Clustering

Professor **Anita Wasilewska**
: CSE 634
DATA MINING

References

Examples: https://docs.google.com/presentation/d/1tjLskhnbLPIzilxHhiv0lO4c-ONd2Yfg_-CztVBfYho/edit#slide=id.g1384764765_0_57

Introduction: http://www3.cs.stonybrook.edu/~skiena/data-manual/lectures/pdf/L20.pdf

https://knowm.org/introduction-to-clustering/

Requirements: http://sungsoo.github.io/2015/05/02/requirements-for-cluster-analysis.html

https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf

Hard & Soft Clustering: https://www.youtube.com/watch?v=ThgJzGYVWzc

Clustering High Dimensionality Data: https://en.wikipedia.org/wiki/Clustering_high-dimensional_data

https://en.wikipedia.org/wiki/Correlation_clustering https://www.youtube.com/watch?v=sIyMAzAHw6I

https://docs.google.com/presentation/d/1lM_LveKhyhiFXRwqns9MGnVAEX9Eey2h7Ue_6j62Rds/edit#slide=id.g36e5e88eb_08

Constraint Based Clustering: https://www.coursera.org/learn/cluster-analysis/lecture/tVroK/6-3-constraint-based-clustering

https://in.pinterest.com/pin/392728029979811507/

Principal Component Analysis: http://www.cse.buffalo.edu/faculty/azhang/data-mining/pca.ppt

PROMPT http://ieeexplore.ieee.org/document/7823676/

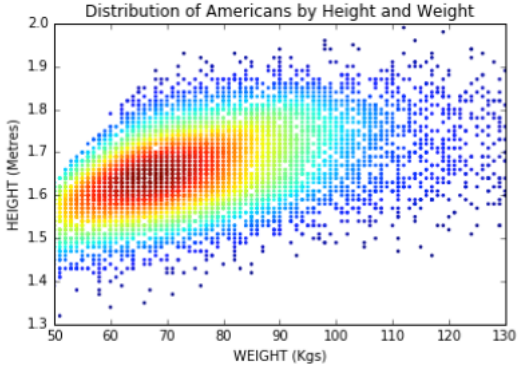
Overview

- Introduction
- Requirements of Clustering in Data Mining
- Distance Metrics
- Supervised and Unsupervised
- Hard and Soft Clustering
- Clustering High Dimensional Data
- Principal Component Analysis
- Constrained based Clustering
- Summary

Introduction

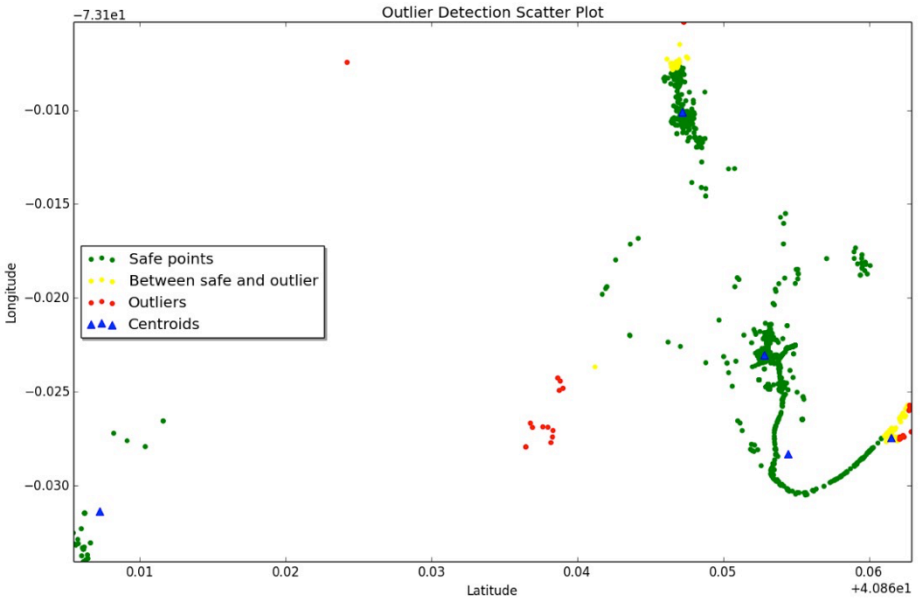
- Clustering is a technique to group data together on the basis of how *(dis)similar* the data is according to a chosen criterion.
- *Motivation:* Researchers all around were facing a general question on how to organize data observed from experiments/researches into structures that convey something meaningful.
- On visualizing the clusters, one could observe why the clusters exist, i.e. the (dis)similarities on which the data have been clustered.
- Helpful in *PATTERN DISCOVERY*.

Examples



Clustering by frequency

Image segmentation in Computer Vision



Outlier detection
(SBU Map)

Source: https://www.coursera.org/learn/cluster-analysis/lecture/LTMGx/1-2-applications-of-cluster-analysis

Introduction contd.

How many Clusters do we see?

- Most common cluster shapes : Compact, Circular.

- Depends upon the eye of the viewer or context of the problem you wish to solve (some may require horizontal/ circular/ vertical or based on specified distance metric)

Source: http://www3.cs.stonybrook.edu/~skiena/data-manual/lectures/pdf/L20.pdf

Requirements for clustering in Data Mining

- Scalability
- Dealing with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Interpretability and usability.

What Is Good Clustering?

