Universidad de Granada

decsai.ugr.es

Modelos de ciencia de datos no numéricos:

Aplicaciones en redes sociales, web y gestión de procesos

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

Proceso de Minería

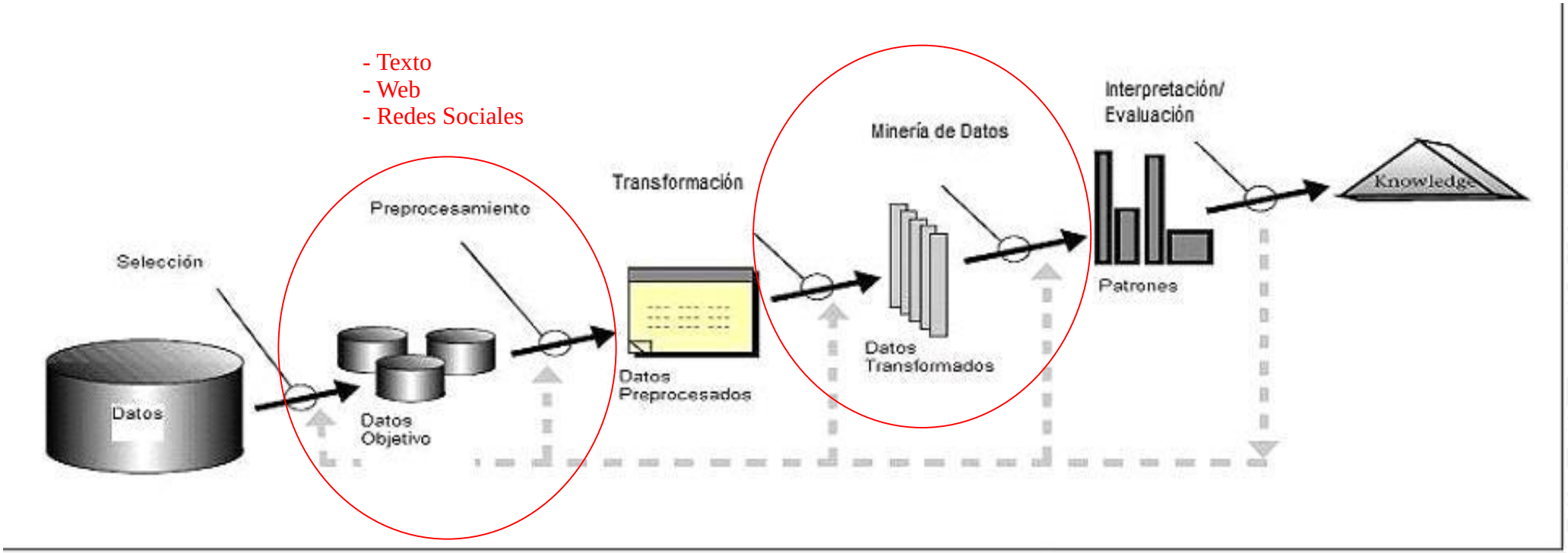
DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

KDT

Minería de Texto y
Minería Web/Redes Sociales

- Texto
- Web
- Redes Sociales



Minería de textos

- Minería de texto descriptiva
  - No tenemos una clasificación a priori. Se analizan patrones en lo que ya ha sucedido

- Minería de texto predictiva
  - Tenemos una clasificación a priori
  - Predice qué puede pasar en el futuro con una ayuda más directa a la toma de decisiones

