الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
جامعة سعيدة د. مولاي الطاهر
كلية التكنولوجيا
قسم: الإعلام الآلي

Mémoire de Master

Spécialité : Réseaux Informatiques et Systèmes Répartis

Thème

La sélection semi-automatique des attributs lors de la mise en correspondance (Matching) pour le couplage d'enregistrements.

Présenté par :

SOUILEM Abdelkrim

BAALI Bendjelloul

Dirigé par :

Dr. BENYAHIA Miloud

Année universitaire 2022-2023

Remerciements

Nous remercions Dieu tout puissant de la patience et de la volonte qu'il nous a donné pour réaliser ce projet de fin d'etude.

Nous tenons a remercier vivement notre encadreur monsieur ' ***BENYAHIA Miloud*** *pour son aide, ses conseils, son encouragement.*

Nous voulons aussi remercier tout ceux qui ont Contribue a Réaliser ce projet;

Nos professeurs qui nous ont enseigné, Nos amis

Les membres de jury de notre travail.

La famille ***SOUILEM*** *et* ***BAALI***

Dédicaces

Je dedie ce travail a tous ceux qui m'ont encouragé et soutenu

A la mémoire de mon père qu'allah le Accueille dans son vaste paradis, ma chère mère, que Dieu la protège.

A mes soeurs et frères

A ma chère épouse, et à mes princesses Ayet, Meriem et Malak

A mes amis

A tous ceux que j'aime

⋆ ⋆ ⋆

SOUILEM

Dédicaces

Je dédie ce modeste travail :

A la mémoire de mon père qu'allah le Accueille dans son vaste paradis, ma chère mère, que Dieu la protège.

A mes frères et sœurs

A mon épouse Madame ***BAALI***

A mes amis

Et à tous ceux qui ont contribué de près ou de loin pour que ce travail soit possible, je vous dis merci.

⋆ ⋆ ⋆

BAALI

Contents

Liste des Figures

Liste des tableaux

Table des sigles et acronymes

DW	Data Warehouse
IBM	International Business Machines
BI	Business Intelligence
OCR	Optical Character Recognition
ETL	Extraction Transformation Loading
DQAF	Data Quality Assessment Framework
FMI	Fonds monetaire international
RL	Record Linkage
BKV	Blocking Key Value
BK	Blocking Key
NYSIIS	New York State Identification Intelligence System
IDE	Integrated Development Environment
JRE	Java Runtime Environment
JDK	Java Development Kit
XML	Extensible Markup Language
HTML	Hypertext Markup Language
PHP	Hypertext Preprocessor

Introduction Générale

Au cours des dernières décennies L'analyse du Big Data est devenu de plus en plus un système d'aide a la décision fiable, en particulier avec l'adoption massive des médias sociaux et des smartphones, qui génèrent une énorme quantité de données que les DW traditionnels ne peuvent pas traiter. Par ailleurs, bien que les systèmes DW et Big Data Analytics soient désormais considérés comme les meilleurs systèmes d'aide a la décision, ces deux technologies peuvent parfois ne pas répondre aux attentes des parties prenantes.
En fait, de nombreuses mauvaises décisions ont été prises, entraînant de mauvaises conséquences, atteignant même des retards et des interruptions de projet en raison de la mauvaise qualité des données. Selon IBM (International Business Machines) [01], la plupart des données stockées dans les bases de données des organisations sont erronées [02]. En conséquence, les parties prenantes peuvent perdre leur confiance dans les systèmes d'aide a la décision et chercher d'autres solutions. Par conséquent, les entreprises étaient plus conscientes de l'importance de la qualité des données. La plupart d'entre eux investissent énormément d'argent pour résoudre les anomalies de leurs données stockées et les rendre utilisables pour en extraire une information.
Les problèmes de qualité des données peuvent apparaître de différentes manières. En effet, il existe des problèmes de complétude (valeurs manquantes), des valeurs en double, des problèmes d'intégrité référentielle, et bien d'autres anomalies.

Dans notre travail, nous nous concentrons sur le problème du couplage d'enregistrements également appelée **Record Linkage (RL)** qui est le processus qui vise à détecter tous les enregistrements qui font référence à la même entité, puis à les fusionner en un seul tuple.
Le moyen idéal d'appliquer le couplage d'enregistrements sur des données modifiées consiste à comparer chaque enregistrement de l'ensemble de données à tous les autres, ce qui correspond au produit cartésien des données. Malheureusement, cela pourrait aboutir à des milliards de comparaisons, ce qui n'est pas raisonnable dans le cas des données volumineuses. Afin de réduire le nombre important de comparaisons, la technique de blocage (Blocking) est utilisée. Habituellement, le blocage consiste à créer un ensemble de blocs. Chaque bloc

regroupera les enregistrements qui partagent une valeur commune nommée ”Blocking Key Value (BKV)” (valeur du champ d'attribut). Ainsi, seuls les enregistrements regroupés dans un même bloc sont à comparer entre eux.
Sur la base des défis RL Nous avons adapté l'algorithme K-Modes comme étape de blocage dont le but est d'améliorer le temps d'exécution et de contrôler le nombre de bloc et le nombre de données par bloc. L'utiliser de l'algorithme K-Modes a montré sa capacité à regrouper des données catégorielles ce qui n'était pas le cas d'autres algorithmes de clustering comme K-Means.
La sélection des attributs pertinents lors de génération des clés de blocage (Bloking Key) BK d'une maniéré semi automatique pour les différentes data set utilisées a prouvé son influence sur la qualité de données obtenu.

Ce document est organisé comme suit :

- Dans le premier chapitre nous avons étudié la qualité des données. Cette étude nous a permis d'avoir un aperçu des approches existantes pour gérer la qualité des données dans la littérature et dans quel domaine une contribution devrait être proposée.

- Le chapitre 2 présente les différentes étapes du couplage d'enregistrements et fournit une étude basée sur les principaux défis de la RL dans la qualité de données.

- Le chapitre 3 présente la description de la méthode proposée et les algorithmes utilises. Les expériences et les résultats de notre application sont présentés et discutées.

- Enfin, nous terminerons ce mémoire par une conclusion générale et quelques perspectives.

Chapitre 1

La qualité des données

1.1 introduction

Les problèmes de qualité des données stockées dans les bases de données et les entrepôts de données sont spécifiques à tous les types de données (structurées ou non structurées) et à tous les domaines d'application : données gouvernementales, commerciales, industrielles ou scientifiques. Il s'agit notamment des valeurs erronées, dupliquées, incohérentes, manquantes, incomplètes, incertaines, obsolètes, inhabituelles ou non fiables dans les données. L'impact des données de non-qualité (ou de mauvaise qualité) sur la prise de décision et son coût financier peut être considérable [03].
Avec la prolifération des sources d'information disponibles et le volume croissant de données potentiellement accessibles, la qualité des données est devenue un enjeu majeur, d'abord au sein des entreprises et, au cours de la dernière décennie, dans le milieu universitaire [04]. Ce chapitre présente un état de l'art sur la qualité des données, nous allons Définir les raisons pour lesquelles les données sont importantes dans le cadre des programmes, D'expliciter « qualité des données », De dresser la liste des sept dimensions de la qualité des données, des problèmes liés à une mauvaise qualité de données et les stratégies susceptibles d'éviter ce type de problème.

1.2 Big Data

Au cours des dernières années, les chercheurs et les entreprises se déférent d'opinion on se qui concerne la définition de terme ' Big Data ' .Certains se concentrent sur le volume des données et les autres se concentrent sur les techniques technologiques utilisées pour analyser et extraire des informations importantes.
Les premières définitions du Big Data ont été proposées par le groupe GARTNER en 2001.

Le groupe a proposé un modèle de définition à trois V : Volume, Vélocité et Variété. Cette définition a été mise à jour plus tard en 2012 par le groupe GARTNER et étendue par IBM en ajoutant un quatrième V qui signifie Veracity.

1.2.1 Volume

Le terme volume est utilisé pour désigner l'énorme quantité de données collectées à partir de différentes sources. L'intérêt de rassembler ces grands ensembles de données et de les analyser est d'extraire le plus de connaissances possible pour les chercheurs et les entreprises [26].

1.2.2 Variety

le terme variété sert à décrire les types des différentes données collectées.l'analyse du Big Data utilise plusieurs types de données. provenant de différentes sources .Il peut être structuré comme des bases de données relationnelles traditionnelles ou des données semi-structurées ou non structurées comme des fichiers texte.

1.2.3 Vélocité

C'est la vitesse à laquelle les données arrivent à une entreprise et le temps qu'il a fallu pour les analyser

1.2.4 Véracité

la véracité [27] est la Crédibilité des données et pertinence pour le public cible. Les entreprises collectent Le big data directement la supervisé par des spécialistes. Donc, il pourrait être plein d'erreurs. La crédibilité des données est alors un point trés important pour tirer les bonnes conclusions. il est également mentionné dans leur définition de la véracité que la confidentialité des données et les préoccupations juridiques doivent être prises en considération.

1.3 Intégration de données

L'intégration des données consiste à combiner des données provenant de diverses sources afin de créer une vue unifiée des jeux de données qui permettent un processus analytique facile [28]. L'intégration des données permet également un accès unifié aux données d'une entreprise, l'utilisateur n'a pas à accéder aux données de chaque département séparément mais le système d'intégration fournira une vue unifiée et épurée des données.

Parmis les défis à relever lors de l'intégration des données est la représentation de schéma différente dans chaque source de données. Les mêmes données peuvent être représentées différemment. Un autre défi important est la fusion des données. Lorsque le système détecte deux enregistrements ou plus faisant référence à la même entité du monde réel, il est important de choisir lequel conserver dans les résultats finaux ou comment fusionner tous les enregistrements en un seul[29].

1.4 Définition de la qualité des données

La qualité des données est la mesure de l'état des données en fonction de divers facteurs : exactitude, exhaustivité, cohérence, fiabilité et actualité. La mesure des niveaux de qualité des données peut aider les organisations à identifier les éventuelles erreurs qui doivent être corrigées et à évaluer si les données présentes dans leurs systèmes informatiques sont suffisantes pour leurs besoins. En d'autres termes, la qualité des données est l'évaluation de la fiabilité et de la pertinence des informations contenues dans un ensemble de données [05].
La qualité des données est une préoccupation importante pour les entreprises et le gouvernement car elle affecte directement la prise de décision et l'efficacité opérationnelle. Des données de mauvaise qualité peuvent entraîner des erreurs, des inefficacités et des coûts pour les organisations. La gestion de la qualité des données implique le développement de processus et de méthodes pour garantir que les données sont de la plus haute qualité, ainsi que la mise en place de mesures pour surveiller et maintenir la qualité des données au fil du temps[06].

1.5 L'importance de la qualité des données

De mauvaises données signifient de graves conséquences potentielles pour l'entreprise, du chaos opérationnel à une mauvaise stratégie commerciale en passant par des analyses floues. Les problèmes de qualité des données peuvent entraîner des coûts supplémentaires, entre autres pertes financières, telles que l'envoi de produits à la mauvaise adresse, la perte d'opportunités commerciales en raison d'un manque d'informations de contact précises ou complètes sur les prospects, ou l'imposition d'amendes aux entreprises pour avoir déposé des plaintes auprès d'informations financières ou réglementaires incorrectes, déclaration de conformité[7].
Nous pensons également que la méfiance des chefs d'entreprise et des managers vis-à-vis de la qualité des données est l'un des principaux obstacles à l'utilisation de l'intelligence décisionnelle (BI, business intelligence) et de l'analytique comme outil d'aide à la décision organisationnelle.

1.6 Comment définir des données de bonne qualité ?

Pour garantie la bonne qualité des données Ces facteurs, quand ils sont respectés, contribuent à produire des jeux de données fidèles et fiables.

1.6.1 L'exactitude

C'est un critère important pour des données de haute qualité. Pour éviter les problèmes de traitement des transactions dans les systèmes d'exploitation et les résultats erronés dans les applications analytiques, vous avez d'abord besoin de données précises. Toute inexactitude dans les données doit être identifiée, documentée et corrigée pour s'assurer que les gestionnaires, les analystes et les autres utilisateurs utilisent des informations de haute qualité[8] .

1.6.2 l'exhaustivité

An d'autre terme La complétude : C'est le degré de présence des valeurs dans une collection de données car les jeux de données doivent contenir tous les éléments nécessaires

1.6.3 la cohérence

C'est le degré de présentation des données dans un même format qui doit être compatibles avec les données précédente ce qui désigne l'absence de conflit entre des valeurs identiques dans des systèmes ou des jeux de données différents[9] .

- **La cohérence interne d'un ensemble de données** signifie que les éléments de données sont basés sur des concepts, des définitions et des classifications compatibles qui peuvent être efficacement combinés.
- **La cohérence entre les ensembles de données** signifie que les données sont basées sur des concepts, des définitions et des classifications communs, et que toute différence est expliquée et peut être justifiée.
- **La cohérence dans le temps** signifie que les données sont basées sur des concepts, des définitions et des méthodes stables dans le temps, et que toute différence est prise en compte et peut être justifiée.
- **La cohérence internationale** signifie que les données sont basées sur des concepts, des définitions, des classifications et des méthodologies communs, et que toute différence est expliquée et justifiée.

1.6.4 l'absence de doublons

L'entité est gérée par plusieurs systèmes d'informations sous des identifiants différents. Alors si les données sont répétées dans les enregistrements des bases de données donc la vue de ces derniers ne sera pas unifiée.

1.6.5 l'actualité des données

C'est-à-dire le fait qu'elles aient été mises à jour si besoin pour rester pertinentes. C'est la mesure dont laquelle l'information est considérée comme vraie et crédible.

1.6.6 la conformité

C'est la conformité aux normes de format des données fixées par l'organisation
Exemple
"Hôpital Ahmed Medaghri" peut apparaître comme "CHU Ahmed Medaghri", "C.H.U Ahmed Medaghri" ou "CHU AM".

1.7 Types (ou causes) des problèmes liés à la qualité des données

1.7.1 Création des données

- Saisie manuelle : Absence de contrôle systématique des formulaires de saisie.
- Saisie automatisée : problèmes de capture OCR (Optical Character Recognition), reconnaissance vocale incomplète, manque de normalisation ou modélisation conceptuelle inadéquate des données[10].
- Entrée de doublons.
- Approximations.
- Contraintes matérielles ou logicielles.
- Erreurs de mesure.
- Corruption des données : violations de la sécurité des données physiques et logiques.

1.7.2 Collecte / import des données

- Corrompre ou détruire des informations par un prétraitement inapproprié[10].
- Perte de données : problèmes de transmission.
- Le programme de conversion de données introduit une erreur..
- Modèles et structures de données inappropriés, spécifications incomplètes ou exigences changeantes dans l'analyse et la conception du système.

1.7.3 Stockage des données

- Absence de métadonnées.
- Absence de mise à jour.
- Modifications ad-hoc.
- Modèles et structures de données inappropriés, spécifications incomplètes ou évolution des besoins dans l'analyse et conception du système[10].
- Contraintes matérielles ou logicielles.

1.7.4 Intégration des données

- L'hétérogénéité des déférents niveaux de qualité et d'agrégation des données causent des problèmes d'intégration[10].
- Problèmes de synchronisation temporelle.
- Systèmes de données non conventionnels.
- Facteurs sociologiques conduisant à des problèmes d'interprétations et d'intégration des données.