- A good clustering method will produce high quality clusters in which:
  - Intra-class (that is, intra-cluster) similarity is high
  - Inter-class similarity is low

- Depends on both the similarity measure used by the method and its implementation
- Measured by its ability to discover some or all of the hidden patterns
- Objective evaluation is problematic: usually done by human / expert inspection

Distance Metrics

- Choosing (dis)similarity measures is a critical step in clustering.
- Often more important than the clustering algorithm.
- Properties of Distance Metric:
  - Non-Negativity - Distance(x, y) >= 0
  - Coincidence - Distance(x, x) = 0
  - Symmetry - Distance(x, y) = Distance(y, x)
  - Subadditivity - Distance(x, y) <= Distance(x, z) + Distance(z, y)

Metrics of Similarity

1. Euclidean Distance
   a. L2 Norm D(A, B) = $\sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$
2. Cosine Distance
   a. D(A, B) = 1 - similarity
   b. Cosine similarity is defined as $\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$
3. Edit Distance
4. Hamming Distance
5. Jaccard Distance
   a. D(A, B) = $1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$

Supervised vs Unsupervised Clustering

Supervised

- Categories or classes are known before training the model
- Predict the class to which a data point belongs.
- Objective is to identify clusters having high probability density of belonging to a class.
- A loss function is used to drive the predictions.

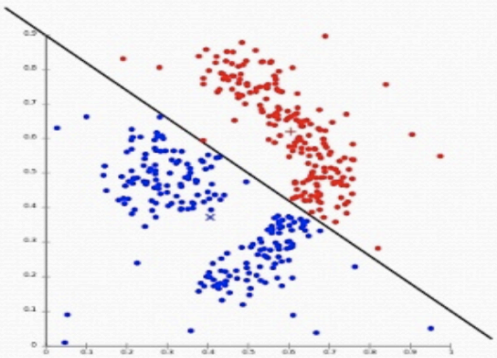
Unsupervised

- Data points are assigned to different classes, without knowing the clusters before training.
- Classification is driven by the properties of the data points without any supervised loss function
- Used mainly for large datasets

Hard vs Soft Clustering

Hard Clustering

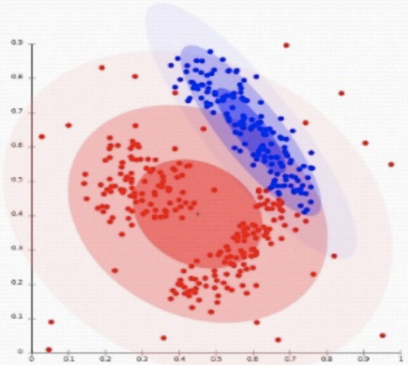
Clusters do not overlap - A data point either belongs to a cluster or not.



each data element belongs to exactly one cluster

Soft Clustering

Clusters may overlap - A data point can belong to more than one cluster

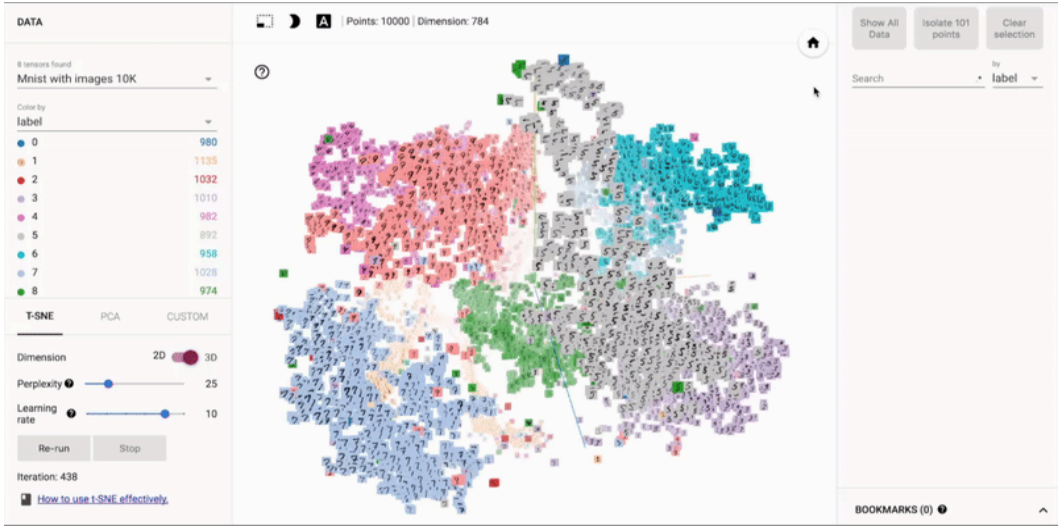


elements can belong to more than one cluster, and associated with each element is a set of membership levels.

Source: https://www.slideshare.net/aselmanb/fuat-a-fuzzy-clustering-analysis-tool

