République Algérienne Démocratique Et Populaire
Ministére de l'Enseignement Supérieur et De La Recherche Scientifique

Université Dr.TAHAR MOULAY SAIDA
Faculté des sciences
Départment Informatique

Mémoire deMaster

Spécialité : Réseaux Informatiques et Systèmes Répartis

Thème

Une méthode hybride pour le couplage d'enregistrements

Présenté par:

AMARI chaimaà

Dirigé par :

Mr . Benyahia Miloud

Annéeuniversitaire2023-2024

Remerciements

Tous d’abord, je remercie Dieu tout puissant de m’avoir donné le courage et la force afin de réaliser ce modeste travail.

je remercie chaleureusement mes parents, mes sœurs pour leurs contributions et leurs soutien constant ; sans leurs aides, leurs conseils et leurs encouragements, ce travail n’aurait vu le jour

je remercie Monsieur Benyahia miloud, mon directeur de recherche, pour sa présence, sa patience, ses orientations, ses corrections et ses suggestions qui m’ont beaucoup aidé dans ma recherche.

Mes remerciements vont également aux membres du jury qui ont accepté de lire et d’évaluer mon travail, ainsi que de participer à cette soutenance.

Dédicace

Je tiens à dédier ce mémoire de fin d'étude

A
Mes très chers parents que Dieu les protège

A
Ma très chère sœurs Fatin et Asmaa

A
Toute ma famille

A
Ma très chère amie et sœur Roufaida

Abstract

Every year, businesses around the world experience considerable losses due to data quality deficiencies. Stakeholders are increasingly aware of the importance of data quality. Significant sums are allocated to improve the quality of stored data. One of the key processes in the field of data quality is record matching (RL). RL (also known as entity reconciliation) is the process of detecting duplicates that refer to the same real entity in one or more datasets. One of the most crucial steps in the RL process is segmentation, which reduces the exponential complexity of the process by dividing the data into a set of blocks. This way, matching is only done between records in the same block. However, choosing the best segmentation keys to split data is a daunting task and in most cases it is done by a domain expert. Several approaches have been proposed in the literature for automatic selection of segmentation keys, but most are based on the existence of reference data, which is not the case for real-world datasets. In this paper, we propose a new unsupervised approach for automatic segmentation key selection. This approach is based on the recently proposed algorithm Grey Wolf Optimizer (GWO) and Bald eagle search (BES) , where we treat the problem as a case of feature selection. The results obtained from experiments on real-world datasets demonstrated the efficiency of our proposal where GWO outperformed existing approaches for feature selection in the literature and returned the best segmentation keys.

KEYWORDS:
Record linkage, blocking keys, blocking, matching, attribute selection, BES, GWO.

ملخص

في كل عام، تعاني الشركات في جميع أنحاء العالم من خسائر فادحة بسبب عيوب جودة البيانات. يدرك أصحاب المصلحة بشكل متزايد أهمية جودة البيانات. يتم تخصيص مبالغ كبيرة لتحسين جودة البيانات المخزنة. إحدى العمليات الرئيسية في مجال جودة البيانات هي مطابقة السجلات (RL). RL (المعروفة أيضًا باسم تسوية الكيان) هي عملية اكتشاف التكرارات التي تشير إلى نفس الكيان الحقيقي في مجموعة بيانات واحدة أو أكثر. إحدى الخطوات الأكثر أهمية في عملية RL هي التجزئة، مما يقلل من التعقيد الأسي للعملية عن طريق تقسيم البيانات إلى مجموعة من الكتل. بهذه الطريقة، تتم المطابقة فقط بين السجلات الموجودة في نفس الكتلة. ومع ذلك، فإن اختيار أفضل مفاتيح التجزئة لتقسيم البيانات يعد مهمة شاقة، وفي معظم الحالات يتم ذلك بواسطة خبير في المجال. تم اقتراح عدة طرق في الأدبيات للاختيار التلقائي لمفاتيح التجزئة، ولكن معظمها يعتمد على وجود بيانات مرجعية، وهذا ليس هو الحال بالنسبة لمجموعات البيانات الحقيقية. في هذا البحث، نقترح طريقة جديدة غير خاضعة للرقابة لاختيار مفتاح التجزئة التلقائي. يعتمد هذا النهج على خوارزميات Bald Eagle Search (BES) وGray Wolf Optimizer (GWO) المقترحة مؤخرًا، حيث نتعامل مع المشكلة كحالة اختيار الميزة. أظهرت النتائج التي تم الحصول عليها من التجارب على مجموعات البيانات الواقعية فعالية اقتراحنا الذي تفوق فيه GWO على أساليب اختيار الميزات الحالية في الأدبيات وأعاد أفضل مفاتيح التجزئة