1.8 Les dimensions de la qualité des données

La qualité est définie comme "l'adaptation aux besoins des utilisateurs". Cette définition est plus large que celles proposées par le passé, quand la qualité était synonyme de précision. Il est maintenant généralement admis que d'autres dimensions sont également importantes.
La qualité de donnée envisage la qualité selon sept dimensions :

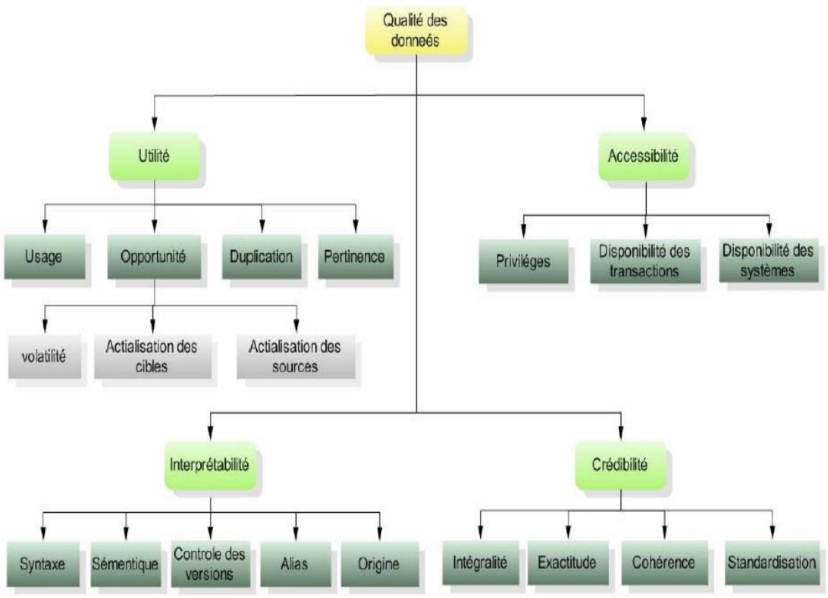

Figure 1.1: Les dimensions de la qualité des données.

1.8.1 L'exactitude

Des données exactes contiennent un minimum d'erreurs et de biais. L'exactitude est également connue sous le nom de validité. Par exemple, si des données sont saisies de manière incorrecte dans le système, des erreurs de transcription peuvent se produire, affectant l'exactitude. Il s'agit généralement d'erreurs accidentelles qui peuvent survenir lorsque des informations sont enregistrées de manière incorrecte ou saisies de manière incorrecte dans une base de données informatique. L'exactitude peut également être affectée par des données incomplètes, tardives ou inexactes. Il peut également être directement modifié par des actions entreprises pour d'autres raisons.

1.8.2 Fiabilité

Les données sont fiables lorsqu'elles sont mesurées et collectées systématiquement et dans le temps. La fiabilité des données dépend de la disponibilité de systèmes d'information avec des protocoles et des procédures cohérents. Pour être fiables, les données ont besoin d'instructions de collecte normalisées et écrites[11] . Le processus de collecte de données du programme ne doit pas varier en fonction de qui l'utilise, sur quel site il est utilisé, quand il est utilisé ou à quelle fréquence il est utilisé. En outre, les procédures de correction des erreurs de données ou de traitement des données manquantes ou incomplètes doivent être cohérentes d'un lieu et d'une période à l'autre.

1.8.3 Exhaustivité

L'exhaustivité signifie que le système d'information saisit tous les individus, services, sites ou autres unités éligibles qu'il a l'intention de mesurer. Les données générées doivent représenter la liste complète des personnes, services, sites et autres unités, et pas seulement un sous-ensemble de la liste. L'exhaustivité est affectée par[12] :

- La mesure dans laquelle les documents sources incluent toutes les informations pertinentes et nécessaires à la préparation du rapport.
- La mesure dans laquelle tous les sites rapportent des informations à un niveau supérieur pour l'agrégation.
- Délai de notification à l'agrégation de données de niveau supérieur. Par exemple, les données d'un site de programme seraient incomplètes si elles n'incluaient pas d'informations sur tous les clients desservis, tous les services fournis aux clients ou toutes les activités réalisées. Les données agrégées d'un programme ne seront pas exhaustives si les données de seulement 90 sites sur 100 sont fournies.

1.8.4 Précision

La précision signifie que les données sont suffisamment détaillées pour mesurer ce que la métrique définit. Par exemple, une métrique idéale pourrait nécessiter des nombres ventilés par sexe de personnes testées pour le VIH. Les systèmes d'information manquent de précision s'ils ne sont pas conçus pour enregistrer le sexe des personnes qui reçoivent des services de dépistage. Lorsque les données sont plus détaillées, elles sont plus précises, ce qui peut avoir un impact positif sur la qualité de la représentation adéquate des données des activités du programme[12] . La précision aide également à répondre aux questions importantes pour les chefs de projet, les directeurs régionaux et les unités nationales et internationales (le cas échéant). Cela suppose que les formulaires de collecte de données sont conçus pour collecter des données exactes et qu'un niveau de détail approprié est communiqué aux niveaux supérieurs.

1.8.5 La temporalité

Les données sont opportunes lorsqu'elles sont transférées au niveau supérieur suivant à temps pour respecter les délais de déclaration. "On time" signifie que les données communiquées auraient pu être utilisées dans un rapport de synthèse préparé au niveau de reporting supérieur. Par exemple, le site de service rend compte du 15 du mois précédent au rapport intermédiaire,

l'intermédiaire rend compte à l'unité de SE le 20 du mois en cours, et cette dernière est prête à faire rapport à la fin du mois[13] . Chacun de ces délais doit être respecté pour que les données soient à jour. La ponctualité est affectée par les facteurs suivants : 1. La fréquence de mise à jour du système d'information sur le projet 2. La rapidité avec laquelle les activités réelles du projet changent 3. Quand l'information est réellement utilisée ou nécessaire La fréquence de fourniture des données doit être suffisamment élevée pour permettre aux directeurs de programme, aux directeurs régionaux et aux directeurs nationaux et internationaux d'utiliser les informations pour prendre des décisions de gestion.

1.8.6 Intégrité

Les données sont intactes lorsque les systèmes d'information sont protégés contre les préjugés délibérés ou la manipulation à des fins politiques ou personnelles. Un examen indépendant des données peut aider à déterminer si l'intégrité des données peut avoir été compromise[13] . Le fait de savoir que les données subiront un contrôle indépendant est susceptible de dissuader une manipulation des données.

1.8.7 Confidentialité

La confidentialité signifie que les clients peuvent être assurés, que leurs données seront conservées conformément aux normes nationales et/ou internationales. Par conséquent, les informations personnelles ne seront pas divulguées de manière inappropriée et les données papier et électroniques seront protégées de manière adéquate. Un autre aspect important est la formation des employés au respect des informations confidentielles et à ne pas les partager avec d'autres clients. Ces mesures protègent la vie privée des clients que nous servons.

1.9 Qualité des données et nettoyage des données

Le processus de nettoyage des données vise à supprimer les données rejetées, obsolètes ou erronées. Des données propres sont un élément essentiel pour obtenir les bonnes informations, rapports et analyses. Dans toute l'organisation, les individus prennent des décisions commerciales en fonction des données qui leur sont présentées. Le nettoyage des données fournit des données de haute qualité qui aident à surmonter les problèmes de fraude et permettent aux organisations de se conformer aux réglementations. Des données de haute qualité sur des entités commerciales clés offrent un canal de croissance pour les entreprises prospères. En utilisant des techniques de nettoyage des données, les organisations peuvent rapidement faire correspondre et identifier les doublons dans les données.

1.10 Résoudre les problèmes de qualité dans les données

Comme le représente la Figure 1.2, on peut classer la plupart des travaux abordant la problématique de la qualité des données selon quatre grands types d'approches complémentaires[13] .

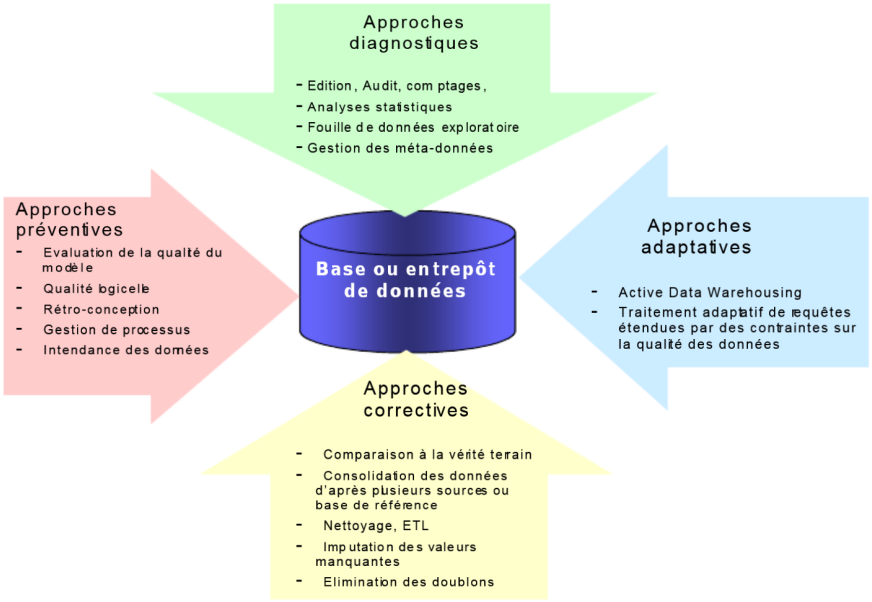

Figure 1.2: Panorama des approches pour l'évaluation et le contrôle de la qualité des données

1.10.1 Les approches préventives

Mettre l'accent sur l'ingénierie des systèmes d'information et le contrôle des processus, en utilisant des techniques d'évaluation de la qualité des modèles conceptuels, du développement de logiciels et des processus utilisés pour le traitement des données.

1.10.2 Les approches diagnostiques

Se concentrer sur les méthodes statistiques, l'analyse et le fouille de données exploratoires pour détecter les anomalies de données.

1.10.3 Les approches correctives

Basées sur des techniques de nettoyage et de consolidation de données et utilisent un langage de manipulation des données étendus et des outils d'extraction et de transformation de données (ETL – Extraction-Transformation-Loading)

1.10.4 Les approches adaptatives (actives)

Typiquement appliqués lors de la médiation ou de l'intégration des données : ils portent sur l'adaptation des traitements (interrogation des données ou opérations de nettoyage), y compris la validation des contraintes de qualité des données dans l'exécution en temps réel.

1.11 techniques de détection/correction des problèmes de qualité des données

Parmi les techniques de détection/correction des problèmes de qualité des données, On présentra dans ce qui suit les techniques les plus couramment utilisées en pratique, dont les coûts respectifs sont estimés dans la Figure 1.3 :

- Soit sur la base de la vérité terrain soit sur la base sur un Sources de données de référence.
- Audit des données.
- Suivi des données
- Nettoyage des données

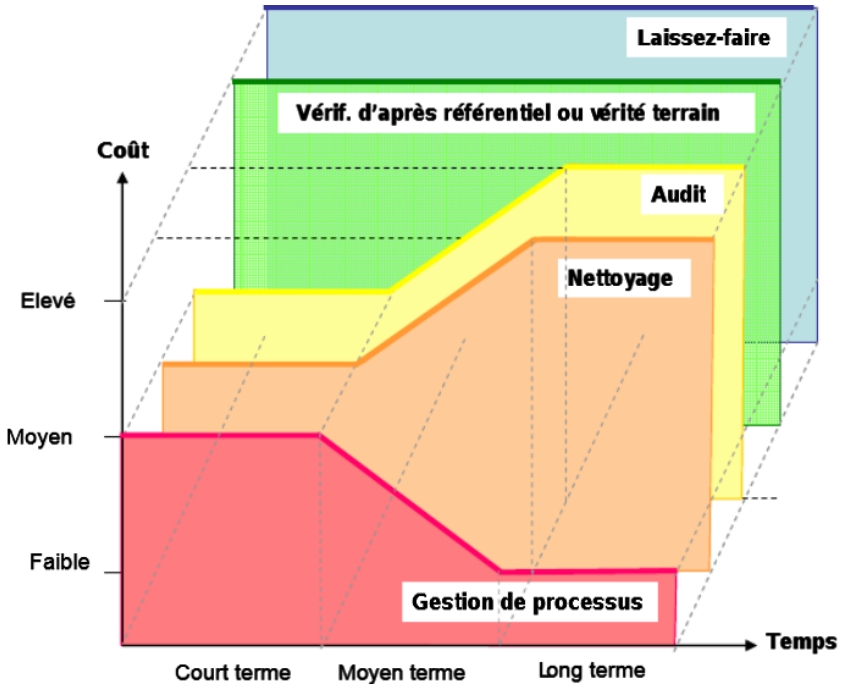

Figure 1.3: Coût approché des approches incluant le coût induit par la non-correction des erreurs

1.11.1 la vérification d'après la vérité-terrain ou d'après une source de données de référence

Cette technique consiste à comparer les valeurs des données à leurs homologues du monde réel (vérification de la vérité). La deuxième méthode, appelée fusion, consiste à comparer deux ou plusieurs bases de données. Les données pertinentes dans la base de données en cours de vérification sont comparées aux données correspondantes dans une autre base de données : les données identiques sont considérées comme correctes, les données incorrectes sont signalées pour enquête et correction éventuelle.

1.11.2 Audit des données

Data Audit met en place des programmes pour vérifier que les valeurs des données satisfont à plusieurs types de contraintes. Ces contraintes se produisent à différents niveaux de la base de données (valeurs, attributs, tuples, relations ou ensembles à chaque niveau). L'audit des données a pour avantage d'être simple à mettre en œuvre. Il peut être conçu en même temps que le modèle conceptuel de données et différents outils d'analyse de données de diagnostic peuvent être utilisés. Cependant, cela ne permet pas une amélioration continue de la qualité

des données. La compilation des données vise l'exhaustivité, c'est-à-dire le respect des règles préalablement définies, mais elle ne garantit en aucun cas l'exactitude des données.

1.11.3 Suivi des données

Le suivi des données est l'échantillage des enregistrements au moment ou ils entrent dans la première procédure de traitement et de les suivre à travers chaque sous-processus jusqu'à ce qu'ils entrent dans la base de données. Les modifications apportées à l'enregistrement au fur et à mesure de son traitement permettent d'élaborer des critères de correction en exploitant la redondance des données.

1.11.4 Nettoyage des données

Le nettoyage des données consiste en un ensemble de transformations conçues pour normaliser le format des données et détecter les paires d'enregistrements qui sont très probablement liés au même objet. Si des données approximativement redondantes sont trouvées et que la correspondance multi-tables calcule des jointures approximatives entre des données dissemblables mais similaires pour permettre leur fusion, une étape de déduplication est appliquée.

1.12 Comment déterminer la qualité des données

En règle générale, pour déterminer le niveau de qualité de ses données, les organisations procèdent d'abord à un inventaire, au cours duquel l'exactitude relative, l'unicité et la validité des données sont mesurées pour établir une ligne de base. Les valeurs de référence ainsi déterminées peuvent être comparées en permanence aux données du système pour identifier et résoudre tout problème de qualité.
Une autre étape courante consiste à créer des règles de qualité des données basées sur les exigences de l'entreprise pour les données opérationnelles et analytiques. Ce type de règle établit le niveau de qualité requis des ensembles de données et détaille les différents éléments qu'ils doivent avoir pour permettre la vérification des attributs de qualité des données tels que l'exactitude et la cohérence[14] .
Une fois les règles en place, l'équipe de gestion des données procède généralement à une évaluation pour mesurer la qualité de l'ensemble de données et consigner les erreurs et autres problèmes. Ce processus peut être répété périodiquement pour maintenir le niveau de qualité le plus élevé possible.
Ces évaluations peuvent être réalisées à l'aide de différentes méthodologies, telles que le Data

Quality Assessment Framework (DQAF) créé par Optum, filiale de santé de UnitedHealth Group, pour formaliser son approche d'évaluation de la qualité des données. Complétude, Actualité, Validité, Cohérence et Complétude, le DQAF propose de mesurer les dimensions de la qualité des données. Optum a mis les détails de ses termes de référence à la disposition de toute organisation nécessitant un modèle.
Le Fonds monétaire international, qui supervise le système monétaire mondial et accorde des prêts aux pays en difficulté économique, a également défini sa propre méthodologie d'évaluation, connue sous le nom de Data Quality Assessment Framework. Le cadre met l'accent sur plusieurs caractéristiques de la qualité des données, notamment l'exactitude, la fiabilité et la cohérence que le FMI exige des pays membres lorsqu'il lui soumet des statistiques[14] .