Técnicas de Minería de Texto

Minería Descriptiva

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

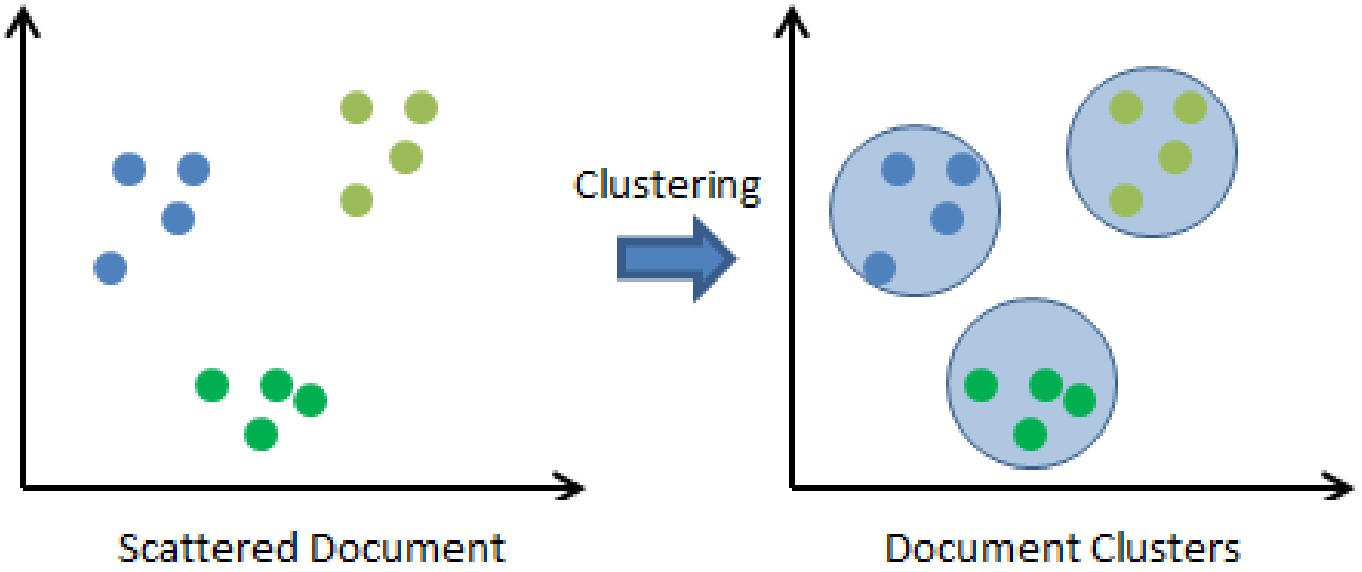
Minería Predictiva

- El Vecino más cercano (Nearest-Neighbour)
- Reglas de decision
- Deep Learning

Clustering

- Jerárquico
  - Clustering conceptual
- Particional
  - k-means

Clustering de documentos



Medidas de similitud entre clusters

Single-link: Similitud máxima entre cualquiera dos documentos

Complete-link: Similitud mínima entre cualquiera dos documentos

Centroide: "intersimilitud media", es decir, media de la similitud entre todos los pares de documentos (pero excluyendo pares de documentos en el mismo cluster). Esto es equivalente a la similitud de los centroides.

Group-average: "Intrasimilitud media", es decir, la similitud media de todos los pares de documentos, incluyendo los pertenecientes al mismo cluster.

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 Clustering

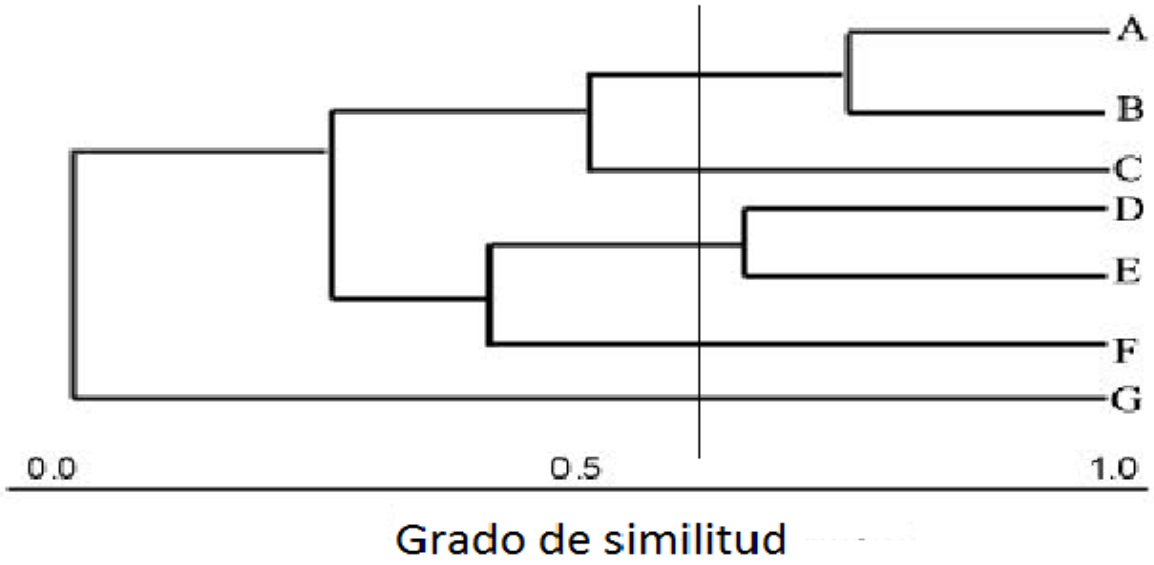
No es necesario conocer a priori el número de clusters

