الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جامعة سعيدة د. مولاي الطاهر

كلية التكنولوجيا

قسم: الإعلام الآلي

Mémoire de Master

Spécialité : Réseaux Informatique et Systèmes Réparties

Thème

La sélection automatique de clé de blocage pour le couplage d'enregistrements

Présenté par :

Mr ABSI Tahar

Mr TENNAH Abdelhamid

Dirigé par :

Mr BENYAHIA Miloud

Promotion 2021 - 2022

Remerciements

En préambule à ce mémoire nous remercions ALLAH qui nous a aidé et nous a donné la patience et le courage durant ces langues années d'étude.

Nous tenons également à exprimer nos vifs remerciements et toutes nos reconnaissances à Notre encadreur Monsieur BENYAHIA Miloud, nous le remercions de nous avoir encadré, orienté, aidé et conseillé.

Nos remerciements vont tout d'abord aux nos parents. Vous trouverez ici le résultat de longues années de sacrifices et de privations. Merci pour les valeurs nobles, l'éducation et le soutien permanent venu de vous.

Nous remercions les membres du jury d'avoir pris le temps de lire et d'évaluer notre travail.

Nous souhaitons adresser encore nos remerciements les plus sincères aux personnes qui nous ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire.

Nous tenons encore à exprimer nos sincères remerciements à tous les enseignants qui nous ont enseigné et qui par leurs compétences nous ont soutenu dans la poursuite de nos études.

Enfin, nous remercions toute personne qui a participé de près ou de loin pour l'accomplissement de ce modeste travail

Absi Tahar& Tennah Abdelhamid

Dédicace

A

Mes chers parents,

Ma chère épouse,

Mes adorables enfants,

Maisem –chihab-safa et marwa

Toute ma famille,

Mes chères amies,

Je dédie ce modeste travail, Qu'ils y trouvent l'expression de ma gratitude et de ma profonde affection.

ABSI Tahar

Dédicace

Ce modeste travail est dédie :

A

- *Mes chers parents.*
- *Toute ma famille.*
- *Tous mes amis.*
- *Toutes les personnes qui m'ont apport´es de l'aide.*

TENNAH Abdelhamid

Résumé

Le couplage d'enregistrements peut être utilisé pour améliorer la qualité et l'intégrité des données, pour permettre la réutilisation des sources de données existantes pour de nouvelles études et pour réduire les coûts et les efforts d'acquisition de données pour les études de recherche. L'étape la plus critique du processus couplage d'enregistrements est le blocage, qui réduit la complexité quadratique du processus en divisant les données en un ensemble de blocs. Cependant, sélectionner les meilleures clés de blocage pour diviser les données est une tâche difficile. L'algorithme K-Modes est utilisé durant l'étape de blocage, K-Modes propose un avantage majeur car l'algorithme traite directement les données catégorielles. La sélection automatique de clés de blocage pour le couplage d'enregistrements est surmontée on utilisant l'algorithme d'optimisation de recherche méta-heuristique récemment proposé pour les pygargues à tête blanche, où le problème est traité comme un cas de sélection de caractéristiques. Les résultats obtenus à partir d'expériences sur des ensembles de données du monde réel ont montré l'efficacité de l'algorithme d'optimisation utilisée.

Mots Clés

Qualité des données, Couplage d'enregistrements, clés de blocage, Indexation, K-Modes, Algorithme d'optimisation.

Abstract

Record linkage can be used to improve data quality and integrity, to enable reuse of existing data sources for new studies, and to reduce data acquisition costs and effort for research studies . The most critical step in the record linkage process is blocking, which reduces the quadratic complexity of the process by dividing the data into a set of blocks. However, selecting the best blocking keys to split data is a difficult task. The K-Modes algorithm is used during the blocking step, K-Modes offers a major advantage because the algorithm directly processes categorical data. The automatic selection of blocking keys for record linkage is overcome using the recently proposed meta-heuristic search optimization algorithm for bald eagles, where the problem is treated as a case of feature selection. Results obtained from experiments on real-world datasets showed the effectiveness of the optimization algorithm used.

Keywords

Data Quality, Record Coupling, Blocking Keys, Indexing, K-Modes, Optimization Algorithm.

ملخص

يمكن استخدام ربط السجلات لتحسين جودة البيانات وسلامتها ، ولتمكين إعادة استخدام مصادر البيانات الحالية للدراسات الجديدة ، ولتقليل تكاليف الحصول على البيانات والجهود المبذولة للدراسات البحثية. الخطوة الأكثر أهمية في عملية ربط السجلات هي الحجب ، مما يقلل من التعقيد التربيعي للعملية عن طريق تقسيم البيانات إلى مجموعة من الكتل. ومع ذلك ، يعد تحديد أفضل مفاتيح الحظر لتقسيم البيانات مهمة صعبة. تُستخدم خوارزمية K-Modes أثناء خطوة الحظر ، وتوفر K-Modes ميزة كبيرة لأن الخوارزمية تعالج البيانات الفئوية مباشرةً. يتم التغلب على الاختيار التلقائي لمفاتيح الحظر لربط التسجيل باستخدام خوارزمية تحسين البحث التلوي المقترحة مؤخرًا للنسور الصلعاء ، حيث يتم التعامل مع المشكلة كحالة اختيار الميزة. أظهرت النتائج التي تم الحصول عليها من التجارب على مجموعات البيانات الواقعية فعالية خوارزمية التحسين المستخدمة.

كلمات مفتاحية :

جودة البيانات ، اقتران التسجيل ، حظر المفاتيح ، الفهرسة ، أوضاع K ، خوارزمية التحسين.

Sommaire

Liste des figures

Liste des tableaux

Introduction

Générale

Introduction Générale

Les mauvaises données coûtent des milliers de milliards de dollars par an rien qu'aux États-Unis. La raison de ces pertes est que divers processus tels que la prise de décision prennent en charge ces mauvaises données chaque jour. Les organisations du monde entier sont désormais plus conscientes de l'importance de la qualité des données dans leurs bases de données, où beaucoup d'argent est investi afin d'améliorer la qualité des données stockées.

Les problèmes de qualité des données peuvent apparaître de différentes manières. En effet, il y a des problèmes d'exhaustivité (valeurs manquantes), de duplication valeurs, problèmes d'intégrité référentielle et bien d'autres anomalies.

Dans ce Mémoire, nous nous concentrons sur le problème de record Linkage(RL) connu sous le nom le couplage d'enregistrements. Record Linkage est le processus qui vise à détecter tous les enregistrements qui se réfèrent à la même entité du monde réel, puis fusionnez-les en un seul.

Le couplage d'enregistrements peut être défini comme un processus en trois étapes où la première étape est **le nettoyage et la normalisation**. L'application du processus RL sur des données sales peut finir par fusionner les mauvais tuples et perdre des informations importantes de nos bases de données. L'étape suivante est «**l'indexation**», elle est considérée comme l'étape la plus importante du processus. Dans cette étape, tous les enregistrements qui représentent une correspondance possible sont regroupés dans un même bloc afin d'être comparés les uns aux autres. La dernière étape consiste à faire **correspondre** les paires d'enregistrements indexés et, par conséquent, nous pouvons obtenir des correspondances, des non-correspondances ou des enregistrements de correspondances possibles.

Cette étude est structurée en chapitres et organisée comme suite :

Chapitre 1 : Cette partie décrit le contexte général dans lequel s'inscrit ce mémoire.

Chapitre 2 : nous donnons un aperçu des solutions existantes proposées dans la littérature et ayant traité le problème de la sélection des clés de blocage.

Chapitre 3 : présente la méthodologie suivie et les outils utilisés pour l'amélioration de l'approche de sélection automatique de clés de blocage, plus précisément en utilisant des métas heuristiques.

Chapitre 4 : Enfin ce chapitre décrira l'environnement de travail et les expérimentations faites. Nous terminerons ce mémoire par une conclusion générale qui servira de base à un futur travail sur le même thème.

Chapitre 1

Contexte

Chapitre 1

Contexte

1.1 Introduction

Les organisations du monde entier perdent une somme énorme à caudé de problèmes de qualité des données. Les parties prenantes sont désormais plus conscientes de l'importance de la qualité des données. Beaucoup d'argent est investi pour améliorer la qualité des données stockées. L'un des principaux processus importants dans le domaine de la qualité des données est le couplage d'enregistrements. Le couplage d'enregistrement est le processus de détection des doublons qui font référence à la même entité réelle dans un ou plusieurs ensembles de données. L'une des étapes les plus importantes du processus RL est le blocage.

1.2 Qualité des données

1.2.1 Définition

La qualité des données est un terme générique décrivant à la fois les caractéristiques des données : complète, fiable, pertinents, cohérente set à jour, il permet de garantir ces caractéristiques par des ensembles des processus [**Boumediene et Wassim, 2015/2016**] La qualité des données consiste à obtenir des données sans duplication, fautes d'orthographe, omissions, modifications redondantes et cohérentes avec la structure. Elle fait également référence à l'utilisation globale d'un ou plusieurs dataset. Les données sont basées sur leur facilité de traitement et d'analyse à d'autres fins. Capacité, généralement traitée par une base de données, un entrepôt de données ou un système d'analyse de données

1.2.2 Les critères de la qualité des données

1. **Les critères intrinsèques**

(a) **L'unicité** : L'unicité est le fait qu'une entité du monde réel ne soit représentée que par un seul et unique objet métier au sein de l'entreprise. Cet objet ne répond donc qu'à un identifiant unique.. L'unicité des données sert aussi à n'avoir qu'une seule description d'un produit donné. Elle contribue alors à l'amélioration de la qualité des données produit [**Rgnier-Pcastaing et al., 2008**].

(b) **L'exactitude** : Une donnée est " exacte " si la valeur des attributs de l'entité concernée est égale à la grandeur qu'elle est censée représenter dans le monde réel. Cette notion englobe donc deux aspects : la précision et la validité [**Rgnier-Pcastaing et al., 2008**].

(c) **La complétude** : La complétude est la présence de valeurs de données significatives pour un ou des attributs, un ou des objets [**Rgnier-Pcastaing et al., 2008**].

(d) **La cohérence** : Cette notion est relative à l'absence d'informations conflictuelles au sein d'un même objet (par exemple, une incohérence serait détectée si un " prix actuel " d'un produit est supérieur au " prix maximum " de ce même produit). Mais cette notion existe aussi au niveau service : les valeurs d'une instance d'un objet métier ne sont pas en conflit avec les valeurs d'une autre instance ou d'une instance d'un autre objet [**Jamm, janvier 2008**].

(e) **L'intégrité** : L'intégrité concerne les relations entre objets. Les relations importantes entre objets sont-elles toutes présentes ? Exemple : toute facture

doit être associée à une commande. Si une facture n'a pas de référence vers une commande, c'est un problème d'intégrité [**Rgnier-Pcastaing et al., 2008**].

