UNIVERSIDAD DE GRANADA

decsai.ugr.es

Minería de Medios Sociales

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

decsai.ugr.es

Bloque II: Minería de Texto y de la Web

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

decsai.ugr.es

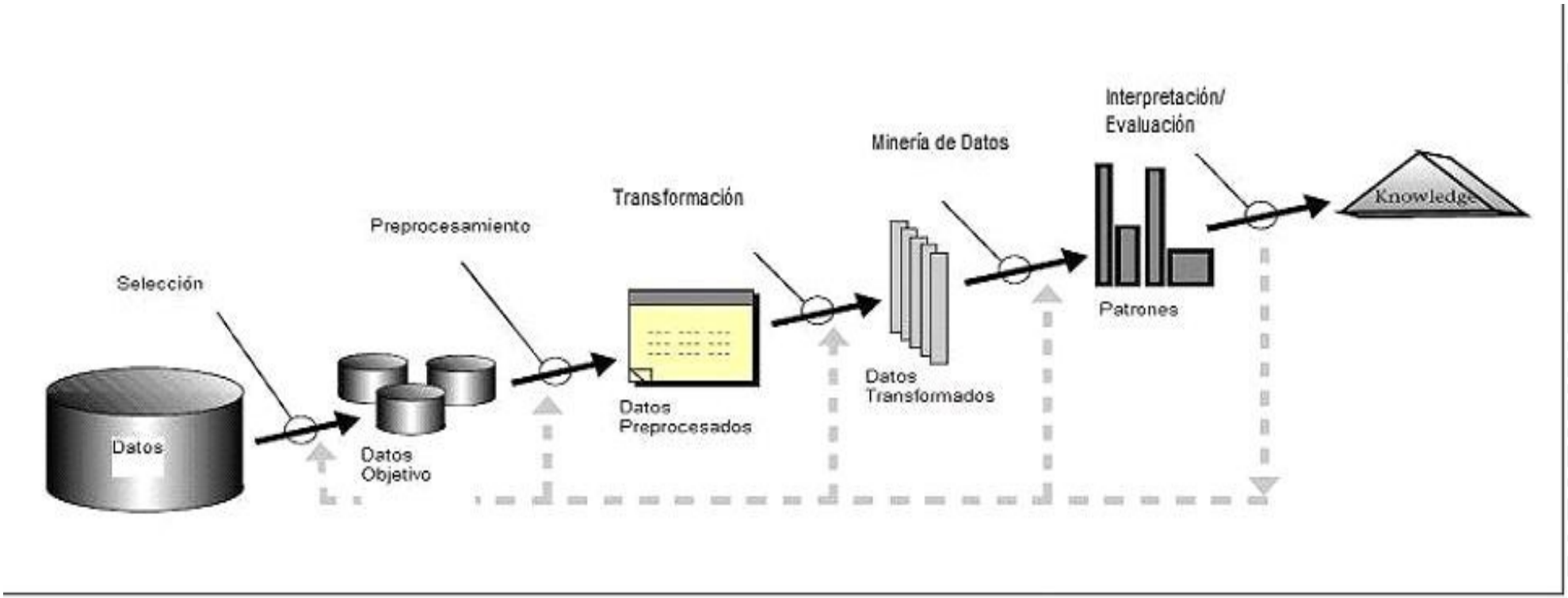
Sesión II.1: Minería de Textos

Departamento de Ciencias de la Computación e Inteligencia Artificial

Conceptos previos: KDD

KDD (Knowledge Discovery in Databases)

Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos.



Conceptos previos: Minería de Datos

Minería de Datos (o Data Mining (DM))

Proceso de descubrimiento eficiente de patrones, desconocidos a priori, en grandes bases de datos.

Relación existente entre KDD y DM:

- DM como fase de KDD
- DM como sinónimo de KDD

De los datos al texto

Motivación

- La gran mayoría de los datos susceptibles de ser procesados en redes sociales, informes de empresa, foros, páginas web, … son textuales.
- Los datos textuales carecen de estructura y homogeneidad. Los datos son simbólicos y los atributos son desconocidos a priori.
- Pueden estar en diferentes lenguajes y fuentes y además estar incompletos, ser irrelevantes o tener ruido.
- Dependen del contexto.

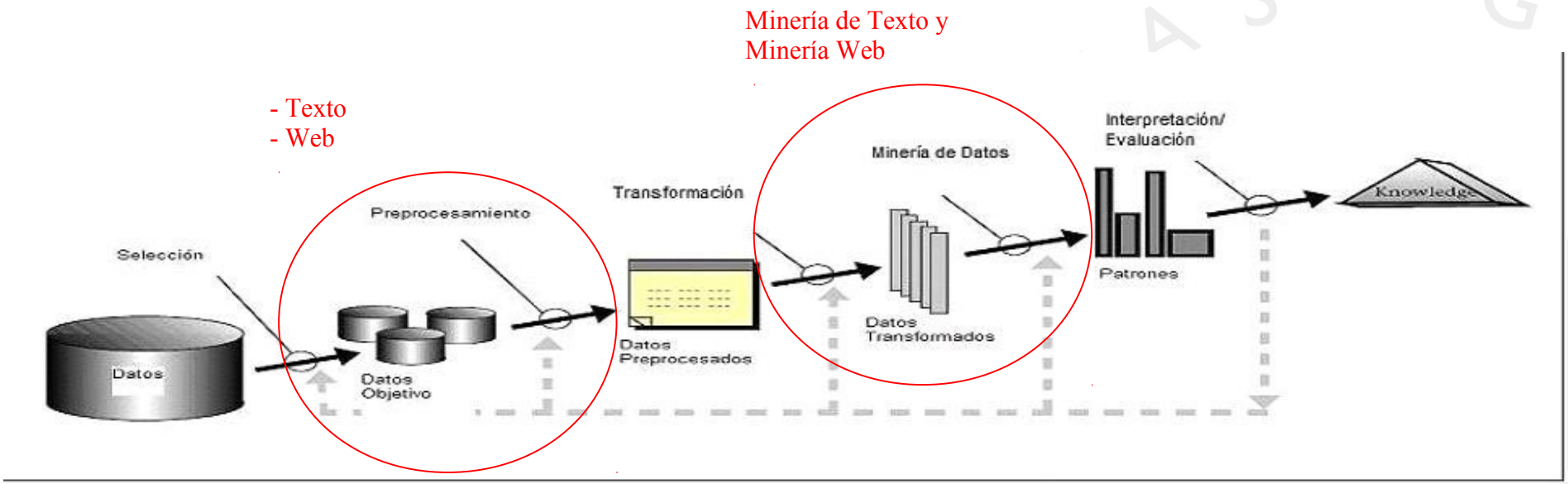
¿TM = KDT?

¿TM = DM con Texto?

KDT

KDT (Knowledge Discovery in Text)

El proceso de descubrir información útil que no está explícitamente en ninguno de los documentos analizados y que aparece cuando los documentos son analizados y relacionados.



KDT versus KDD

KDD

- Entender el dominio de la aplicación
- Seleccionar el conjunto de datos objetivo
- Limpiar, preprocesar y transformar los datos
- Desarrollo del modelo y construcción de la hipótesis
- Selección y ejecución de algoritmos de minería de datos
- Interpretación de resultados y visualización

KDT

- El usuario define conceptos interesantes
- Los textos se obtienen via RI o manualmente
- Los textos y conceptos se representan en una Forma Intermedia
- Identificación de conceptos en la colección de textos
- Algoritmos de Minería de textos
- Interpretación de resultados mediante un humano

KDT versus RI

- La Recuperación de Información (RI) ayuda a KDT para la recuperación de documentos y su preprocesamiento.

