	0
Text 3	0.47712	0	0	0	0.1760913	0	0.1760913	0	0	0	0	0.477121

Other Considerations

- The TF-IDF formula can also be modified to normalize the term frequency, and to avoid IDF value as 0
- Words like “work”, “working”, “worker” are considered to be separate words, even though they convey similar meanings
- Stemming: Strips the word down to the word root
- Even after removing stop words and performing stemming, the dimensionality of text data remains huge
  - Latent Semantic Indexing (LSI)
  - Locality Preserving Indexing (LPI)

Web Mining

Challenges in Web Data Mining

- Huge amount of data
  - Seems infeasible to mine and store information
  - Only a small fraction of this data is actually relevant or useful
- Lack of unifying structure
- Highly dynamic

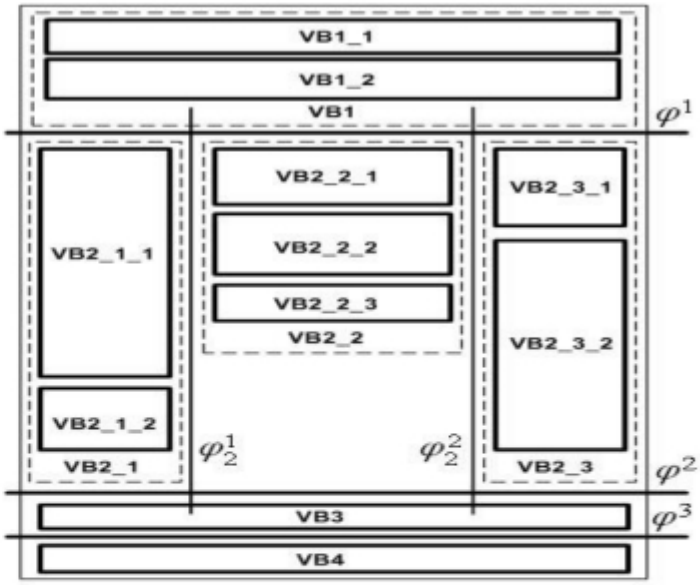
Keyword based Search Engines

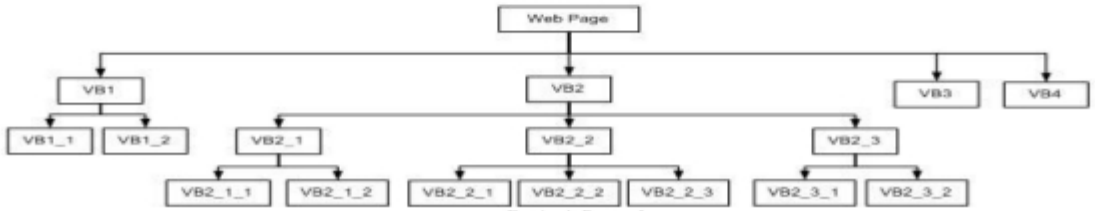
- Index-based Web search engines search the Web, index Web pages, and build and store huge keyword-based indices
  - Help locate sets of Web pages containing certain keywords
- Drawbacks:
  - Breadth of topics can lead to a huge number of document entries returned by a search
  - Many documents that are highly relevant to a topic may not contain keywords defining them – "polysemy"
- More advanced techniques needed, beyond normal text mining
  - Web content mining
  - Web structure mining
  - Web usage mining

Mining the Web Page Layout Structure

- The DOM structure of a Web page is a tree structure, where every HTML tag in the page corresponds to a node in the DOM tree.
- Due to the flexibility of HTML syntax, many Web pages do not obey the W3C HTML specifications, which may result in errors in the DOM tree structure.
- VIsion-based Page Segmentation (VIPS) Algorithm
  - Aims to extract the semantic structure of a Web page based on its visual presentation.
  - Tree structure: each node in the tree corresponds to a block.
  - Each node will be assigned a value (Degree of Coherence) to indicate how coherent is the content in the block based on visual perception
  - Compared with DOM-based methods, the segments obtained by VIPS are more semantically aggregated.

