الجمهورية الجزائرية الديمقراطية الشعبية وزارة التعليم
العالي والبحث العلمي
جامعة سعيدة د. مولاي الطاهر
كلية الرياضيات و الإعلام الآلي و الاتصالات السلكية و اللاسلكية
قسم: الإعلام الآلي

Mémoire de Master en informatique

Spécialité : Réseaux informatiques et systèmes réparties (RISR)

T h è m e

Utilisation des algorithmes méta-heuristiques pour la résolution d'entité

▪ **Présenté par :**

MOKHTARI Toumia

SADOUKI Hafsa

▪ **Dirigé par :**

Mr BENYAHIA Miloud

Année universitaire 2024-2025

Remerciements

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

Nous tenons à remercier vivement notre encadreur pour son aide, ses conseils, son encouragement.

Nous voulons aussi remercier tous ceux qui ont Contribue à Réaliser ce projet ;

Nos professeurs qui nous ont enseigné,

Nos amis

Les membres de jury de notre travail.

Les familles SEDDOUKI et BOUOUJA

Dédicace

Je dédie ce travail :

A ma mère

A mon père

Aux familles « SEDDOUKI » et « BOUOUJA »

A mes amis Toumia, Hanane et Mokhtaria

A ceux que j'aime

Et à toute personne qui a contribué à la réalisation de ce mémoire.

Hafsa

Dédicace

Je dédie ce modeste travail :

A mes chers parents, Pour leur patience illimitée, leur encouragement continu, leur aide, en témoignage de mon profond amour et respect pour leurs grands sacrifices

A la famille « MOKHTARI »

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.

Toumia

خلاصة

عند تحسين مفاتيح الحظر .في سياق البيانات الضخمة، ي مثل تحديد الكيانات تحدي"ا رئيسي"ا للحفاظ على جودة البيانات
) مح ' سن)GWOبحث النسر الأصلع(، و BES وتجميع الكيانات المتشابهة، ي عد' تطبيق أساليب الاستدلال الفوقي، مثل
تحسين نقار الخشب(، فعا "لا بشكل خاص في توليد مفاتيح الحظر نظ "را لتعقيد قواعد البيانات WPOالذئب الرمادي(، و وحجمها الهائل

ت م 'كن هذه الخوارزميات، المستوحاة من الطبيعة، من استكشا ,,ف دقي ,,ق لمجال الحلول، وتضمن أدا "ء فائق"ا من حيث الدقة والاسترجاع
والكفاءة الحسابية .ويهدف تطبيق هذه الأساليب إلى تحسين موثوقية أنظمة المعلومات وسرعتها ودقتها، مع تلبية متطلبات بيئات البيانات واسعة النطاق.

الكلمات المفتاحية: جودة البيانات؛ حل الكيان؛ البيانات الضخمة؛ BES؛ GWO؛ WPO؛ مفتاح الحظر.

Abstract

In the context of Big Data, entity identification is a major challenge for maintaining data quality. When optimizing blocking keys and grouping similar entities, the implementation of metaheuristics such as BES (Bald Eagle Search), GWO (Grey Wolf Optimizer), and WPO (Woodpecker Optimization) is particularly effective in generating blocking keys given the complexity and sheer volume of databases.

These algorithms, inspired by nature, enable judicious exploration of the solution space and ensure superior performance in terms of precision, recall, and computational efficiency. The implementation of these methods aims to improve the reliability, speed, and accuracy of information systems, while meeting the requirements of large-scale data environments.

Keywords: Data quality; entity resolution; RE; Big Data; BES; GWO; WPO; Blocking key.

.

Résumé

Dans le cadre du Big Data, l'identification des entités représente un enjeu majeur pour maintenir la qualité des données.

Dans le cadre de l'optimisation des clés de blocage et du regroupement d'entités similaires, la mise en œuvre de métaheuristiques telles que BES (Bald Eagle Searche), GWO (Grey Wolf Optimizer) et WPO (Woodpecker Optimization) est particulièrement efficace dans la génération des clés de blocage face à la complexité et au volume considérable des bases.

Ces algorithmes, qui s'inspirent de la nature, permettent une exploration judicieuse de l'espace des solutions et assurent des performances supérieures en matière de précision, rappel et efficacité de calcul.

L'implémentation de ces méthodes vise à améliorer la fiabilité, la rapidité et l'exactitude des systèmes d'information, tout en répondant aux attentes des contextes de données à grande échelle.

Mots-clés : Qualité des données ; résolution d’entités ; RE ; Big Data; BES ; GWO; WPO; Clé de

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

BES	**B**ald **E**agle **S**earche
GWO	**G**rey **W**olf **O**ptimizer
WPO	**W**ood**p**ecker **O**ptimization
ER	**E**ntity **R**esolution
UTL	**E**xtraction **T**ransformation **L**oading
SVM	**S**upport **V**ectors **M**achines
IDE	**I**ntegrated **D**eveloppement **E**nvironnement
CDDL	**C**ommon **D**evelopment and **D**istribution **L**icense
XML	**E**Xtensible **M**arkup **L**anguage
HTML	**H**yper**T**ext **M**arkup **L**anguage
OS	**O**perating **S**ystem
SDK	**S**oftware **D**evelopment **K**it

Introduction Générale

La résolution d'entités a pour objectif de détecter et de rassembler automatiquement des enregistrements qui référencent la même entité dans diverses bases de données. Ce processus s'avère complexe en raison de : l'immensité des données, la diversité des rédactions (par exemple : noms, adresses) et le besoin d'harmoniser la précision avec le coût en termes de calcul.Face à cette situation, ce souci devient vite complexe et onéreux à résoudre par des techniques exactes]20[.

C'est à ce point que l'importance des métaheuristiques se révèle pleinement. Elles proposent une méthode efficace pour explorer intelligemment l'ensemble des solutions envisageables. À l'inverse des méthodes exactes, qui se révèlent inapplicables pour de vastes jeux de données.

Pour ce mémoire, notre attention s'est portée sur la résolution d'entités grâce à l'utilisation des trois algorithmes métaheuristique. BES (Bald Eagle Search), GWO (Grey Wolf Optimizer)et WPO (Woodpecker Optimization) qui offrent la possibilité de trouver des solutions de qualité supérieure dans un délai raisonnable, sans nécessiter l'examen exhaustif de toutes les combinaisons.

Ces algorithmes, en s'appuyant sur leur aptitude à concilier exploration (recherche de nouvelles solutions) et exploitation (perfectionnement des solutions déjà en place), optimisent efficacement la production des clés de blocage : une phase cruciale dans la résolution d'entité. En perfectionnant les clés de blocage, ils réussissent à minimiser le nombre de comparaisons superflues entre différents enregistrements, tout en préservant ou renforçant la précision et le rappel du processus d'appariement.

De plus, les métaheuristiques possèdent une flexibilité et une adaptabilité élevées, ce qui leur permet de se conformer aux particularités de divers types de données ou standards d'évaluation sans avoir besoin de changements significatifs. Ce qui les rend indispensables pour résoudre des problématiques complexes.

Pour conclure, nous terminerons ce mémoire avec une conclusion générale.

Le mémoire est organisé en trois chapitres :

- Dans le premier chapitre, nous avons examiné la qualité des données. Cette recherche nous a fourni une vue d'ensemble des méthodes actuelles de gestion de la qualité des

données présentées dans la littérature, ainsi que l'axe où une contribution devrait être apportée.

- Le deuxième chapitre décrit les diverses phases de la résolution d'entités et propose une analyse centrée sur les principaux enjeux de la RE en matière de qualité des données.

- Le chapitre 3 expose la présentation de la méthode suggérée ainsi que les algorithmes employés. Nous présentons et discutons les expériences et les résultats de notre application.
  .

Nous terminons par une conclusion générale, la bibliographie et les annexes.

CHAPITRE 01 : La qualité des données

1. Introduction

Dans le contexte du Big data La qualité des données est un enjeu majeur. En fait, les systèmes méga données génèrent, traitent et stockent de grandes quantités de données provenant de multiples sources et souvent hétérogène. La gestion de la qualité de ces données devient encore plus importante, car des données de mauvaise qualité peuvent compromettre la fiabilité des analyses, des décisions et des résultats d'une organisation.

Ce chapitre présente les dernières technologies en matière de qualité des données. Nous définirons les raisons pour lesquelles les données sont importantes dans un programme, expliquerons la « qualité des données », énumérerons les dimensions de la qualité des données, les problèmes associés à une mauvaise qualité des données et les stratégies permettant d'éviter de tels problèmes.

.

2. Big Data « Méga Données »

Gartner définit le Big Data comme une source d'informations à haut volume et à grande vitesse, et/ou une grande diversité, nécessitant de nouvelles formes de traitement pour améliorer la prise de décision, la découverte d'informations et l'optimisation des processus [Gartner, 21].

Le Big Data est un ensemble de technologies et de méthodes permettant de collecter, stocker et traiter de grandes quantités de données hétérogènes à grande vitesse et d'en tirer profit. Ces données peuvent être structurées, semi-structurées ou non structurées. De nombreux travaux de recherche décrivent les caractéristiques du Big data à travers le modèle 5V présenté dans la figure 1]19[.

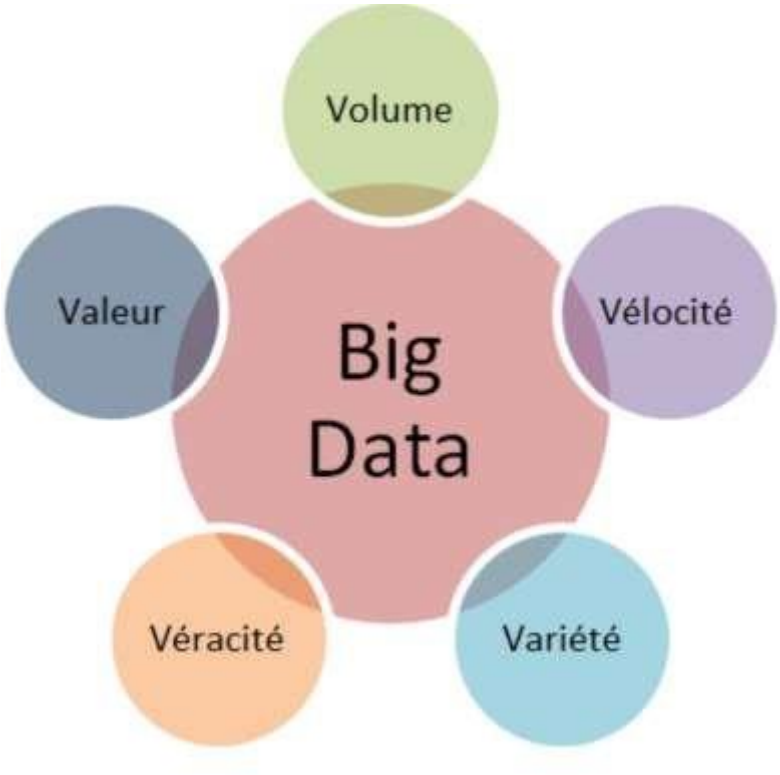

Figure 1 *:* Modèle 5V de la Big Data

Le volume de données fait référence à la quantité massive de données générées par seconde]1[;

Les données jouent un rôle important dans la mise en œuvre d'outils et de modèles de stockage et de traitement des données. **La vélocité** fait référence à la rapidité avec laquelle les données sont générées et traitées.

