Máster en Ciencia de Datos y e I.C.
Sistemas de Recuperación de Información y de Recomendación

Juan Manuel Fernández Luna - jmfluna@decsai.ugr.es
Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada

v

Modelos avanzados de RI

Sumario

- 1ª parte: realimentación de relevancia.
- 2ª parte: agrupamiento documental.
- 3ª parte: clasificación documental.

1ª Parte

Realimentación de relevancia

Sumario

- 1ª parte: realimentación de relevancia.
- 2ª parte: agrupamiento documental.
- 3ª parte: clasificación documental.

1ª Parte

Realimentación de relevancia

Motivación

Cuando buscamos por “coche” no recuperamos documentos que tratan de “autos” o “automóviles” ¿Estamos perdiendo documentos relevantes?

Necesidad de mejorar los resultados

¿Cómo?

Técnicas de modificación de la consulta

- Métodos globales: expansión de consulta.
- Métodos locales: realimentación por relevancia y pseudo realimentación.

Motivación

Cuando buscamos por “coche” no recuperamos documentos que tratan de “autos” o “automóviles” ¿Estamos perdiendo documentos relevantes?

Necesidad de mejorar los resultados

¿Cómo?

Técnicas de modificación de la consulta

- Métodos globales: expansión de consulta.
- Métodos locales: realimentación por relevancia y pseudo realimentación.

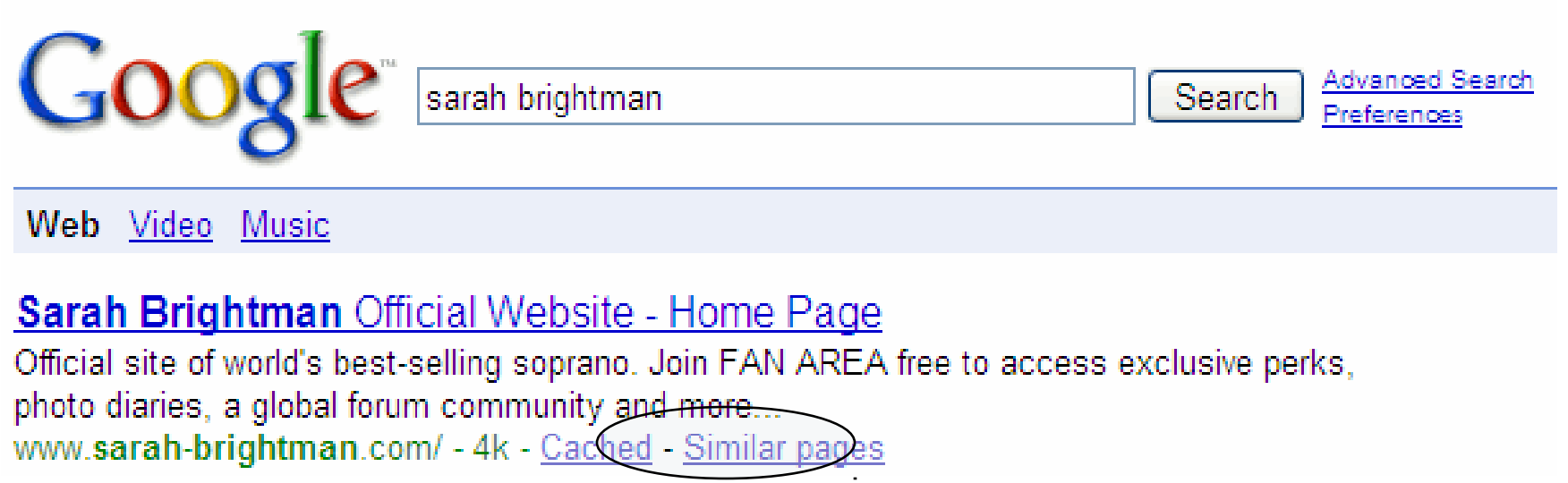
Sumario

- Realimentación por relevancia.
- El proceso.
- Algoritmo de Rochio
- Suposiciones de la realimentación.
- Problemas.
- Evaluación.
- Pseudo realimentación.
- Expansión de consultas.
- Expansión basada en tesauros.

Realimentación de relevancia

- Realimentación por parte del usuario sobre el conjunto de documentos inicialmente recuperados:
  - El usuario formula una consulta simple y corta.
  - El usuario marca algunos resultados como relevantes o no relevantes.
  - El sistema calcula una mejor representación de la necesidad de información basada en la realimentación.
  - Varias iteraciones.

Ejemplos


Google
sarah brightman
Search
Advanced Search
Preferences
Web
Video
Music
Sarah Brightman Official Website - Home Page
Official site of world's best-selling soprano. Join FAN AREA free to access exclusive perks, photo diaries, a global forum community and more...
www.sarah-brightman.com/ - 4k - Cached - Similar pages

Ejemplos

Browse | Search | Prev | Next | Random

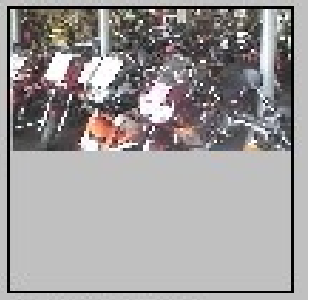
(144473, 16458)
0.0
0.0
0.0

(144457, 252140)
0.0
0.0
0.0

(144456, 262857)
0.0
0.0
0.0

(144456, 262863)
0.0
0.0
0.0

(144457, 252134)
0.0
0.0
0.0



(144483, 265154)
0.0
0.0
0.0

(144483, 264644)
0.0
0.0
0.0

(144483, 265153)
0.0
0.0
0.0

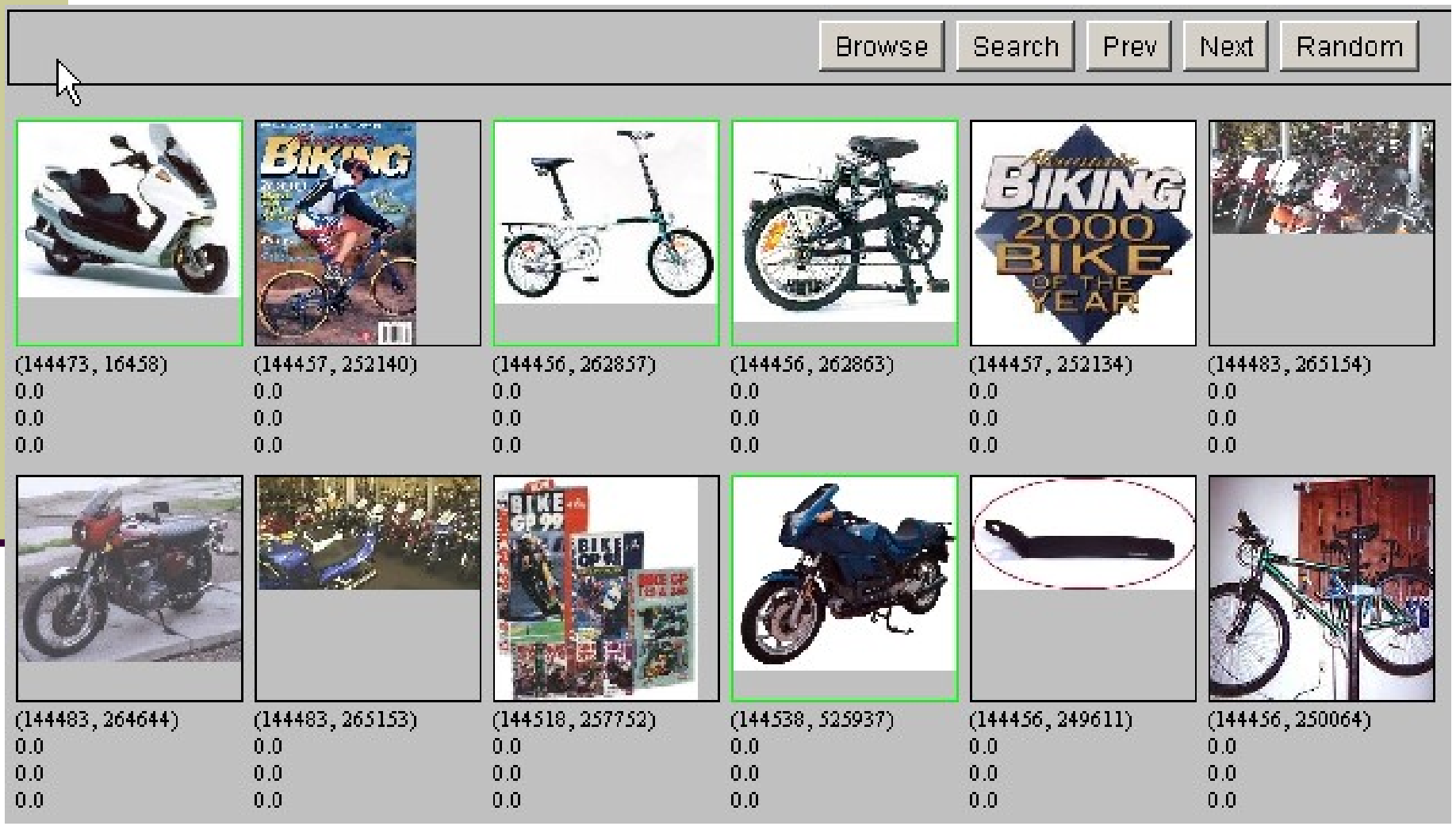
(144518, 257752)
0.0
0.0
0.0

(144538, 525937)
0.0
0.0
0.0

(144456, 249611)
0.0
0.0
0.0

(144456, 250064)
0.0
0.0
0.0

Ejemplos


Browse
Search
Prev
Next
Random
BIKING
BIKING 2000 BIKE OF THE YEAR
(144473, 16458)
0.0
0.0
0.0
(144457, 252140)
0.0
0.0
0.0
(144456, 262857)
0.0
0.0
0.0
(144456, 262863)
0.0
0.0
0.0
(144457, 252134)
0.0
0.0
0.0
(144483, 265154)
0.0
0.0
0.0
(144483, 264644)
0.0
0.0
0.0
(144483, 265153)
0.0
0.0
0.0
(144518, 257752)
0.0
0.0
0.0
(144538, 525937)
0.0
0.0
0.0
(144456, 249611)
0.0
0.0
0.0
(144456, 250064)
0.0
0.0
0.0

Ejemplos

Browse | Search | Prev | Next | Random

(144538, 523493)
0.54182
0.231944
0.309876

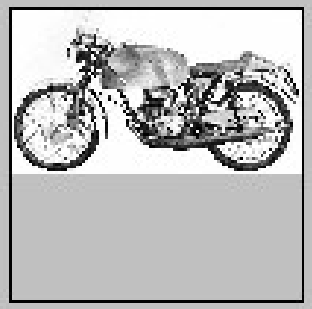
(144538, 523835)
0.56319296
0.267304
0.295889

(144538, 523529)
0.584279
0.280881
0.303398

(144456, 253569)
0.64501
0.351395
0.293615

(144456, 253568)
0.650275
0.411745
0.23853

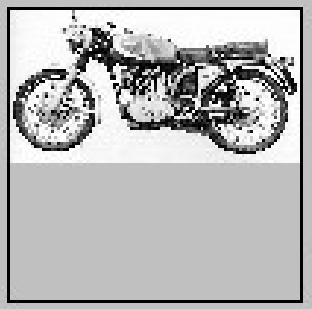
(144538, 523799)
0.66709197
0.358033
0.309059



(144473, 16249)
0.6721
0.393922
0.278178

(144456, 249634)
0.675018
0.4639
0.211118

(144456, 253693)
0.676901
0.47645
0.200451



(144473, 16328)
0.700339
0.309002
0.391337

(144483, 265264)
0.70170796
0.36176
0.339948

(144478, 512410)
0.70297
0.469111
0.233859

Sumario

- Realimentación por relevancia.
- El proceso.
- Algoritmo de Rochio
- Suposiciones de la realimentación.
- Problemas.
- Evaluación.
- Pseudo realimentación.
- Expansión de consultas.
- Expansión basada en tesauros.

El procedimiento de la realimentación

- Consulta inicial:

  New space satellite applications

  \+ 1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
  \+ 2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
  3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
  4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
  5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
  6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
  7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada
  \+ 8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

- El usuario marca los relevantes con '+'.

El procedimiento de la realimentación

Consulta expandida:

- 2.074 new
- 30.816 satellite
- 5.991 nasa
- 4.196 launch
- 3.516 instrument
- 3.004 bundespost
- 2.790 rocket
- 2.003 broadcast
- 0.836 oil
- 15.106 space
- 5.660 application
- 5.196 eos
- 3.972 aster
- 3.446 arianespace
- 2.806 ss
- 2.053 scientist
- 1.172 earth
- 0.646 measure

El procedimiento de la realimentación

2 1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan

1 2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer

3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own

4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit

5. 0.492, 12/02/87, Telecommunications Tale of Two Companies

6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use

7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers

8 8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

Sumario

- Realimentación por relevancia.
- El proceso.
- Algoritmo de Rochio
- Suposiciones de la realimentación.
- Problemas.
- Evaluación.
- Pseudo realimentación.
- Expansión de consultas.
- Expansión basada en tesauros.

Algoritmo de Rocchio

- El algoritmo de Rocchio se basa en el modelo del espacio vectorial para general una consulta de realimentación.
- Rocchio busca la consulta q_{opt} que maximiza:

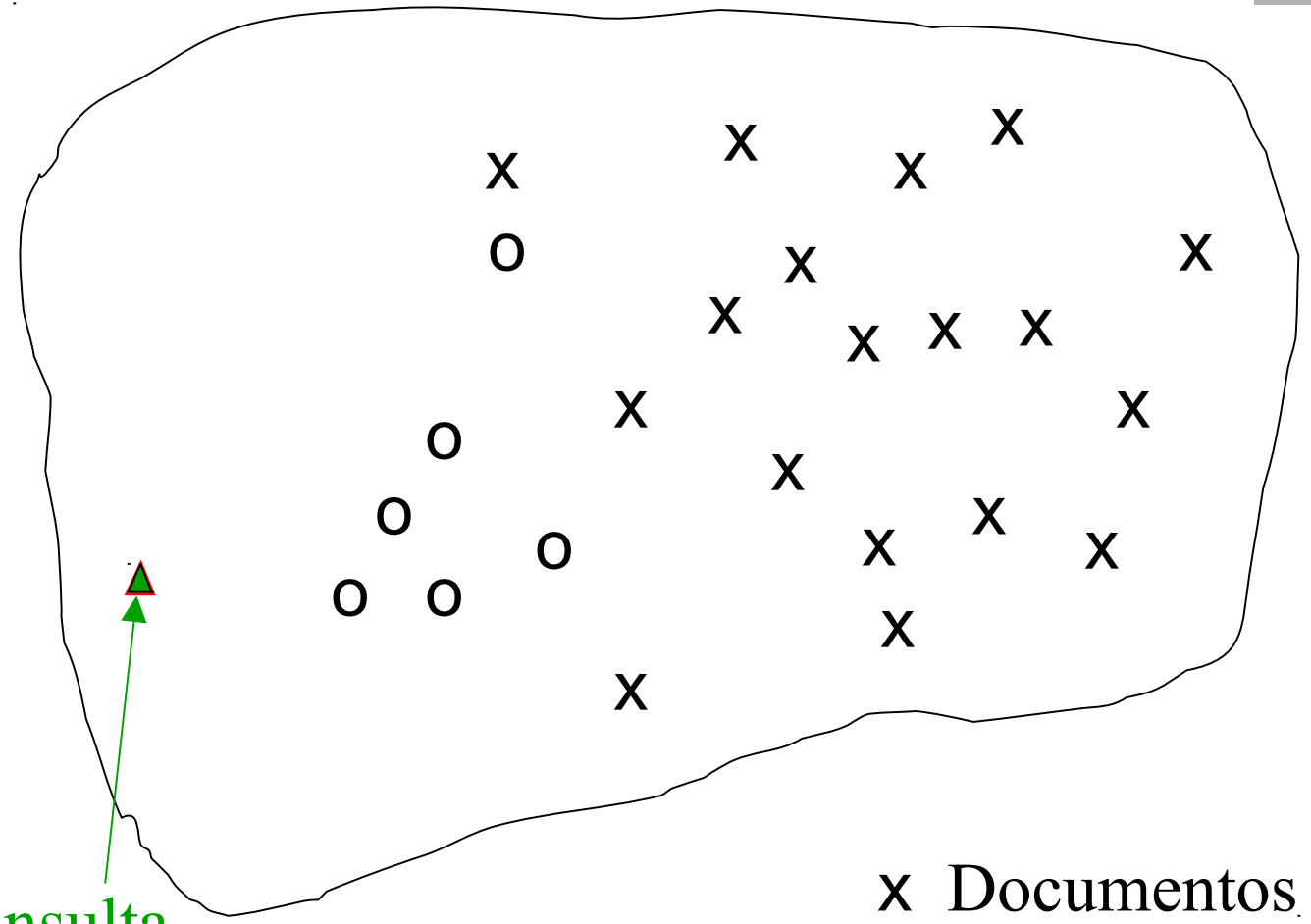
$$q_{opt} = \arg\max_{\vec{q}} [\cos(\vec{q}, \mu(C_r)) - \cos(\vec{q}, \mu(C_{nr}))]$$

- Intenta separar los documentos relevantes e irrelevantes.