2. Les critères de services

(a) **L'actualité** : Une valeur de donnée est à jour si elle est correcte en dépit d'un écart possible avec la valeur exacte, due à des changements liés au temps ; une donnée est périmée à la date tsi elle est incorrecte à cette date mais était correcte aux instants précédant t. L'actualisation est le degré mesurant à quel point une donnée en question est à jour (par exemple, l'âge ne devient obsolète qu'à la date anniversaire)[**Rgnier-Pcastaing et al., 2008**].

(b) **L'accessibilité** : Est la dimension qualité qui concerne la facilité d'accès aux données. Cela signifie que les services de données sont calibrés en fonction de leur utilisation et qu'ils existent souvent aussi bien en mode événement (déclenché à chaque mise à jour), qu'en mode requête (à la demande d'un processus consommateur) ou en mode batch pour des synchronisations en masse (pour le décisionnel par exemple) [**Jamm, janvier 2008**].

(c) **La pertinence** : La pertinence est la dimension qualité qui définit l'utilité d'une donnée. Une donnée peut être accessible mais tellement détaillée que de nombreux attributs de l'objet proposé sont inutiles aux processus consommateurs. Une donnée doit être adéquate à son usage. Les services de donnée seront d'autant mieux utilisés que la granularité d'information dispensée correspondra aux besoin [**Jamm, janvier 2008**].

(d) **La compréhensibilité** : La compréhensibilité est la dimension qualité associée à la question : " cette donnée est-elle compréhensible ? ". Une donnée est compréhensible si chaque utilisateur, chaque processus, chaque application trouve facilement la bonne information parmi les attributs disponibles d'un objet. C'est le cas si celui-ci est clair et que l'alignement sémantique de

l'ensemble des concepts entre tous les dépositaires (humains ou informatiques) a été réalisé et documenté [**Jamm, janvier 2008**].

1.2.3 l'importance de la qualité de données

La qualité des données ne consiste pas seulement à aider les organisations à charger les bonnes données dans leurs systèmes d'information. Elle permet d'éliminer les données erronées ou les données en double. Le nettoyage des données devient une étape importante dans l'intégration des informations dans le système. La gestion de la qualité des données est la capacité de fournir des données fiables pour répondre aux besoins commerciaux et techniques des utilisateurs. Il est mesuré en termes d'exactitude, de cohérence, d'unicité, d'exhaustivité et de disponibilité. Il s'agit d'une méthode de gestion de l'information conçue pour gérer et comparer des données entre différents systèmes d'information ou bases de données d'une entreprise. Habituellement, il s'agit de convertir des données de qualité en informations utiles essentielles à l'organisation.

1.2.4 Principaux problèmes du non qualité des données

Les problèmes des données ne naissent pas de nulle part, les causes de la non qualité des données sont connues : On trouve les problèmes techniques ou les problèmes humains. Ces problèmes s'accumulent avec le temps, depuis la création, durant la manipulation et jusqu'à l'exploitation et l'analyse [Berti-Equille, 2006]

1. **Création des données :**

- Entrée manuelle : absence de vérifications systématiques des formulaires de saisie
- Entrée automatique : problèmes de capture OCR, de reconnaissance de la parole Incomplétude, absence de normalisation ou inadéquation de la

modélisation conceptuelle des données : attributs peu structurés, absence de contraintes d'intégrité pour maintenir la cohérence des données

- Entrée de doublons
- Approximations
- Contraintes matérielles ou logicielles
- Erreurs de mesure
- Corruption des données : faille de sécurité physique et logique des données

2. **Collecte/import des données :**

- Destruction ou mutilation d'information par des prétraitements inappropriés.
- Perte de données : buffer over flows, problèmes de transmission.
- Absence de vérification dans les procédures d'import massif.
- Introduction d'erreurs par les programmes de conversion de données.

3. **Stockage des données :**

- Absence de méta-données.
- Absence de mise à jour et de rafraîchissement des données obsolètes ou répliquées.
- Modifications ad-hoc.
- Modèles et structures de données inappropriés, spécifications incomplètes ou évolution des besoins dans l'analyse et conception du système.
- Contraintes matérielles ou logicielles.

4. **Intégration des données :**

- Problèmes d'intégration de multiples sources de données ayant des niveaux de qualité et d'agrégation divers.
- Problèmes de synchronisation temporelle.
- Systèmes de données non conventionnels.
- Facteurs sociologiques conduisant à des problèmes d'interprétations et d'intégration des données.

5. **Recherche et analyse des données :**

- Erreur humaine.
- Contraintes liées à la complexité de calcul.
- Contraintes logicielles, incompatibilité.
- Problèmes de passage à l'échelle, de performances et de confiance dans les résultats.
- Approximations dues aux techniques de réduction des grandes dimension

1.2.5 Approches générales pour détecter et corriger les problèmes de qualité des données

Comme le représente la figure 1.1, on peut classer la plupart des travaux abordant la problématique de la qualité des données selon quatre grands types d'approches complémentaires [**Berti-Equille, 2006**]

— Les approches préventives centrée sur l'ingénierie des systèmes d'information et le contrôle des processus avec des techniques permettant d'évaluer la qualité des modèles conceptuels, la qualité des développements logiciels et celle des processus employés pour le traitement des données,

— Les approches diagnostiques centrées sur des méthodes statistiques, d'analyse et de fouille de données exploratoire permettant de détecter des anomalies sur les données,

— Les approches correctives centrées sur des techniques de nettoyage et de consolidation de données et utilisant des langages de manipulation des données étendus et des outils d'extraction et de transformation de données (ETL Extraction-Transformation-Loading)

— Les approches adaptatives ou actives appliquées généralement lors de la médiation ou de l'intégration des données : elles sont centrées sur l'adaptation des traitements (requêtes ou opérations de nettoyage sur les données) de telle façon que ceux-ci incluent à l'exécution en temps-réel la vérification des contraintes sur la qualité des données.

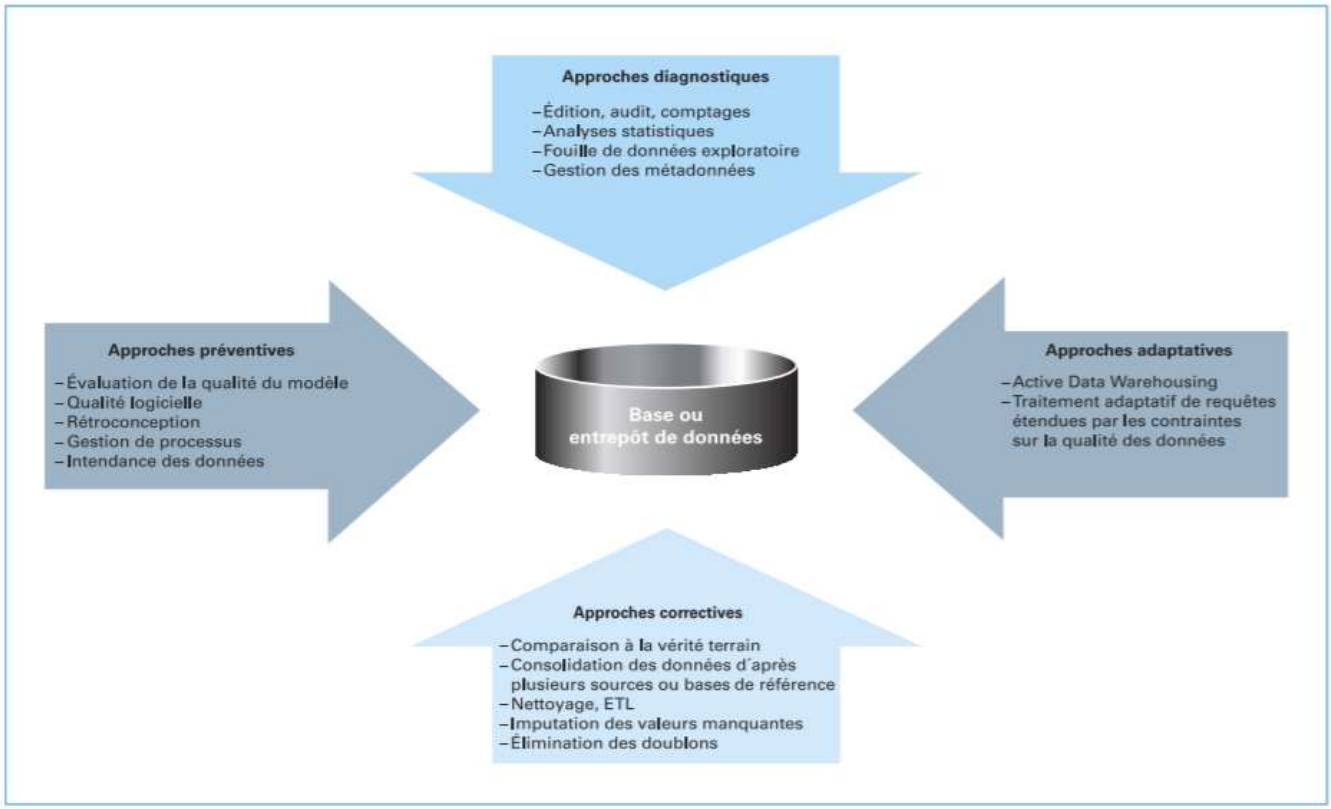

Figure 1.1 – Panorama des approches pour l'évaluation et le contrôle de la qualité des données.

1.3 Record Linkage

1.3.1 Définition

Record Linkage (RL), également connu sous le nom de couplage d'enregistrements, est le processus qui vise à identifier les enregistrements qui font référence à la même entité du monde réel.

Les techniques de record linkage sont utilisées pour relier les enregistrements de données relatifs aux mêmes entités, telles que les patients ou les clients. Le record linkage peut être utilisé pour améliorer la qualité et l'intégrité des données, pour permettre la réutilisation des sources de données existantes pour de nouvelles études et pour réduire les coûts et les efforts d'acquisition de données pour les études de recherche.

1.3.2 Méthodologie du Record Linkage

Il existe deux méthodes principales de record linkage :

Déterministe : il est déterminé par le nombre d'identifiants correspondants. il génère des liens en fonction du nombre d'identificateurs individuels qui correspondent parmi les ensembles de données disponibles. on dit que deux enregistrements correspondant via une procédure de couplage d'enregistrements déterministe si tous ou certains identificateurs sont identiques. Le couplage d'enregistrements déterministe est une bonne option lorsque les entités des ensembles de données sont identifiées par un identifiant commun, ou lorsqu'il existe plusieurs identifiants représentatifs (par exemple, nom, date de naissance et sexe lors de l'identification d'une personne) dont la qualité des données est relativement haute.

- **Probabiliste** : il est déterminé par la probabilité d'un certain nombre d'identifiants correspondents. Le couplage d'enregistrements probabilistes tente de relier deux éléments d'information ensemble à l'aide de plusieurs clés, éventuellement non uniques. Par exemple, dans une étude basée sur un registre, les événements de la maladie peuvent être liés aux données de mortalité en utilisant des combinaisons de nom et prénom non uniques.