- **Aglomerativo:** Bottom-up. Cada elemento es inicialmente un cluster
- **Divisivo:** Top-Down. Todos los elementos están inicialmente en el mismo cluster

El clustering jerárquico crea una jerarquía en forma de árbol binario (dendograma)

Hierarchical Clustering

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



Hierarchical Agglomerative Clustering (HAC)

- Se comienza con cada caso como un cluster individual.
- En cada paso, se combina el par de clusters más cercano hasta que sólo quede uno (o k).
- HAC asume una medida de similitud para determinar la similitud de dos clusters
- AGNES (1990),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 dendograma

Hierarchical Divisive Clustering (HDC)

- Se parte de un cluster con todos los documentos.
- En cada paso, los clusters se dividen en los dos clusteres menos similares.
- Se repite hasta que los clusteres formados son distintos unos de otros.
- DIANA (1990)

Clustering Particional: k-means

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

Clustering Particional: k-means

Algoritmo k-means

1.- Distribuir todos los documentos en k clusters

2.- Calcular el vector medio para cada cluster

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

4.- Mover todos los documentos a los clusters 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 cluster.
- Cada documento tiene un grado de pertenencia a cada clúster

Clustering difuso

Fuzzy C-Means

nd número de documentos

nc número de clusters

di es el documento *i*

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

ce_j es el centroide del cluster *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*:

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

Clustering difuso

Fuzzy C-Means (FCM)

PROS

- El coeficiente de *fuzziness* mide cuánto se pueden solapar entre sí los clusters
- 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 clusters inicial
- La bondad depende de la inicialización de los clusters.
- Hay que establecer un *cutoff* 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 cluster 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:

- cada clúster se asigna a la clase más frecuente en el clúster.
- La precisión de esta asignación se mide contando el número de documentos asignados y dividiendo por el número de documentos.
- El clustering perfecto tiene una pureza de 1 y el peor tiene una pureza de 0.

Concepto de Regla Asociación

I : conjunto de items (itemset)

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 –A Priori TID, ECLAT, FP-Growth

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

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

Ejemplo de Regla de Asociación

Supongamos que estamos analizando un conjunto de artículos de investigación. Si extraemos reglas de asociación, podemos recuperar una regla que sea:

(Machine learning, neural networks -> Deep learning) CF=0.9

Esta regla indica que los documentos que nombran machine learning y neural networks también mencionan Deep Learning, reflejando así la relación que hay entre estos conceptos.

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 $\Leftrightarrow d_i \Leftrightarrow \tau_i \in T$

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

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

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 $\Leftrightarrow d_i \Leftrightarrow \tilde{\tau}_i \in FT$