- Tienen diferentes objetivos:

  RI: Optimizar la consulta, la búsqueda y la recuperación de documentos.

  KDT: Extraer conocimiento no explícito en los datos.

KDT versus EI

- La Extracción de Información (EI) localiza unidades de texto relevantes para el usuario.
- La EI transforma un documento escrito en Lenguaje Natural en una representación estructurada o basada en slots.
- La EI puede utilizarse en la etapa de preprocesamiento de KDT.
- Elementos básicos que la EI puede obtener:
  - Entidades
  - Atributos
  - Hechos
  - Eventos

UNIVERSIDAD DE GRANADA

decsai.ugr.es

Preprocesamiento

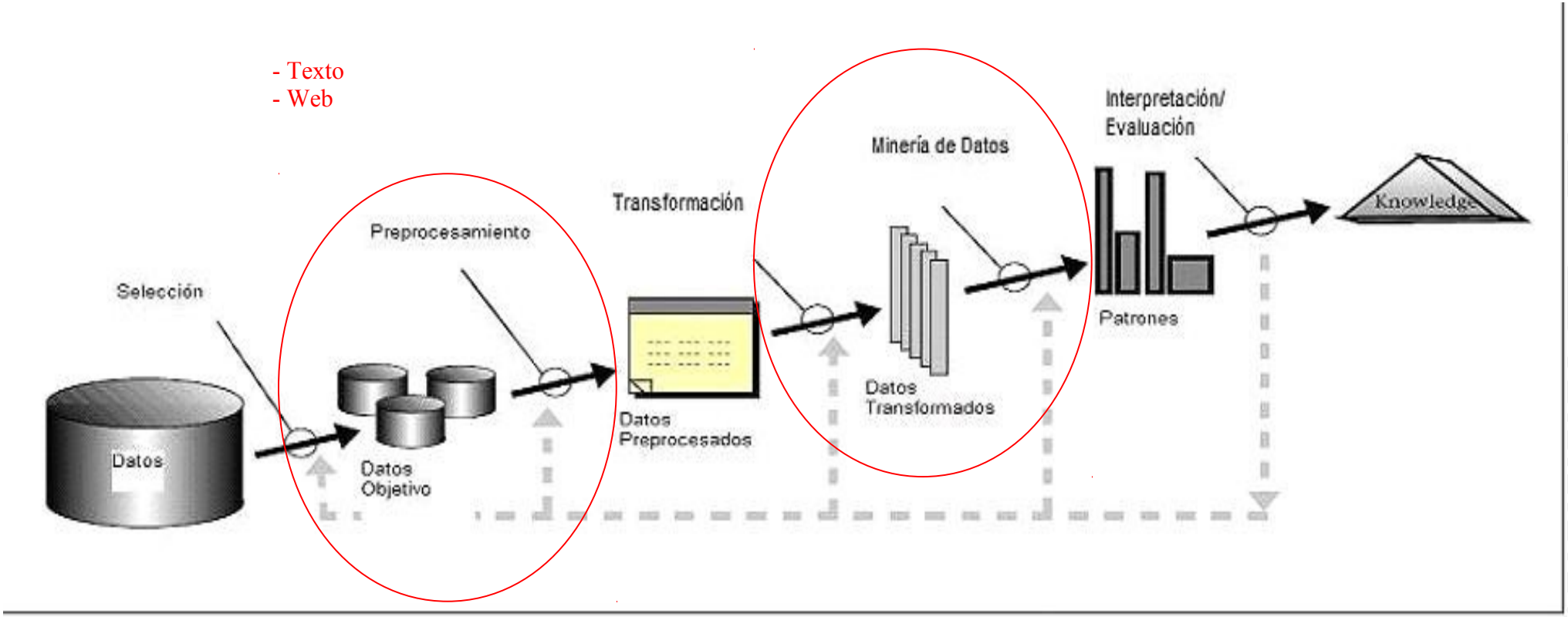
DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

KDT

Minería de Texto y
Minería Web

- Texto
- Web



Formas Intermedias

- Una **Forma Intermedia** es un modelo de representación del conocimiento capaz de expresar el contenido implícito del texto de una forma computable mediante un algoritmo o un programa.

- La técnica para conseguir la forma intermedia determina el tipo de información a conseguir en el proceso de descubrimiento

- Por ejemplo:

Pre-procesamiento	Representación	Descubrimiento
Categorización	Vector de términos representativos	Relaciones entre los términos
Análisis de Textos completos	Secuencias de palabras	Patrones del lenguaje
Extracción de información	Tabla de base de datos	Relaciones entre entidades

Formas Intermedias

- **Algunos ejemplos de formas intermedias …**

- ❑ Términos
- ❑ Conceptos
- ❑ Bolsas de palabras
- ❑ Taxonomía de términos
- ❑ Frases de texto multi-términos
- ❑ Términos de consultas (Consultas enriquecidas)
- ❑ Grafo directo acíclico
- ❑ Documentos prototipo
- ❑ Gráficos semánticos

Formas Intermedias

- Palabras Clave:

  DIFUSOS, TEORIA DE CONJUNTOS DIFUSOS, SISTEMAS NEURO-DIFUSOS, MODELADO DIFUSO, SISTEMAS DIFUSOS, IMPLEMENTACION SISTEMAS DIFUSOS, CIRCUITOS INTEGRADOS PARA EL CONTROL DIFUSO Y NEURO-DIFUSO, COMPORTAMIENTOS DIFUSOS, CONJUNTOS DIFUSOS, NÚMEROS DIFUSOS, PUNTO DIFUSO, METODOS DE RAZONAMIENTO DIFUSOS, CONTROLADOR DIFUSO, OBJETOS DIFUSOS, SISTEMAS DIFUSOS EVOLUTIVOS.

- Términos:

  controlador, circuits, fuzzy, numbers, difuso, sistemas, …

- Conceptos:

  fuzzy, fuzzy-logic, controladores difusos, sistema difusos, ...

Técnicas de pre-procesamiento

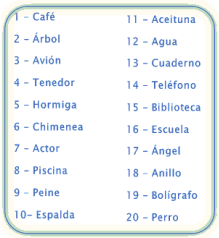
- Análisis de Texto completo
- Categorización
- Técnicas de PLN (Procesamiento de Lenguaje Natural)
  - Etiquetado de parte del discurso
  - Tokenización
  - Lematización
- Técnicas de EI
  - Categorización
  - Adquisición de patrones léxicos-sintácticos
  - Extracción automática de términos
  - Localización de trozos de texto
- Técnicas de RI
  - Indexación

Fases del pre-procesamiento

Recolección de documentos

Estandarización de documentos

Tokenización



Lematización

Lema	Categoría
sal	nombre
salir	verbo

Ejemplo:
http://www.molinolabs.com/lematizador.html

¿stemming?