1.13 l'avantage de disposer de données de qualité

D'un point de vue financier, le maintien d'un haut niveau de qualité des données permet aux organisations de dépenser moins d'argent pour identifier et corriger les données erronées dans leurs systèmes. Les entreprises peuvent également éviter les erreurs opérationnelles et les pannes dans leurs processus commerciaux, ce qui peut augmenter les dépenses d'exploitation et réduire les revenus.
Une bonne qualité des données améliore également la précision des applications analytiques, ce qui peut conduire à de meilleures décisions commerciales qui stimulent les ventes, améliorent les procédures internes et offrent aux organisations un avantage concurrentiel.
Des données de qualité facilitent également l'utilisation de tableaux de bord de veille stratégique et d'outils d'analyse. Lorsque les données analytiques sont fiables, en effet, les utilisateurs sont plus enclins à les utiliser qu'à prendre des décisions basées uniquement sur leur intuition ou leurs propres calculs.

1.14 Conclusion

Ce chapitre a été consacré à dresser un état de l'art de la qualité des données. Nous avons donné une définition de la qualité de donnée et les déférentes raisons pour laquelle elle est importante, ensuite nous avons cité les dimensions et les problèmes liés à une mauvaise qualité de données et comment les corriger. Dans le chapitre suivant nous allons introduire le Couplage d'enregistrement qui est L'un des principaux processus dans le domaine de la qualité des données.

Chapitre 2

Couplage d'enregistrements

2.1 Introduction

L'un des principaux processus importants dans le domaine de la qualité des données est le processus de couplage d'enregistrements (RL). L'objectif du couplage d'enregistrements est d'identifier les tuples qui font référence à la même entité du monde réel et de les fusionner en un seul. RL aide à améliorer considérablement la qualité des données en supprimant tous les enregistrements faisant référence à la même entité de monde réel Dans ce chapitre nous allons présenter un état de l'art sur le couplage d'enregistrement et l'algorithme K-Modes.

2.2 Définition

Le couplage d'enregistrements est le processus par lequel des enregistrements ou des unités provenant de différentes sources de données sont réunis dans un seul fichier à l'aide d'identifiants non uniques, tels que des noms, des dates de naissance, des adresses et d'autres caractéristiques. L'idée initiale du couplage d'enregistrements remonte aux années 1950, puis cette technique a été appliquée par des personnes issues d'un large éventail de domaines, tels que l'entreposage de données et l'intelligence de gestion, la recherche historique, ainsi que la pratique et la recherche médicales[15] .

2.3 Types de couplage

Il existe deux types de couplage d'enregistrements : l'appariement exact et l'appariement statistique. L'appariement exact se divise en deux sous-types : le couplage d'enregistrements déterministe et le couplage d'enregistrements probabiliste, tel qu'illustré à la figure ci-dessous[6].

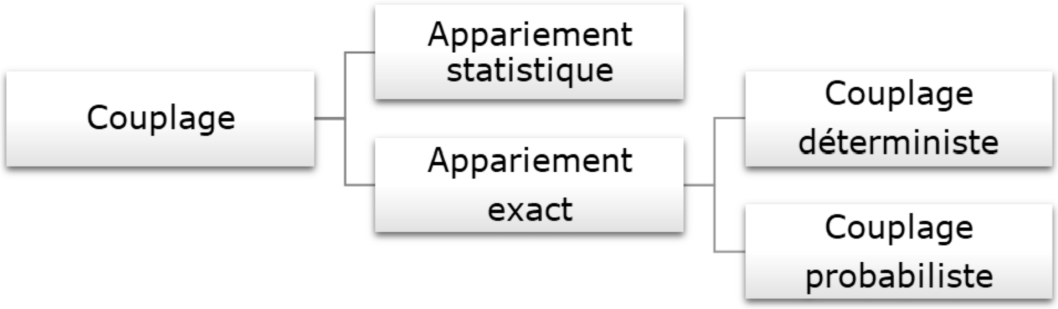

Figure 2.1: Type de couplage d'enregistrements

2.3.1 Appariement statistique

Le but de l'appariement statistique est de créer un fichier qui reflète la répartition sous-jacente de la population. Les enregistrements combinés ne correspondent pas nécessairement à la même entité, telle qu'une personne ou une entreprise. Les documents correspondants peuvent avoir des unités différentes mais faire référence à la même population. Supposons que la relation des variables dans la population est similaire à la relation dans le fichier. Cette méthode est principalement utilisée dans les études de marché et est rarement utilisée par les agences statistiques officielles.

2.3.2 Appariement exact

L'objectif d'une correspondance exacte est de corréler les informations d'un enregistrement particulier dans un fichier avec les informations d'un fichier secondaire pour créer un fichier contenant les informations correctes pour chaque enregistrement. Le couplage se fait au niveau de l'enregistrement, par exemple entre les enregistrements de mortalité et les recensements.

2.3.2.1 Couplage d'enregistrements déterministe

Il s'agit de la forme la plus simple de couplage d'enregistrements, qui produit des liens basés sur des identifiants ou des variables communes parmi les sources de données disponibles. Il arrive souvent qu'il n'existe pas de variable unique exempte d'erreurs, présente sur la majorité des données et ayant un pouvoir de discrimination suffisant. Seule une combinaison de variables sera capable de discriminer entre deux enregistrements. C'est une technique souvent utilisée par les agences de statistiques officielles. Statistique Canada utilise cette méthode pour construire ses registres d'entreprises, d'adresses et de population, ce qui implique de multiples opérations d'enquête par la suite[16] .

2.3.2.2 Couplage d'enregistrements probabiliste

Il s'agit d'un autre type d'appariement exact. Comme dans l'autre cas, il n'y a pas d'identifiant unique disponible pour l'appariement. Contrairement à l'appariement déterministe, l'appariement probabiliste peut compenser si les informations sont incomplètes ou sujettes à erreur. Les enregistrements qui ne concordent pas totalement pour chaque variable peuvent être reliés entre eux pour constituer un ensemble de paires potentielles. Un score est alors calculé pour chaque paire potentielle. Ensuite, un statut de couplage est attribué à chaque paire potentielle sur la base du score[16] . Elles peuvent être non supervisées ou supervisées.

A- Méthodes non supervisées

Ces méthodes sont intéressantes quand il n'existe pas de données annotées. Un modèle théorique de couplage d'enregistrements qui offre des méthodes statistiques pour estimer les paramètres de couplage et les taux d'erreurs a été proposé par Fellegi et Sunter dans [30]. Ce modèle est décrit ci-après.

Modèle de Fellegi et Sunter

Définition: Le modèle définit les quantités suivantes pour fonctionner :

- M-probabilité : c'est la probabilité de couplage (matching) estimée pour qu'une paire d'attributs quelconque d'un champ donné soit similaire sachant que la paire d'enregistrements correspondante est un vrai couplage (i.e. les deux enregistrements correspondent en réalité);
- U-probabilité : c'est la probabilité de non couplage (unmatching) estimée pour qu'une paire d'attributs quelconque d'un champ donné soit similaire sachant que la paire d'enregistrements correspondante est un faux couplage, i.e. la probabilité que les deux enregistrements se couplent par hasard;
- ratio de couplage d'attributs : le ratio de couplage d'une paire d'attributs correspondants à un champ ayant une probabilité de couplage M et une probabilité de non couplage U est calculé par :
  - pour un accord sur l'attribut : $\log(\frac{M}{U})$;
  - pour un désaccord sur l'attribut : $\log(\frac{1-M}{1-U})$;
- ratio de couplage d'enregistrements R : le ratio de couplage d'une paire d'enregistrements est calculé par la somme des ratios de couplage des différentes paires d'attributs qui le composent;

– le couplage d'enregistrements se fait par rapport aux seuils Tλ et Tμ comme suit :

- si R $\prec$ Tλ alors non couplage (rejet);

- si T$\mu \prec$ R $\prec$ Tλ alors possible couplage (indécis);

- si R $\succ$ Tμ alors couplage (acceptation).

Problèmes posés. Les problèmes posés par ce modèle se traduisent par 3 questions :

– Comment définir M et U ?

– Comment fixer les seuils Tλ et Tμ ?

– Comment déterminer les mesures de similarité entre champs ?

Interprétation. Pour l'évaluation, nous pouvons définir les vrais et les faux positifs (resp. négatifs) à l'aide de ces considérations de couplage (resp. non couplage) et de liaison (resp. non liaison) (voir Tableau 2.1). En effet, si nous avons décidé de coupler (resp. ne pas coupler) une paire d'enregistrements alors si cette paire est liée (resp. n'est pas liée) en réalité, alors il s'agit d'un vrai positif (resp. vrai négatif) sinon il s'agit d'un faux positif (resp. faux négatif).

Table 2.1: Classification des résultats de couplage

	Couplage	Non couplage
Liaison	vrai positif (TP)	faux positif (FP)
Non liaison	faux négatif (FN)	vrai négatif (TN)

Plusieurs méthodes se sont basées sur le modèle de Fellegi et Sunter pour le couplage d'enregistrements. A titre d'exemple, Winkler a détaillé dans [31] une technique qui se base sur l'algorithme "espérance-maximisation" pour estimer les paramètres du modèle et optimiser les règles de couplage.

Dans le cas des bases de données volumineuses (par exemple, des dizaines de milliers d'enregistrements), le couplage d'enregistrements consiste à comparer tous les enregistrements deux à deux en calculant le produit cartésien, ce qui n'est pas efficace en temps et en mémoire. Pour résoudre ce problème, les auteurs dans [32] proposent une amélioration du modèle de Fellegi et Sunter consistant à séparer la base de données en des groupements d'enregistrements (blocs ou fenêtres), en se basant sur de simples heuristiques qui aident à éliminer les paires d'enregistrements clairement non liés.

B- Méthodes supervisées

Dans ce type de méthodes, les auteurs traitent le problème de couplage d'enregistrements se référant à la même entité comme un problème de classification supervisée. En effet, ils proposent de représenter chaque paire d'enregistrements à l'aide d'un vecteur de caractéristiques décrivant les similarités entre les paires d'attributs. Ces caractéristiques peuvent être binaires (par exemple, les attributs "nom" correspondent), discrètes (par exemple, les n-premiers caractères du "prénom" correspondent) ou continues (par exemple, la mesure de Levenshtein entre les prénoms). Ainsi, la comparaison entre paires d'enregistrements conduit à les ranger dans les 3 classes : couplage, non couplage ou couplage possible. Les classifieurs utilisés déterminent, à partir des données annotées, les conditions de couplage entre les enregistrements, à savoir la détermination des champs pertinents et les seuils de décision pour la similarité entre les attributs.

Nous pouvons citer à titre d'exemple les travaux dans [33] qui proposent une méthode basée sur deux niveaux d'apprentissage supervisé employant un SVM. Le premier niveau représente la comparaison de paires d'attributs en utilisant un apprentissage avec la mesure de Levenshtein (similarité au niveau attribut) et le deuxième niveau représente la comparaison des paires d'enregistrements en utilisant un apprentissage sur la similarité (similarité au niveau enregistrement).

Les auteurs dans [34] utilisent un apprentissage actif pour sélectionner les paires d'enregistrements les plus informatives. Ainsi, l'utilisateur est sollicité pour annoter ces paires d'enregistrements informatives comme paires liées ou non liées afin d'entraîner les classifieurs. Ces derniers génèrent des règles de classification qui croisent des attributs et des mesures de similarité entre ces attributs.

Les auteurs dans [35] proposent de normaliser certains attributs (ceux qui désignent des noms et des adresses postales) pour effectuer convenablement la comparaison de ces attributs dans le processus de couplage d'enregistrements. Cette normalisation est effectuée en se basant sur un résultat de segmentation des attributs (par exemple un attribut adresse est segmenté en : numéro de voie, type de voie, nom de voie). Pour la segmentation, un modèle de Markov caché est utilisé avec comme observations, les mots de l'attribut, et comme états cachés, les segments. Cette méthode nécessite suffisamment de données annotées sous forme de chaînes de caractères segmentées pour l'apprentissage.

2.4 Les étapes du processus de couplage d'enregistrements

Le processus Record Linkage (RL) peut être défini par un processus en trois étapes [17]tel qu'il est illustré dans la figure 2.2

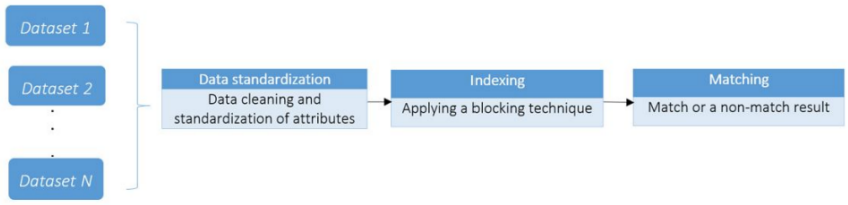

Figure 2.2: Principales étapes du processus de couplage d'enregistrements

2.4.1 Nettoyage et normalisation

La première étape est la standardisation des données. En fait, il a été prouvé dans des recherches antérieures que laisser des ensembles de données sans normalisation d'attributs, et la détection d'anomalies de schéma peut conduire à une mauvaise conclusion et même fusionner les mauvais tuples[18]. Par exemple, l'attribut qui représente le nom d'une personne peut apparaître dans un ensemble de données sous la forme Nom complet et dans un autre sous la forme de deux attributs (Prénom et Nom). Ainsi, la normalisation des données est une étape importante dans le processus RL.

2.4.2 L'indexation

La deuxième étape du processus RL est l'indexation, le but de cette étape est d'identifier les tuples qui seront comparés les uns aux autres lors de l'étape de mise en correspondance. La meilleure façon de le faire en termes de précision est l'approche naïve en comparant chaque enregistrement à tous les autres. Mais, bien sûr, cela pourrait aboutir à un nombre inacceptable de comparaisons. Par exemple, la comparaison de deux bases de données avec 2 millions d'enregistrements chacune peut 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".

2.4.2.1 BLOCAGE

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 Blocking Key-Value (BKV).

Une clé de blocage peut être choisie comme 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 aussi être choisie avec la concaténation de plusieurs attributs comme les quatre premiers caractères du Prénom et le Code postal de l'attribut adresse. La décision de sélectionner quel attribut ou groupe de attributs seront utilisés comme clé de blocage est très important car les blocs qui seront créés dépendent de cette décision. Ainsi, choisir l'attribut le moins error-prone comme BK est très important et nous avons parfois besoin de l'intervention d'un expert pour cette décision. Une fois l'étape d'indexation effectuée, seuls les enregistrements d'un même bloc sont comparés entre eux.

A- Blocage standard

La première technique de blocage proposée est le blocage standard [19] . L'idée derrière cela est de regrouper tous les tuples qui partagent une clé-valeur de blocage commun dans le même bloc. Une clé de blocage peut être sélectionnée comme un attribut unique ou une concaténation de plusieurs. Ce faisant, seuls les tuples qui se trouvent dans le même bloc peuvent être comparés les uns aux autres. L'inconvénient majeur du blocage standard est que la génération d'un BKV basé sur un éventuel attribut erroné (comme les noms et les adresses) peut conduire à une mauvaise correspondance. Pour éviter ce problème, les attributs les moins sujets aux erreurs doivent être sélectionnés [20] ou en utilisant plusieurs clés de blocage.

