

# Chapitre 1

## Etat de l'art

Le clustering est le regroupement de plusieurs données similaires en un seul groupe appelé cluster. L'analyse de cluster permet ainsi d'identifier des groupes de données relativement homogènes sur la base de leur similarité pour des caractéristiques données ce qui dans notre cas peut par exemple être le type d'emploi occuper en fonction des filières suivies par d'anciens étudiants. Cependant analyser différents profils d'individus peut représenter des difficultés techniques importantes, c'est pourquoi cette première section du document présentera les différentes solutions possibles pouvant apporter une réponse à cette difficulté de catégorisation des parcours.

### 1.1 Les types de clustering

#### 1.1.1 Clustering hiérarchique

Parmi les différents types de clustering existant [?], le premier étudié sera le clustering hiérarchique. Très utilisé comme outil d'analyse de données, l'idée principale du clustering hiérarchique est de construire un arbre binaire fusionnant de façon successive les groupes de points similaires. L'un des avantages de cette méthode est tout d'abord l'apport de l'arbre qui permet d'avoir une vision globale des données traitées. De plus, cette méthode de clustering possède ses propres outils de visualisation qui sont le dendrogramme et la classification double. Le dendrogramme permet d'illustrer l'arrangement des clusters (figure 1.1) tandis que la classification double est une technique d'explorations de données non-supervisée permettant de segmenter simultanément les lignes et les colonnes d'une matrice. Les avantages du clustering hiérarchique sont sa facilité d'implémentation dans des algorithmes tel que K-Means en plus de fournir une représentation comme dit précédemment. Cependant sa complexité le rend inefficace sur de larges jeux de données [?]. De plus, la première injection de données ainsi que leur ordre à un fort impact sur le résultat final. En outre, il n'est pas possible de défaire ou modifier les étapes précédentes du traitement, c'est à dire qu'une fois une instance assignée à un cluster, il n'est plus possible de la déplacer pour effectuer d'éventuelles modifications ou corrections [?]. Dans notre cas la base de CV utilisée n'étant pas de taille importante le clustering hiérarchique reste une méthode applicable. Cependant la problématique à résoudre

## Dendrogram



FIGURE 1.1 – Exemple de dendrogramme

est comment gérer efficacement les filières intégrant plusieurs domaines tel que la filière MIASHS de Nanterre qui possède une dimension mathématiques et une informatique. Les données étant représentées sous forme d'arbre cela entrainerait une répétition au niveau des résultats.

### **1.1.2 Clustering partionné**

LOREM IPSUM

## **1.2 Les algorithmes**

LOREM LE IPSUM

### **1.2.1 K-means**

### **1.2.2 Agglomerative Hierarchical**

### **1.2.3 DBSCAN**

## **1.3 Les types de clusters**

Le but du clustering étant de trouver des groupes d'objets présentant des similarités définies en fonction de l'objectif recherché. Il existe toutefois une multitude de types de cluster qui seront étudiés au sein de cette section chacun avec ses avantages et inconvénients en fonction de notre cas avant de statuer sur le type qui sera utilisé pour le reste de ce document.

### **1.3.1 Well-Separated**

### **1.3.2 Prototype-Based**

### **1.3.3 Graph-Based**

### **1.3.4 Well-Separated**