République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Borj Bou Arréridj
Faculté des Mathématiques et d'Informatique
Département d'informatique

MEMOIRE

Présenté en vue de l'obtention du diplôme

Master en informatique

Spécialité : Réseaux & Multimédia

THEME

Une approche améliorée d'appariement de données

Présenté par :

Mehennaoui Nadhir

Naili Moussa

Soutenu publiquement le : 19/06/2024

Devant le jury composé de:

Président : MOUSSAOUI ALI

Examinateur : BOUMAZA FARID

Encadreur : BENMSSAHEL ILYAS

2023/2024

Dédicace

À ce précieux mémoire qui couronne mon parcours de Master 2, symbole de mes efforts, de mes découvertes et de mon engagement. Que chaque page soit le témoignage de mon apprentissage et de ma passion pour ce domaine. Merci à tous ceux qui m'ont soutenu et inspiré durant ce voyage académique.

Remerciement

Avant tout nous tenons à remercier le dieu le tout puissant de nous avoir aidé et nous donné santé et courage et patiente afin de réaliser ce travail.

Nous tenons à remercier et exprimer notre profonde gratitude et reconnaissance envers notre encadreur ***Mr ILYAS BENMESSAHEL*** *pour son soutien durant cette recherche ainsi que sa patiente et sa guidance et tout ce qui nous a apporté comme savoir pour réussir ce travail.*

Nos remerciements s'étendent aussi à tout le personnel et les enseignants de notre département ainsi que toute personne ayant contribué de près ou de loin durant tout notre cycle universitaire.

A la fin nous tenons à remercier plus particulièrement nos amies, nos familles qui nous ont soutenu et encourager sans cesse durant toutes ces années.

Merci à tous !!

Résumé

L’appariement de données (AD) représente le processus identifiant une même entité réelle. De différentes approches d’AD ont été proposées dans la littérature mais la plupart d'entre elles ne tiennent pas compte du contrôle de la taille des blocs.

Dans ce mémoire, nous proposons une amélioration de L’approche RL basée sur les K-Modes, dont laquelle une nouvelle technique de contrôle de taille de bloc est introduit en tant que post-étape vers le blocage. Les différentes expérimentations qui ont été faites sur un ensemble de données du monde réel montrent des résultats satisfaisants, où la plupart des doublons ont été détectés.

Abstract

Record Linkage (RL) is the process of identifying the records that refers to the same real-world entity. Several RL approaches were proposed in the literature but most of them were introduced without a bloc's sizes controlling technic. In this thesis, we propose an enhanced K-Modes-based RL approach, in which a new bloc size mechanism is introduced as a post-step to blocking. The experiments that have been done on a real-world dataset show satisfying results where most of the duplicate records were detected.

ملخص

عملية مطابقة البيانات تمثل العملية التي تحدد كيانًا واحدًا من الواقع يمثل في مصادر مختلفة. تم اقتراح عدة نهج لـ AD في الأدب، لكن معظمها لا يأخذ في اعتباره التحكم في حجم الكتل.

في هذه الرسالة، نقدم تحسينًا لنهج AD القائم على K-Modes. نقدم تقنية جديدة للتحكم في حجم الكتل كخطوة إضافية نحو الحجب. تظهر التجارب المختلفة التي أُجريت على مجموعة بيانات من العالم الحقيقي نتائج مرضية، حيث تم اكتشاف معظم الازدواجيات بنجاح.

Table des matières

Liste des abréviations

AD	L'appariement de données
RL	Record linkage
BK	Blocking Key
BKV	Blocking Key Value
BD	Base de Donnée
DQ	Data Quality
DW	Data Warehouse
JS	Jaccard Similarity
JWS	Jaro Winkler Similarity
NYSIIS	New-York State Identification Intelligence System

Liste des figures

Liste des tableaux

Chapitre 01 : Introduction Générale

1.1. Contexte

De nos jours et dans les systèmes d'information, le nombre de données stockées augmente de plus en plus et prend une grande importance au niveau des organisations du monde entier. Cette augmentation engendre de différents problèmes au sein de ses entreprises.

Les problèmes de ces entreprises peuvent influencer leur progrès au sein de leur travail administrativement et même financièrement car elles perdent de grosse somme afin de pouvoir trouver des solutions pour ces anomalies, sans oublier la perte de temps [2]. L'une de ses anomalies est appelé « la mauvaise qualité de données ».

La qualité des données est un concept très important, avoir des données correctes et cohérentes est indispensable dans nos jours. Il existe plusieurs techniques pour avoir une donnée de bonne qualité dans la littérature, parmi ces techniques on distingue.

appariement des données RL est le processus utilisé pour détecter les doublons dans une ou plusieurs bases de données. basés sur l'algorithme k-modes a été proposé récemment dans la littérature [1], dans ce mémoire nous allons voir de plus près la méthode k-modes et nous allons l'améliorer afin d'avoir comme sorti un nombre de clusters pas forcement similaire au nombre en entrée, ainsi qu'avec des tailles normales contrairement au k-modes traditionnel qui donne des blocs avec des tailles non équilibrées.

1.2. Objectifs

Plusieurs techniques de appariement des données ont été proposées dans la littérature, mais la plupart d'entre elles ne disposent pas de mécanisme de contrôle de la taille des blocs générés, ce qui est une condition très importante dans le domaine de la RL en temps réel. Dans ce mémoire, nous proposons un mécanisme pour contrôler les tailles de bloc générées par L'appariement de données basé sur le k-modes. Les expérimentations faites sur une base de données du monde réel

montrent des résultats satisfaisants où la plupart de ces enregistrements en double ont été détectés tout en respectant les tailles de blocs.

1.3. Structure du rapport

Ce plan de travail est conçu pour guider la réalisation d'une étude approfondie sur l'intégration de données et l'analyse statistique. Le Chapitre 1, "Introduction", présentera le contexte, les objectifs et la méthodologie de recherche adoptée. Le Chapitre 2 se concentrera sur la "Qualité de données (Data quality)", examinant les normes de qualité des données, les sources potentielles de biais et les stratégies d'amélioration. Le Chapitre 3 explorera "L'appariement de données & Clustering", en détaillant les techniques utilisées pour identifier et lier des enregistrements similaires à partir de différentes sources de données. Le Chapitre 4, "Contribution", discutera des contributions théoriques ou pratiques de l'étude à l'avancement du domaine. Enfin, le Chapitre 5 présentera les "Résultats" obtenus à partir de l'application des méthodes discutées dans les chapitres précédents, avec une analyse approfondie des conclusions et des implications.

Chapitre 02 : Qualité de données

2.1. Introduction

Même si les systèmes BD et DW ont prouvé leur supériorité au fil des années, ils peuvent parfois ne pas répondre aux attentes des parties prenantes ou ne pas prendre les bonnes décisions. En effet, de nombreux projets DW ont a été annulé en raison de problèmes de qualité des données (DQ).

Problèmes de QD : il peut apparaître de différentes manières, comme des valeurs manquantes, des enregistrements en double.[1] [11] ou l'intégrité référentielle problèmes. Des données de mauvaise qualité entraînent des pertes estimées à environ 3 000 milliards de dollars par an rien qu'aux États-Unis [2]. Par ailleurs, le Data Ware Housing Institute aux États-Unis a également mentionné que 15 à 20 % des données stockées les données de la plupart des entreprises sont de mauvaise qualité [3]. Par conséquent, les dirigeants des entreprises peuvent perdre confiance dans les systèmes DW et recherchez d'autres solutions car les problèmes de DQ peuvent augmenter le coût du Projets d'entrepôt de données. Les impacts sur la qualité des données sur une entreprise typique peuvent être résumés en trois classes principales (niveau opérationnel, niveau tactique et niveau stratégique) [11].

- Impact sur les opérations

La mauvaise qualité des données a un impact significatif sur le niveau opérationnel. Cela peut provoquer l'insatisfaction des clients et augmenter le coût d'un projet. L'objectif principal d'une entreprise est de faire sûr que tous leurs clients sont satisfaits des services proposés, mais des erreurs de données peuvent empêcher cela. Par exemple, une erreur dans l'adresse d'un client entraînera un problème dans la livraison du produit et cela signifie un insatisfait client. D'un autre côté, une mauvaise qualité des données augmente les coûts opérationnels, Étant donné que le traitement des erreurs de données, leur détection et leur correction coûtent plus cher du temps et plus de budget (des millions de dollars).

- Impact au niveau tactique

Une mauvaise qualité des données est plus dangereuse le niveau tactique. Prendre une décision basée sur des données erronées peut conduire une entreprise à de graves problèmes. La mise en œuvre d'un entrepôt de données est C'est également plus difficile avec des données de mauvaise qualité. Les données stockées doivent être nettoyé et corrigé en premier ce qui entraîne des frais supplémentaires, cela peut affecter le budget du projet DW de manière significative. Un autre effet des mauvaises données sur Le niveau tactique est la difficulté de restructurer et de mettre en place les bons les données au bon endroit et au bon moment pour servir les clients. Des données médiocres peuvent également accroître la méfiance entre les organisations.

- Impacts stratégiques

Chaque entreprise a ses propres stratégies et planifications. Lorsqu'une stratégie est déployée, des plans spécifiques doivent être mis en place et modifiés en fonction des résultats obtenus. Si les résultats sont erronés, tardifs ou Sur la base de données médiocres, l'exécution de la stratégie planifiée est une tâche ardue.

2.2. Définition Qualité de données

La qualité de données renvoie à la capacité d'une entreprise de s'assurer que les données de son système d'information sont correctes et ne peuvent pas être erronées à travers le temps [3].

Il s'agit de fournir des données correctes, complètes, à jour tout en mettant en place des indicateurs peu coûteux compréhensibles et facile à communiquer. Une donnée est dite de qualité si et seulement si elle répond aux exigences de son utilisation.

2.3. Critères

Figure 1 Critères de data qualité

Cette figure nous montre les différents critères que doit avoir une donnée, ci-dessous l'explication de ces critères :

- L'unicité : est appelé aussi l'exhaustivité, c'est le fait que la donnée ne représente qu'une seule entité du monde réel, elle répond à un seul identifiant unique. Par exemple dans une entreprise un employé ne peut pas être représenté par plusieurs données mais avec une seule.

- La complétude : « est-il possible d'identifier l'information ainsi que de la manipuler ? » est le fait que les attributs aient des valeurs significatives.

- L'exactitude : cette notion englobe deux aspects : la précision et la validité, on dit qu'une donnée est exacte si et seulement si la valeur des attributs de l'entité concernée est égale à la grandeur qu'elle est censée représenter dans le monde réel.

- La conformité : cette notion permet de définir des contraintes réglementaires auxquelles la donnée est sous l'obligation de respecter. Si une donnée répond à ces dernières donc elle répondra au contrôle de la qualité afin d'éliminer les valeurs inexactes.

- La cohérence : ce critère représente l'absence d'informations en conflit. Dans le cas d'un référentiel produit, on constatera une incohérence lorsque le prix de vente effectif de ce produit est supérieur à son prix maximum autorisé à la vente.
- L'intégrité : concerne les relations entre objets, étudie l'existence de toute les relations.
- La fraicheur : est nécessaire pour avoir une bonne vision à un instant t et pour prendre les bonnes décisions au bon moment d'où ce critère représente le rapport entre les données et le temps.
- L'accessibilité : « peut-on accéder et utiliser facilement les données ? », représente la facilité d'utilisation de données.
- La pertinence : s'assurer que ces données sont utiles.
- La compréhensibilité : s'assurer que les données sont compréhensibles par les utilisateurs.

2.4. Gestion de la qualité des données

Geiger définit la gestion de la qualité des données comme le processus qui définit politiques et attribu des rôles afin que les données soient collectées, stockées et diffusées [3]. Il a également mentionné que les affaires et la technologie les groupes doivent coopérer afin de parvenir à une gestion de la qualité des données.

Dans la littérature, il existe plusieurs études visant à identifier différentes anomalies dans les données et leurs schémas de description (définitions). Ces anomalies peuvent être classées en deux groupes : les métadonnées et anomalies de données.