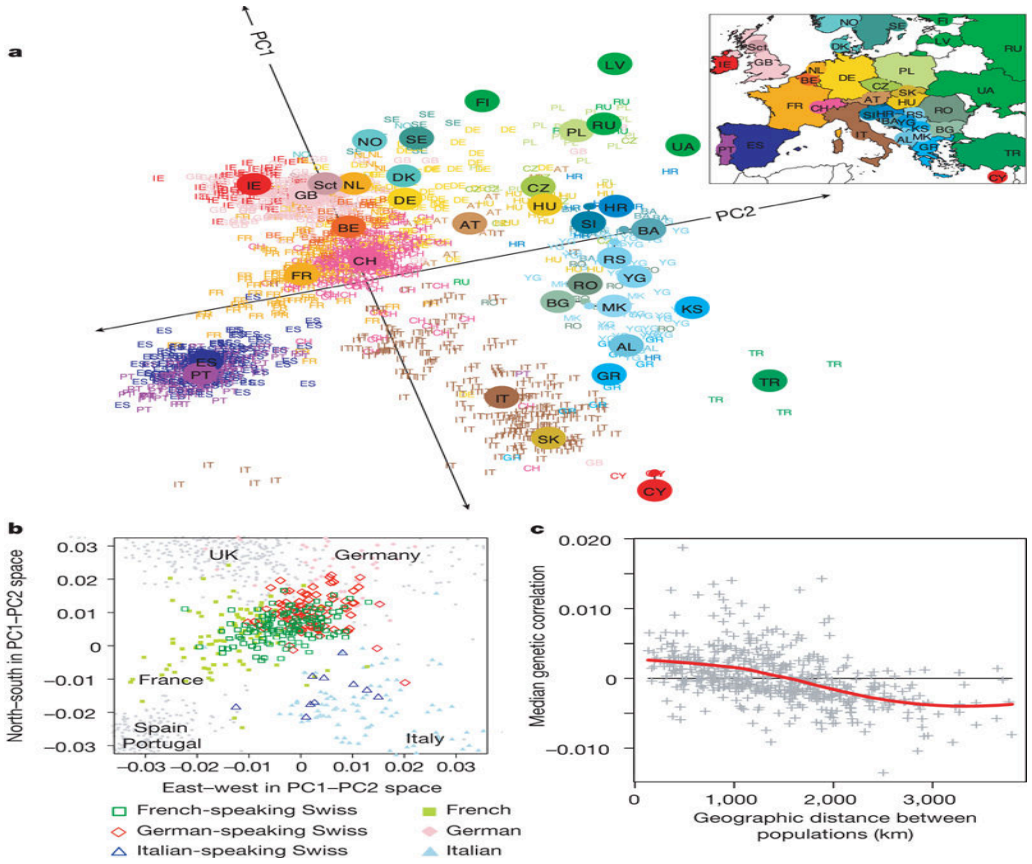
Clustering High Dimensionality Data

- Problems
  - ***"Curse of Dimensionality"*** - Hard to think, Impossible to visualize, Intractable complete enumeration
  - *Analyze* - Analytical techniques "generalized"
  - *Understand* - Get insights from the dataset
    - Pattern analysis
    - Prediction



- Approaches
  - Subspace clustering
  - Projected clustering
  - Hybrid clustering
  - Correlation clustering

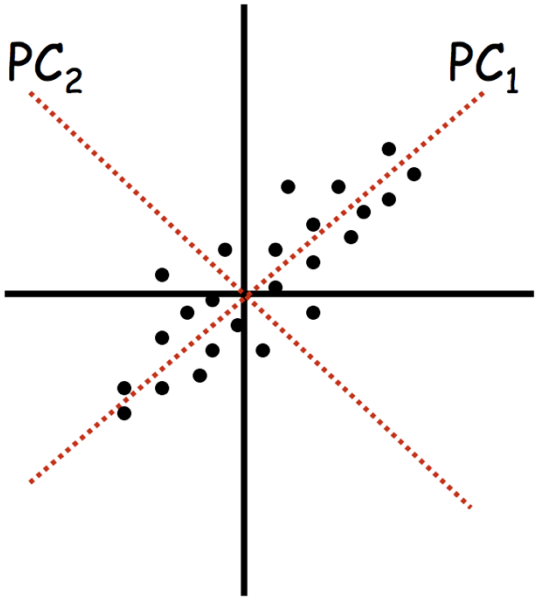
Sources: Open sourcing the Embedding Projector: a tool for visualizing high dimensional data



Source: Genes mirror geography within Europe: John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens& Carlos D. Bustamante, Year: 2008

Principal Component Analysis

- An technique used to reduce the dimensionality of the data set to 2D or 3D
  - Reduce number of dimensions in data
  - Retain as much variation as possible
  - Linear transformation of the original variables
  - Find Principal components (PC's) that are uncorrelated and ordered
- Example applications
  - Face recognition
  - Image compression
  - Gene expression analysis



Sources: http://www.cse.buffalo.edu/faculty/azhang/data-mining/pca.ppt

PCA: one attribute

- Question: how much spread is in the data along the given axis? (distance to the mean)
- Variance=Standard deviation^2

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}$$

Temperature
42
40
24
30
15
18
15
30
15
30
35
30
40
30

Sources: http://www.cse.buffalo.edu/faculty/azhang/data-mining/pca.ppt

PCA: Two dimensions

Covariance: measures the correlation between X and Y

- Cov(X,Y)=0: independent
- Cov(X,Y)>0: move same dir
- Cov(X,Y)<0: move opposite dir

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

X=Temperature	Y=Humidity
40	90
40	90
40	90
30	90
15	70
15	70
15	70
30	90
15	70
30	70
30	70
30	90
40	70

Sources: http://www.cse.buffalo.edu/faculty/azhang/data-mining/pca.ppt

Covariance Matrix

- Contains covariance values between all possible dimensions (=attributes):

$$C^{nxn} = (c_{ij} \mid c_{ij} = \mathrm{cov}(Dim_i, Dim_j))$$

- Example for three attributes (x,y,z):

$$C = \begin{pmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) & \mathrm{cov}(x,z) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) & \mathrm{cov}(y,z) \\ \mathrm{cov}(z,x) & \mathrm{cov}(z,y) & \mathrm{cov}(z,z) \end{pmatrix}$$

Sources: http://www.cse.buffalo.edu/faculty/azhang/data-mining/pca.ppt

EigenValues and EigenVectors

- Vectors **x** having same direction as $A\mathbf{x}$ are called *eigenvectors* of A (A is an n by n matrix).
- In the equation $A\mathbf{x}=\lambda\mathbf{x}$, λ is called an *eigenvalue* of A.