La Variété fait référence à l'hétérogénéité des données et de leurs sources. Les données structurées contiennent un format de données unifié qui peut être facilement géré via un système de gestion de base de données relationnelle. En revanche, les données non structurées et semi-structurées sont plus difficiles à analyser et nécessitent une infrastructure de traitement plus avancée]1[.

La Véracité fait référence à la cohérence et à la fiabilité des données. Les données peuvent être classées selon leur degré de qualité (données de bonne ou mauvaise qualité, ou données de qualité indéfinie si leur qualité ne peut être évaluée)]2[.

Enfin, **la valeur** fait référence à l'utilité des données dans la prise de décision et aux avantages qui peuvent découler de ces grandes quantités de données]2[. Il s'agit donc de l'une des caractéristiques les plus pertinentes du Big Data car elle reflète les avantages pour l'organisation.

3. Définition de la qualité de donnée

La qualité des données fait référence à l'état et à l'exhaustivité des données utilisées dans un système d'information, une base de données ou une application, garantissant que les décisions basées sur ces données sont fiables et pertinentes. Une bonne qualité des données permet une gestion optimale, une analyse précise et une prise de décision éclairée.

La qualité des données est une évaluation de la fiabilité et de la pertinence des informations contenues dans un ensemble de données [3]. Des données de mauvaise qualité peuvent entraîner des erreurs, des inefficacités et une augmentation des coûts pour une organisation. La gestion de la qualité des données implique de développer des processus et des méthodes pour garantir que les données sont de la plus haute qualité et de prendre des mesures pour surveiller et maintenir la qualité des données au fil du temps [4].

4. Les dimensions de la qualité des données

Étant donné que de nombreuses définitions différentes de la qualité ont été proposées dans la littérature, il existe de nombreux types de dimensions.

Le tableau 1.1 illustre certaines dimensions de la qualité des données. Du point de vue de la recherche, les dimensions de la qualité doivent être adaptées au type de données, aux caractéristiques des données et aux processus clés par lesquels les données sont manipulées.

Tableau 1 : les sept dimensions de la qualité de données [Wan98 ; JV97]

Dimension	**Définition**
La temporalité	La temporalité fait simplement référence au temps qui s'écoule entre un changement de l'état du monde réel et un changement de l'état du système d'information.
Intégrité	Lorsque les systèmes d'information sont protégés contre les biais intentionnels ou les manipulations à des fins politiques ou personnelles, les données restent intactes.
La précision	Les données sont exactes lorsque les valeurs des données stockées dans la base de données correspondent aux valeurs réelles.
Exhaustivité	Par exhaustivité, on entend que le système d'information enregistre tous les individus, services, sites ou autres unités admissibles qu'il prévoit d'évaluer. Les informations produites doivent refléter l'intégralité des individus, services, emplacements et autres

	entités, et non se limiter à une portion de la liste.
L'exactitude	Exactitude ou validité équivaut à avoir Des données précises comportent un nombre minimal d'erreurs et de préjugés. Des données manquantes, récentes ou incorrectes peuvent influencer la précision.
Confidentialité	La confidentialité signifie que les clients peuvent être assurés que leurs données seront conformes aux normes nationales et/ou internationales. Par conséquent, les informations personnelles ne seront pas divulguées de manière inappropriée.
Fiabilité	Il s'agit de la capacité de la fonction à conserver un certain niveau de performance lorsqu'elle est mise en œuvre dans des circonstances spécifiques.

4 Objectifs de la qualité des données

Sur le plan financier, préserver une excellente qualité de données aide les organisations à réduire leurs dépenses pour détecter et rectifier les erreurs dans leurs systèmes.

De plus, les entreprises peuvent prévenir les fautes opérationnelles et les arrêts dans leurs procédures commerciales, ce qui peut entraîner une hausse des coûts d'exploitation et une diminution des revenus. Une qualité de données optimale augmente aussi l'exactitude des applications analytiques, pouvant mener à des décisions commerciales plus judicieuses qui propulsent les ventes, perfectionnent les processus internes et confèrent aux entreprises un atout compétitif. L'utilisation de tableaux de bord stratégiques et d'instruments d'analyse est également facilitée par la qualité des données. En effet, quand les données d'analyse sont dignes de confiance, les utilisateurs sont davantage disposés à s'en servir plutôt qu'à se fier uniquement à leur intuition ou à leurs propres calculs pour prendre des décisions.

5 Les problèmes liés à la qualité des données

Les soucis liés à la qualité des données peuvent engendrer plusieurs causes, notamment :

- **Erreurs humaines** : Des fautes lors de la saisie, telles que des coquilles, des erreurs lors de la transcription ou des oublis.
- **Systèmes ou processus qui ne fonctionnent pas correctement** : Des systèmes mal élaborés, des instruments de recueil inappropriés ou des procédures de gestion des données peu performantes peuvent conduire à des erreurs relatives à la qualité.

- **Variabilité des sources de données :** Les informations issues de diverses sources peuvent manquer de cohérence à cause de formats divergents, de conventions distinctes ou d'informations saisies de manière disparate.
- **Mise à jour et maintenance insuffisants** : Si les données ne sont pas périodiquement actualisées ou épurées, elles risquent de devenir dépassées, erronées ou incomplètes.
- **Problèmes d'intégration** : Lors de l'extraction de données provenant de diverses bases de données ou systèmes, des erreurs peuvent survenir pendant le processus d'intégration (par exemple, des erreurs d'association ou des conflits de format).

6. Comment y remédier

Comme illustré dans la figure 2, la majorité des études traitant de la question de la qualité des données peuvent être regroupées en quatre catégories principales d'approches complémentaires [5].

Figure 2 : Vue d'ensemble des approches pour évaluer et gérer la qualité des données.

6.1 Les approches préventives

Insister sur l'ingénierie des systèmes d'information et le contrôle des processus, en recourant à des méthodes d'évaluation de la qualité des modèles conceptuels, du développement de logiciels et des procédures utilisées pour le traitement des données.

6.2 Les approches de diagnostic

Axé sur les techniques statistiques, l'examen et l'exploration de données pour identifier les irrégularités dans les données.

6.3 Les approches correctives

Elles reposent sur des méthodes d'épuration et de consolidation de données, et s'appuient sur un langage étendu de manipulation des données ainsi que sur des outils d'extraction et de transformation de données (ETL– Extraction-Transformation-Loading).

6.4 Les approches adaptatives (actives)

Habituellement mis en œuvre lors de la médiation ou de l'intégration des données : ils concernent l'ajustement des processus (interrogation de données ou opérations de nettoyage), y compris la vérification du respect des normes de qualité des données lors de l'exécution en temps réel.

7. Coût de non qualité

Il est crucial de résoudre les problèmes liés à la qualité des données pour assurer l'intégrité, la fiabilité et la pertinence des données employées dans les systèmes d'information. Des erreurs, des inefficacités et des prises de décisions incorrectes peuvent découler d'un jeu de données de mauvaise qualité. Voici des méthodes et des phases précises pour régler les soucis de qualité des données.

8. Détection / correction des problèmes de qualité des données.

Nous allons présenter ci-après les méthodes les plus populaires en pratique pour la détection et la correction des problèmes de qualité des données :

- Basé soit sur la réalité du terrain, soit sur une source de données de référence.
- Audit des données.
- Suivi des données.
- Nettoyage des données.

8.1 La validation basée sur des données de référence ou une source de vérité :

Cette méthode implique une comparaison entre les valeurs des données et leurs équivalents dans le monde réel (vérification de la véracité). La seconde approche, connue sous le nom de

fusion, implique la comparaison de deux ou plusieurs bases de données. On procède à la comparaison entre les données pertinentes de la base de données actuellement en cours d'examen et les données équivalentes présentes dans une autre base : les données identiques sont jugées exactes, tandis que les données fausses sont notifiées pour une enquête et une éventuelle rectification.

8.2 Vérification des données

Data Audit établit des programmes destinés à contrôler que les valeurs des données respectent diverses sortes d'exigences. Ces restrictions se manifestent à divers échelons de la base de données (valeurs, attributs, tuples, relations ou ensembles à chaque niveau). L'audit des données présente l'avantage d'être facile à appliquer. On peut le développer simultanément avec le schéma conceptuel de données, et divers outils d'analyse de données diagnostiques peuvent être employés. Toutefois, cela n'autorise pas une progression constante de la qualité. L'objectif de la collecte des données est d'assurer une complétude, c'est-à-dire de se conformer aux règles préétablies, Toutefois, elle ne garantit pas du tout la précision des données.

8.3 Suivi des données

Le suivi des données consiste à prélever des échantillons des enregistrements lorsqu'ils sont intégrés dans la première étape de traitement, puis à les tracer à travers chaque sous-processus jusqu'à leur insertion dans la base de données. L'élaboration de critères de correction par l'exploitation de la redondance des données est rendue possible grâce aux modifications apportées à l'enregistrement au cours de son traitement.

8.4 Nettoyage des données

Le nettoyage des données représente une série de modifications visant à standardiser le format des données et à identifier les paires d'enregistrements qui sont sans doute associées au même élément. Si des données à peu près redondantes sont détectées et que l'alignement multi-table génère des jointures approximatives entre des données dissemblables mais similaires pour permettre leur fusion, une étape de déduplication est appliquée.

9. L’importance de la qualité des données

Pour faire simple, plus la qualité des données est élevée, meilleur sera le résultat. La santé de vos données influence directement l'efficacité de nombreux cadres essentiels qui soutiennent votre structure organisationnelle. Assurer la précision de vos données vous donne la possibilité de consolider directement les instruments que vous employez pour leur gestion et leur analyse. Si votre gouvernance des données est inefficace, il est peu probable qu'elle puisse assurer une conformité totale ou appliquer correctement les contrôles d'accès si vos données sont pleines d'erreurs et d'incohérences. On observe la même tendance en matière de protection des données. Des données inexactes, comportant des erreurs et des informations absentes,

compliquent la tâche de vos équipes en charge des données pour repérer les activités suspectes ou isoler les menaces.

10. L’intégration des données

La qualité des données assure leur précision, leur intégralité, leur cohérence et leur mise à jour régulière, ce qui les rend appropriées pour l'application prévue. La notion d'intégrité des données, pour sa part, est un principe plus vaste qui couvre la précision et la sécurité des données dans leur globalité.

