


Clase 5 (B) - Agrupamiento de textos

Marcelo Errecalde^{1,2}

¹Universidad Nacional de San Luis, Argentina 

²Universidad Nacional de la Patagonia Austral, Argentina 

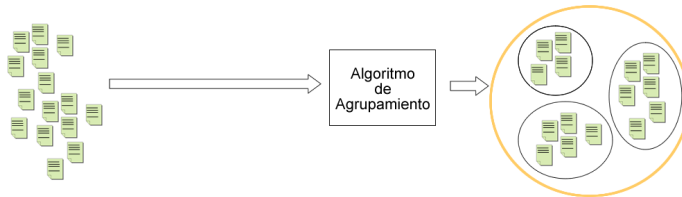


Curso: Minería de Textos
Facultad de Informática - Universidad Nacional de La Plata
23 al 27 de Septiembre de 2019

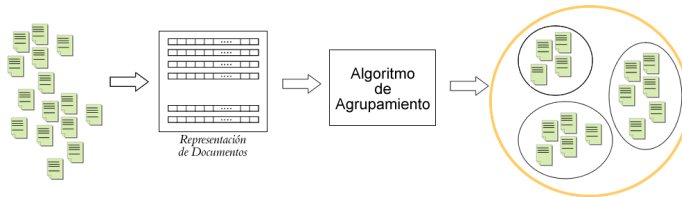
Resumen

- 1 **Agrupamiento de documentos**
 - ¿Qué es el Análisis de Clusters?
 - Tipos de Clustering
 - Medidas de Similitud
 - Validación de los agrupamientos

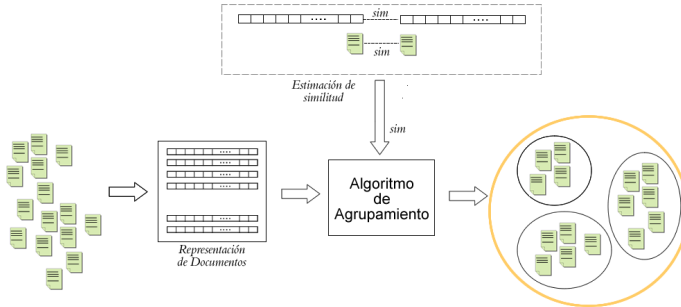
Breve descripción del agrupamiento de documentos



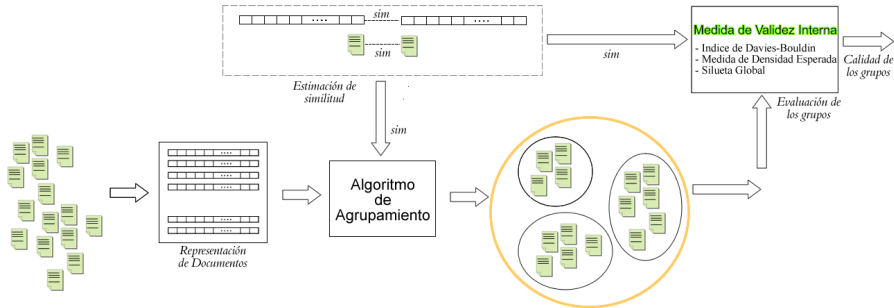
Breve descripción del agrupamiento de documentos



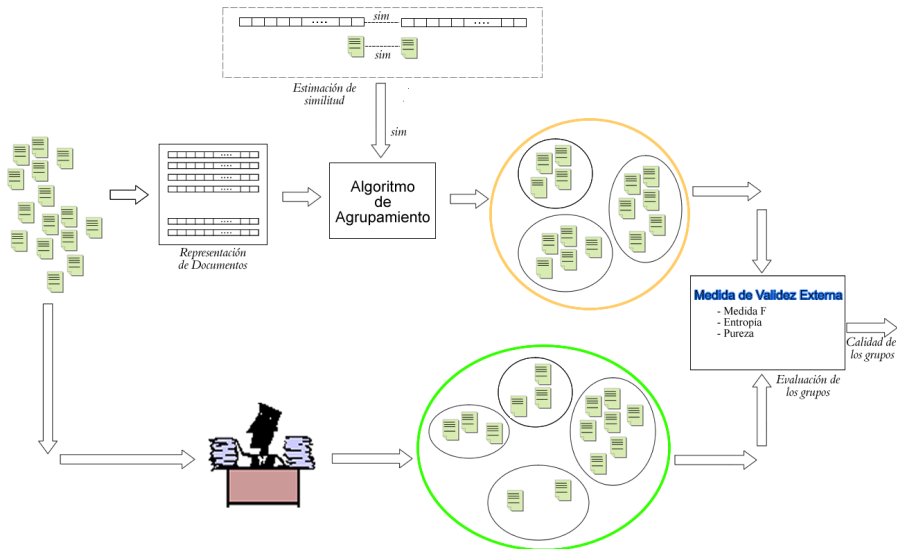
Breve descripción del agrupamiento de documentos