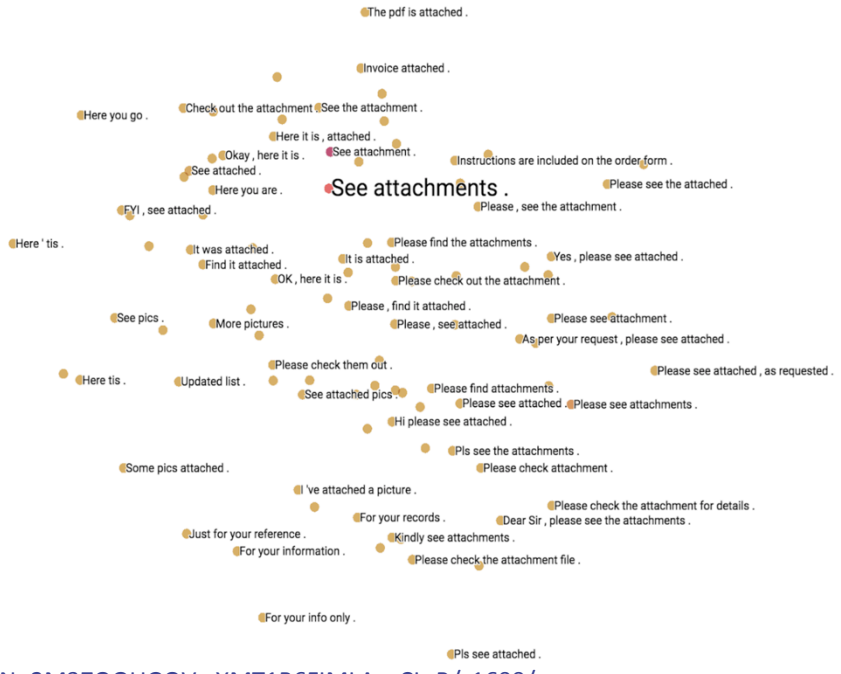
$$\begin{pmatrix}2 & 3\\ 2 & 1\end{pmatrix}x\begin{pmatrix}3\\ 2\end{pmatrix}=\begin{pmatrix}12\\ 8\end{pmatrix}=4x\begin{pmatrix}3\\ 2\end{pmatrix}$$

- $A\mathbf{x}=\lambda\mathbf{x} \Leftrightarrow (A-\lambda I)\mathbf{x}=0$
- How to calculate **x** and :
  - Calculate $det(A-\lambda I)$, yields a polynomial (degree n)
  - Determine roots to $det(A-\lambda I)=0$, roots are eigenvalues λ
  - Solve $(A-\lambda I)\mathbf{x}=0$ for each λ to obtain eigenvectors **x**

Sources: http://www.cse.buffalo.edu/faculty/azhang/data-mining/pca.ppt

Principal Components

- principal component (PC1)
  - The eigenvalue with the largest absolute value will indicate that the data have the largest variance along its eigenvector, the direction along which there is greatest variation
- principal component (PC2)
  - the direction with maximum variation left in data, orthogonal to the PC1
- In general, only few directions manage to capture most of the variability in the data.



Sources: http://www.cse.buffalo.edu/faculty/azhang/data-mining/pca.ppt
https://1.bp.blogspot.com/-5vEgY1mh1cA/WEfRAmER3iI/AAAAAAAABco/beMK-6LNq2M37QOUGQVwXMT1B6FIMLAxgCLcB/s1600/image00.png

Steps of PCA

- Let $\overline{X}$ be the mean vector(taking the mean of all rows)
- Adjust the original data by the mean X' = X – $\overline{X}$
- Compute the covariance matrix C of adjusted X
- Find the eigenvectors and eigenvalues of C.
- For matrix *C, v*ectors **e** (=column vector) having same direction as *C***e** :
  - *eigenvectors* of *C* is **e** such that *C***e**= λ **e**,
  - λ is called an *eigenvalue* of *C*.
- *C***e**= λ **e** $\Leftrightarrow$ (*C*- λ I)**e**=0

Sources: http://www.cse.buffalo.edu/faculty/azhang/data-mining/pca.ppt

Transformed Data

- Eigenvalues λ_j corresponds to variance on each component j
- *Thus, sort by* λ_j
- Take the first p eigenvectors $\mathbf{e_i}$ where p is the number of top eigenvalues
- These are the directions with the largest variances

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{ip} \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_p \end{pmatrix} \begin{pmatrix} x_{i1} - \overline{x_1} \\ x_{i2} - \overline{x_2} \\ \dots \\ x_{in} - \overline{x_n} \end{pmatrix}$$

Sources: http://www.cse.buffalo.edu/faculty/azhang/data-mining/pca.ppt

Example

X1	X2	X1'	X2'
19	63	-5.1	9.25
39	74	14.9	20.25
30	87	5.9	33.25
30	23	5.9	-30.75
15	35	-9.1	-18.75
15	43	-9.1	-10.75
15	32	-9.1	-21.75
30	73	5.9	19.25