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{d_j \in C_r} \vec{d}_j - \frac{1}{|C_{nr}|} \sum_{d_j \notin C_r} \vec{d}_j$$

- Problema: No conocemos la relevancia real de los documentos.

Algoritmo de Rocchio



Consulta óptima

x Documentos no relevantes

o Documentos relevantes

Algoritmo de Rocchio

- En la práctica:

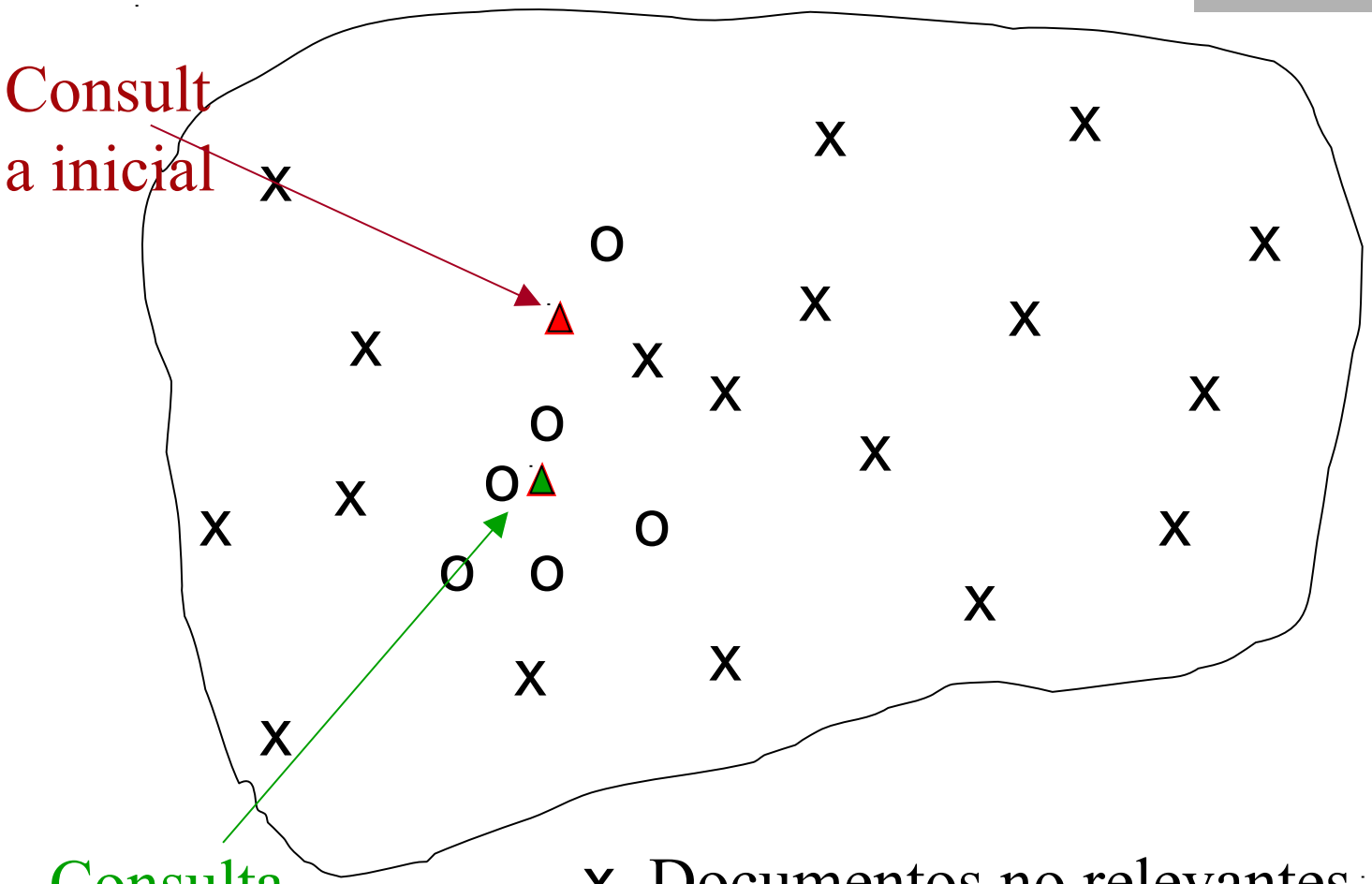
$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{d_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} \vec{d}_j$$

- *D_r = Vectores del conjunto de documentos relevantes conocidos.*
- *D_{nr}* = Vectores del conjunto de documentos no relevantes conocidos.
  - Diferentes de los conjuntos C_r y C_{nr}
- *q_m* = consulta modificada; *q_0* = vector de la consulta original; *α,β,γ*: pesos (elegidos a mano o empíricamente)
- La consulta nueva se mueve hacia los documentos relevantes y se aleja de los irrelevantes.

Algoritmo de Rocchio

- Equilibrio entre α vs. β/γ : Si tenemos muchos documentos juzgados, entonces β/γ más altos.
- Algunos pesos del vector pueden ser negativos.
  - Se ignoran asignándoles un cero.

Algoritmo de Rocchio



Algoritmo de Rochio

- Se puede modificar la consulta basándonos en relimentación por relevancia y aplicar el modelo vectorial.
- Se emplean sólo los documentos marcados.
- La realimentación puede mejorar el recall y la precisión.
- La realimentación es más útil para incrementar el recall en situaciones en qeu sea importante.
  - Se espera que los usuarios revisen los resultados y se tomen su tiempo para iterar.

Sumario

- Realimentación por relevancia.
- El proceso.
- Algoritmo de Rochio
- Suposiciones de la realimentación.
- Problemas.
- Evaluación.
- Expansión de consultas.
- Expansión basada en tesauros.

Suposiciones de la realimentación

- A1: El usuario tiene suficiente conocimiento para formular la consulta inicial.
- A2: Se considera correcto que
  - La distribución de términos en documentos relevantes será similar.
  - La distribución de términos en documentos no relevantes será diferente de aquellos no relevantes.

Sumario

- Realimentación por relevancia.
- El proceso.
- Algoritmo de Rochio
- Suposiciones de la realimentación.
- Problemas.
- Evaluación.
- Pseudo realimentación.
- Expansión de consultas.
- Expansión basada en tesauros.

Problemas de la realimentación

- Las consultas largas son difíciles para los motores.
  - Tiempos de respuesta largos para el usuario.
  - Alto costo computacional para el SRI.
  - Solución parcial:
    - Sólo reponderar aquellos términos principales.
      - Quizá los 20 primeros por frecuencia del término.
- Los usuarios a menudo son reticentes a ofrecer realimentación.
- A veces es difícil entender el por qué fue recuperado un documento después de aplicar realimentación.

Sumario

- Realimentación por relevancia.
- El proceso.
- Algoritmo de Rochio
- Suposiciones de la realimentación.
- Problemas.
- Evaluación.
- Pseudo realimentación.
- Expansión de consultas.
- Expansión basada en tesauros.

Evaluación de la realimentación

- Se usa q_0 y se calcula la gráfica de precisión y recall.
- Se emplea q_m para calcular la misma gráfica.
  - Evaluación con todos los documentos de la colección:
    - Mejoras espectaculares ... ¡¡Se hace trampa!!
    - Se debe parcialmente a que los documentos conocidos se situarán arriba de la ordenación.
    - Se debe evaluar con respecto a los documentos no vistos del usuario.
  - Usar los documentos de la colección residual (el conjunto de documentos menos aquellos juzgados como relevantes).
    - Medidas menores con respecto a las obtenidas con la consulta original.
    - Evaluación más realista.
    - El rendimiento relativo puede ser comparado.
- Empíricamente, una vuelta de realimentación es a menudo útil. Dos, marginal mente útil.

Sumario

- Realimentación por relevancia.
- El proceso.
- Algoritmo de Rochio
- Suposiciones de la realimentación.
- Problemas.
- Evaluación.
- Pseudo realimentación.
- Expansión de consultas.
- Expansión basada en tesauros.

Pseudo realimentación de relevancia

- Automatiza la parte manual de la realimentación.
- Algoritmo:
  - Recuperar una lista ordenada de documentos a partir de una consulta inicial.
  - Se ausme que los k más altos son relevantes.
  - Realizar la realimentación (Rocchio).
- En general, funciona muy bien, pero...
- Puede ir muy mal para algunas consultas.
- Varias iteraciones produce “query drift”.
- ¿Por qué?

Sumario

- Realimentación por relevancia.
- El proceso.
- Algoritmo de Rochio
- Suposiciones de la realimentación.
- Problemas.
- Evaluación.
- Pseudo realimentación.
- Expansión de consultas.
- Expansión basada en tesauros.

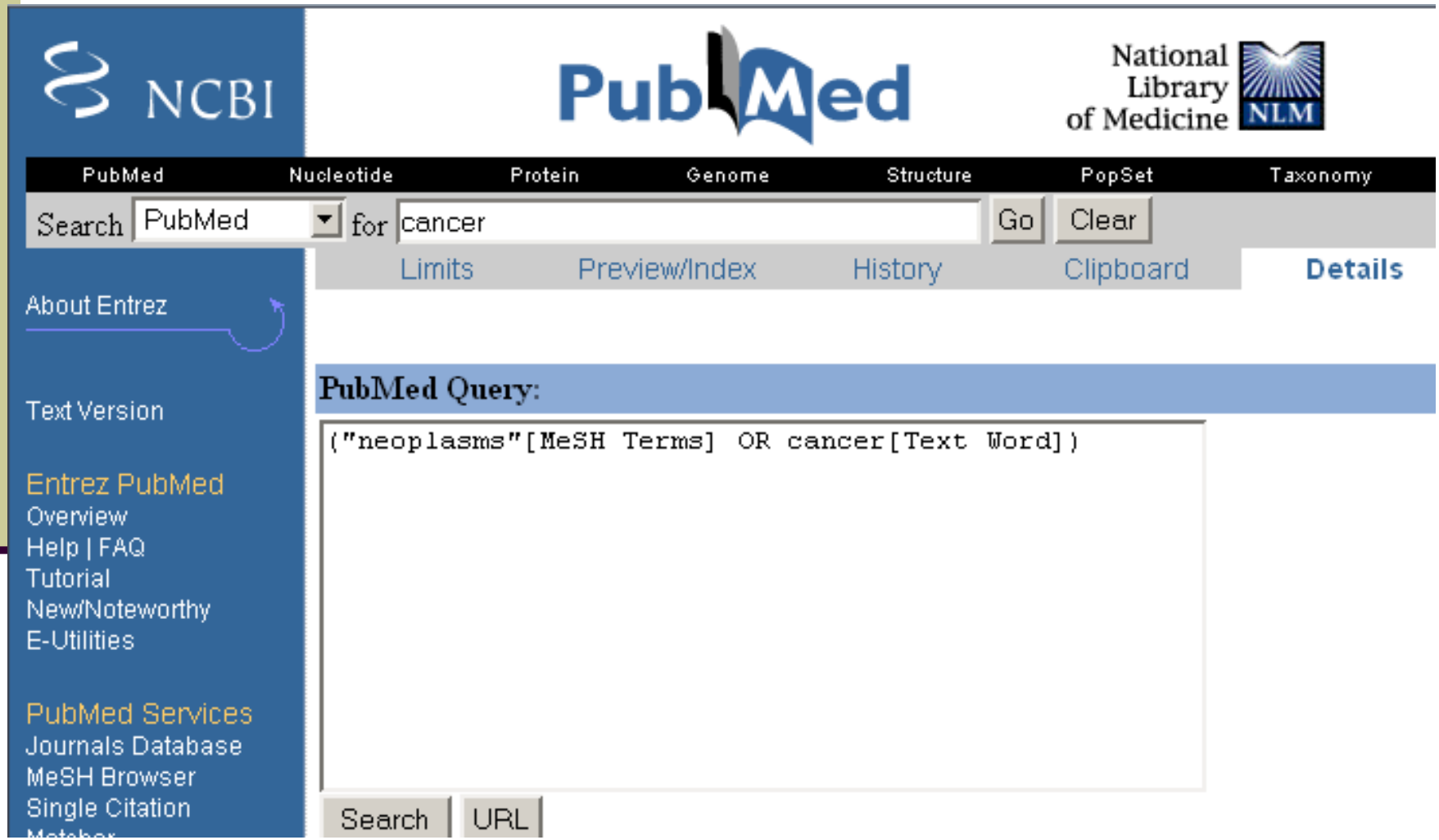
Expansión de consultas

- Realimentación:
  - Reponderación de términos de la consulta original.
  - Inclusión de nuevos términos en la consulta.

Expansión de consultas

- Uso de un tesauro.
  - Por ejemplo, MedLine: physician, syn: doc, doctor, MD, médico.
- Análisis global (estático: a partir de los documentos de la colección).
  - Tesauros construidos automáticamente (co-ocurrencias).
  - Refinamientos basados en minería de logs (común en la web).
- Análisis local (dinámico: a partir de los documentos del conjunto de resultados).

Expansión de consultas



Sumario

- Realimentación por relevancia.
- El proceso.
- Algoritmo de Rochio
- Suposiciones de la realimentación.
- Problemas.
- Evaluación.
- Pseudo realimentación.
- Expansión de consultas.
- Expansión basada en tesauros.

Expansión basada en tesauros

- Para cada término t de la consulta, expandir la consulta con sinónimos y palabras relacionadas con t en el tesauro.
  - Gato → felino.
- Se suelen ponderar más bajo a los términos añadidos que a los originales.
- Incrementa el recall.
- Ampliamente usados en entornos científicos y de ingeniería.
- Puede decrementar la precisión, sobre todo con términos ambíguos:
  - “interest rate” → “interest rate fascinate evaluate”
- La construcción manual del tesauro es costosa y su actualización.

Construcción de tesauros

Intento de generar un tesauro analizando de forma automática los documentos de la colección.

- Concepto fundamental: similitud entre dos palabras.
- Definición 1: Dos palabras son similares si co-ocurren con palabras similares.
- Definición 2: Dos palabras son similares si ocurren en una relación gramatical concreta con las mismas palabras.
  - Por ejemplo, las manzanas y peras se puede recolectar, pelar, comer y prepara, por tnato deben ser similares.
- Las relaciones basadas en co-ocurrencia son más robustas, y las gramaticales más precisas.