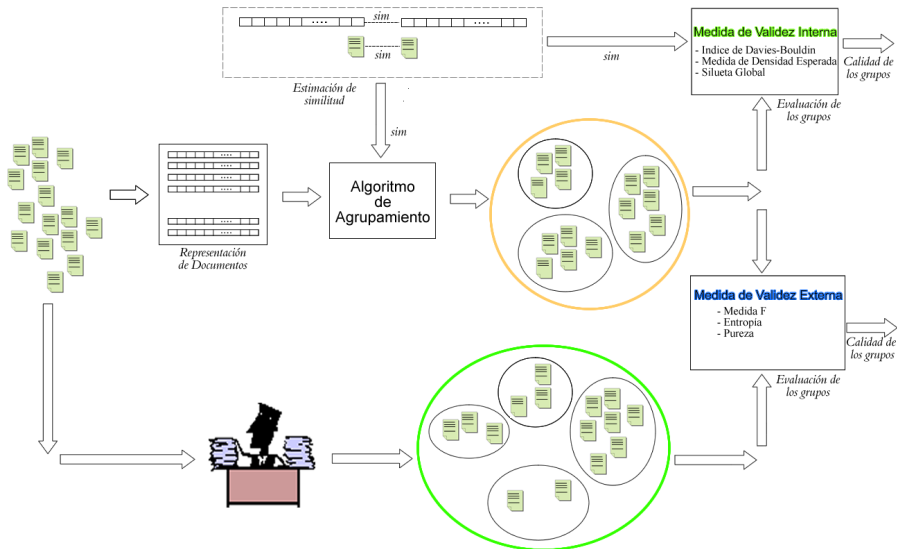
Breve descripción del agrupamiento de documentos



Breve descripción del agrupamiento de documentos



Breve descripción del agrupamiento de documentos



Definición

Análisis de Clusters

Proceso que divide los datos en **grupos** (**clusters**) que tienen un **significado**, que son **útiles**, o ambos.

- Grupos **significativos** \Rightarrow Los grupos deberían capturar la **estructura natural** de los datos.
- Grupos **útiles** \Rightarrow Los grupos sirven de base para otras técnicas de análisis y procesamiento de datos.

Análisis de Clusters

Grupos **significativos**

- Estos grupos mejoran nuestro **entendimiento** de los datos y las clases subyacentes.
- Rol fundamental en Biología, Recuperación de Información, Meteorología, Psicología y Medicina y Negocios.

Grupos **útiles**

- Hincapié en encontrar **prototipos de clusters** (objetos de datos representativos de los otros objetos del cluster).
- Rol fundamental en **resumir** grandes conjuntos de datos, **compresión** de imagen y sonido y búsqueda NN eficiente.

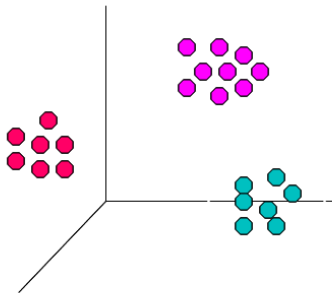
Encontrar grupos de objetos tal que los de un mismo grupo sean **similares** (o estén **relacionados**) y sean **diferentes** (o estén **poco relacionados**) con los objetos de los otros grupos.