Mean1=24.1
Mean2=53.8

$$C = \begin{pmatrix} 75 & 106 \\ 106 & 482 \end{pmatrix}$$

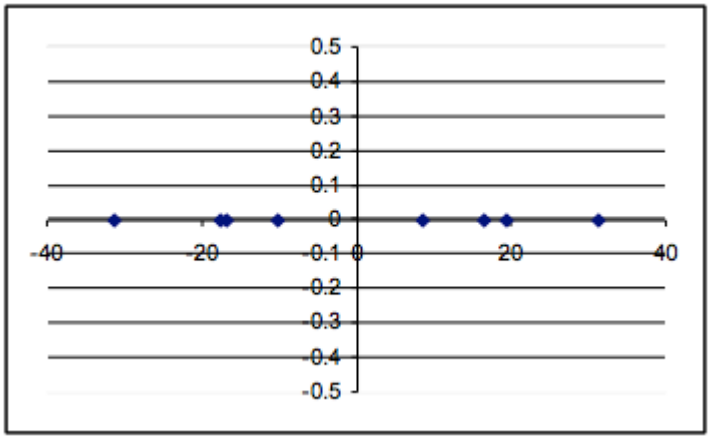
- Eigenvectors:
- e1=(-0.98,-0.21), l1=51.8
- e2=(0.21,-0.98), l2=560.2
- Thus the second eigenvector is more important!

Sources: http://www.cse.buffalo.edu/faculty/azhang/data-mining/pca.ppt

Keeping only one dimension

- We keep the dimension of e2=(0.21,-0.98)
- We can obtain the final data as

$$y_i = \begin{pmatrix} 0.21 & -0.98 \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} = 0.21 * x_{i1} - 0.98 * x_{i2}$$



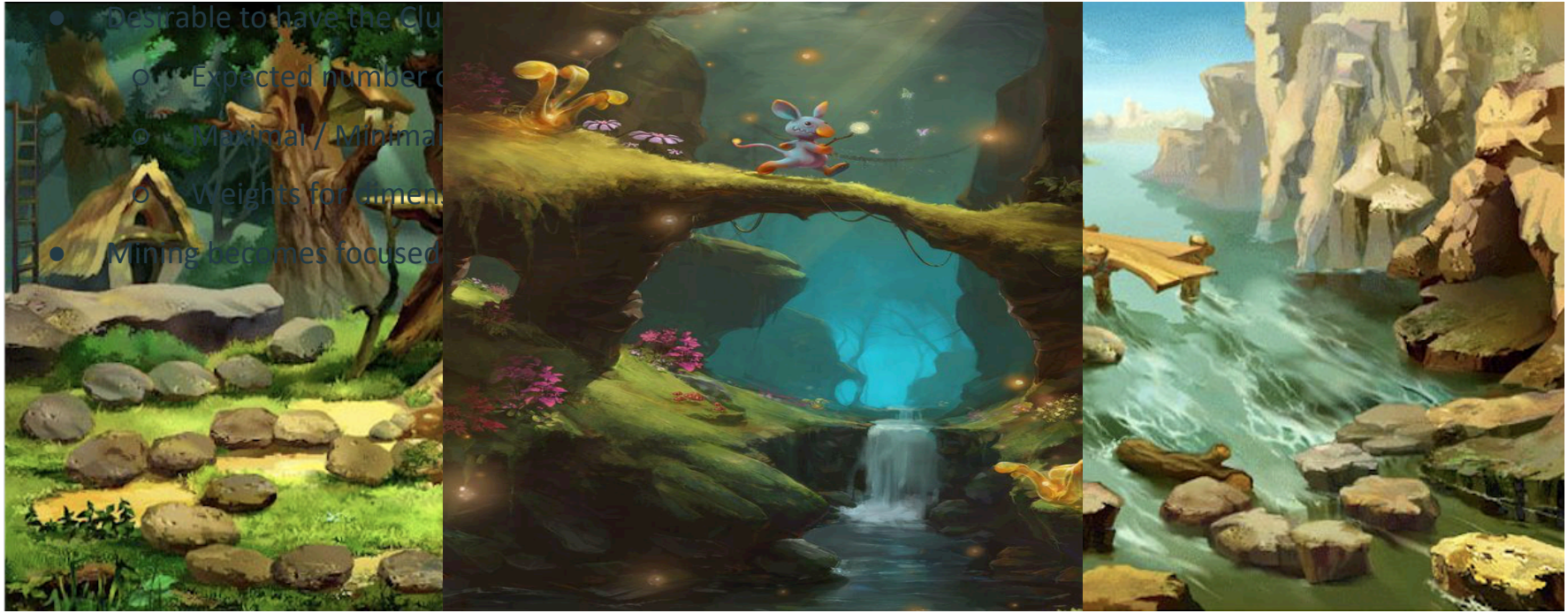
yi
-10.14
-16.72
-31.35
31.374
16.464
8.624
19.404
-17.63

Sources:

http://www.cse.buffalo.edu/faculty/azhang/data-mining/pca.ppt

Constraint Based Clustering

- Constraint based Clustering – finds clusters that satisfy user-specified preferences or constraints.
  - Desirable to have the clu
    - Expected number
    - Maximal / Minimal
    - Weights for dimen
  - Mining becomes focused



Source: *https://in.pinterest.com/pin/392728029979811507/*

Categories of Constraints

- Constraints on Individual objects
  - Ex: Luxury mansions worth over a million dollars
  - Processed through selection
- Constraints on the selection of Clustering parameters
  - Number of clusters, radius, MinPts
- Constraints on distance or similarity functions
  - Different measures for specific attributes / Objects
  - Weighting process – Clustering with obstacle objects
- User specified constraints on properties of individual clusters
  - Clusters satisfy given properties