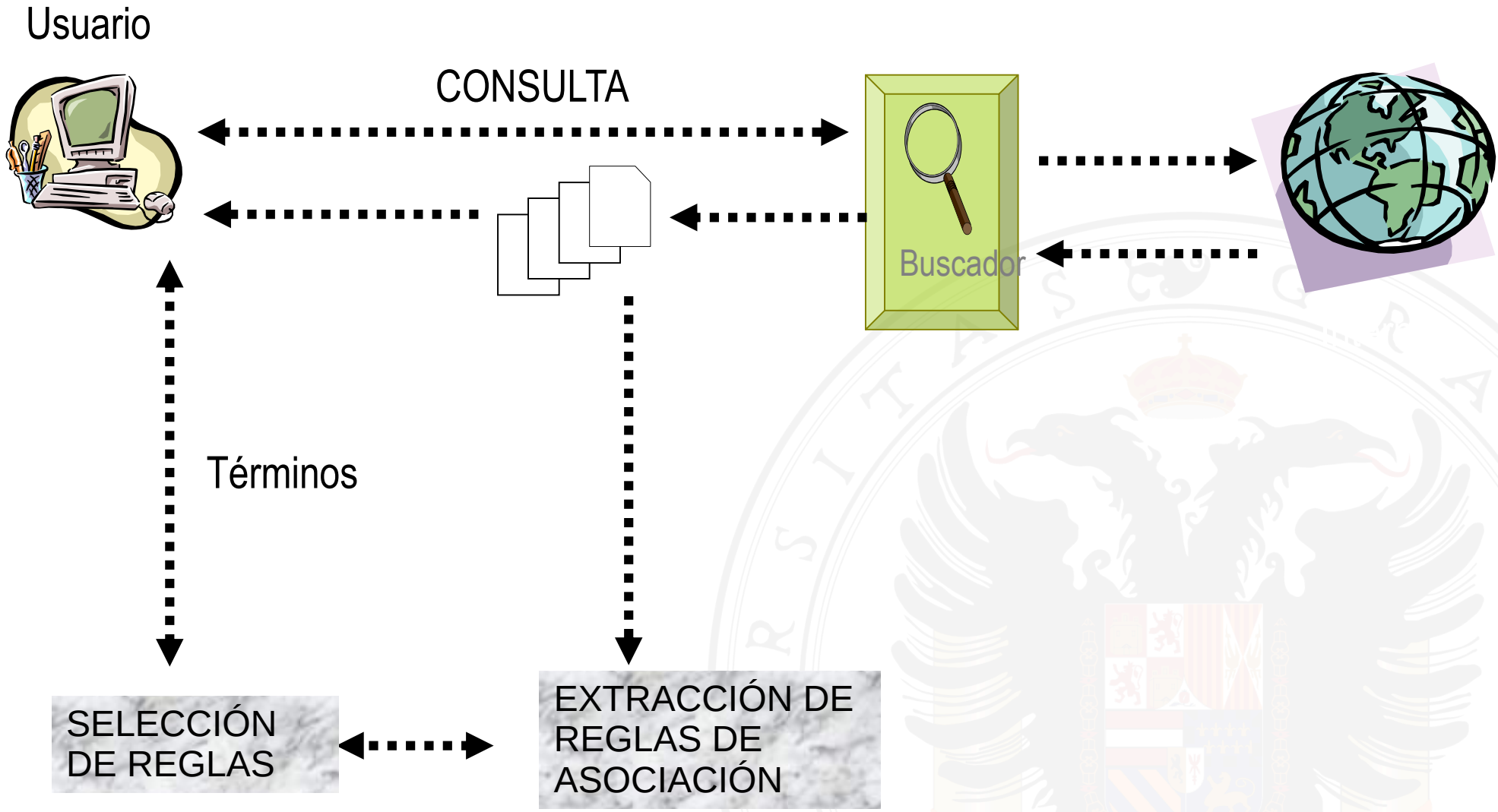
$$FT = \{d_1, \ldots, d_n\}$$

Ejemplo: Minería de Texto para el refinamiento de consultas en Recuperación de Información

Minería de texto para refinamiento de consultas

- 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)

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

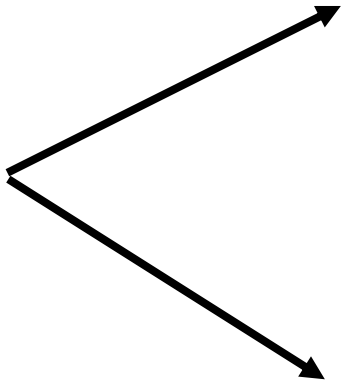
Representación del Texto

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 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 ó de forma automática, los términos en las reglas se añaden a la consulta (basándose en un proceso de especialización o generalización de consultas)
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→categoría

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

Ejemplo Experimental

Consulta a un buscador 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

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 (X) 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 almacenadas 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.
- Se pueden hacer dos tipos de detecciones: Detección restrospectiva y Detección On-line

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 clusters
3. Si el grado de similitud mayor supera un umbral, entonces la noticia se añade al cluster correspondiente y el vector prototípico se actualiza

   En otro caso, la noticia forma un nuevo cluster
4. 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 cluster existente.