الكلمات الدالة

ربط السجل، وحجب المفاتيح، والحجب، والمطابقة، واختيار السمة.

Résumer

Chaque année, les entreprises du monde entier subissent des pertes considérables en raison de défauts de qualité des données. Les parties prenantes sont de plus en plus conscientes de l'importance de la qualité des données. Des sommes importantes sont allouées pour améliorer la qualité des données stockées. L'un des processus clés dans le domaine de la qualité des données est l'appariement des enregistrements (RL). RL (également connu sous le nom de réconciliation d'entités) est le processus de détection des doublons faisant référence à la même entité réelle dans un ou plusieurs ensembles de données. L'une des étapes les plus cruciales du processus RL est la segmentation, qui réduit la complexité exponentielle du processus en divisant les données en un ensemble de blocs. De cette façon, la correspondance n'est effectuée qu'entre les enregistrements du même bloc. Cependant, choisir les meilleures clés de segmentation pour diviser les données est une tâche ardue et, dans la plupart des cas, elle est effectuée par un expert du domaine. Plusieurs approches ont été proposées dans la littérature pour la sélection automatique des clés de segmentation, mais la plupart reposent sur l'existence de données de référence, ce qui n'est pas le cas des jeux de données réels. Dans cet article, nous proposons une nouvelle approche non supervisée pour la sélection automatique des clés de segmentation. Cette approche est basée sur les algorithmes récemment proposés Grey Wolf Optimizer (GWO) et Bald EagleSearch (BES), où nous traitons le problème comme un cas de sélection de fonctionnalités. Les résultats obtenus à partir d'expériences sur des ensembles de données du monde réel ont démontré l'efficacité de notre proposition dans laquelle GWO a surpassé les approches existantes de sélection de caractéristiques dans la littérature et a renvoyé les meilleures clés de segmentation.

MOTS CLÉS:
Record linkage,clés de blocage,blocage,Matching,sélection des attributes, BES, GWO.

Table des matiéres:

Liste des figures:

Liste des tables:

Introduction générale

1-Introduction :

À l’ère numérique actuelle, le volume mondial de données a atteint des niveaux sans précédent, estimés à environ 610 zettaoctets – un chiffre qui souligne la nature omniprésente des données dans notre monde actuel. Alors que les organisations génèrent et collectent des données à partir de diverses sources dans divers formats, il existe un besoin crucial d'intégration des données pour permettre une analyse efficace des données et une extraction d'informations. Cependant, l'intégration des données peut être un processus long en raison de problèmes de qualité des données, tels que des valeurs manquantes ou en double, et de problèmes d'intégrité référentielle.

Les parties prenantes reconnaissent de plus en plus l’importance de la qualité des données et investissent des ressources substantielles pour améliorer la qualité des données stockées. Le couplage d'enregistrements (RL) est une tâche cruciale dans la qualité des données, qui implique l'identification des enregistrements qui correspondent à la même entité du monde réel dans différentes sources de données. Lorsque RL est appliqué à une seule base de données, on parle de processus de déduplication. Ces dernières années, RL a gagné du terrain dans divers domaines à de nombreuses fins, notamment la préservation de la vie privée, la suppression des citations bibliographiques en double, la comparaison de prix et la détection de fraude.

2-Problématique :

La meilleure façon de détecter tous les tuples qui font référence à lamême entité du monde réel est de comparer chacun d'entre eux dansl'ensemble de données à tous les autres etessayez de résoudre les problèmes d'exhaustivité (valeurs manquantes),de duplication valeurs, problèmes d'intégrité référentielle et bien d'autres anomalies.