Diccionario de la colección

Stemming

- Análisis morfológico (inflectional stemming): Eliminar diferentes formas de la misma palabra.
- Por ejemplo:

  - Plural y Singular: libro/libros

  - Tiempos verbales: ayudo, ayudaré, ayudamos, …
- En otros idiomas no es tan sencillo debido a:

  - Irregularidades en los verbos: seek/sought (buscar/buscado en inglés), ageben/agegeben (declarar/declarado en alemán)

Stemming

Palabra	Raíz
abaco	abac
abajo	abaj
abandera	abander
abandona	abandon
abandonada	abandon
abandonadas	abandon
abandonado	abandon
abandonados	abandon
abandonamos	abandon
abandonan	abandon
abandonar	abandon
abandonarlo	abandon
abandonaron	abandon
abandono	abandon

Ejemplo Algoritmo Porter en inglés online: http://9ol.es/porter_js_demo.html

Diccionario

Documentos/ Términos	Término 1	Término 2	Término 3	…	Término M
Documento 1	p_{11}	p_{12}	p_{13}	…	p_{1m}
Documento 2	p_{21}	p_{22}	p_{23}	…	p_{2m}
Documento 3	p_{31}	p_{32}	p_{33}	…	p_{3m}
…					
Documento N	p_{n1}	p_{n2}	p_{n3}	…	p_{nm}

p_{ij} es el peso del término en el documento y puede estar basado en:

- Esquemas binarios de presencia/ausencia
- Frecuencias normalizadas
- tf-idf (frecuencia del término/frecuencia inversa del documento)

$$tfidf(t_i) = tf(t_i) * idf(t_i)$$

$$idf(t_i) = \log(\frac{N}{df(t_i)})$$

Diccionario

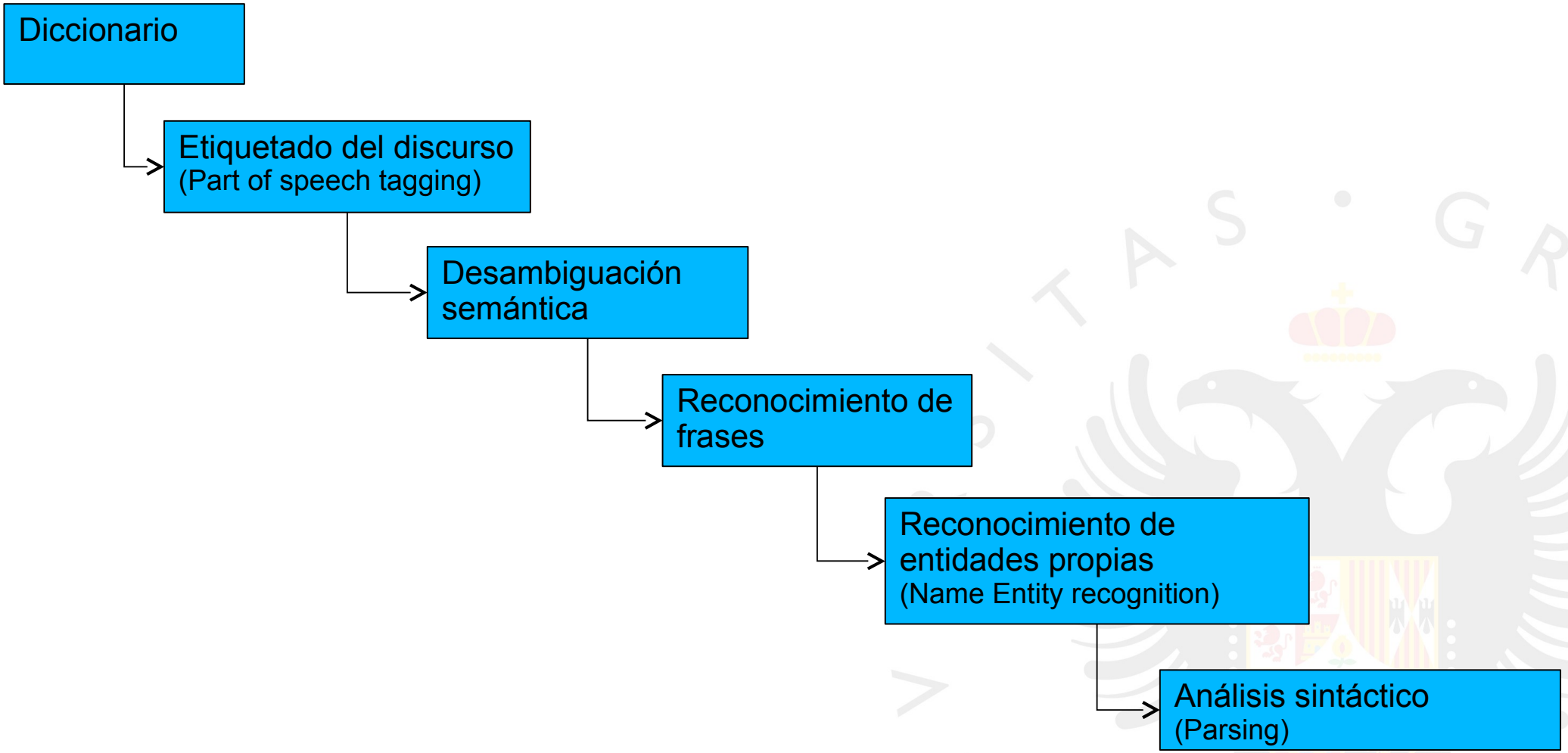
Técnicas de reducción de diccionario:

- Diccionario local
- Palabras de parada (stopwords)
- Palabras frecuentes
- Selección de características (hacer diccionarios locales a una categoría)
- Reducción de tokenes: lematización, sinónimos.

A tener en cuenta: los diccionarios reducidos basados en este tipo de técnicas mejoran las técnicas y procesos que los utilizan.

También se pueden considerar conjuntos de términos multi-palabra (n-gramas).

Fases de pre-preprocesamiento (II)



Part of Speech Tagging (POS)

Ejemplo:

Esto es un ejemplo para probar

Esto: Artículo
Es: Verbo
Un: Artículo
Ejemplo: Sustantivo
Para: Preposición
Probar: Verbo

Desambiguación semántica

- Basada en conocimiento:
  - Diccionarios
  - Tesauros
  - Ontologías
- Basados en corpus:
  - Etiquetados
  - Gran cantidad de ejemplos

Fuente: Marco, A. M. (2004). Universidad Politécnica de Valencia Desambiguación en procesamiento del lenguaje natural mediante técnicas de aprendizaje automático.

Name Entity Recognition

Encontrar y clasificar nombres en un texto:

La Guardia Civil ha detenido en Diezma a Juan López y María Clavero de Bollullos, con antecedentes policiales en 2017, como presuntos autores de cuatro robos en vehículos

Persona: Juan López, María Clavero

Fecha: 2017

Localización: Diezma, Bollullos

Organización: Guardia Civil

Preprocesamiento - Forma Intermedia

PREPROCESAMIENTO	FORMA INTERMEDIA	DESCUBRIMIENTO
Técnicas de Procesamiento de LN	Bolsa de palabras	Reglas de Asociación
Taxonomía de términos	Términos	Generalización de Reglas de Asociación
Episodios	Episodios	Reglas de Episodios
Taxonomía de términos	Jerarquías de Conceptos	Reglas de Asociación
- Tokenización - Etiquetado de parte del discurso - Lemmatizaciones	Direct Acydic Graph	
	Conceptos o términos	Reglas de Asociación cualitativas
	Representación Basada en Modelos de Recuperación de Información	- Clustering / Visualización de Documentos - Reglas de Asociación - Modelado predictivo (Modelos de Clasificación)
Indexación	Estructura Indexada (conjunto de palabras clave)	Reglas de Asociación
- Etiquetado de parte del discurso - Extracción de términos	Documentos Prototípicos	Clustering de Términos

Preprocesamiento – Ejemplo

Se desea realizar data warehousing en una base de datos médica

¿Cómo procesar campos de texto en una base de datos relacional?