L'intégration des données offre aussi un accès centralisé aux informations d'une société, l'utilisateur n'a pas besoin de se connecter aux données de chaque département individuellement, mais le système d'intégration proposera une perspective consolidée et affinée des données [6]. L'un des défis à surmonter lors de l'intégration de données est la variété des schémas présents dans chaque source de données. On peut présenter les mêmes données de différentes manières. Un autre enjeu crucial est la fusion des données. Quand le système identifie plusieurs enregistrements correspondant à la même entité réelle, il est essentiel de décider lequel garder dans les conclusions finales ou comment regrouper tous les enregistrements en un unique [7].

La mise en place d'un système d'intégration de données complet est compliquée et requiert une coopération entre différents champs de recherche. La figure 3 représente les trois éléments majeurs d'un système d’intégration de données. L’alignement des schémas facilite la résolution des disparités entre eux en identifiant les éléments de schémas correspondants dans diverses sources de données, notamment les attributs équivalents.

L'entité résolution et la fusion de données servent à surmonter les disparités au niveau des instances. L'identification d'entités facilite la détection de descriptions d'entités qui se réfèrent aux mêmes entités du monde réel dans des contextes où les identifiants d'entité ne sont pas accessibles. En fin de compte, la fusion des données offre une solution au problème des attributs conflictuels et permet de regrouper les descriptions d'entités similaires en une unique description.

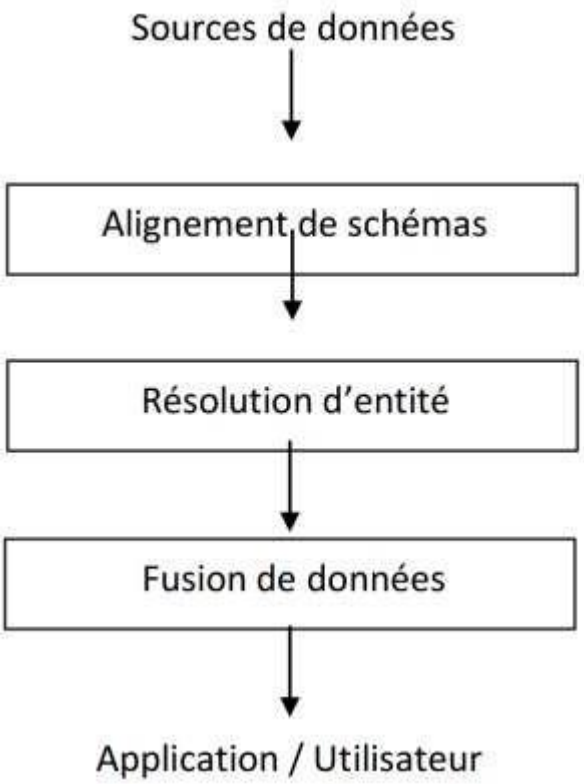

Figure 3 : Éléments clés d'un système d'intégration de données.

11. Conclusion

Ce chapitre a été dédié à l'établissement d'une revue de la littérature sur la qualité des données. Nous avons expliqué ce qu'est la qualité des données ainsi que les diverses raisons de son importance, puis nous avons détaillé les dimensions et les enjeux associés à une mauvaise qualité de données, tout en proposant des solutions pour remédier à ces problèmes. Dans le chapitre suivant nous allons introduire la résolution d'entité qui est L'un des principaux processus dans le domaine de la qualité des données.

CHAPITRE02 : Résolution d’entités

1. Introduction

À travers le globe, d'énormes volumes de données sont recueillis et conservés, avec d'autres qui s'accumulent quotidiennement. Ces informations capturent l'univers dans lequel nous évoluons ainsi que les attributs et propriétés en constante évolution des individus, des lieux et des objets qui nous entourent.

Dans ce système global de gestion des données, les institutions rassemblent de manière autonome des bases d'informations qui se superposent concernant la même entité du monde réel. Chaque organisation adopte sa propre méthode pour structurer et classer les données qu'elle possède

Les organisations et les institutions s'efforcent d'extraire des leçons significatives de ces données non raffinées. Des méthodes d'analyse sophistiquées on été développé pour identifier des motifs dans les données et en tirer leur signification et même tenter de prédire l'avenir. L'efficacité de ces algorithmes est conditionnée par la qualité et l'abondance des données qui les nourrissent. En fusionnant les données de diverses organisations, on peut fréquemment constituer un jeu de données plus fourni et plus exhaustif, qui permettra d'extraire des conclusions plus pertinentes.

Le processus d'assemblage de ces divers ensembles de données pour constituer des ensembles de données plus complets sur le monde dans lequel nous évoluons est appelé résolution d'entités. Dans ce chapitre, nous exposerons un panorama de l'art sur la résolution d'entités et les algorithmes méta heuristiques employés pour ce processus.

2. Résolution d'entités

Entity Resolution ou la résolution d'entité est le processus qui vise à détecter et à associer diverses données mentionnant une même entité dans plusieurs ensembles de données. Cela implique l'examen et la mise en parallèle de propriétés et de spécificités pour établir si une entité est distincte, ainsi que pour vérifier si diverses sources font référence à la même entité. Les buts sont divers]8[:

- Éliminer les répétitions et les données superflues. Cela pourrait potentiellement diminuer les coûts en rationalisant les opérations de manière globale ; l'information étant centralisée à un seul emplacement, cela permet de minimiser les interventions manuelles, facilitant ainsi la gestion des modifications dans les données.
- Optimiser la qualité générale des données pour appuyer l'analyse des données et le processus de décision.
- Établir une vision cohérente des données pour simplifier leur intégration et consolidation, tout en répondant potentiellement à des normes de régulation.

3. Techniques de résolution d'entités

Les méthodes de résolution d'entités comprennent différentes stratégies destinées à identifier et à connecter avec exactitude les enregistrements qui se réfèrent aux mêmes entités du monde réel. On peut catégoriser ces techniques en diverses méthodes, chacune offrant des bénéfices et des applications spécifiques]9[.

3.1. Techniques d'appariement déterministe

3.1.1 Exact Matching (Correspondance exacte)

C'est le cas où les enregistrements sont considérés comme identiques s'ils sont exactement identiques sur certains champs clés (par exemple, le nom, l'adresse ou le numéro d'identification).

Phonetic Matching(Correspondance phonétique)

3.1.2 Phonetic Matching (Correspondance phonétique)

Correspondance des enregistrements basée sur la similarité phonétique des noms ou d'autres informations textuelles. Des algorithmes standards comprennent Soundex ou Metaphone, qui transforment les mots selon leur énonciation]10[.

3.1.3 Token-based Matching (Correspondance basée sur des jetons)

C'est une question de fractionner un texte en unités (qu'il s'agisse de mots ou de locutions) et de confronter ces unités d'un enregistrement à l'autre. On utilise fréquemment les mesures de similarité de Jaccard ou de similarité cosinus]10[.

3.2 Techniques d'appariement probabiliste

3.2.1 Modèle de Fellegi et Sunter

Assignez des chances aux correspondances possibles en prenant en compte l'accord et le désaccord des valeurs d'attributs. Cette approche permet de gérer l'incertitude liée à la correspondance et est couramment employée dans l'appariement probabiliste des enregistrements]11[.

3.2.2 La similarité de Jaccard

Évalue la similarité entre des groupes d'éléments. Lorsqu'il est appliqué à des attributs définissant des actifs (comme les mots contenus dans un document), il permet de repérer les enregistrements correspondants sur la base du recoupement des éléments.

3.2.3 Blocage et fenêtrage

Segmentez le jeu de données en sections ou fenêtres basées sur des critères précis, diminuant ainsi la quantité de paires d'enregistrements à analyser. Cela peut grandement augmenter l'efficacité des calculs.

3.3 Techniques basées sur l'apprentissage automatique

3.3.1 Apprentissage supervisé

On entraîne des modèles à partir de données étiquetées pour déterminer si deux enregistrements correspondent à la même entité. Les algorithmes standards incluent les machines à vecteurs de support (SVM) et les forêts aléatoires]12[.

3.3.2 Apprentissage non supervisé

Utilise des algorithmes de clustering pour regrouper des enregistrements similaires. Le clustering hiérarchique, la méthode K-means et DBSCAN sont des exemples d'approches non supervisées]12[.

4 Défis majeurs dans la résolution d'entités

4.1 Insuffisance de noms distincts

Pour commencer, le premier défi est de reconnaître l'unicité des noms ou des étiquettes. L'usage récurrent du même nom pour diverses entités dans le monde réel pose clairement un problème pour distinguer les différentes identités.

4.2 Incohérences dans les conventions de nommage

Les noms sont consignés selon diverses méthodes et formats de données. Il arrive parfois que les noms soient présentés dans leur forme complète, mais il est fréquent que des abréviations soient utilisées ou que certaines portions moins importantes du nom soient négligées]13[. Sur le plan mondial, les méthodes d'attribution des noms présentent une diversité remarquable. Les prénoms peuvent se trouver au début ou à la fin d'un nom, et les noms de famille peuvent être présents ou absents.Le nom de famille peut aussi différer selon le genre et la situation matrimoniale de l'individu. Les noms peuvent être exprimés dans divers alphabets ou systèmes de caractères, ou traduits de manière différente d'une langue à l'autre.

4.3 Incohérences présentes dans l'enregistrement des données

Un nom peut être entendu uniquement au cours d'une conversation, sans possibilité de vérifier l'orthographe exacte ou peut être mal écrit à cause de la précipitation. Il arrive souvent que les noms ou les labels soient mal saisis lors de la saisie manuelle ou omis par inadvertance. La saisie de données à l'échelle mondiale peut provoquer des discordances de translittération entre différentes écritures, ou des fautes de transcription lors de l'entrée orale.

5 Les étapes du processus de résolution des entités

Le processus de résolution d'entités (ER) peut être défini par un processus en six étapes tel qu'il est illustré dans la figure 4]14[.

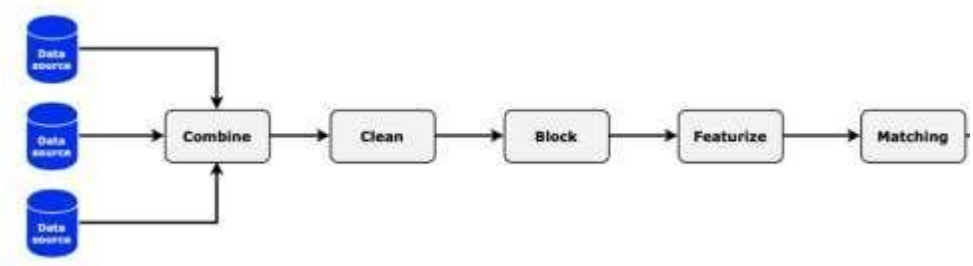

Figure 4 : Étapes de résolution d'entité

5.1 Combiner (Combine)

La première phase implique la collecte de toutes les données et leur intégration dans un schéma uniforme. Dans l'idéal, il devrait y avoir une base de données universelle regroupant toutes les données sous un même schéma. Toutefois, dans la réalité, il est courant que divers ensembles de données soient conservés dans différents systèmes et formats.