La manipulation de données hétérogènes a été confrontée aux fait que les descriptions étaient très pauvres, voire manquantes. Des métadonnées telles car les contraintes et les commentaires étaient quasiment inexistants. En conséquence, Il était difficile de trouver la sémantique des données intégrées [13] [9].

Les anomalies de données peuvent être classées en trois catégories principales, à savoir (1) les anomalies intra-colonnes (telles que valeurs nulles, valeurs aberrantes, anomalies syntaxiques et sémantiques - échec de respecter les expressions régulières (-), (2) anomalies inter-colonnes (dépendances sémantiques incluant dépendances fonctionnelles, exactes ou approximatives ou

dépendances fonctionnelles conditionnelles.) et enfin (3) anomalies interlignes (doublons et similaires).

Nous en donnerons ici un aperçu synthétique. Nous avons étudié les fonctionnalités d'outils de gestion de la qualité des données et d'outils ETL tels que Talend, Pentaho, Nadif et Katara

La comparaison repose sur plusieurs critères. Ceux-ci représentent des fonctionnalités liés à la qualité des données tels que (1) Analyse statistique des données : fonctions qui fournir des statistiques simples, par exemple le nombre de lignes, le nombre de valeurs nulles et le nombre de valeurs distinctes et uniques. Statistiques

Les types de données (texte, numérique et date) sont utilisés pour analyser les caractéristique des colonnes, telles que les longueurs minimale, maximale et moyenne ; (2) Le transformations nécessaires sur les valeurs des données lors de l'intégration : groupes les fonctions de transformation des dates et des nombres ; et (3) les doublons et similaire : différents algorithmes de calcul des distances de similitudes sont mis en œuvre.

Ces outils offrent la possibilité de réaliser des statistiques et des transformations de données. Ils permettent d'éliminer les doublons. Cependant, ce sont les utilisateurs qui doit les guider dans ce processus de nettoyage. L'utilisateur doit savoir structures de données et sémantique pour corriger les anomalies et les valeurs incohérentes.

Ces actions correctives doivent être initiées par l'utilisateur. Aucune aide ne lui est apportée. Par exemple, Pentaho Data Intégration et Data Cleaner ne permettent pas de vérifier les dépendances fonctionnelles, alors que Talend Data La qualité oui, mais c'est l'utilisateur qui doit avoir connaissance des données schéma et les dépendances à vérifier. Il convient de noter que non les outils corrigent les erreurs causées par la violation des dépendances fonctionnelles. Nous ont ainsi identifié les faiblesses à améliorer et les fonctionnalités à développer pour contribuer au développement de nouveaux outils qui ne obliger l'utilisateur à connaître les structures et la sémantique des données manipulées à partir des sources et nous permettre d'assister à la correction de tous types d'anomalies. La redécouverte des métadonnées devient alors notre objectif pour mieux résoudre les problèmes de qualité des données. Le but est de découvrir le sens et contraintes

qui pourraient être définies dans chaque colonne, les relations entre les colonnes et d'en déduire les colonnes clés afin de mieux réaliser la déduplication.

2.5. Dimensions de la qualité des données

Afin de donner une valeur mesurée pour la qualité des données, un ensemble de critères de qualité des données les dimensions doivent être utilisées. Batini et Scannapieco [10] se partagent les majeures dimensions de la qualité des données en deux groupes : principale et secondaire.

Les principales dimensions DQ comprennent l'exactitude, l'exhaustivité, l'actualité et Cohérence. Les dimensions secondaires comprennent l'accessibilité, la capacité d'interprétation et d'autres dimensions liées au temps. De plus, une ou deux métriques sont définis pour chaque dimension [10].

La dimension de précision peut être définie syntaxiquement et sémantiquement. La plupart des méthodologies de qualité des données ne prennent en compte que la précision syntaxique. Il est défini comme la proximité d'une valeur V avec un élément du domaine D. Plusieurs métriques existent dans la littérature pour mesurer la syntaxe Précision telle que la distance d'édition, des sons similaires (comme Soundex et NYSIIS) et transposition de caractères. L'exhaustivité est généralement représentée par la présence de la valeur nulle dans la collecte de données où La valeur existe dans le monde réel. Nous pouvons trouver différents types de complétude Comme : valeur, attribut, tuple et exhaustivité de la relation. L'actualité, l'actualité et la volatilité sont les dimensions de qualité des données liées au temps les plus importantes. La monnaie fait référence au taux auquel le les données stockées sont mises à jour avec. Habituellement, la monnaie est mesurée à l'aide du métadonnées de la dernière mise à jour. La volatilité est liée au type de données, elle est faible si les données sont stables comme l'attribut prénom et élevées si les données changent aime souvent les e-mails. L'actualité représente l'adéquation des données pour un certain problème à un instant fixe. La cohérence représente le degré de violation des règles sémantiques prédéfinies. Aujourd'hui, les données peuvent être qualifiées par volume, variété, véracité, vitesse et valeur.

2.6. Quatre piliers pour la gestion de la qualité des données

Afin de gérer la qualité des données de manière organisée, quatre principaux il faut suivre les étapes. Ces étapes ont été nommées par Jonathan comme étant les quatre piliers pour la gestion de la qualité des données [3]. La première étape est le profilage des données, le but est de se faire une idée sur les données stockées et leur qualité par rapport aux spécifications de qualité. Données le profilage dans le monde des entrepôts de données est appelé « analyse du système de sources ». Il permet au responsable qualité d'avoir une idée de l'exactitude et de la l'exhaustivité des données

stockées. Le profilage des données aide également à détecter valeurs en double en cas de données trop nombreuses que prévu.

Une fois le profilage des données effectué, le responsable qualité peut avoir une idée sur les problèmes de qualité des données stockées. Quatre solutions sont possibles pour traiter les données concernées :

(1) Supprimez les données dans le cas où le problème est difficile à résoudre. (2) Acceptez les données où l'erreur dans le les données peuvent être ignorées. (3) Correction des données. Par exemple, dans le cas des valeurs en double, nous pouvons sélectionner un tuple et supprimer les autres et enfin (4) insérer une valeur par défaut, notamment en cas de valeurs manquantes. L'intégration des données est également un processus important pour atteindre un niveau élevé de données qualité. En particulier, lorsque les données sont collectées à partir de plusieurs sources, la même entité du monde réel peut être représentée différemment au niveau de chaque source. Cela permet de supprimer les enregistrements en double, ce qui améliore la qualité des données stockées. L'augmentation des données joue également un rôle important dans la gestion de la qualité des données. Elle développe des informations significatives sur les clients en intégrant des données externes provenant de tiers avec les données stockées.

2.7. Coût de la non qualité

Il existe deux types de coûts de non qualité :

- **Les coûts de non qualités internes :**

Ces coûts représentent les frais qui viennent à travers un disfonctionnement, à la non-qualité d'un produit qui ne répondra pas aux exigences des utilisateurs, organisations, c'est à dire avant sa livraison au client. Ces frais sont de différents catégories : les rebuts, les coûts de réparation ou de remplacement des machines défectueuses…etc.

- **Les couts de non qualité externes**

Ces coûts représentent aussi les frais relatifs à un produit ou service non conforme aux exigences qualité mais après avoir quitté l'entreprise, c'est-à-dire après sa livraison au client. Entrent dans cette catégorie les frais qu'on pouvait éviter si le livrable avait satisfait le client. Par exemple : le traitement des réclamations client, le remplacement des produits défectueux, les éventuels pénalités, dommages et intérêts.

2.8. Causes de la non qualité

Mal appréhender l'environnement de données : C'est le fait d'avoir une méconnaissance sur la nature des données ou bien une ignorance de la définition de la variété des données qui peut conduire à une analyse inadéquate ou pire, à la propagation d'interprétations incorrecte / inexacte. D'où bien connaître ses données permet d'éviter beaucoup d'erreurs.

Informations incomplètes : Dans les statiques, le manque de données ou valeurs se produit lorsqu'aucune valeur n'est exprimée pour une variable pour une observation donnée. A titre d'exemple : quand un utilisateur oublie de remplir un champ.

Erreurs typographiques et données inexactes : Une donnée inexacte est toute donnée présentant des problèmes de conformité ou d'authenticité : adresse incomplète, Nom mal orthographié, valeur qui n'a rien à voir avec la variable qui la contient. Toutes les erreurs peuvent être corrigées dans la plupart des cas, mais encore faut-il pouvoir les détecter.

Format inconsistant / problème de type : Dans le cas où le format de stockage des données utilisé est incohérent, les systèmes d'analyse ou de stockage d'informations peuvent ne pas être en mesure de les interpréter correctement.

Doublons : Couramment la plupart des organisations souffrent du problème de la redondance de données. Qui peut être créé du fait de la duplication d'information collecter à partir de différentes sources, ou bien par une erreur humaine ou en ajoutant des données au lieu de les mettre à jour.

Unité de mesure / langages multiples : Du fait de l'utilisation de divers langages, et de codes ainsi que les unités de mesure différentes, le risque d'apparition du problème de compatibilité augmente. Pour cela il faut toujours penser à la conversion avant la compilation.

Les valeurs aberrantes : En statistique, les valeurs aberrantes sont des valeurs qui s'écartent significativement de la distribution d'une variable. Il s'agit d'une observation anormale qui s'écarte des données bien structurées.

Erreur de transformation : Il est courant en science des données de s'appuyer sur certaines transformations mathématiques avant la modélisation des données. Normaliser les valeurs des variables pour passer de variables catégorielles à des variables continues ou indicatrices. Ces transformations sont souvent liées à des hypothèses sur les données ou à des limitations de l'algorithme que vous essayez d'utiliser.

Problèmes de définition : Il est toujours important de pouvoir décrire précisément les variables présentes dans l'ensemble de données. Le cas échéant, il faudra demander, dans le cas où la définition d'une variable n'est pas assez précise.

Questions de conformité : En définitive, cela peut sembler évident pour certains, mais lors du traitement des données, il est important de s'assurer que l'entreprise, ses dirigeants et ses employés

respectent les normes légales et éthiques. Ainsi, on permet à son entreprise d'éviter les risques financiers, juridiques et de réputation qu'encourt une organisation lorsqu'elle ne respecte pas les lois, règlements, conventions ou simplement une morale ou des obligations.

2.9. Les approches de la qualité de données management

2.9.1 Les approches préventives :

Elles sont des méthodes axées sur la maitrise et le contrôle de la conception des systèmes d'informations complexes, utilisant des techniques d'évaluation de la qualité du model conceptuel [3].

Afin de garantir une meilleure qualité de données, il faut choisir le processus de développement de logiciels le plus qualitative.

2.9.2 Les approches diagnostics :

Ses procédures se concentrent sur les techniques statistiques, analytiques, et exploratrices des données afin de souligner les anomalies dans les données [3].

2.9.3 Les approches adaptatives :

Appelée aussi les approches actives, couramment employées dans le processus de collection, de transformation, et d'intégration des données appuyant sur le choix du traitement le plus adéquat en termes des requêtes ou bien d'opération de nettoyage pour engendrer la validation en temps réel des contraintes de qualité des données lors de l'exécution.

2.9.4 Les approches correctives :

L'approche corrective s'articule autour de techniques de nettoyage et d'intégration de données et utilise des langages étendus de manipulation de données et des outils d'extraction et de transformation de données (ETL, extract-transform-load).