Texto original ⇨ Forma intermedia de representación

Preprocesamiento – Ejemplo

Análisis de registros multifrase (identificar separadores)

Limpiar registros de palabras vacías

Substituir acrónimos

Construir diccionario de datos

Calcular frecuencias de términos

Calcular itemsets frecuentes

Obtener forma intermedia

UNIVERSIDAD DE GRANADA

decsai.ugr.es

Proceso de Minería

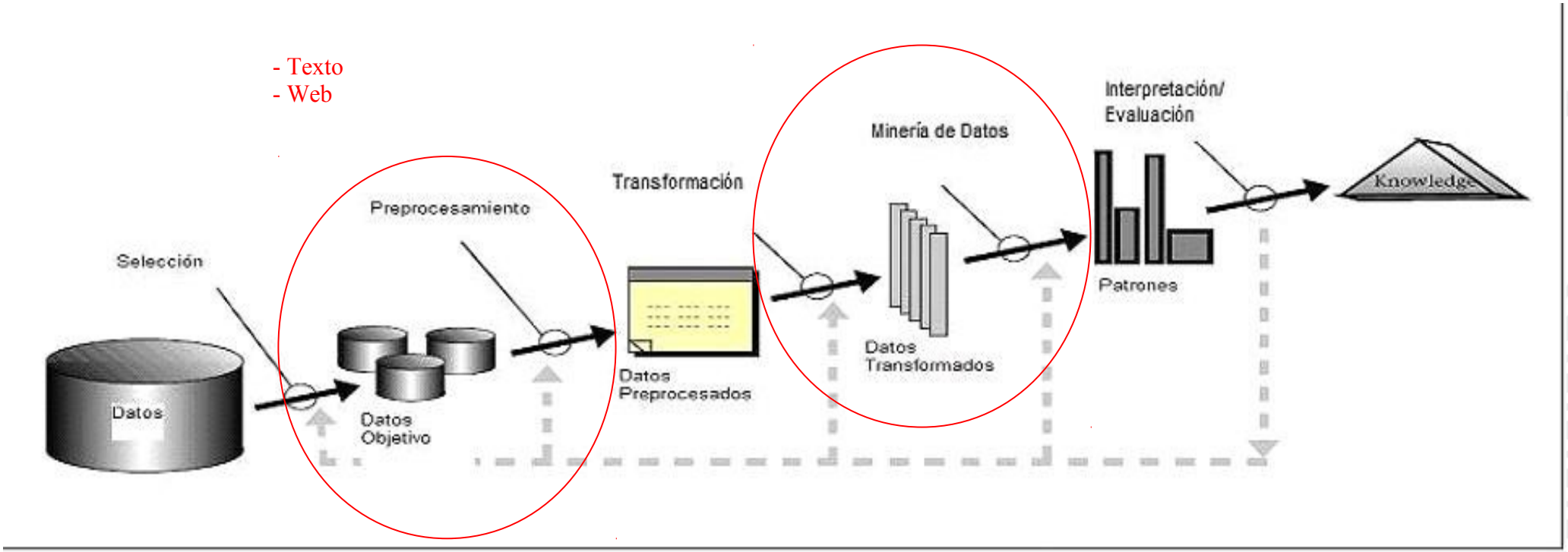
DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

KDT

Minería de Texto y
Minería Web

- Texto
- Web



Minería de textos

- Minería de texto descriptiva
  - No tenemos una clasificación a priori

- Minería de texto predictiva
  - Tenemos una clasificación a priori
  - Más intuitiva que la minería de datos descriptiva porque entendemos las palabras

Técnicas de Minería de Texto

Minería Descriptiva

- Clustering de documentos (Clustering difuso)
- Clustering Conceptual
- Reglas de Asociación (Reglas de asociación difusas)

Minería Predictiva

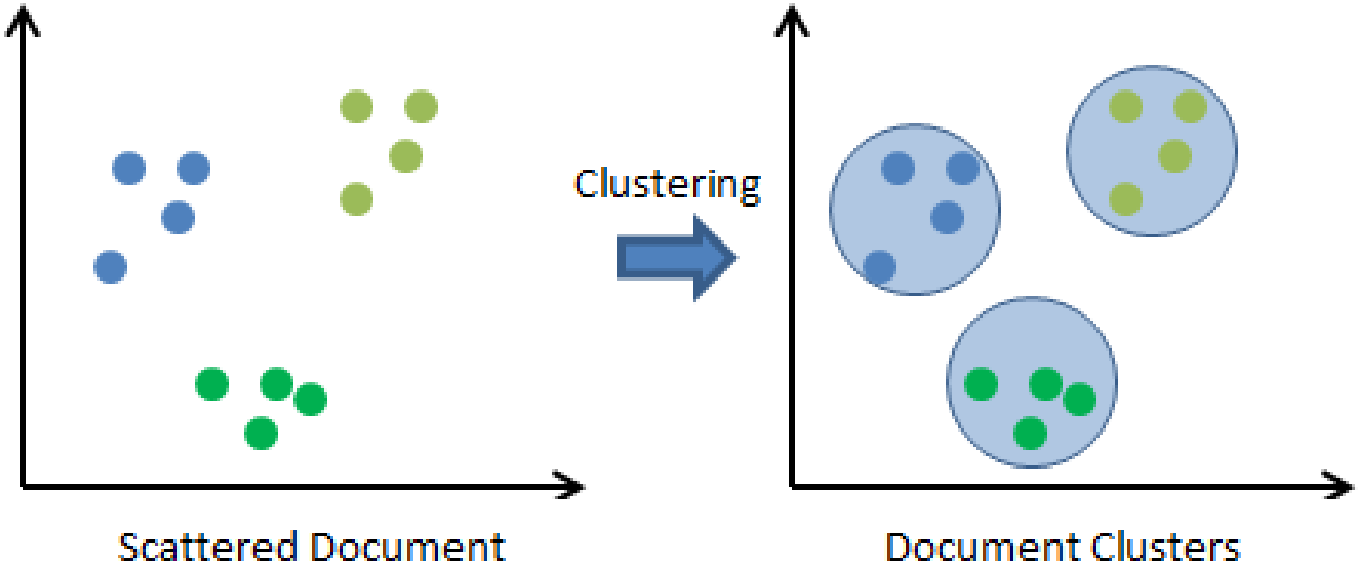
- El Vecino más cercano (Nearest-Neighbour)
- Reglas de decisión

Clustering

- Jerárquico (Aglomerativo (HAC) - Divisivo)
  - Clustering conceptual

- Particional
  - k-means (media)
  - k-medoids (mediana)
  - k-modes (moda) Apropiado para atributos categóricos

Clustering de documentos



Medidas de distancia entre documentos

Medidas clásicas de distancia

- Distancia Euclídea
- Distancia Manhattan
- Distancia Minkowski
- Correlación

Medidas de similitud entre documentos

- Coeficiente de Jaccard
- Medida del coseno

Hierarchical Agglomerative Clustering (HAC)

- No es necesario conocer a priori el número de clústeres
- HAC crea una jerarquía en forma de árbol binario (dendrograma)
- HAC asume una medida de similitud para determinar la similitud de dos clústeres
- BIRCH (1996), ROCK (1999), CHAMELEON (1999)

Hierarchical Agglomerative Clustering (HAC)

Algoritmo de HAC

1.- Cada documento se asigna a un clúster separado.