1.3.3 Les étapes de Record Linkage

Le record linkage peut être défini comme un processus en trois étapes :

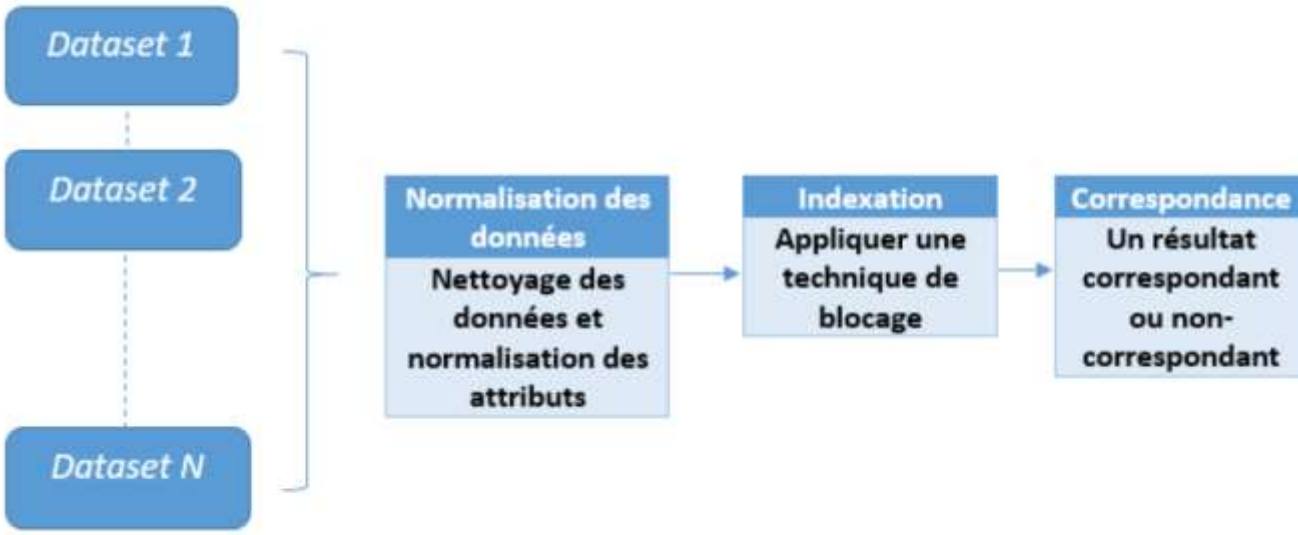

Figure 1.2 – Les étapes de Record Linkage.

1. **Nettoyage et normalisation** : appliquer le processus RL à une donnée corrompue peut aboutir à la fusion des mauvais tuples et à la perte d'informations importantes dans nos base de données. Par exemple, l'attribut d'adresse peut être représenté dans une base de données sous la forme d'un champ unique, mais dans une autre sous forme de plusieurs champs (code postal, rue, ville, etc.). Par conséquent, afin de faciliter le processus RL, la normalisation du champ d'adresse doit être effectuée avant de démarrer le processus RL.

2. **L'indexation** : elle est considérée comme l'étape la plus importante du processus. L'indexation est l'endroit où tous les enregistrements représentant une correspondance possible sont regroupés dans le même bloc afin d'être comparés les uns aux autres.La technique d'indexation la plus utilisée est le «blocage».

3. **La mise en correspondance** des paires d'enregistrements indexés : le résultat peut être l'une des trois (correspondances, non-correspondances, correspondances possibles). Dans le cas du troisième résultat, nous avons besoin de correspondances, correspondances possibles).

Dans le cas du troisième résultat, nous avons besoin de l'intervention d'un expert du domaine pour décider si les paires d'enregistrements représentent la même entité du monde réel.

1.4 Le blocage

1.4.1 Définition

Le blocage est la technique la plus utilisée dans l'étape de l'indexation.
Le blocage est le processus qui divise le dataset en un ensemble de blocs.
Tous les tuples affectés au même bloc partagent une valeur commune appelée la valeur de clé de blocage (BKV).

Une clé de blocage peut être choisie comme un attribut unique. Par exemple tous les enregistrements qui partagent la même valeur pour l'adresse d'attribut sont affectés au même bloc. Sinon, une clé de blocage peut également être choisie avec la concaténation de plusieurs attributs comme les quatre premiers caractères du prénom et le code postal de l'attribut d'adresse.

BK	Name	Address	City	Phone	TYpe
Losangelos310/246-1501	Amine morton's	435 s.la ceinega blv	Los angelos	310/246-1501	American
Studiocity818 /762-1221	Art's delicatessen	12224 ventura blvd	Studio city	818 /762-1221	American

Figure 1.3 – Exemple de clé de blocage.

Deux paramètres importants contrôlent les performances d'une bonne technique de blocage :

- **La valeur de la clé de blocage (BKV)** : Une clé de blocage peut être formée en utilisant un champ (attribut) ou une concaténation de plusieurs parties d'un ensemble de champs. Par exemple, un BKV peut être formé à l'aide de la valeur Prénom ou il peut être formé par la concaténation des trois premiers caractères du champ Prénom et du code postal du champ d'adresse.Tableau 1 montre un exemple de blocage de clés générées à partir du dataset restaurant. Deux clés de blocage ont été générées. Le premier (BK1) est l'encodage phonétique Soundex du nom du restaurant concaténé avec le numéro de téléphone. Le second (BK2) est le codage phonétique NYSIIS du nom du restaurant concaténé avec le numéro d'adresse.

- **Le nombre de clés de blocage** : l'utilisation de plus d'une seule clé de blocage peut améliorer l'efficacité des techniques de blocage puisque les valeurs des clés de blocage contiendront plus d'informations sur l'enregistrement. Par conséquent, sélectionner les meilleurs champs pour former un bon BKV et le nombre de clés de blocage est une étape très importante pour obtenir le meilleur résultat du processus

1.4.2 L'objectif de blocage

Le blocage a deux objectifs principaux :
1. Le nombre d'appariements candidats générés doit être petit pour minimiser le nombre de comparaisons détaillées à l'étape de record linkage.

2. L'ensemble candidat ne doit pas omettre d'éventuelles correspondances vraies, puisque seules les paires d'enregistrements de l'ensemble candidat sont examinées en détails lors du record linkage.

Ces objectifs de blocage représentent un compromis. D'une part, le but du record linkage est de trouver tous les enregistrements correspondants.

1.5 Conclusion

De nos jours, avec les développements technologiques, les entreprises stockent de plus en plus de donnée. Malheureusement les travaux de maintenance et de la qualité des données sont souvent négligés, pourtant les données de mauvaise qualité constituent un facteur de cout important.

Les données de mauvaise qualité peuvent donc avoir des effets significativement négatifs sur l'efficacité d'une organisation.

Dans ce chapitre, on a commencée par définir qualité des données et leur concept, par la suite on abordera les conséquences de la non-qualité. Ensuite on a expliqué le couplage d'enregistrements. Et finalement on a parlé de la technique de blocage.

Chapitre 2

Etat de l’art

Chapitre 2

Etat de l'art

2-1 Introduction

Dans ce chapitre, on va présenter un état de l'art sur les approches existent dans la littérature qui ont accordé une attention aux couplage d'enregistrements et un aperçu des solutions existantes proposées dans la littérature et ayant traité le problème de la sélection automatique des clés de blocage.

2.2 Record Linkage

Les deux étapes les plus importantes du processus de couplage d'enregistrements : les étapes d'indexation et La mise en correspondance. L'indexation est une étape critique du processus de RL. En fait, au cours de cette étape, le nombre de comparaisons est réduit en éliminant autant que possible les paires d'enregistrements sans correspondance. [**Christen, b**].

La technique d'indexation la plus utilisée est le «**blocage**», il est utilisé depuis les premières applications du record linkage [P.Fellegi et B.Sunter]. Le blocage consiste à créer un ensemble de blocs de manière à ce que tous les tuples du même bloc partagent la même valeur de clé de blocage [**P.Fellegi et B.Sunter**].

La clé de blocage peut être sélectionnée comme un attribut unique ou une combinaison de plusieurs attributs. Bien sûr, il existe une approche naïve dans laquelle chaque enregistrement est comparé à tous les autres ; par conséquent, dans le cas d'une grande base de données, il en résulte des milliards de comparaisons.

Plusieurs approches de blocage ont été proposées dans la communauté du record linkage : Le premier est le blocage traditionnel, qui regroupe les enregistrements qui partagent une valeur de clé de blocage similaire dans le même bloc [**A.Jaro**]. De cette manière, seuls les enregistrements appartenant au même groupe sont comparés les uns aux autres. Une autre approche proposée est le quartier triés [**Hernàndez et Stolfo**]. Il consiste à générer les clés de blocage et à trier les enregistrements par ordre alphabétique.

Une fois le tri terminé, une fenêtre glissante est déplacée sur les enregistrements. Pour chaque itération, seuls les enregistrements de la même plage de fenêtres sont comparés les uns aux autres. Cette approche a été étendue plus tard dans [**Christen, c**] en utilisant un tableau d'indexation inversé et dans[Yan et al.] les auteurs ont utilisé une fenêtre à changement dynamique.

L'indexation Q-gram est également une puissante approche d'indexation. L'idée derrière cela, est de diviser le BKV en sous-chaînes de taille Q, puis de sélectionner un nombre (fixé par l'utilisateur) de ces sous-chaînes et de les concaténer pour former les nouvelles valeurs de clés de blocage. Dans [**McCallum et al.**] , les auteurs ont proposé une technique d'indexation basée sur l'algorithme de clustering Canopy. Il se compose de deux étapes principales. La première étape consiste à diviser les données en petites fractions de données appelées auvents en utilisant des méthodes peu coûteuses telles que le blocage d'index inversé. Une fois les auvents créés, la deuxième étape consiste à exécuter un algorithme de clustering classique dans chaque auvent associé à une métrique de distance coûteuse comme la distance d'édition.

Une autre technique d'indexation est l'indexation basée sur un tableau de suffixes. Cette approche a été proposée pour la première fois dans [**A et K**], l'idée de base de cette approche est de générer un certain nombre de suffixes à partir des valeurs des clés de blocage avec une longueur minimale fixée par l'utilisateur et de les insérer dans un tableau d'indexation inversé. L'utilisation d'un tableau d'indexation inversé donne la possibilité d'insérer le même enregistrement dans plus d'un bloc comme l'approche d'indexation Q-gram. Cette approche génère (c - lm + 1) suffixe à partir d'une valeur de clé de blocage d'un caractère "c" et d'un minimum longueur suffisante de "lm" . Cette approche a ensuite été étendue dans [**A et K**] avec la possibilité de fusionner les enregistrements appartenant à un suffixe similaire bloquant les clés après avoir mesuré la similitude entre elles. Pour plus d'informations sur toutes les techniques d'indexation qui existent dans la littérature, une enquête a été publiée par Christen dans [**Christen, b**].