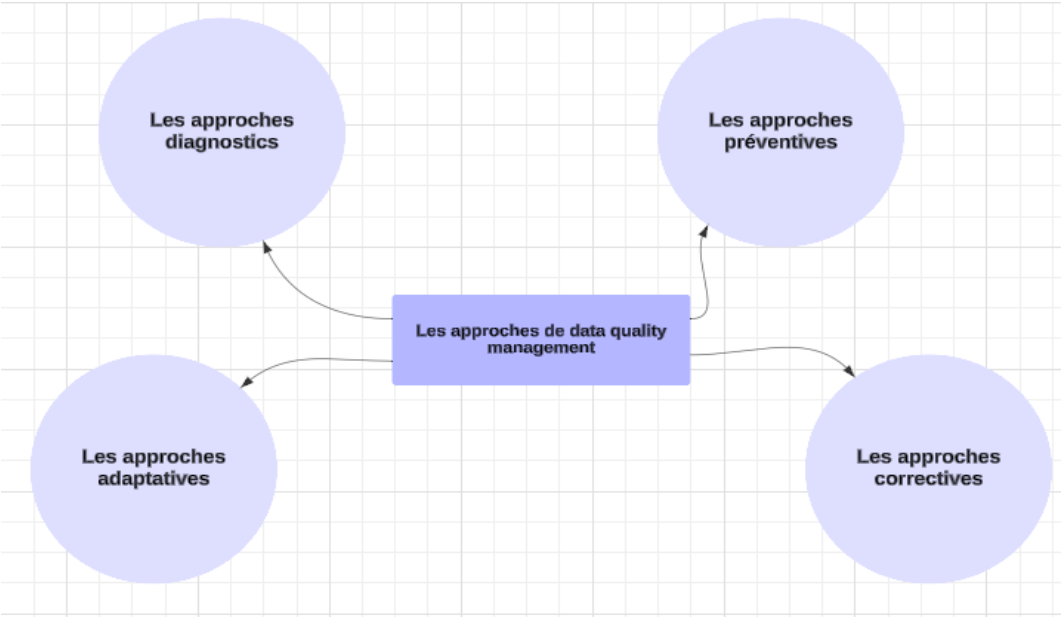

Figure 2: Les approches de data quality management

2.10 C'est quoi l'ETL :

ETL est l'acronyme d'Extraction, Transformation et du chargement. Dont il représente l'un des outils propres à la technologie informatique intergicielle [13]. Ses outils ETL Illustrent les différents traitements appliqués sur les données d'une manière automatique. En suivant son un processus bien défini, qui commence par extraire les informations primaires (ou bien dite pertinentes) des données source, Ensuite il les transforme en choisissant un format adéquat par

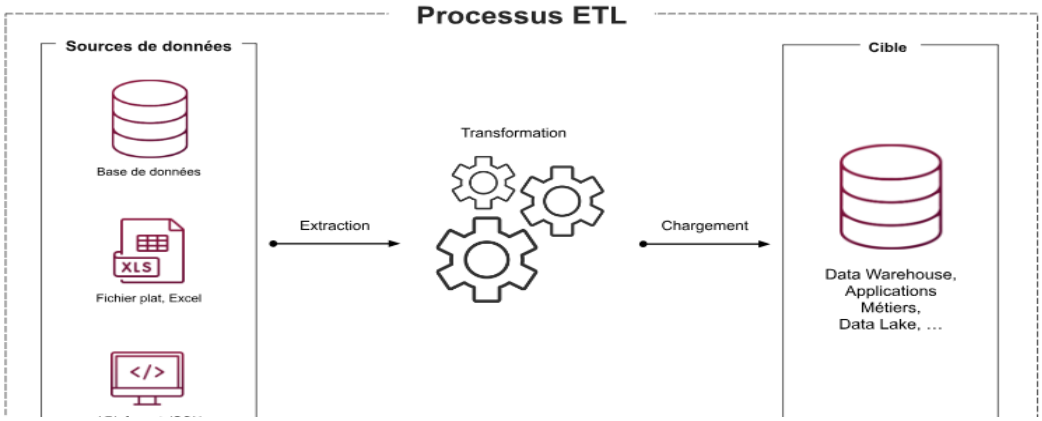

Figure 3: Processus ETL

rapport aux exigences des entreprises, au final il clôture le processus en les chargeant en données cibles [6].

Un ancien projet étudiant Crée entre 2006 et 2009, GeoKettle est un ancien projet étudiant québécois qui s'est développé à partir de Kettle (Pentaho Data Integration). En 2009, la compagnie Spatialytics a été créée pour poursuivre le développement des projets GeoKettle et GeoMondrian. GeoKettle est une version de Pentaho Data Integration spécialisée dans le traitement des données géospatiales. Avec son interface graphique et son approche metadata driven, ce logiciel reprend le fonctionnement de l'ETL Pentaho. Il s'installe sur Windows, Mac OS, Linux et Solaris et peut extraire des données issues de plus de 35 database (Oracle, MySQL, etc.... GeoKettle constitue un ETL complet et entièrement gratuit rivalisant avec produits propriétaire.

Scriptella est un ETL codé en Java proposé en version gratuite. Cette solution a vu le jour 2004. Elle est principalement utilisée pour réaliser des projets de migration de données et des opérations d'intégration à partir de bases croisées. Le point noir de cet outil est l'absence d'interface graphique.

Knowage, plus connu sous son ancien nom SpagoBi, est une suite de logiciels dédiés à l'intégration et la création de reporting. L'éditeur propose aujourd'hui une double licence communautaire et commerciale après de nombreuses années entièrement gratuites. Désormais les fonctionnalités avancées de l'ETL sont payantes. Avec une interface graphique et un développement sur Java Knowage est une solution d'intégration ergonomique.

2.11. Les outils de la qualité de données(nettoyage) :

Parmi les outils utilisés pour assurer une meilleure qualité de données on a [7] :

- Le profilage : Est utilisé afin de déterminer les critères de qualité de la donnée en délimitant les paramètres d'actions.
- Le monitoring : Représente l'analyse de la qualité des données après avoir appliqué le profilage.
- La standardisation : C'est le fait de fixer une convention adéquate à la donnée.
- Le nettoyage : est la correction des données erronées.

- Le matching : représente la recherche relative d'une même entité afin de détecter les doublons et éviter l'incohérence.

- L'enrichissement : c'est l'ensemble des appels aux applications externes pour compléter les données.

- La surveillance : C'est le suivit de l'évolution des données au fil du temps pour déterminer leurs désorientations et la violation des règles.

Quelques exemples de logiciels utilisés pour garantir la qualité des données :

L'outil	La description
Talend	Talend fournit des représentations graphiques et statistiques permettant d'identifier rapidement les problèmes de qualité, ainsi que les modèles et à détecter les anomalies. Cet outil facilite le nettoyage, la normalisation et le profilage des données du système. ET peut également résoudre les problèmes de qualité des données au fur et à mesure. Il dispose d'une interface libre-service pratique pour les utilisateurs professionnels et techniques. Talend garantit que des données fiables sont toujours disponibles pendant l'intégration, améliorant ainsi les performances de l'entreprise et réduisant les coûts. Le Talend Confidence Score intégré fournit des indices de confiance instantanés, exploitables et interprétables pour distinguer les ensembles de données nettoyés des données à nettoyer. Cet outil enrichit vos données en les combinant avec des détails provenant de sources externes, telles que l'identification authentique de l'entreprise ou les codes postaux.

OpenRefine	Connue avant sous le nom de Google Refine, OpenRefine est un outil puissant utilisé pour travailler avec des données encombrées, nettoyer et convertir des données d'un format à un autre.
ZoomInfo OperationsOS	ZoomInfo OperationsOS fournit des données accessibles flexibles et de premier ordre pour vous aider à accélérer votre activité. Sa précision d'entrée, son taux de correspondance et son taux d'entrée inégalés offrent la plus grande fiabilité des données.

Table 1: Exemples d'outils

2.12. Conclusion :

Vu l'importance de l'amélioration de la qualité de données circulant au sein des entreprises, ses dernières doivent automatiquement suivre des méthodes. A titre d'exemple le L'appariement de données , le Matching qui seront bien détaillé dans le chapitre suivant.

Chapitre 03 : Le L'appariement de données et Clustering

3.1 Introduction :

Étant donné que les bases de données sont rarement complètement valides et formellement cohérentes, la même entité peut apparaître dans plusieurs enregistrements et aussi souvent de manière différente et imparfaite. C'est ce qu'on appelle la duplication et l'incohérence dans la base de données.

Le terme couplage d'enregistrements est utilisé pour indiquer la procédure de regroupement des informations

à partir de deux ou plusieurs enregistrements1 censés appartenir à la même entité. Le couplage d'enregistrements est utilisé pour lier données provenant de plusieurs sources de données ou pour rechercher des doublons dans une seule source de données. Dans l'ordinateur science, le couplage d'enregistrements est également connu sous le nom de couplage de données. La correspondance des données ne restreint pas la structures de données aux enregistrements. Dans ce document, le terme couplage d'enregistrements est utilisé tandis que le terme, la « correspondance des données » est également satisfaisante.

L'idée du couplage d'enregistrements a été introduite au milieu des années 1900. Pour autant que nous le sachions, Dunn [1946] a été le premier à utiliser le terme couplage d'enregistrements dans son concept de « Livre de vie ». Un livre de vie est un livre personnel qui commence à la naissance et se termine à la mort. Registres des principaux événements de la vie, comme le mariage et l'obtention du diplôme, remplissez les pages. Dans ce contexte, le couplage d'enregistrements est le processus d'assembler les pages de la personne en un volume. Après cette introduction conceptuelle de couplage d'enregistrements, Newcombe et al. [1959] ont commencé à étendre ce concept de couplage d'enregistrements dans un manière mathématique.

Dans le couplage d'enregistrements, les attributs de l'entité (stockés dans un enregistrement) sont utilisés pour relier deux ou plusieurs enregistrements. Les attributs peuvent être des identifiants (uniques) d'entité, mais également des attributs tels que le (sur)nom, le sexe, date de naissance et couleur des cheveux. Si l'identifiant de deux enregistrements2 est identique, alors les enregistrements (très probablement) appartiennent à la même entité. En général, le couplage d'enregistrements est considéré comme le processus de relier les enregistrements pour lesquels ces identifiants uniques ne sont pas disponibles. Les données doivent être liés en fonction d'attributs

ayant moins de pouvoir distinctif tels que le (nom de famille), le sexe, la date de naissance et la couleur des cheveux.

De plus, il est possible que différentes bases de données (avec une certaine structure logique) codent (modèlent) la même entité ou ensemble d'entités de différentes manières. Par conséquent, cela provoque des incohérences entre les bases de données.

Pour résoudre ses problématiques il existe plusieurs processus : L'appariement de données , Matching, et le Clustering.

Durant ce chapitre nous allons répondre aux questions suivantes :

- C'est quoi le L'appariement de données , les étapes à suivre, les problèmes rencontrer lors de l'application de la solution, ainsi que les approches existantes dans la littérature.
- C'est quoi le Matching, ses approches et le Phonetic encodeing ?
- C'est quoi le clustering, les différents algorithmes existants et une comparaison entre les différentes approches ?

3.2 L'appariement de données :

3.2.1 Définitions :

Le L'appariement de données est le processus consistant à combiner des enregistrements ou des unités provenant de diverses sources de données dans un seul fichier à l'aide d'identificateurs non uniques tels que des noms, des dates de naissance, des adresses et d'autres caractéristiques. Également appelé l'appariement de données, jointures de données, résolution d'entité et de nombreux autres termes en fonction du domaine dans lequel ils sont utilisés [5].

L'idée originale de la liaison d'enregistrements revient aux années 1950, juste après, cette technique a été utilisée par des personnes dans divers domaines, notamment : Entreposage de données et intelligence de gestion, recherche historique, pratique et recherche médicale.

Les couplages sont utilisés depuis longtemps dans la recherche statistique et l'élaboration de données administratives. Le couplage d'enregistrements peut être utilisé pour créer des cadres d'échantillonnage, supprimer les doublons de fichiers, fournir des informations supplémentaires pour faciliter la manipulation des données ou combiner des fichiers pour examiner les relations entre deux éléments de données ou plus dans différents fichiers.

