

**Unsupervised Extraction and
Retrieval**

Information Retrieval and Access



24/03/2008

Pilar Sopena Suils

Table of Contents

1 INTRODUCTION	3
2 UNSUPERVISED CLASIFICACION	4
3 UNSUPERVISED CHARACTERISTICS	5
4 CLUSTERING	6
5 SELF-ORGANIZED MAPS	7
6 BIBLIOGRAPHY	8

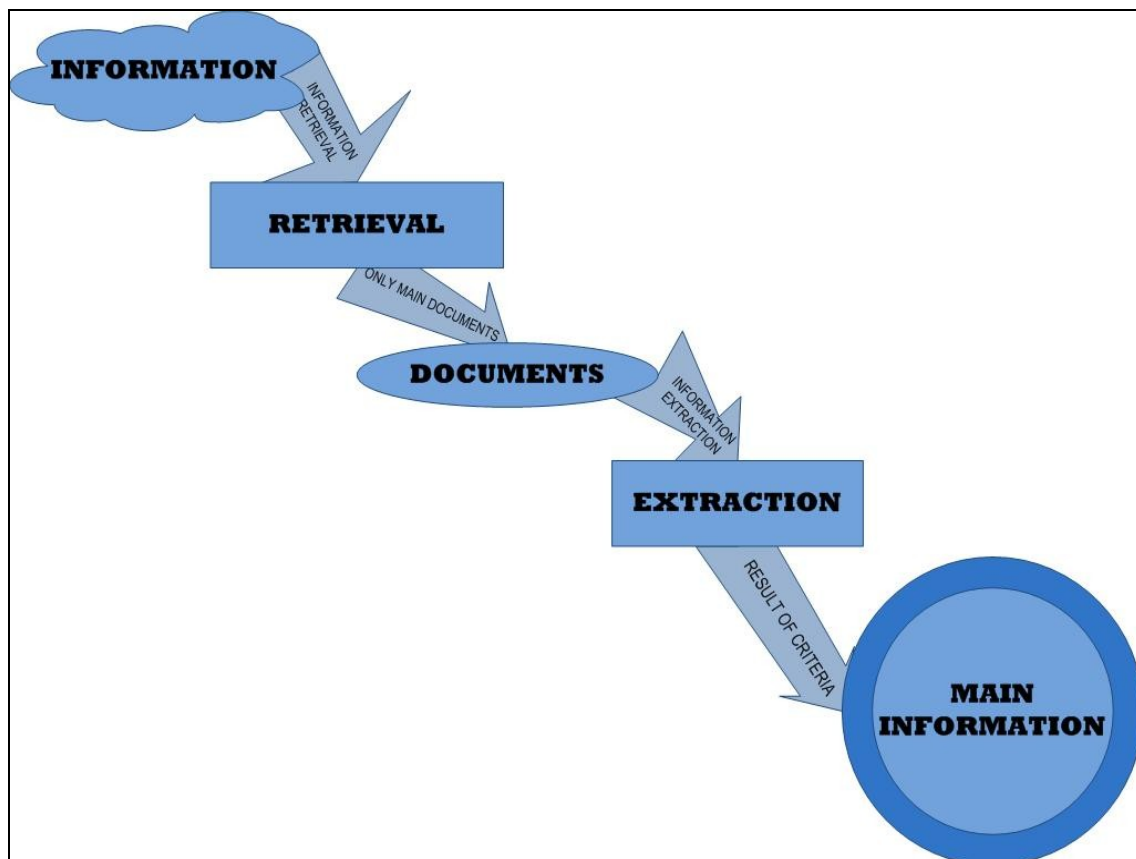
1 INTRODUCTION

In this site can find the **concepts link to unsupervised learning** and its application for **information retrieval and extraction**.

Information Extraction vs. Information Retrieval

- **Information extraction** is a, manual or automatic, **search of words, paragraphs or parts of text** contains important searched information and it is presented structured.
- **Information retrieval** is a, manual or automatic, **search what documents are related to the search** though the contents or characteristics of them.

Although the **concepts** are very similar, it **isn't the same**. What are they relate? The **information extraction needs an information retrieval**, this is the relationship.



2 UNSUPERVISED CLASSIFICATION

Unsupervised classification is a division of data **without a learning procedure**; it is an **automatic process** which extracts common features or relationships. This type of classification is **also called self-adaptive** because the division of data is modified to fit the incoming data set.

The **modification** are occurred **while** come in the data even when are a **model of the in data**.

The **modes** of search are:

- **Likeness level among concepts**, this mode looks for close concepts. In retrieval information means compare concepts and structure information.
- **Features and relationships extraction**, features of concepts and the relationships among them are sought. Similar concepts should make similar results.
- **Main feature analysis and extraction**, this method chooses singularity and fundamental features in concepts to have enough information for the retrieval, this is important because it reduces the amount of search and the volume of initial data.
- **Modelling to group and search the prototype** that symbolizes the whole group. Unsupervised classification is a division without a teacher, that is, an automatic process that extract usual characteristics or relationships. That kind of classification, also called self-adaptive because the modification are occurred while come in the data even when are a model of the in data.

3 UNSUPERVISED CHARACTERISTICS

Basic model of unsupervised learning is:

- **Hebb's rule:** The connectivity between two connected neurons is greater if both are activated and less if they are not.
- **Side interaction:** Every neuron has an excited connection with near neighbours and inhibited with the far ones, this model imitates the brain.

Learning and training is based on the next concepts:

- Self-adaptive: the algorithm gets adaptation through the modification of its parameters to solve the problem.
- Competition: Based in encouraging the winners against the losers.

Advantages and disadvantages of unsupervised learning (UL) are:

- UL may find an **innovative solution**.
- UL may **not** find a **good solution**.
- It is **difficult to create models and select the parameters**.

4 CLUSTERING

Clustering is **classification of objects into different groups**. For the classification the **distances between objects** and the all characteristics of them are used.

The **distances** more used are:

- **Euclidean**, the more usual.
- **Manhattan**.
- **Hamming**.

Objects are grouped using distance calculation, the **objects nearer of each other** belong the **same group**.

Clustering is often used in **information retrieval and information extraction** for making groups of **near concepts**, so the result of any process performed over the group should be similar to the result obtained from one member. **Existing commercial applications** are:

- **iResearch Report**. A system of information extraction and document summing-up.
- **News Feed Researcher**. An extension for the previous system for news and feeds.

5 SELF-ORGANIZED MAPS

Self-organizing maps, also called **Kohonen maps**, are a kind of neuronal network that are based on **unsupervised** learning and competition. The neurons are placed on 2-D squared net, although greater dimensions may be used.

The **topology of the net** is composed of the elements:

- **The input layer** which propagates the data to the next layer.
- **Kohonen map** neurons layer, typically a 2-D squared layer. Neurons compete for getting the best output. The winner neuron inhibits another near neurons for strengthening the learning.

The **Main application** is to **group the information of unknown characteristics**.

6 BIBLIOGRAPHY

Bibliography used for create this page:

- [Information Extraction](#)
- [Information Retrieval](#)
- [Unsupervised Learning](#)
- [Clustering](#)
- [Course of connectional systems](#)