5.2 Nettoyage et normalisation (Clean)

La seconde phase implique de garantir que les données sont bien nettoyées et présentées dans un format homogène. Il se peut que chaque source de données exige des procédures distinctes pour parvenir à un modèle uniforme, cependant, toutes les données doivent subir les mêmes processus de nettoyage ; c'est pourquoi nous le réalisons après la fusion. L'une des premières étapes consiste à s'assurer que les données utilisent le même codage. Notre objectif est non seulement de garantir l'absence de caractères inhabituels, mais également de veiller à ce que les langues respectent une traduction standardisée. Une méthode fréquemment utilisée est de tout transformer en ASCII. Par conséquent, il est possible que les valeurs Müller et Muller finissent par être identique. Et Mueller, que peut-on dire ? En outre, il est recommandé de prévenir les problèmes de casse en optant soit pour des lettres majuscules, soit pour des lettres minuscules.

5.3 Blocage (Blocking)

L'étape du blocage consiste à minimiser la quantité de comparaisons entre les enregistrements en les classant selon des critères partagés. Ceci contribue à rendre le processus de résolution plus efficient et plus prompt. Par exemple :

• **Blocage par caractéristiques partagées** : À l'instar de la consolidation d'enregistrements possédant des codes postaux semblables ou des appellations de villes identiques.

• **Indexation** : L'usage d'index pour faciliter la recherche de registres similaires.

• **Segmentation** : Fractionner les données en sous-groupes (blocs) afin de minimiser l'espace

de recherche. Cela diminue le nombre de couples d'enregistrements à analyser, puisqu'on ne met en comparaison que les enregistrements se trouvant dans les mêmes blocs.

5.4 Featurize (Calcul de la Similarité)

Après le verrouillage des données, il est nécessaire de calculer la similarité entre les couples d'enregistrements. Ceci permet d'identifier si deux enregistrements correspondent à la même entité réelle]15[.

• **Indicateurs de similarité :** Application de diverses métriques pour juger la ressemblance entre les enregistrements. Ces actions peuvent englober :

- Distance de Levenshtein (également connue sous le nom de distance d'édition).
- Mesure de similarité de Jaccard.
- Similitude cosinus.
- Jaro-Winkler.
- Correspondance exacte.

5.5 Matching

Après avoir comparé les paires et déterminé leur similarité, une décision doit être prise pour chaque paire d'enregistrements :

• **Correspondance** : Les enregistrements sont jugés se rapporter à la même entité réelle.

• **Non- Correspondance** : Les enregistrements sont jugés comme représentant des entités distinctes.

• **Correspondance possible ou incertaine** : Lorsque le degré de similarité est élevé mais que des doutes subsistent, une révision manuelle ou l'emploi d'un modèle probabiliste pourraient être requis.

5.6 Clustering

Après l'achèvement du processus de linkage, nous pouvons noter la création de regroupements sur le graphique. Chaque cluster est une entité distincte dérivée de diverses sources de données. Un identifiant distinct est attribué à chaque regroupement, et les attributs des enregistrements issus de diverses sources sont fusionnés pour la création d'un enregistrement de référence. Si un conflit survient concernant la valeur d'un attribut, nous pouvons recourir à une hiérarchie de préférences des sources de données pour décider de la source à partir de laquelle obtenir cette information. Il est bénéfique de maintenir l'association entre cet identifiant de référence et les identifiants des enregistrements issus de diverses sources de données.

5.7 Mesurer la performance

Les méthodes statistiques peuvent nous guider dans la manière d'évaluer et de fusionner tous les indices révélés par la comparaison des caractéristiques individuelles.

En examinant deux enregistrements, on peut envisager quatre situations différentes. Le tableau 2 énumère les diverses combinaisons de décisions concernant la correspondance et la véracité sur le terrain.

Tableau 2 : Classification des correspondances

Tu décides	Vérité de terrain	Instance de
Correspondance	Correspondance	Vrai positif (TP)
Correspondance	Ne correspond pas	Faux positif (FP)
Ne correspond pas	Correspondance	Faux négatif (FN)
Ne correspond pas	Ne correspond pas	Vrai négatif (TN)

6 Utilisation des algorithmes méta-heuristiques pour la résolution d'entité

Recourir à des algorithmes méta-heuristiques pour résoudre des entités représente une démarche novatrice susceptible d'accroître l'efficacité et la précision de ce processus. Ceci est particulièrement vrai dans des situations où les techniques traditionnelles basées sur des règles ou des comparaisons exactes ne suffisent pas. Les techniques méta-heuristiques constituent des approches heuristiques qui visent à dénicher des solutions de qualité supérieure, bien qu'approximatives, pour résoudre des problèmes complexes dans des espaces de recherche étendus et souvent peu définis.

7 Métaheuristique

L'heuristique est une méthode d'optimisation qui a pour but de trouver une « bonne » solution à un problème en un temps raisonnable, mais sans garantie sur la validité ou l'optimalité de la solution ainsi fournie [BS+12]. Parmi les heuristiques certaines sont adaptables à un grand nombre de problèmes différents sans changements majeurs dans l'algorithme, on parle alors de méta-heuristiques. Le qualificatif "méta » fait référence à une combinaison de un ou plusieurs heuristiques. Avant de pouvoir être appliquée à la résolution d'un problème particulier, quelques transformations (mineures en général) sont nécessaires, Pour lesquels elles s'adapteront avec plus ou moins de facilité à chaque problème[BS+12]. Les Métaheuristiques contient un composant d'exploration (aussi appelée « diversification ») qui permet de rechercher de nouvelles solutions dans l'espace de recherche, et un composant d'exploitation (également appelée "intensification") qui utilise les résultats obtenus lors de la phase d'exploration, afin de sélectionner le sous-espace de recherche le plus prometteur, et de

"plonger" vers l'optimum local le plus proche [MBF11]. Les métaheuristiques sont apparues dans les années 1980 et forment une famille d'algorithmes d'optimisation visant à résoudre des problèmes d'optimisation difficile. Durant les vingt dernières années, les métaheuristiques ont reçu un intérêt grandissant et ont montré leur efficacité dans de vastes domaines d'application en résolvant de nombreux problèmes d'optimisation, ce sont généralement des problèmes des données incomplètes, ou capacité de calcul limitée. Plusieurs d'entre elles sont souvent inspirées par des systèmes naturels dans de nombreux domaines tels que : la biologie (algorithmes évolutionnaires et génétiques), l'éthologie (algorithmes de colonies de fourmis) et aussi la physique (recuit simulé)[Tal09].

8 Pourquoi Utiliser des Méta-heuristiques pour la Résolution d'Entité ?

La résolution d'entité est un problème complexe et souvent mal défini, où des erreurs peuvent se produire en raison de la diversité des formats de données, de la variation dans la saisie des informations ou du bruit dans les données. Les algorithmes méta-heuristiques peuvent offrir plusieurs avantages dans ce contexte :

• **Exploration de vastes domaines de recherche** : Les algorithmes méta-heuristiques sont capables d'explorer efficacement une multitude de solutions potentielles pour l'appariement d'enregistrements, surtout en présence d'une grande variabilité dans les données.

• **Évasion des minima locaux** : les méta-heuristiques, comme le recuit simulé ou les algorithmes génétiques, permettent d'éviter de se retrouver dans des solutions sous-optimales locales.

• **Adaptabilité** : Ces algorithmes sont capables de s'adapter à des données complexes et mal structurées. Ils peuvent fonctionner dans des environnements dynamiques et évoluer avec de nouvelles informations ou différentes normes de comparaison.

• **Optimisation multi-objectifs** : les algorithmes métaheuristiques permettent une optimisation multi-objectifs si la résolution d'entité nécessite la coordination de plusieurs critères (par exemple, précision, rappel, réduction des faux positifs).

• **Robustesse aux erreurs et aux inexactitudes** : ces algorithmes peuvent gérer des données bruitées ou incomplètes, ce qui est courant lors de la résolution d'entités.

9 Avantages et limites de l'utilisation des métaheuristiques

Avantage :

• **Efficacité** : Ils permettent de trouver des solutions de haute qualité en explorant efficacement de grands espaces de solutions.

• **Flexibilité** : ils s'adaptent à une variété de problèmes de résolution d'entités.

• **Robustesse** : ces algorithmes peuvent gérer des données bruitées et incomplètes.

Limite :

• **Temps de calcul** : les algorithmes métaheuristiques peuvent être plus lents à converger vers une solution optimale que les méthodes exactes.

• **Complexité de mise en œuvre** : les métaheuristiques peuvent être plus difficiles à mettre en œuvre et nécessitent souvent un réglage fin des paramètres.

• **Résultats non garantis** : Les solutions trouvées ne sont pas nécessairement optimales et peuvent dépendre fortement des paramètres de l'algorithme.

10. Grey Wolf Optimizer (GWO)

10.1. Définition

L'algorithme d'optimisation métaheuristique Grey Wolf Optimizer (GWO) a été suggéré par Mirjalili et ses collaborateurs en 2014.

Il reproduit le comportement social des loups gris (hiérarchie et tactiques de chasse) afin d'identifier les solutions les plus efficaces dans un domaine d'exploration]16[.

Les loups gris possèdent :

- Une hiérarchie sociale rigoureuse :
  Alpha (α) : Le leader (la solution optimale),
  Beta (β) : Le second en commandement (la deuxième meilleure option),
  Delta (δ) : En troisième position,
  Omega (ω) : Les autres participants.

- Une approche de chasse en trois phases :
  1. Poursuite de la proie,
  2. Encerclage de la proie,
  3. Assaut final.

10.2. Mode de fonctionnement

Deux mécanismes principaux sous-tendent le GWO :
Exploration (chercher de manière extensive) :
• Les loups se déplacent pour explorer diverses régions de l'espace de recherche.
• Lorsque le paramètre A est élevé.
Exploitation (focalisation sur une zone) :
• Les loups se dirigent vers la proie optimale (solution).
• Lorsque le paramètre A est faible.

En mathématique

• Encerclement :

Nous représentons la distance entre le prédateur et sa proie :

$$D = |C * X_p(t) - X(t)|$$

$$X(t+1) = X_p(t) - A * D$$

Xp(t) : Position de la proie (solution idéale),

X(t) : Position actuelle du loup,

A et C : Coefficients de régulation.

Attaque :

La valeur de A se diminue **linéairement** de 2 à 0 au fil des itérations :

$$A = 2 * a * r1 - a$$

$$C = 2 * r2$$

Où r1 et r2 sont des nombres aléatoires [0,1].

