Clustering

Prof. Dra. Silvia Schiaffino
ISISTAN

sschia@exa.unicen.edu.ar

Clustering: Concepto

- Cluster: un número de cosas o personas similares o cercanas, agrupadas

- Clustering: es el proceso de particionar un conjunto de objetos (datos) en un conjunto de sub-clases con cierto significado

Clustering: Concepto

- Clustering es un proceso de aprendizaje no supervisado:
  - Las clases no están predefinidas sino que deben ser descubiertas dentro de los ejemplos
- Primariamente es un método descriptivo para interpretar un conjunto de datos
- Particionar ejemplos de clases desconocidas en subconjuntos disjuntos de clusters tal que:
  - Ejemplos en un mismo cluster sean altamente similares entre sí
  - Ejemplos en diferentes clusters sean altamente disimiles entre sí

Aplicaciones

- Reconocimiento de patrones
- Procesamiento de imágenes
  - Reconocer áreas con cierta característica de tierra (GIS)
- Investigación de mercado
  - Detectar diferentes tipos de clientes según su forma de compra
- Web mining
  - Categorización de documentos
  - Clustering en Web logs

- Como preprocesamiento para otras técnicas de data mining

Características

- Un buen método de clustering debería identificar clusters que sean tanto compactos somo separados entre sí. Es decir, que tengan:
  - Alta similaridad intra-cluster
  - Baja similaridad inter-cluster
- Un buen método debiera descubrir algunos o todos los patrones ocultos en los datos
- La calidad del método de clustering depende de la medida de similitud
- La forma de representación del cluster también permite evaluar al método y afecta al resultado

Hard clustering vs. Soft clustering

- El clustering típicamente asume que cada instancia pertenece a exactamente un cluster, hard clustering. No permite incertidumbre acerca de la membrecia de una instancia a cada cluster en un conjunto de clusters

- Soft clustering asigna probabilidades de pertenencia de una instancia a más de un cluster

Principales técnicas de clustering

- Existen diferentes enfoques:
  - Algoritmos de particionamiento
  - Algoritmos jerárquicos
  - Métodos basado en densidad (BD espaciales)
  - Métodos basado en grilla
  - Métodos basado en modelo

Alg. Basados en Particionamiento

Construyen una partición del conjunto de datos *D* de *n* objetos en un conjunto de *k* clusters

- Dado un k, intentan encontrar una partición de k clusters que optimiza el criterio de particionamiento
  - Global optimal: enumerar exhaustivamente todas las particiones
  - k-means [McQueen67]: cada cluster es representado por su centro del cluster
  - K-modes [Huang98]: idem para datos categóricos
  - K-medoids or PAM (Partition around medoids) [Kaufman87]: cada cluster se representa por uno de los objetos del cluster

K-Means

- Asume que las instancias son vectores de valores reales
- Los clusters se basan en centroides o centros de gravedad, o la media de las instancias en el cluster c:

$$\mu(c) = \frac{1}{|c|} \sum_{x \in c} x$$

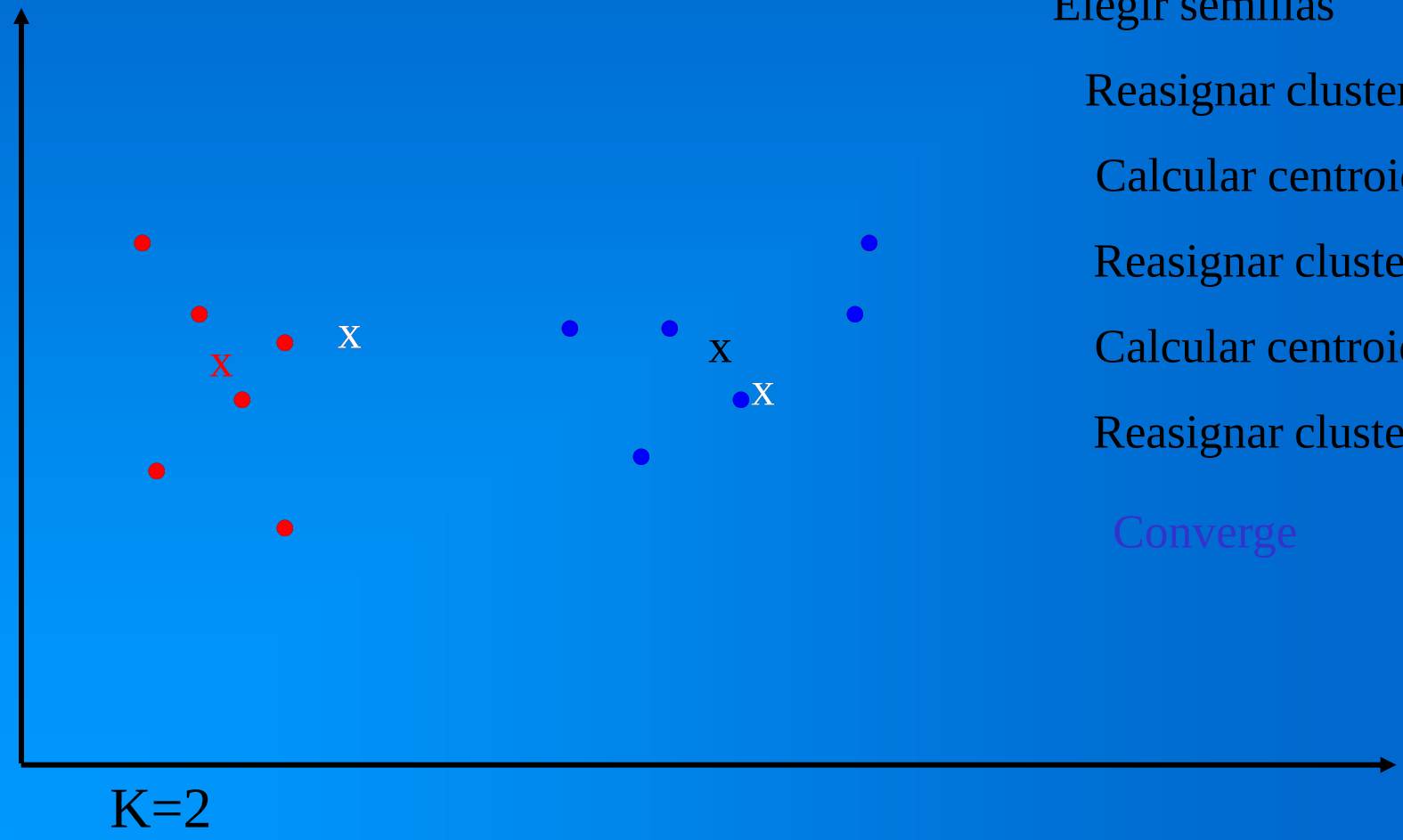
- Las instancias se reasignan a clusters en base a su distancia a los centroides

K-Means

Dado k (número de clusters) y el conjunto de datos n:

1. Arbitrariamente elegir k objetos como centros iniciales de cluster (semillas)
2. Repetir:
    3. (re)asignar cada objeto al cluster con el cual el objeto sea más similar, basándose en el valor medio de los objetos del cluster
    4. Actualizar los valores medios del cluster (centroides), es decir calcular el valor medio de los objetos para cada cluster
5. Hasta que no se produzcan cambios (convergencia)