2.- Repetidamente, unir los dos clústers más similares

3.- Hasta que haya sólo un clúster

4.- Representar la jerarquía en un árbol binario (dendrograma)

Medidas de similitud entre clústeres

HAC

Single-link: Similitud máxima entre cualquiera dos documentos de los clústeres comparados

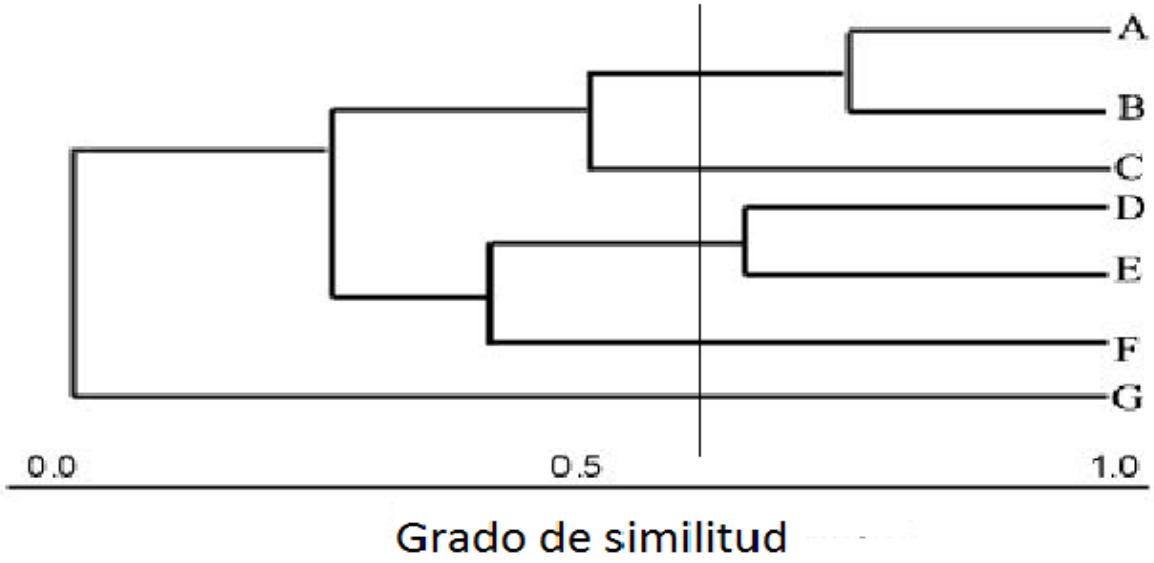
Complete-link: Similitud mínima entre cualquiera dos documentos de los clústeres comparados

Centroide: "intersimilitud media", es decir, media de la similitud entre todos los pares de documentos de cada uno de los clústeres comparados. Esto es equivalente a la similitud de los centroides.

Group-average: "Intrasimilitud media", es decir, la similitud media de todos los pares de documentos, que formarán el nuevo clúster tras la unión.

Hierarchical Agglomerative Clustering (HAC)

- El dendograma se lee del nivel más dividido hacia el menos.
- La línea transversal del dendrograma nos indica la similitud para ese nivel.
- Podemos cortar el dendrograma en un punto de corte para conseguir un clustering plano.



Clustering Particional: k-means

- Es necesario conocer a priori el número de clústeres (k)
- Para poder escoger el número de clústeres se puede minimizar el error cuadrático entre el documento y la media del clúster para todos los documentos de cada clúster.
- Se parte de un único clúster que se va dividiendo en diferentes clústeres en base a la similitud entre documentos

Clustering Particional: k-means

Algoritmo k-means

1.- Distribuir todos los documentos en k clústeres

2.- Calcular el vector medio para cada clúster

3.- Comparar el vector de cada documento con el vector medio de cada clúster y encontrar el más similar

4.- Mover todos los documentos a los clústeres con vectores más similares.

5.- Si ningún documento se ha movido a un nuevo clúster, entonces parar;

Si no, ir a paso 2.

Clustering difuso

Motivación

- El clústering jerárquico y el k-medias genera una partición donde cada documento puede pertenecer sólo a un clúster.
- El clustering difuso permite que los documentos se asignen a más de un clúster.
- Cada documento tiene un grado de pertenencia a cada clúster

Clustering difuso

Fuzzy C-Means

nd - número de documentos

nc - número de clústeres

di - es el documento *i*

uij - es el grado de pertenencia del documento *i* en el clúster *j*

ce_j - es el centroide del clúster *j*

Clustering difuso

Fuzzy C-Means (FCM)

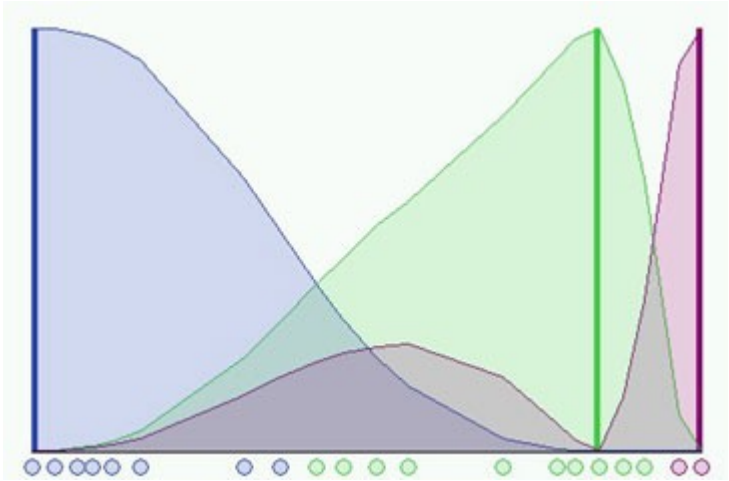
1.- Se calculan los u_{ij} aleatoriamente para cada documento a cada cluster.

2.- Calcular el centroide ce_i para cada cluster i

3.- Para cada iteración, minimizar la función J:

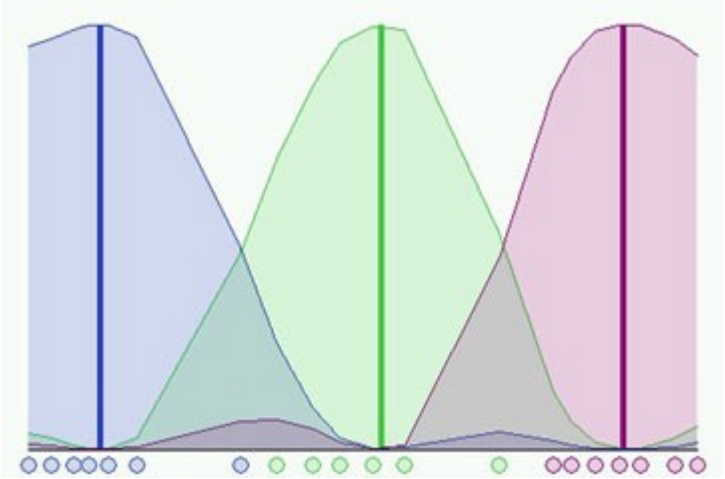
$$J = \sum_{i=1}^{nd} \sum_{j=1}^{nc} u_{ij} \|d_i - ce_i\|^2$$

Clustering difuso

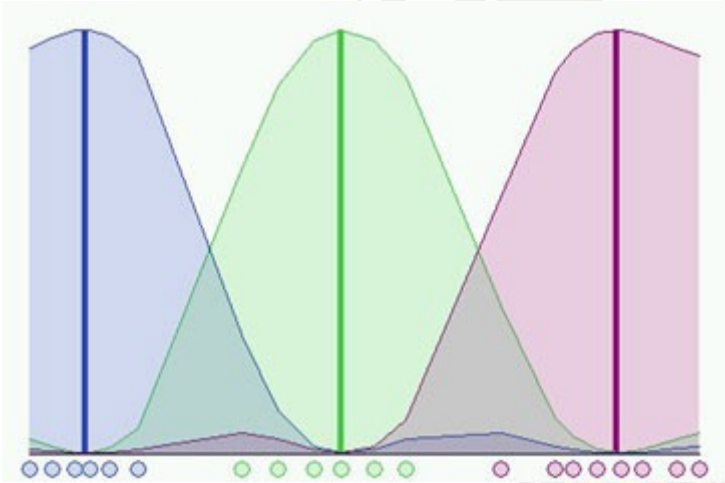


Asignación Aleatoria de Centroides

3 Clústeres



Parada = 0.3; Iteraciones 8



Parada = 0.01; Iteraciones 37

https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html

Clustering difuso

Fuzzy C-Means (FCM)