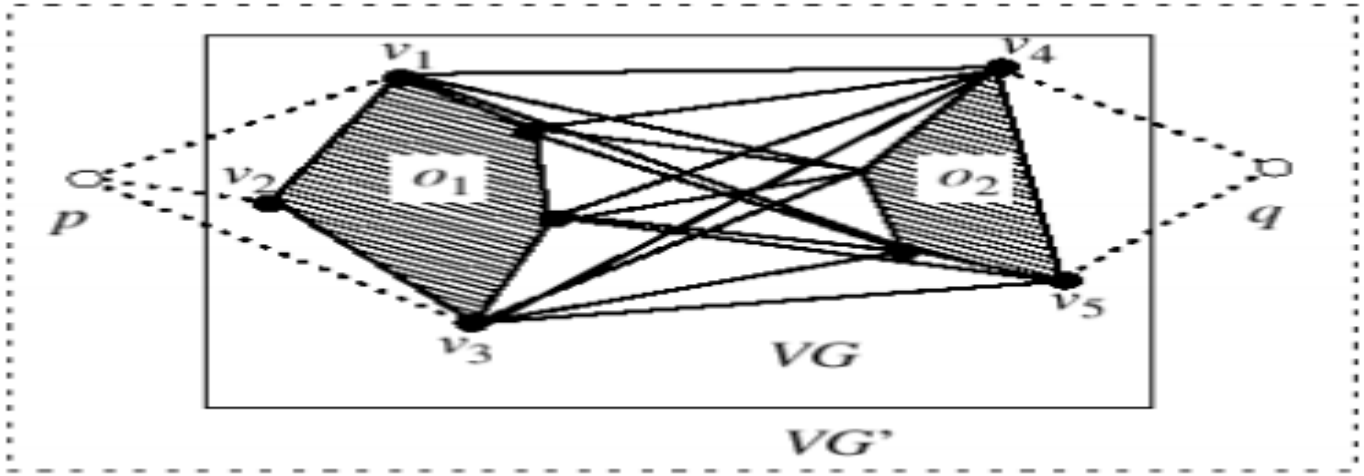
Source: *https://in.pinterest.com/pin/392728029979811507/*

Clustering with Obstacle objects

- City – rivers, lakes, bridges, roads etc
- Obstacles must be avoided
- Distance function between objects must be re-defined
  - Straight line distance is meaningless
- When using a partitioning approach – distance calculation with obstacles becomes expensive
  - k-means – not suitable as cluster centre may lie on an obstacle
  - Distance between objects can be determined using triangulation

Clustering with Obstacles

- Point p is visible from q in region R if straight line between p and q does not intersect any obstacle
- Visibility graph - VG
  - Each vertex of the obstacle has a corresponding node
  - Edge between two vertices only if they are visible to each other
  - Additional points can be added and paths can be determined



Source https://www.slideshare.net/Krish_ver2/36-constraint-based-cluster-analysis

User-Constrained Cluster Analysis

- Example: Relocation package delivery centres
  - N customers : high-value and ordinary customers
  - Determine locations for k service stations
  - Constraints
    - Each station should server
      - At least 100 high value customers
      - At least 5000 ordinary customers
- Constrained Optimization problem
  - Direct Mathematical approach is expensive.

User-Constrained Cluster Analysis

- Micro-Clustering
- Instead of points can work on micro-clusters
- Initially find a partition of k-groups satisfying given constraints
- Iteratively refine solution
  - Move m customers from cluster Ci to Cj if Ci has at least m surplus customers
  - Movement done if total sum of distances (objects – Centers) is reduced
  - Can be directed by selecting promising points
  - Deadlock has to be avoided (constraint cannot be satisfied).

Summary

What you should remember.

- Need for Clustering? Why is it useful?
- Different types of Cluster Algorithms.
- Applications of Clustering
- Features of good clustering approach
- Importance of clustering in data mining
- Different distance metrics
- Clustering High Dimensional Data
- Principal Component Analysis for Dimensionality Reduction
- Constrained Cluster Analysis

Part 2

PROMPT: Personalized User Tag Recommendation for Social Media Photos Leveraging Personal and Social Contexts

Authors:

- Rajiv Ratn Shah: National University of Singapore
- Anupam Samanta: IIT Dhanbad, India
- Deepak Gupta: IIT Dhanbad, India
- Et.al.

Published in **2016 IEEE** International Symposium on Multimedia (**ISM),** San Jose, CA, USA.

Date of Conference: 11-13 Dec. 2016

Date Added to IEEE *Xplore*: 19 January 2017 (http://ieeexplore.ieee.org/document/7823676/)

Goal

- Number of photos on social media platforms has increased rapidly (*e.g.*, Flickr has over ten billion photos)
- Requires an automatic tag recommendation system for an efficient multimedia search and retrieval
- User tags are very helpful in providing several significant multimedia-related applications such as a landmark recognition
- **PROMPT**, that recommends personalized tags for a given photo leveraging personal and social contexts of user

Dataset

- Yahoo Flickr Creative Commons 100 Million (**YFCC100m**) dataset, a collection of 100 million media records from Flickr.

- We used 28 million photos for the train set and 46,700 photos for the test set.

- For each image in the dataset contains metadata such as user tags, spatial and temporal information, image_id, user_id, image_web_link and visual_tags.