Une fois l'indexation terminée, les enregistrements indexés seront comparés les uns aux autres en utilisant une technique d'appariement et décideront si les paires d'enregistrements représentent la même entité du monde réel ou non. Généralement, la valeur de correspondance est normalisée entre la plage de [0,1] où 1 représente une correspondance exacte et 0 une non-correspondance totale [**Christen, a**].

Plusieurs algorithmes de correspondance de chaînes existent dans les littératures ; certains d'entre eux appartiennent à la famille des encodages phonétiques comme (Soundex et phonex [**Holmes et McCabe**], phoenix [**Gadd**], NYSIIS et Double-Metaphone [**Philips**]). D'autres appartiennent à la famille de recherche de motifs comme l'algorithme Edit-distance qui est défini dans [**lev**] comme le nombre d'insertions, de suppressions et de substitutions pour transformer une chaîne en une autre. Il est généralement mis en oeuvre à

l'aide d'un programme dynamique. Plus d'informations sur les techniques d'appariement quittées peuvent être trouvées dans [**Elmagarmid et al.**]. Une autre approche a été proposée pour détecter les doublons en utilisant des techniques d'apprentissage automatique comme dans [**OUHAB et al.**] où les auteurs ont utilisé l'algorithme SVM pour classer la paire d'enregistrements comme correspondances ou non.

2.3 Approches de blocage

Plusieurs approches de blocage ont été proposées par la communauté RL. Chacun d'eux dépend d'une manière ou d'une autre d'un bon choix de clé de blocage :

- La première technique de blocage proposée est le blocage standard. Il consiste à regrouper tous les enregistrements qui partagent la même clé de blocage dans le même groupe. La sélection du meilleur attribut comme clé de blocage est donc très cruciale pour cette approche.

- L'indexation Q-gram L et al. [**a**] est également une approche de blocage très populaire qui dépend du choix initial de la clé de blocage. Dans cette approche, la valeur de la clé de blocage est divisée en un ensemble de Q-gams. Ensuite, ces Q-grammes obtenus sont utilisés comme nouvelles clés de blocage pour former les nouveaux blocs.

— Une autre approche de blocage populaire qui dépend de la sélection initiale de la clé de blocage est l'indexation basée sur un tableau de suffixes [**A et K**] où un ensemble de suffixes est généré à partir des valeurs de clé de blocage sélectionnées à partir de nouveaux blocs.

— L'approche des quartiers triés (sorted neighborhood) [**M.A et S.J**] commence également par la génération des clés de blocage. Une fois que cela est fait, les enregistrements sont triés en fonction de leurs BK générés et une fenêtre glissante, de tailleW, se déplace sur les enregistrements. Tous les

enregistrements qui sont dans la même plage de fenêtres sont comparés les uns aux autres. Cette approche a été améliorée plus tard dans [**Yan et al.**].

— Le blocage basé sur les K-Modes proposé dans [**H.N et al.**] dépend également du choix des clés de blocage initiales, puisque les données sont regroupées en utilisant uniquement les clés de blocage comme attributs de clustering au lieu d'utiliser tous les attributs de l'ensemble de données.

— La déduplication peut également être utilisée pour supprimer des fichiers identiques du stockage Big Data, [Y et al.] a proposé une approche de déduplication qui permet de supprimer des fichiers médicaux identiques contenant les mêmes données sur la base de la cryptographie.

— [**D.C et al.**] a proposé une nouvelle approche de blocage qui permet le contrôle des tailles de bloc générées, l'approche proposée commence par la réduction de bloc qui supprime les enregistrements qui ont la cooccurrence moyenne la plus faible. Ensuite, tous les blocs qui ont une taille supérieure à une valeur prédéfinie sont divisés en blocs plus petits tandis que le bloc de très petite taille est fusionné en fonction de la similitude des attributs .

2.4 Clé de blocage automatique

La plupart des approches de blocage dépendent de la sélection initiale de la clé de blocage, proposer une solution de sélection de clé de blocage automatique est l'une des priorités les plus importantes de la communauté RL. Ces dernières années, la communauté Record Linkage a proposé plusieurs approches de sélection automatique des clés de blocage. Certaines de ces approches nécessitent l'existence d'un dataset de référence puisqu'elles sont basées sur des algorithmes d'apprentissage supervisé comme [**Vogel et Naumann**],[**M et al., b**], [**M et C.A**].

— [**Vogel et Naumann**] ont proposé une approche automatique pour la sélection des clés de blocage basée sur les combinaisons d'unigramme en tant

que clés de blocage générées [**Vogel et Naumann**]. La première étape de cette approche consiste à générer toutes les combinaisons d'uni-gramme possibles. Pour chaque combinaison, un algorithme de détection de doublons est exécuté sur un dataset de référence. Après chaque exécution, si le nombre de doublons détectés est acceptable, la qualité globale de la clé de blocage est calculée. Toutes les clés de blocage sélectionnées à partir de cette étape seront stockées et triées en fonction de leur qualité. Une fois la première étape effectuée, le processus de record linkage est exécuté sur un dataset de test à l'aide de la liste des clés de blocage triées de l'étape précédente et le meilleur BK satisfaisant aux critères d'arrêt choisis est sélectionné.

— [**M et al., b**] ont proposé de générer un ensemble de prédicats de blocage. Chaque prédicat peut être spécifié pour un attribut du dataset . De cette manière, le problème de la sélection automatique des clés de blocage consiste à savoir comment sélectionner le meilleur sous-ensemble de prédicats de blocage qui détecte le plus de doublons possibles dans le dataset. Les auteurs ont utilisé le problème de couverture d'ensemble rouge-bleu pour sélectionner les meilleurs prédicats où les lignes du haut et du bas sont pour les paires positives et négatives, tandis que la ligne du milieu représente tous les prédicats de blocage générés.

2.5 Sélection des attributs

Nous avons un ensemble d'approches proposées qui ne nécessitent pas l'existence d'un dataset standard .

— [**Kejriwal et Miranker**] ont proposé une approche non supervisée pour la sélection des clés de blocage. Ils ont utilisé le critère de discrimination des pêcheurs pour sélectionner les meilleures clés de blocage après la génération d'un dataset faiblement étiqueté. En effet, ils ont traité la situation comme un problème de sélection de fonctionnalités.

— [**B et P**] ont proposé une autre approche non supervisée pour la sélection automatique des clés de blocage sur la base de trois critères principaux. Le premier est la couverture clé. C'est le nombre de paires d'enregistrements couvertes par chaque clé de blocage. Le second est la taille de bloc, qui est le nombre d'enregistrements dans le même bloc où ils ont fixé une taille de bloc maximale car ils traitent le problème de la résolution d'entité en temps réel. Le dernier est la distribution des blocs qui représente la variance. Les résultats obtenus ont montré que leur approche peut retourner des clés de blocage de bonne qualité dans un temps raisonnable pour RL en temps réel.

— [**J et Q**] ont proposé une approche d'apprentissage par schéma de blocage actif . Ils ont utilisé des techniques d'apprentissage actif pour sélectionner le meilleur schéma de blocage. Leur approche se compose de deux étapes principales. La première est l'échantillonnage actif. Pendant cette phase, un taux d'équilibre est calculé pour chaque schéma de blocage sur un dataset. Le but est de minimiser le taux d'équilibre. La deuxième étape est la ramification active. C'est là qu'un schéma de blocage local est recherché sur l'ensemble d'apprentissage. Les expériences sur quatre datasets du monde réel montrent de bons résultats, en particulier en ce qui concerne le rapport de réduction.

— [**M et al., a**] a proposé une approche de sélection automatique des clés de blocage qui traite spécifiquement des dataset en langue arabe, les auteurs ont choisi d'étendre le travail de B et P en ajoutant d'autres types de clés de blocage générées en utilisant des stems et IsExactStem en tant que fonction dans l'étape de génération des BK puisque les stems ont prouvé leur efficacité dans le cas de la langue arabe.

2.6 Sélection des attributs à l'aide d'une métaheuristique

La sélection des fonctionnalités est considérée comme un problème NP-Hard. L'utilisation de méthodes exactes risque de présenter une grande complexité de calcul. Par conséquent, dans l'apprentissage automatique, le problème de la sélection des fonctionnalités est résolu à l'aide de méthodes stochastiques telles que l'heuristique et la méta-heuristique. Il existe trois stratégies de sélection des fonctionnalités : Filtre, Hybride et Intégré [G et F].
Dans les approches utilisant le filtre, les entités sont triées selon des critères de classement où toutes les entités qui ont un score inférieur à un seuil défini sont éliminées. Plusieurs méthodes de classement ont été proposées dans la littérature comme les critères de corrélation [**I et A**] et les critères conditionnels basés sur l'information mutuelle [**F**]. Cependant, les méthodes wrapper sont considérées comme plus efficaces que les filtres puisqu'elles sont basées sur un algorithme de recherche guidé par un classificateur pour mesurer la fonction objective. La plupart des approches existantes utilisant le wrapper sont basées sur des méta-heuristiques comme algorithme de recherche puisque les métaheuristiques conviennent pour résoudre les problèmes NP-Hard.

Plusieurs algorithmes métaheuristiques ont été utilisés pour résoudre le problème de la sélection des attributs. Certains d'entre eux ont utilisé directement une méta-heuristique et d'autres ont essayé de combiner deux ou plusieurs méta-heuristiques pour améliorer à la fois l'exploration et l'exploitation.

— Les algorithmes génétiques (AG) sont parmi les premières métaheuristiques utilisées pour résoudre le problème de la sélection d'attributs [**Bala.J et al.**]. En effet, les entités peuvent être utilisées comme population initiale pour Les algorithmes génétiques, de telle sorte que chaque sous-ensemble d'entités est composé de valeurs binaires (1 ou 0). La valeur 1 signifie que la ième entité est

incluse dans le sous-ensemble d'entités et 0 signifie que l'entité n'est pas incluse. Pour mesurer l'adéquation d'un sous-ensemble actuel de fonctionnalités, l'algorithme ID3 est utilisé pour calculer la précision de la classification qui a été utilisée comme fonction de fitness pour l'algorithmes génétique. Les résultats obtenus ont montré que la combinaison de L'algorithme génétique et ID3 donne de meilleurs résultats que les approches traditionnelles de classement des attributs.

Des algorithmes génétiques ont également été utilisés pour la génération automatique de clés dans d'autres domaines tels que la cryptographie [**S et al.**].

— une nouvelle approche cryptographique basée sur l'ADN et l'algorithme génétique pour l'étape de génération de clé. Ils ont choisi d'utiliser une technique de croisement K-Point avec une fonction de fitness qui mesure le caractère aléatoire de chaque clé générée.