PROS

- El coeficiente de fuzziness mide cuánto se pueden solapar entre sí los clústeres
- Esto hace que el FCM sea más rápido ya que en base a este coeficiente podemos establecer el solapamiento

Clustering difuso

Fuzzy C-Means (FCM)

CONTRAS

- Hay que saber el número de clústeres inicial
- La bondad depende de la inicialización de los clústeres.
- Hay que establecer un punto de corte para la función de pertenencia
- El FCM no es un algoritmo determinístico

Clustering conceptual

Motivación

- El clustering k-medias encuentra problemas cuando los atributos no son numéricos, debido al cálculo de distancia entre los elementos
- El clustering conceptual (Michalski, 1983) se basa en un clustering 'cualitativo' frente a otro 'cuantitativo', formando los conceptos como agrupación de elementos con atributos similares.

Clustering conceptual

- Encuentra descripciones de características para cada concepto (clase)
- Produce un esquema de clasificación para un conjunto de objetos sin etiquetas
- COBWEB (1987), CLASSIT, AUTOCLASS (1996)
  - Clustering jerárquico en forma de árbol de clasificación
  - Cada nodo se refiere a un concepto y contiene una descripción probabilística para ese concepto
- No muy recomendable para conjuntos grandes de datos.

Bondad del clustering

- La bondad de un método de clúster depende tanto de la medida de similitud como del método y de su implementación
- La bondad de un método de clúster se puede evaluar en base a:
  - La similitud intra-clases sea alta
  - La similitud inter-clases sea baja

Evaluación del clustering

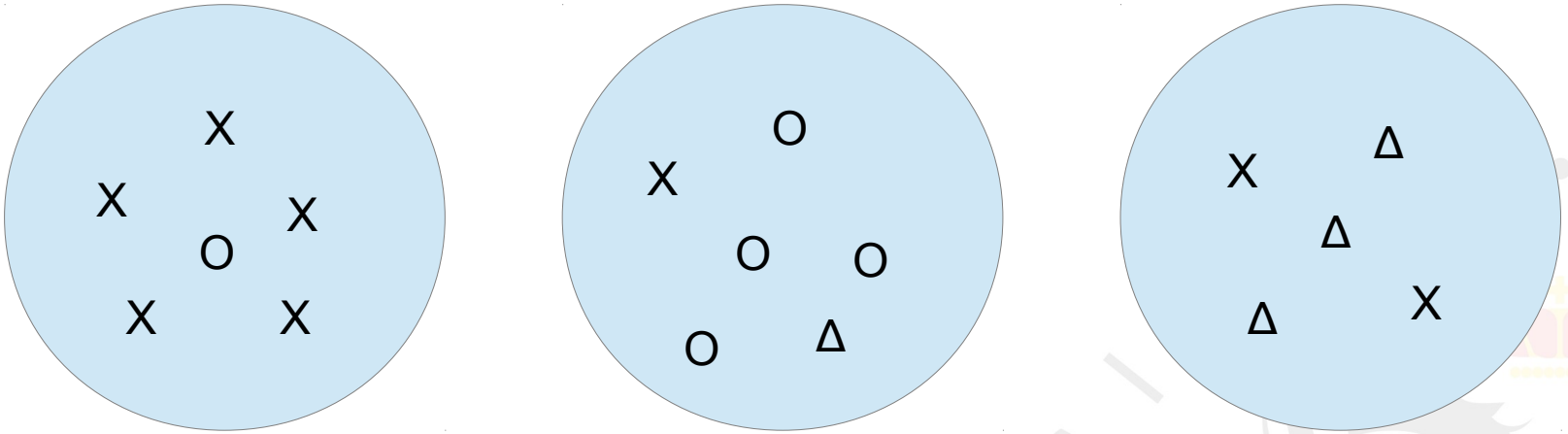
Purity

Para calcular la pureza:

- Para cada clúster se selecciona la clase más frecuente en el clúster.
- Se calcula cuántos de los elementos del clúster pertenecen a esa clase.
- La precisión de esta asignación se mide contando el número de documentos asignados a la clase mayoritaria de cada uno de los clústeres correspondientes y dividiendo por el número total de documentos.
- El clustering perfecto tiene una pureza de 1 y el peor tiene una pureza de 0.

Evaluación del clustering

Purity



N=17

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

purity = (1/17) x (5 + 4 + 3) = 0,71

Concepto de Regla Asociación

I : conjunto de items

T : conjunto de transacciones que contienen items de *I*

$I_1 \Rightarrow I_2$ una regla que significa que la aparición de I_1 en *T* implica la aparición de I_2 en *T*

$$I_1, I_2 \subseteq I \qquad I_1 \cap I_2 = \phi$$

Concepto de Regla de Asociación

Soporte: Porcentaje de transacciones conteniendo un itemset

$$Supp(I_1 \Rightarrow I_2) = supp(I_1 \cup I_2)$$

Confianza: Mide la fuerza de la regla

$$Conf(I_1 \Rightarrow I_2) = \frac{supp(I_1 \cup I_2)}{supp(I_1)}$$

Concepto de Regla de Asociación

Algoritmo A Priori

Encontrar el conjunto de itemsets con soporte por encima de *minsupp* (itemsets frecuentes)

Generar las reglas, descartando aquellas por debajo del umbral *minconf*

Transacciones Difusas

I : conjunto de items $I_o \subseteq I$

$\tilde{\tau}$: una transacción difusa (conjunto difuso no vacío) $\tilde{\tau} \subseteq I$

$\tilde{\tau}(I_0)$: grado de pertenencia de I_0 a $\tilde{\tau}$

FT-set: Conjunto de transacciones difusas con pares $(\tilde{\tau}_j, I_0)$ donde:

$$\tilde{\tau}(I_0) = \min_{i \in I_0} \tilde{\tau}(i)$$

Transacciones de Texto

$D = \{d_1, \ldots, d_n\}$: colección de documentos

$I = \{t_1, \ldots, t_m\}$: conjunto de términos asociados con pesos

Transacción de texto $\Leftrightarrow d_i \Leftrightarrow \quad \tau_i \in T$

$$W = \{w_1, \ldots, w_m\},\ w_i \in \{0,1\},\ i = 1,\ldots,m$$

$$T = \{d_1, \ldots, d_n\}$$

Transacciones de texto difusas

$D = \{d_1, \ldots, d_n\}$: colección de documentos

$I = \{t_1, \ldots, t_m\}$: conjunto de términos con pesos asociados

$W = \{w_1, \ldots, w_m\}$ se calcula mediante *tf-idf* normalizado o frecuencia normalizada

Transacción de Texto Difusa $\Leftrightarrow d_i \Leftrightarrow \tilde{\tau}_i \in FT$

$$FT = \{d_1, \ldots, d_n\}$$

Minería predictiva

- Se intentan predecir los valores de una o varias variables a partir de un conjunto de datos.
- Los datos deben estar etiquetados a priori (a qué clase pertenecen)
- Clasificación de términos y documentos (supervisado)

El Vecino más cercano (Nearest-Neighbour)

Algoritmo

1.- Calcular la similitud del nuevo documento con todos los documentos en la colección

2.- Seleccionar los *k* documentos que son más similares al nuevo documento

3.- La salida es la etiqueta más frecuente en los *k* documentos seleccionados.