Mise à jour de la position :

Chaque loup ajuste sa position en se basant sur **Alpha**, **Beta** et **Delta** :

$$X1 = X - alpha - A1 * D - alpha$$

$$X2 = X - beta - A2 * D - beta$$

$$X3 = X - delta - A3 * D - delta$$

Nouvelle position $X = (X1 + X2 + X3)/3$

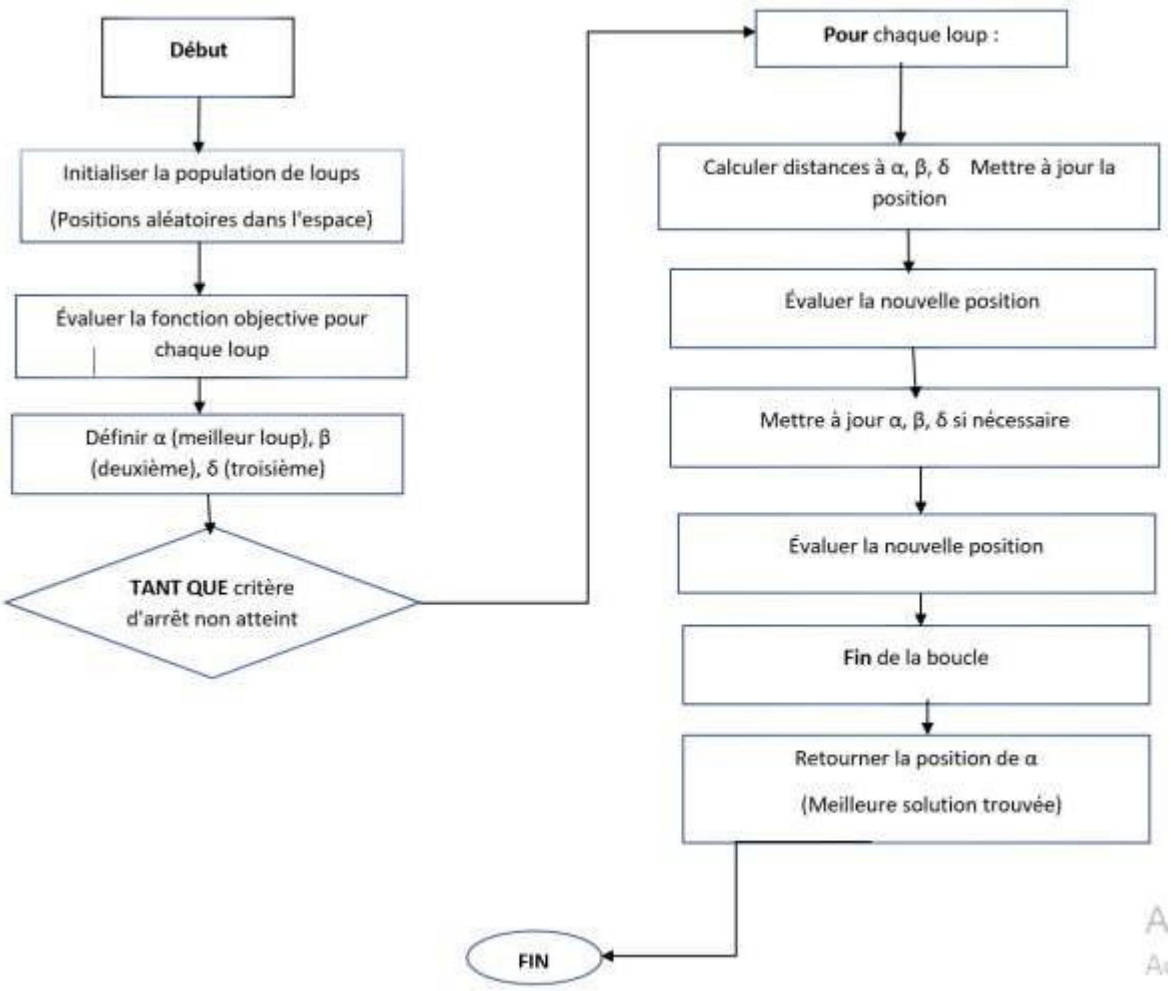

***Figure 5** : Organigramme de comportement de GWO*

11. Bald Eagle Search (BES)

11.1. Définition

L'algorithme d'optimisation métaheuristique Bald Eagle Search (BES) s'inspire du mode de chasse de l'aigle chauve. Askarzadeh l'a présenté en 2020. Idée centrale]17[:

L'aigle :

- Parcourt de vastes territoires pour localiser une zone idéale (exploration),
- Choisit la meilleure zone identifiée (sélection),
- Exploite cette zone pour attraper sa proie (exploitation).

Inspiration venant de la nature

Dans son milieu naturel, l'aigle chauve manifeste trois types de comportements :

- Exploration en hauteur : Rechercher divers lieux prometteurs.

- Sélectionner une zone fertile : L'endroit où il suppose que la proie se trouve.
- Immersion pour la capture : Se concentrer et se diriger rapidement vers l'objectif.

BES répète ces étapes dans la quête de solutions optimales au sein d'un espace spécifique.

11.2. Mode de fonctionnement

L'algorithme BES se divise en trois étapes majeures :

Étape 1 — Exploration
• Les aigles parcourent et explorent différentes régions de l'espace d'étude.
• Ils produisent des emplacements aléatoires.
Objectif : Explorer largement sans se limiter trop rapidement.

Étape 2 - Sélection
• Chaque aigle évalue les régions qu'il a explorées.
• Il sélectionne les meilleures options en fonction de la valeur de la fonction objective.
Objectif : Axer les recherches sur des régions prometteuses. Phase 3 - Mise en œuvre (Attaque)

Étape 3 - Exploitation

• L'aigle se dirige rapidement vers la zone optimale.
• Il effectue une recherche locale pour préciser l'emplacement.
Objectif : Affiner et dénicher la solution précise.

En mathématique

Exploration

$$X_{new} = X_{best} + rand * (X_{rand} - X_{best})$$

Où X_{rand} est une position aléatoire.

Sélection :

Choisir la position avec la **meilleure valeur** de fonction objective.

Exploitation :

$$X_{new} = X_{current} + a * (X_{best} - X_{current}) + b * (X_{rand} - X_{current})$$

Où a et b sont des coefficients aléatoires entre 0 et 1.

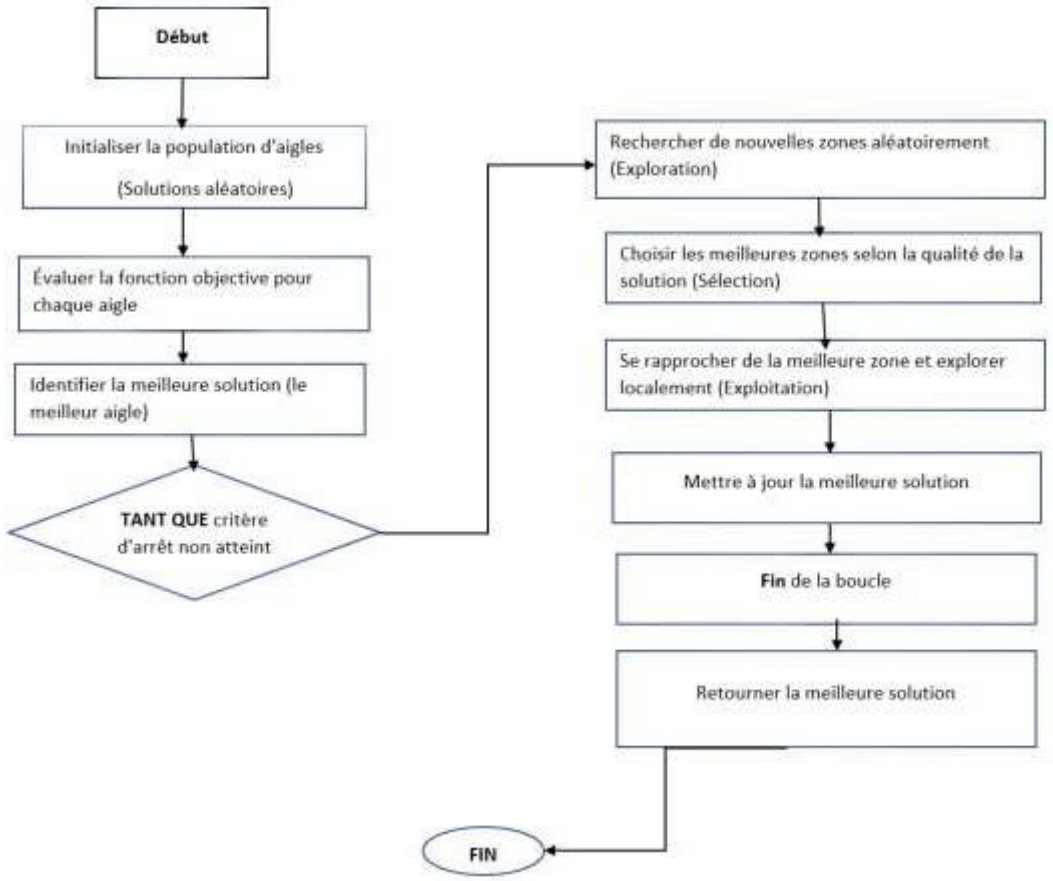

Figure 6 : Organigramme de comportement de BES

12. Woodpecker Optimization Algorithm (WOA)

12.1. Définition

(WOA) est une métaheuristique basée sur le comportement alimentaire des pics-verts. On l'a suggéré récemment (en 2021) pour apporter des solutions à des problèmes d'optimisation complexes]18[.

Dans la nature, un pivert :

- Recherche des troncs d'arbres,
- Juge si le creusage est opportun,
- Fore avec vigueur pour dénicher la nourriture dissimulée (insectes, larves).

WOA reproduit ce processus en trois phases majeures lors de la quête de solutions optimales.

12.2. Mode de fonctionnement

L'algorithme WOA se décompose en trois étapes majeures :

Étape 1 — Exploration (Recherche)

• Le pic se déplace de branche en branche (dans des zones aléatoires).
• Il recherche des objectifs prometteurs.
• Élaboré par la création de nouvelles solutions au hasard.
Étape 2 - Évaluation (Choix)
• Le sommet juge la qualité de chaque arbre.
• Il détermine si cela vaut la peine de poursuivre ou d'explorer ailleurs.

Étape 3 - Forage (Exploitation)
• Une fois une cible appropriée identifiée, il procède à un forage en profondeur.
• Cela imite l'amélioration locale autour de la solution optimale identifiée.

En mathématique

Recherche :

$$X_{new} = X_{current} + random_{step} * (X_{best} - X_{current})$$

Exploration des alentours.

Sélection :

Sélection de la cible optimale basée sur sa qualité.

Perçage (Exploitation) :

$$X_{new} = X_{current} + attack_{strength} * (X_{target} - X_{current})$$

Où :

$attack_{strength}$ est un coefficient simulant la force de perçage,

X_{target} est une cible très prometteuse.