Les tableaux 2.2 et 2.3 montrent une démonstration du blocage standard.

Table 2.2: Exemple de blocage standard avec l'adresse comme clé de blocage 1

ID	Prénom	Nom	Ville	Age
R1	Abde Elkrim	Souilem	Saida	27
R2	Abdelkrim	Souilem Tahi	Saida	1993
R3	Bendjelloul	Baali	Sidi Bel Abbes	1993
R4	Ben djelloul	Bali	SBA	1976
R5	Benjelloul	Baaly	Sidi Bel abbes	44

Table 2.3: Exemple de blocage standard avec l'adresse comme clé de blocage 2

Blocs (Clé de Blocage)	Enregistrements
Bloc 1 (Saida)	R1,R2
Bloc 2 (Sidi Bel Abbes)	R3,R5
Bloc 3 (SBA)	R4

B- Indexation Q-gram

Une autre technique de blocage puissante est l'indexation Q-gram [21] . Dans cette approche, la valeur-clé de blocage est divisée en un ensemble de sous-chaînes d'une taille Q, puis un certain nombre de ces sous-chaînes sont concaténées pour former une nouvelle valeur de clé de blocage. Toutes les nouvelles chaînes obtenues seront utilisées comme clés de blocage. De cette façon, le même enregistrement peut être affecté à différents blocs.

C- Indexation basée sur un tableau de suffixes

L'indexation par tableau de suffixes [22] est une autre technique d'indexation qui existe dans la littérature. Il consiste à générer un ensemble de suffixes d'une longueur minimale choisie par l'utilisateur à partir de chaque clé de blocage. Une fois cela fait, tous les suffixes générés seront utilisés comme de nouvelles clés de blocage, ce qui permet l'insertion du même enregistrement dans un bloc différent de la même manière que l'indexation Q-Gram. Cette approche a été étendue dans [23] . Les auteurs ont proposé l'idée de fusionner deux blocs si la similarité entre leurs Id-Suffixes atteint un seuil fixe.

D- Quartier trié (Sorted neighborhood)

Dans [24] , les auteurs ont proposé une autre approche d'indexation appelée l'approche des quartiers triés. La première étape de cette approche consiste à générer les clés de blocage pour tous les attributs puis à les trier par ordre alphabétique selon leur BKV. Une fois cela fait, une fenêtre glissante de taille W est déplacée sur les enregistrements. A chaque étape, les enregistrements qui se trouvent dans la plage d'habillage de fenêtre sont comparés les uns aux autres. L'utilisation d'une fenêtre glissante peut réduire le nombre de comparaisons pour chaque enregistrement à (2W - 1) [25] .

Le tableau 2.4 montre une illustration de l'approche par voisinage trié. Les mêmes enregistrements de tableau 2.1 sont utilisés dans cet exemple. La taille de la fenêtre dans cet exemple est w=3.

Table 2.4: Exemple 1 de Sorted neighborhood avec une taille de fenêtre de w =3.

Position de la fenêtre	Les valeurs de Clé de Blocage (Ville)	ID
1	Saida	R1
2	Saida	R2
3	SBA	R4
4	Sidi Bel Abbes	R3
5	Sidi Bel Abbes	R5

Table 2.5: Exemple 2 de Sorted neighborhood avec une taille de fenêtre de w =3.

Windows range	paires
1-3	(R1,R2),(R1,R4),(R2,R4)
2-4	(R2,R4),(R2,R3),(R4,R3)
3-5	(R4,R3),(R4,R5),(R3,R5)

La liste finale des paires candidates est : (R1,R2), (R1,R4), (R2, R4), (R2, R3), (R3, R4), (R4, R5), (R3, R5).

2.4.3 Matching

La troisième et dernière étape du processus de couplage d'enregistrement consiste à faire correspondre les enregistrements indexés qui se trouvent dans le même bloc et à décider s'ils représentent la même entité du monde réel ou pas.La valeur de correspondance est normalisée entre 0 et A où 1 représente une correspondance exacte et 0 une noncorrespondance totale. La correspondance peut être effectuée à l'aide d'un ensemble de fonctions de similarité de chaînes qui existent dans la littérature [Levenshtein 1966] ou en utilisant un algorithme d'apprentissage automatique pour classer l'enregistrement comme correspondant ou non correspondant .

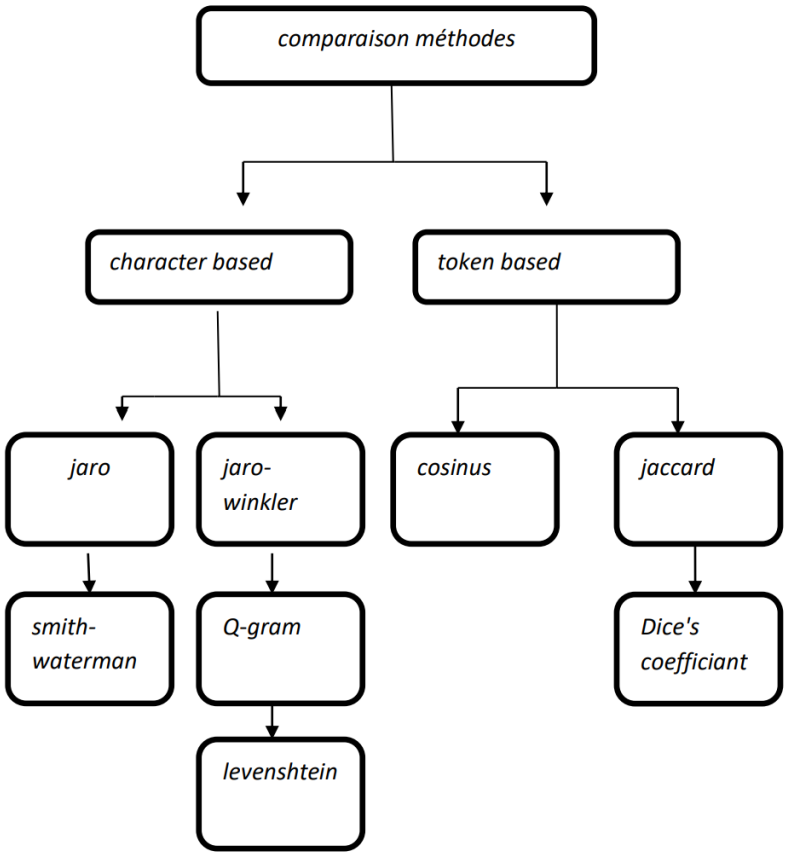

Figure 2.3: Types de méthodes de Matching

2.4.3.1 Encodage phonétique

La première famille de techniques de Matching est la l'encodage phonétique Une variété d'algorithmes de Matching existe dans la littérature.

A- Soundex

Soundex est considéré comme l'une des fonctions d'encodage phonétique les plus efficaces. Il transforme les chaînes selon 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 respectant les règles suivantes : (0 pour les 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 [26] . 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 en Y, DT-RT RD-NT-ND à D).

2.4.3.2 Pattern finding

La deuxième famille de techniques de Matching est la recherche de modèles. Une variété d'algorithmes de Matching existe dans la littérature.

A- Éditer la distance

La distance d'édition est également connue sous le nom de distance de Levenshtein. Il a été proposé en 1965 par Vladimir Levenshtein. Il est considéré comme l'une des métriques les plus utilisées pour mesurer la similarité entre deux chaînes. Généralement, il est défini comme le nombre d'insertions, de suppressions et de mises à jour afin de transformer une chaîne en une autre. Pour mieux comprendre, l'exemple suivant montre une démonstration de la façon de calculer les coûts de passage d'un mot à l'autre.

Word 1	Word 2	Operation	Cost
I		Delete (I)	1
N	E	Substitute (E)	1
T	X	Substitute (X)	1
E	E	Comparison	0
	C	Insert (C)	1
N	U	Substitute (U)	1
T	T	Comparison	0
I	I	Comparison	0
O	O	Comparison	0
N	N	Comparison	0
		Sum	5

Figure 2.4: Exemple Edit distance

Dans l'exemple présenté dans la figure 2.4, on voit que la distance d'édition entre les deux mots (Intention, Exécution) est la somme des coûts pour transformer la première chaîne en la seconde qui est égale à Cinq. Les étapes décrites dans l'exemple ne sont pas la seule solution pour transformer la première chaîne en la seconde mais sont celles qui coûtent le moins cher.

B- Jaro-Winkler

Le Jaro-Winkler est une métrique de similarité de chaîne qui a été proposée par William E. Winkler en 1990 comme une extension de la distance Jaro. Afin de mesurer la similarité Jaro-Winkler entre deux chaînes, la première étape consiste à mesurer la similarité Jaro traditionnelle qui est définie comme suit :

$$Jaro - Sim(s_1, s_2) = \begin{cases} 0, & m = 0 \\ \frac{1}{3}(\frac{m}{s1} + \frac{m}{s2} + \frac{m-t}{m}) & sinon \end{cases} \tag{2.1}$$

Où:

– s représente la longueur de la chaîne.

– m représente le nombre de caractères communs entre les séquences comparées 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 commençant par le même préfixe L pour une longueur maximale de Quatre. La similarité Jaro-Winkler est définie comme suit :

$$JaroWinlker - Sim(s_1, s_2) = Jaro - Sim(s_1, s_2) + LP(1 - Jaro - Sim(s_1, s_2)) \quad (2.2)$$

Où:

– JaroSim (s1,s2) est la similarité Jaro entre les chaînes.

– L est la longueur du préfixe.

– P est un facteur d'échelle (une constante qui prend généralement la valeur 0,1).

C- Distance de Jaccard

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

$$Jaccard(A, B) = \frac{A \cap B}{A \cup B} \quad (2.3)$$

Une fois cela fait, la distance de Jaccard n'est obtenue que par la soustraction du coefficient de Jaccard à 1.

$$Jaccarddistance(A, B) = 1Jaccard(A, B) \quad (2.4)$$

2.5 Conclusion

Les nombreuses applications du couplage d'enregistrements dans divers domaines ont rendu la littérature RL très riche. Dans ce chapitre, nous avons essayé de couvrir la plupart des travaux importants qui existent pour résoudre les problèmes d'une mauvaise qualité des données. Nous avons donné une vue globale dur le couplage d'enregistrement et on a fermé le chapitre par l'algorithme K-Modes pour traiter les données de type non numériques dans le Big Data. Le chapitre suivant sera dédié à la conception et l'implémentation de notre application.

Chapitre 3

Implémentation et Expérimentation

3.1 Introduction

Dans ce chapitre, nous décrivons l'environement et les technologies utilisées pour l'implémentation de notre application , puis nous présentons notre contribution dans le domaine du couplage d'enregistrements. Une solution est proposée pour chacun des défis abordés dans le chapitre précédent. Nous présentons une nouvelle technique de blocage basée sur l'algorithme K-Modes comme étape d'indexation. Ensuite la solution proposée sera évaluée afin de montrer sa performance et son efficacité et les résultats obtenus seront discutés. Enfin et nous présentant quelques captures d'écran.

3.2 L'approche proposé

3.2.1 Introduction

Dans ce qui suit nous donneront une explication détaillées de chaque étape de procussus de couplage d'enregistrement proposé on commençant par le chargement de l'ensemble de données jusqu'a l'étape d'évaluation.

La figure 3.5 montre les étapes de la solution proposé avec l'indexation basé sur l'algorithme K-Modes.

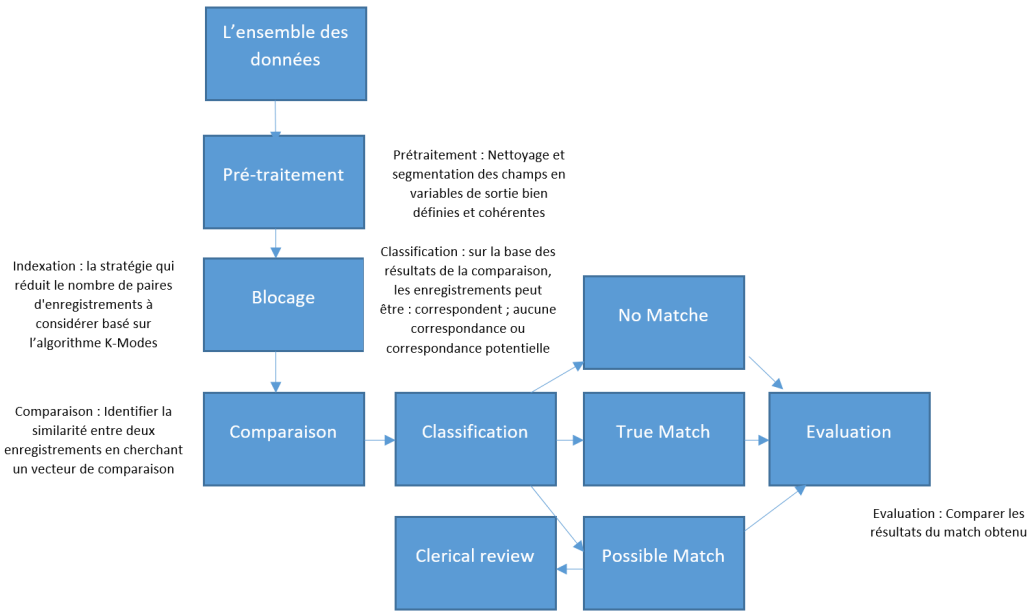

Figure 3.1: Shéma globale de l'approche proposé

3.2.2 Description de la méthode utilisée

Notre système est composé de 4 étapes :

- Exploration et traitement des données ;
- Génération des clés de blocage ;
- Indexation basé sur l'algorithme K-Modes ;
- Matching.

3.2.2.1 Exploration et traitement des données

Afin d'évaluer notre solution proposée. Un ensemble de données du monde réel a été utilisé. Cet ensemble de données a été sélectionné car il a été utilisé pour évaluer la plupart des approches de RL existantes dans la littérature. Il s'agit d'un ensemble de données de couplage d'enregistrements avec des informations sur différents restaurants.

3.2.2.1.1 Format Une trame de données avec 5 attributs : Nom, Adresse, Ville, Téléphone et Type. Cet ensemble de données comprend 533 restaurants de la base de données Fodors et 331 enregistrements de la base de données Zagat. Il est approprié pour effectuer divers types de couplage d'enregistrements et peut être évalué par des méthodes de couplage

d'enregistrements standard.
La figure 3.6 montre le format de l'ensemble de données utilisée dans notre application :

name	addr	city	phone	type
arnie morton's of chicago	435 s. la cienega blv.	los angeles	3102461501	american
arnie morton's of chicago	435 s. la cienega blvd.	los angeles	3102461501	steakhouses
art's delicatessen	12224 ventura blvd.	studio city	8187621221	american
art's deli	12224 ventura blvd.	studio city	8187621221	delis
hotel bel-air	701 stone canyon rd.	bel air	3104721211	californian
bel-air hotel	701 stone canyon rd.	bel air	3104721211	californian
cafe bizou	14016 ventura blvd.	sherman oaks	8187883536	french
cafe bizou	14016 ventura blvd.	sherman oaks	8187883536	french bistro
campanile	624 s. la brea ave.	los angeles	2139381447	american
campanile	624 s. la brea ave.	los angeles	2139381447	californian
chinois on main	2709 main st.	santa monica	3103929025	french
chinois on main	2709 main st.	santa monica	3103929025	pacific new wave
citrus	6703 melrose ave.	los angeles	2138570034	californian
citrus	6703 melrose ave.	los angeles	2138570034	californian

Figure 3.2: L'ensemble de données Restaurant.