Tamaño de la Ventana: Especifica el máximo número de clusters (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=0.8, Con=0.9]

(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

$$\{ \; t_1 \Rightarrow t_2 \quad t_1 \Rightarrow c_2$$

Ejemplo 2:

c_1	...	c_n

$$\{ \; c_1 \Rightarrow c_2$$

Ejemplo 3:

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

$$\{ \; t1 \Rightarrow r2 \quad t1 \Rightarrow c2 \quad c1 \Rightarrow r2$$

Ejemplo 4:

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

$$\{ \; d1 \Rightarrow d2$$

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

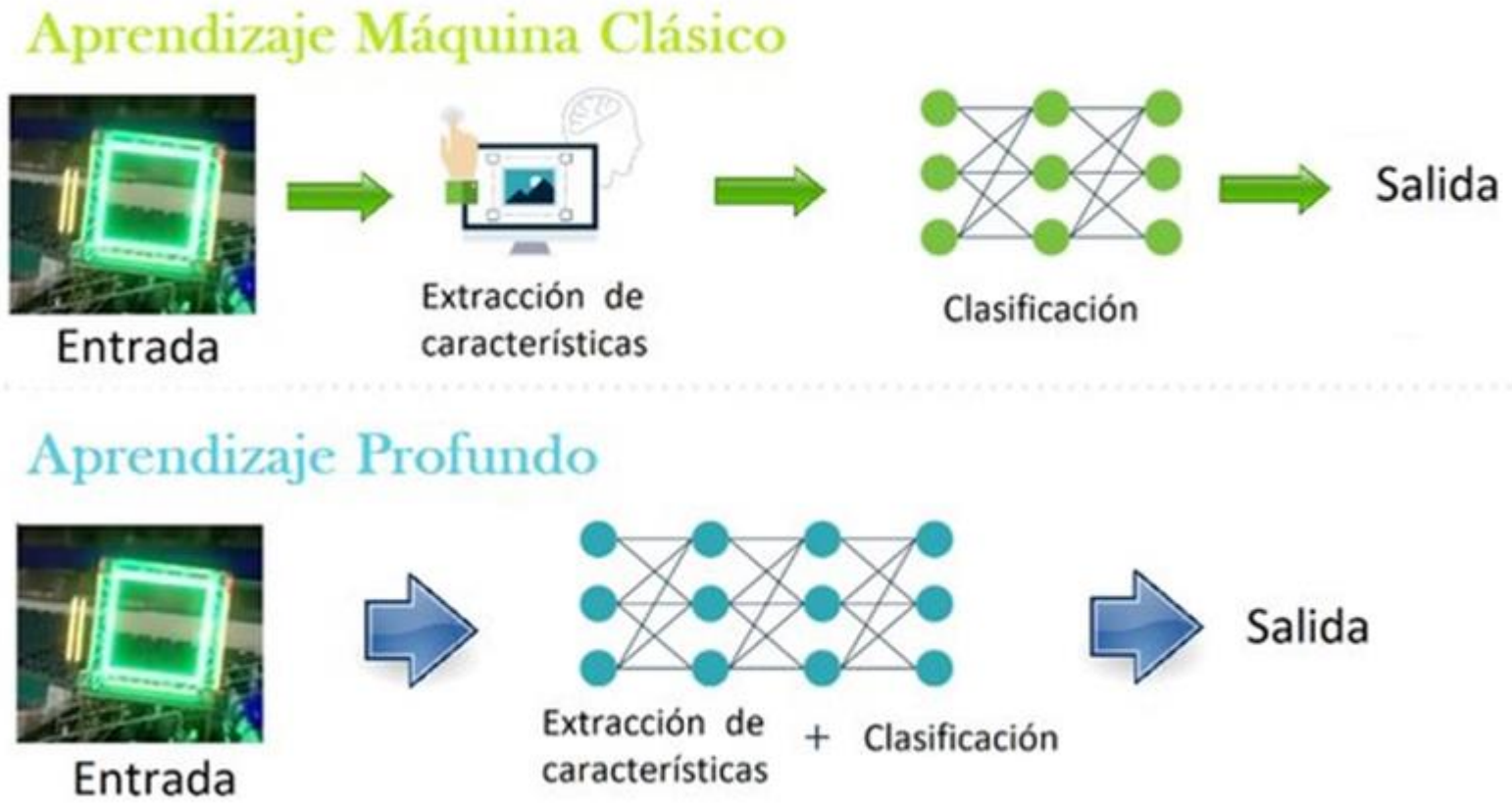
Algoritmo de inducción de reglas para obtener un conjunto cobertura de reglas

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.

Deep Learning

- Aprendizaje automático basado en redes neuronales en diversas capas de entrada, ocultas y de salida.