K-Means



Elegir semillas

Reasignar clusters

Calcular centroides

Reasignar clusters

Calcular centroides

Reasignar clusters

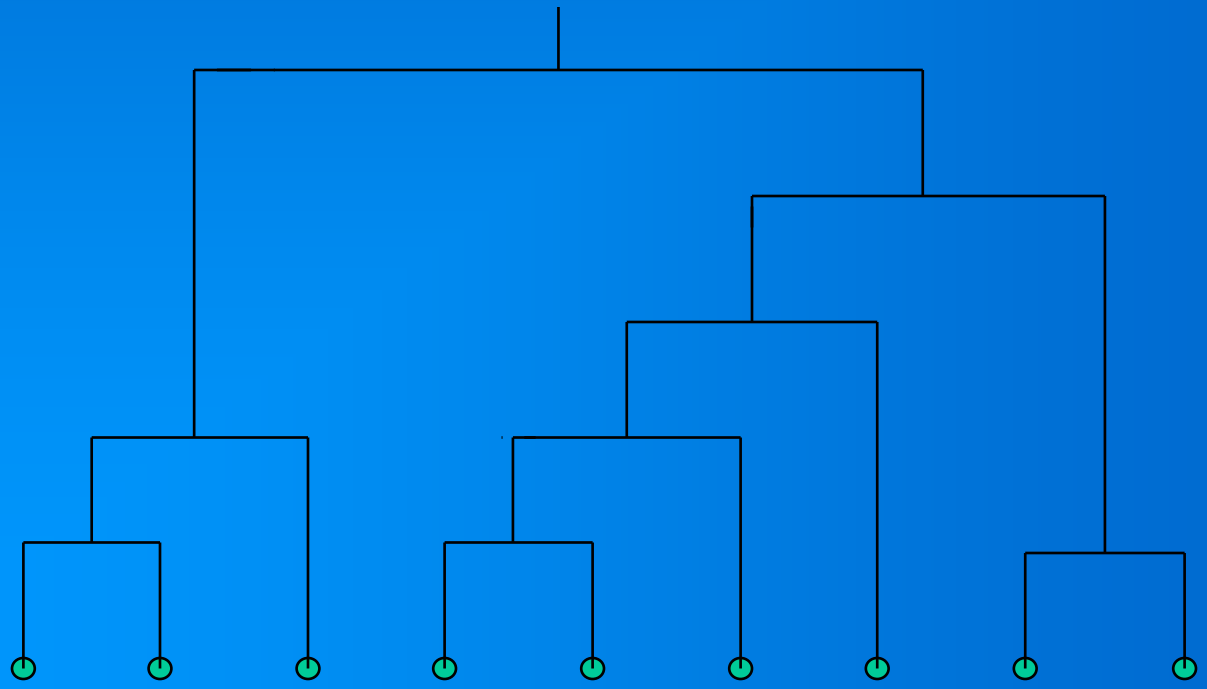
Converge

K-Means

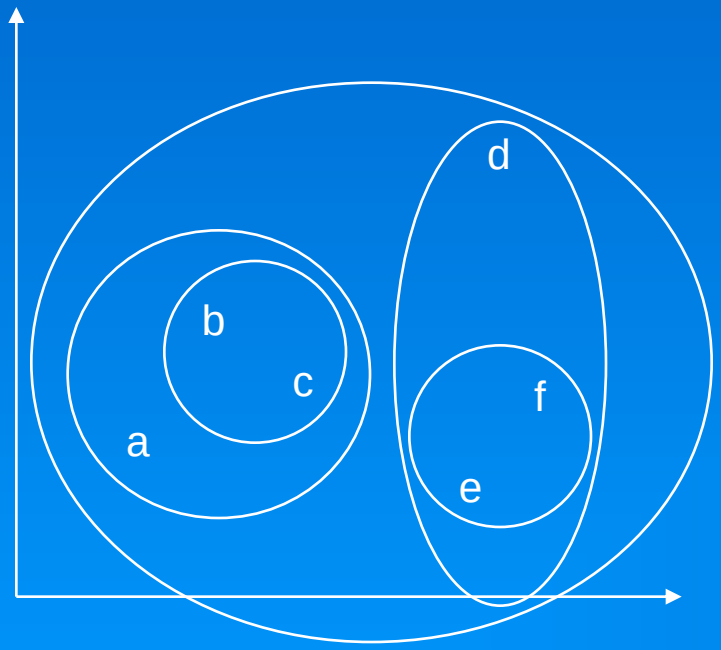
- Ventajas:
  - Entre los algoritmos de particionamiento es eficiente
  - Implementación sencilla
- Desventajas:
  - Necesito conocer *k* de antemano
  - Sensible a ruido
  - El resultado puede variar en base a las semillas elegidas al inicio
  - Algunas semillas pueden resultar en una tasa de convergencia menor
  - La selección de semillas se puede basar en heurísticas o resultados obtenidos por otros métodos
  - Puede caer en mínimos locales
  - No trata datos nominales (K-Modes)

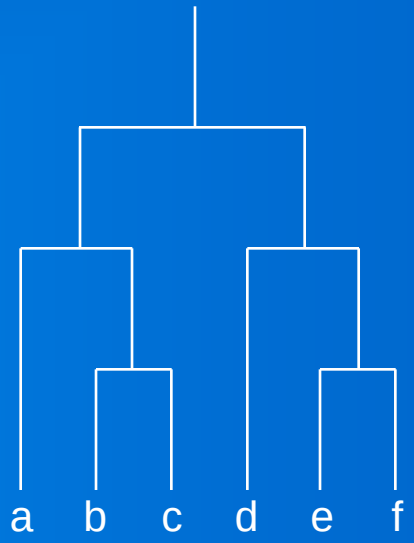
Clustering Jerárquico

- Un dendograma muestra como se mezclan los clusters de manera que cortando el dendograma en diferentes niveles se consiguen differentes clusters



Representaciones



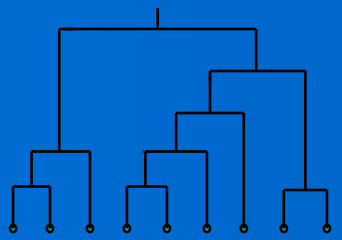


Clustering Jerárquico

- Construye un árbol binario o dendograma a partir de un conjunto de ejemplos:
  - *Aglomerativo* (*bottom-up*) métodos que comienzan con cada ejemplo en un cluster diferente y combinan iterativamente los clusters para formar clusters mayores (Ej. AGNES, Agglomerative Nesting [Kaufman90])
  - *Divisivo* (*top-down*) métodos que comienzan con todos los ejemplos en un mismo cluster y los separan en clusters más chicos (Ej. DIANA, Divisive Analysis [Kaufman90])

Clustering Jerárquico

- Clustering Aglomerativo:
  - Asume una función de similitud para determinar la similitud de dos instancias
  - Comienza con todas las instancias en un cluster separado (cada instancia es un cluster) y junta los dos clusters que son más similares en cada paso hasta que queda un único cluster o se cumple un criterio de terminación
    - Entre todos los cluster existentes determinar los dos clusters ci y cj que son más similares
    - Reemplazar ci y cj por un único cluster ci $\cup$ cj
  - Crea un árbol binario