3.2.2.1.2 Notation Nous présentons la notation utilisée dans le reste de ce chapitre. Tout d'abord, l'ensemble de données est noté D. Dans notre travail, nous supposons que Les données ont été importées un fichier.arf.
Un jeu de données Data est défini comme suit :
Data = {B1, B2, . . . , Bi}. Chaque Bi $\in$ D est composé d'un ensemble de tuples et défini comme: B = {t1, t2, . . . , tj} . Chaque ti $\in$ B est composé d'un ensemble d'attributs A = (Att1, Att2, . . . , Attk).

3.2.2.2 Génération des clés de blocages

La première étape de notre processus d'indexation basé sur K-Modes proposé est la génération des clés de blocage. Puisque nous considérons que l'ensemble de données existe déjà dans un système de fichier.arf. Ensuite, pour chaque tuple tj dans un bloc Bi , une ou plusieurs clés de blocage sont générées. L'idée derrière la génération de clés de blocage est de les utiliser comme attributs de clustering au lieu d'utiliser tous les attributs de l'ensemble de données, cela peut réduire considérablement le temps d'exécution et maintenir de bons résultats de clustering puisque les clés de blocage contiennent les informations les plus importantes sur les enregistrements.
La figure 3.7 montre un passage du célèbre jeu de données Restaurant qui est utilisé pour tester la plupart des techniques de blocage. Dans cet exemple, la clé de blocage est formée

par la concaténation de Nom du restaurant et du numéro de téléphone.

BK 2	name	addr	city	phone	type
le select2128751993	le select	507 columbus ave. between 84th and 85th sts.	new york	2128751993	american
le central4153912233	le central	453 bush st.	san francisco	4153912233	french
dining room3102755200	dining room	9500 wilshire blvd.	los angeles	3102755200	californian
sanppo4153463486	sanppo	1702 post st.	san francisco	4153463486	asian
evergreen cafe2127443266	evergreen cafe	1288 1st ave. at 69th st.	new york	2127443266	asian

Figure 3.3: Exemple de génération de clé de blocage.

3.2.2.3 Indexation basé sur l'algorithme K-Modes

Nous avons basée sur l'algorithme K-Modes dans l'étape d'indexation . Le clustering K-Modes divise l'ensemble des données en un ensemble de clusters sans chevauchement ; chacun contient des enregistrements qui peuvent faire référence à la même entité du monde réel. La plupart des enregistrements qui se trouvent dans le même cluster mais ne représentent pas la même entité seront ignorés lors de la phase de Matching. Enfin, tous 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.

3.2.2.3.1 Introduction : L'algorithme K-Modes a été introduit pour la première fois en 1998 par HUANG [Huang 1998]. Il a été proposé comme une extension du célèbre algorithme de clustering classique K-Means. Le principal avantage de K-Modes est sa capacité à traiter directement des données catégorielles ce qui n'était pas le cas avec K-Means qui n'accepte que des caractéristiques numériques. L'utilisation de K-Modes au lieu de K-Means nous a aidés à éviter la conversion numérique des données, qui est une tâche coûteuse, en particulier lorsqu'il s'agit de très grands ensembles de données, comme dans le cas du Big Data.

3.2.2.3.2 Les étapes de K-Modes : L'algorithme K-Modes est composé de 4 étapes principales :

- Tout d'abord, chaque enregistrement sera représenté sous la forme 'clé, valeur' où la clé est l'ID de cluster et la valeur l'ID d'enregistrement ;
- Dans un deuxième temps, la dissemblance est mesurée entre chaque objet et les modes des clusters avec les équation 3.1 et 3.2. Chaque objet sera affecté à son nouveau cluster le plus proche avec la mise à jour du cluster-ID dans la paire 'clé, valeur' ;

- Troisièmement, tous les enregistrements avec le même identifiant de cluster en tant que clé sont combinés, puis calculent le nouveau mode de chaque cluster ;
- Les étapes 2 et 3 sont répétées jusqu'à ce qu'aucun objet ne modifie son affectation.

3.2.2.3.3 Les avantages de l'algorithme K-Modes

- Traite les ensembles d'apprentissage catégoriques,
- Simple, rapide,
- Converge après quelques itérations

3.2.2.3.4 Les inconvénients de l'algorithme K-Modes

- Fait face au problème de la non-unicité du mode du cluster,
- Le choix des k modes initiaux est aléatoire.

3.2.2.3.5 Fonctionnement de l'algorithme K-Modes : D'après (HUANG, 1998) l'algorithme K-Modes est un processus en 4 étapes :

- Choisir k modes parmi les objets formant ainsi k clusters ;
- (Ré)affecter chaque objet O au cluster Ci tel que d(O,Modei) est minimal ;
- Mettre à jour le mode de chaque cluster ;
- Aller à l'étape (2) jusqu'à stabilisation des objets.

3.2.2.3.6 La distance de K-Modes : L'algorithme k-modes utilise le matching simple comme mesure de dissimilarité.

On a deux objets X1 et Y1 ayant des valeurs catégoriques: X1=(x_{11}, x_{12},...,x_{1m}) et Y1=(y_{11}, y_{12},...,y_{1m})
On a m attributs , Le matching simple est défini:

$$d(X1, Y1) = \sum_{t=1}^{m} \delta(x_{1t}, y_{1t}). \tag{3.1}$$

$$\delta(x_{1t}, y_{1t}) = \begin{cases} 0 & Si\, x_{1t} = y_{1t} \\ 1 & Si\, x_{1t} \neq y_{1t} \end{cases} \tag{3.2}$$

On a deux cas extrêmes:

- d=0 : si tous les attributs sont similaires
- d=m : si tous les attributs sont dissimilaires.

3.2.2.3.7 La mise à jour des modes Méthode à base des fréquences:

- La valeur qui se répète le plus souvent est gardée
- En cas d'égalité de nombre d'occurrence: choix aléatoire.

3.2.2.4 Matching

- La dernière étape du processus RL consiste à faire correspondre les enregistrements indexés qui se trouvent dans le même bloc. Pour classer l'enregistrement comme correspondant ou non correspondant La correspondance peut être effectuée à l'aide d'un ensemble de fonctions de similarité de chaînes comme « edit distance » , « Jaro-Winkler », « Jaccard distance » ou avec le codage phonétique en transformant une chaîne en un code qui représente la façon dont la chaîne est prononcée. Une variété d'algorithmes de codage phonétique existe dans la littérature (Soundex et phonex, phoenix, NYSIIS et Double-Metaphone).

La valeur de correspondance est normalisée entre 0 et 1 où 1 représente une correspondance exacte et 0 une non-correspondance totale.

3.3 Outils et environnement de développement

3.3.1 Environnement de développement

L'environnement de développement est un facteur important qui doit être détaillé pour connaître dans quelles situations, le même travail peut être reproduit.
La stratégie proposée dans le cadre de ce travail a été implémentée et testée dans l'environnement suivant :

- **Caractéristiques matérielles et logicielles du PC utilisé** : Nous avons développé notre application sur une machine DELL avec un processeur Intel(R) Core (TM) i5 - 8350U CPU, une vitesse de 1.70 Ghz et une capacité mémoire de 8 GB. Système d'exploitation Windows 11 professionnel de 64bits.
- **Langage utilisé** : Java.
- **IDE utilisé** : NetBeans.

3.3.2 Outils de développement

3.3.2.1 NetBeans IDE

NetBeans IDE est un environnement de développement intégré gratuit et à code source ouvert destiné au développement d'applications sous Windows, Mac, Linux et Solaris. L'environnement IDE simplifie le développement d'applications Web, d'entreprise, de bureau et mobiles utilisant les plates-formes Java et HTML5. Il offre également une assistance pour le développement d'applications PHP et C/C++.
NetBeans constitue par ailleurs une plateforme qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)). L'IDE NetBeans s'appuie sur cette plateforme.
Figure 3.1 présente fenêtre de programmation Sur Netbeans :

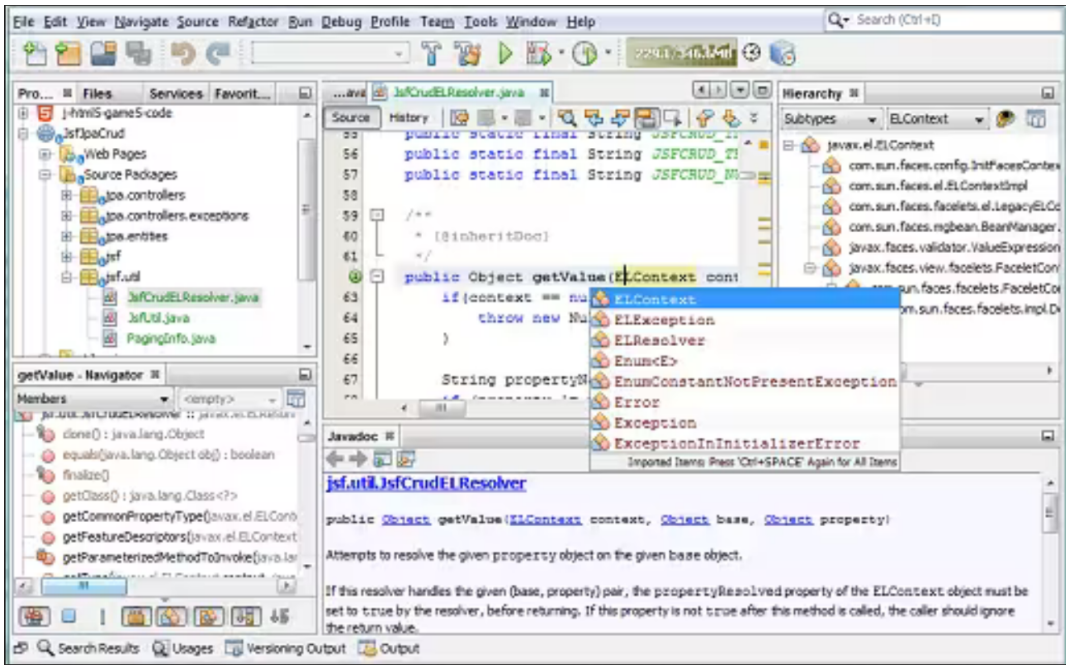

Figure 3.4: fenêtre de programmation Sur Netbeans

3.3.2.2 Langage de programmation (Java)

Java est un langage de programmation et une plateforme informatique qui ont été créés par Sun Microsystems en 1995. Beaucoup d'applications et des sites Web ne fonctionnent pas si Java n'est pas installé et leur nombre ne cesse de croître chaque jour. Java est rapide, sécurisé et fiable. Des ordinateurs portables aux centres de données, des consoles de jeux aux super ordinateurs scientifiques, des téléphones portables à Internet, la technologie Java est présente sur tous les fronts.

C'est un langage de programmation à usage général, évolué et orienté objet dont la syntaxe est proche du C. Ses caractéristiques ainsi que la richesse de son écosystème et de sa communauté lui ont permis d'être très largement utilisé pour le développement d'applications de types très disparates. Java est notamment largement utilisée pour le développement d'applications d'entreprises et mobiles.

3.3.3 Environnement Java

Java est un langage interprété, ce qui signifie qu'un programme compilé n'est pas directement exécutable par le système d'exploitation mais il doit être interprété par un autre programme, qu'on appelle interpréteur.
Figure 3.2 présente l'architecture exécutable Code java :

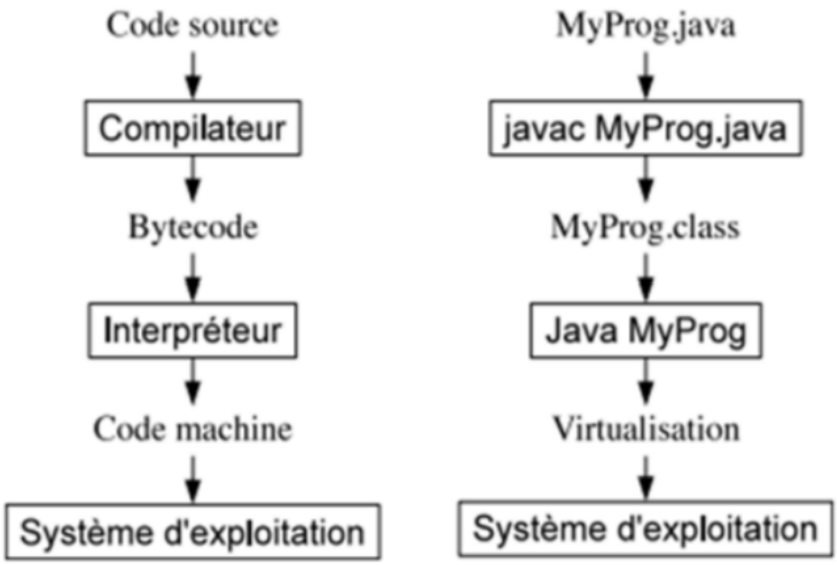

Figure 3.5: architecture exécutable Code java

3.3.4 JavaFX

3.3.4.1 JavaFX Intégration

JavaFX est une bibliothèque graphique intégrée dans le JRE et le JDK de Java. Oracle la décrit comme « The Rich Client Platform », c'est-à-dire qu'elle permet de réaliser des interfaces graphiques évoluées et modernes grâce à de nombreuses fonctionnalités, telles que

les animations, les effets, la 3D, l'audio, la vidéo, etc. Elle a de plus l'avantage d'être dans le langage Java, qui permet de réaliser des architectures avec des paradigmes objet, et aussi de pouvoir utiliser le typage statique. Dans ce premier tutoriel, nous allons voir ensemble un rapide historique de la bibliothèque pour ensuite découvrir les fondamentaux qui sont les classes « Stage », « Scene », « Application » et le « threading » associé, pour finir nous verrons les « Node » avec un exemple d'utilisation du « scene graphe ». Cette présentation ne fait pas dans le bling-bling, même si JavaFX est doué pour cela, en préférant se focaliser sur les concepts primordiaux d'une telle bibliothèque. Bien comprendre ces basiques vous aidera bien à commencer pour ensuite pouvoir faire des interfaces de qualité et peut-être spectaculaires.