3.2.2 Les types de couplage :

Il existe deux types de couplage d'enregistrement : l'appariement exact et l'appariement statistique. Ce dernier peut être divisé en deux sous-types : le couplage d'enregistrements déterministe et le couplage d'enregistrements probabiliste. Comme suit [5] :

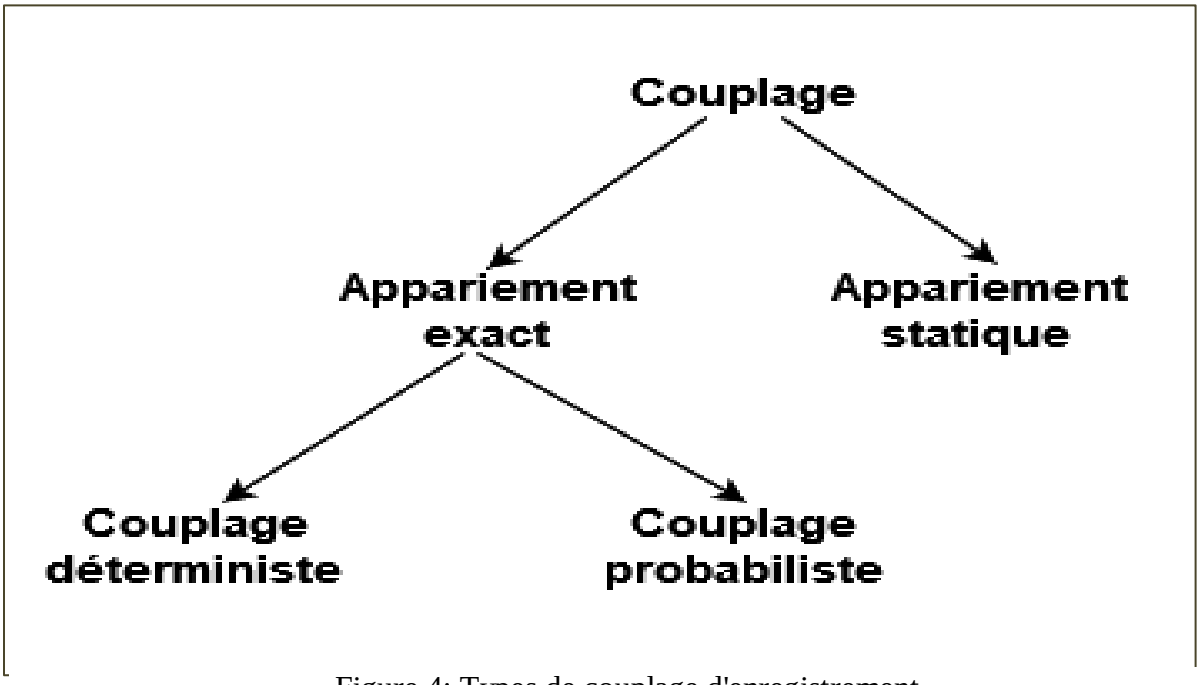

Figure 4: Types de couplage d'enregistrement

- **Appariement statistique :**

L'appariement statistique a pour but de créer un fichier qui reflète la répartition sous-jacente de la population. Les enregistrements fusionnés ne correspondent pas nécessairement à la même entité. Les fichiers appariés peuvent avoir des unités différentes mais faire référence à la même population. Les relations entre les variables au sein de la population sont supposées être similaires à celles du fichier. Cette méthode est principalement utilisée dans les études de marché et rarement par les bureaux officiels de statistiques.

- **Appariement exact :**

Le but d'appariement exact est d'associer des informations sur des enregistrements spécifiques dans un fichier avec des informations dans un autre fichier secondaire pour créer un fichier unique qui contient les informations correctes pour chaque enregistrement. La liaison se fait au niveau de l'enregistrement. A titre d'exemple, le lien entre la date du décès et le recensement.

- **Couplage déterministe :**

Il s'agit de la forme la plus simple de couplage d'enregistrements, générant des liens basés sur des identificateurs ou des variables communes entre les sources de données disponibles. Il arrive souvent qu'aucune des variables ne soit sans erreur, présente dans la plupart des données et suffisamment discriminante. Seule la combinaison des variables distingue les deux enregistrements. Il s'agit d'une technique souvent utilisée par les agences statistiques officielles. Statistique Canada utilise cette méthode pour constituer ses registres des entreprises, des adresses et de la population, qui impliquent alors de multiples opérations d'enquête.

- **Couplage probabiliste :**

Il s'agit d'un autre type de correspondance exacte. Comme dans l'autre cas, aucun identifiant unique n'est disponible pour correspondre. Contrairement à l'appariement déterministe, l'appariement probabiliste peut compenser si les informations sont incomplètes ou éventuellement erronées. Les enregistrements qui ne correspondent pas complètement à chaque variable peuvent être liés ensemble pour former un ensemble de paires potentielles. Les scores sont ensuite calculés pour chaque paire potentielle. Ensuite, un état d'appariement est attribué à chaque paire potentielle en fonction du score.

3.2.3 Le flux de travail de records linkage :

Comme mentionné dans l'introduction, dans le domaine de la qualité des données, le processus d'identification des enregistrements qui représentent la même entité du monde réel et choisir lequel conserver ou comment les fusionner est connu sous le nom de le processus de

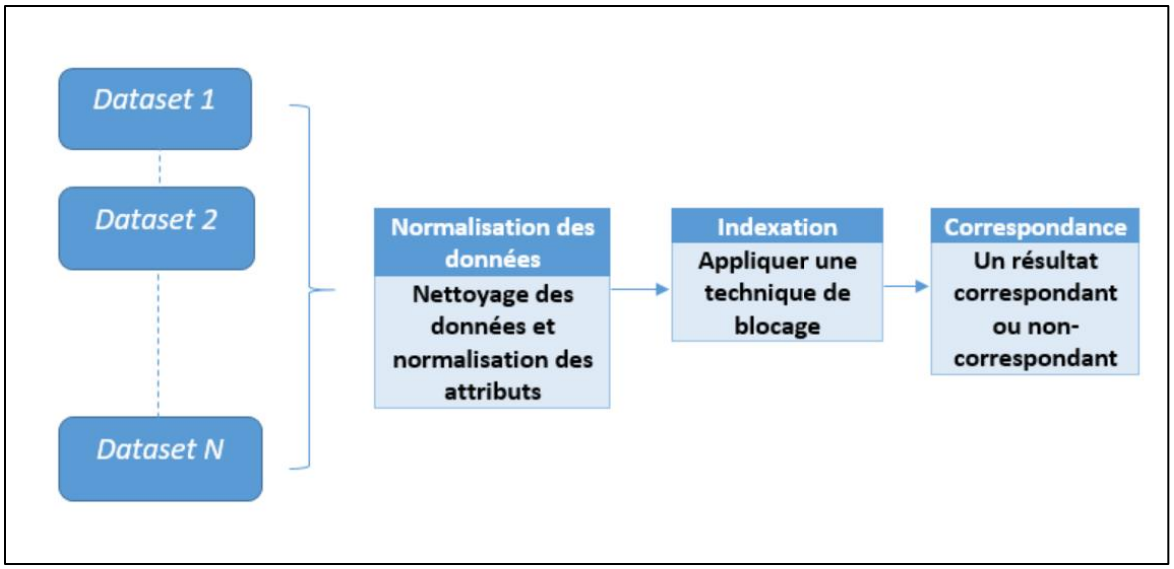

Figure 5: Le flux de travail de records linkage

couplage d'enregistrements. D'après la littérature, le processus RL peut être défini par un processus en trois étapes [15] comme le montre la figure(5) :

- **La première étape est la Normalisation des données :**

En fait, il a été prouvé dans des recherches antérieures [Clark 2004] que laisser des ensembles de données sans normalisation d'attributs, et la détection d'anomalies de schéma peut conduire à de mauvaises conclusions et même à la fusion des mauvais tuples. Par exemple, l'attribut qui représente le nom d'une personne peut apparaître dans un ensemble de données sous forme de nom complet et dans un autre comme deux attributs (Prénom et Nom). Donc les données la normalisation est une étape importante dans le processus RL.

- **La deuxième étape du processus RL est l'indexation :**

Le but de cette étape est : pour identifier les tuples qui seront comparés les uns aux autres lors de la mise en correspondance étape. Le moyen idéal d'y parvenir en termes de précision est l'approche naïve en comparant chaque enregistrement à tous les autres. Mais bien sûr, cela pourrait finir par un nombre inacceptable de comparaisons. Par exemple, en comparant deux bases de données contenant chacune 2 millions d'enregistrements peuvent aboutir à 412 comparaisons et la plupart de ces comparaisons aboutiront à un résultat sans correspondance. Par conséquent, l'indexation vise à réduire le nombre de comparaisons.

La technique d'indexation la plus utilisée dans la communauté RL est connue sous le nom de"**BLOCAGE**".

Le blocage est le processus qui divise l'ensemble de données en un ensemble de blocs.

Tous les tuples affectés au même bloc partagent une valeur commune connue sous le nom de valeur-clé de blocage (BKV). Une clé de blocage peut être choisie comme attribut unique.

Par exemple, tous les enregistrements partageant la même valeur pour les adresses d'attribut sont affectées au même bloc. Sinon, un blocage.

La clé peut également être choisie avec la concaténation de plusieurs attributs comme les quatre premiers caractères du Prénom et du Code Postal du attribut d'adresse. La décision de sélectionner quel attribut ou groupe de les attributs seront utilisés comme clé de blocage est très important car le les blocs qui seront créés dépendent de cette décision. Alors, en choisissant le moins attribut sujet aux erreurs car BK est très important et parfois nous avons besoin d'intervention d'un expert pour cette décision. Une fois l'étape d'indexation effectuée, seuls les enregistrements d'un même bloc sont comparés les uns aux autres. Plusieurs des techniques de blocage ont été proposées dans la littérature, La clé de blocage peut être un attribut unique ou bien elle peut être la concaténation de plusieurs attributs.

Une bonne technique de blocage est contrôlée à travers ces deux paramètres : « la valeur-clé de blocage » et « le nombre de clés de blocage ».

Exemple : « Attribut unique » :

Nom	Num téléphone	Adresse
Moussa	056567	BBA
Nadir	078768	Setif

Exemple : « Attributs concaténés » :

La clé de blocage est un constitué de : 3 lettres de l'attribut Nom + Num téléphone

BKV	Nom	Num téléphone	Adresse
Mou056567	Moussa	056567	BBA
Nad078768	Nadir	078768	Setif

- **La dernière étape du processus RL consiste à :**

faire correspondre les enregistrements indexés qui sont dans le même bloc. La correspondance peut être effectuée en utilisant un ensemble de fonctions de similarité de chaîne qui existent dans la littérature [Levenshtein 1966] ou en utilisant un algorithme d'apprentissage automatique pour classer l'enregistrement comme correspondant ou non [7].

3.2.4 Les défis liés au couplage d'enregistrements (challenges) :

Comme la plupart des tâches de nettoyage des données, l'arrivée de l'ère du Big Data met nouveaux défis pour le processus de couplage d'enregistrements. La plupart des traditionnels les techniques de blocage sont incapables de gérer l'énorme quantité de données générées chaque jour. Chacun des V mentionnés dans la définition du Big Data dans les sections précédentes, apporte de nouveaux défis qui doivent être relevés par le processus RL.

L'aspect volume du Big Data a rendu le processus de couplage d'enregistrements plus complexe. Plus il y a de données, plus il y a de comparaisons, ce qui rend le blocage des approches est plus important pour réduire au maximum les nombre de comparaisons. La véracité du Big Data constitue la première étape de

le processus RL est plus crucial. Les données sont collectées à partir de diverses sources avec différents schémas. La normalisation de ces données est très importante pour obtenir de bons résultats. L'application du processus RL sur un Big Data set peut aboutir à un résultat élevé du temps de calcul, qui n'est pas acceptable dans le cas de problèmes BD, dans lequel le temps est un

axe très important. Améliorer les approches traditionnelles de RL et les rendre plus évolutives est une question très importante à aborder.

La vitesse de BD. De nos jours, les données sont collectées à partir de différentes sources sans se demander s'il a une valeur commerciale ou non. Étant donné que la collecte de Big Data n'est pas supervisée par des spécialistes (chercheurs scientifiques et agences), la sélection des données les plus précieuses à analyser est également une étape très importante étape.

3.2.5 Domaines d'application :

Au cours des dernières décennies, le couplage d'enregistrements a été utilisé dans divers domaines et pour plusieurs raisons. RL peut être intégré à tout processus qui doit intégrer des données provenant de différentes sources dans lesquelles les valeurs en double doivent être supprimées afin d'en extraire des conclusions correctes des données collectées. RL a été utilisé dans des domaines tels que les soins médicaux, analyser les données du recensement, détecter la criminalité et la fraude et préserver la vie privée.