Clustering Jerárquico

- Asume una función de similitud que determina la similitud de dos instancias: *sim*(*x*,*y*)
  - Por ejemplo la similitud del coseno o coeficiente de correlación de Pearson
- Asume una función de similitud que determina la similitud de dos clusters conteniendo multiples instancias:
  - Single Link: similitud de los dos elementos del cluster más similares
  - Complete Link: similitud de los dos elementos del cluster menos similares
  - Group Average: promedio de similitudes entre los elementos del cluster

Similitud entre elementos

Distancia euclideana

$$\vec{x} = \begin{bmatrix} x_1 \\ \mathrm{M} \\ x_N \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_1 \\ \mathrm{M} \\ y_N \end{bmatrix}$$

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{n=1}^{N} (x_n - y_n)^2}$$

Similitud entre elementos

Similitud del coseno

$$\vec{x} = \begin{bmatrix} x_1 \\ \text{M} \\ x_N \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_1 \\ \text{M} \\ y_N \end{bmatrix}$$

$$C_{\text{cosine}}(\vec{x}, \vec{y}) = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i \times y_i}{\|\vec{x}\| \times \|\vec{y}\|}$$

$\vec{x} = \vec{y}$ $\qquad +1 \geq \text{Correlación del coseno} \geq -1 \qquad \vec{x} = -\vec{y}$

Similitud entre elementos

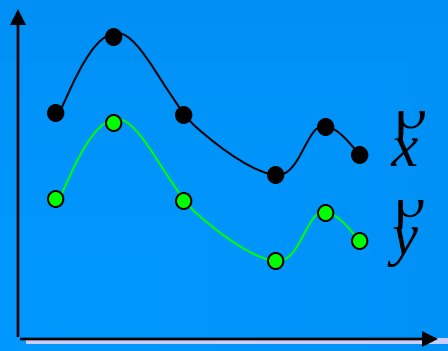
Coeficiente de correlación de Pearson

Dos vectores $\vec{x} = \begin{bmatrix} x_1 \\ M \\ x_N \end{bmatrix}$ e $\vec{y} = \begin{bmatrix} y_1 \\ M \\ y_N \end{bmatrix}$

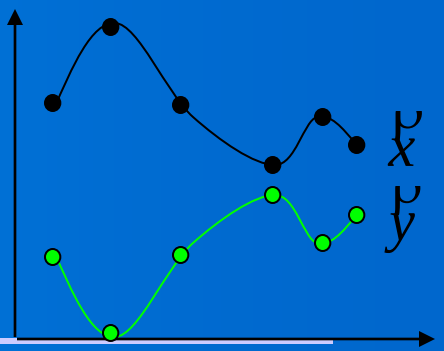
$$C_{pearson}(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{N}(x_i - m_x)(y_i - m_y)}{\sqrt{[\sum_{i=1}^{N}(x_i - m_x)^2][\sum_{i=1}^{N}(y_i - m_y)^2]}}$$