Figure 3.3 Illustre un projet Java FX Main :

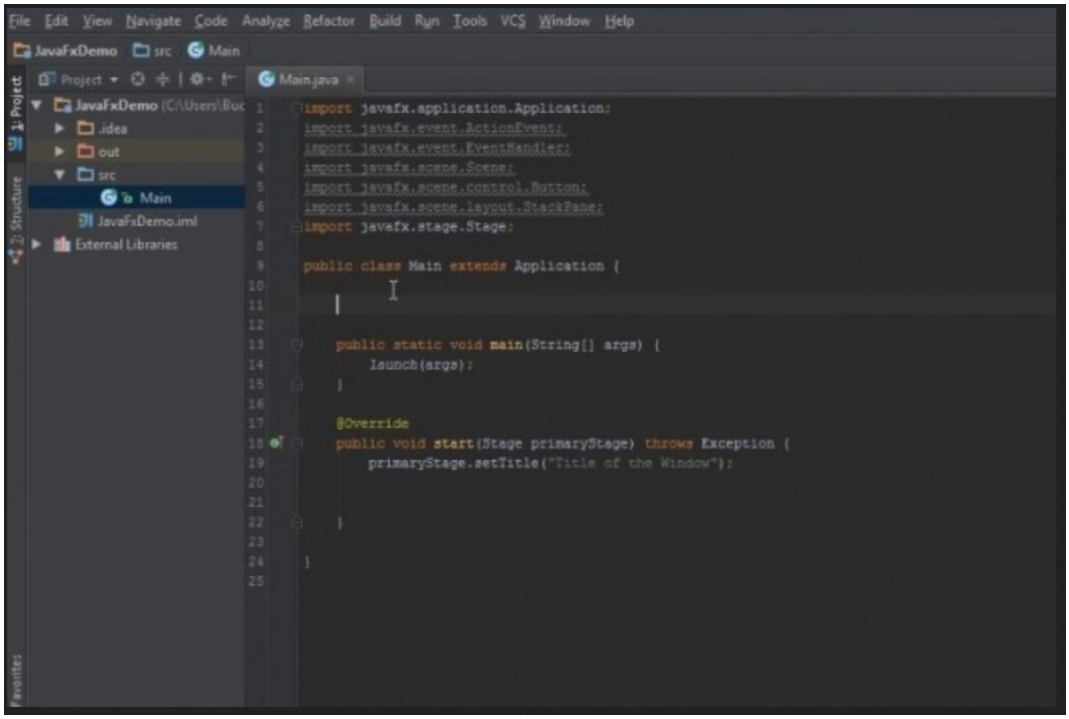

Figure 3.6: projet Java FX Main

3.3.4.2 Scene Builder

JavaFXSceneBuilder (Scene Builder) vous permet de concevoir rapidement des interfaces utilisateur d'application JavaFX en faisant glisser un composant de l'interface utilisateur d'une bibliothèque de composants de l'interface utilisateur et en le déposant dans une zone d'affichage du contenu. Le code FXML de la mise en page de l'interface utilisateur que vous créez dans l'outil est automatiquement généré en arrièreplan.

Scene Builder peut être utilisé comme un outil de conception autonome, mais il peut également être utilisé avec des IDE Java pour que vous puissiez utiliser l'IDE pour écrire, construire et exécuter le code source du contrôleur que vous utilisez avec l'interface utilisateur de votre appli cation. Bien que SceneBuilder soit plus étroitement intégré à l'EDINetBeans, il est également intégré aux autres EDI Java décrits dans ce document. L'intégration vous permet d'ouvrir un document FXML à l'aide de Scene Builder, d'exécuter les exemples Scene Builder et de générer un modèle pour le fichier source du contrôleur.
Figure 3.4 présente l'utilisation Java FX Scene Builder :

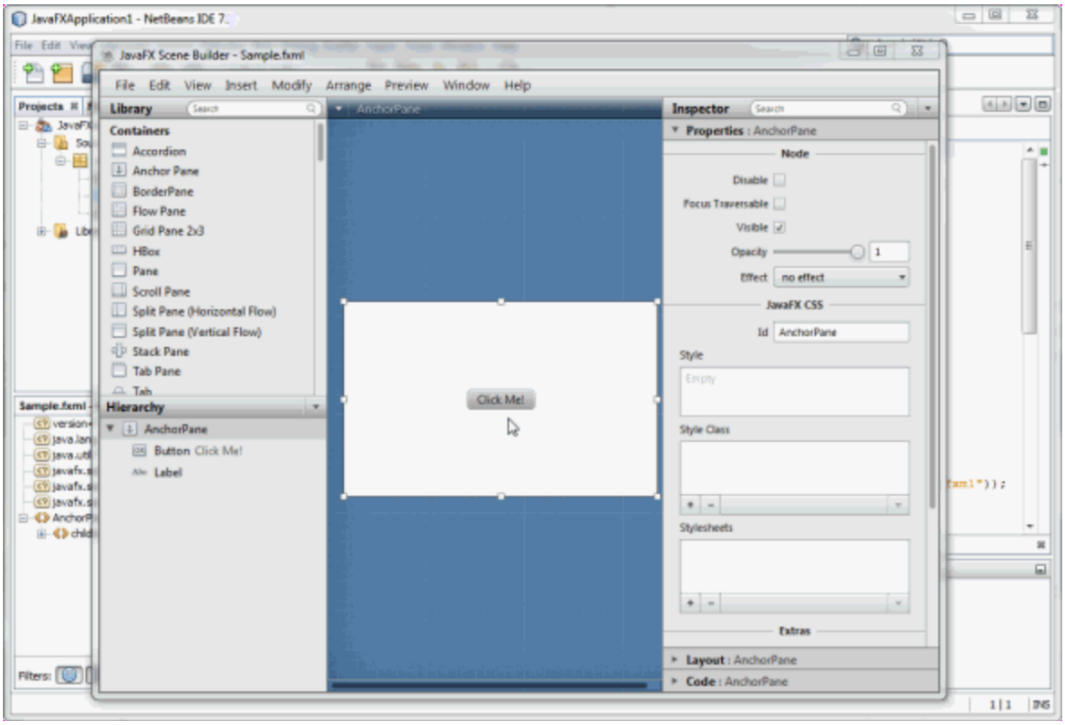

Figure 3.7: Utilisation Java FX Scene Builder

3.4 Implémentation et expérimentation

3.4.1 Présentation de L'application

L'application de Matching qu'on a développé en se basent sur K-Modes lors de la conception est la suivante :

3.4.1.1 Interface d'accueil

La figure 3.12 montre La page d'accueil de notre application qui nous permet de charger l'ensemble de données utilisé dans notre système

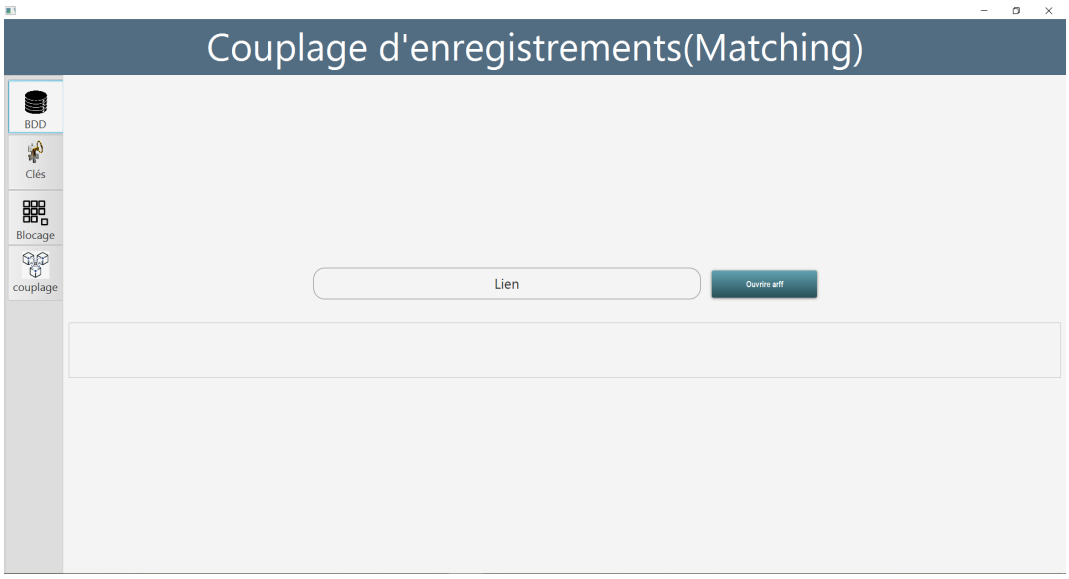

Figure 3.8: Page d'accueil de l'application

3.4.1.2 Sélection de fichier data set.arff

L'ensemble de données Restaurant est chargé dans un fichier.arf

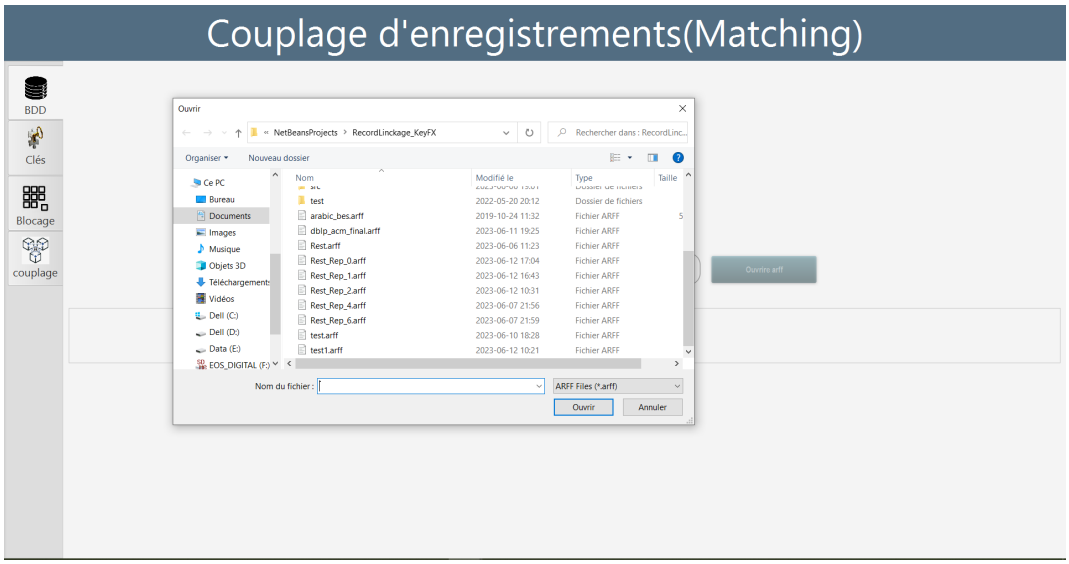

Figure 3.9: Sélection de fichier data set.arff

3.4.1.3 Chargement de l'ensemble de données dans l'application

La figure 3.14 montre le chargement de Data set Restaurant dans notre application afin qu'on puisse travailler avec :

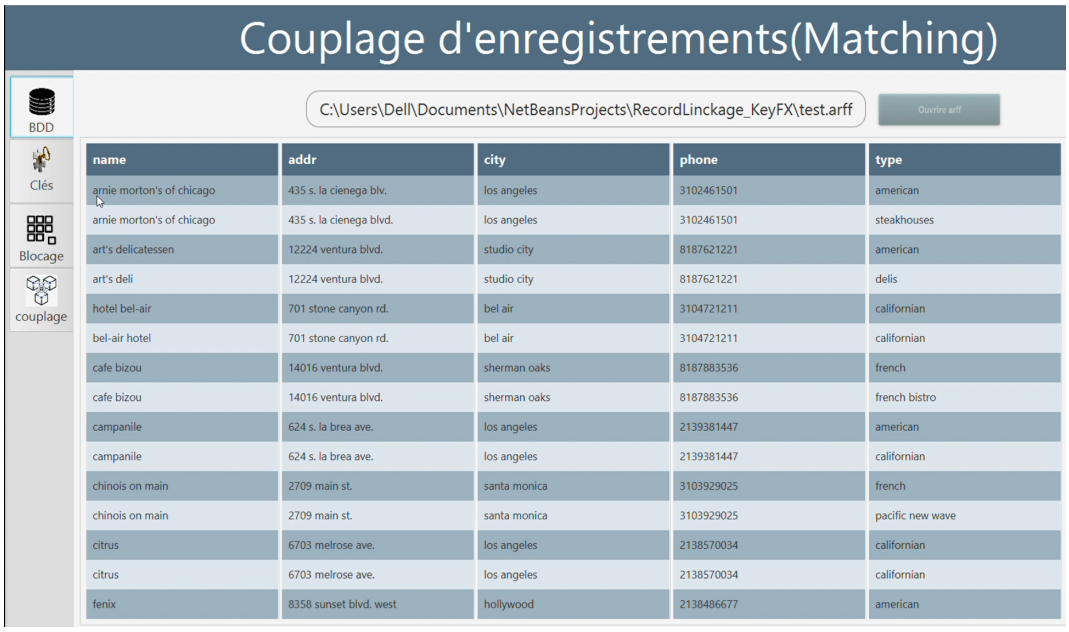

Figure 3.10: Chargement de l'ensemble de données dans l'application

3.4.1.4 Génération des clés de blocage

La figure 3.15 nous permet de créer les clés de blocage. Comme le montre la figure on peut créer plusieurs clés soit par leur valeur dans l'ensemble de données soit par les techniques de Matching Soundex et NYSIIS. La clé peut être une valeur d'un seul attribut ou une concaténation de deux comme il est montré dans la figure 3.2.

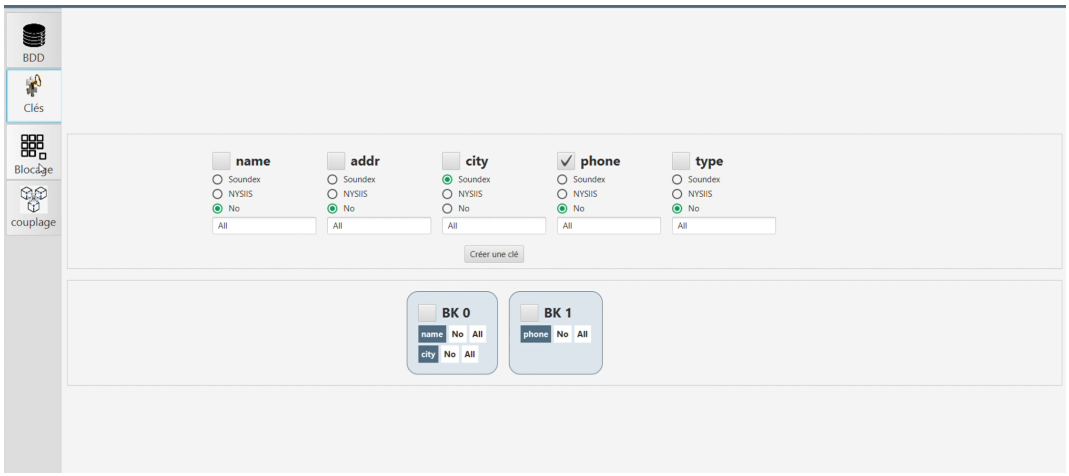

Figure 3.11: Interface de création des clés de blocage

3.4.1.5 Interface de Création des Blocks

La figure 3.17 montre l'interface qui nous permet de choisir le nombre de bloc qu'on veut avoir pour chaque test :

Figure 3.12: Création Les Blocks

3.4.1.6 Information des blocs

La figure 3.16 montre les information principale utilisé pour chaque bloc :

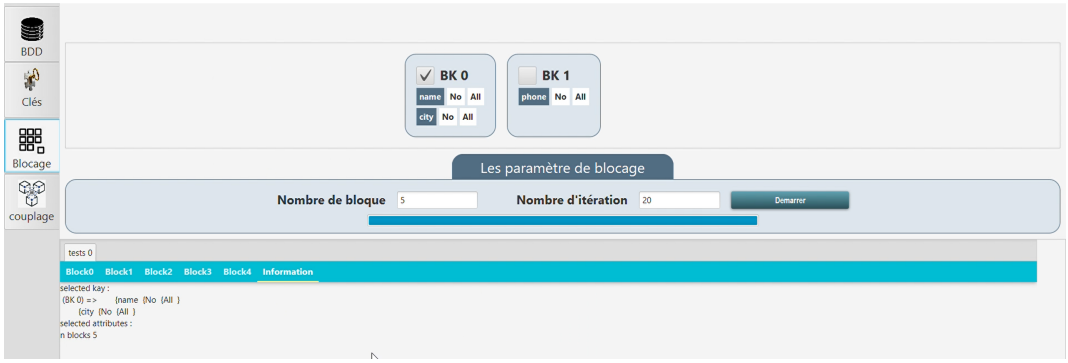

Figure 3.13: Information des blocs

3.4.1.7 information de détail de matching

La figure 3.18 montre les information détaillés de couplage pour deux enregistrement :

Figure 3.14: information de détail de matching

3.4.1.8 Résultat de Matching

La figure 3.19 montre les résultat obtenues après le Matching.