Fuente: DOI: 10.13140/RG.2.2.16637.31207

Deep Learning: Red neuronal convolucional para clasificación

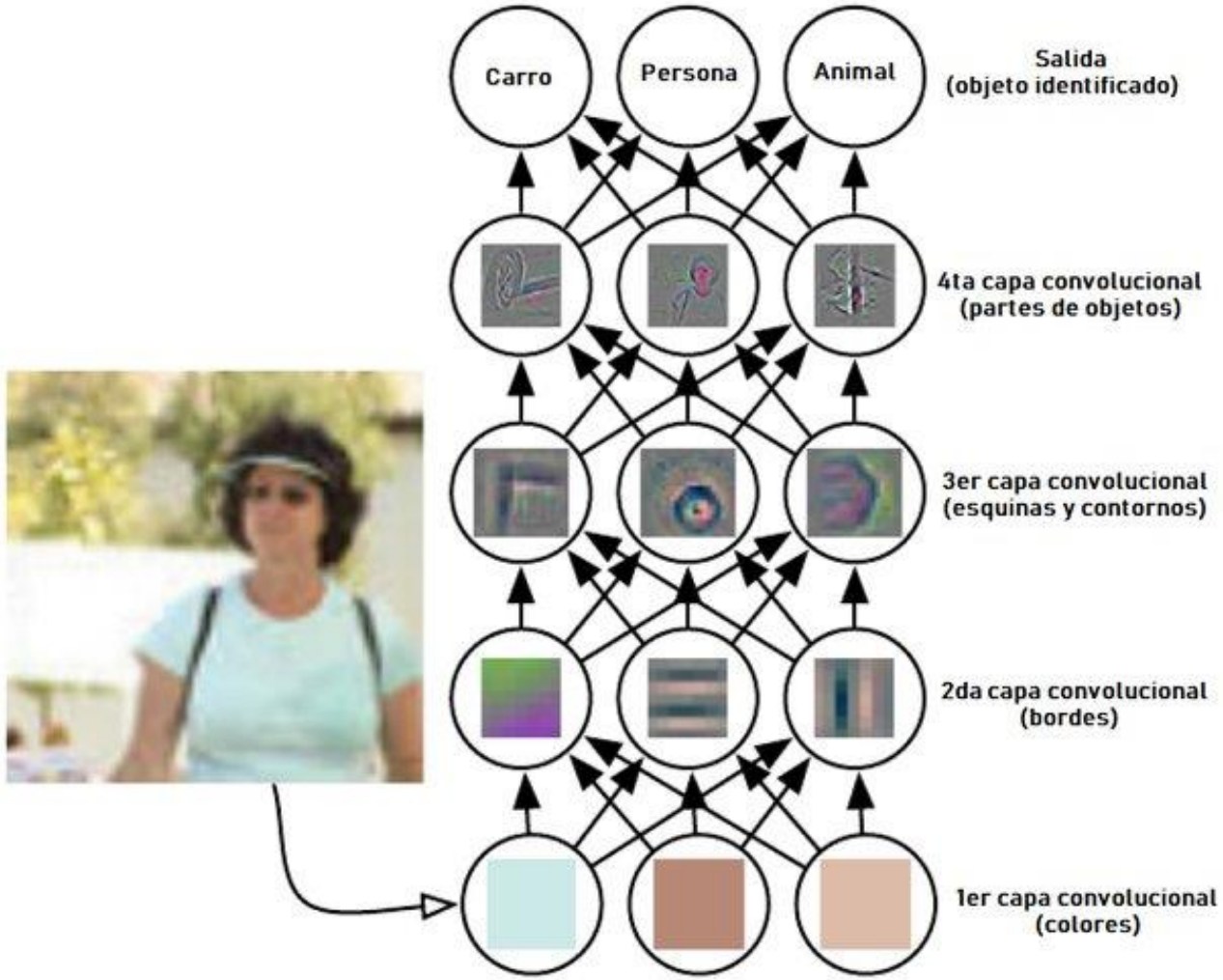
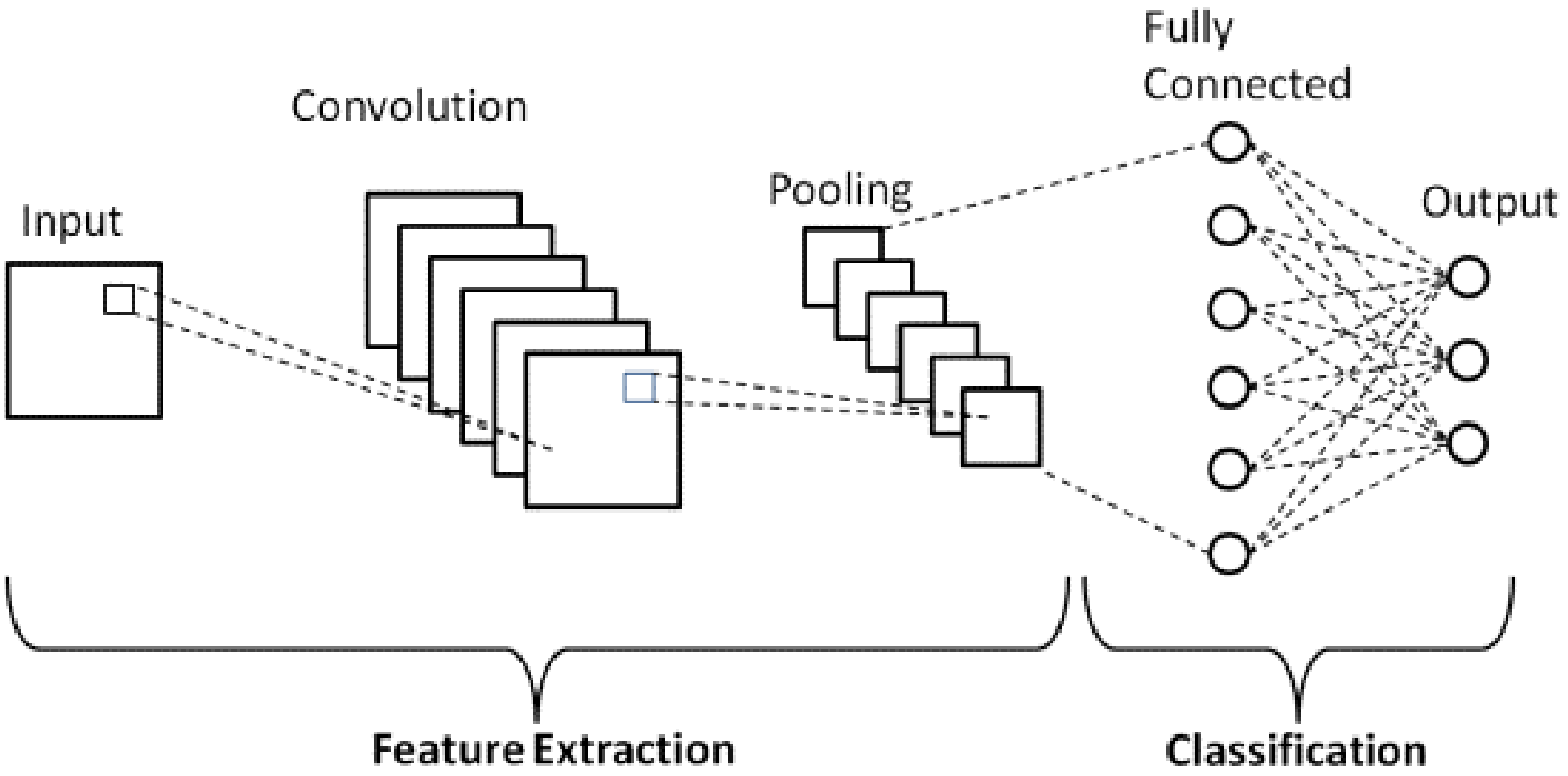