Aux États-Unis, chaque année, le National Center for Health Statistics demande à un ensemble d'hôpitaux sélectionnés pour soumettre une liste de tous les patients sortants et toutes les visites en ambulatoire. Leur objectif est de relier les données collectées auprès des hôpitaux avec d'autres données collectées telles que le Nation Death Index afin d'obtenir des informations sur la mortalité à l'hôpital après décharge, pour ce faire, ils utilisent le procédé RL. Un autre exemple est la liaison données des patients provenant de différents services afin d'améliorer la qualité des services dans les hôpitaux.

Le couplage d'enregistrements est également utilisé par les éditeurs en ligne afin de détecter les publications en double dans leur stockage et également détecter le plagiat dans les documents stockés dans leurs bases de données. Les entreprises commerciales également profiter du processus RL en détectant et en supprimant les doublons entités à partir de leurs bases de données afin de mieux adresser leur publicité investissements et de réduire le budget dédié.

3.3 Matching :

3.3.1 Définition :

Comme mentionné ci-dessus, la dernière étape du processus de couplage d'enregistrements est faire correspondre les enregistrements indexés et décider si deux paires comparées représentent la même entité du monde réel ou non. Généralement :

- La valeur correspondante est entre [0.1]

• 1 est une correspondance exacte

• 0 n'est pas une correspondance

Dans cette section, un bref aperçu est fourni sur les techniques d'appariement existantes dans la littérature.

Dans la littérature, il existe deux familles de techniques d'appariement. La première l'un est le codage phonétique. L'idée de cette technique est de transformer une chaîne en un code qui représente la façon dont la chaîne est prononcée. Une variété d'algorithmes de codage phonétique existent dans la littérature (Soundex et phonex, phénix, NYSIIS et Double-Metaphone). La deuxième correspondance La famille des techniques est la recherche de modèles. L'idée principale de ces techniques consiste à mesurer la similarité entre deux mots sans aucune transformation en utilisant un ensemble de mesures de similarité de chaînes telles que la distance d'édition.

a. Distance d'édition :

Cette distance est également connue sous le nom de distance de Levenshtein. Elle a été proposée en 1965 par Vladimir Levenshtein. Il est considéré comme l'un des les métriques les plus utilisées afin de mesurer la similarité entre deux codes. Généralement, il est défini comme le nombre d'insertions, de suppressions et mises à jour c'est-à-dire toute modification faite afin de transformer une chaine en une autre. Pour mieux comprenez-le, l'exemple suivant montre une démonstration de la façon dont pour calculer les coûts de passage d'un mot à un autre.

Exemple :

Dans cette exemple la distance d'édition est égale à 6

Mots1	Mots 2	Operation	cout
I	D	Comparaison entre « i » et « d »	1
D	I	Comparaison entre « d » et « i »	1
T	S	Comparaison entre « t » et « s »	1
S	T	Comparaison entre « s » et « t »	1
Z		Suppression de « z »	1
A	A	Comparaison entre « a » et « a »	0
N	N	Comparaison entre « n » et « n »	0
C	C	Comparaison entre « c » et « c »	0
E	E	Comparaison entre « e » et « e »	0

	S	Insertion de « s »	1
		Somme = distance	6

Table 2 : Exemple de distance d'édition

Dans l'exemple ci-dessus, nous pouvons voir que la distance d'édition entre le deux mots (idtszance , dist ance) est la somme des coûts de transformation la première chaîne dans la seconde qui est égale à 6. Les marches qui sont représentés dans l'exemple ne sont pas la seule solution pour transformer la première chaîne dans la seconde mais sont celles qui coûtent le moins cher.

a. Distance de Jaro-Winkler :

Le Jaro-Winkler est une métrique de similarité de chaînes de caractères proposée parWilliam E. Winkler en 1990 dans le prolongement de la distance Jaro. Pour mesurer la similarité Jaro-Winkler entre deux codes, Dont son résultat est représenté par un 0 ou 1 tout dépend de la similarité ou la dissimilarité. la première étape est de mesurer la similarité de la tradition Jaro qui est définie comme :

$$\boldsymbol{Jaro\_Sim}(\boldsymbol{c1}, \boldsymbol{c2}) = \frac{1}{3} \times \left(\frac{m}{|c1|} + \frac{m}{|c2|} + \frac{m-t}{m} \right)$$

- (c) représente la longueur de la chaîne.
- (m) représente le nombre de caractères communs entre les comparé des séquences avec le même indice.
- (t) représente le demi-nombre de transpositions afin d'améliorer la métrique précédente, William E. Winkler utilise une échelle de préfixe P afin de mettre en favoris les chaînes qui commencent par le même préfixe L pour une longueur maximale de quatre. La similitude Jaro-Winkler est défini comme suit :

$$\mathbf{JaroWinlker\_Sim(c1,c2) = Jaro\_Sim(c1,c2) + LP(1 - Jaro\_Sim(c1,c2))}$$

- Jaro_Sim(c1,c2) est la distance de Jaro entre les chaînes.
- L est la longueur du préfixe commun (maximum 4 caractères)
- P est un facteur d'échelle (une constante qui prend généralement la valeur 0,1).

Exemple :

Soient deux chaines de caractères c1 : wissem , c2 : wissam nous allons dresser leur table de correspondance

	w	i	s	s	a	m
w	1	0	0	0	0	0
i	0	1	0	0	0	0
s	0	0	1	0	0	0
s	0	0	0	1	0	0
e	0	0	0	0	0	0
m	0	0	0	0	0	1

Table 3:Exemple Distance de Jaro-Winkler

- S = 5, |c1|=6, |c2|=6
- $\left[\frac{\text{Max}(6,6)}{2}\right] - 1 = 2$ L'éloignement maximum pour le critère de correspondance est 2

$$\boldsymbol{dj} = \frac{1}{3} \times \left(\frac{5}{6} + \frac{5}{6} + \frac{5-0}{5}\right)$$

$$\boldsymbol{dj} = \frac{1}{3} \times \left(\frac{5+5}{6} + \frac{5}{5}\right)$$

$$\boldsymbol{dj} = \frac{1}{3} \times \left(\frac{10 \times 5}{6 \times 5} + \frac{5 \times 6}{5 \times 6}\right)$$

$$\boldsymbol{dj} = \frac{1}{3} \times \left(\frac{80}{30}\right)$$

$$\boldsymbol{dj} = \frac{8}{9} \cong 0.8888888889$$

$$dw = 0.88889 + \left(2 \times 0.1(1 - 0.88889)\right)$$

$$dw = 0.88889 + (0.2 \times 0.11111)$$

$$\boldsymbol{dw \cong 0.911112}$$

c. Distance de Jaccard :

La distance de Jaccard est généralement utilisée pour mesurer la similarité entre deux ensembles d'échantillons qui peuvent être le cas de Strings. Afin de mesurer la distance de Jaccard, il faut d'abord calculer le coefficient de Jaccard qui est défini comme :

$$\textbf{Jaccard(A, B) = A} \cap \textbf{B A} \cup \textbf{B}$$

Une fois cela fait, la distance de Jaccard est obtenue uniquement par la soustraction du coefficient de Jaccard de 1

$$\textbf{Jaccard distance(A, B) = 1 − Jaccard(A, B)}$$

d. Distance de Hamming :

La distance de Hamming est une mesure de la différence entre deux chaînes de caractères de même longueur. Contrairement à la distance de Jaro-Winkler, elle ne produit pas un résultat continu entre 0 et 1, mais plutôt un nombre entier qui représente le nombre de positions où les symboles correspondants diffèrent.

Pour calculer la distance de Hamming entre deux chaînes vous pouvez suivre une méthode mathématique simple :

Assurez-vous que les deux chaînes ont la même longueur. Si elles sont de longueurs différentes, la distance de Hamming n'est pas définie.

Parcourez les deux chaînes caractère par caractère, en comparant les caractères correspondants à la même position.

Chaque fois que les caractères correspondants diffèrent, incrémentez un compteur de distance.

Une fois que vous avez parcouru toutes les positions dans les deux chaînes, le compteur de distance représente la distance de Hamming entre les deux chaînes.

Mathématiquement, cela peut être exprimé comme suit :

$$\text{distance\_de\_Hamming} = \sum_{i=1}^{n} \begin{cases} 1, si\ s1[i] \neq s2[i] \\ 0, sinon \end{cases}$$

Où n est la longueur des chaînes $s1$ et $s2$, et $s1[i]$ et s2[i] représentent les caractères correspondants à la position i dans les deux chaînes.

Cette méthode fournit un moyen simple et efficace de quantifier la différence entre deux chaînes de caractères, ce qui est utile dans divers domaines tels que la détection et la correction d'erreurs en transmission de données ou l'évaluation de la similarité entre des séquences génétiques.

3.3.2. Encodage phonétique (Phonetic Encoding):

Soundex est considéré comme l'un des encodages phonétiques les plus efficaces les fonctions.

Il transforme les chaînes selon la manière dont elles sont prononcées. ils peuvent être comparés les uns 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 :

- Conserver la première lettre de la chaine.
- Remplacer toutes les consonnes en respectant les règles suivantes :

A, E, H, W, O, U, Y, I par 0 ; B, F, P, V par 1 ; C, G, K, Q, S, X, Z par 2 ; D, T par 3 ; L par 4 ; M, N par 5 ; R par 6

- Dans le cas où la chaine est trop courte l'algorithme remplace les trois premiers nombres après le premier caractère par des zéros.

D'où cet algorithme codifie chaque donnée de cette manière et si le résultat est le même donc les données sont similaires.

NYSIIS (New York State Identification Intelligence System):

Est un algorithme d'encodage phonétique développé en 1970 dans le cadre du système d'identité de l'État de New York. Utilisé afin de détecter les mots similaires.

NYSIIS a la même idée et le même objectif que l'algorithme Soundex. La différence est que NYSIIS renvoie un code composé de lettres qui estce n'est pas le cas de Soundex. L'algorithme NYSIIS augmente la précisionde 2,7 % par rapport à Soundex [16]. Les bases des règles 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 à FF, SCH à SSS) et les derniers caractères (EE-IE à Y, DT-RT RD-NT-ND à D).

3.4. Clustering :

3.4.1. Définition :

Appelé le partitionnement de données est une méthode qui vise à diviser un ensemble de données en différents paquets homogènes de manière est ce que chaque sou ensemble partage des caractéristiques communes.

Le clustering est utilisé notamment lorsqu'il est coûteux d'étiqueter les données. C'est néanmoins un problème mal défini mathématiquement : différentes métriques et différentes représentations des données aboutiront à différents regroupements sans qu'aucun ne soit nécessairement meilleur qu'un autre. Ainsi la méthode de clustering doit être choisie avec soin en fonction du résultat attendu et de l'utilisation prévue des données [5].

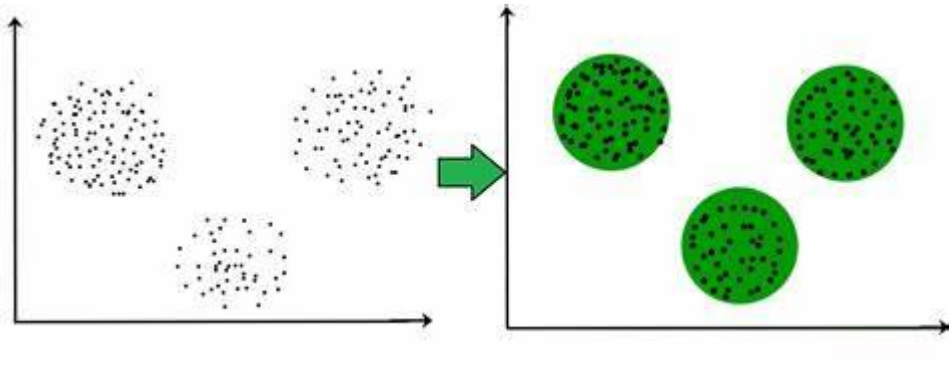

Figure 6: Exemple de Clustering

Avoir une méthode de clustering est soumis à de multiples exigences :

- Possibilité de découvrir tout ou une partie des clusters cachés.
- Similitude intra-cluster et dissimilarité inter-cluster.
- Capacité a géré différents types d'attributs.
- Peut gérer le bruit et les valeurs aberrantes.
- Prut gérer une haute dimensionnalité.
- Evolutif interprétable et utilisable.