— L'algorithme Firefly (FFA) est une autre méta-heuristique utilisée dans le domaine de la sélection d'attributs [E et al.]. Dans cette approche, le FFA est utilisé pour sélectionner les meilleurs attributs et le "K-nearest neibor" est utilisé comme classificateur pour minimiser une fonction de fitness donnée. Cette approche a ensuite été améliorée en proposant un FFA récursif où les caractéristiques les mieux sélectionnées lors de la première exécution de FFA, forment un nouveau dataset comprenant uniquement ces caractéristiques les mieux sélectionnées [N et Z]. Ensuite, la FFA pour la sélection d'entités est exécutée sur le nouveau dataset pour rechercher une meilleure solution. Les résultats obtenus ont montré une amélioration en termes de précision.

D'autres approches combinaient au moins deux méta-heuristiques pour améliorer à la fois l'exploration et l'exploitation.

— [**Nagaveni et N.**] ont proposé une solution hybride utilisant l'algorithme d'optimisation des colonies de fourmis (ACO) et l'algorithme Cuckoo Search (CS) pour la sélection d'attributs.

— [**N et al.**] ont essayé trois approches méta-heuristiques hybrides différentes pour voir comment l'hybridation pouvait améliorer la sélection des caractéristiques (i) le premier, les auteurs ont introduit la fonction de croisement de l'algorithme GA dans l'algorithme d'optimisation de la recherche Penguins (PeSOA). Le but était d'éviter la convergence rapide du PeSOA.(ii) La deuxième approche hybride était une combinaison des algorithmes Firefly (FA) et Deferential Evolution (DE) [L et al., b]. Le but était d'utiliser l'avantage de la recherche locale donnée par l'algorithme DE. (iii) Le troisième test de l'approche hybride a consisté à introduire les fonctions de mutation et de clonage du système immunitaire artificiel pour créer de nouvelles chauves-souris pour la phase d'initialisation de l'algorithme BAT. Les résultats obtenus de ces travaux ont montré que l'utilisation d'une seule méta-heuristique fonctionne mieux que les approches hybrides, notamment en ce qui concerne la grande complexité des approches d'hybridation. Les dernières méthodes sont les méthodes intégrées qui combinent des filtres et des wrappers

2.7 Conclusion

On peut déduire à partir des approches discutées que :

- La plupart des approches de blocage dépendent de la sélection initiale de la clé de blocage.
- proposer une solution de sélection de clé de blocage automatique est l'une des priorités les plus importantes de la communauté RL.
- La communauté Record Linkage a proposé plusieurs approches de sélection automatique des clés de blocage.
- Le problème de la sélection des clés de blocage peut être considéré comme un problème de sélection de fonctionnalités où toutes les clés de blocage possibles sont supposées être générées et l'objectif est de sélectionner le meilleur sous-ensemble de ces clés qui peut accélérer les performances de l'approche RL .
- Plusieurs algorithmes méta-heuristiques ont été utilisés pour résoudre le problème de la sélection des attributs . Certains d'entre eux ont utilisé directement une méta-heuristique et d'autres ont essayé de combiner deux ou plusieurs métaheuristiques

Chapitre 3

Expérimentation

Chapitre 3

Expérimentations

3.1 Introduction

Dans ce chapitre, on va présenter un état de l'art sur les approches existent dans la littérature qui ont accordé une attention au couplage d'enregistrements et un aperçu des solutions existantes proposées dans la littérature et ayant traité le problème de la sélection automatique des clés de blocage.

3.2 Couplage d'enregistrements basé sur les K-modes

Dans cette méthode, les données sont regroupées en blocs en utilisant uniquement les clés de blocage comme attributs de clustering, qui sont dans notre cas les fonctionnalités sélectionnées.

3.2.1 Présentation

K-Modes a été proposé pour la première fois en 1998 par HUANG [**HUA-15-**]. Cet algorithme est considéré comme une extension de l'algorithme de clustering classique K-Means, il a été proposé afin de regrouper des données catégorielles ce qui n'est pas le cas avec l'algorithme K-Means qui n'accepte que des attributs numériques. Bien sûr, il existe l'algorithme de clustering hiérarchique classique qui traite à la fois des données catégorielles et numériques, mais sa complexité quadratique le rend inadapté au clustering de grands ensembles de données. L'utilisation de K-Modes dans notre cas nous a permis d'éliminer l'étape de conversion de données numériques qui prenait du temps et était une étape nécessaire à faire avec l'algorithme k-means. L'algorithme ce base sur trois points principaux: (i) Mesure de dissimilarité simple (démontrée dans l'équation 1) afin de faire correspondre les objets. (ii) Utilisation de modes à la place des moyens et (iii) Méthode basée sur la fréquence pour mesurer le mode d'un ensemble.

Pour la sélection des modes initiaux, deux techniques sont utilisées. La première permettre à affecter les K premiers enregistrements distincts comme modes initiaux. La seconde consiste à mesurer les fréquences pour toutes les catégories de chaque attribut et à les trier par ordre décroissant en fonction de leurs fréquences. Une fois cela fait, nous attribuons les premières catégories fréquentes aux premiers k-modes initiaux.

3.2.2 Principe d'utilisation

Dans notre travail, nous utilisons l'algorithme K-Modes pour regrouper les données en blocs. Chaque bloc contiendra des correspondances possibles.

Le clustering n'utilisera que les clés de blocage générées à l'étape précédente comme attributs de clustering, au lieu d'utiliser tous les attributs de l'ensemble de données, afin de gagner du temps et parce que les clés de blocage contiennent les informations les plus importantes sur les enregistrements.

Une fois le clustering terminé et avant de passer à l'étape de correspondance, nous exécutons le filtrage adaptatif présenté dans []. Les auteurs ont proposé une nouvelle approche afin de réduire le nombre de comparaison de paires d'enregistrements une fois le blocage effectué, en ignorant la comparaison entre les paires qui sont dans le même cluster mais sont considérées comme des correspondances improbables. Cette approche a été proposée après avoir observé que toutes les méthodes de blocage peuvent générer de très gros blocs.

La première technique de filtrage utilisée dans cette approche est le «filtrage long», où deux enregistrements sont déclarés comme des correspondances peu aimables si la différence entre la longueur de leur variable de filtrage est supérieure à une valeur prédéfinie K.

La deuxième technique est le filtrage par comptage où chacune des variables de filtrage des deux enregistrements comparés est divisée en un ensemble de bi-grammes, puis le nombre de bi-grammes communs entre les deux chaînes est comparé à C_{min} où C_{min} = max(s1; s2) - 2k + 1 avec k représente la distance d'édition entre les deux variables de filtrage. Si le nombre de bi-grammes communs est inférieur à Cmin, les enregistrements sont déclarés comme des correspondances peu aimables.

Les expériences sur des ensembles de données du monde réel et sur des ensembles de données synthétiques [], ont montré que les techniques de filtrage ont réduit le nombre de comparaisons à 80% par rapport au blocage traditionnel même lorsqu'il s'agit de petits blocs.

La dernière étape de notre approche consiste à faire correspondre les paires d'enregistrements dans chaque cluster. Nous avons choisi d'utiliser un ensemble de métriques de similarité basées sur les caractères car elles sont conçues pour traiter des erreurs typographiques [8], ce qui est le cas des jeux de données les plus réels. Plusieurs métriques pour la correspondance de chaînes existent dans la littérature ont été évaluées (distance d'édition, similitude Jaro Winkler, similitude Jaccard, distance Smith-Waterman, Q-Grams et plus).

3.3 Algorithme d'optimisation de recherche de BES (Bald Eagle Search) pour la sélection automatique des clés de blocage

3.3.1 Présentation

L'algorithme d'optimisation BES (Bald Eagle Search) a été récemment proposé pour résoudre les problèmes d'optimisation. Il a été construit en utilisant à la fois des techniques d'essaim et d'évolution (**Alsattar et al., 2019**).

Notre objectif est d'adapter l'algorithme BES au problème de sélection de clé de blocage. Ce dernier peut en effet être modélisé comme un problème de sélection de caractéristiques. La population initiale est un groupe de sous-ensembles de fonctionnalités, c'est-à-dire des clés de blocage. Ainsi, chaque membre d'une population représente un sous-ensemble concurrent de clés de blocage. De plus, les sous-ensembles n'ont pas nécessairement les mêmes tailles ou éléments. Les populations sont régénérées à l'aide de l'algorithme BES.

L'aptitude de chaque membre d'une population est calculée en utilisant l'approche RL proposée dans (Benkhaled et al., 2019) de manière globale. Dans la méthode utilisée, K-Modes est utilisé comme étape d'indexation en regroupant les données en utilisant uniquement les clés de blocage qui sont le sous-ensemble de fonctionnalités actuellement sélectionné.

3.3.2 Principe d'utilisation

Les meilleures clés de blocage en termes de fitness sont celles dans lesquelles K-Modes regroupe les enregistrements les plus dupliqués lorsqu'ils sont utilisés comme attributs de clustering. En conséquence, la fonction de fitness est le paramètre de complétude de la paire (PC). Le PC mesure le nombre de doublons détectés par une approche RL en utilisant les touches de blocage sélectionnées.

La méthode utilisée peut être résumée par les points suivants :

• Générer toutes les listes de clés de blocage possibles.
• Initialisez la première population qui est un sous-ensemble d'entités aléatoires de la liste des clés de blocage précédemment générée.
• Exécutez l'algorithme BES pour les itérations T sur la population précédemment générée dans une méthode wrapper avec l'exhaustivité de la paire comme fonction de fitness.
• Le meilleur membre de la dernière population est sélectionné comme meilleur sous-ensemble d'entités à utiliser comme clés de blocage.

A) Prétraitement

La liste des clés candidates est celle à partir de laquelle la population initiale de l'algorithme BES sera sélectionnée au hasard. Avant de générer la liste des clés candidates, une étape essentielle de prétraitement ne peut être négligée. Il s'agit, en fait, de nettoyer l'ensemble A. en d'autres termes, il faut éliminer les attributs de mauvaise qualité de l'ensemble A. Deux paramètres sont utilisés pour calculer la qualité globale d'un attribut. Premièrement, l'exhaustivité représente le pourcentage de valeur nulle concernant les attributs spécifiés (**Pipino et al., 2002**). Nous avons utilisé la mesure NBC (complétude basée sur zéro) où la complétude est mesurée à l'aide de l'équation (1). En utilisant cette méthode, la valeur 1 représente le meilleur résultat et 0 le pire. Tous les attributs qui ont une valeur d'exhaustivité inférieure au seuil prédéfini sont éliminés à partir de la génération de la liste des clés de blocage candidates.

$$\mathbf{Completness}(\boldsymbol{Attj}) = \mathbf{1} - \frac{\boldsymbol{nombre\ de\ valeurs\ nulles\ dans\ Attj}}{\boldsymbol{Nombre\ d'instances}} \qquad (\mathbf{1})$$

Le deuxième paramètre est la cardinalité d'un attribut. La cardinalité représente le nombre de valeurs distinctes pour un attribut spécifié. Dans le processus RL, les attributs à très faible cardinalité ne conviennent pas pour être utilisés comme clés de blocage. Par exemple, l'utilisation de l'attribut sex comme clé de blocage divise les données en seulement 2 blocs (M / F). Par conséquent, dans notre approche, les attributs à très faible cardinalité sont éliminés de la génération de la liste des clés de blocage candidates.