- También conocida como **clasificación no supervisada**.
- Areas conectadas (pero no iguales) al Análisis de Cluster
 - **Particionado** \implies usualmente relacionado al particionado de **grafos** en **subgrafos**.
 - **Segmentado** \implies división de grupos mediante técnicas muy simples (ejemplo: segmentado de imágenes basado en el color y la intensidad de los pixels, o de personas de acuerdo a su ingreso)

Definición (más operativa)

Análisis de Clusters

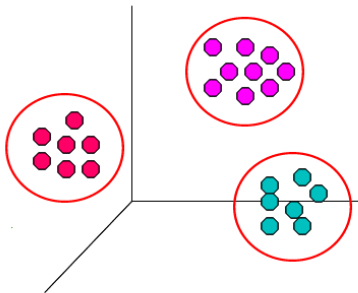
Encontrar grupos de objetos tal que los de un mismo grupo sean similares (o estén relacionados) y sean diferentes (o estén poco relacionados) con los objetos de los otros grupos.



Definición (más operativa)

Análisis de Clusters

Encontrar **grupos** de objetos tal que los de un mismo grupo sean similares (o estén relacionados) y sean diferentes (o estén poco relacionados) con los objetos de los otros grupos.

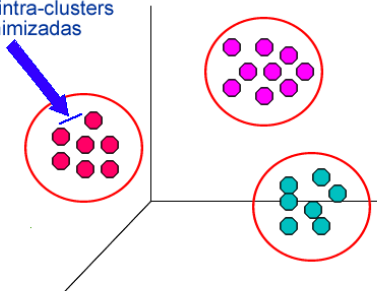


Definición (más operativa)

Análisis de Clusters

Encontrar grupos de objetos tal que los de un **mismo grupo** sean **similares** (o estén **relacionados**) y sean diferentes (o estén poco relacionados) con los objetos de los otros grupos.

Distancias intra-clusters
son minimizadas

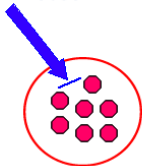


Definición (más operativa)

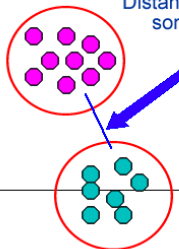
Análisis de Clusters

Encontrar grupos de objetos tal que los de un mismo grupo sean similares (o estén relacionados) y sean **diferentes** (o estén **poco relacionados**) con los objetos de los **otros grupos**.

Distancias intra-clusters
son minimizadas



Distancias inter-clusters
son maximizadas



¿Qué es el Análisis de Clusters?

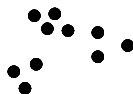
La noción de cluster es ambigua....



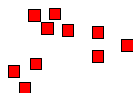
¿Cuántos Clusters?

¿Qué es el Análisis de Clusters?

La noción de cluster es ambigua....



¿Cuántos Clusters?



Dos Clusters

¿Qué es el Análisis de Clusters?

La noción de cluster es ambigua....



¿Cuántos Clusters?



Seis Clusters



Dos Clusters

¿Qué es el Análisis de Clusters?

La noción de cluster es ambigua....



¿Cuántos Clusters?



Seis Clusters



Dos Clusters



Cuatro Clusters

Tipos de clusterings (agrupamientos)

Un **clustering** (**agrupamiento**) es un conjunto de clusters.

Principal **distinción** entre tipos de agrupamientos.

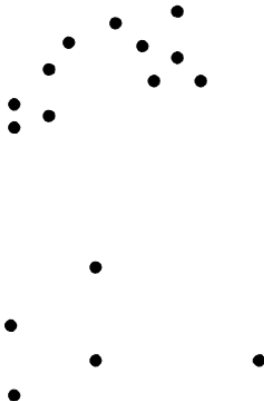
Clustering Particional

Los objetos de datos se dividen en subconjuntos (clusters) no solapados, tal que cada objeto pertenece a exactamente un subconjunto.

Clustering Jerárquico

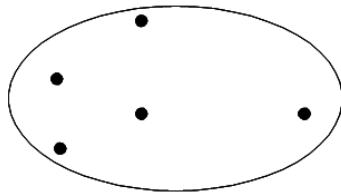
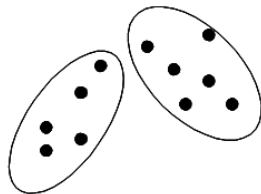
Conjunto de clusters anidados organizados como un árbol jerárquico.