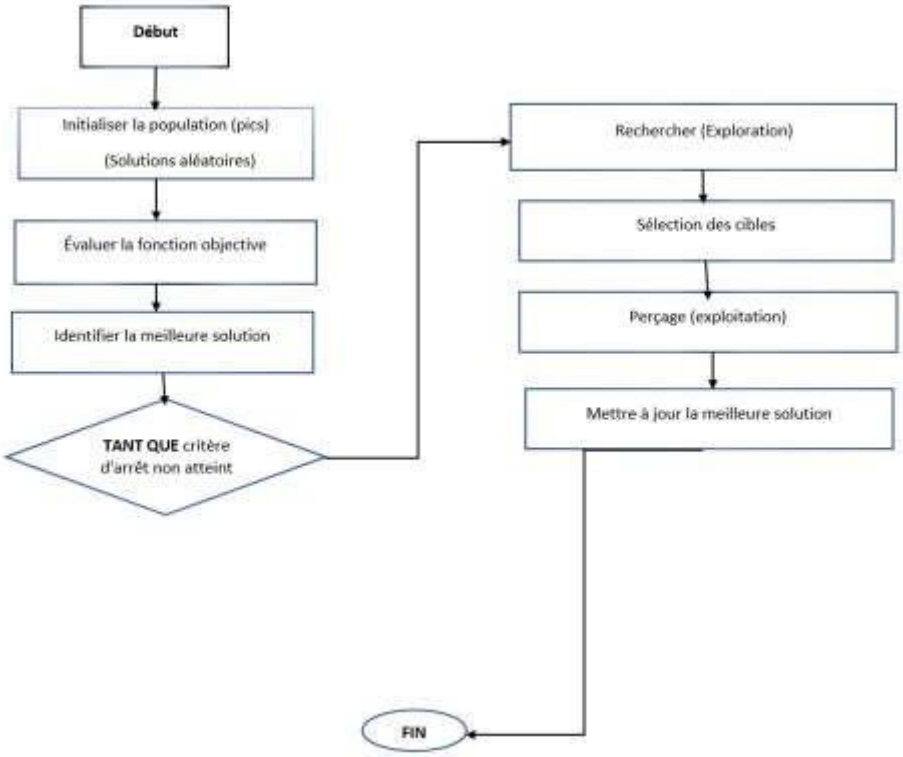

Figure 7 : Organigramme de comportement de WPO

13. Conclusion

La résolution d'entité est un composant essentiel dans la gestion des données à travers différents secteurs. En facilitant la connexion d'informations issues de diverses sources tout en assurant leur singularité et leur cohérence, elle intervient de manière significative dans l'optimisation de la qualité des données, l'intégration des systèmes ainsi que dans le processus de prise de décisions informées. Néanmoins, ce processus demeure compliqué et exige le recours à des outils et méthodes sophistiqués, surtout dans des situations impliquant d'importants volumes de données et une variété de formats.

Recourir aux algorithmes méta-heuristiques pour résoudre les problèmes d'entité constitue une méthode efficace pour gérer les cas complexes de correspondance des enregistrements. Ces techniques aident à améliorer l'exactitude, la souplesse et la productivité du processus de résolution, notamment dans des environnements caractérisés par une grande diversité et un large volume de données.

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 sur la résolution d'entité et on a fermé le chapitre par les algorithmes méta-heuristiques pour traiter la résolution d'entités dans le Big Data. Le chapitre suivant sera dédié à la conception et l'implémentation de notre application.

CHAPITRE03 : CONCEPTION ET IMPLEMENTATION

1. Introduction

Dans ce chapitre, nous détaillons le contexte et les technologies mises en œuvre pour le développement de notre application, avant de mettre en lumière notre apport dans le secteur de la résolution d'entités. Pour chaque défi traité dans le précédent chapitre, une solution est suggérée. Nous introduisons une nouvelle méthode de résolution basée sur les métaheuristiques. Trois algorithmes distincts seront employés dans la contribution à la résolution d'entités : BES pour Bald Eagle Search, WP pour WoodPecker et GWO pour Grey Wolf Optimization. Par la suite, l'apport de ces trois derniers dans le domaine de résolution d'entités sera examiné pour démontrer leur performance et leur efficacité. On discutera des résultats obtenus. Pour finir, nous vous montrons quelques captures d'écran.

2. Description de méthode utilisée

Dans ce qui suit nous donneront une explication détaillées de chaque étape de contribution des trois algorithmes (BES,GWO,WPO)dans le processus de résolution d'entités proposé on commençant par le chargement de l'ensemble de données jusqu'à l'étape d'évaluation.

2.1. Utilisation de l'algorithme BES dans la résolution d'entité

L'algorithme de résolution d'entité à base de BES est décrit ci-dessous

Données d'entrée : collection d'entrées « Restaurants »
Début
Initialiser la Population P ;
Produire des clés de blocage de manière aléatoire ;
Pour chaque solution **Faire**
blocage des données en fonction des clés suggérées ;
Regrouper les records dont les valeurs sont similaires basées sur les clés ;
Fin Pour ;
Pour chaque Bloc généré **Faire**
Utiliser le « clustering K-Modes » pour diviser le bloc en Clustres ;
Calculer « rappel, Précision,Score F1» baséée sur Match/no Match ;
Trouver la solution optimale Locale ;
Fin Pour ;
Tant que NbIter est inférieur à IterMax **Faire**
Ajustez les clés de blocage suggérées ;
Conservez les solutions les plus performantes ;
Affiner les solutions prometteuses ;
Fin Tant que ;
Conserver la meilleure combinaison de clés produite par BES.
Regroupement définitif avec **K-Modes ;
Fin ;

2.2. Utilisation de l'algorithme GWO dans la résolution d'entités

L'algorithme de résolution d'entité à base de GWO est décrit ci-dessous

Données d'entrée : collection d'entrées « Restaurants »
Début
Initialiser la Population (meute des loups) ;
Produire des clés de blocage de manière aléatoire ;
Pour chaque Loup **Faire**
blocage des données en fonction des clés suggérées ;
Regrouper les records dont les valeurs sont similaires basées sur les clés ;
Fin Pour ;
Pour chaque Bloc généré **Faire**
Utiliser le « clustering K-Modes » pour diviser le bloc en Clustres ;
Calculer « rappel, Précision,Score F1» baséée sur Match/no Match ;
Fin Pour ;
Identifier les trois meilleurs loups (Alpha , Beta , Delta) ;
Tant que NbIter est inférieur à IterMax **Faire**
Mettre à jour les positions des loups;
Réévaluer les nouvelles solutions.;
Mettre à jour les nouveaux Alpha, Beta, Delta;
Fin Tant que ;
Conserver la meilleure combinaison de clés produite par GWO.
Regroupement définitif avec **K-Modes ;
Fin ;

2.3. Utilisation de l'algorithme WPO dans la résolution d'entités

Données d'entrée : collection d'entrées « Restaurants »
Début
Initialiser la Population (pics-verts) ;
Produire des clés de blocage de manière aléatoire ;
Pour chaque pic-vert **Faire**
blocage des données en fonction des clés suggérées ;
Regrouper les records dont les valeurs sont similaires basées sur les clés ;
Appliquer **K-Modes clustering** sur chaque bloc obtenu ;
Evaluer la solution ;
Fin Pour ;
Identifier les meilleurs individus ;
Tant que NbIter est inférieur à IterMax **Faire**
Les pics-verts cherchent plusieurs nouveaux arbres;
Trouve des solutions prometteuses.;
Réévaluer les nouvelles solutions;
Mettre à jour les meilleurs arbres trouvés ;
Fin Tant que ;
Conserver la meilleure combinaison de clés produite par WPO ;
Regroupement définitif avec **K-Modes ;
Fin ;

3. Environnement de travail

Dans cette section, nous présenterons l'environnement logiciel de notre travail :

- Système d'exploitation : windows10
- Langage de programmation : Java

4. Outils de développement

4.1. NetBeans IDE

NetBeans est un environnement de développement intégré (IDE) pour Java, placé en open source par Sun en juin 2000 sous licence CDDL (Common Development and Distribution License). En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, XML et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages web).

NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X et Open VMS.

NetBeans est lui-même développé en Java, ce qui peut le rendre assez lent et gourmand en ressources mémoires.

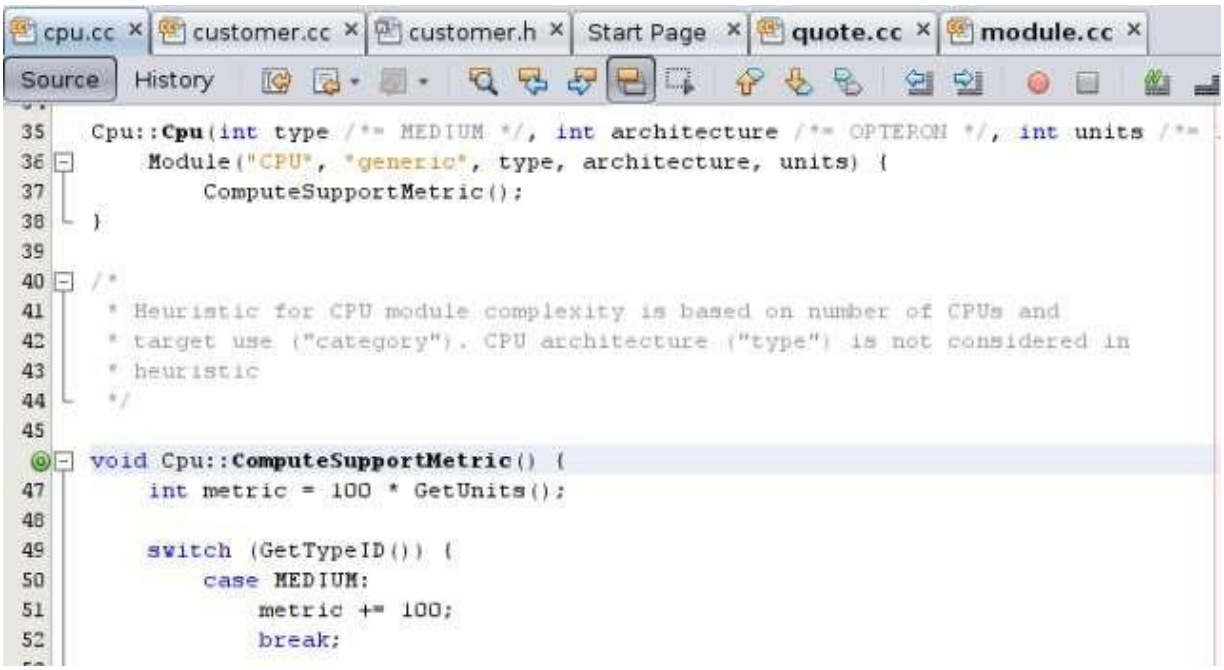

Figure 8 : fenêtre de programmation Sur Netbeans

4.2. Scene Builder

Scene Builder est un outil interactif de conception d'interface graphique pour JavaFX. Il permet de créer des interfaces utilisateurs rapidement et sans avoir besoin de coder ; il en résulte des

fichiers au format FXML qui sont ensuite chargés par le programme pour afficher une interface graphique à l'utilisateur.

Développé initialement par Oracle et sous le nom JavaFX Scene Builder, son code source a été publié en open source à partir de sa version 2.0.

Depuis, le logiciel est principalement développé et soutenu par la société Gluon.