3.4.2 Algorithmes de clustering :

Il existe plusieurs types d'algorithmes de clustering, ils sont classés entre hiérarchiques et non hiérarchiques comme illustré dans cette..image:

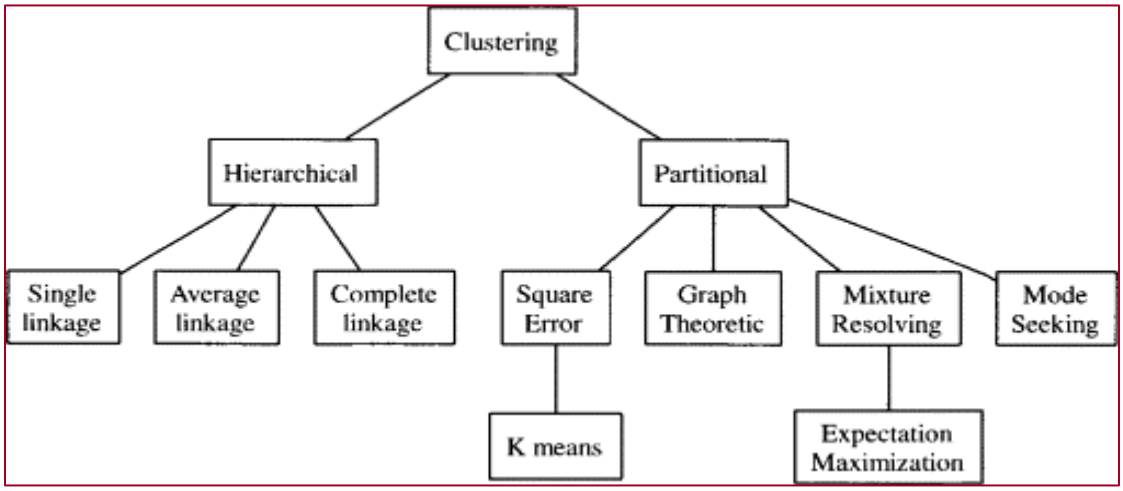

Figure 7: Algorithme de clustering

Les clés d'enregistrements des clusters C1 est celui de E1, C2 est E2, C3 est E3

Nous allons commencer la comparaison entre les clusters et les enregistrements, nous aurons cette première itération :

	C1	**C2**	**C3**
Itération 01	E1 (BEAb1993) E4 (LAIm1988) E9 (LAIm1988)	E2 (TIHa1993) E6 (THHa1993) E7 (KHHa1993) E8 (TIHa1993)	E3 (BEHa1993) E5 (BEHa1993)

Après cette itération nous devrions choisir de nouveau clusters et pour le faire il faut voir les enregistrements ou la clé de blocage est la même qui se répète, comme deuxième itération on aura :

	LAIm1993	**TIHa1993**	**BEHa1993**
Itération 02	E1(BEAb1993) E3 (BEHa1993) E5(BEHa1993) E7 (KHHa1993)	E2 (TIHa1993) E6(THHa1993) E8 (TIHa1993)	E4(LAIm1988) E9(LAIm1988)

Après chaque itération on répète la même condition pour trouver les nouveaux clusters jusqu'à ce que le résultat soit le même à chaque fois dans cet exercice on s'arrête dans la deuxième itération car la prochaine sera égale à celle-ci d'où le résultat final est dans le tableaux.

Tableau comparatif entre les approches de clustering :

Approche	Avantages	Inconvénients

Partitionnement (k-Means)	- La simplicité. - La complexité de l'algorithme est calculée en O (k.n) - Adapté aux données énormes. - Il est fait pour les données numériques.	- Le nombre de clusters doit être fixé au début. - Le résultat dépend du choix des centroides. - Les données bruitent (isolées) sont négligés.
Partitionnement (k-Modes)	- Il traite les objets isolés. - La robustesse avec la présence du bruit - Spécifié pour les données catégoriques.	- **Difficulté de détermination du nombre optimal de clusters (K).**

Table 3: tableau comparatif entre les algorithmes de clustring

3.5 Conclusion :

Dans ce chapitre nous avons vu de plus près les méthodes utilisées pour limiter les erreurs au niveau des bases de données et pour assurer la qualité de données, parmi ces dernières l'approche la plus connu est le couplage d'enregistrements (L'appariement de données) ainsi que le Clustering, nous avons choisi d'utilisé l'algorithme k-Modes.

Chapitre 04 : Contribution

4.1. Introduction :

Dans ce chapitre, nous présentons nos contributions au domaine du L'appariement de données . Une solution est proposée pour chacun des défis abordés dans le chapitre d'introduction, nous commençons par présenter une nouvelle approche de détection de doublons récemment proposée dans la littérature 'Le couplage d'enregistrement basé sur le K-Modes'. Ensuite, nous présentons notre contribution qui représente une amélioration de l'approche de RL proposé dans [1]. la nouvelle méthodologie est proposé de contrôler les tailles des blocs générés. Cette méthodologie, rend le RL basé sur K-Modes adapté aux applications en temps réel telles que le problème du RL en temps réel.

4.2 Algorithme K-modes :

Est une technique non supervisée, automatique utilisé pour regrouper un ensemble de données dans un nombre spécifié de clusters en fonction de leurs attributs catégoriels [8].

4.2.1 Les étapes :

1. Sélectionner K modes initiaux un pour chaque cluster.

2. Allouer un objet au cluster le plus proche.

3. Utiliser la méthode basée sur la fréquence pour mettre à jour les modes du cluster après chaque attribution.

4. Répéter les étapes 2 et 3 jusqu' à ce qu'aucun objet n'ait changé de cluster après un test de cycle complet de l'ensemble de données.

4.2.2 Un exemple d'application de l'algorithme :

Dans cet exercice nous allons faire une application de l'algorithme sur la base de données :

Nom	**Prénom**	**Datede naissance**	**Numéro de téléphone**
BEKHALED	Abdhamid	15/01/1993	+213777775896
MEHENNAOUI	Nadir	01/01/1993	+213665655665
BEN KHALED	Hamid	15-02-1993	0777-77-58-96
LABANI	Imad	25-05-1988	0771677689
BENKHALED	Hamad	15-01/1993	0777775896
MEHENAOUI	Nadir	01011993	0665655665
KHALED	Hamid	15011993	0777775896
MEHENNAOUI	Nadir	0101993	0665-65-56-65
LARBANI	Imad	25051988	0771-67-76-89

Table 4: Base de données initiale

La première étape consiste à normaliser la date de naissance ainsi que le numéro de téléphone c'est-à-dire les mettre sous la même forme, donc on aura la base suivante après normalisation :

Nom	**Prénom**	**Datede naissance**	**Numéro de téléphone**
BEKHALED	Abdhamid	15/01/1993	0777-77-58-96
MEHENNAOUI	Nadir	01/01/1993	0665-65-56-65
BEN KHALED	Hamid	15/01/1993	0777-77-58-96
LABANI	Imad	25/05/1988	0771-67-76-89
BENKHALED	Hamad	15/01/1993	0777-77-58-96
MEHENNAOUI	Nadir	01/01/1993	0665-65-56-65
KHALED	Hamid	15/01/1993	0777-77-58-96
MEHENNAOUI	Nadir	01/01/1993	0665-65-56-65
LARBANI	Imad	25/05/1988	0771-67-76-89

Table 5: Base de données avec normalisation

La deuxième étape consiste à définir trois clusters aléatoirement, pour cet exercice nous avons choisi :

C1 :

BKHAKED	HAMID	02/05/1999	0665-66-69-20

C2 :

MEHENNAOUI	NADIR	18/5/1993	0765-66-69-20

C3 :

BEN KHALED	HAMID	14/08/1993	0765-86-69-20

Après avoir choisi ces derniers, nous devons faire la comparaison avec les autres enregistrements en utilisant la méthode distance d'édition (edit distance).

Pour le faire il faudra définir une clé de blocage pour chaque enregistrement, pour le faire nous avons choisi comme clé de blocage :

Blocking key : 2(Nom) + 2(Prénom) + année de naissance

Les clés de blocage de tous les enregistrements :

Enregistrement	ClédeBlocage
E1	BEAb1993
E2	MEHa1993
E3	BEHa1993
E4	LAIm1988
E5	BEHa1993
E6	MENa1993
E7	KHHa1993
E8	MENa1993
E9	LAIm1988

Table 6: les clés de blocage de chaque enregistrement

Exemple du K-mode :

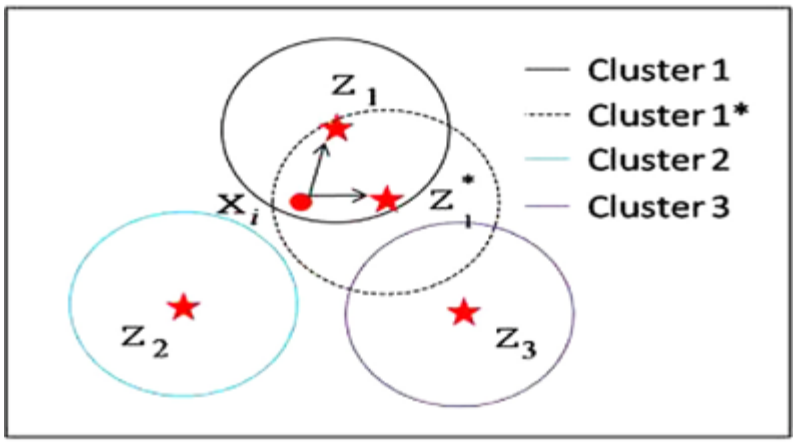

Figure 8: Example du K-Mode

3.4 L'appariement de données basé sur l'algorithme k-modes :

Dans cette section, nous présentons une nouvelle approche évolutive dans le domaine de doublons détectés. Notre approche est basée sur l'algorithme K-Modes qui s'occupe de la division des données en sous-ensembles appelés les clusters. Chacun de ses derniers regroupe les enregistrements permettant référencer une entité réelle [1].

Une fois le clustering est effectué, le filtrage adaptatif est exécuté en post-traitement étape. La plupart des enregistrements qui se trouvent dans le même cluster mais ne représentent pas la même entité sera ignorée lors de la phase de matching. Enfin, tout les enregistrements restants du même bloc sont comparés les uns aux autres à l'aide d'un ensemble de métriques de similarité de chaînes qui existent déjà dans la littérature.

Le mécanisme de cette approche est le suivant :

En entrée, on a un ensemble de données qui doit être normalisé, c'est-à-dire que toutes les données doivent respecter un format de données précis. Dans cet ordre, l'algorithme k- modes prend en charge la division des données dans les k clusters.

Les clés de bloc seront définies pour être en mesure de répartir les enregistrements dans différents blocs et de passer à la deuxième étape du RL, qui est Matching.

À ce niveau-là des métriques de similarité (Jaro-Winkler, Jaccard Similarity...) seront appliqué sur toutes les correspondances de chaque bloc pour pouvoir extraire les vraies et les fausses correspondances.

4.4 Méthode proposées :

En étudiant l'algorithme existant, RL basé sur le k-modes, on remarque que la taille des blocs est ignorée. Hormis qu'elle augmente le temps d'exécution, à titre d'exemple la recherche d'une info précise dans plusieurs lignes, ceci va prendre énormément de temps vu que le nombre de comparaisons est important.

Ceci a fait qu'on propose une amélioration sur l'algorithme précèdent en appliquant un control sur les tailles des blocs générés par le k-modes. Pour concrétiser cette solution on a proposé deux étapes principales.

Pour contrôler la taille des blocs, notre approche utilise comme entrée les blocs générés par le RL basé sur le k-modes.

Ensuite, tous les blocs qui ont une taille supérieure à un seuil MaxSize prédéfini par l'utilisateur sont divisée en nouveaux blocs à l'aide du blocage standard avec une autre clé de blocage.