Clustering Particional



Puntos Originales

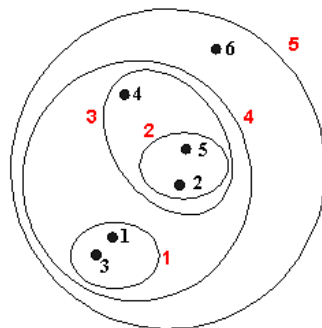
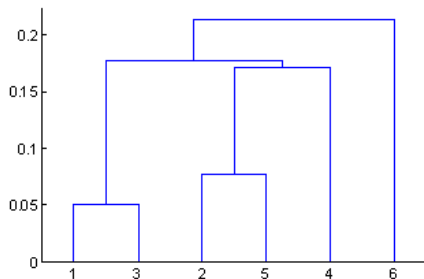
Clustering Particional



Puntos Originales

Un Clustering
Particional

Clustering Jerárquico



Otras distinciones de los agrupamientos

Exclusivo vs no exclusivo (NE)

En agrupamientos NE los objetos de datos pueden pertenecer a múltiples clusters.

Difuso vs no difuso

Agrupamiento **difuso**:

- Un objeto pertenece a cada cluster con un peso $w_i \in [0, 1]$
- Pesos deben sumar 1.
- Clustering probabilístico tiene características similares.

Parcial vs completo

En agrupamientos parciales algunos puntos pueden quedar sin clasificar.

Medidas de Similitud de Documentos

- Componente fundamental de cualquier algoritmo de clustering.
- Si $d_1, d_2 \in \mathcal{D}$ son (representaciones de) documentos
- Una función de similitud φ , es un mapping

$$\varphi : \mathcal{D} \times \mathcal{D} \mapsto [0, 1]$$

tal que:

- 1 valores de $\varphi(d_1, d_2)$ cercanos a 1, indican que los documentos d_1 y d_2 son similares.
- 2 valores de $\varphi(d_1, d_2)$ cercanos a 0, indican poca similitud entre d_1 y d_2 .

Medidas de Similitud basadas en Conjuntos

Idea

Dos documentos $d_i, d_j \in \mathcal{D}$ son representados por los conjuntos D_i, D_j de sus términos. Las similitudes se basan en distintas ponderaciones de la intersección de conjuntos.

Ejemplos:

- Coeficiente de **Jaccard**: $\varphi_{jacc}(D_i, D_j) = \frac{|D_i \cap D_j|}{|D_i \cup D_j|}$.
- Coeficiente de **“dice”**: $\varphi_{dice}(D_i, D_j) = \frac{|D_i \cap D_j|}{|D_i| + |D_j|}$.
- Coeficiente de **solapamiento**: $\varphi_{over}(D_i, D_j) = \frac{|D_i \cap D_j|}{\max(|D_i|, |D_j|)}$.

Medidas de Similitud Geométricas

Idea

Dos documentos $d_i, d_j \in \mathcal{D}$ son comparados usando sus representaciones vector \vec{d}_i, \vec{d}_j , y su similitud se estima en base a la amplitud del ángulo formado por ambos vectores.

Ejemplo:

La función de similitud coseno $\varphi_{\cos} : \mathbb{R}^m \times \mathbb{R}^m \mapsto [0, 1]$

$$\varphi_{\cos}(\vec{d}_i, \vec{d}_j) = \frac{\langle \vec{d}_i, \vec{d}_j \rangle}{||\vec{d}_i|| \cdot ||\vec{d}_j||}$$

Validación de los grupos (o agrupamientos)

Evaluación (o validación) de grupos

Parte **fundamental** aunque poco explorada del análisis de grupos (cluster analysis).

Incluye

- Determinar la **tendencia de clustering**.
- Determinar el **número correcto de clusters**.
- Evaluar **cuan bien** los resultados del análisis de clusters (AC) se ajustan a los datos **sin** referencia a información externa.
- Comparar los resultados del AC con resultados conocidos **externamente**.
- Comparar dos conjuntos de clusters para determinar cual es mejor.

Medidas de Validación de agrupamientos (MVA)

Las MVA's se dividen en 3 grandes grupos

Internas (o no supervisadas)

Miden las “bondades” de la estructura de un agrupamiento sin recurrir a ningún tipo de información externa. Estas medidas (o índices) suelen ser referenciados como **internas** dado que sólo usan información presente en el conjunto de datos.