Construcción de tesauros

- Ejemplo de tesauro:

word	ten nearest neighbors
absolutely	absurd whatsoever totally exactly nothing
bottomed	dip copper drops topped slide trimmed slig
captivating	shimmer stunningly superbly plucky witty
doghouse	dog porch crawling beside downstairs gazec
Makeup	repellent lotion glossy sunscreen Skin gel p
mediating	reconciliation negotiate cease conciliation p
keeping	hoping bring wiping could some would othe
lithographs	drawings Picasso Dali sculptures Gauguin
pathogens	toxins bacteria organisms bacterial parasite
senses	grasp psyche truly clumsy naive innate awl

Construcción de tesauros

Problemas de la construcción automática de tesauros:

- La calidad de las relaciones es un problema normalmente.
- La ambigüedad del término puede introducir términos que son erróneos o irrelevantes:
  - “Apple computer” → “Apple red fruit computer”
- Problemas:
  - Falsos positivos: Palabras que parence ser similares pero no lo son.
  - Falsos negativoss: Palbras que parecen ser diferentes pero son similares.
- Como los términos suelen estar muy correlados, la expansión no incluye muchos documentos nuevos.

2ª y 3ª Partes

Agrupamiento y clasificación documental

Motivación

- Desde el punto de vista del aprendizaje existen dos grandes campos:
  - Aprendizaje supervisado:
    - descubrimiento de patrones en los datos que relacionan atributos con un atributo objetivo (clase a la que pertenecen). Estos patrones se utilizan para predecir los valores de dicho atributo para datos futuros.
  - Aprendizaje no supervisado:
    - Los datos no tienen ese atributo objetivo. Exploración de los datos para encontrar la estructura intríseca.

Motivación

- Aprendizaje supervisado:
  - descubrimiento de patrones en los datos que relacionan atributos con un atributo objetivo (clase a la que pertenecen). Estos patrones se utilizan para predecir los valores de dicho atributo para datos futuros.

Motivación

- Aprendizaje no supervisado:
  - Los datos no tienen ese atributo objetivo. Exploración de los datos para encontrar la estructura intrínseca.

Motivación

- En recuperación de información:
- Clasificación o categorización documental:
  - Cada documento pertenece a una clase conocida.
  - Nuevos documentos se tienen que situar en uno de los grupos existentes de manera correcta.
- Agrupamiento documental:
  - Los documentos no están etiquetados y los grupos deben ser descubiertos.

2ª Parte

Agrupamiento documental

Sumario

- ¿Qué es el agrupamiento documental?
- Aplicaciones del agrupamiento.
- La base: “The cluster hypothesis”.
- Componentes del proceso.
  - Representación de documentos.
  - Medida de asociación.
  - Representación de los grupos.
  - Algoritmos de agrupamiento.
  - Evaluación del agrupamiento.

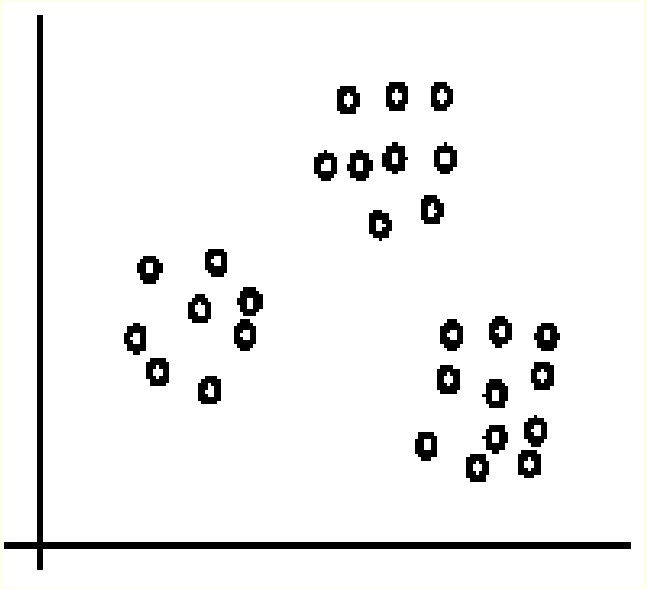
¿Qué es el agrupamiento?

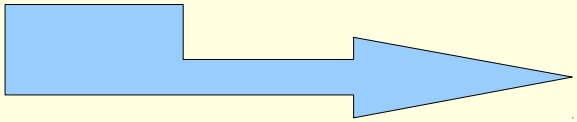
- El agrupamiento (cluster analysis, clustering) es una técnica estadística cuyo objetivo es el de agrupar objetos similares.
- Objetos similares entre sí, se disponen en un mismo grupo (cluster), y los que son muy diferentes, en otros grupos, lejos de los primeros.
- Esta técnica es un ejemplo de aprendizaje no supervisado, ya que no se da información a priori sobre la pertenencia de los objetos a clases.

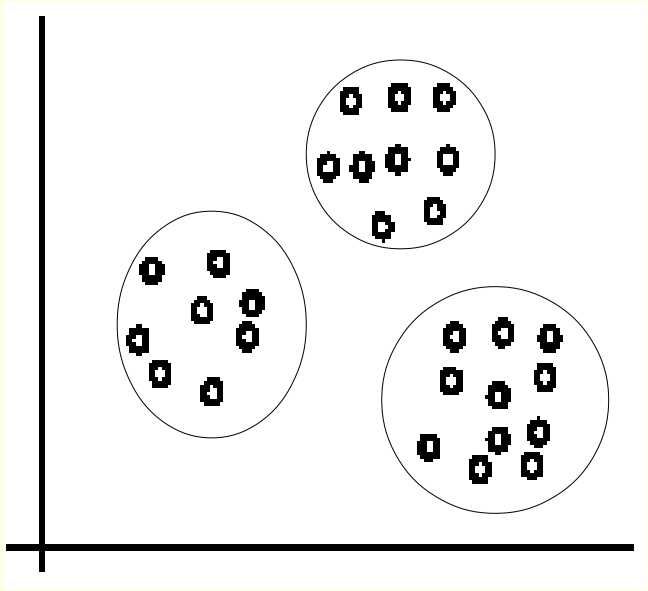
¿Qué es el agrupamiento?

- En Recuperación de Información, esos objetos son documentos (también pueden ser términos).
- La agrupación se hace teniendo en cuenta los términos que contiene dichos documentos.
- Documentos en el mismo grupo son similares, mientras que aquellos que son diferentes, se disponen en grupos distintos.

¿Qué es el agrupamiento?







Sumario

- ¿Qué es el agrupamiento documental?
- Aplicaciones del agrupamiento.
- La base: “The cluster hypothesis”.
- Componentes del proceso.
  - Representación de documentos.
  - Medida de asociación.
  - Representación de los grupos.
  - Algoritmos de agrupamiento.
  - Evaluación del agrupamiento.

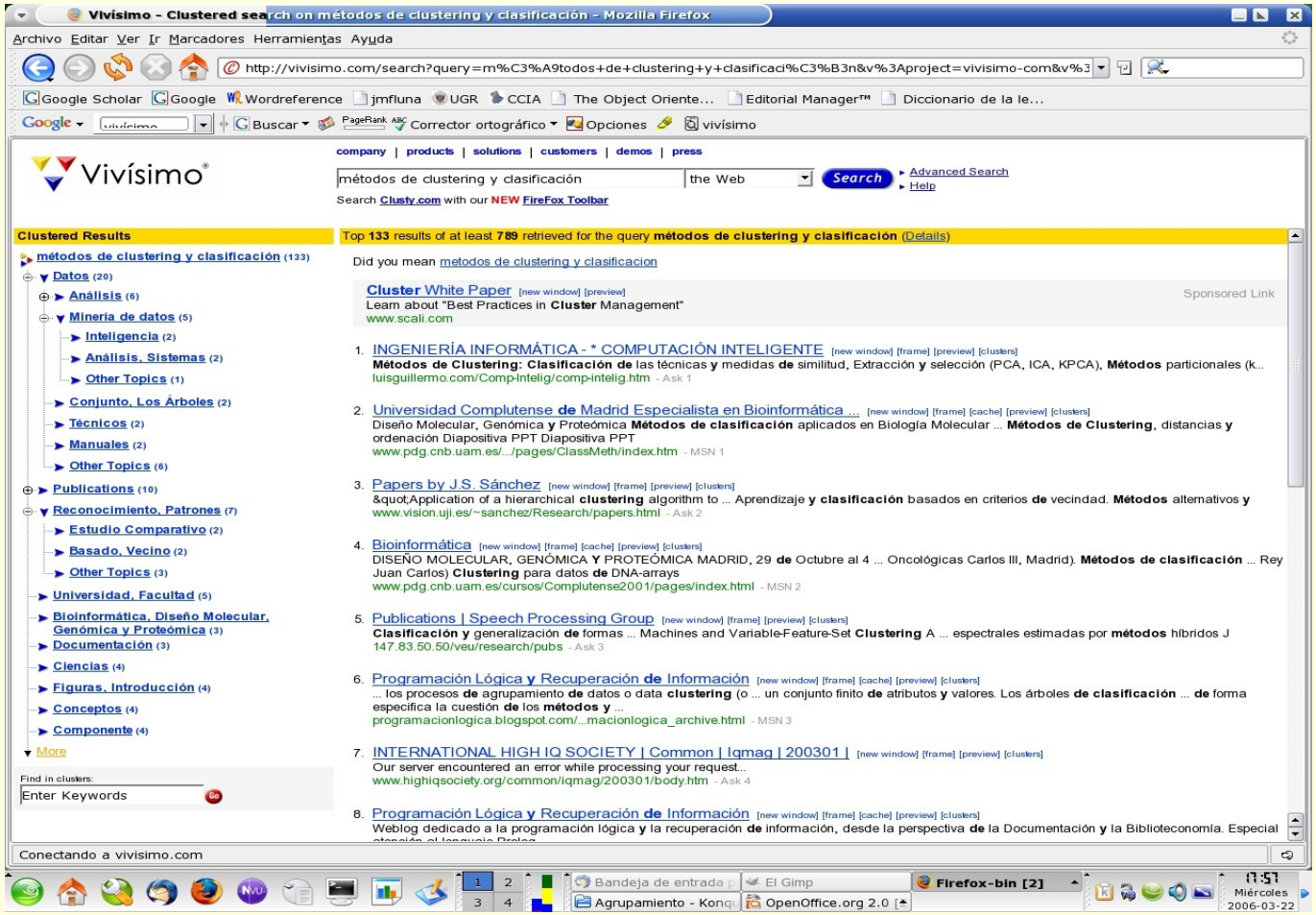
Aplicaciones del agrupamiento

- Navegación y visualización de conjuntos de documentos recuperados.
- Análisis de colecciones completas.
- Mejora de la efectividad de recuperación de los motores de búsqueda (modelo de recuperación basado en agrupamiento).

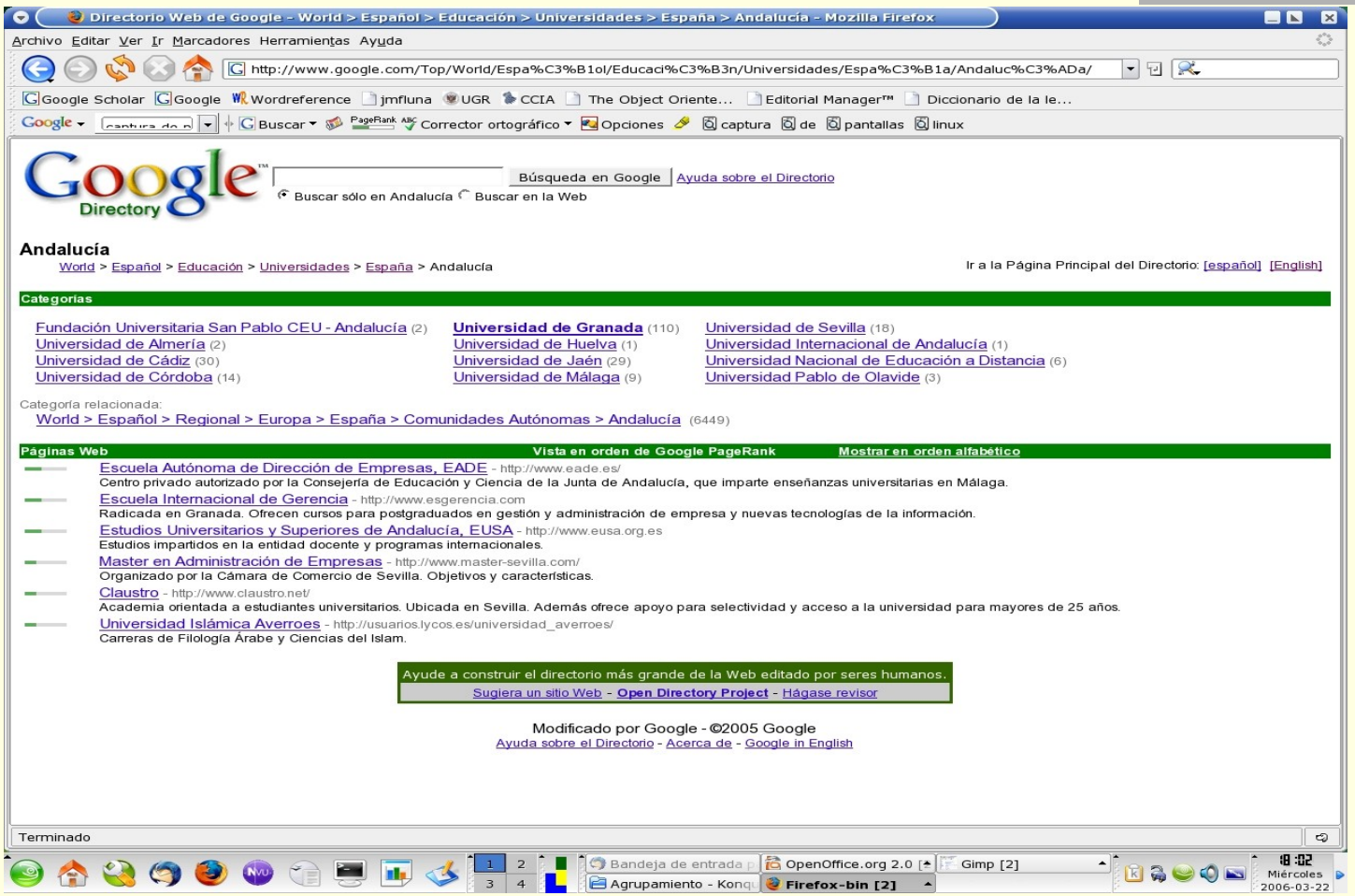
Aplicaciones del agrupamiento

- En este sentido...
  - Si una colección esté bien dividida en grupos, sólo deberemos buscar en aquel que contenga los documentos relevantes.
  - La búsqueda en colecciones más pequeñas debería incrementar la eficiencia y la eficacia.

Aplicaciones del agrupamiento



Aplicaciones del agrupamiento



Sumario

- ¿Qué es el agrupamiento documental?
- Aplicaciones del agrupamiento.
- La base: “The cluster hypothesis”.
- Componentes del proceso.
  - Representación de documentos.
  - Medida de asociación.
  - Representación de los grupos.
  - Algoritmos de agrupamiento.
  - Evaluación del agrupamiento.

La base del agrupamiento

- “The cluster hypothesis” (van Rijsbergen, 1979):
  - Documentos similares tienden a ser relevantes a la misma consulta.

    ó
  - Documentos relevantes tienden a ser más parecidos entre ellos que con los no relevantes.
  - Fácilmente comprobable empíricamente.