$$m_x = \frac{1}{N}\sum_{n=1}^{N} x_n$$

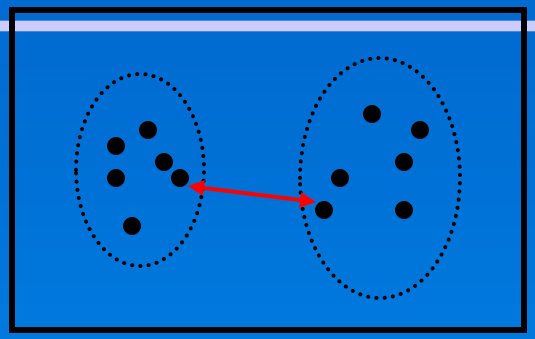
$$m_y = \frac{1}{N}\sum_{n=1}^{N} y_n$$



$+1 \geq$ Pearson Correlation ≥ -1



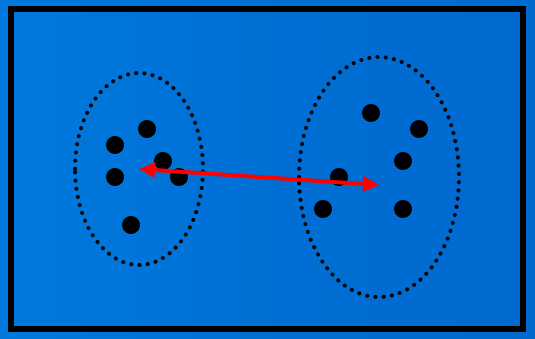
Similitud entre clusters

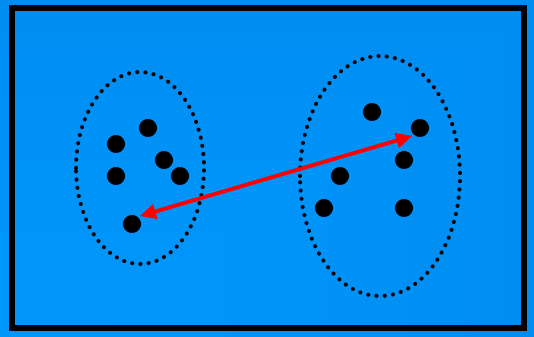


Single Link

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

Average Link



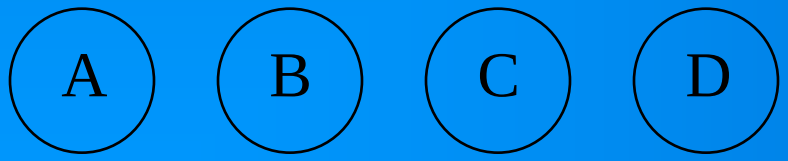


Complete Link

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

Clustering de documentos

Documentos iniciales

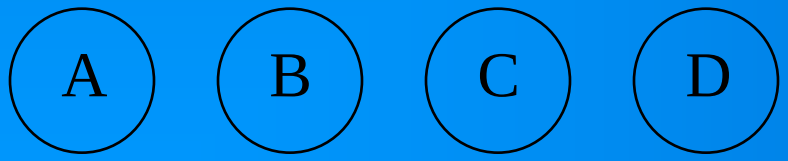


Matriz de Distancias

Dist	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

Clustering de documentos

Documentos iniciales

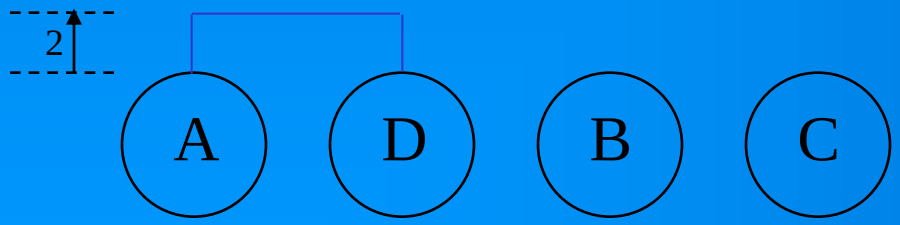


Matriz de Distancias

Dist	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

Clustering de documentos

Clusters actuales

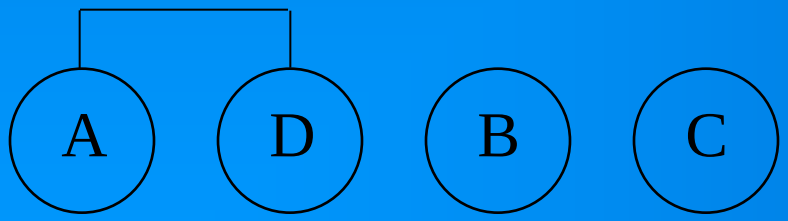


Matriz de Distancias

Dist	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