Ejemplo: Búsqueda de documentos en la web

Reglas de decisión

- Una vez clasificados un conjunto de documentos, ¿cuáles son las reglas que nos permiten obtener esa clasificación?
- Aquellos patrones que permiten obtener los ejemplos positivos.

  - Ejemplos positivos: Los documentos que se deben de recuperar ante una determinada cadena de búsqueda

  - Ejemplos negativos: Los documentos que no se deben de recuperar ante dicha cadena

- Cuando un documento nuevo llega, se clasificará atendiendo a estas reglas.
- Los algoritmos de obtención de reglas de decisión suelen ser complejos y muy ineficientes.

Reglas de decisión

Algoritmo de inducción de reglas para obtener un conjunto cobertura de reglas

1.- Ir construyendo una frase F hasta que los falsos errores positivos sean 0, añadiendo palabras que minimicen el error.

2.- Guardar F como la próxima regla R. Eliminar los documentos cubiertos por F, y continuar con el paso 1 hasta que se cubran todos los documentos.

Evaluación de la minería predictiva

Precision, Recall y F-measure

$$precision = \frac{número\ de\ predicciones\ correctas\ positivas}{número\ de\ predicciones\ positivas}$$

$$recall = \frac{número\ de\ predicciones\ correctas\ positivas}{número\ de\ documentos\ positivos}$$

$$F - measure = \frac{2}{\frac{1}{precision} + 1/recall}$$

Tipos de Minería de Texto

- Descubrimiento de Tendencias
- Descubrimiento de Eventos
- Descubrimiento de Asociaciones

Tipos de Minería de Texto

Descubrimiento de Tendencias

Se estudian cambios bruscos en la frecuencia de determinados términos o frases en documentos y correos electrónicos

Fines de marketing, generalmente.

Se asocian momentos temporales a transacciones.

Varias transacciones ordenadas temporalmente se denominan *patrón secuencial*

Se incluyen restricciones temporales para determinar periodos en los que analizar el patrón.

Tipos de Minería de Texto

Ejemplo:

Conocer si en un determinado periodo de tiempo aparece el nombre de una persona en Twitter con una determinada frecuencia, indica que su popularidad es alta

Tipos de Minería de Texto

Descubrimiento de Eventos

Aplicado en general a las noticias transmitidas por canales.

Las noticias son almacenanadas por las agencias en texto plano o semiestructurado de etiquetas tal como SGML.

Se tratan de identificar eventos previamente no identificados en una colección de noticias

Se suele utilizar métodos de clustering basados en las restricciones temporales

Dichas restricciones suelen ir de una a cuatro semanas.

Tipos de Minería de Texto

Detección Retrospectiva:

Se realiza sobre colecciones acumuladas de noticias almacenadas en una lista

Clustering incremental

Basado en el modelo de espacio de vectores

Algoritmo

1. Se ordenan las noticias cronológicamente en una lista
2. Se saca la noticia más reciente de la lista y se calcula la similitud mediante el coseno con todos vectores prototípicos de los clústeres
3. Si el grado de similitud mayor supera un umbral, entonces la noticia se añade al clúster correspondiente y el vector prototípico se actualiza

   En otro caso, la noticia forma un nuevo clúster

1. Repetir los pasos 2-3 hasta que la lista esté vacía.

Tipos de Minería de Texto

Detección On-line:

La noticia se procesa cuando llega

Los algoritmos de clustering se basan en umbrales:

Umbral de detección: Especifica el mínimo valor de similitud requerido por el sistema para estar seguro de que la noticia pertenece a un nuevo evento

Umbral de clustering: Especifica el mínimo valor requerido por el sistema para añadir la noticia como un nuevo elemento de un clúster existente.

Tamaño de la Ventana: Especifica el máximo número de clústeres (u otra medida) disponible con la que comparar la noticia actual

Tipos de Minería de Texto

Descubrimiento de Asociaciones

Resolver preguntas que implican directamente asociaciones entre términos

P.ej: Consulta: “Encuentra todas las asociaciones entre fresas y cualquier ciudad de España”

Resultado: (fresas, naranjas)⇒Almería

[Sop=0.7, Con=0.78]

(fresas, kiwies)⇒Ciudad Real

[Sop=0.2, Con=0.3]

Aplicaciones con asociaciones

	Transacciones de Texto	Reglas de Asociación
Ejemplo 1:	t_1 … t_n c_1 … c_m f_1 … f_k	$t_1 \Rightarrow t_2$ $t_1 \Rightarrow c_2$
Ejemplo 2:	c_1 … c_n	$c_1 \Rightarrow c_2$
Ejemplo 3:	t_1 … t_n c_1 … c_m	$t1 \Rightarrow t2$ $t1 \Rightarrow c3$ $c1 \Rightarrow c2$

Aplicaciones con asociaciones

Transacciones de Texto

Reglas de Asociación

Ejemplo 4:

D_1	t_1	...	t_n						
D_2				c_1	...	c_m			
D_3							r_1	...	r_f

$$\begin{cases} t1 \Rightarrow r2 \\ t1 \Rightarrow c2 \\ c1 \Rightarrow r2 \end{cases}$$

Ejemplo 5:

Colección 1	$T_1 =$	d_1	...	d_m
Colección 2	$T_2 =$	d_1	...	d_r
Colección 3	$T_3 =$	d_1	...	d_s

$$\begin{cases} d1 \Rightarrow d2 \end{cases}$$

(C. Justicia, M.J. Martín-Bautista, D. Sánchez)