Nous continuons à diviser les blocs générés avec une taille supérieure au seuil MaxSize jusqu'à ce qu'aucun bloc B avec une taille $|B| > MaxSize$ reste.

Une fois la phase de division terminée, nous commençons à fusionner tous les blocs qui ont une taille inférieure à une valeur prédéfinie MinSize.

Chaque bloc bi est fusionné avec le prochain plus petit bloc bj s'ils vérifient les conditions suivantes :

- $|bi| + |bj| \leq MaxSize$.
- La similarité entre les modes de bi et bj est maximisée

4.4.1 La division (fractionnement) des blocs :

L'algorithme en dessous représente la première étape de notre approche, qui est la division des blocs. En entrée, nous avons la taille maximale des blocs MAXsize et l'ensemble de blocs générés par le RL basé sur le k-modes (B +) avec une taille supérieure à MAXsize.

Le principe est très simple, la fonction « diviser » commence par la création des nouveaux blocs Bnew en utilisant le blocage standard jusqu'à ce qu'il ne reste plus de bloc dans la liste des blocs B+. Ensuite, les blocs nouvellement générés sont assignés vers des listes selon leurs tailles :

- Si la taille du bloc dépasse MAXsiz, il sera ajouté de nouveau à la liste B+.
- Si sa taille est inférieure à MINsize, il sera ajouté à la liste Bmin .
- Sinon il reste dans la liste des blocs avec la taille correcte.

Algorithme diviser

1 : **Entrées :**

- La taille maximale des blocs : MAX_{size}
- La taille minimale des blocs : MIN_{size}
- Les blocs B^+ générés par le K-Modes avec une taille $\geq MAX_{size}$
- Une liste supplémentaire des clés de blocage : K

2 : **Sorties :** des blocs B^{min} avec une taille $\leq MIN_{size}$
3 : **Fonction** diviser
4 : **TantQue (**$B^{min} \neq$ nul) **faire**
5 : **Pour chaque** (membre B dans B^+) **faire**
6 : B^{new} = blocage-standard (B, new K)
7 : **Pour chaque** (membre C dans B^{new}) **faire**
8 : **Si** ($C \geq MAX_{size}$) **faire**
9 : B^+. Insert (C)
10 : **Sinon Si** ($C \leq$ MINsize) **faire**
11 : B^{min}.Insert (C)
12 : **Fin Si**
13 : **Fin Pour**
14 : **Fin Pour**
15 : **Fin TantQue**
16 : Retourner B^{min}
17 : **Fin Fonction**

4.4.2. La fusion des blocs :

L'algorithme en dessous représente la deuxième étape de notre approche qui fonctionne en triant et fusionnant les petits blocs jusqu'à ce qu'ils atteignent une taille acceptable. Il combine des blocs trop petits pour les rendre plus grands et vérifie après chaque fusion si la taille du nouveau bloc est acceptable. Si un bloc reste trop petit après une tentative de fusion, il est réinséré dans la liste pour une future fusion. Ce processus continue jusqu'à ce que tous les blocs aient une taille acceptable

Son but est de fusionner tous les blocs qui ont une taille inférieure à MINsize en respectant deux condition :

- La taille du nouveau bloc généré après la fusion ne dépasse pas le seuil MAXsize.
- La similarité entre les deux modes des blocs générés est maximisée.

Algorithme fusioner

---

1 : **Entrées :**

- La taille maximale des blocs : MAX_{size}
- La taille minimale des blocs : MIN_{size}
- Les blocs B^{-} générés par le K-Modes avec une taille $\leq MIN_{size}$
- Les blocs générés après l'étape de la fusion : B^{min}
- Les blocs $B^{correct}$ générés par le K-Modes avec une taille acceptable
- Une liste supplémentaire des clés de blocage : K

2 : **Sorties :** des blocs $B^{correct}$ avec une taille acceptable
3 : **Fonction** fusion
4 : B^{merged} = Sorted (B^{-} U B^{min})
5 : **TantQue (**$B^{merged} \neq$ NUL) **Faire**
6 : **Pour chaque** (membre B dans B^{merged}) **faire**
7 : B_j = FindClosest (B, B^{merged} U $B^{correct}$)
8 : B_{ij} = MergeBlocks (B, B_j)
9 : **Si** ($B_{ij} \leq MIN_{size}$) **faire**
10 : B^{merged}.Insert (B_{ij})
11 : **Sinon**
12 : $B^{correct}$.Insert (B_{ij})
13 : **Fin Pour**
14 : **Fin TantQue**
15 : **Retourner** $B^{correct}$
16 : **Fin Fonction**

4.4.3. L'Algorithme générale :

L'algorithme général suit une séquence logique de préparation, division et fusion pour transformer les blocs générés par K-Modes en blocs de taille acceptable. La préparation consiste à classer les blocs par taille, la division s'occupe des blocs trop grands, et la fusion ajuste les blocs trop petits. Le résultat final est une collection de blocs ayant tous une taille comprise entre MINsize et MAXsize

L'approche générale peut être résumée par l'algorithme général suivant qui fait appel au deux fonctions précédemment discutées.

Algorithme générale

1 : **Entrées :**

- La taille maximale des blocs : MAX_{size}
- La taille minimale des blocs : MIN_{size}
- Les blocks générés par K-Modes : BK.
- Une liste des clés de blocage : K

2 : **Sortie :** des blocs avec une taille acceptable

3 : **Début**

4 : B^{-}, B^{+}, $B^{correct}$ = prepare (BK, MAX_{size}, MIN_{size}

5 : B^{min} = Diviser (B^{-}, B^{min}, MAX_{size}, MIN_{size}, K)

6 : $B^{correct}$ = Fusion (B^{-}, B^{min}, MAX_{size}, MIN_{size}, K)

7 : **Retourner** ($B^{correct}$)

16 : **Fin**

Chapitre 05 : Validation et Résultats

5.1. Introduction :

Afin de pouvoir mettre en œuvre notre approche qui est étudié et amélioré dans les chapitres précédents nous avons rapporté un ensemble de données du monde réel. On a choisi d'utilisé un jeu de données parmi ceux qui ont été utilisé pour évaluer la plupart des approches du L'appariement de données s existants dans la littérature.

Le dataset utilisé est « restaurant », il contient des enregistrements rapportés des restaurants des États Unis par les deux guides : Fodor et Zagat.

Dans ce jau de données il contient 864 enregistrements et 112 enregistrements dupliqués.

Example des attribus du Base De Donnes « restaurant » :

name_phone_id	city_phone	phone_id	name	addr	city	phone	type	class
A655310 2461501	LASANG 435	31024 61501	arnie morton's	435 s. la cienega blv.	los angeles	310/246-1501	american	0

5.2. Environnement matériel et logiciel :

5.2.1 Environnement matériel :

- Processeur : Intel® Core™ i5 -7200U CPU @ 2,50 GHz 2,70 GHz.
- Type de système : système d'exploitation 64 bits processeur x 64.
- Mémoire installée (RAM) : 8,00 Go (7,82 Go utilisable).

5.1.2 Environnement logiciel :

Nous avons utilisé comme logiciel de programmation l'environnement de développement nommé éclipse. Ce dernier et une ide libre et intégrée et lancer par IBM, il permet de créer des

projets de développements en utilisant n'importe quel langage de programmation mais il est principalement écrit en java.

Figure 9: Logo Eclipse

5.2.3. Langage de programmation :

Pour le language de programmation nous avons utilisé le : java qui est un langage de programmation orienté objet, il est connu depuis 1995, il est l'un des langages les plus connu et utilisé à travers le monde entier.

5.3. Métriques d'évaluation :

Pour évaluer la performance d'une approche proposée, trois paramètres qui sont généralement utilisés et qui sont :

Ration de réduction :

Elle est utilisée après avoir appliqué une approche/clustering pour certaines vérifications, elle est utilisée pour mesurer dans quelle mesure la technique de blocage a réussi à réduire le nombre de comparaisons entre les enregistrements.

La formule pour calculer RR est la suivante :

$$RR = 1 - \frac{nombres\ de\ comparaison\ après\ le\ blocage}{nombres\ de\ comparaison\ avant\ le\ blocage}$$

PC (pair completeness) :

Est une équation utilisée après avoir appliqué l'algorithmes du clustering pour calculer le nombre de valeurs dupliqués, sa formule :

$$PC = \frac{nombre\ de\ paire\ d'enregistrements\ détéctés}{nombre\ de\ paire\ d'enregistrements\ en\ double\ dans\ le\ dataset}$$

F-score :

Est utilisé pour contrôler le compromis entre PC et RR, elle est définie comme la moyenne harmonique entre ces deux dernières métriques, sa formule :

$$f_{mesure} = 2 * \frac{RR * PC}{RR + PC}$$

5.4. Résultats :

5.4.1. Le nombre des doublons détectés (Avant et après) :

En lançant l'étape du Matching avant et après notre contribution (Avec les trois métriques de similarités : distance d'édition, Jaro-Winkler et Jaccard)

a. Avant la division et la fusion des clusters : 70-120 :

		Jaccard		Jaro-Winkler		Edit distance		Hamming Distance	
K	RR	PC	F-Score	PC	F-Score	PC	F-Score	PC	F-Score
5	**0,74**	**0,96**	**0,84**	**0.99**	**0,84**	**0.97**	**0,83**	**0.95**	**0.85**
8	**0,88**	**0,92**	**0,90**	**0,92**	**0,90**	**0,91**	**0,89**	**0.92**	**0.91**
10	**0,91**	**0,86**	**0,88**	**0,88**	**0,89**	**0,87**	**0,89**	**0.86**	**0.88**

Table 7: Nombre des doublons détectés avant la division et le fusionnement

b. Après la division et la fusion des clusters :

		Jaccard		Jaro-Winkler		Edit distance		Hamming Distance	
K	RR	PC	F-Score	PC	F-Score	PC	F-Score	PC	F-Score
5 =>9	**0,89**	**0,91**	**0,90**	**0,89**	**0,89**	**0,88**	**0,88**	**0.87**	**0.88**
8 => 8	**0,87**	**0,90**	**0,88**	**0,91**	**0,88**	**0,87**	**0,87**	**0.91**	**0.89**
10 => 9	**0,88**	**0,92**	**0,89**	**0,92**	**0,89**	**0,89**	**0,88**	**0.90**	**0.88**

Table 8: Nombre des doublons détectés après la division et le fusionnement

Ces résultats indiquent les performances de la méthode de détection de doublons avant et après l'application de la division et de la fusion des clusters, en utilisant différentes mesures de similarité.

En examinant les tableaux fournis, dans ce qui suit, une comparaison des performances avant et après la division et la fusion des clusters pour chaque mesure de similarité :

• Pour la mesure Jaccard :

Avant la division et la fusion des clusters : Les valeurs de K (5, 8, 10) montrent généralement un taux élevé de RR (Rappel Réciproque) et de F-Score, indiquant que le nombre de comparaisons est plus élevé, ainsi que le nombre de doublons détectés.

Après la division et la fusion des clusters : Les valeurs de K (5 => 9, 8 => 8, 10 => 9) montrent des performances légèrement inférieures en termes de taux de RR et de F-Score, indiquant que le nombre de comparaisons est moins élevé, ainsi que le nombre de doublons détectés.

• Pour la mesure Jaro-Winkler :

Avant la division et la fusion des clusters : Les valeurs de K (5, 8, 10) montrent généralement un taux élevé de RR et de F-Score, indiquant que le nombre de comparaisons est plus élevé, ainsi que le nombre de doublons détectés.

Après la division et la fusion des clusters : Les valeurs de K (5 => 9, 8 => 8, 10 => 9) montrent des performances légèrement inférieures en termes de taux de RR et de F-Score, indiquant que le nombre de comparaisons est moins élevé, ainsi que le nombre de doublons détectés.

• Distance d'édition :

Avant la division et la fusion des clusters : Les valeurs de K (5, 8, 10) montrent généralement un taux élevé de RR et de F-Score, indiquant que le nombre de comparaisons est plus élevé, ainsi que le nombre de doublons détectés.