- **2656493496** **17230000@N02** eef+llc **2008-06-21 19:53:57.0** 1215713339 NIKON+COOLPIX+L3 080621_gardaitaly_039 **garda,italy,lake** **http://www.flickr.com/photos/17230000@N02/2656493496/** http://farm4.staticflickr.com/3079/2656493496_9ae8b540ac.jpg Attribution-NonCommercial-NoDerivs License http://creativecommons.org/licenses/by-nc-nd/2.0/ 3079 4 9ae8b540ac e3ef024a94 jpg 0

Methodology

Steps of recommending user tags by PROMPT for a social media photo is summarized as follows:

- It determines a group of users from the train set with 259,149 unique users, having similar tagging behavior as the user of the photo, based on cosine similarity.
- The candidate sets of photos and tags for the photo are computed from the selected user group.
- Relevance scores are computed for candidate tags using our proposed approaches. Finally, the top five user tags are recommended to the photo.

Methodology contd..

Determine a group of users who have similar tagging behavior as the given user

- In this study, we consider the 1,540 most frequent user tags from the YFCC100M dataset.
- We construct a 1,540-dimensional feature vector, called, the UTB vector, to represent a user's tagging behavior.
- We cluster users and their photos in the train set with 28 million photos into several groups based on cosine similarities among UTB vectors during pre-processing
- Moreover, we construct a 1,540-dimensional feature vector for a given photo, called, the photo description (PD) vector.
- Compute the photo's N nearest semantically similar neighbors.
- UTB and PD vectors help PROMPT to find an appropriate set of candidate photos and tags for the given photo.

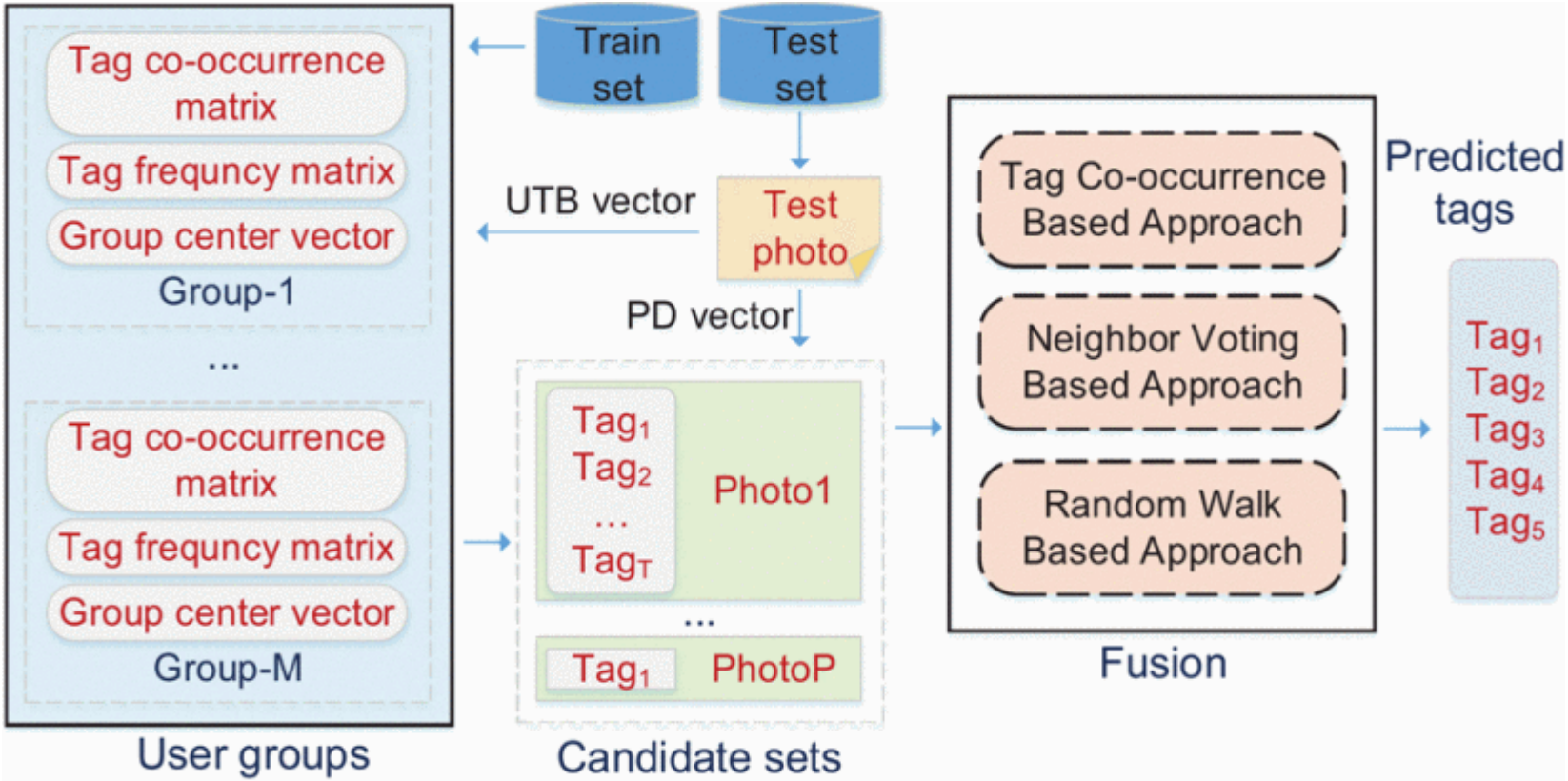


Figure. System overview of the PROMPT system.

