

Extracción y Recuperación de Información no Supervisada

Recuperación y Acceso a la Información



24/03/2008

Pilar Sopena Suils

Índice de Contenidos

1 INTRODUCCIÓN	3
2 CLASIFICACIÓN NO SUPERVISADA.....	4
3 CARACTERÍSTICAS DE NO SUPERVISIÓN	5
4 CLUSTERING	6
5 MAPAS AUTO-ORGANIZATIVOS	7
6 BIBLIOGRAFÍA	8

1 INTRODUCCIÓN

En esta página se podrán encontrar los enlaces a conceptos relacionados con el **aprendizaje no supervisado** y su aplicación a la **extracción y recuperación de la información**.

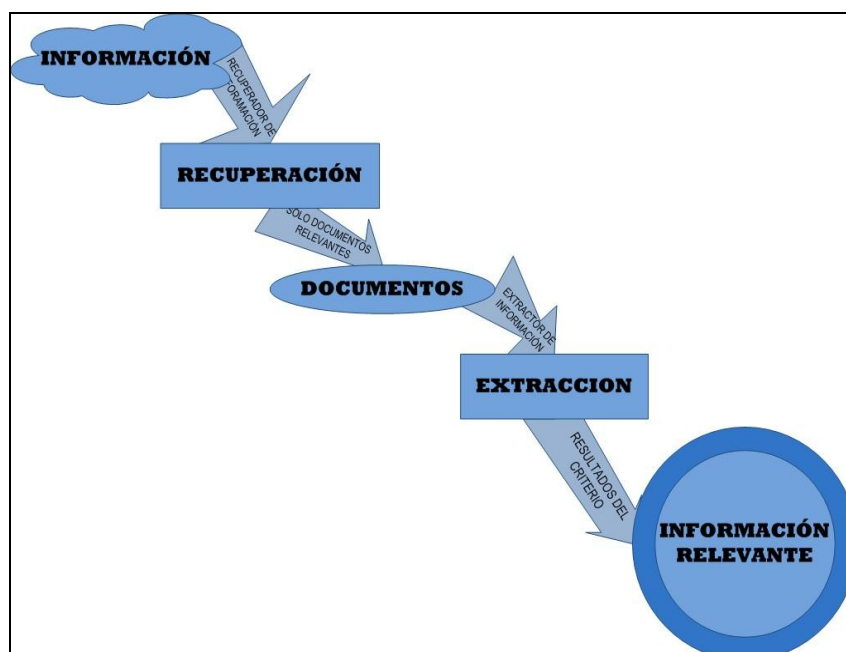
Los conceptos de la **Recuperación de la Organización** que se van a ver van a ser:

- **Extracción vs. Recuperación**
- **Aprendizaje no Supervisado y sus características**
- **Clustering**
- **Mapas de Kohonen o Auto-organizativos**

Extracción de la Información vs. Recuperación de la Información

- La **extracción de la información** (*IE en inglés Information Extraction*) es la **búsqueda** manual o automática de **palabras, párrafos ó trozos de textos** que contengan información relevante buscada y presentarla de forma estructurada.
- La **recuperación de la información** (*IR en inglés Information Retrieval*) es la búsqueda manual o automática de **qué documentos cumplen el criterio** a través de su contenido o características.

Aunque los términos parecían lo mismo vemos que no lo son, **¿pero en qué se relacionan?** Para poder realizar una **extracción** es necesaria una **previa recuperación**, he aquí la relación.



2 CLASIFICACIÓN NO SUPERVISADA

La **clasificación no supervisada** es realizar una división **sin contar con un tutor** o información de apoyo, es decir, se realiza el proceso automático extrayendo características comunes y/o relaciones existentes. Este tipo de clasificación también se le denomina **auto-adaptativa** ya que se va modificando de forma que se adecua al conjunto de datos de entrada del que parte.

Los **modos** de esta **clasificación** son búsqueda de:

- **Nivel de semejanza entre conceptos** buscando la cercanía de los mismos. Aplicada a la recuperación de información significa buscar comparación entre los conceptos para estructurar la propia información.
- **Extracción de características y sus relaciones**, indagar en las características que existen en los conceptos y sus relaciones. Dichos conceptos parecidos deberían dar los mismos resultados.
- **Analizar y extraer principales características**, escoger las peculiaridades necesarias y fundamentales que nos den la suficiente información; para la recuperación es importante ya que se restringe los datos a tener en cuenta para realizar la recuperación.
- **Modelización**, tratar de agrupar y buscar el prototipo que represente a todo un grupo.

3 CARACTERÍSTICAS DE NO SUPERVISIÓN

El **modelo básico** en el que se basa la **no supervisión**:

- **Regla de Hebb** la conexión entre dos neuronas conectadas es mayor si ambas están activadas y menor si no lo están.
- **Interacción lateral** cada neurona tiene relación excitatoria con un vecindario cercano e inhibitoria con uno lejano imitando el cerebro.

El **aprendizaje y entrenamiento** se basa en los siguientes conceptos:

- **Auto-adaptación**: a través de la aplicación de los algoritmos se van modificando sus parámetros para resolver el problema.
- **Competición**: potenciar a los ganadores en detrimento de los perdedores

Las ventajas y desventajas de la no supervisión son:

- Encontrar la **solución innovadora**.
- **No** encontrar una **buena solución**.
- **Dificultad** en la **modelización y parametrización**.

4 CLUSTERING

El **clustering o agrupamiento** es partir el grupo de entrada en **pequeños grupos**, para este particionamiento se usan **distancias**. Tiene en cuenta todas las **características de los datos**.

Las distancias más utilizadas son:

- **Euclídea**, la más común.
- **Manhattan**.
- **Hamming**.

A través del **cálculo de distancias** se **agrupan los elementos** de acuerdo a los más cercanos según el cálculo.

El **clustering** es utilizado en la **recuperación y extracción de información** para realizar agrupamientos de **conceptos cercanos** de manera que tratando el grupo el resultado sea similar. Las **aplicaciones comerciales** existentes son:

- **iResearch Report** es un sistema de extracción de información y realizador de resúmenes de documentos.
- **News Feed Researcher** extensión del anterior para las noticias y los feeds.

5 MAPAS AUTO-ORGANIZATIVOS

Los **mapas auto-organizativos** también denominados mapas de **kohonen**, son un tipo de **red neuronal** que se basa en el **aprendizaje no supervisado y la competitividad**. Las neuronas están dispuestas en una red cuadrada 2-D, también las hay de más dimensiones.

La **topología** de la red se compone de:

- **Capa de entrada** que propaga los datos de entrada a la siguiente
- **Mapa kohonen** capa de neuronas típicamente cuadrada de 2 dimensiones, entre las que compiten por obtener la salida. La neurona ganadora inhibe a las cercanas para reforzar el aprendizaje.

La mayor **aplicación** es **agrupar la información** entorno a **características desconocidas**.

6 BIBLIOGRAFÍA

La bibliografía utilizada para realizar esta página es:

- [Extracción de la Información](#)
- [Recuperación de la Información](#)
- [Aprendizaje supervisado](#)
- [Clustering](#)
- [Curso de sistemas conexionistas](#)
- [Departamento de et-evannai UC3M](#)