Externas (o supervisadas)

Miden el grado de concordancia entre la estructura de los grupos descubiertos y alguna estructura externa al conjunto de datos (de ahí su nombre).

Relativas

Compara agrupamientos o grupos particulares usando alguna de las dos medidas previas.

Medidas de Validez Internas (MVI)

Las diferentes MVIs intentan identificar propiedades estructurales específicas de los agrupamientos como **cohesión**, **separación**, **densidad** o alguna combinación de estas propiedades.

- La familia de **índices de Dunn**
- el **índice de Davies-Bouldin**
- el **coeficiente de Silueta** (Silhouette Coefficient)
- la **Medida-Λ**
- la **Medida de Densidad Esperada** $\bar{\rho}$

MVIs, cohesión y separación

Las MVIs, suelen expresar la validez de un cluster global de K clusters como:

$$validez_{total} = \sum_{i=1}^K w_i validez(C_i)$$

y la función de *validez* suele ser alguna forma de **cohesión**, **separación** o una **combinación** de éstas.

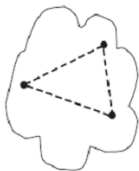
Cohesión

Mide cuan estrechamente relacionados están los objetos en un cluster.

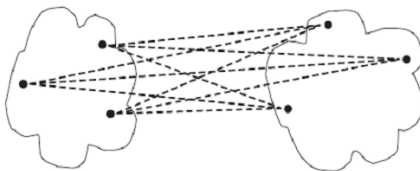
Separación

Mide cuán distintos (bien-separados) está un cluster de otro.

Cohesión y separación basada en grafos



Cohesión



Separación

$$cohesion(C_i) = \sum_{x \in C_i, y \in C_i} proximidad(x, y)$$

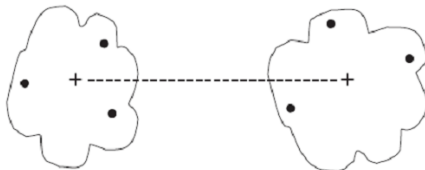
$$separacion(C_i, C_j) = \sum_{x \in C_i, y \in C_j} proximidad(x, y)$$

la función de *proximidad* puede ser *similitud*, *dis-similitud* (o distancia) o una función simple de estas cantidades.

Cohesión y separación basada en prototipos



Cohesión



Separación

$$cohesion(C_i) = \sum_{x \in C_i} proximidad(x, c_i)$$

$$separacion(C_i, C_j) = proximidad(c_i, c_j)$$

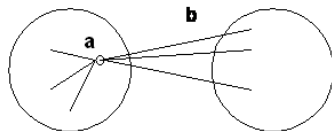
$$separacion(C_i) = proximidad(c_i, c)$$

Una MVI informativa: el Coeficiente de Silueta

Componente fundamental de esta medida: fórmula para determinar el coeficiente de silueta de un objeto arbitrario i :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

with $-1 \leq s(i) \leq 1$.



- $a(i)$ es la distancia promedio de i a los restantes objetos de su cluster.
- $b(i)$ es la distancia promedio de i a todos los objetos del cluster más cercano.
- Se busca que $s(i)$ sea tan **cercano a 1** como sea posible

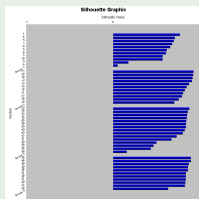
Una MVI informativa: el Coeficiente de Silueta

Combina ideas de **cohesión** y **separación**, pero para puntos individuales, grupos y agrupamientos

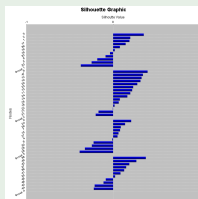
Puedo calcular la silueta de:

- un **grupo**: es el promedio de los coeficientes de silueta de sus objetos.
- un **agrupamiento**: es el promedio de los coeficientes de silueta de sus grupos.

Agrupamiento bueno



Agrupamiento malo



Medidas de Validez Externas (MVE)

Las MVEs evalúan un agrupamiento usando las medidas clásicas para evaluar un modelo de clasificación (categorización supervisada)

- Entropía
- Pureza
- Precisión y Recall
- Medida F (F -measure)