B) Génération de clés de candidat

Une fois que les attributs de mauvaise qualité sont éliminés; pour chaque ensemble de données D, différentes clés de blocage peuvent être générées en fonction du domaine de l'ensemble de données et du type d'attributs. Nous avons utilisé un ensemble de fonctions F pour générer des clés candidates telles que First4Chars (Attributes), Concatenation (), Soundex (Attribute), Last4Chars (Attribute) et NYSIIS (Attribute).

BK	Name	Address	City	Phone	TYpe
Losangelos310/246- 1501	Amine morton's	435 s.la ceinega blv	Los angelos	310/246-1501	American
Studiocity818 /762- 1221	Art's delicatessen	12224 ventura blvd	Studio city	818 /762-1221	American

Tableau 3.1 : Exemple clés de blocage

Le tableau 2 présente certaines des différentes fonctions utilisées pour générer la liste des clés de blocage possibles. D'autres fonctions spécifiques ont été utilisées pour chaque ensemble de données ne sont pas mentionnées dans le tableau. Par exemple, «Extract-Number ()» est une fonction utilisée pour extraire le numéro du restaurant du champ d'adresse dans le cas du jeu de données du restaurant.

Fonction	Description
Soundex (attribut), **NYSIIS (attribut)**	Soundex et NYSIIS sont tous deux des algorithmes de codage phonétique (Holmes et McCabe2002) qui transforment une chaîne en une présentation alphanumérique de la façon dont elle est prononcée .
First_N_Chars (attribut)	Extrayez les N premiers caractères d'un champ d'attribut.
Last_N_Chars (attribut)	Extrayez les N premiers caractères d'un champ Attributaire.
Numérique (attribut)	Extrayez la valeur numérique d'une chaîne.
Remove-SP (attribut)	Supprimez les caractères spéciaux d'une chaîne.
Valeur exacte (attribut)	Utilisez la valeur d'attribut sans modification.

Tableau 3.2 Les fonctions utilisées pour la génération des clés de blocage.

A- Soundex

Soundex est considéré comme l'une des fonctions d'encodage phonétique les plus efficaces. Il transforme les chaînes en fonction de leur prononciation afin qu'elles puissent être comparées les unes aux autres sans tenir compte des fautes d'orthographe. En utilisant Soundex, des noms comme ALLAN et ALLEN sont tous deux représentés avec le même code "A450", ce qui facilite la correspondance entre les deux noms. Les principales étapes de Soundex sont :

— Conservez la première lettre de la chaîne.
— Remplacez toutes les consonnes en suivant les règles suivantes : (**0** pour le caractères A, E, H, I, O, U, W, Y. **1** pour les caractères B,F,P,V. **2** pour C,G,J,K,Q,S,X,Z. **3** pour D,T. **4** pour L et **5** remplace M,N. **6** remplace le caractère R.
— Dans le cas où la chaîne est trop courte, l'algorithme complète les trois nombres après le premier caractère par des zéros.

b- **NYSIIS** (New York State Identification Intelligence System) NYSIIS a la même idée et le même objectif que l'algorithme Soundex. La différence est que NYSIIS renvoie un code composé de lettres ce qui n'est pas le cas de Soundex. L'algorithme NYSIIS augmente la précision de 2,7 % par rapport à Soundex [Rajkovic & Jankovic 2007]. Les règles de base de l'algorithme NYSIIS sont la transformation des premiers caractères où :(MAC est remplacé par MCC et KN devient NN, K en C, PH-PF en FF, SCH en SSS) et les derniers caractères (EE- IE à Y, DT-RT-RD-NT-ND à D).

3.3.3 Bald Eagle Searche Pseudo Code

```
Algorithm 1. BES for feature selection.
Inputs: Dataset D, Number of iterations T, Number of
          solutions M, BES parameters: c1, c2,α, R.
Outputs: The best subset of blocking keys.
Function BES
    Initialize a random population of M member.
    Initialize the best solution: Pbest = All features.
    While (i <= T) do
    For each (member P in pop)) do
              Pnew = Calculate (Pnew) using equation 1.
          IF (f (Pnew) > F (P)) do
               P = Pnew
          IF (f (Pnew) > F (Pbest)) do
               P = Pnew
          End IF
          End IF
          End For
   For each (member P in pop)) do
              Pnew = Calculate (Pnew) using equation 2.
          IF (f (Pnew) > F (P)) do
               P = Pnew
          IF (f (Pnew) > F (Pbest)) do
               P = Pnew
          End IF
          End IF
          End For
        For each (member P in pop)) do
              Pnew =Calculate (Pnew) using equation 3.
          IF (f (Pnew) > F (P)) do
               P = Pnew
          IF (f (Pnew) > F (Pbest)) do
               P = Pnew
          End IF
          End IF
          End For
     End While.
     Return Pbest.
End Function.
```

La première étape consiste à générer la population initiale qui est un sous-ensemble sélectionné au hasard de clés de blocage à partir de la liste générée précédemment. Ensuite, nous commençons la phase de sélection d'espace (lignes 8-16), où pour chaque membre de la population, nous calculons un nouvel objet (Pnew) à partir de celui-ci en utilisant l'équation 2. α est une variable qui prend une valeur comprise entre 1,5 et 2. R est un nombre aléatoire généré avec R ∈ [0,1] et Pi désigne la position actuelle (Alsattar et al.2019).

Une fois Pnew calculé, ses nouvelles fonctionnalités obtenues sont nettoyées, toutes les valeurs continues sont arrondies en nombres entiers et tous les nombres entiers qui sont en dehors de la plage de [0,N] (N est le nombre de clés de blocage générées) sont remplacés par un fonction aléatoire (index de clé de blocage).

L'étape suivante consiste à mesurer la forme physique à l'aide des fonctionnalités de Pnew et à la comparer aux performances de P et de Pbest si elle est meilleure qu'eux. Ensuite, ils sont tous les deux remplacés par Pnew.

Dans la phase de recherche spatiale (lignes 17-25), pour chaque membre de la population, un nouvel objet Pnew est calculé à l'aide de l'équation 3. X et Y sont deux nombres aléatoires pour représenter le mouvement en spirale de l'aigle vers la zone de descente. Enfin, le swooping (lignes 26-34), c'est là que l'aigle cherche sa cible finale en spirale. Pour chaque objet de la population, Pnew est calculé à l'aide de l'équation 4.

$$\boldsymbol{P_{new} = P_{best} + \alpha * r * (P_{mean} - P_i)} \quad \mathbf{2}$$

$$\boldsymbol{P_{new} = P_{best} + y(i) * (P_i - P_{i+1}) + x(i) * (P_i - P_{mean})} \quad \mathbf{3}$$

$$\boldsymbol{P_{new} = rand * P_{best} + x(i) * (P_i - C_1 * P_{mean}) + y(i) * (P_i - C_2 * P_{best})} \quad \mathbf{4}$$

Dans l'ensemble, l'algorithme général d'optimisation de la recherche du pygargue à tête blanche peut être résumé dans l'organigramme suivant (Figure 3.)

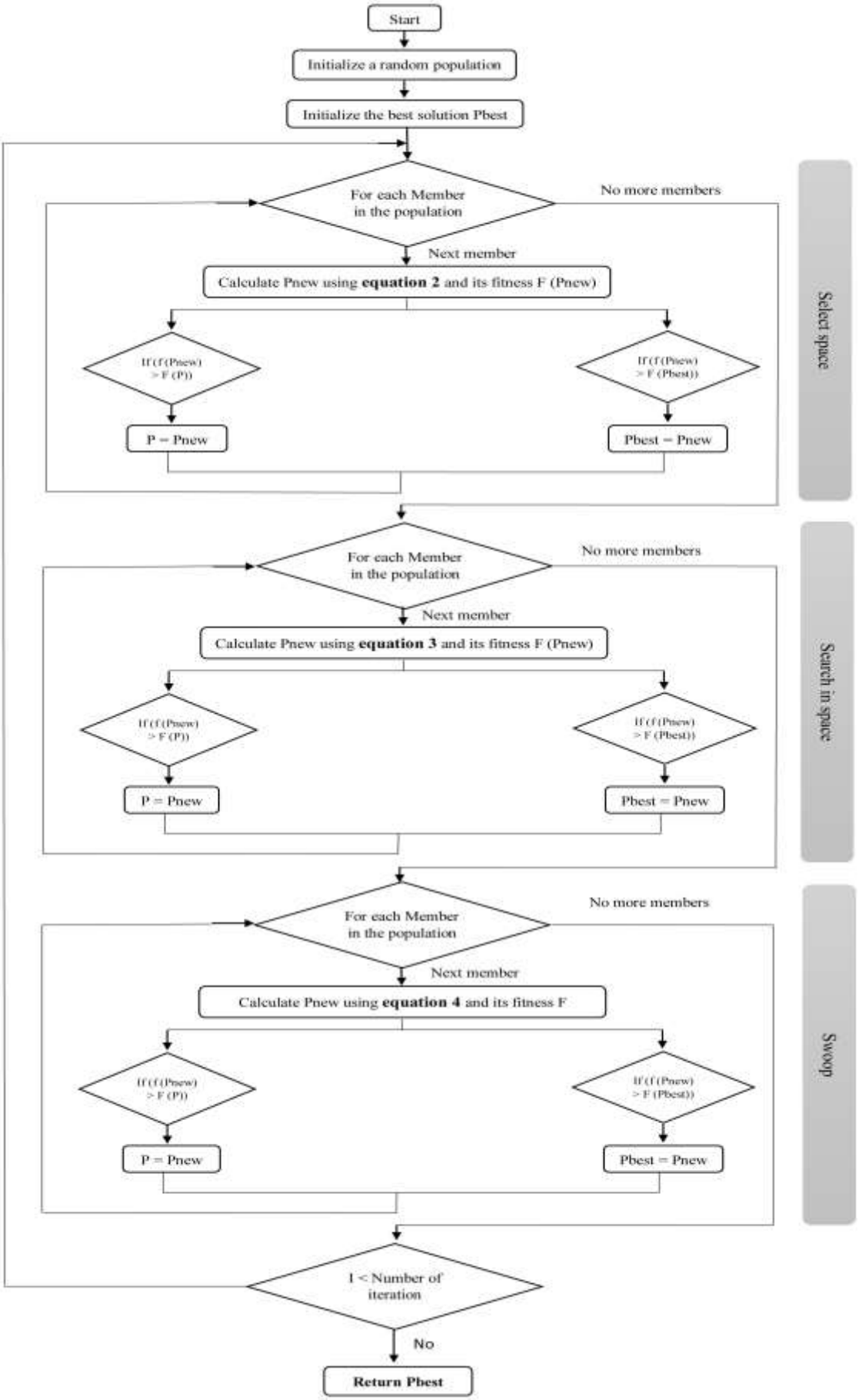

Figure 3.1 Diagramme de flux BES

Une fois le regroupement effectué, la correspondance entre les enregistrements d'un même cluster est effectuée à l'aide d'une métrique de similarité de chaîne telle que la similarité de Jaro-Winkler. Le nombre de valeurs en double détectées est exprimé à l'aide du paramètre Pair Completeness (PC) (équation 5). PC, dans ce cas, est utilisé comme fonction de fitness dans l'algorithme BES.