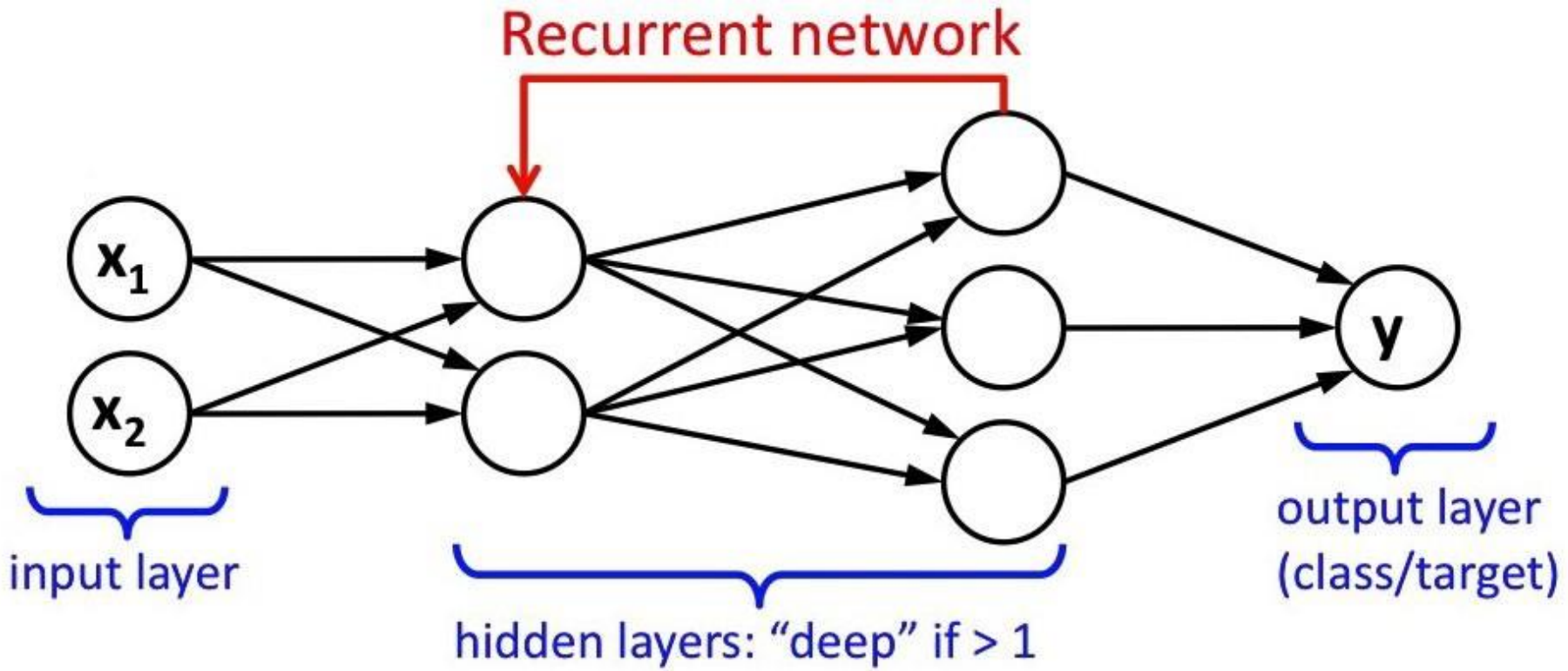
Imagen obtenida de https://www.deeplearningbook.org/