Clustering de documentos

Clusters actuales

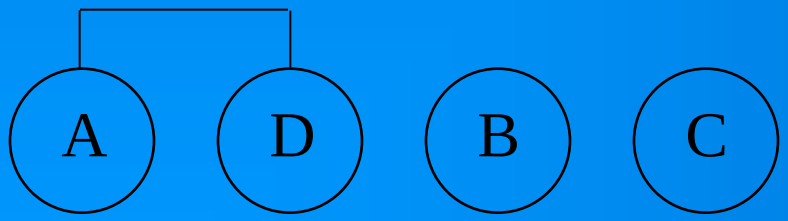


Matriz de Distancias

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Clustering de documentos

Clusters actuales

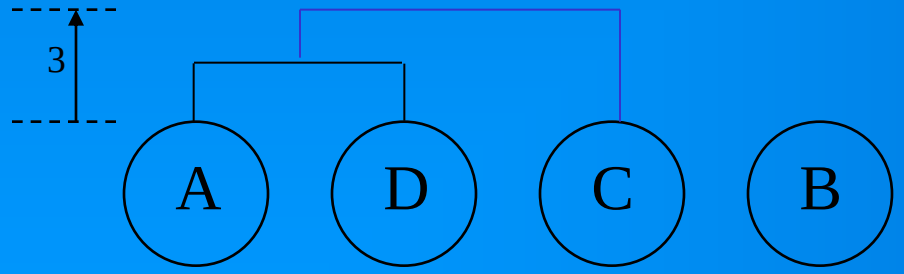


Matriz de Distancias

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Clustering de documentos

Clusters actuales

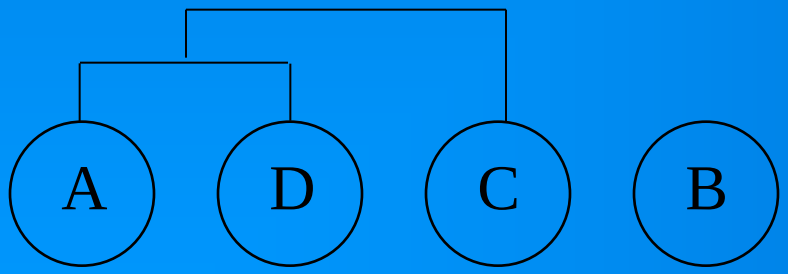


Matriz de Distancias

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Clustering de documentos

Clusters actuales

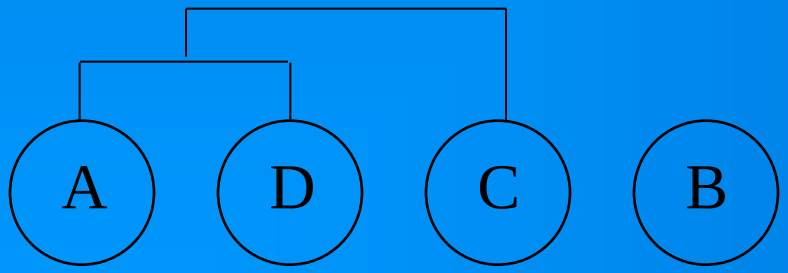


Matriz de Distancias

Dist	AD C	B		
AD C		10		
B				

Clustering de documentos

Clusters actuales

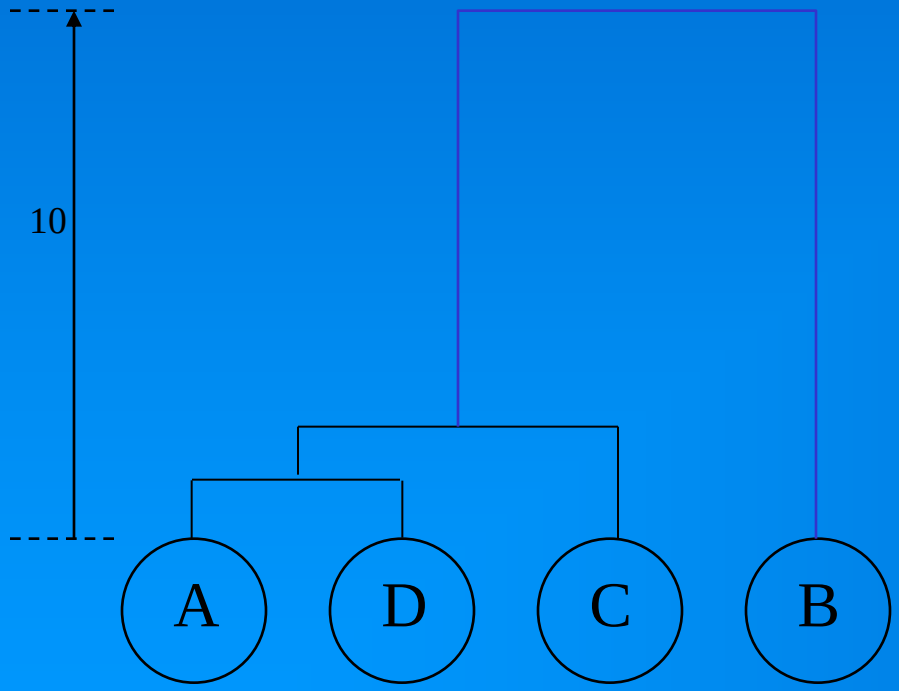


Matriz de Distancias

Dist	AD C	B		
AD C		10		
B				

Clustering de documentos

Clusters actuales

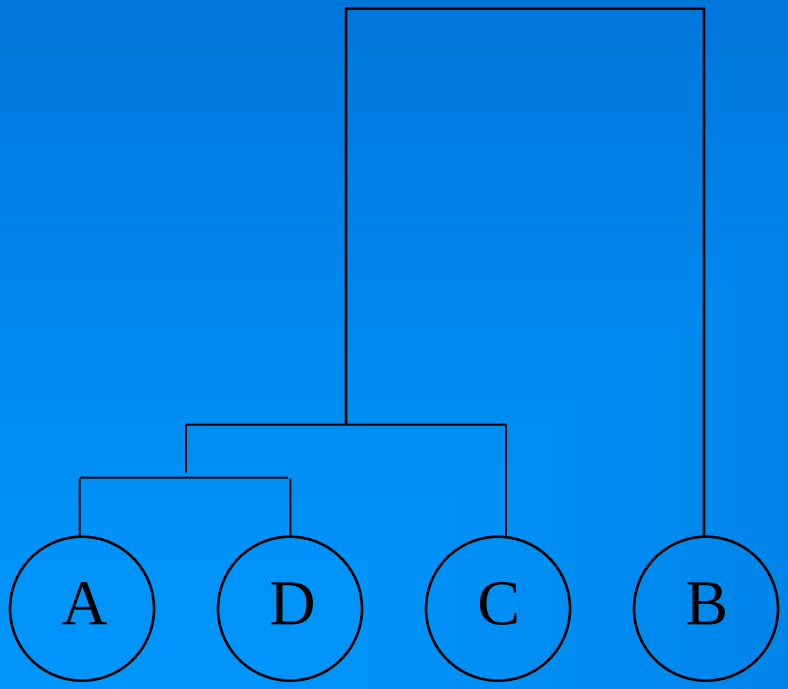


Matriz de Distancias

Dist	AD C	B		
AD C		10		
B				

Clustering de documentos

Clusters actuales

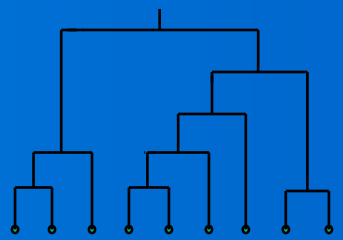