Sumario

- ¿Qué es el agrupamiento documental?
- Aplicaciones del agrupamiento.
- La base: “The cluster hypothesis”.
- Componentes del proceso.
  - Representación de documentos.
  - Medida de asociación.
  - Representación de los grupos.
  - Algoritmos de agrupamiento.
  - Evaluación del agrupamiento.

Representación de los documentos

- Típicamente, documentos se representan como vectores de términos ponderados.
- Se suele realizar el "preprocesado" de documentos clásico en R.I.:
  - Eliminación de palabras vacías.
  - Segmentación.
  - Selección de los mejores términos.
- Los esquemas de ponderación utilizados se suelen basar en el tf·idf.

Sumario

- ¿Qué es el agrupamiento documental?
- Aplicaciones del agrupamiento.
- La base: “The cluster hypothesis”.
- Componentes del proceso.
  - Representación de documentos.
  - Medida de asociación.
  - Representación de los grupos.
  - Algoritmos de agrupamiento.
  - Evaluación del agrupamiento.

Medidas de asociación

- Permiten establecer el grado de asociación entre cada par de documentos de la colección.
- Los tipos de medidas más ampliamente utilizados son:
  - Asociación o similitud.
  - Disimilaridad.

Medidas de asociación

- Similitud:
  - Cuantas más coincidencias haya entre dos documentos D_i, D_j, (fundamentalmente, términos de indexación), más alta será la similitud ($Sim(D_i,D_j)$).
  - Normalmente:
    - $0 \leq Sim(D_i,D_j) \leq 1$
    - $Sim(D_i,D_i)=1$
    - $Sim(D_i,D_j) = Sim(D_j,D_i)$

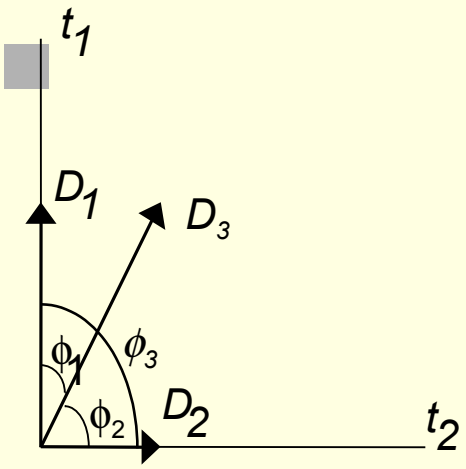
Medidas de asociación

- Coeficiente de Dice:
  - $Sim(D_i, D_j) = 2|D_i \cap D_j| / (|D_i| + |D_j|)$
- Coeficiente de Jaccard:
  - $Sim(D_i, D_j) = |D_i \cap D_j| / (D_i \cup D_j)$
- Distancia Euclídea:

$$Sim(D_i, D_j) = \sqrt{(D_{i1} - D_{j1})^2 + (D_{i2} - D_{j2})^2 + \ldots + (D_{ir} - D_{jr})^2}$$

Medidas de asociación

- Medida del coseno:
  - Calcula el coseno del ángulo que forman los vectores en el espacio de documentos.



$D_1 = (t_1, t_1) \qquad D_2 = (t_2) \qquad D_3 = (t_1, t_1, t_2)$

$Sim(D_1, D_3) = \cos(\phi_1)$

$Sim(D_1, D_2) = \cos(\phi_3) = 0$

$$\cos(D_i, D_j) = \frac{\sum_{k=1}^{m} d_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{m} d_{ik}^2 \cdot \sum_{k=1}^{m} d_{jk}^2}}$$

Sumario

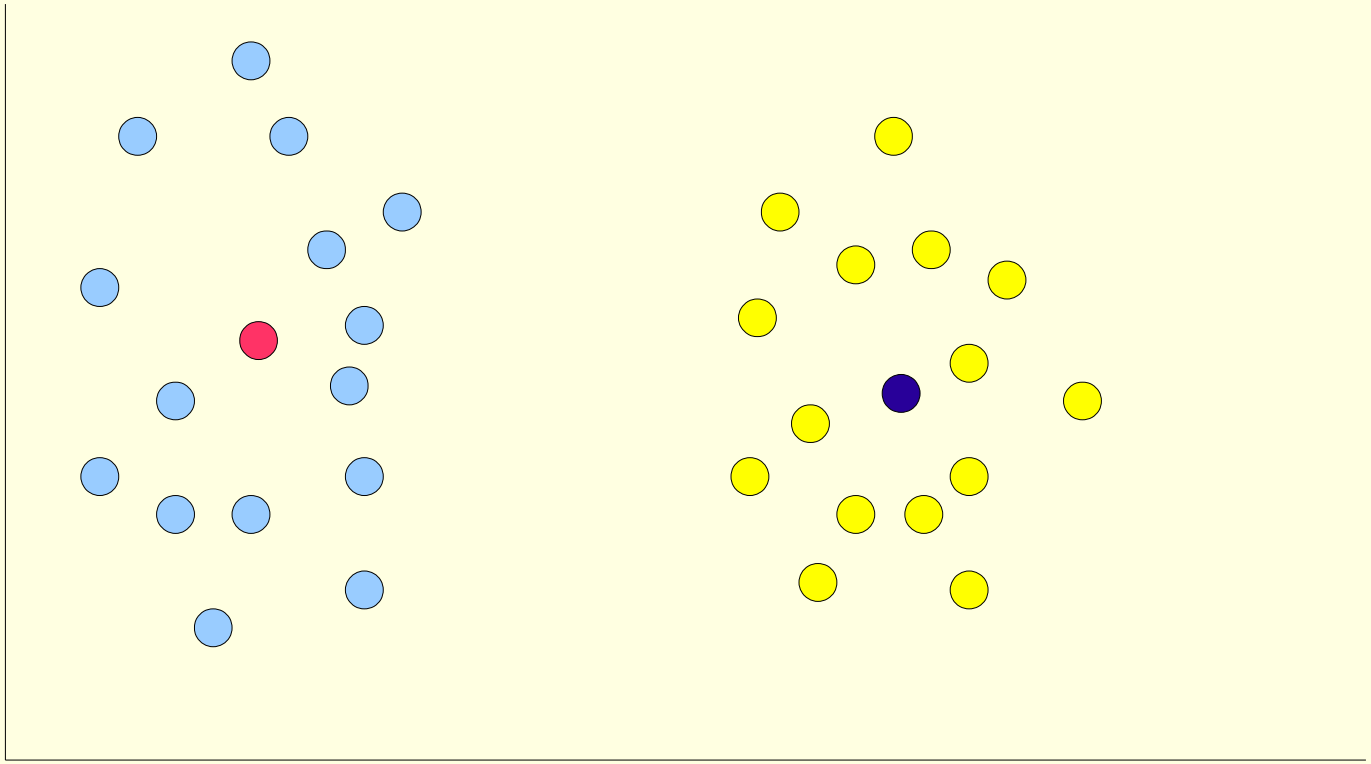
- ¿Qué es el agrupamiento documental?
- Aplicaciones del agrupamiento.
- La base: “The cluster hypothesis”.
- Componentes del proceso.
  - Representación de documentos.
  - Medida de asociación.
  - Representación de los grupos.
  - Algoritmos de agrupamiento.
  - Evaluación del agrupamiento.

Representación de los grupos

- Centroide del grupo (cluster centroid).
  - Normalmente se construye como un vector medio de todos los vectores de los documentos.
  - El peso de un término en el centroide es la media del peso del término en los documentos del grupo.
  - El centroide se puede considerar como un documento con la máxima similitud a todos los documentos del grupo.
  - Debe ser descriptivo del contenido del grupo.
  - Debe distinguir su grupo del resto.

Representación de los grupos

- Centroide del grupo (cluster centroid):



Sumario

- ¿Qué es el agrupamiento documental?
- Aplicaciones del agrupamiento.
- La base: “The cluster hypothesis”.
- Componentes del proceso.
  - Representación de documentos.
  - Medida de asociación.
  - Representación de los grupos.
  - Algoritmos de agrupamiento.
  - Evaluación del agrupamiento.

Métodos de agrupamiento

- Clasificación atendiendo a las relaciones entre grupos:
  - Algoritmos que generan grupos ordenados (los grupos se establecen entorno a una jerarquía).
    - Jerárquicos.
  - Algoritmos que generan grupos no ordenados (los grupos aparecen totalmente desconectados entre sí).
    - De particiones.

Métodos de agrupamiento

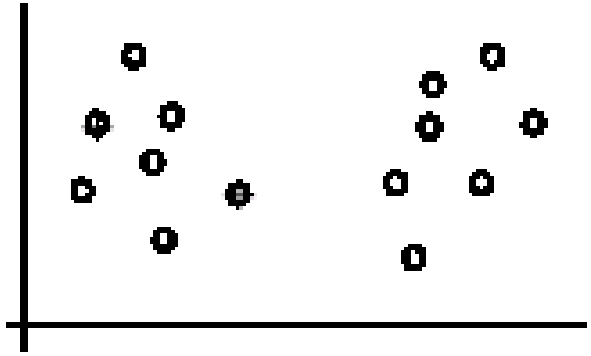
- Métodos de particionamiento:
  - Dado un conjunto de documentos, el objetivo es agruparlos en k conjuntos (crear una partición de k grupos),
    - Tal partición debe ser óptima bajo algún criterio.
    - Se usan funciones de la similitud de los documentos dentro del grupo, y la distancia entre grupos.

Métodos de agrupamiento

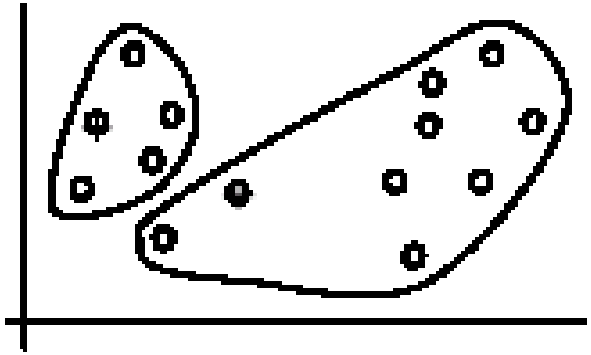
- Ejemplo: Algoritmo de las K medias.

- 1) Elegir aleatoriamente K documentos (semillas), que actuarán como centroides iniciales.
- 2) Asignar cada documento al grupo cuyo centroide esté más cercano.
- 3) Recalcular los centroides.
- 4) Si el criterio de finalización se cumple, entonces, parar. En caso contrario, ir a 2.
- Criterios de parada: particiones o centroides constantes.

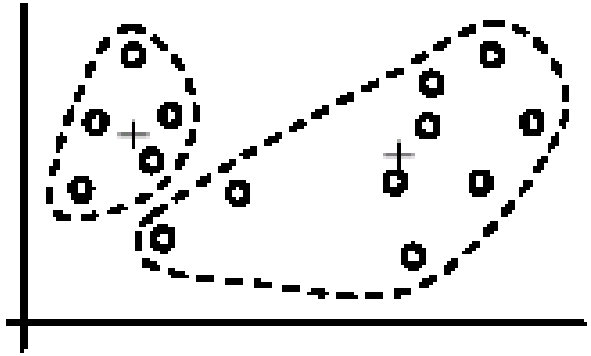
Métodos de agrupamiento



Elección aleatoria de los k centroides (k=2)

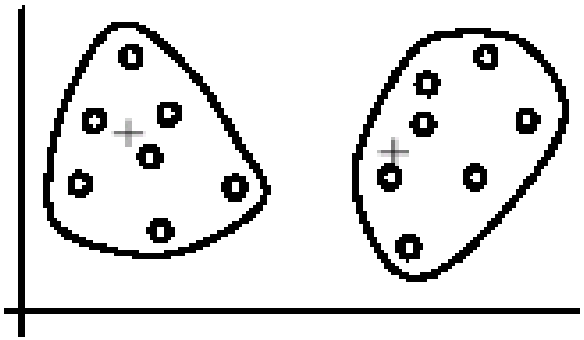


Iteración 1: asignación de documentos.

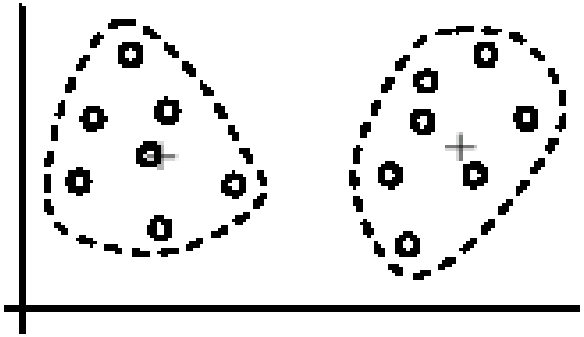


Relocalización de centroides

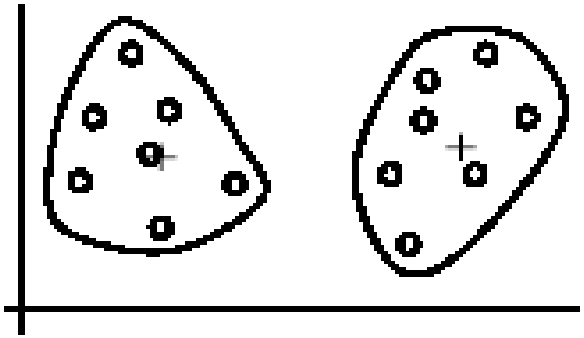
Métodos de agrupamiento



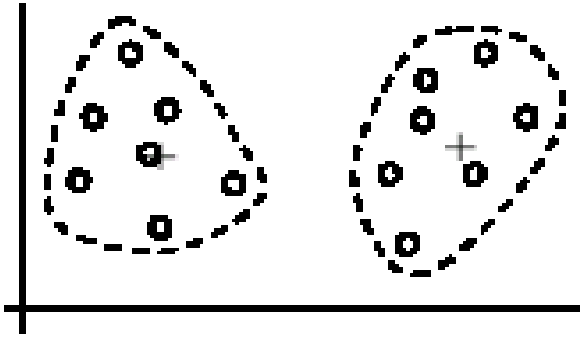
Iteración 2: asignación de documentos.



Relocalización de centroides



Iteración 3: asignación de documentos.



Relocalización de centroides

Métodos de agrupamiento

- Ejemplo: Algoritmo de pasada única.

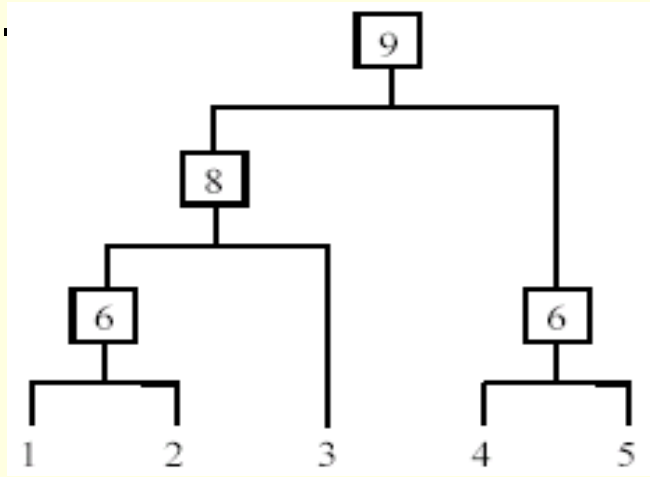
- 1) Crear un primer grupo con un documento elegido al azar.
- 2) Para cada documento, calcular la similitud con cada centroide.
  - 2.1) Si similitud > umbral, entonces añadir el documento al grupo correspondiente y recalcular el centroide.
  - Sino, crear un nuevo grupo con ese documento como centroide.

Métodos de agrupamiento

- Ventajas de los métodos de particionamiento:
  - Muy rápidos: orden de eficiencia linea con respecto al número de documentos.
  -
- Inconvenientes:
  - Muchos parámetros por definir (número de grupos, tamaño de éstos, criterio de pertenencia,...).
  - El orden de los documentos puede cambiar el agrupamiento.

