Clustering de Documents Texte

Pr T.RACHAD

QFM – 2024/2025

Plan

1. Notion de clustering
2. Types de clustering
3. Applications du clustering
4. Processus du clustering
5. Algorithmes de clustering

Notion de Clustering

- Le clustering, partitionnement ou regroupement de données est une méthode d'analyse de données non supervisée qui permet de catégoriser des objets en plusieurs groupes intitulés clusters.

- Les données d'entrée ne sont pas labélisées et doivent être subdivisées en plusieurs sous-ensembles significatifs qui regroupent des objets similaires.

- Un bon regroupement devrait regrouper des objets similaires et séparer les objets différents.

Applications du Clustering

- Le Clustering est très utile pour plusieurs applications:
  - Extraction de documents,
  - Segmentation d'images
  - Classification
  - …

- En text mining le clustering sert à explorer une collection massive de documents texte pour regrouper ceux qui sont similaires.

Types de clustering

- Le **clustering plat** (flat clustering) produit une partition unique d'un ensemble d'objets en groupes disjoints. Le nombre de clusters est prédéfini manuellement.

- le **clustering hiérarchique** produit une série imbriquée de partitions. C’est l’algorithme qui décide sur le nombre de clusters.

- Le **clustering dur** (Hard clustering) dans lequel chaque document est membre exactement d’un seul cluster.

- Le **clustering souple** (soft clustering) dans lequel l'affectation d'un document est une distribution sur tous les clusters.

Types de clustering

- Le **clustering agglomératif** commence avec chaque objet dans un cluster séparé et fusionne successivement les clusters jusqu'à ce qu'un critère d'arrêt soit satisfait.

- Le **clustering divisifs** commence par un seul cluster contenant tous les objets et effectue le fractionnement jusqu'à ce qu'un critère d'arrêt soit satisfait.

Processus du Clustering

- Le clustering peut contenir les étapes suivantes:
  1. Récupération du dataset
  2. Appliquer les prétraitements necessaires
  3. Extraction et sélection des éléments clés
  4. Définition des métriques adéquats permettant de mesurer les similarités entre les documents (fonction de similarité)
  5. Appliquer un algorithme de clustering.
  6. Description des clusters (Data abstraction)
  7. Evaluation.

Prétraitements

- Le clustering de documents texte en considérant tout le texte contenu dans chaque document est une tâche très complexe .

- Les documents doivent être convertis en une structure plus simple à traiter par des machines. À savoir une structure vectorielle en fonction d'un vocabulaire fixe:
  - One hot vector
  - Bag of words
  - TF-IDF
  - Word2Vec
  - Doc2Vec

Métriques de similarité

- Avant de décider sur un algorithme de clustering, il faut choisir la métrique adéquate qui permettra de mesurer les similarités entre les documents.

- Les métriques le plus utilisées sont:
  - la distance Euclidienne
  - La distance de Minkowski
  - La similarité de cosine
  - Indice de jaccard
  - …

Algorithmes de Clustering

- Les algorithmes de clutering courants sont:
  - K-means (dur, plat)
  - the EM-based mixture resolving (souple, plat, probabiliste)
  - HAC (hiérarchique, agglomératif).

Description du Clustering

- Une description significative et concise du cluster est sollicitée pour permettre par la suite un traitement automatique ultérieur ou pour aider les utilisateurs à interpréter les regroupements générés.

- Il existe de nombreuses possibilités pour générer automatiquement des étiquettes de cluster:
  - Le titre du document central ou plusieurs titres de document typiques peuvent être utilisés.
  - Plusieurs mots communs aux documents du cluster peuvent être affichés.
  - Une phrase nominale distinctive, si elle peut être trouvée, est probablement la meilleure étiquette.

Evaluation du Clustering

- La mesure la plus utilisée est la Pureté

- Supposons {L1, L2,. . . , Ln} sont les classes de documents étiquetées manuellement, et {C1, C2,. . . , Cm} sont les clusters renvoyés par le processus de clustering. Pureté (Ci) = $\max_j$ (| Lj ∩Ci | / | Ci |),

- D'autre mesures: entropie, informations mutuelles…