$$\boldsymbol{PC} = \frac{nombre\ de\ paires\ d'enregistrements\ détectées.}{nombre\ de\ paires\ d'enregistrements\ en\ double\ dans\ dataset} \qquad \mathbf{5}$$

D'autres mesures sont utilisées pour mesurer la performance d'une approche de couplage d'enregistrements. Le taux de réduction (RR) (équation 6) est utilisé pour mesurer dans quelle mesure la technique de blocage a réussi à réduire le nombre de comparaisons. La mesure F (équation 7) est utilisée pour contrôler le compromis entre RR et PC.

$$\boldsymbol{RR} = 1 - \frac{nombre\ de\ paires\ d'enregistrements\ détectées.}{nombre\ de\ paires\ d'enregistrements\ en\ double\ dans\ dataset} \qquad \mathbf{6}$$

$$\boldsymbol{f_{Measure}} = 2 * \frac{RR * PC}{RR + PC} \qquad 7$$

3.4 Implémentation et expérimentation

3.4.1 Environnement de travail

A) Environnement matériel

- Processeur : 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 2.42 GHz
- Mémoire installée (RAM) Mémoire installée (RAM) : 8,00 Go
- Système d'exploitation : Windows 11 Famille Unilingue

B) Langage de programmation

Parmi les différents langages de programmation existant dans le monde de développement, nous avons choisi le JAVA. C'est un langage de programmation orienté objet, développé par Sun Microsys-tems. Il permet de créer des logiciels compatibles avec de nombreux systèmes d'exploitations (Windows, Linux, Macintosh, Solaris). Java donne aussi la possibilité de développer des programmes pour téléphones portables et assistants personnels. Enfin, ce langage peut être utilisé sur internet pour des petites applications intégrées à la page web (applet) ou encore comme langage serveur (jsp).

La technologie Java est indissociable du domaine de l'informatique

C) Environnement logiciel

Eclipse est un IDE (un environnement de développement intégré) conçu avec des fonctionnalités permettant de simplifier le développement d'applications Java. Cet IDE est réputé multi-langage, multiplateforme et extensible par des greffons ou plug-ins. Il est avant tout conçu pour le langage Java, mais ses nombreux greffons en font un environnement de développement de choix pour de nombreux autres langages de programmation (C/C++, Python, PHP, Ruby. . .).

Toutes les fonctionnalités qu'on peut attendre de ce genre de logiciel sont présentes ou existent sous forme de greffons (coloration syntaxique, complétion, débugger, gestion de projets, intégration aux gestionnaires de versions. . .). Eclipse est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X et Open VMS. Eclipse est lui-même développé en Java, ce qui peut le rendre assez lent et gourmand en ressources mémoires 3.

Une collection d'algorithmes d'apprentissage automatique pour les tâches d'exploration de données. Il contient des outils pour la préparation des données, la classification, la régression, la mise en cluster, l'exploration de règles d'association et la visualisation. Weka est un logiciel open source distribué sous licence GNU General .

3.4.2 Testes et résultats

Pour le teste nous avons utilisé un célèbre dataset du monde réel "Restaurant" et "datasets arabe". Ces datasets sont généralement utilisé pour évaluer l'efficacité des méthodes de couplage d'enregistrements **[J et Q]**, **[Kejriwal et Miranker]**, **[Christen, b]**, **[H et al., a]**,**[H et al., b]**. Le dataset restaurant contient des enregistrements sur les restaurants collectés auprès des guides Fodor et Zagat. Le tableau 3.2 résume les informations sur les ensembles de données utilisés.

Datasets	Type	Les enregistrements	Nombre attributs
Restaurant	**Déduplication**	**864**	**5**
Dataset	**Déduplication**	**3000**	**3**

Tableau 3.3 – datasets utilisé dans les tests.

On a commencé par tester le dataset « restaurant ». Nous avons d'abord utilisé trois clés de blocage sélectionnées manuellement qui sont (BK1 : Soundex(named)+phone_number, BK2 : NYSIIS(City), BK3 : phone_number).

Les résultats obtenus sont très bons, le jeu de données du restaurant comprend 112 valeurs en double. En général, les résultats obtenus sont satisfaisants. Maintenant on passe au traitement de notre problème principal qui est la sélection automatique des clés de blocage. Comme nous l'avons mentionné ci-dessus, les résultats obtenus précédemment ont été obtenus à l'aide de trois clés de blocage sélectionnées manuellement.

La question est maintenant : existe-t-il une autre combinaison de clés de blocage qui peut nous donner de meilleurs résultats ? Pour cela, il faut réfléchir à une approche qui sélectionne automatiquement les clés de blocage sans l'intervention des humains. Pour cela on a utilisé l'algorithme BES de méta-heuristiques pour faire face à ce problème.

Figure 3.2 Interface principal de l'application

La première série d'expériences a été réalisée sur le dataset restaurants, il contient cinq attributs sur chaque restaurant (nom, destinataire, ville, téléphone et type).

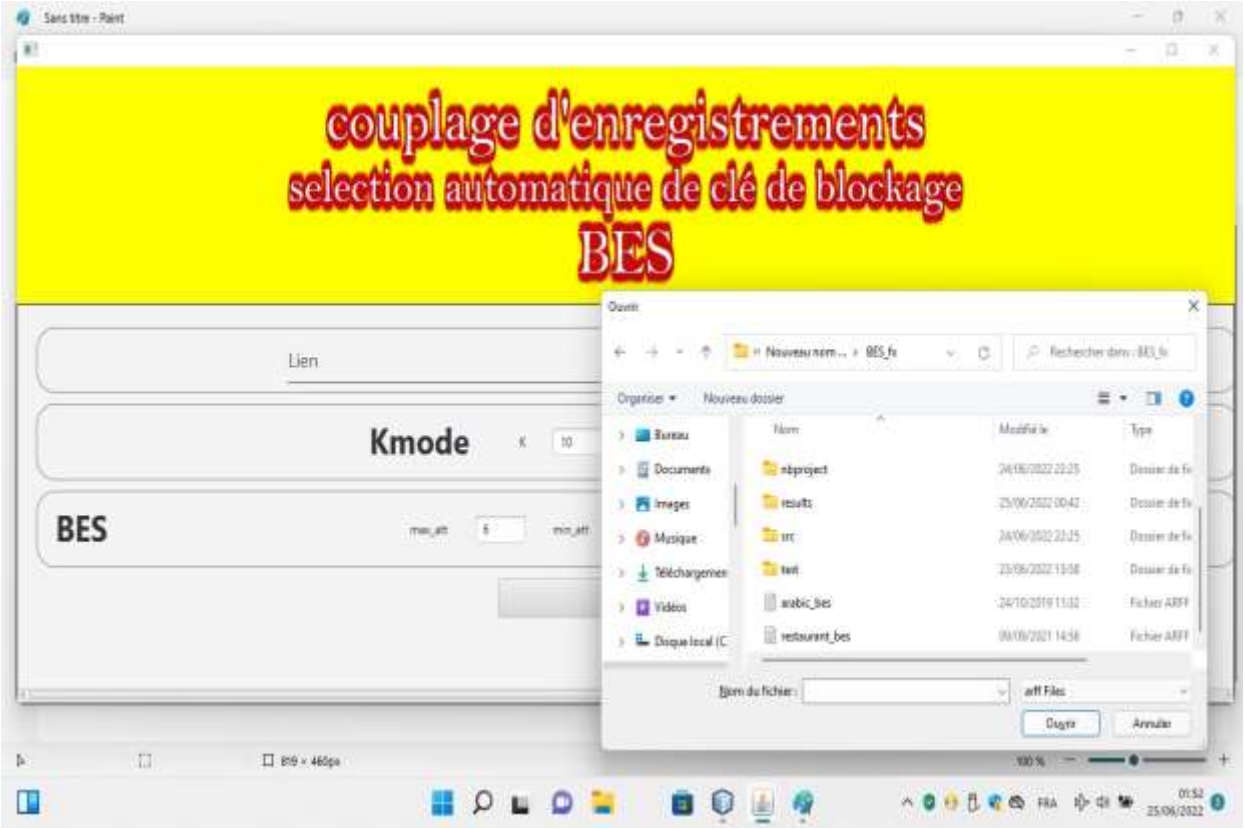

Figure 3.3 Chargement dataset restaurant

Plusieurs clés de blocage possibles ont été générées, y compris {[name_phone_id], [city_phone], [name], [phone]}, {[city_phone], [name], [city]} , {[name_phone_id], [First2Chars (name_ phone_id) + First2Chars (city_phone)],[First2Chars (name) + First2Chars (addr)], [First2Chars (city) +[First2Chars (phone)], [First3Chars (name_phone_id) + [First3Chars (city_phone)]+ [First3Chars (city)]}. Les clés sont générées à l'aide de la fonction **Soundex** et NYSIIS abordées dans la section précédente.