Convolutional Neural Network (CNN)



Fuente: https://www.upgrad.com/blog/basic-cnn-architecture/

Recurrent Neural Network (RNN)



Fuente: https://towardsdatascience.com/implementation-of-rnn-lstm-and-gru-a4250bf6c090

Ejemplo RNN para minería predictiva texto

Clasificación de texto para revisión de películas de IMDB

- Conjunto de datos: 50.000 reseñas de diferentes películas
- Modelo para predecir si la reseña es negativa o positiva
- Problema de clasificación binaria

Fuente: https://ichi.pro/es/clasificacion-de-texto-con-rnn-108823858105073

Ejemplo RNN para minería predictiva texto

- Las críticas de una película no son uniformes: Reseñas de 4/5 palabras hasta 17/18 palabras.
- Se pasa a una longitud uniforme mediante representación vectorial
- La posición de una palabra se aprende del texto y de las palabras que la rodean

sentence=['Fast cars are good', 'Football is a famous sport',
'Be happy Be positive']

After padding:

[[364 50 95 313 0 0 0 0 0 0]
[527 723 350 333 722 0 0 0 0 0]
 [238 216 238 775 0 0 0 0 0 0]]

Ejemplo RNN para minería predictiva texto

Etapas de la RNN:

1.- Avance hacia la capa oculta y predicción

2.- Comparación de la predicción con el valor real utilizando la función de pérdida (cuanto menor el valor, mejor será el modelo)

3.- Utiliza los valores de error en la retropropagación y calcula el gradiente para cada nodo (para ajustar los pesos de la red)

4.- Función de activación: tangente hiperbólica para mantener el valor entre -1 y 1

Evaluación de la minería predictiva

Precision, Recall y F-measure

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

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

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