Matriz de Distancias

Dist	AD CB			
AD CB				

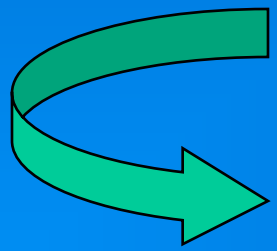
Clustering Jerárquico

- Ventajas:
  - No se necesita conocer el número de clusters k
- Desventajas:
  - No puede recuperarse de decisiones incorrectas
  - Computacionalmente costoso



Ejemplo de uso: Categorización de texto

- Categorización de Textos
  - Asignar a un conjunto de documentos $D=\{d_1,d_2,...,d_n\}$ un conjunto de categorías $C=\{c_1,c_2,...,c_m\}$



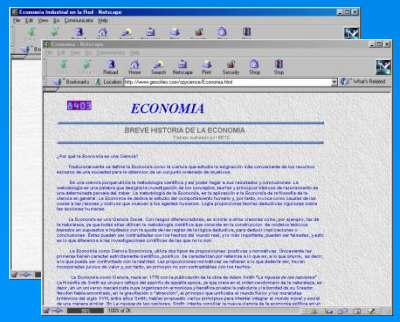
Categorización de Jerárquica de Textos

- Las categorías poseen relaciones semánticas de manera que documentos que pertenecen a una categoría c_i también lo hacen a una categoría padre c_j

Clusters representan jerarquías de temas

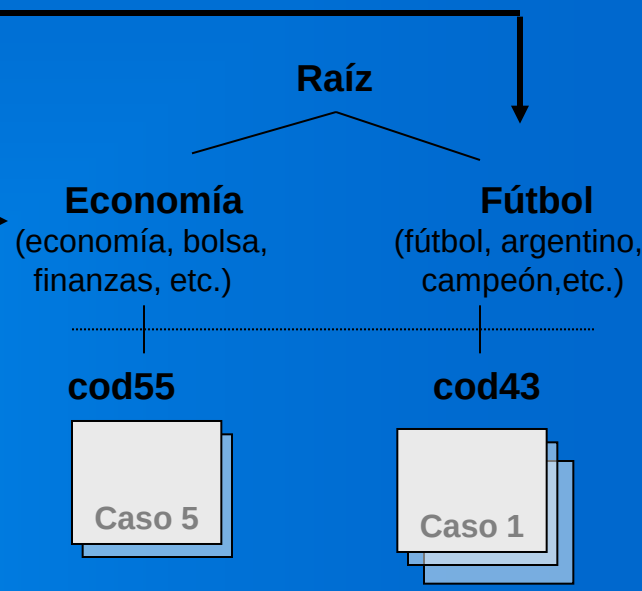
cod55

Fútbol, 0.98
Argentino, 0.81
Campeón, 0.72
Eliminatoria, 0.57
América, 0.43
Club, 0.31
Competición, 0.20
...



cod43

Economía, 0.96
Bolsa, 0.84
Finanzas, 0.82
Dólar, 0.67
Mercado, 0.41
Mercosur, 0.51
Investigación, 0.10
...