Figure 3.4 Les Blocs (Clusters) génères pour dataset restaurant

Figure 3.5 correspondance (Matching) pour dataset restaurant

Itérations	Clés de blocage sélectionnées
2	{[city_phone],[phone_id], [name],[addr],[city]}
4	{[First4Chars(City)]+[First4Chars(Phone)], [Soundex(City)+Phone number], [Name]}
8	{[Phone Number], [First4Chars (Name) + First4Chars (ZipCode)], [First4(Name)+First4(Type)], [Soundex (City)]}

Tableau 3.4 – les meilleures clés de blocage sélectionnées pour dataset restaurant

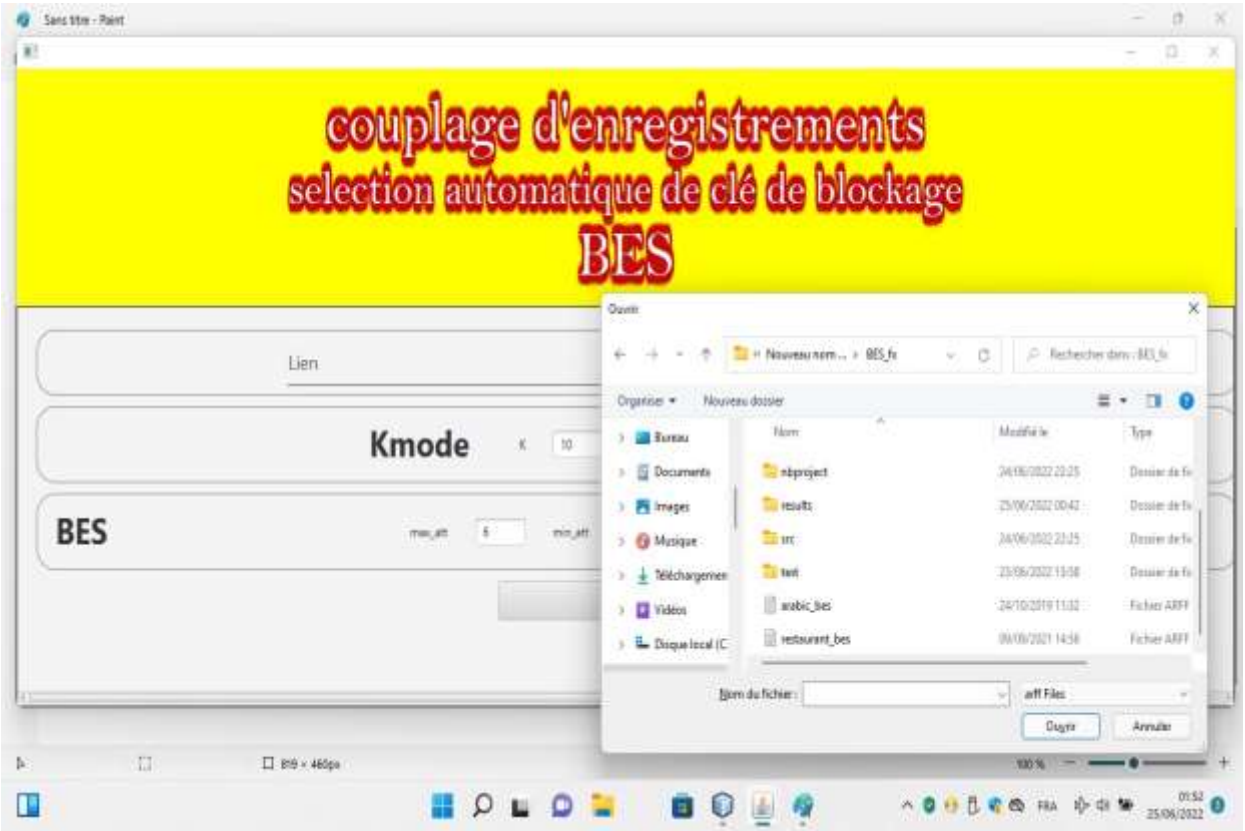

Figure 3.6 Chargement dataset arabe

Figure 3.7 Les Blocs (Clusters) génères pour dataset arabe

Figure 3.8 correspondance (Matching) pour dataset arabe

3.5 Conclusion

Dans ce chapitre, nous avons présenté l'étude expérimentale qui a été réalisée pour évaluer les algorithmes présents.

Des résultats encourageants sont obtenus avec l'expérience concernant le couplage d'enregistrements basé sur les K-Modes dans l'etape d'indexation (création des blocs).

Les résultats obtenus sont prometteurs et ont également montré l'efficacité de l'utilisation de l'algorithme Bald Eagle Search dans la sélection automatique des clés de blocage, même avec un faible nombre d'itérations.

Conclusion

Générale

Conclusion Générale

Notre étude est concentrée sur la sélection des attributs, plus particulièrement, la sélection des clés de blocage qui est un élément essentiel dans ce domaine. Nous avons utilisé l'algorithme BES pour la sélection automatique des clés de blocage.

L'avantage majeur de l'algorithme k-modes réside dans le fait qu'il traite directement avec les données catégorielles, nous n'avons donc pas à les convertir en données numériques.

L'algorithme BES a finalement été appliqué à datasets du monde réel. Les résultats ont montré une amélioration par rapport à la sélection manuelle, montrant l'applicabilité de l'algorithme proposé dans la résolution de problèmes réels.

Bibliographie

Bibliographie

Ahmed.K Elmagarmid, Panagiotis.G Ipeirotis, et Vassilios S Verykios. Duplicate record detection : A survey. IEEE Transactions on knowledge and data engineering, 19(1) :1–16,.

Aizawa. A et Oyama. K. A fast linkage detection scheme formultisource information integration. Dans In null, page 30–39. IEEE.

Alian M, Awajan A, et Ramadan B. Unsupervised learning blocking keys technique for indexing arabic entity resolution. International Journal of Speech Technology, 22(3) :621–628, a.

Andrew McCallum, Kamal Nigam, et Lyle H Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. Dans Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, page 169–178. ACM.

Bala.J, Huang. J, Vafaie. H, et DeJong. K et Wechsler. H. Hybrid learning using genetic algorithms and decision trees for pattern classification. in IJCAI, 1 :719–724.

Bassour Boumediene et Abbar Riadh Wassim. Dtectionde doublons. Thése de master Universit Djilali Liabes ,Sid Bel Abess, 2015/2016.

Benkhaled H.N, Berrabah D, et Boufares F. A novel approach to improve the record linkage process. Dans En *2019* IEEE *6*th International Conference on Control, Decision and Information Technologies, page 1504–1509. IEEE.

Bilenko M, Kamath B, et Mooney R.J. Adaptive blocking : Learning to scale up record linkage. in sixth international conference on data mining. Dans Sixth International Conference on Data Mining (ICDM'*06*, page 87–96. IEEE, b.

Chandrashekar G et Sahin F. An introduction to variable and feature selection, guyon, isabelle and elisseeff, andré,. Journal of machine learning research, 3(1) :16–28. numéro : mars, pages : 1157–1182, année : 2003. Informatique et génie électrique,.

David Holmes et M.Catherine McCabe. Improving precision and recall for

soundex retrieval. Dans Information Technology : Coding and Computing, *2002*. Proceedings. International Conference on, page 22–26. IEEE.

Dif N et Elberrichi Z. An enhanced recursive firefly algorithm for informative gene selection. International Journal of Swarm Intelligence Research (IJSIR, 10(2) :21–33.

Dif N, Attaoui M, et Elberrichi Z. Gene selection for microarray data classification using hybrid meta-heuristics. Dans International Symposium on Modelling and Implementation of Complex Systems,

Emary E, Zawbaa H.M, Ghany K.K.A., et A.E.and Pârv B Hassanien.

Firefly optimization algorithm for feature selection. Dans Actes de la 7e Conférence des Balkans sur l'informatique, page 26. ACM.

Fleuret F. Fast binary feature selection with conditional mutual information. Journal of Machine learning research, 5(novembre) : 1531–1555.

Franck Rgnier-Pcastaing, Michel Gabassi, et Jacques Finet. Enjeux et mthodes de la gestion des donnes. Paperback, 2008.

Gravano L, Ipeirotis P.G, Jagadish H.V., Koudas N, Muthukrishnan S, et Srivastava D. Approximate string joins in a database (almost) for free. VLDB, 1 :491–500, a.

Guyon I et Elisseeff A. An introduction to variable and feature selection. Journal of Machine learning research, 3(mars) :1157–1182.

Hernández M.A et Stolfo S.J. The merge/purge problem for large databases. Dans ACM Sigmod Record, 24 :127–138.

Ivan P.Fellegi et Alan B.Sunter. A theory for record linkage. Journal of the American Statistical Association, 64(328) :,1183–1210,.

Jamm. DES DONNES QUALIT :Exploitez le capital de votre organisation. livre blanc, janvier 2008.

Jona J.et Nagaveni et N. Ant-cuckoo colony optimization for feature selection in digital mammogram. Journal pakistanais des sciences biologiques, 17(2) :266.

Kalsi S, Kaur H, et Chang V. Dna cryptography and deep learning using genetic algorithm with nw algorithm for key generation. Journal of medical

systems, 42(1) :17.

Köpcke H, Thor A, et Rahm E. Evaluation of entity resolution approaches on real-world match problems. Proceedings of the VLDB Endowment, 3(1-2) :484–493, a.

Köpcke H, Thor A, et Rahm E. Learningbased approaches for matching web data entities. IEEE Internet Computing, 14(4) :23–31, b.

Laure Berti-Equille. Qualité des données. Techniques de l'ingénieur. Informatique, 2006.

Matthew A.Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. Journal of the American Statistical Association, 84(406) :414–420,.

Mauricio A Hernàndez et Salvatore J Stolfo. The merge/purge problem for large databases. Dans ACM Sigmod Record, volume 24, page 127–138. ACM.

MichelsonMet Knoblock C.A. Learning blocking schemes for record linkage. AAAI, 6 :440–445.

Muro C, Escobedo Rand Spector L, et Coppinger R. Wolf-pack (Canis lupus) hunting strategies emerge from simple rules in computational simulations. Behav Processl, volume 88.

Nascimento D.C, Pires C.E.S., et Mestre D.G. Exploiter la cooccurrence de bloc pour contrôler la taille des blocs pour la résolution d'entité. Knowledge and Information Systems, *62*(*1*), *359-400*, 62(1) : 359–400. page 119–132. Springer.

Peter Christen. A comparison of personal name matching : Techniques and practical issues. Dans Data Mining Workshops, *2006*. ICDM Workshops *2006*. Sixth IEEE International Conference on, page 290–294. IEEE, a.

Peter Christen. A survey of indexing techniques for scalable record linkage and deduplication. IEEE transactions on knowledge and data engineering, 24(9) :,1537–1555,, b.

Peter Christen. Towards parameter-free blocking for scalable record linkage, c.

Pipino L.L, Lee Y.W, Wang, et R.Y. Data quality assessment. Communications of the ACM, 45(4) :211–218.

Ramadan B et Christen P. Unsupervised blocking key selection for real-time entity resolution. Dans Pacific-Asia Conference on Knowledge Discovery and Data Mining, page 574–585. Springer.

Shao J et Wang Q. Active blocking scheme learning for entity resolution. Dans Pacific-Asia Conference on Knowledge Discovery and Data Mining, page 350–362, Cham. Springer.

Su Yan, Dongwon Lee, Min-Yen Kan, et Lee C Giles. Adaptive sorted neighborhood methods for efficient record linkage. Dans Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, page 185–194. ACM.

TN Gadd. Phonix : The algorithm. Program, 24(4) :363–366,.

Tobias Vogel et Felix Naumann. Automatic blocking key selection for duplicate detection based on unigram combinations. Dans Proceedings of the International Workshop on Quality in Databases (QDB.

OUHAB Abdelkrim, MALKI Mimoun, BERRABAH Djamel, et BOUFARES Faouzi. An unsupervised entity resolution framework for english and arabic datasets. International Journal of Strategic Information Technology and Applications (IJSITA, 8(4) :16–29,.

Yang Y, Zheng X, Guo W, Liu X, et Chang V. Privacy-preserving smart iot-based healthcare big data storage and self-adaptive access control system. Information sciences, 479 :567–592.

Webographie

1. https ://www.futura-sciences.com/tech/definitions/internet-java-485/

2. https ://bit.ly/2I9RGeg

3. https ://www.techno-science.net/definition/5346.html

4. https ://www.cs.waikato.ac.nz/ml/weka/