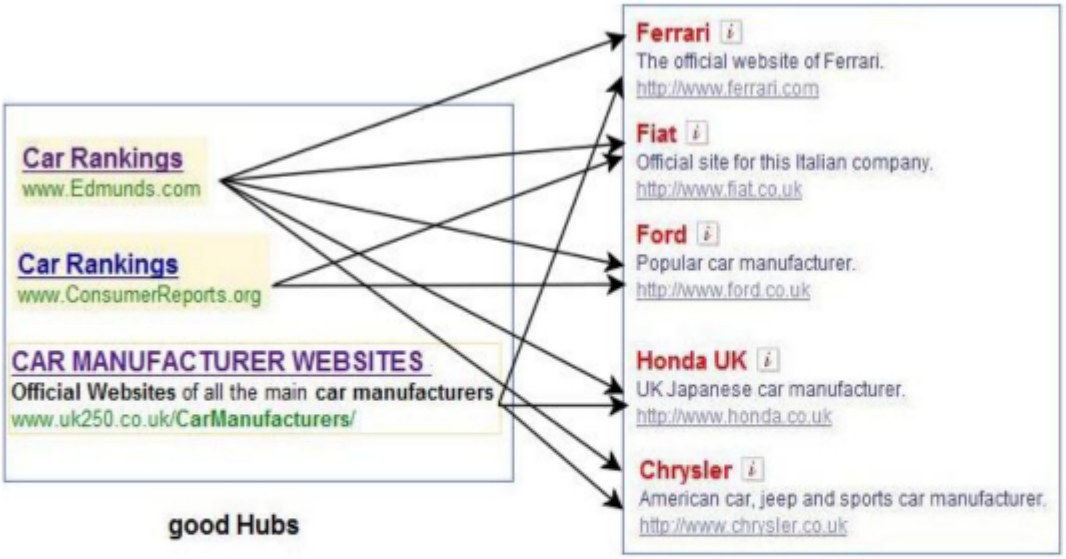

YAHOO! SHOPPING
AUCTIONS
Books & Comics
Auctions > Books & Comics
Search
Sell Stuff
Categories
Antique & Rare
Audio & Large Print
Bestsellers
Children & Young Adult
Classic Titles
Comics
Cooking, Food & Wine
First Editions
Foreign Language
Lesbian, Gay & Bisexual
Literature & Fiction
Magazines
Non Fiction
Poetry
Self Help
Travel, Reference & Education
Other
Or find in:
Yahoo! Shopping
Yahoo! Classifieds
Popular Searches
Harry Potter
Stephen King
Spider Man
Superman
Rolling Stone
Superheroes
Spider-Man
Superman, Batman, Incredible Hulk
X-Men, Green Lantern, Avengers
Captain America, Fantastic Four
Flash Gordon, Wonder Woman
Star Wars
Darth Vader
Star Wars, Empire Strikes Back
Return of the Jedi, Phantom Menace
Attack of the Clones
Obi-Wan Kenobi, Luke, Yoda, R2D2
Darth Maul, Queen Amidala, Anakin
Featured Auctions
Promote your auctions here for a daily fee.
Inside Yahoo!
Ask Yahoo! - Literature
Books Shopping
Ebooks Shopping
Yahoo! Book Clubs
Yahoo! Message Boards
Yahoo! News
Oprah's Picks
Copyright © 2002 Yahoo! Inc. All rights reserved.


VB1_1
VB1_2
VB1
VB2_2_1
VB2_2_2
VB2_2_3
VB2_2
VB2_3_1
VB2_3_2
VB2_1_1
VB2_1_2
VB2_1
VB2_3
VB3
VB4
φ^1
φ_2^1
φ_2^2
φ^2
φ^3


Web Page
VB1
VB2
VB3
VB4
VB1_1
VB1_2
VB2_1
VB2_2
VB2_3
VB2_1_1
VB2_1_2
VB2_2_1
VB2_2_2
VB2_2_3
VB2_3_1
VB2_3_2

Identifying Authoritative Web Pages

- When an author of a Web page creates a hyperlink pointing to another Web page, this can be considered as the author's endorsement of the other page.
  - The collective endorsement of a given page by different authors on the Web may indicate the importance of the page and may naturally lead to the discovery of authoritative Web pages.
- A hub is one or a set of Web pages that provides collections of links to authorities.
- How can we use hub pages to find authoritative pages
  - HITS (Hyperlink-Induced Topic Search) algorithm
  - First, HITS uses the query terms to collect a starting set of from an index-based search - root set.
  - Some of them should contain links to most of the prominent authorities.
  - The root set can be expanded into a base set by including all of the pages that the root-set pages link to and all of the pages that link to a page in the root set
  - Second, a weight-propagation phase is initiated. This iterative process determines numerical estimates of hub and authority weights.



good Hubs

good Authorities

A higher authority weight occurs if the page is pointed to by pages with high hub weights.

A higher hub weight occurs if the page points to many pages with high authority weights.

Query: Top automobile makers

Ref: http://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html

Web Usage Mining

- Web server usually registers a log entry for every access of a Web page.
  - Includes the URL requested, the IP address from which the request originated, and a timestamp.
- Weblog data need to be cleaned, condensed, and transformed in order to retrieve and analyze significant and useful information
- Data mining can be performed on Weblog records to find association patterns, sequential patterns, and trends of Web accessing
- With the use of such Weblog files, studies have been conducted on analyzing system performance, improving system design by Web caching, Web page prefetching, and Web page swapping; understanding the nature of Web traffic; and understanding user reaction and motivation.