Métodos de agrupamiento

- Métodos jerárquicos:
- El resultado es una estructura arbórea denominada dendograma.



- Valores numéricos: similitud en los niveles en que se forman los grupos.
- Nivel 1: un único grupo.

Métodos de agrupamiento

- Tipos de Métodos jerárquicos:
- Aglomerativos:
  - Construyen el árbol a partir de los documentos individuales, mezclando los pares de grupos más similares y finalizando cuando todos los grupos se han mezclado en uno único.
- Divisivos:
  - Parten del conjunto completo en un único grupo y van creando grupos a partir de él, dividiendo en varios grupos hijos, de manera recursiva, hasta que se forman grupos formados por documentos individuales.

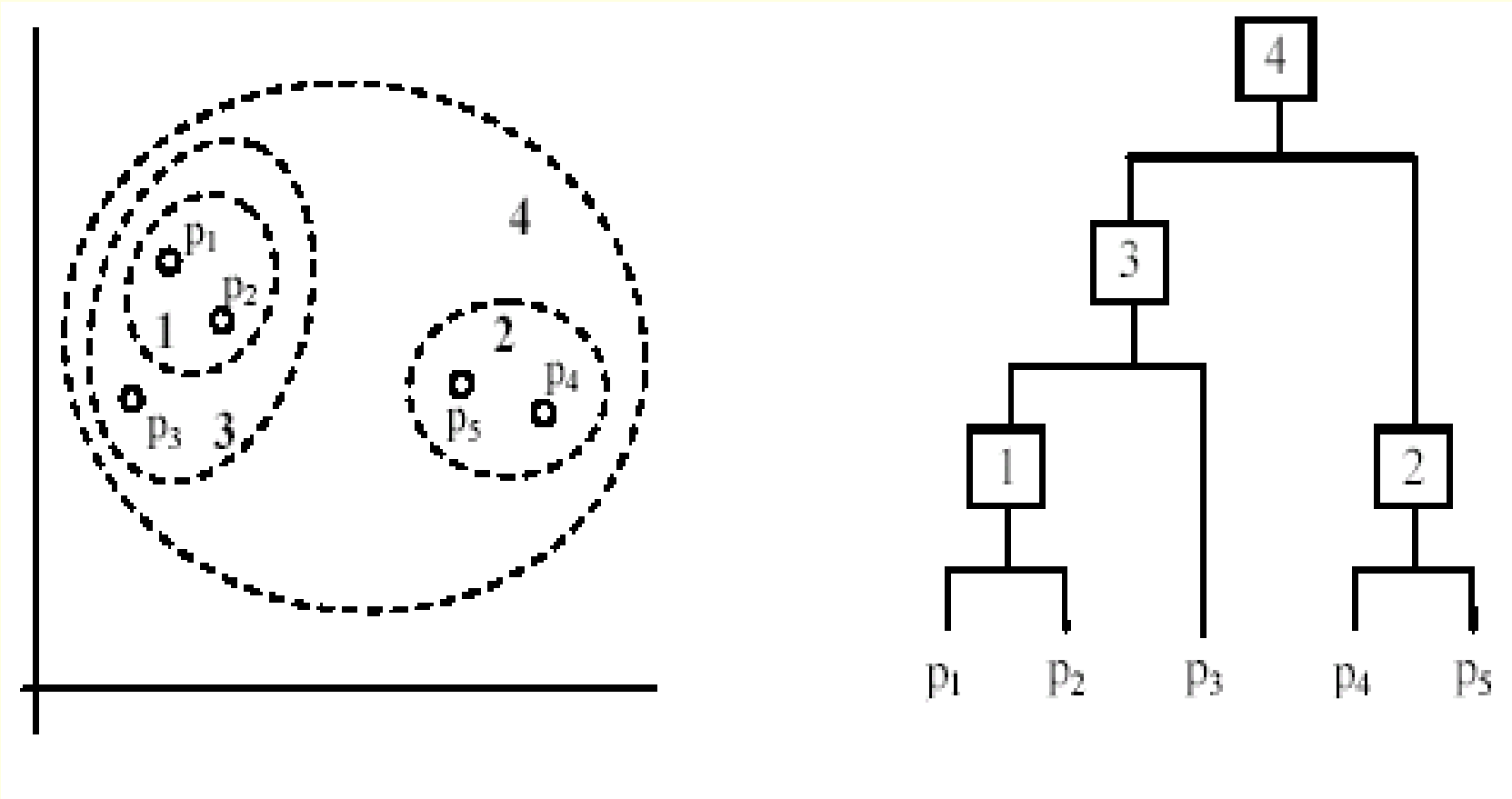
Métodos de agrupamiento

- Métodos de agrupamiento jerárquico aglomerativo:

- 1) Calcular las similitudes entre los documentos.
- 2) Crea un grupo con los dos documentos más cercanos.
- 3) Redefinir las similitudes entre el nuevo grupo y el resto de documentos, o grupos. El resto de valores permanecen inalterados.
- 4) Repetir 2) y 3) hasta que todos los documentos estén en un único grupo.

Métodos de agrupamiento

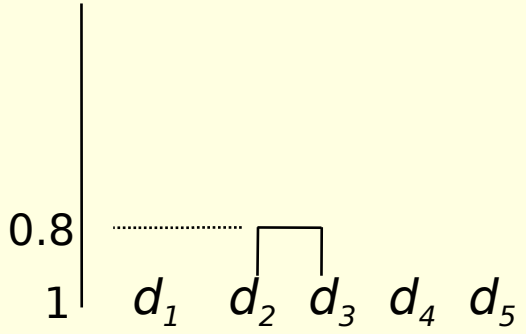
- Ejemplo:



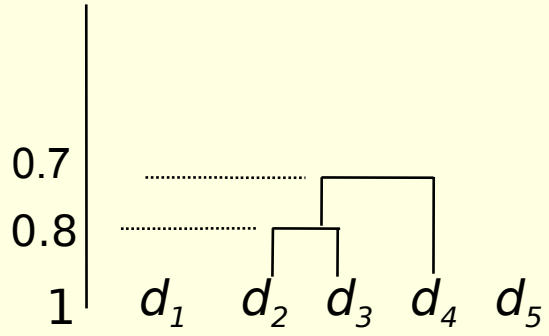
Métodos de agrupamiento

- Ejemplo:

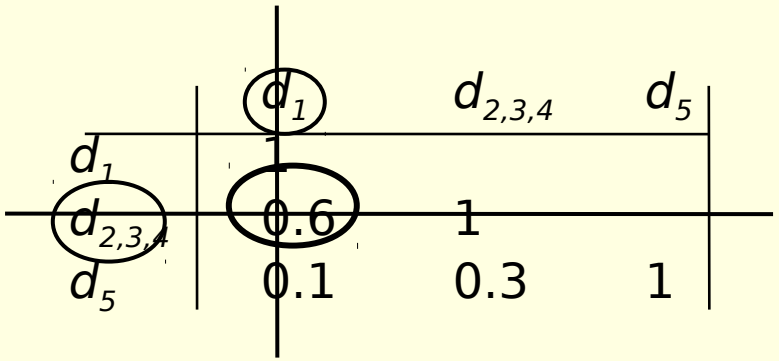
	d_1	d_2	d_3	d_4	d_5
d_1	1				
d_2	0.6	1			
d_3	0.4	0.8	1		
d_4	0.1	0.5	0.7	1	
d_5	0.1	0.2	0.2	0.3	1

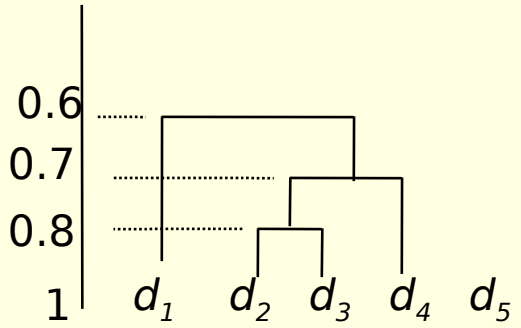


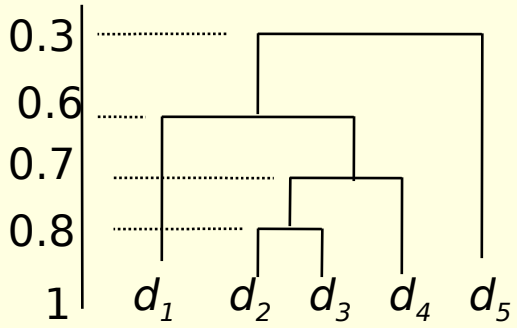
	d_1	$d_{2,3}$	d_4	d_5
d_1	1			
$d_{2,3}$	0.6	1		
d_4	0.1	0.7	1	
d_5	0.1	0.2	0.3	1



Métodos de agrupamiento







Métodos de agrupamiento

- Métodos para calcular la distancia entre grupos:

- Enlace único (single link):
  - La similitud entre dos grupos es la máxima de las similitudes entre todos los pares de documentos, tales que uno está en un grupo y otro en el segundo.
  - Cada documento de un grupo será más similar, como mínimo, a otro miembro de su mismo grupo, que a cualquier miembro de otro.

Métodos de agrupamiento

- Métodos para calcular la distancia entre grupos:

- Enlace completo (complete link):
  - La similitud entre dos grupos es la mínima de las similitudes entre todos los pares de documentos, tales que uno está en un grupo y otro en el segundo.
  - Cada documento de un grupo será más similar al documento más diferente de su grupo que al más diferente de cualquier otro grupo.

Métodos de agrupamiento

- Métodos para calcular la distancia entre grupos:

- Enlace medio (average link):
  - La similitud entre dos grupos es la media de las similitudes entre todos los pares de documentos, tales que uno está en un grupo y otro en el segundo.

- Método del centroide:
  - La similitud entre dos grupos es la similitud entre sus centroides.

Métodos de agrupamiento

- Eficiencia:

- Cálculo de todos las similitudes entre pares: $O(n^2)$.
- El resto de operaciones son el cálculo entre los grupos recientemente formados y el resto de grupos.

  - Todos los algoritmos son como mínimo $O(n^2)$.
  - Enlace simple: $O(n^2)$.
  - Enlace completo y medio: $O(n^2 \log n)$.

Métodos de agrupamiento

- Algunas consideraciones:

- El enlace simple produce grupos grandes pero no cohesionados.
- No es muy efectivo en términos de precisión y recall.

- El enlace completo, justo al contrario, grupos pequeños y muy cohesionados.
- Generalmente buenos grupos pero pocos.

- Enlace medio, en medio de los dos.

Métodos de agrupamiento

- Ventajas de los aglomerativos:
  - Propiedades teóricas.
  - Uso extensivo en el pasado.
  - Representación ordenada y jerárquica de colecciones, útil para la búsqueda.
  - Fáciles de implementar.

- Inconvenientes:
  - Los requerimientos de espacio y tiempo son un problema para colecciones grandes.
  - La actualización requiere un agrupamiento completo.

Sumario

- ¿Qué es el agrupamiento documental?
- Aplicaciones del agrupamiento.
- La base: “The cluster hypothesis”.
- Componentes del proceso.
  - Representación de documentos.
  - Medida de asociación.
  - Representación de los grupos.
  - Algoritmos de agrupamiento.
  - Evaluación del agrupamiento.

Evaluación del agrupamiento

- Encuestas a usuarios.
- Técnicas estadísticas para medir la calidad.
- Utilizar el modelo de recuperación basado en agrupamiento:
  - Dada una consulta, recuperar grupos completos o parte de ellos.
  - Los documentos pueden estar o no ordenados dentro del grupo.
  - El concepto de centroide del grupo es muy importante.
  - Evaluar la recuperación.

Retos actuales del agrupamiento

- Selección de atributos.
- Selección del método de agrupamiento.
- Generación de los representantes de grupos.
- Validación de la calidad del agrupamiento.
- Actualización de la estructura de grupos.
- Costos computacionales.

Ejemplo software de agrupamiento

- Carrot2:

  http://project.carrot2.org

Bibliografía básica

- van Rijsbergen, C.J., *Information Retrieval*. London: Butterworths, 2nd Edition, 1979;
  - http://www.dcs.gla.ac.uk/~keith/Preface.html
  -
- Tombros, A., PhD Thesis, 2002. Chapter 3 http://www.dcs.qmul.ac.uk/~tassos/publications.html

3ª Parte

Clasificación textual

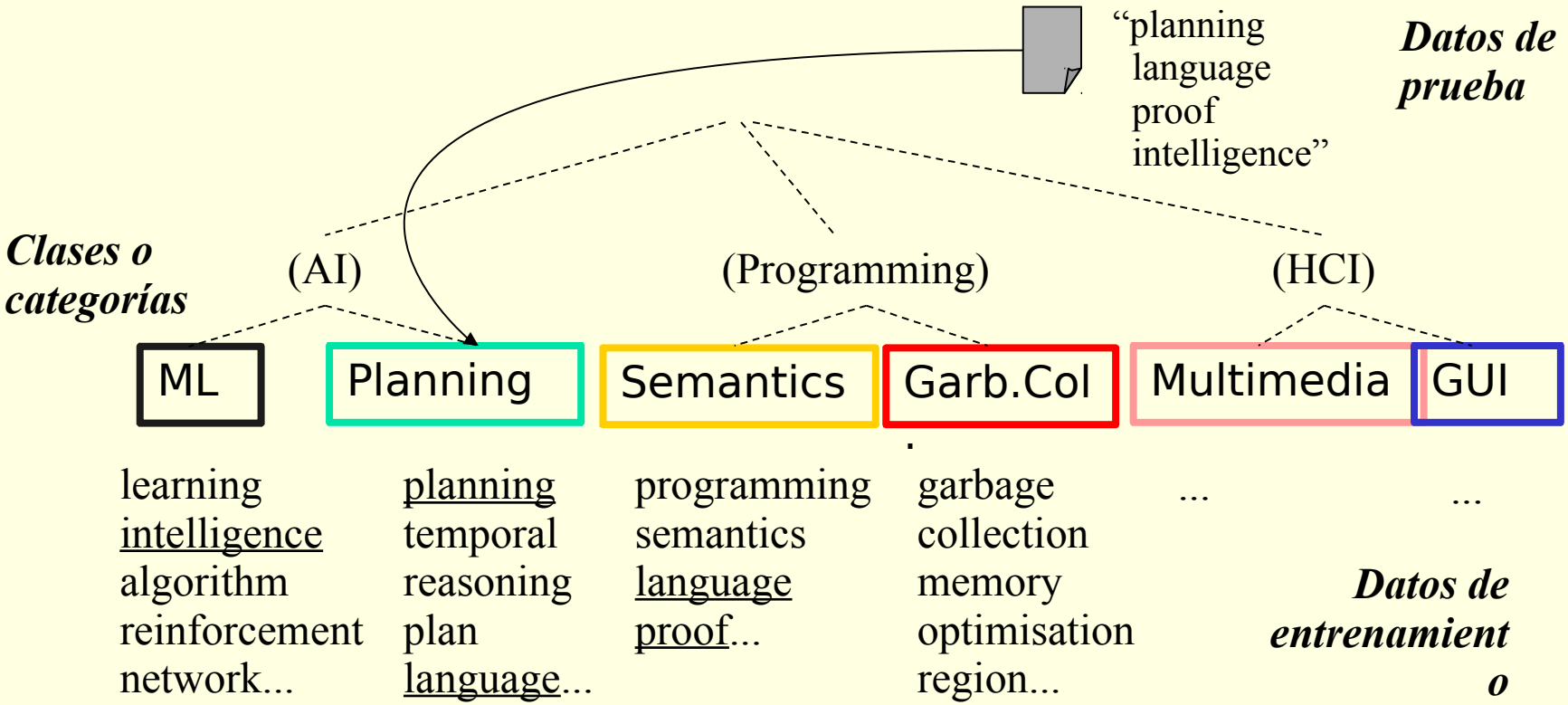
Sumario

- Definición de clasificación documental.
- Tipos de clasificaciones.
- Aplicaciones de la clasificación.
- Clasificación y aprendizaje automático.
- Indexación y reducción de la dimensionalidad.
- Métodos de clasificación.
- Evaluación de los clasificadores.