Comme on peut voir 3 résultats déférentes peuvent être obtenues. True Matche pour une correspondance entre deux enregistrements, No matche pour le non correspondance, possible Matche pour un résultat égal au seuil défini dans notre application.

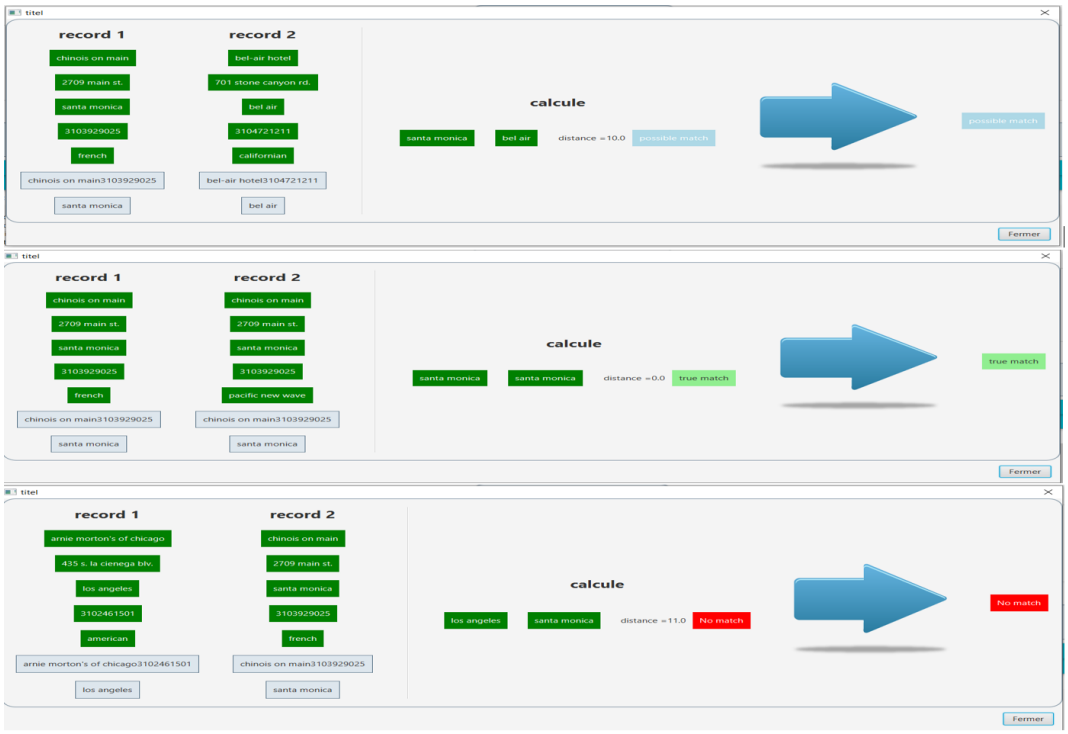

Figure 3.15: Résultats de Matching

3.4.1.9 Résultat statiques de Matching

La figure 3.20 montre les résultats statiques de Matching

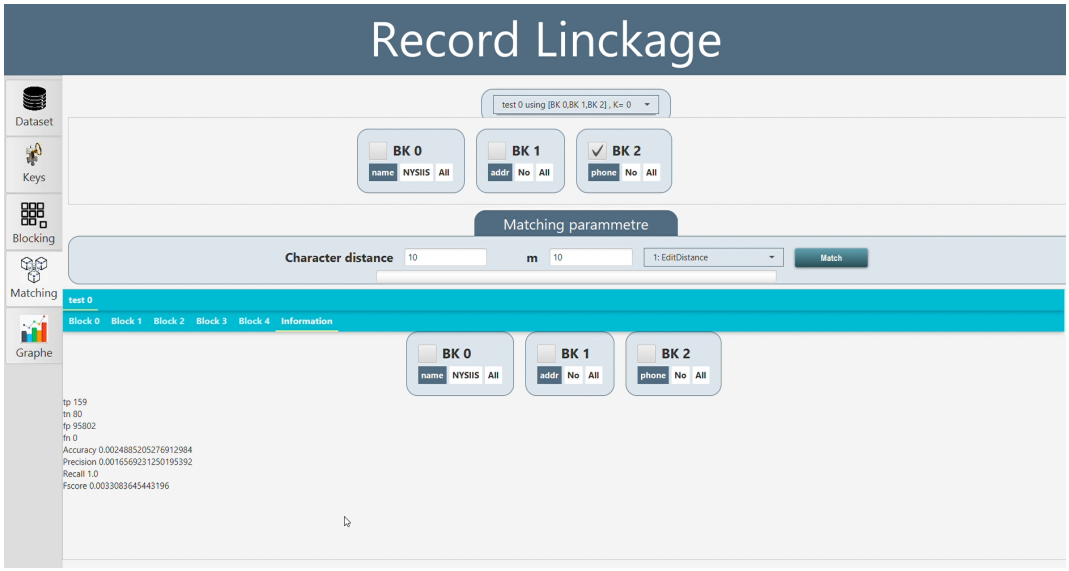

Figure 3.16: résultats statiques de Matching

3.4.1.10 Résultat graphique de Matching

La figure 3.21 montre les résultats graphiques de Matching

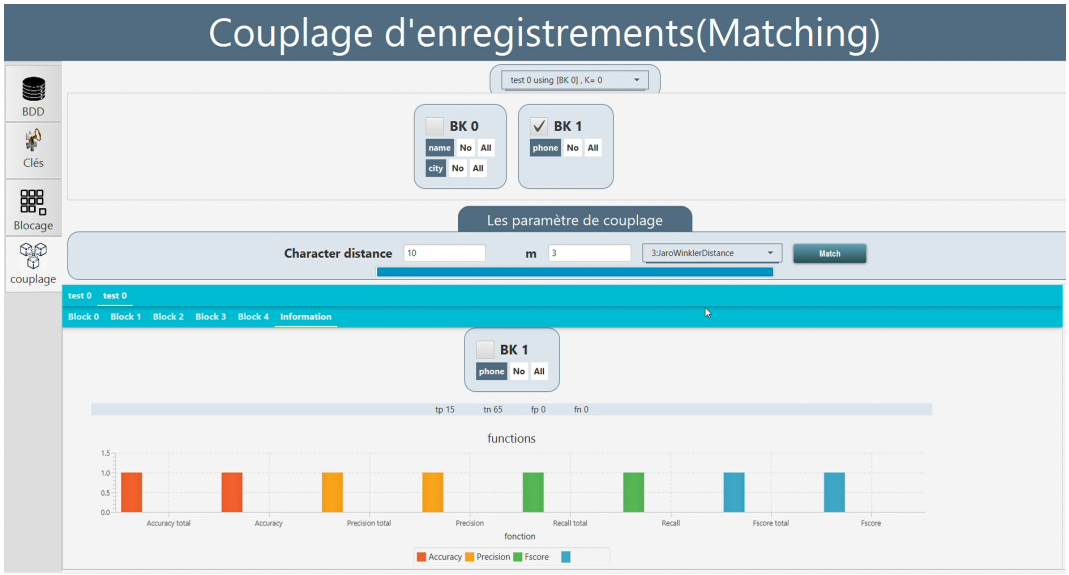

Figure 3.17: résultats graphiques de Matching

3.4.2 Evaluation

Dans la communauté Record Linkage, quatre paramètres principaux sont utilisés pour mesurer la performance d'un processus RL.

3.4.2.1 Accuracy

Cette métrique est utilisée pour mesurer à quel point la comparaison est exacte.

$$A = \frac{Tp + Tn}{Tp + Fp + Fn + Tn} \tag{3.3}$$

3.4.2.2 Précision

Cette métrique est utilisée pour mesurer la précision des comparaisons.

$$P = \frac{Tp}{Tp + Fp} \tag{3.4}$$

3.4.2.3 Recall

Cette métrique est utilisée pour mesurer Le ratio de liens correctement prédits à partir de toutes les correspondances vraies.

$$R = \frac{Tp}{Tp + Fn} \tag{3.5}$$

3.4.2.4 F-Score

F-Score est utilisé pour mesurer la moyenne harmonique entre les deux paramètres précédents.

$$F = \frac{2 * P * R}{P + R} \tag{3.6}$$

Avec :

Tp (Vrais positifs) : paires qui apparaissent dans le même cluster à la fois dans la vérité terrain et dans la prédiction. Connu sous le nom de vrais matchs.

Tn (Vrais négatifs) : paires qui apparaissent dans différents clusters à la fois dans la vérité terrain et dans la prédiction. Connu sous le nom de véritables non-correspondances.

Fp (Faux positifs): paires qui apparaissent dans le même cluster dans la prédiction mais dans des clusters différents dans la vérité terrain. Connu sous le nom de fausses correspondances.

Fn (Faux négatifs) : paires qui apparaissent dans le même cluster dans la vérité terrain mais dans des clusters différents dans la prédiction. Connu sous le nom de faux non-matchs ou matchs manqués.

3.4.3 Résultats généraux

Après avoir divisé l'ensemble de données Restaurant en un ensemble de bloc, on a calculer la correspondance des enregistrement dans chacun (Matching) d'où 3 métriques de calcule de similarité sont utilisé : Editer la distance ; distance de Jacard et la métrique de similarité Jaro-winkler . Dans notre exemple nous avons opté pour Editer distance avec m = 2 (seuil de distance)

Les résultats des expériences sont représentés dans le tableau 3.1

Table 3.1: Les résultats des expériences

Nb-Clustre	A	P	R	F
5	0 ,7999	0 ,8784	0 ,7584	0 ,7984
10	0 ,8599	0 ,9812	0 ,7952	0 ,8081
15	0 ,8998	0 ,9812	0 ,8405	0 ,8958
20	0 ,9498	0 ,6603	0 ,9050	0 ,9276
25	0 ,9998	0 ,4491	1,0	0 ,9994

Tout d'abord, nous avons évalué le taux d'exactitude de notre proposition de blocage basée sur l'algorithme K-Modes. Nous avons fait varier le nombre de clusters de 5 à 25 clusters et mesuré les résultats obtenus concernant l'accuracy à chaque fois.Comme nous pouvons le voir à partir des résultats obtenus sur la table 3.1, Accuracy continue de s'améliorer avec l'augmentation du nombre de clusters. Et c'est tout à fait normal car en divisant les données en plusieurs blocs (clusters), nous réduisons davantage le nombre de comparaisons d'enregistrements.

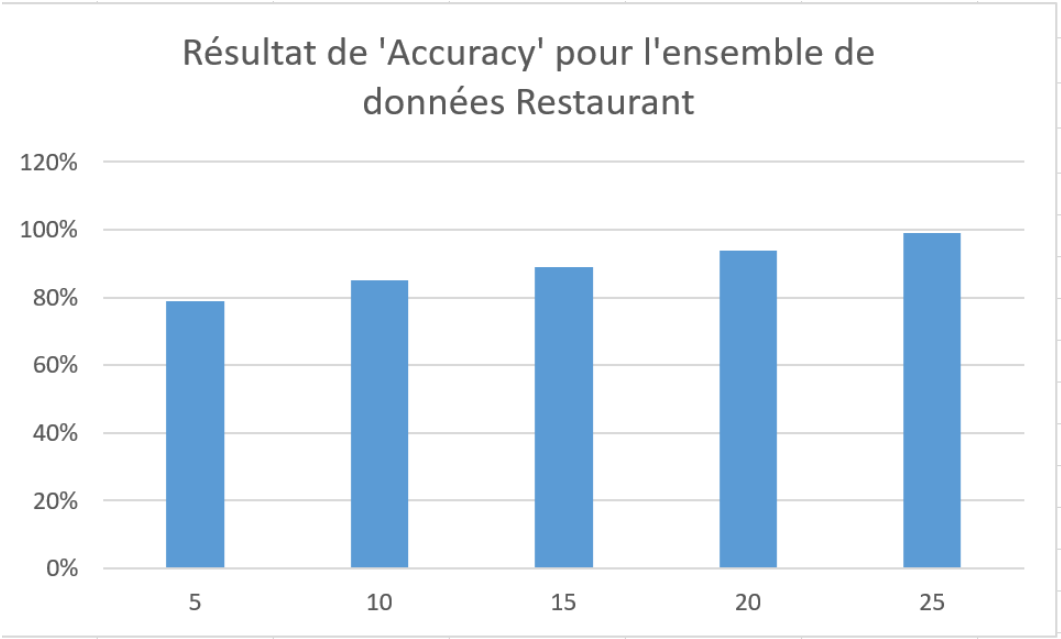

Figure 3.18: Résultat de la métrique Accuracy

Pour les expériences de précision, nous avons fait varier le nombre de clusters de 5 à 25 et nous avons mesuré Le ratio de liens correctement prédits parmi toutes les prédictions de liens positifs qui est P en utilisant la métrique de similarité Modifier la distance (ED)). Le seuil 2 Edits pour ED. Comme nous pouvons le voir, le résultat obtenu le plus élevé est avec

10 et 15 clusters où 0 ,9812 de précision est détecté. Ca s'explique par le fait qu'avec plus de clusters on perd plus de vrais appariements car certains vrais appariements peuvent être attribués à différents clusters et ne seront pas comparés entre eux.

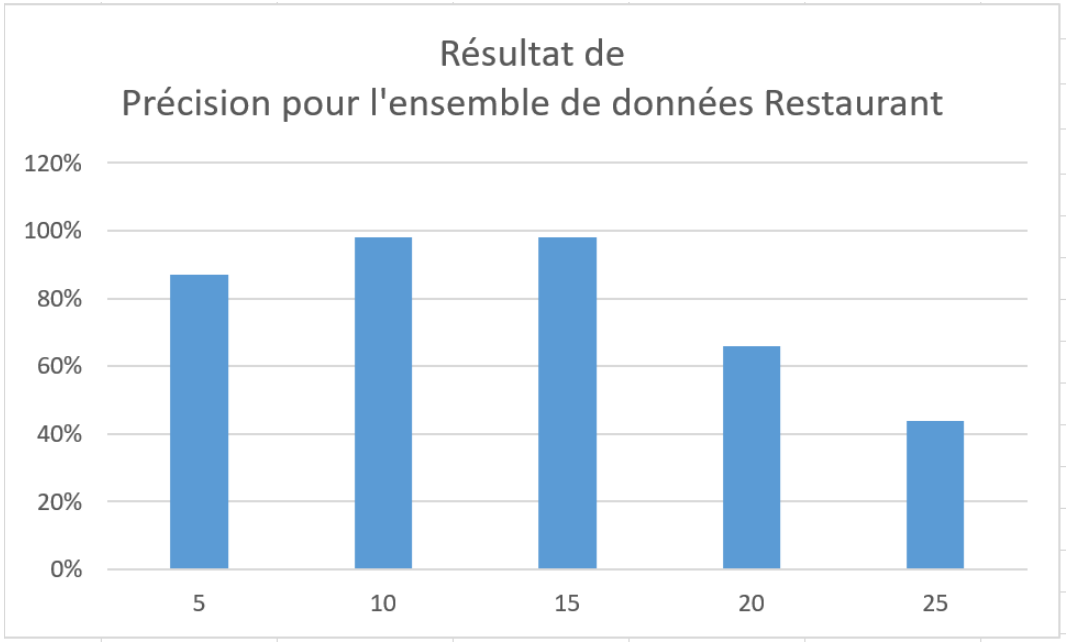

Figure 3.19: Résultat de la métrique Précision

Pour le Recall et le Fmesure on voit qu'avec plus de clustre les résultats seront plus élevés Ca s'explique par le fait que plus il y'aura plus de clustre moins qu'il y'aura des comparaisons.