Ejemplo: Minería de Texto para el refinamiento de consultas en RI

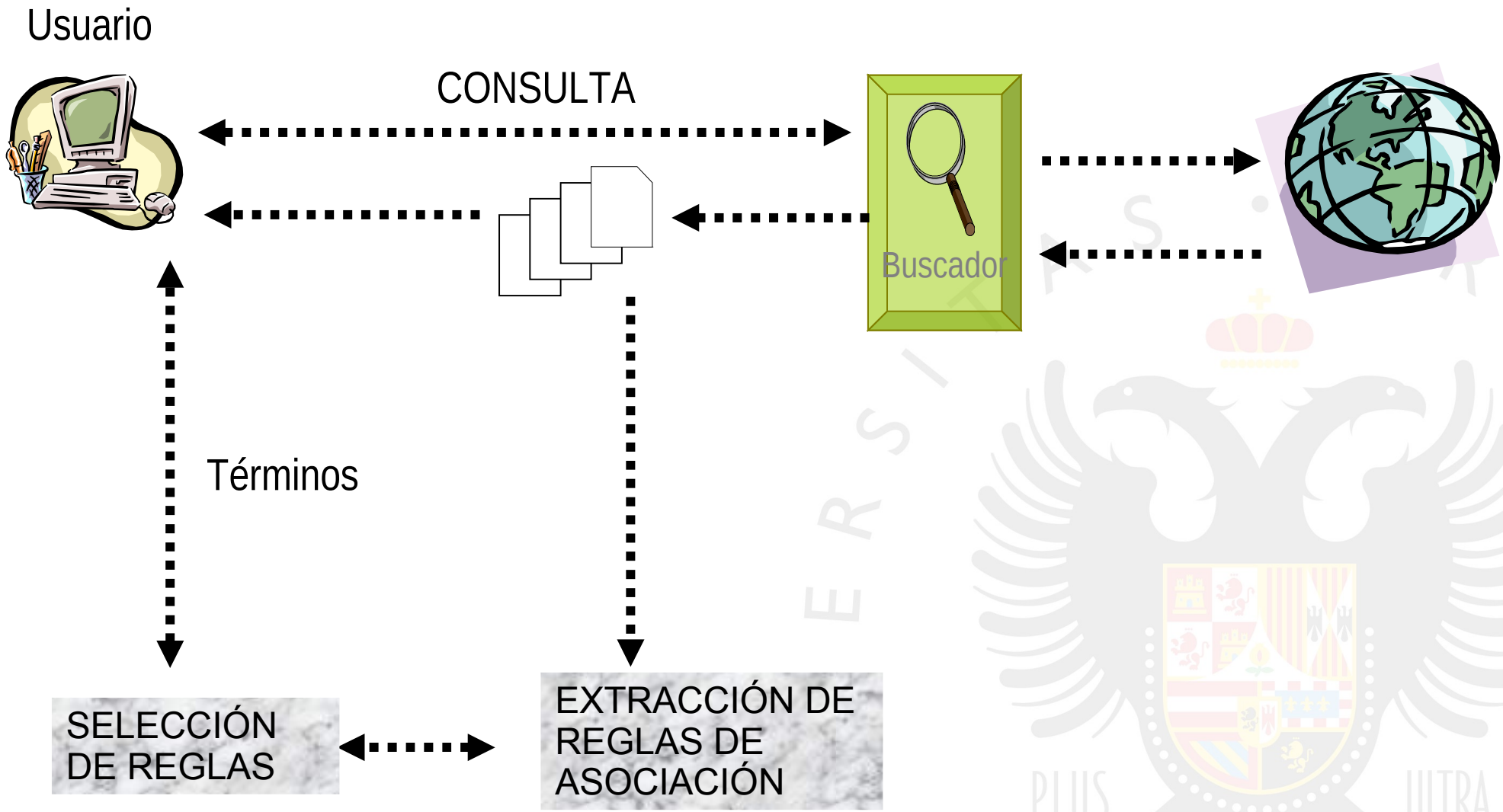
Contenidos

- Refinamiento de Consultas mediante Reglas de Asociación Difusas
- Representación del Texto
- Transacciones en Texto y Reglas de Asociación
- Refinamiento de consultas automático y semi-automático
- Selección de reglas y categorización
- Un ejemplo práctico

Recuperación de información en la web

- Problema: El usuario no encuentra la información necesaria en la web
  - Indexación desconocida
  - Falta de conocimiento sobre el vocabulario
- Solución: Refinamiento de consultas
  - Automático
  - Semi-automático

Minería de texto para refinamiento de consultas



Minería de texto para refinamiento de consultas

Ayuda a los usuarios con la construcción de consultas

Se aplican técnicas de minería mediante la extracción de reglas de asociación

Los términos en el antecedente/consecuente de la regla se pueden añadir a la consulta

El usuario puede ver la lista de términos y seleccionar los mejores (retroalimentación)

Minería de texto

¿ Información no estructurada ?

$\Downarrow$

¿Representaciones del texto?

Bolsa de palabras, índice, reglas de asociación generalizadas, frases de texto multi-término (episodios),…

Representación del texto

Términos de indexación

Esquema de frecuencias o esquema tf-idf (términos que ocurren frecuentemente en un documento pero infrecuentemente en la colección)

Obtención de términos mediante ficheros directos como en Recuperación de Información

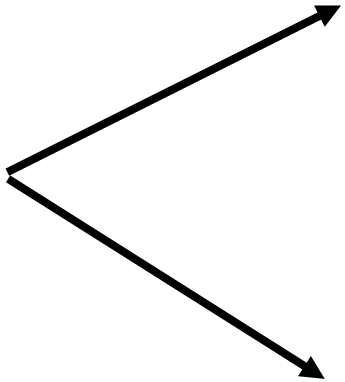
Representación del Texto

1. $D = \{d_1, \ldots d_n\}$: colección de documentos
2. Extraer todos los términos para cada $d_i \in D$
3. Eliminar palabras inservibles
4. Lematizar
5. Conjunto de términos $\{t_1, \ldots, t_k\} \in S$ y sus pesos $\{w_1, \ldots, w_k\}$ para cada documento

Procedimiento de refinamiento de consultas

$Q = \{q_1, \ldots, q_k\}$: una consulta

$P = \{p_1, \ldots, p_k\}$: pesos asociados



Refinamiento Automático de consultas

Refinamiento Semi-automático de consultas

Refinamiento automático de consultas

1. El usuario realiza una consulta inicial a un sistema de recuperación de información
2. Se recupera una lista inicial ordenada de los documentos recuperados
3. Se construyen transacciones difusas de texto y se extraen las reglas de asociación
4. Los términos en las reglas se añaden a la consulta (basándose en un proceso de especialización o generalización de consultas)
5. Se pregunta otra vez al sistema con la consulta refinada

Refinamiento semi-automático de consultas

1. El usuario realiza la consulta al sistema
2. Se recupera una lista inicial ordenada de los documentos recuperados
3. Se construyen transacciones difusas de texto y se entraen las reglas de asociación
4. Los términos de las mejores reglas se muestran al usuario, el cuál selecciona los términos más apropiados
5. El usuario pregunta de nuevo al sistema con la consulta refinada

Generalización y especialización

Generalización de una consulta:

Los términos que aparecen en el consecuente de la regla se añaden a la consulta

Especialización de una consulta:

Los términos que aparecen en el antecedente de la regla se añaden a la consulta

Selección de reglas

Criterios adicionales al soporte y la confianza o los factores de certeza

Reglas de la forma:

término→términoConsulta : Para restringir la consulta añadiendo el antecedente

término1, término2, ...→términoConsulta: Para añadir todos los términos del antecedente como término único

términoConsulta→término: Para sugerir al usuario un término que quizás es más general o se usa más en el vocabulario de indexación y puede hacer recuperar más documentos si se pregunta la web de nuevo

Categorización

Se puede contar con una categorización previa de los documentos

Las clases se añaden a las transacciones como items

Pueden aparecer reglas del tipo

$$término \rightarrow categoría$$

que indican que los documentos en los que aparece ese término pueden ser clasificado en esa categoría.

Ejemplo Experimental

http://www.alltheweb.com

Consulta en español con resultados en español:
fresas

Número de documentos recuperados: 61.000

Nos quedamos con los 100 primeros

Obtenemos 832 términos

100 transacciones con 832 items

Ejemplo Experimental

Cinco categorías:

- Clase I: Fresadoras industriales
- Clase M: Fresas como frutas
- En clase M distinguimos dos subclases:
  - Clase F: Fresas como producto para cultivar y comerciar
  - Clase C: Fresas en recetas de cocina
- Clase X: Ninguna de las clases anteriores

Ejemplo Experimental

Nivel de las reglas: 5

Umbral de soporte: 5% excepto para el esquema TFIDF, que es un 2%

Caso Crisp: 87954 reglas

Esquema difuso de frecuencia: 68 reglas

Esquema difuso TFIDF: 3686 reglas

Ejemplo Experimental

Algunas reglas obtenidas que sugieren al usuario nuevo vocabulario sobre la búsqueda

frontales→fresas (CF=1)

herramientas→fresas (CF=0.7)

Ejemplo Experimental

Reglas con categorías y factor de certeza 1:

frontales→Clase I

herramientas→Clase I

accesorios→Clase I

brocas→Clase I