Definición de clasificación textual

- Dados:
  - Conjunto fijo de categorías: $C=\{c_1,c_2,\ldots, c_n\}$
  - Un documento $d_j \in$ *D, de la colección.*
- Objetivo:
  - Asignar un valor booleano a cada par $<d_j, c_i> \in D \times C$.
    - Verdadero → d_j se clasifica en c_i.
    - Falso → caso contrario.
  - Función de clasificación $\Phi: D \times C \rightarrow \{V, F\}$

Definición de clasificación textual



De una presentación sobre clasificación de Christopher Manning. Universidad de Stanford.

Sumario

- Definición de clasificación documental.
- Tipos de clasificaciones.
- Aplicaciones de la clasificación.
- Clasificación y aprendizaje automático.
- Indexación y reducción de la dimensionalidad.
- Métodos de clasificación.
- Evaluación de los clasificadores.

Tipos de clasificaciones

- Etiquetado simple vs. multietiquetado:
  - Única etiqueta: exactamente una categoría se asigna a cada documento.
  - Multietiquetado: cualquier número de etiquetas, desde 0 to |C|, se pueden asignar a cada documento.
  - Clasificación binaria: tipo especial de del primer tipo, donde $d_j \in D$ se asigna a c_i o a su complementaria (spam – no spam).
- El caso binario es más general que el multietiquetado, porque el problema de asignar varias etiquetas a un documento, se pude ver como varios problemas binarios, uno por categoría existente.

Tipos de clasificaciones

- Uso de clasificadores:
  - Dado un documento $d_j \in D$, queremos encontrar todas las categorías $c_i \in C$ bajo las que podríamos clasificarlo (clasificación centrada en el documento - document-pivoted).
    - Útil cuando los documentos llegan en diferentes momentos de tiempo, por ejemplo, el filtrado de mensajes de correo electrónico.
  - Dada una categoría $c_i \in C$, queremos encontrar todos los documentos $d_j \in D$ tales que quedan clasificados bajo la categoría (clasificación centrada en la categoría - category-pivoted).
    - Útil cuando se se añaden nuevas categorías a C.

Tipos de clasificadores

- Categorización estricta vs. ordenación de categorías:
  - Se toma una decisión estricta acerca de las categorías bajo las que se debe clasificar un documento, ó
  - se ordenan las categorías basándonos en su grado de adecuación, permitiendo que sea un humano el que tome finalmente la decisión.
  - Este último tipo puede originar un sistema de categorización interactivo, útil en aplicaciones críticas.

Tipos de clasificadores

- Clasificación jerárquica vs. plana:
  - En la primera, se establecen relaciones entre las propias clases, mientras que en la segunda, no existen.

Sumario

- Definición de clasificación documental.
- Tipos de clasificaciones.
- Aplicaciones de la clasificación.
- Clasificación y aprendizaje automático.
- Indexación y reducción de la dimensionalidad.
- Métodos de clasificación.
- Evaluación de los clasificadores.

Aplicaciones de la clasificación

- Indexación automática de sistemas de recuperación booleanos:
  - A cada documento se le asigna uno o varios términos de indexación a partir de un vocabulario controlado.
  - Las entradas de dicho vocabulario se pueden ver como categorías.
  - Visión centrada en el documento.

Aplicaciones de la clasificación

- Organización documental:
  - Por ejemplo, en un periódico, previamente a la publicación de anuncios, éstos pueden ser clasificados en “venta de coches”, “venta de casas”,...

Aplicaciones de la clasificación

- Filtrado de texto:
  - Clasificación de un flujo de documentos entrantes, dependiendo de su relevancia para el usuario. (Por ejemplo, una agencia de noticias, que manda unas noticias a unos clientes y otras, a otros).
  - Normalmente binario (relevante – no relevante).
  - Es común tener un perfil del usuario, donde éste establezca sus gustos, y que pueda ser actualizado, explícitamente por el propio usuario o implícitamente por el sistema (Filtrado adaptativo).

Aplicaciones de la clasificación

- Desambiguación del significado de las palabras:
  - ¿banco = institución financiera o asiento?
  - Desambiguación: dada la ocurrencia ambigua de una palabra en un texto, asignarle su significado correcto.
  - Los contextos de las ocurrencias de las palabras se pueden ver como documentos, mientras que los significados como categorías.
  - Se dispone de un conjunto de “documentos” asignados a las “categorías” correctas, y se intenta encontrar el significado correcto de nuevas palabras en un contexto.

Aplicaciones de la clasificación

- Clasificación jerárquica de páginas web:
  - ... bajo un catálogo jerárquico, o directorio, por ejemplo, Yahoo!.
  - Los usuarios pueden encontrar más sencillo navegar por la jerarquía.
  - Peculiaridades:
    - La naturaleza hipertextual es útil (se pueden utilizar los enlaces entre páginas).
    - La estructura jerárquica de las categorías puede ser también útil: el problema de la clasificación se puede descomponer en problemas más simples tomando en consideración las ramas de los nodos internos.

Sumario

- Definición de clasificación documental.
- Tipos de clasificaciones.
- Aplicaciones de la clasificación.
- Clasificación y aprendizaje automático.
- Indexación y reducción de la dimensionalidad.
- Métodos de clasificación.
- Evaluación de los clasificadores.

Clasificación y aprendizaje automático

- El enfoque actual con el que se afronta la clasificación documental es mediante el aprendizaje automático (machine learning).
- Objetivo: construir un clasificador para una clase c_i observando las características de los documentos asignados a ella (aprendizaje).
- A partir de estas propiedades, un documento nuevo, no observado, podrá ser clasificado o no bajo c_i.
- Aprendizaje supervisado.

Clasificación y aprendizaje automático

- Conjunto inicial: $\Omega=\{d_1, \ldots, d_{|\Omega|}\} \subset D$ *de documentos* preclasificados bajo clases del conjunto $C=\{c_1, \ldots, c_{|C|}\}$.
- Antes de construir el clasificador, se suele hacer una división del conjunto anterior en dos:
  - Conjunto de aprendizaje: $TS=\{d_1, \ldots, d_{|TS|}\}$ – El clasificador se construye observando las características de estos documentos.
  - Conjunto de prueba: $Te=\{d_{|TS|+1}, \ldots, d_{|\Omega|}\}$ – Usado para probar el rendimiento del clasificador: cada documento de este subconjunto se pasa al clasificador y su salida se compara con la real.
  - Ningún documento en Te puede participar en la construcción del clasificador.

Clasificación y aprendizaje automático

- Validación cruzada:
  - El conjunto T se divide en k subconjuntos disjuntos: Te_1, …, Te_k.
  - Se aplican los métodos de clasificación a los conjunto de entrenamiento y de prueba, iterativamente.
  - La efectividad final del clasificador es algún tipo de agregación de los resultados individuales.

Sumario

- Definición de clasificación documental.
- Tipos de clasificaciones.
- Aplicaciones de la clasificación.
- Clasificación y aprendizaje automático.
- Indexación y reducción de la dimensionalidad.
- Métodos de clasificación.
- Evaluación de los clasificadores.

Indexación y reducción de la dimensionalidad

- Los clasificadores necesitan poner la colección documental en un formato adecuado para su procesamiento: indexación.
- Documentos representados mediante vectores de términos (características) ponderados.
- Se usan esquemas de ponderación comunes a los utilizados en R.I.
  - Normalmente basados en *tf-idf.*

Indexación y reducción de la dimensionalidad

- El gran número de términos asignados a un documento, una vez realizada la indexación, puede llegar a ser un problema para los clasificadores.
- Reducción de la dimensión del problema.
- Peligro: pérdida de información sobre la semántica de los documentos.

Indexación y reducción de la dimensionalidad

- Reducción de la dimensionalidad:
  - Local: para cada categoría, c_i, se elige un subconjunto de términos para realizar la clasificación bajo c_i.
  - Global: se elige el subconjunto para todas las categorías a la vez.

Indexación y reducción de la dimensionalidad

- Términos resultantes:
  - Por selección: un subconjunto del conjunto original.
  - Por extracción: los términos utilizados no son del mismo tipo que los del conjunto original. Se obtienen a partir de combinaciones o transformaciones de los términos originales.

Indexación y reducción de la dimensionalidad

- Selección de términos:
  - Basado en la frecuencia de los términos:
    - Se eliminan términos infrecuentes.
    - Términos que aparecen en x documentos, también se eliminan (por ejemplo, x=1,...,3).
    - Eliminación de palabras vacías.

Indexación y reducción de la dimensionalidad

- Selección de términos:
  - Basado en teoría de la información:
    - Factor de asociación DIA.
    - Chi-cuadrado.
    - Ganancia de información.
    - Información mutua esperada.
    - ...
    - Intentan capturar la intuición de que los mejores términos para una categoría se distribuyen de manera diferente en los conjuntos de ejemplos positivos o negativos para c_i.

Indexación y reducción de la dimensionalidad

- Extracción de términos:
  1) Método de extracción.
  2) Método para convertir los documentos originales a la nueva representación.
  3)
  - Agrupamiento de términos: se intenta agrupar términos con una proximidad semántica muy alta, sustituyéndolos por el representante de la clase.
  - Latent semantic indexing.

Sumario

- Definición de clasificación documental.
- Tipos de clasificaciones.
- Aplicaciones de la clasificación.
- Clasificación y aprendizaje automático.
- Indexación y reducción de la dimensionalidad.
- Métodos de clasificación.
- Evaluación de los clasificadores.

Métodos de clasificación

- Algoritmo de Rocchio.
- K vecinos más cercanos (KNN).
- Árboles de decisión (DT).
- Naïve Bayes (NB).
- Support Vector Machine (SVM)
- Algoritmos de votación.

Métodos de clasificación

- Rocchio:
  - Representando los documentos del conjunto de entrenamiento como vectores, construye un representante para cada clase, c_i.

$$C_i = \alpha * centroide_{C_i} - \beta * centroide_{\overline{C_i}}$$

  - Dado un documento de prueba, calcula la similitud entre éste y los centroides de cada clase.
  - Se asigna a aquella clase con máxima similitud.

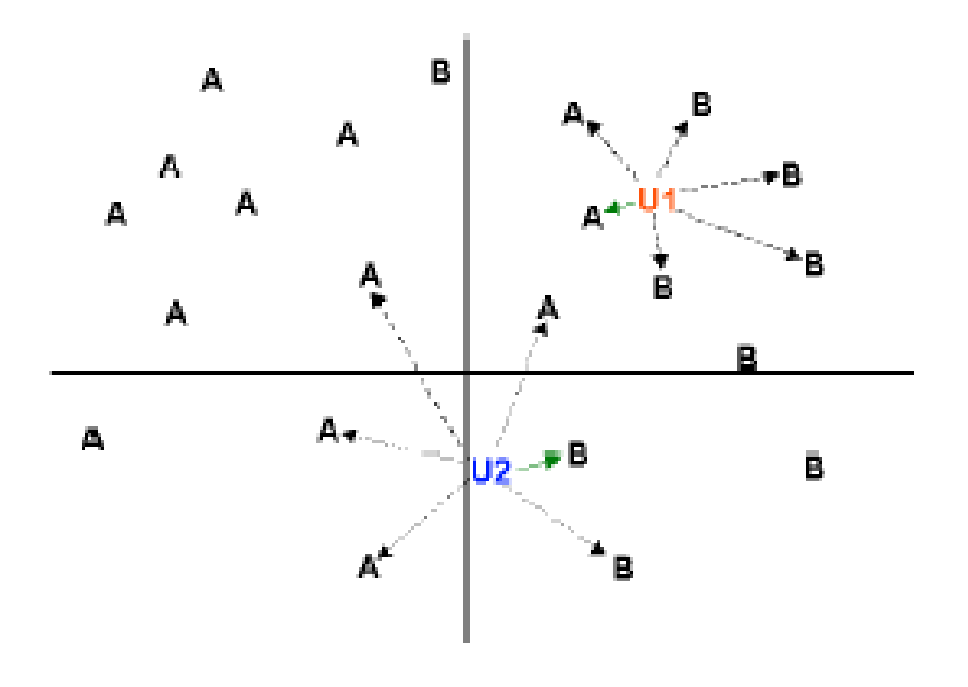
Métodos de clasificación

- Ventajas:
  - Fácil de implementar.
  - Rápido en el aprendizaje.
  - Basado en en el mecanismo de realimentación por relevancia de R.I.

- Desventajas:
  - Bajo rendimiento en la clasificación.
  - Combinación lineal: demasiado simple.
  - Las constantes se determinan empíricamente.

Métodos de clasificación

- K vecinos más cercanos:
- Principio: los documentos que están cerca en el espacio pertenecen a la misma clase.

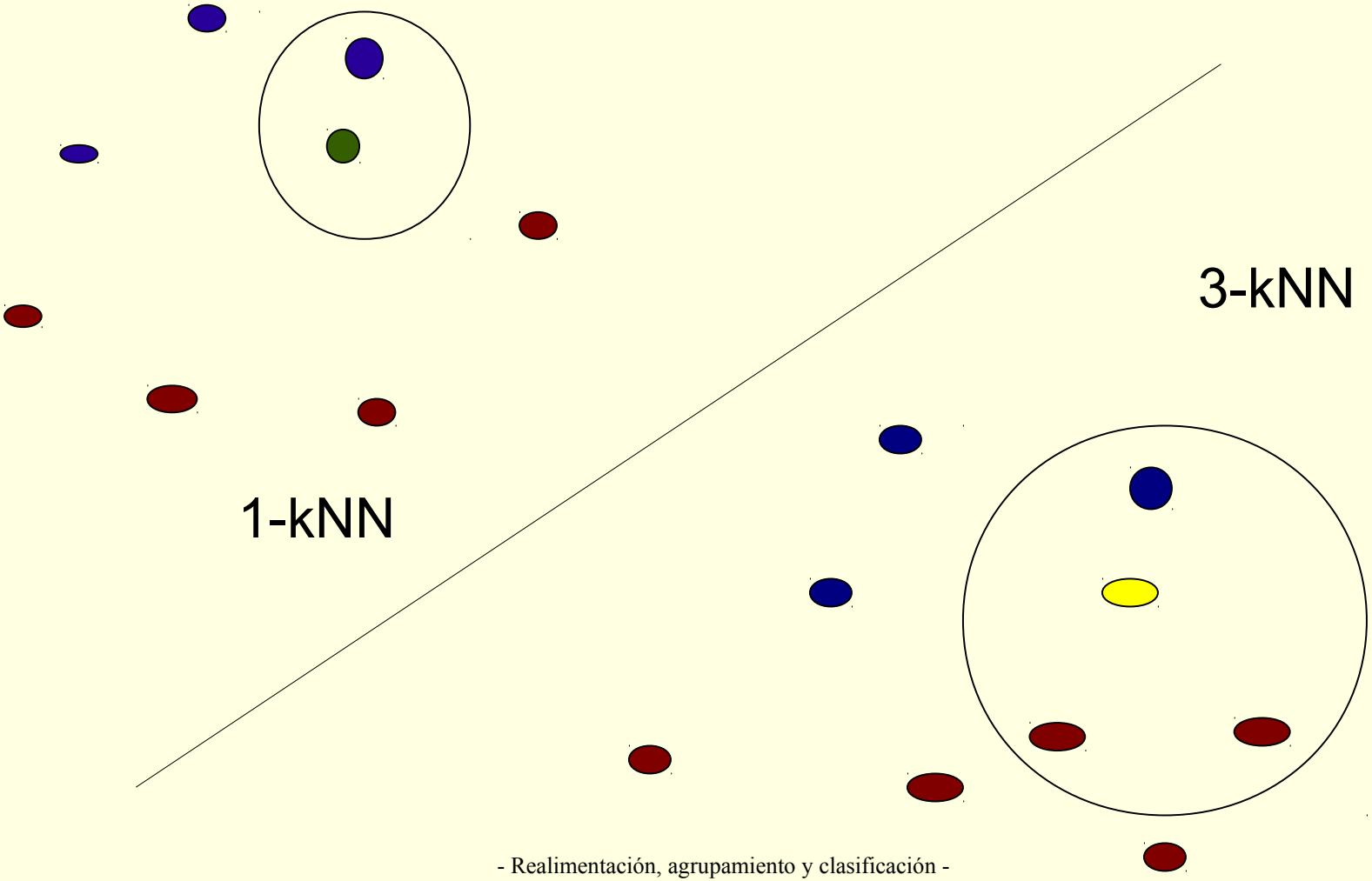


Obtenido de la presentación de JY Nie del curso IFT6255.