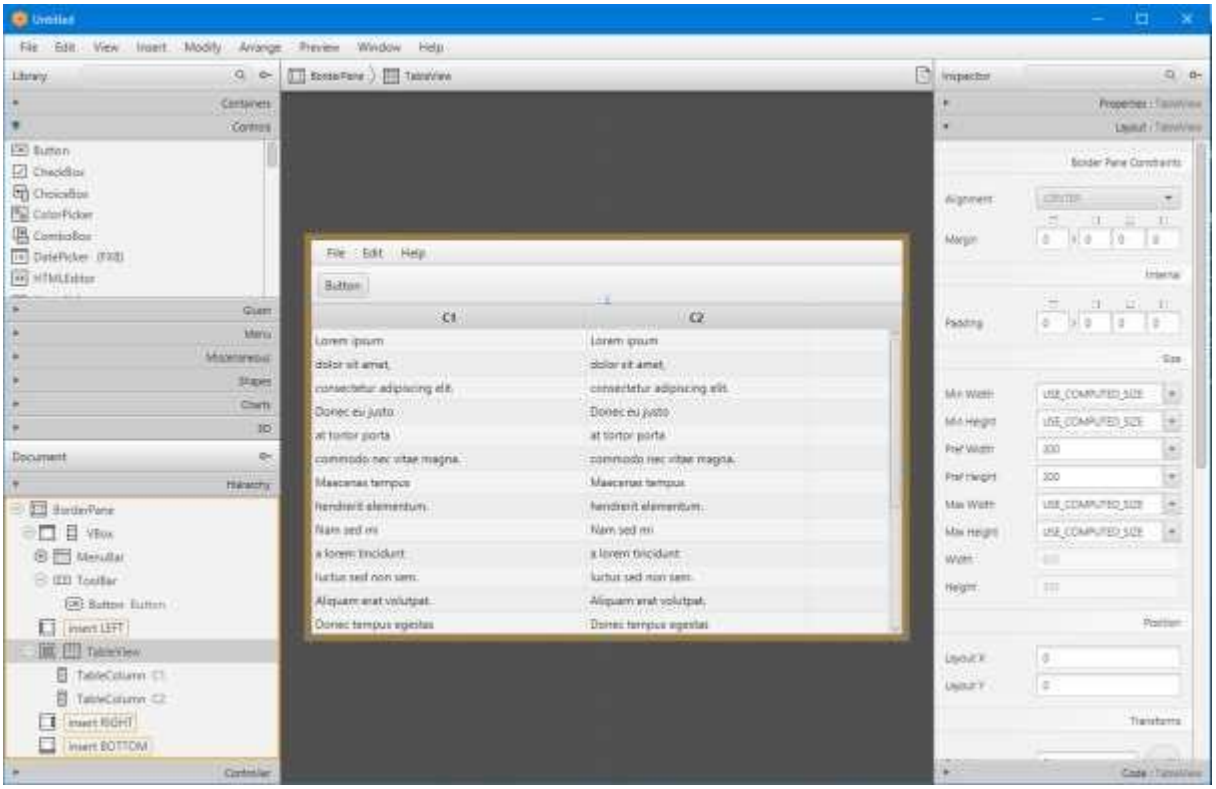

Figure 9 : Utilisation Java FX Scene Builder

4.3. JAVA

Java est un langage de programmation largement utilisé pour coder des applications web. Il a été fréquemment choisi parmi les développeurs depuis plus de deux décennies, des millions d'applications Java étant utilisées aujourd'hui. Java est un langage multiplateforme, orienté objet et centré sur le réseau, qui peut être utilisé comme une plateforme à part entière. Il s'agit d'un langage de programmation rapide, sécurisé et fiable qui permet de tout coder, des applications mobiles aux logiciels d'entreprise en passant par les applications de big data et les technologies côté serveur.

4.4. JavaFX

JavaFX est une technologie créée par Sun Microsystems qui appartient désormais à Oracle.Avec l'apparition de Java 8 en mars 2014, JavaFX devient la bibliothèque de création d'interface graphique officielle du langage Java, pour toutes les sortes d'application (applications mobiles, applications sur poste de travail, applications Web), le développement de son prédécesseur Swing étant abandonné (sauf pour les corrections de bogues).

JavaFX contient des outils très divers, notamment pour les médias audio et vidéo, le graphisme 2D et 3D, la programmation Web, la programmation multi-fils etc. Le SDK de JavaFX étant désormais intégré au JDK standard Java SE,

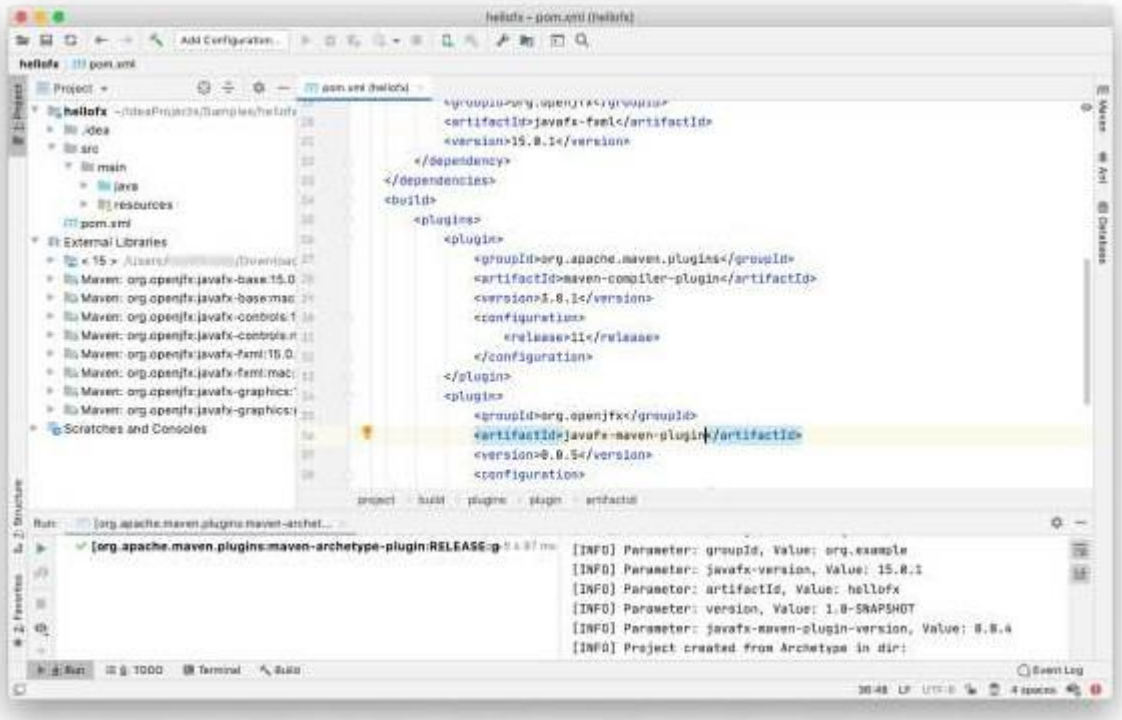

Figure 10 : projet Java FX

5. Implémentation

Présentation de L'application

L'application de résolution d'entitées qu'on a développé en se basent sur BES,GWO et WPO lors de la conception est la suivante :

Interface d'accueil

La figure 11 montre La page d'accueil de notre application qui nous permet de charger l'ensemble de données utilisé dans notre système ainsi que l'application des trois algorithmes utilisés et un accès direct à l'ensemble d'évaluation.

Figure 11 : Page d'accueil de l'application

Chargement de Dataset

Le dataset utilisé dans notre application est « Restaurants ».Une trame de données avec 5 attributs : Nom, Adresse, Ville, Télé phone 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 standard. La figure 12 montre L'ensemble de données Restaurant chargé dans un fichier.arf

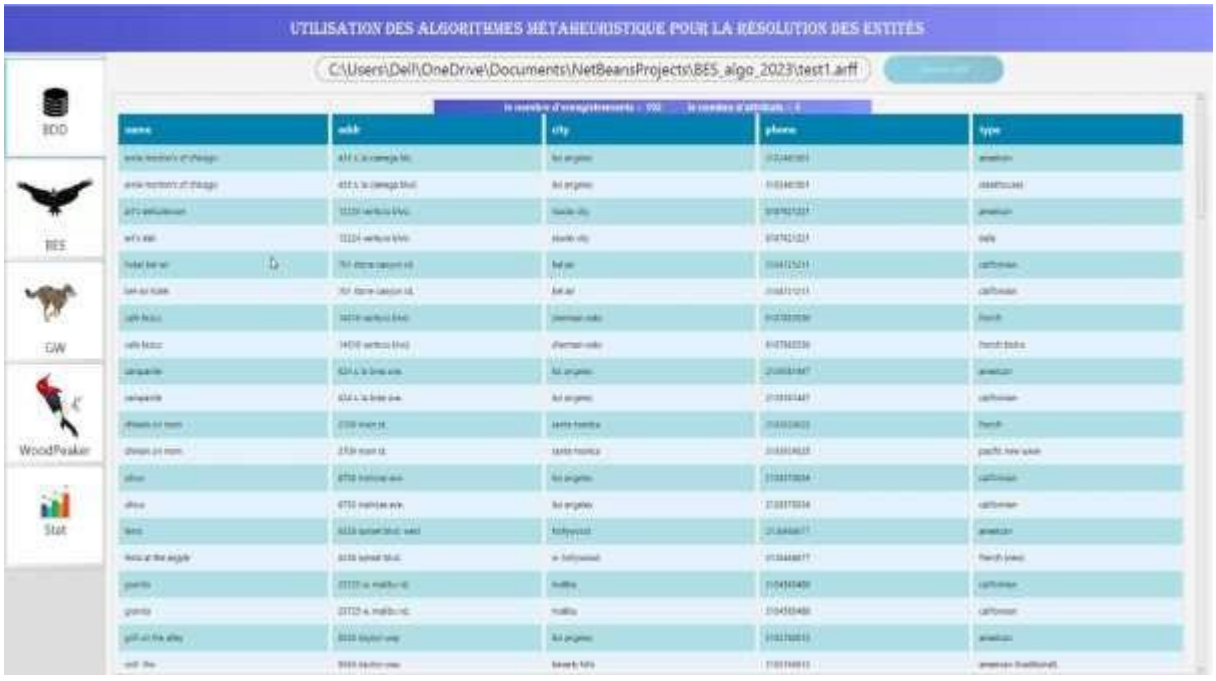

Figure 12 : Sélection de fichier arf

Génération des clés de blocage

La figure 13 montre la génération des clés de blocage aléatoirement par les trois algorithmes utilisée dans notre application.

