# Déduplication sur des types d'attributs hétérogènes

Loujain Liekah*, Yacine Gaci*, George Papadakis**

* LIRIS - University of Claude Bernard Lyon 1, Villeurbanne, France
loujain.liekah5@gmail.com, ey_gaci@esi.dz
** National and Kapodistrian University of Athens, Athenes, Greece
gpapadis@di.uoa.gr

**Résumé.** La déduplication est une tâche qui consiste à reconnaître plusieurs représentations d'un même objet du monde réel. La majorité des solutions existantes se concentrent sur les données textuelles et souvent négligent les attributs booléens et numériques, tandis que le problème des valeurs manquantes n'est pas suffisamment couvert. Les solutions supervisées ne peuvent être appliquées sans un nombre adéquat d'exemples étiquetés, ce qui implique des processus d'étiquetage coûteux en temps. Nous proposons dans ce papier D-HAT, un pipeline non supervisé qui est intrinsèquement capable de traiter des types d'attributs de haute dimension, épars et hétérogènes. Au cœur de ce pipeline se trouvent : (i) une nouvelle fonction de matching qui résume efficacement les signaux de correspondance multiples, et (ii) *MutMax*, un algorithme de regroupement glouton qui désigne comme doublons les paires ayant un score de matching mutuellement maximal. Nous évaluons D-HAT sur cinq datasets réels, et démontrons que notre approche surpasse significativement l'état de l'art.

## 1 Introduction

Le présent article est un résumé de l'article publié dans la conférence *Advanced Data Mining and Applications* (Liekah and Papadakis, 2022). L'intégration d'ensembles de données qui se chevauchent et se complètent est un processus courant qui crée des connaissances nouvelles et précieuses (Chen et al., 2014). Une tâche importante de l'intégration consiste à identifier les données qui représentent la même entité du monde réel, comme les produits, les instituts ou les patients. Cette tâche est appelée *déduplication* (Dong and Srivastava, 2015), matching d'entités (Konda et al., 2016), résolution d'entités (Papadakis et al., 2020b) ou couplage d'enregistrements. La déduplication améliore la qualité des données en réparant et en conservant les sources de données (Fan et al., 2014), en réduisant la taille du stockage et en préparant les données pour les applications en aval (Dong and Srivastava, 2015).

La majorité des solutions pour la déduplication sont basées sur le calcul de scores de similarité par paire à partir d'un ou plusieurs attributs (Christophides et al., 2021). Les méthodes *non supervisées* créent un graphe de similarité, où les nœuds correspondent aux enregistrements et où les arêtes sont pondérés par les scores de matching des nœuds adjacents (Hassanzadeh et al., 2009). Le graphe est ensuite divisé en groupes de telle sorte que tous les nœuds de chaque