Métodos de clasificación

- Se calcula la similitud entre el documento a clasificar y los documentos de entrenamiento (distancia euclídea o coseno).
- Se seleccionan los k vecinos más cercanos.
- Se asigna el documento a la clase que contenga más vecinos.
- Diferentes formas de contar los votos de los vecinos.

Métodos de clasificación



Métodos de clasificación

- Ventajas:
  - Efectivo y simple.
  - No tiene parámetros.
  - Se consideran más características locales que Rocchio.
- Inconvenientes:
  - Tiempo de clasificación largo.
  - Difícil de encontrar el valor óptimo de k.
  - El rendimiento depende de la medida de similitud.

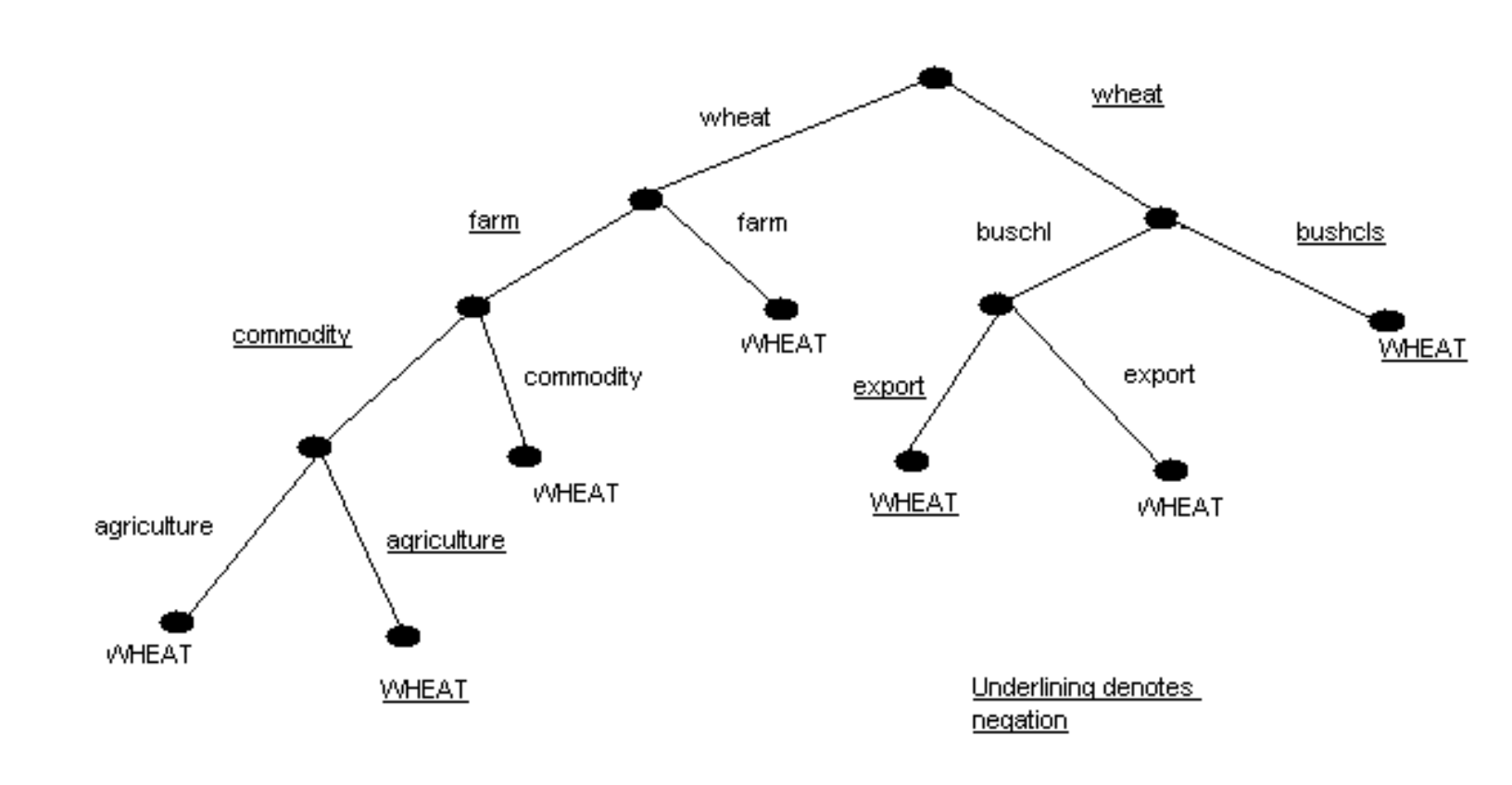
Métodos de clasificación

- Árboles de decisión:

- El nodo raíz contiene todos los documentos.
- Cada nodo interno representa un subconjunto de documentos separados de acuerdo a un atributo.
- Cada arco está etiquetado con un predicado que puede ser aplicado al atributo del padre.
- Cada nodo hoja está etiquetada con una clase.

Métodos de clasificación

- Árboles de decisión:



Métodos de clasificación

- Procedimiento de partición recursivo a partir del nodo raíz.
- El conjunto de documentos queda separado en subconjuntos de acuerdo con un atributo.
- Se utiliza el atributo más discriminativo primero (normalmente, mayor ganancia de información).

Métodos de clasificación

- Ventajas:
- Fácil de entender.
- Fácil generación de reglas.
- Se reduce la complejidad del problema.
- Inconvenientes:
- Tiempo de entrenamiento relativamente alto.
- Un documento sólo está conectado con una rama.
- Una vez que se comete un error en un nivel alto, cualquier subárbol es incorrecto.
- Puede sufrir de sobreentrenamiento.

Métodos de clasificación

- Naïve Bayes:
- Basado en el teorema de Bayes:

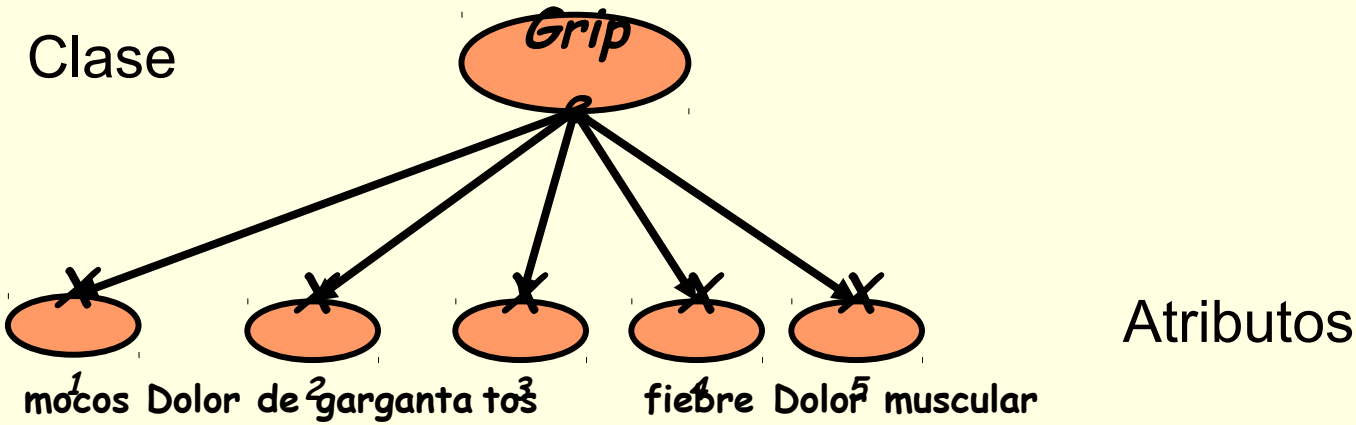
$$P(c_i|d_j)=\frac{P(c_i)P(d_j|c_i)}{P(d_j)}=P(c_i)P(d_j|c_i)$$

  - $P(c_i|d_j)$: probabilidad de que d_j pertenezca a c_i.
  - $P(d_j)$: probabilidad de que un documento seleccionado al azar tenga el vector d_j como su representación (ignorado, ya que es constante).
  - $P(c_i)$: probabilidad de que un documento seleccionado aleatoriamente pertenezca a c_i.
  - $P(d_j|c_i)$: cálculo costoso ya que hay muchos documentos.

Métodos de clasificación

- Naïve Bayes:
- Se supone que los términos son independientes a la hora de calcular $P(d_j|c_i)$.
- Y por tanto, $$P(d_j|c_i)=\prod_{k=1}^{|T|} P(w_{kj}|c_i) \quad (2)$$
- $P(w_{kj}|c_i)$ es fácil de calcular para cada término y para cada clase a partir del conjunto de entrenamiento.
- Al igual que $P(c_i)$.

Métodos de clasificación



- Suposición de independencia condicional de las características dada la clase:

$$P(X_1,\ldots,X_5|C)=P(X_1|C)\cdot P(X_2|C)\cdot\cdots\cdot P(X_5|C)$$

Métodos de clasificación

- Estimación de probabilidades: un enfoque sencillo basado en ocurrencias en el conjunto de entrenamiento.

$$P(c_i) = \frac{N(C=c_i)}{N}$$

$$P(x_j|c_i) = \frac{N(X_j=x_j, C=c_i)}{N(C=c_i)}$$

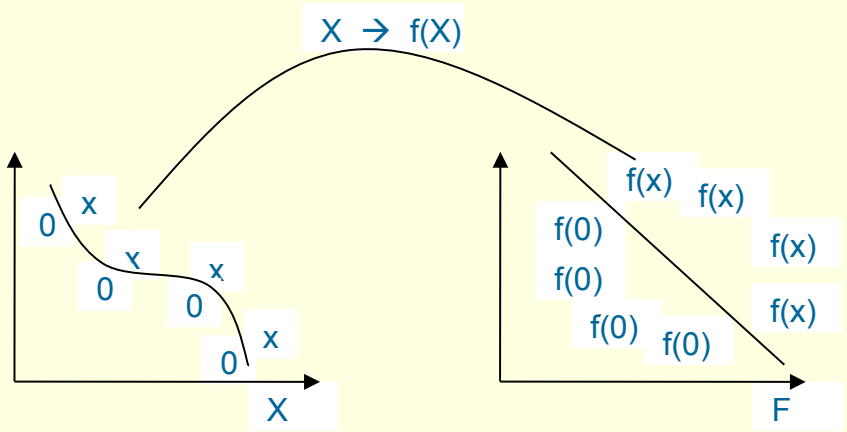
Métodos de clasificación

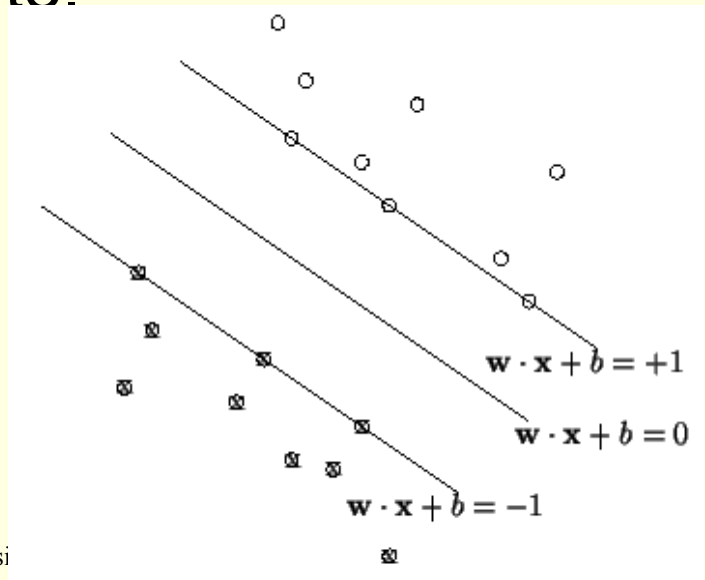
- Ventajas:
  - Modelo simple y fácil de implementar.
  - Eficiente en términos de espacio y tiempo.
- Inconvenientes:
  - La suposición de independencia es demasiado fuerte.
  - No es muy buen en términos de clasifiación.

Métodos de clasificación

- Support Vector Machines:

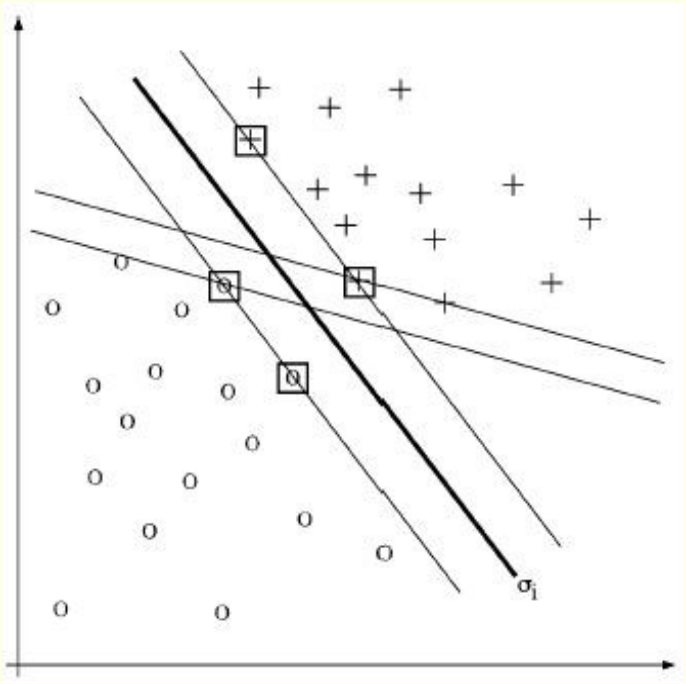
- Encuentra la “línea” que separa el hiperplano que maximiza el margen entre los ejemplos positivos y negativos de entrenamiento.





Métodos de clasificación

- Support Vector Machines:



Gráfica de [Seb02]

- Las líneas representan superficies de decisión.
- La superficie σ_1 es la mejor posible.
  - El elemento intermedio del conjunto más amplio de superficies de decisión paralelas.
  - Mínima distancia a cualquier ejemplo de entrenamiento máxima.
  - Los cuadrados indican los vectores de soporte, es decir, el conjunto de ejemplos de entrenamiento usados en la decisión.

Métodos de clasificación

- Ventajas:
  - Clasificación muy efectiva.
  - Pueden trabajar con datos de dimensinalidad alta.
  - No se suele necesitar reducción de dimensionalidad.
- Inconvenientes:
    - Puede llegar a ser costosas, pero existen algoritmos muy eficientes.
    - Selección de parámetros y del kernel.

Métodos de clasificación

- Algoritmos por votación:

- Uso de múltiples clasificadores (pobres) combinados en un único (de calidad).