Figure 13 : Génération aléatoire des clés de blocage par les trois algorithmes

Création des Blocs

La figure 14 montre la création des blocs avec l'algorithme BES

Figure 14 : Création des Blocs

Résultat de Matching

La figure 15 montre les résultats obtenus 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 15 : Résultat de Matching

6. Résultats et évaluations

La matrice de confusion

La matrice de confusion est une méthode d'évaluation des performances qui permet de visualiser et d'analyser les résultats d'un modèle de classification. Elle affiche le nombre de **vrais positifs (VP)**, **vrais négatifs (VN)**, **faux positifs (FP)** et **faux négatifs (FN).**

Dans le domaine de résolution d'entitées, quatre critères majeurs sont employés pour évaluer l'efficacité d'une procédure RL.

Accuracy: (Tous corrects / Tous) = Vrai positif + vrai négatif / Vrai positif + vrai négatif + Faux Positif + Faux Nigatif.

Precision :(Vrais positifs / positifs prévus) = Vrai positif / (Vrai positif + Faux positif).

Recall:(Vrais positifs / tous les positifs réels) = Vrai positif / Vrai positif + faux négatif.

Score F1: score F1 = (2 * Precision * Recall) / (Precision + Recall).

Où :

- Vrai positif (TP) : résultat pour lequel l'approche prédit correctement l'appartenance d'un pixel à la partie tumeur.
- Vrai négatif (TN) : résultat pour lequel l'approche prédit correctement le non appartenance d'un pixel à la partie tumeur.
- Un faux positif (FP) est un résultat où le modèle prédit incorrectement l'appartenance d'un pixel.
- Un faux négatif (FN) est un résultat où le modèle prédit incorrectement le non appartenance. [54]

Figure 16 : Evaluation de l'algorithme BES

Figure 17 : Evaluation de l'algorithme GWO

Figure 18 : Evaluation de l'algorithme WPO

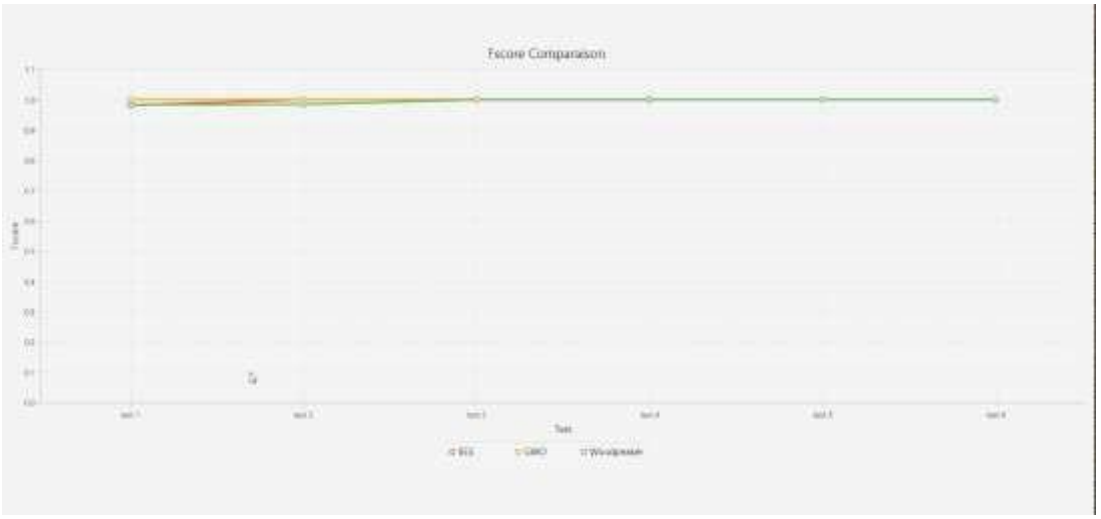

Figure 19 : Résultat de la métrique Fscore

On peut observer les performances des trois algorithmes intégrés dans notre application, à savoir BES, GWO et WPO, dans les graphiques ci-dessus. Ces figures se réfèrent à des indicateurs d'évaluation tels que l'exactitude, la précision, le rappel et le Fscore. Voici un examen des conclusions :

- L'exactitude mesure la capacité du modèle à identifier de façon précise les entités similaires.
- La précision évalue le pourcentage de prédictions positives qui s'avèrent correctes.
- Le rappel évalue le taux de véritables positifs qui ont été correctement identifiés.
- Le Fscore est un indicateur équilibré qui combine précision et rappel.

De manière générale, les trois algorithmes semblent présenter des résultats quasi identiques dans tous les aspects examinés, tels que la précision, le rappel, l score F et l'exactitude.

7. Conclusion

Dans ce chapitre, nous avons présenté le jeu de donnée qui a servi à l'évaluation de notre contribution, nous avons présenté trois algorithmes métaheuristiques de résolution d'entités tout en donnant les étapes principales de chacun d'eux.

Dans la deuxième partie nous avons exposé les résultats obtenues .nous avons fini par une présentation de quelque capture d'écrans de l'application.

Conclusion générale

L'augmentation fulgurante des volumes de données, souvent désignée sous le terme de Big Data, a rendu la résolution d'entités à la fois plus complexe et plus vitale que jamais. Il est essentiel de bien identifier, rassembler et combiner des enregistrements qui représentent la même entité afin de garantir la qualité, l'uniformité et la fiabilité des systèmes d'information.

L'emploi de métaheuristiques a prouvé son efficacité face à l'ampleur, la variabilité et l'imperfection des données qui posent des défis. Dans ce cadre, les algorithmes BES, WPO et GWO ont eu un rôle déterminant.

Ces algorithmes, tirés de comportements observés dans la nature, favorisent une exploration astucieuse de l'espace des solutions pour produire des clés de blocage optimales et par conséquent raffiner le processus d'appariement et de regroupement des entités.

En adoptant ces méthodes, on peut obtenir des résultats de grande qualité tout en diminuant considérablement le temps de calcul, un aspect crucial dans un contexte Big Data.

Par conséquent, l'association du Big Data, de la résolution d'entités et des métaheuristiques pave la voie vers des systèmes plus intelligents, rapides et fiables, en mesure de gérer des quantités de données toujours croissantes avec une exactitude renforcée.

Bibliographie

]1[GemaBello-Orgaz, Jason J Jung, and David Camacho. Social big data: Recent achieve ments and new challenges. Information Fusion, 28:45–59, 2016.

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

]3[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.

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

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

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

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

]8[(2009). Swoosh: a generic approach to entity resolution. The VLDB Journal The International Journal on Very Large Data Bases, 18(1), 255- 276.

]9[hristen, P. (2012a). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science Business Media.

]10[Thor, A., & Rahm, E. (2007, January). MOMA-A Mapping-based Object Matching System. In CIDR (pp. 247-258).

]11[Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.

]12[Leitão, L., Calado, P., & Weis, M. (2007, November). Structure-based inference of XML similarity for fuzzy duplicate detection. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (pp. 293-302). ACM.

]13[Christen, P. (2012a). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & Business Media.

]14[P. Christen, "Febrl - an open source data cleaning, deduplication and record linkage system with a graphical user interface," Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08, pp. 1065–1068, 2008.

]15[Reyes-Galaviz, O. F., Pedrycz, W., He, Z., & Pizzi, N. J. (2017). A supervised gradient-based learning algorithm for optimized entity resolution. Data & Knowledge Engineering, 112, 106-129.

]16[Mirjalili, S., Mirjalili, S. M., & Lewis, A.*Grey Wolf Optimizer*. Advances in Engineering Software, 69, 2014.

]17[Fathollahi-Fard, A. M., Hajiaghaei-Keshteli, M., & Tavakkoli-Moghaddam, R.*Bald Eagle Search: A New Nature-Inspired Metaheuristic*. Computers & Industrial Engineering, 2020.

]18[Fathollahi-Fard, A. M., Hajiaghaei-Keshteli, M., & Tavakkoli-Moghaddam, R. *Bald Eagle Search: A New Nature-Inspired Metaheuristic*. Computers & Industrial Engineering, 2020.

]**19[Deng, W., et al.***An Enhanced Woodpecker Mating Optimization Algorithm and Its Application in Engineering Problems*. IEEE Access, 2021.

]20[Chen, X., Li, X., & Wang, J. *Big Data and Entity Resolution: Challenges and Opportunities*.In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM), 2015

]20[**Christen, P., & Vatsalan, D.***Flexible and Extensible Generation and Corruption of Personal Data*.Springer, Transactions on Data Privacy, 201

خلاصة

عند تحسين مفاتيح الحظر والتجميع في سياق البيانات الضخمة، يمثل تحديد الكيانات تحديا رئيسيا للحفاظ على جودة البيانات (محسن الدئب الرمادي)GWOبحث النسر الأصلع(، و BES وتجميع الكيانات المتشابهة، بعد' تطبيق أساليب الاستدلال الفوقي، مثل (تحسين نقار الخشب)، فعالا بشكل خاص في توليد مفاتيح الحظر نظرا لتعقيد قواعد البيانات WPOالذئب الرمادي(، و وحجمها الهائل

تمكن هذه الخوارزميات، المستوحاة من الطبيعة، من استكشاف دقيق لمجال الحلول، وتضمن أداء فائقا من حيث الدقة والاسترجاع والكفاءة الحسابية .ويهدف تطبيق هذه الأساليب إلى تحسين موثوقية أنظمة المعلومات وسرعتها ودقتها، مع تلبية متطلبات بيئات البيانات واسعة النطاق.

الكلمات المفتاحية: جودة البيانات؛ حل الكيان؛ البيانات الضخمة؛ BES؛ GWO؛ WPO؛ مفتاح الحظر.

Abstract

In the context of Big Data, entity identification is a major challenge for maintaining data quality. When optimizing blocking keys and grouping similar entities, the implementation of metaheuristics such as BES (Bald Eagle Search), GWO (Grey Wolf Optimizer), and WPO (Woodpecker Optimization) is particularly effective in generating blocking keys given the complexity and sheer volume of databases.

These algorithms, inspired by nature, enable judicious exploration of the solution space and ensure superior performance in terms of precision, recall, and computational efficiency. The implementation of these methods aims to improve the reliability, speed, and accuracy of information systems, while meeting the requirements of large-scale data environments.

Keywords: Data quality; entity resolution; RE; Big Data; BES; GWO; WPO; Blocking key.

Résumé

Dans le cadre du Big Data, l'identification des entités représente un enjeu majeur pour maintenir la qualité des données.

Dans le cadre de l'optimisation des clés de blocage et du regroupement d'entités similaires, la mise en œuvre de métaheuristiques telles que BES (Bald Eagle Searche), GWO (Grey Wolf Optimizer) et WPO (Woodpecker Optimization) est particulièrement efficace dans la génération des clés de blocage face à la complexité et au volume considérable des bases.

Ces algorithmes, qui s'inspirent de la nature, permettent une exploration judicieuse de l'espace des solutions et assurent des performances supérieures en matière de précision, rappel et efficacité de calcul.

L'implémentation de ces méthodes vise à améliorer la fiabilité, la rapidité et l'exactitude des systèmes d'information, tout en répondant aux attentes des contextes de données à grande échelle.

Mots-clés : Qualité des données ; résolution d’entités ; RE ; Big Data; BES ; GWO; WPO; Clé de