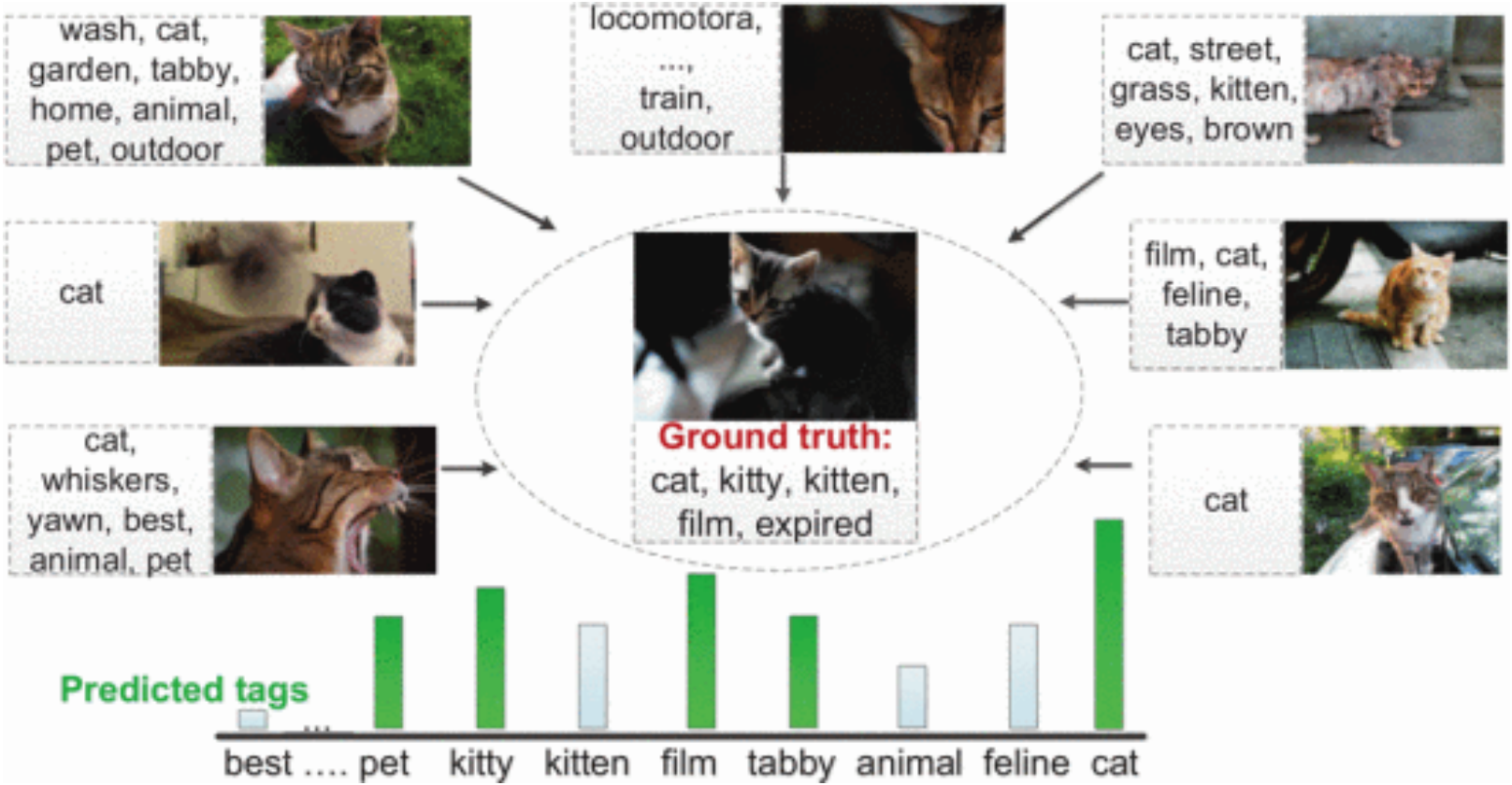


Figure. Selecting nearest images from a cluster for a given image

Evaluation

- Tags in T fulfill the following criteria:
  - Valid English dictionary words
  - Do not refer to persons, dates, times or places
  - Appear frequently with photos in the train and test sets
  - Different tenses/plurals (tags) of the same word (an already added tag in *T*) are not considered.
- Recommended tags for a given photo in the test set are evaluated based on the following three metrics:
  - Precision@*K, i.e.*, proportion of the top K predicted tags that appear in user tags of the photo,
  - Recall@K, *i.e.*, proportion of the user tags that appear in the top K predicted tags
  - *Accuracy@K, i.e.*, 1 if at least one of the top K predicted tags is present in the user tags, 0 otherwise
- PROMPT is tested for the following values of *K:*1, 3, and 5.

Numbers

	Comparison	K = 1	K = 3	K = 5
Accuracy@K	Type-1	0.410	0.662	0.746
	Type-2	0.422	0.678	0.763
Precision@K	Type-1	0.410	0.315	0.251
	Type-2	0.422	0.326	0.262
Recall@K	Type-1	0.062	0.142	0.188
	Type-2	0.064	0.147	0.197

Figure. Results for the top K predicted tags

Results

User tag list (ground truth):
tent, black, beach, hotel, resort, tourist, holiday

Predicted tag list:
beach, sea, coast, shore, nature

User tag list (ground truth):
cat, kitty, kitten, film, expired

Predicted tag list:
cat, film, kitty, pet, tabby

User tag list (ground truth):
surf, ocean, coast, sunset, foam, silhouette

Predicted tag list:
beach, coast, water, ocean, sunset

Thank you.

Team Number - 6