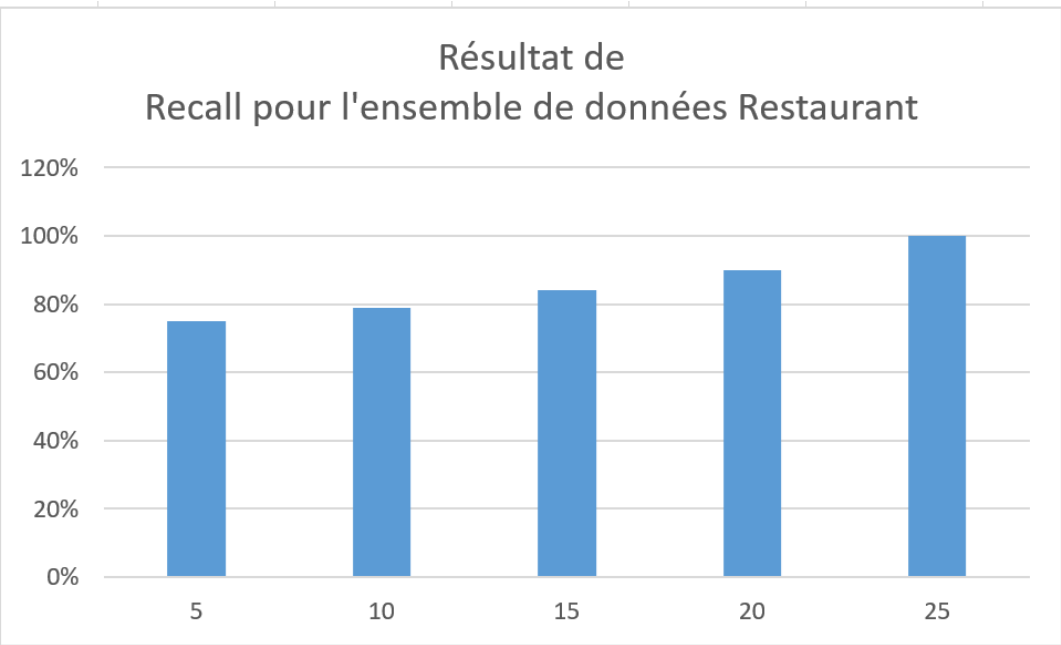

Figure 3.20: Résultat de la métrique recall

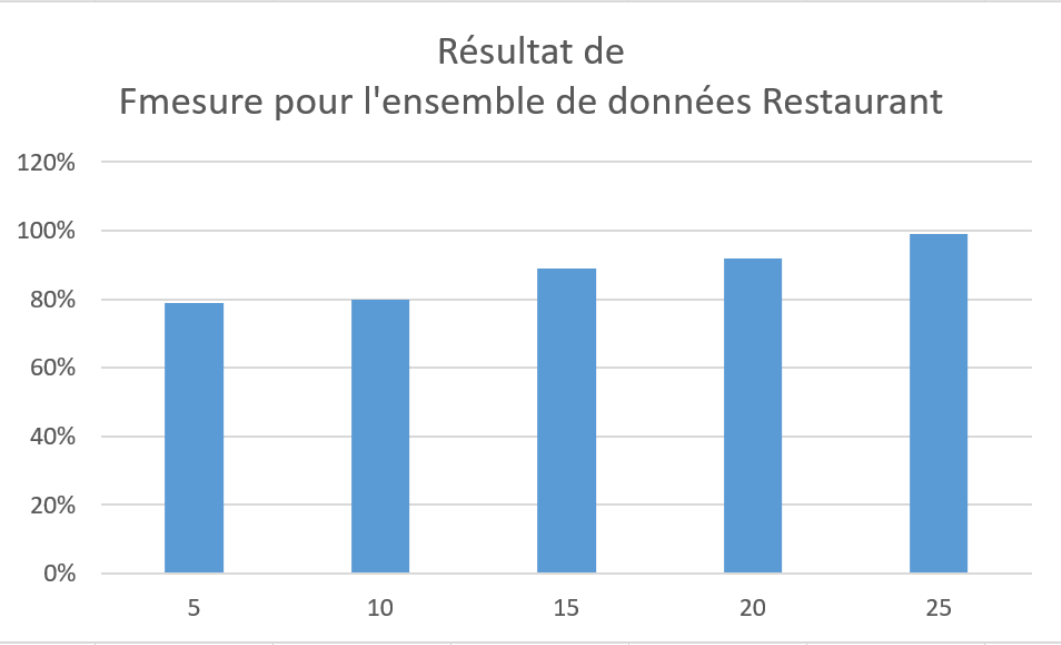

Figure 3.21: Résultat de la métrique Fmesure

3.4.4 Discussion des résultats

Les résultats de l'expérience concernant le RL basé sur les K-Modes montrent l'efficacité de notre proposition. Les résultats obtenus ont montré la grande efficacité et l'évolutivité de la solution proposée par rapport aux problèmes de couplage d'enregistrement. D'où Nous avons constaté que notre contribution remplisse les critères pour lesquels elle est conçue.

3.5 Conclusion

Dans ce chapitre, nous avons présenté le jeu de donnée qui a servi à l'évaluation de notre contribution : nous avons présenté notre contribution aux principaux défis du couplage d'enregistrements. Tout d'abord, pour réduire l'espace de recherche et éviter de comparer chaque enregistrement de l'ensemble de données à tous les autres pour la correspondance. Une technique de blocage a été proposée, l'idée était d'utiliser l'algorithme K-Modes pour regrouper les données dans un ensemble de blocs en utilisant uniquement les clés de blocage comme attributs de regroupement. Cela rassemblera tous les enregistrements similaires dans le même cluster où la correspondance n'est effectuée qu'entre les paires d'enregistrements du même bloc. Ensuite, nous avons présenté l'étude expérimentale fait pour évaluer notre contribution. Dans un premier temps, nous avons défini les métriques utilisées pour évaluer notre travail. Ensuite, Les résultats de l'expérience concernant le RL basé sur les K-Modes ont été montrés. Et on a fini le chapitre par la présentation et l'implémentation de notre application.

Conclusion Générale

A l'ère du Big Data, les données sont générées chaque jour a un rythme explosif. Prendre des décisions basées sur des données collectées de partout sans veiller a leur qualité peut avoir un impact négatif sur plusieurs aspects (Finance, business, réputation). Selon le groupe Gartner, 40 °/. des initiatives commerciales n'ont pas atteint leurs objectifs en raison de la mauvaise qualité des données. Pour passer des problèmes de la mauvaise qualité des données résultant de l'intégration de données venant de déférentes sources distribués, autonome et hétérogène comme que les erreurs manuelles des différentes techniques sont utilisées.

Les problèmes de qualité des données peuvent apparaître de différentes manières : valeurs manquantes, valeurs en double, problèmes d'intégrité référentielle et bien d'autres encore.

L'impact de la mauvaise qualité des données sur les systèmes d'aide à la décision, les problèmes liés à la mauvaise qualité des données qui peuvent apparaître de différentes manières : valeurs manquantes, valeurs en double, problèmes d'intégrité référentielle et bien d'autres encore nécessite le développement et l'implémentation des systèmes et des applications pour les détecter et les corriger.

La méthode proposée qui est basée sur l'algorithme K-Modes comme étape d'indexation et la sélection semi-automatique des attributs pour la génération des clés de blocage a prouvé ces performances après avoir évalué sur l'ensemble de data set utilisée. Les résultats obtenus ont montré la grande efficacité et l'évolutivité de la solution proposée.

perspectives

Comme travaux futurs, nous nous intéressons aux points suivants :

- Proposer une approche de sélection automatique des attributs est un défi omniprésent dans la communauté RL qui vise à donner les meilleures clés de blocage possibles.
- Il est possible de repondérer les résultats finals et de les corriger pour les faux non-appariements.
- Proposer une approche de filtrage pour réduire encore plus l'espace de comparaison.

Bibliographie

[1] Thomas C Redman. Bad data costs the us $3 trillion per year. *Harvard Business Review*, 22:11–18, 2016.

[2] Jonathan G Geiger. Data quality management, the most critical initiative you can implement. *Data Warehousing, Management and Quality, Paper*, pages 098–29, 2004.

[3] Vic Barnett, Toby Lewis, et al. *Outliers in statistical data*, volume 3. Wiley New York, 1994.

[4] Carlo Batini, T Catarci, and M Scannapiceco. A survey of data quality issues in cooperative information systems. In *Pre-conference ER tutorial*, 2004.

[5] Louardi Bradji and Mahmoud Boufaida. Adaptation des techniques de l'extraction des connaissances à partir des données (ecd) pour prendre en charge la qualité des données. 2017.

[6] Abdelkrim OUAHAB. *Qualité de données pour l'intégration de données*. PhD thesis, Université de Sidi Bel Abbès-Djillali Liabes, 2019.

[7] Laure Berti-Équille. La qualité des données comme condition à la qualité des connaissances: un état de l'art. *Revue des Nouvelles Technologies de l'Information*, 2004.

[8] NOMS SOUS-SPECIFIES. Memoire de master.

[9] Faouzi Boufares and A Ben Salem. Heterogeneous data-integration and data quality: Overview of conflicts. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 867–874. IEEE, 2012.

[10] Richard Y Wang, Veda C Storey, and Christopher P Firth. A framework for analysis of data quality research. *IEEE transactions on knowledge and data engineering*, 7(4):623–640, 1995.

[11] Mourad Kezai. *Optimisation automatique des performances dans les entrepôts de données: étude comparative*. PhD thesis, 2018.

[12] MJ Carey, S Ceri, P Bernstein, U Dayal, C Faloutsos, JC Freytag, G Gardarin, W Jonker, V Krishnamurthy, MA Neimat, et al. Data-centric systems and applications. *Italy: Springer*, 2006.

[13] Hamid Naceur Benkhaled and Djamel Berrabah. Data quality management for data warehouse systems: State of the art. *JERI*, 2019.

[14] Peter Christen et al. Towards parameter-free blocking for scalable record linkage. 2007.

[15] Mahfoudh Ghalem Abdelkadir Regba. Intégration de données: Approche semi-automatique pour la mise en correspondance de schémas de bases de données hétérogènes. 2018.

[16] Peter Christen. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering*, 24(9):1537–1555, 2011.

[17] David E Clark. Practical introduction to record linkage for injury research. *Injury Prevention*, 10(3):186–191, 2004.

[18] Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[19] Lifang Gu and Rohan Baxter. Adaptive filtering for efficient record linkage. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 477–481. SIAM, 2004.

[20] Luis Gravano, Panagiotis G. Ipeirotis, Hosagrahar Visvesvaraya Jagadish, Nick Koudas, Shanmugauelayut Muthukrishnan, Lauri Pietarinen, and Divesh Srivastava. Using q-grams in a dbms for approximate string processing. *IEEE Data Eng. Bull.*, 24(4):28–34, 2001.

[21] Akiko Aizawa and Keizo Oyama. A fast linkage detection scheme for multi-source information integration. In *International Workshop on Challenges in Web Information Retrieval and Integration*, pages 30–39. IEEE, 2005.

[22] Timothy De Vries, Hui Ke, Sanjay Chawla, and Peter Christen. Robust record linkage blocking using suffix arrays. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 305–314, 2009.

[23] Mauricio A Hernández and Salvatore J Stolfo. The merge/purge problem for large databases. *ACM Sigmod Record*, 24(2):127–138, 1995.

[24] P Rajkovic and D Jankovic. Adaptation and application of daitch-mokotoff soundex algorithm on serbian names. In *XVII Conference on Applied Mathematics*, volume 12, 2007.

[25] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304, 1998.

[26] Gema Bello-Orgaz, Jason J Jung, and David Camacho. Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45–59, 2016.

[27] Arvind Sathi. *Big data analytics*. Mc Press, 2012.

[28] AnHai Doan, Alon Halevy, and Zachary Ives. *Principles of data integration*. Elsevier, 2012.

[29] Jens Bleiholder and Felix Naumann. Data fusion. *ACM computing surveys (CSUR)*, 41(1):1–41, 2009.

[30] WE Winkler. The state of record linkage and current research problems. statistics of income division, internal revenue service publication r99/04, 1999.

[31] Mikhail Bilenko and Raymond J Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48, 2003.

[32] Sheila Tejada, Craig A Knoblock, and Steven Minton. Learning domain-independent string transformation weights for high accuracy object identification. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 350–359, 2002.

[33] Tim Churches, Peter Christen, Kim Lim, and Justin Xi Zhu. Preparation of name and address data for record linkage using hidden markov models. *BMC Medical Informatics and Decision Making*, 2(1):1–16, 2002.

ملخص

يُعد ربط السجل مشكلة مهمة في جودة البيانات، إنها عملية الكشف عن جميع السجلات التي تشير إلى نفس كيان العالم الحقيقي، ثم دمجها في مجموعة واحدة.

.من أجل تقليل العدد الكبير من المقارنات ، تتمثل تقنية الحظر في إنشاء مجموعة من الكتل تشترك في قيمة مشتركة تسمى قيمة مفتاح الحظر
بناءً على تحديات ربط السجل، قمنا بتكييف خوارزمية ك- مود كخطوة حظر والغرض منها هو تحسين وقت التنفيذ والتحكم في عدد الكتل وعدد البيانات لكل كتلة.

أظهرت الطريقة المقترحة أن اختيار السمات أو الخصائص ذات الصلة عند إنشاء مفاتيح الحظر قد أثبت تأثيره على جودة البيانات التي تم الحصول عليها.

تشمل خصائص السمات التي تؤثر على قرار الاختيار مستوى الأخطاء في قيم السمات وعدد (وتوزيع) قيم السمة ، أي محتوى معلومات السمة.

الكلمات الرئيسية: جودة البيانات ، ربط السجل ، خوارزمية ك- مود ، الحجب ، مفتاح الحجب

Abstract

Record linkage (RL), is an important issue for data quality. It is the process of detecting all records that refer to the same real-world entity and then merging them into a single tuple. In order to reduce the large number of comparisons, the technique of blocking (Blocking) consists in creating a set of blocks which share a common value named value of blocking key (Bloking Key Value) BKV.
Based on the challenges of RL, we have adapted the K-Modes algorithm as a blocking step whose purpose is to improve the execution time and to control the number of blocks and the number of data per block.
The proposed method has shown that the selection of relevant attributes when generating blocking keys has proven its influence on the quality of data obtained.
Attribute characteristics that affect the selection decision include the level of errors in attribute values and the number (and distribution) of attribute values, i.e. the information content of the attribute.

Keywords: Data quality, record linkage, K-Modes algorithm, blocking, blocking key.

Résumé

Le couplage d'enregistrement également appelée **Record Linkage (RL)** est un enjeu important pour la qualité de données. C'est le processus qui vise à détecter tous les enregistrements qui font référence à la même entité du monde réel, puis à les fusionner en un seul tuple.
Afin de réduire le nombre important de comparaisons, la technique de blocage (Blocking) consiste à créer un ensemble de blocs qui partagent une valeur commune nommée valeur de clé de blocage (Bloking Key Value) BKV.
Sur la base des défis de RL Nous avons adapté l'algorithme K-Modes comme étape de blocage dont le but est d'améliorer le temps d'exécution et de contrôler le nombre de bloc et le nombre de données par bloc.
La méthode proposée a montré que la sélection des attributs pertinents lors de génération des clés de blocage a prouvé son influence sur la qualité de données obtenues.
Les caractéristiques d'attributs qui affectent la décision de sélection comprennent le niveau d'erreurs dans les valeurs d'attribut et le nombre (et la distribution) des valeurs d'attribut, c'est-à-dire le contenu informationnel de l'attribut.

Mots clés : Qualité des données, couplage d'enregistrement, L'algorithme K-Modes, Blocage, Clé de blocage.