  - Se disponen de varios clasificadores base.
  - La salida de cada clasificador se combina con la del resto y se genera una salida final.

Métodos de clasificación

- Ejemplo: Adaboost.

- Ventajas:
  - Muy efectivos.
  - Robustos al ruido.

- Inconvenientes:
  - Requieren más cómputo y memoria.

Sumario

- Definición de clasificación documental.
- Tipos de clasificaciones.
- Aplicaciones de la clasificación.
- Clasificación y aprendizaje automático.
- Indexación y reducción de la dimensionalidad.
- Métodos de clasificación.
- Evaluación de los clasificadores.

Evaluación de clasificadores

- Necesitamos conocer el rendimiento de los clasificadores. Normalmente se lleva a cabo experimentalmente, más que analíticamente.
- Comparando ese rendimiento bajo un mismo banco de datos se pueden establecer los mejores clasificadores.
- Efectividad del clasificador: una función que mide cuántas decisiones correctas ha realizado.

Evaluación de clasificadores

- Necesitamos bancos de datos diseñados para tal fin.
- Colecciones para clasificación:
  - Conjunto de documentos de aprendizaje con sus clases correspondientes.
  - Conjunto de documentos de prueba con sus clases correspondientes.
- Ejemplos: Reuters 21578, 22173, Reuters RCV1

Evaluación de clasificadores

- ¿Qué podemos evaluar?
- Eficiencia:
  - Tiempos de entrenamiento y prueba, así como requerimientos de espacio.
- Eficacia:
  - Habilidad de tomar las decisiones de clasificación correctas.
- Objetivo:
  - Conseguir una alta capacidad de clasificación a la vez que eficiencia computacional.

Evaluación de clasificadores

- Eficiencia:

- Rara vez utilizada, aunque muy importante para clasificadores reales.
- Difícil de comparar porque los entornos cambian.
- Eficiencia en el entrenamiento: tiempo medio para construir un clasificador por categoría con el conjunto de entrenamiento.
- Eficiencia de la clasificación: tiempo medio de clasificación de un nuevo documento.

Evaluación de clasificadores

- Eficacia:

- Una vez construido el clasificador utilizando el conjunto de prueba, se evalúa la eficiencia utilizando el conjunto de test, calculando, para cada categoría c_i:
  - TP_i: verdaderos positivos.
  - FP_i: falsos positivos.
  - TN_i: verdaderos negativos.
  - FN_i: falsos negativos.

Evaluación de clasificadores

- TP_i: verdaderos positivos para c_i:
  - El conjunto de documentos que, tanto el clasificador como los juicios almacenados en el conjunto de prueba, se clasifican bajo c_i.
  -
- FP_i: falsos positivos para c_i:
  - El conjunto de documentos que el clasificador clasifica bajo c_i, pero el conjunto de prueba indica lo contrario.

Evaluación de clasificadores

- TN_i: verdaderos negativos para c_i:
  - El conjunto de documentos que, tanto el clasificador como los juicios almacenados en el conjunto de prueba, no pertenecen a c_i.
  -
- FN_i: falsos negativos para c_i:
  - El conjunto de documentos que el clasificador no clasifica bajo c_i, pero el conjunto de prueba indica lo contrario, que debían ser clasificados como c_i.

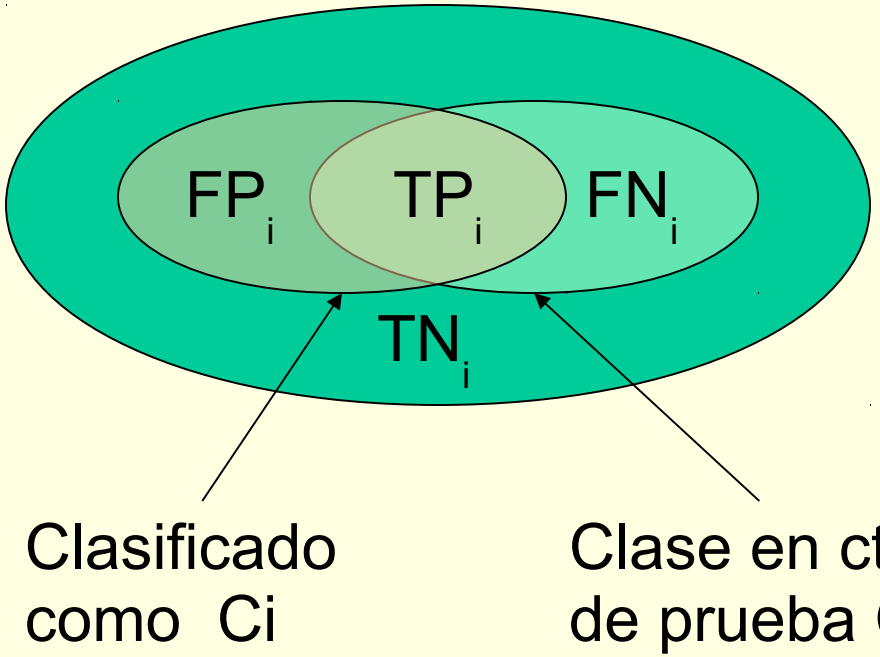
Evaluación de clasificadores

- Precisión para c_i:

$$p_i = \frac{TP_i}{TP_i + FP_i}$$

- Recall para c_i:

$$r_i = \frac{TP_i}{TP_i + FN_i}$$



Evaluación de clasificadores

- Se necesitan estimadores de precisión y recall para la colección completa.

- Dos métodos:
  - "Microaveraging": contar los verdaderos positivos y falsos positivos de todas las clases. La precisión y el recall se calcula utilizando los valores globales.
  - "Macroaveraging": media de la precisión (recall) de las categorías individuales.

Evaluación de clasificadores

microaveraging

$$P = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FP_i)}$$

$$R = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FN_i)}$$

macroaveraging

$$P = \frac{\sum_{i=1}^{|C|} P_i}{|C|}$$

$$R = \frac{\sum_{i=1}^{|C|} R_i}{|C|}$$

Evaluación de clasificadores

- Ambas medidas pueden dar resultados muy diferentes, si las categorías tienen diferente grado de generalidad (número bajo de ejemplos de entrenamiento positivos).
- La habilidad de comportarse bien con categorías poco generales será resaltada por la medida macroaveraging.
- La decisión dependen de la aplicación.

Evaluación de clasificadores

- Combinación de ambas medidas:
  - Precisión y recall no tienen sentido de manera aislada.
  - Si clasificamos correctamente cada documento bajo su categoría, entonces recall=1, pero la precisión sería muy baja.
  - Valores altos de precisión se obtendrían pagando el precio de valores bajos de recall.

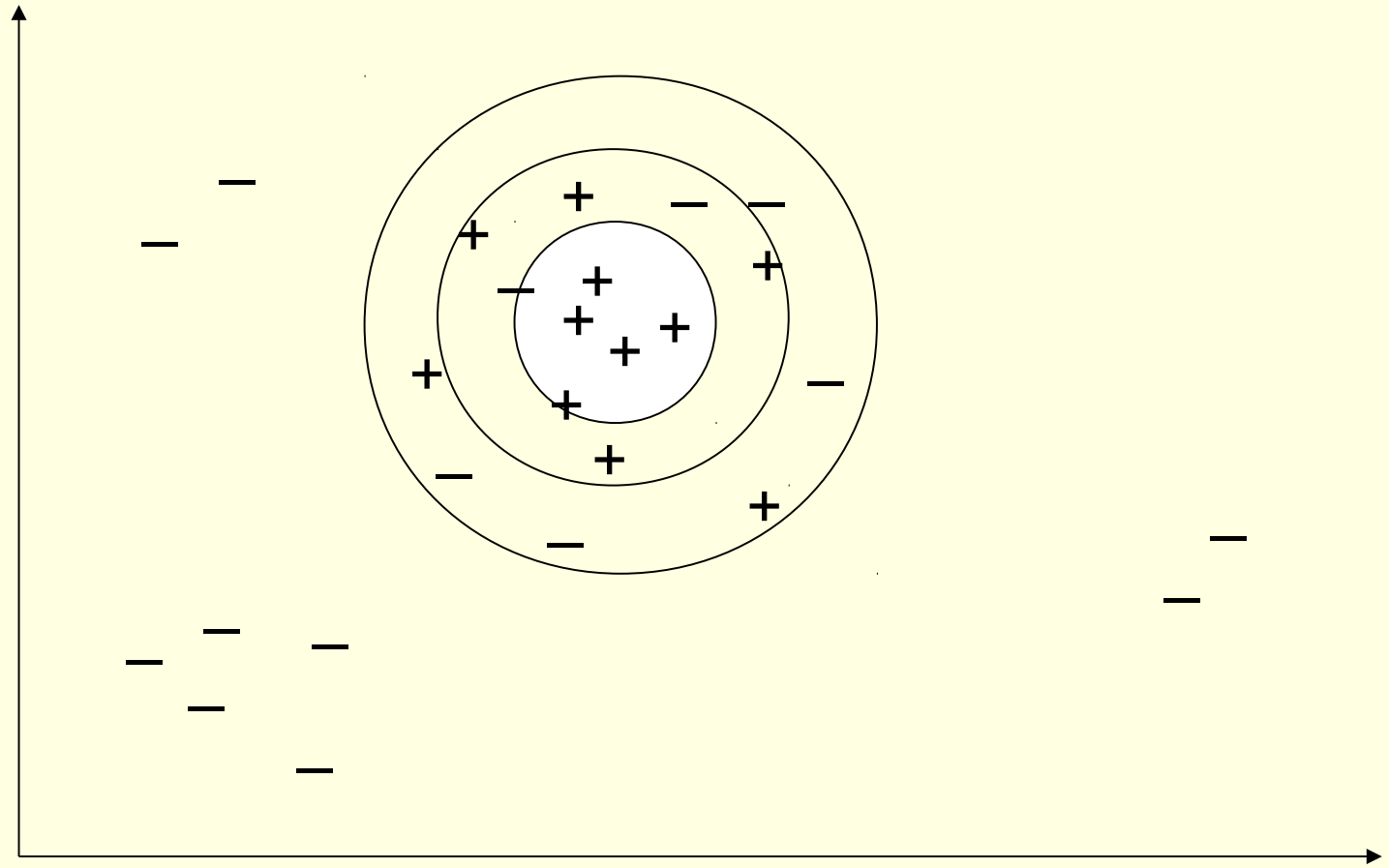
Evaluación de clasificadores

- Un clasificador debe ser evaluado con una medida que combine precisión y recall:
  - Media de los 11 puntos de precisión.
  - El punto “breakeven”.
  - Medida F1.

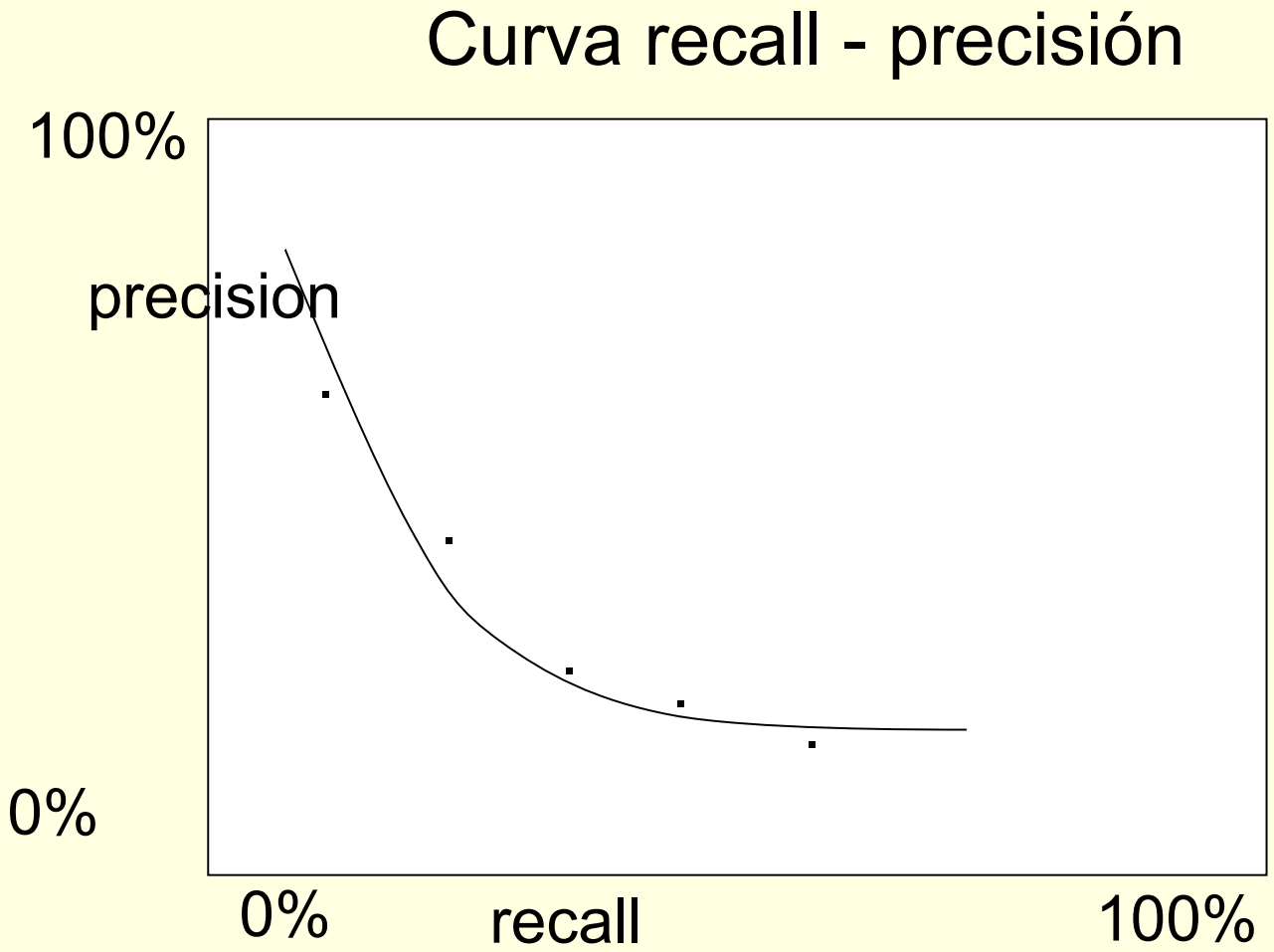
Evaluación de clasificadores

- Media recall para los 11 puntos de precisión:
  - Se calcula la precisión para cada valor 0.0, 0.1, 0.2,..., 1.0 de recall para cada categoría.
  - El valor medio será la medida de rendimiento.

Evaluación de clasificadores



Evaluación de clasificadores



Evaluación de clasificadores

- Punto “Breakeven”:
  - Se calcula la precisión en función del recall.
  - Breakeven es el valor para el que la precisión es igual que el recall.

Evaluación de clasificadores

- Medida F_β:

$$F_\beta = \frac{(\beta^2 + 1)\, pr}{\beta^2\, p + r}$$

  - β da importancia a p o a r. Lo normal β = 1.

- Medida F_1:

$$F_1 = \frac{2\, pr}{p + r}$$

Evaluación de clasificadores

- Precisión (accuracy):

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

  - No se utiliza mucho en clasificación textual.
  - Un valor alto en el denominador la hace insensible a variaciones en el número de decisiones correctas (TP + TN).

Evaluación de clasificadores

- ¿Qué clasificador es el mejor?
-
  - Normalmente K-NN y SVM tienen un rendimiento bueno.
  - Naïve Bayes... rendimiento aceptable, simple y rápido.
  - El rendimiento depende de factores experimentales como las características del conjunto de documento, el número de ejemplos de entrenamiento por categoría, etc.

Esto esto todo...

Gracias por la atención prestada