Après la division et la fusion des clusters : Les valeurs de K (5 => 9, 8 => 8, 10 => 9) montrent des performances légèrement inférieures en termes de taux de RR et de F-Score, indiquant que le nombre de comparaisons est moins élevé, ainsi que le nombre de doublons détectés.

• Hamming Distance :

Avant la division et la fusion des clusters : Les valeurs de K (5, 8, 10) montrent généralement un taux élevé de RR et de F-Score, indiquant que le nombre de comparaisons est plus élevé, ainsi que le nombre de doublons détectés.

Après la division et la fusion des clusters : Les valeurs de K (5 => 9, 8 => 8, 10 => 9) montrent des performances légèrement inférieures en termes de taux de RR et de F-Score, indiquant que le nombre de comparaisons est moins élevé, ainsi que le nombre de doublons détectés.

En général, il semble que la détection de doublons et le nombre de comparaisons avant la division et la fusion des clusters présentent des performances légèrement supérieures ou similaires à celles après.

Cependant, il est important de noter que les différences observées peuvent varier en fonction de l'ensemble de données spécifique et des seuils de similarité utilisés.

Mais une chose à retenir :

- Si RR est élevé, signifie que le nombre de comparaison est réduit.
- Si PC est faible, signifie que le taux de détection de doublons à diminuer.

5.4.2. Changement dans les valeurs de MaxSize et MinSize :

En lançant l'étape du Matching avec de différents seuils de MaxSize et MinSize en utilisant Jaccard Similarité, comme suit :

	MaxSize et MinSize								
	50-100			70-120 (9)			120-150		
K	RR	PC	F-Score	RR	PC	F-Score	RR	PC	F-Score
5	**0,90**	**0,86**	**0,87**	**0,89**	**0,89**	**0,89**	**0,82**	**0,91**	**0,86**
8	**0,91**	**0,85**	**0,87**	**0,87**	**0,91**	**0,88**	**0,79**	**0,90**	**0,84**
10	**0,91**	**0,86**	**0,88**	**0,88**	**0,92**	**0,89**	**0,80**	**0,89**	**0,84**

Table 9 : Résultats du changement dans les valeurs de MaxSize et MinSize

• Pour les RR (Rappel Réciproque) et PC (Précision) :

Les performances varient en fonction des valeurs de K et des configurations de MaxSize et MinSize, sans montrer de tendance claire.

Dans certains cas, une configuration de MaxSize plus élevée peut conduire à un taux de comparaison plus élevé, tandis qu'une configuration de MinSize plus élevée peut conduire à augmenter le nombre de doublons détectés. Cependant, ces différences ne sont pas cohérentes dans tous les scénarios.

• Pour la métrique F-Score :

La configuration 70-120 (9) donne généralement un taux de résultats plus élevé pour toutes les valeurs de K (5, 8, 10). Les F-Scores obtenus sont respectivement de 0,87, 0,88 et 0,89.

La configuration 50-100 donne des F-Scores légèrement moins élevés pour toutes les valeurs de K.

La configuration 120-150 donne des F-Scores variables, mais généralement inférieurs aux deux autres configurations.

Ces résultats indiquent que la configuration 70-120 semble être la meilleure parmi les trois options proposées en termes de F-Score, car le nombre de comparaisons est le plus faible.

5.4.3. Nombres des clusters selon les valeurs de MaxSize et MinSize :

En lançant l'étape du Matching avec de différents seuils de (MaxSize /MinSize) et nombre des blocks générés par le K-Modes en utilisant Jaccard Similarité, comme suit :

Nombres des blocks générés par le K-Modes	70-120	50-100	120-150
5	**9**	**10**	**6**
8	**8**	**11**	**5**
10	**9**	**11**	**5**

Table 10 : Nombres des clusters selon les valeurs de MaxSize et MinSize

On peut observer que les nombres de clusters varient en fonction des valeurs de MaxSize et MinSize. Voici quelques observations possibles :

• Pour K = 5, la configuration 50-100 génère le plus grand nombre de clusters (10), suivie de près par la configuration 70-120 (9). La configuration 120-150 produit le plus petit nombre de clusters (6).

• Pour K = 8, la configuration 50-100 génère le plus grand nombre de clusters (11), tandis que les configurations 70-120 et 120-150 produisent respectivement 8 et 5 clusters.

• Pour K = 10, les configurations 50-100 et 70-120 produisent le plus grand nombre de clusters (11), tandis que la configuration 120-150 génère le plus petit nombre de clusters (5).

Ces résultats mettent en évidence l'impact des valeurs de MaxSize et MinSize sur le nombre de clusters générés. Des valeurs plus élevées de MaxSize et MinSize tendent à conduire à un nombre plus élevé de clusters, tandis que des valeurs plus basses peuvent réduire le nombre de clusters formés.

5.4.4. Comparaison du temps d'exécution :

Ces résultats nous montrent l'évolution du temps d'exécution de l'étape de Matching avant et après les deux phases division et fusion.

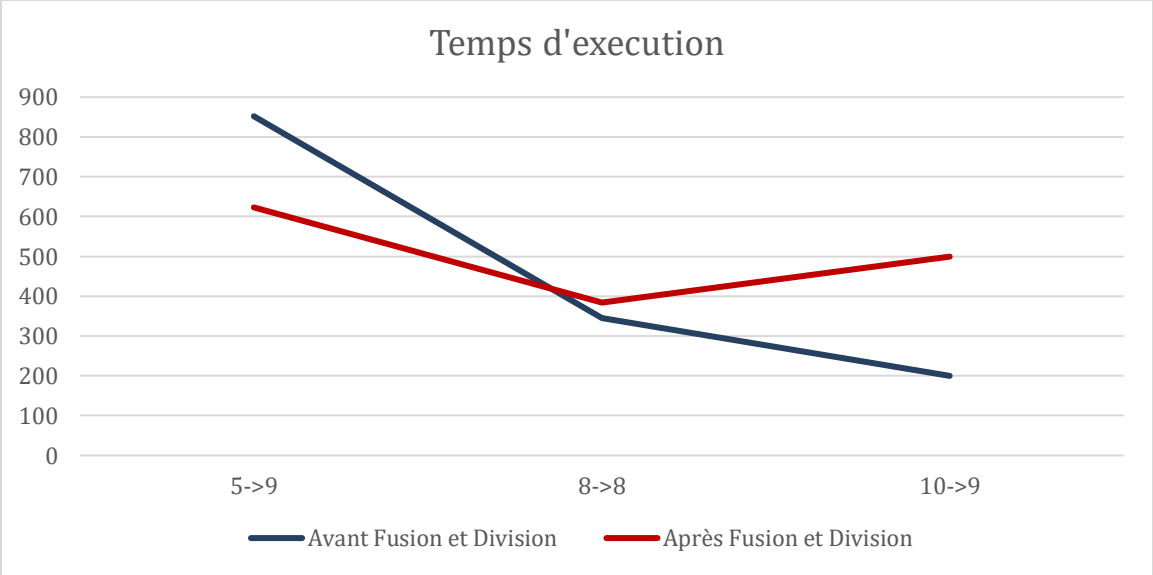

Figure 10: Courbe qui représente l'évolution du temps d'exécution

Dans cette courbe nous remarquons qu'avant la fusion et la division le temps d'exécution diminue, de plus le nombre de clusters est grand le Matching ne prend pas beaucoup de temps.

D'où le temps d'exécution change selon le nombre de clusters en sortie Plus le nombre de clusters en sortie est grand le temps d'exécution s'améliore.

6. Conclusion générale

La qualité de donnée est un concept très important et qui a un très grand impact sur le développement des entreprises c'est pour ça que ces dernières visent toujours à s'assurer que son système d'information est de qualité et cherchent des méthodes pour assurer cette qualité.

Précédemment nous avons vu quatres chapitres dont le premier s'étale sur la définition de la qualité de données ainsi que ses critères, les causes de la non qualité ainsi que d'autre points la concernant.

Dans le deuxième chapitre nous avons présenté une des méthodes de la littérature qui permet de détecter les doublons dans une base de données qui est le couplage d'enregistrement (L'appariement de données) ainsi que les différents algorithmes de clustering.

Le chapitre contribution représente ce que nous avons apporté en plus ou comme amélioration pour le L'appariement de données basé sur l'algorithme k-modes.

L'amélioration que nous avons apportée à ce dernier est le contrôle de la taille des blocs des clusters en sortie de l'approche précédente, nous avons ajouté deux méthodes qui sont : la fusion et la division des blocs selon une taille Maxsize et Minsize que nous avons initialisés.

Dans le dernier chapitre pour finir nous avons expérimenté nos méthodes selon différentes métriques d'évaluation et nous avons constaté que notre contribution améliore beaucoup le L'appariement de données et aide à détecter le plus grand nombre possible de doublons.

Comme futurs travaux nous proposons de trouver une méthode qui permet de déterminer automatiquement les valeurs de Maxsize et Minsize correspondantes, afin de résoudre le cas ou il reste toujours un cluster dans LMn dont sa taille est inférieure à Minsize.

REFERENCES :

[1] Benkhaled, h. N.,Berrabah, d., & Boufares, f. (2019, April). A novel approach to improve the record linkage process. In 2019 6th international conference on control, decision and information technologies (codit) (pp. 1504-1509). IEEE.

[2] Thomas c Redman. bad data costs the us $3 trillion per year. harvard business review, vol. 22, pages 11–18, 2016.

[3] Jonathan Geiger. Data quality management, the most critical initiative you can implement. Data warehousing, management and quality, paper, pages 098–29, 2004.

[4] Kralil mMeriem & Sehini Khadîdja, Mémoire de Master « La sélection automatique d'une clé de blocage a l'aide d'une méta heuristique », année 2020, Université de Sidi Bel Abbes.

[6] BENKHALED, H. N. (2021). La qualité des données dans le contexte Big Data (Doctoral dissertation), Université de Sidi Bel abbés

[7] Abdelkrim Ouhab, Mimoun Malki, Djamel Berrabah et Faouzi Boufares. An unsupervised entity resolution framework for english and arabic datasets. International Journal of Strategic Information Technology and Applications (IJSITA), vol. 8, no. 4, pages 16–29, 2017

[8] hexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, vol. 2, no. 3, pages 283–304, 1998

[9] AÏCHA BEN SALEM AND AL. SEMANTIC RECOGNITION OF A DATA STRUCTURE IN BIG-DATA. JOURNAL OF COMPUTER AND COMMUNICATIONS, VOL. 2, NO. 09, PAGE 93, 2014.

[10] CARLO BATINI AND MONICA SCANNAPIECO. DATAAND INFORMATION QUALITY. SPRINGER INTERNATIONAL PUBLISHING., PAGE 43, 2016.

[11] THOMAS C REDMAN. THE IMPACT OF POOR DATA QUALITY ON THE TYPICAL ENTERPRISE. COMMUNICATIONS OF THE ACM, VOL. 41, NO. 2, PAGES 79–82, 1998

[12] El Akkaoui, Z., & Zimányi, E. (2009, November). Defining ETL worfklows using BPMN and BPEL. In Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP (pp. 41-48

[13] HOUDA ZAIDI, FAOUZI BOUFARES AND YANN POLLET. NETTOYAGE DE DONNEES GUIDE PAR LA SEMANTIQUE INTER-COLONNES. IN EGC, PAGES 549–550, 2016.

[14] VLADIMIR I LEVENSHTEIN. BINARY CODES CAPABLE OF CORRECTING DELETIONS, INSERTIONS, AND REVERSALS. IN SOVIET PHYSICS DOKLADY, VOLUME 10, PAGES 707–710, 1966.

[15] Peter Christen. A survey of indexing techniques for scalable record linkage and deduplication. IEEE transactions on knowledge and data engineering, vol. 24, no. 9, pages 1537–1555, 2011.

[16] Rajkovic and D Jankovic. Adaptation and application of Daitch-Mokotoff Soundex algorithm on Serbian names. In XVII Conference on Applied Mathematics, volume 12, 2007.

[17] Mauricio A Hernández and Salvatore J Stolfo. The merge/purge problem for large databases. ACM Sigmod Record, vol. 24, no. 2, pages 127–138, 1995.