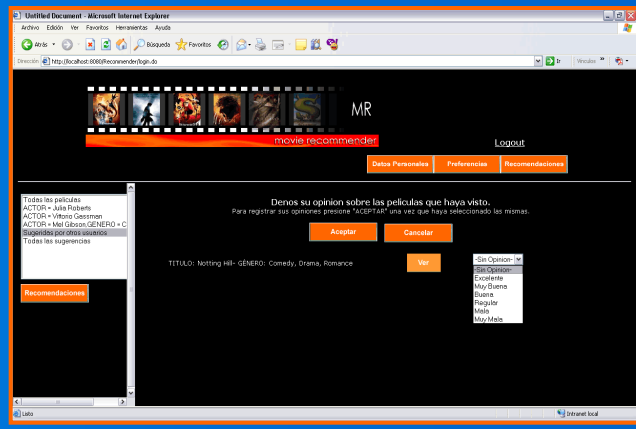


Ejemplo: Comunidades de usuarios en sistemas de recomendación

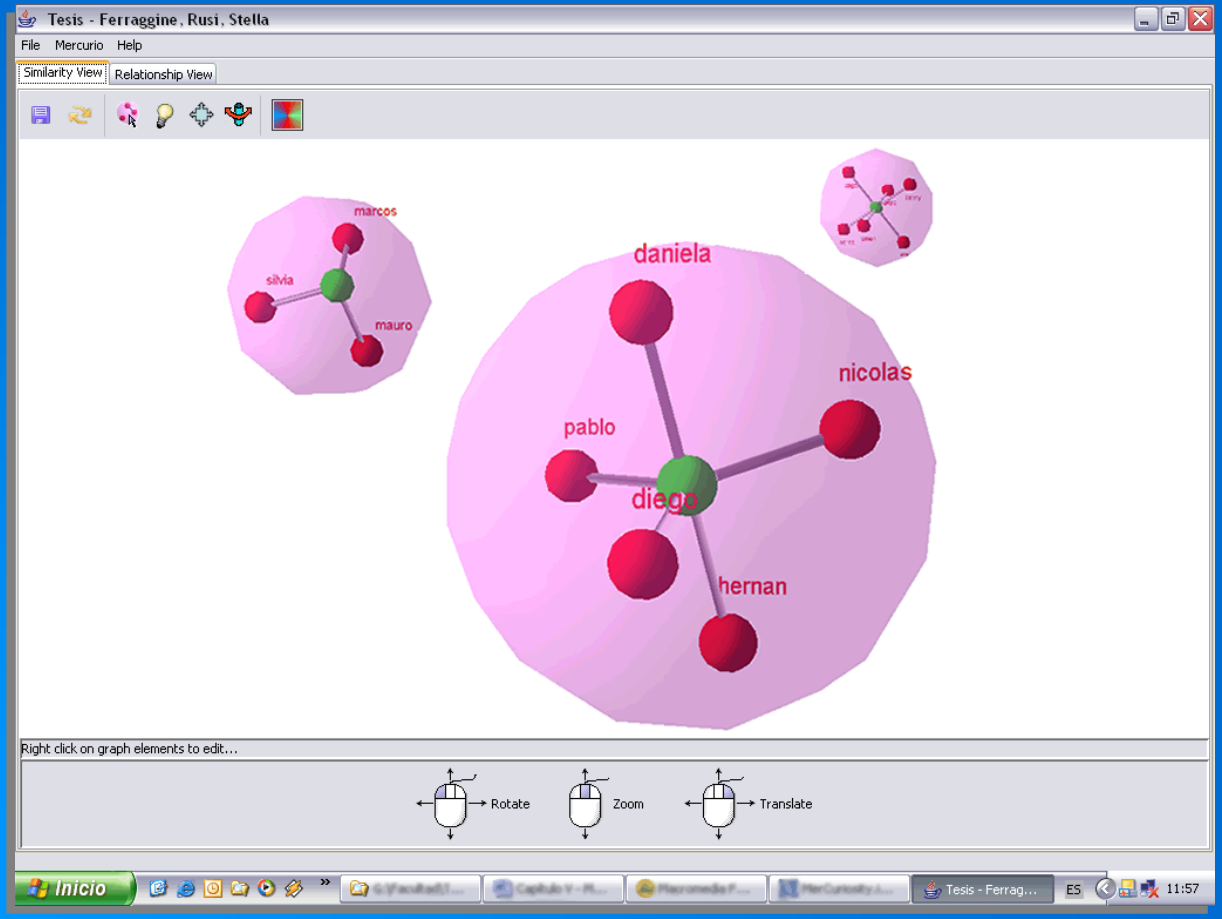
- MovieRecommender: sistema de recomendación de películas que utiliza filtrado colaborativo

- Sugiere películas basándose en la similitud de intereses del usuario actual con otros usuarios “cercanos”

- Análisis de diferentes algoritmos de clustering para armar comunidades de usuarios similares



Ejemplo: Comunidades de usuarios en sistemas de recomendación



Otros algoritmos de clustering

- Basados en densidad
  - DBSCAN [Ester et al, 96]
  - OPTICS [Ankerst et al, 99]
  - DENCLUE [Hinnenburg & Keim, 98]
  - CLIQUE [Agrawal et al 98]
- Basados en Grilla
  - STING [Wang et al 97]
  - BAN-Clustering/GRIDCLUS [Schikuta97]
  - Wave-Clustering [Sheikholeslami et al 98]
  - CLIQUE [Agrawal et al 98]
- Datos categóricos
  - ROCK [Guha99]

Otros algoritmos de clustering

- Basados en modelo
  - Redes Neuronales
    - SOM: Self organizing maps [Kohonen 81]
  - Machine Learning
    - COBWEB (clustering conceptual) [Fisher 87]
  - Estadísticos
    - Gaussian Mixture Model [Raftery 93]
    - Autoclass (Bayesiano) [Cheeseman 96]

Preguntas?