3-Organisation du mémoire

Cette étude est structurée en chapitres et organisée comme suite :

Chapitre 1 : La qualité des données :

Nous allons présenter la qualité des données, nous donnerons un aperçu sur la qualité et nous parlerons sur les critères qui définissent la qualité des données,les problèmes de la non-qualité des données.

Chapitre 2 : Le couplage d'enregistrement

Nous allons présenter le Recorde Linkage, L'indexation et le blocage et l'implémentation évaluons Algorithme de K-Mods , BES et GWO qui résolue le problème de la sélection automatique des clés de blocage, et la mise en correspondance des paires d'enregistrements indexés (Matching).

Chapitre 3 : Implémentation et Exprérimentation :

Nous allons présenter l'environnement de travail et explication de notre application.

Chapitre01:La qualité des données

1-1Introduction

La qualité des données est un aspect essentiel de la stratégie de gestion des données de toute organisation, et elle devient encore plus importante dans le contexte du couplage d'enregistrements. Le couplage d'enregistrements est le processus d'identification et de liaison des enregistrements qui font référence aux mêmes entités dans différentes sources de données. Cependant, ce processus peut être sujet à des erreurs dues à des facteurs tels que des données manquantes ou incohérentes, ce qui peut entraîner de fausses correspondances ou des liens manqués entre les enregistrements. Ces erreurs peuvent avoir des conséquences importantes, notamment une analyse de données inexacte et une mauvaise prise de décision. Par conséquent, garantir une qualité élevée des données est essentiel pour un couplage efficace des enregistrements et pour tirer des conclusions valides à partir des données couplées. La mise en œuvre des meilleures pratiques en matière d'évaluation de la qualité des liens nécessite un engagement et un partage d'informations accrus entre les collecteurs de données, les éditeurs de liens et les analystes de données. Cela comprend la génération et le partage d'informations pertinentes pendant le couplage et l'utilisation de mesures de qualité de couplage appropriées pour évaluer le niveau d'erreur de couplage et son impact sur l'analyse. En donnant la priorité à la qualité des données et en mettant en œuvre des pratiques efficaces de couplage d'enregistrements, les organisations peuvent garantir que leurs données sont exactes, fiables et adaptées à leur objectif.

1-2-La qualité des données:

1-Définition :

La qualité des données fait référence à l'état d'un ensemble de valeurs de variables qualitatives ou quantitatives. Une qualité élevée des données signifie que les données sont exactes,complètes, fiables,pertinentes et opportunes, ce qui les rend adaptées à leur utilisation prévue dans les opérations, la prise de décision et la planification.Les principaux attributs de la qualité des données comprennent l'exactitude, la cohérence, l'exhaustivité, la fiabilité, la pertinence,l'actualité et la validité.

1-3-Les critères de la qualité des données :

1-3-1-Les critères intrinsèques de qualité:

1-Exactitude :L'exactitude fait référence au degré auquel les données représentent correctement l'objet ou l'événement du monde réel qu'elles sont censées décrire. Les données précises sont exemptes d'erreurs, de distorsions et de biais. C'est un critère essentiel pour la qualité des données car il garantit que les décisions et les actions basées sur les données sont valides et fiables.[1]

2-Complétude : L'exhaustivité fait référence à la mesure dans laquelle les données comprennent toutes les informations nécessaires pour soutenir leur utilisation prévue. Les données complètes sont exemptes de valeurs manquantes ou incomplètes et fournissent une vue complète de l'objet ou de l'événement qu'elles décrivent. L'exhaustivité est cruciale pour la qualité des données, car elle garantit que les décisions et les actions basées sur les données sont bien informées et complètes.

3-Cohérence :La cohérence fait référence au degré auquel les données sont présentées de manière uniforme et standardisée, en suivant les mêmes règles et conventions dans différentes sources et contextes. Des données cohérentes sont exemptes de contradictions, d'ambiguïtés et de divergences et facilitent l'intégration, la comparaison et l'analyse des données. La cohérence est essentielle à la qualité des données car ellegarantit que les décisions et les actions basées sur les données sont comparables et fiables.

4-Actualité :L'actualité fait référence au degré auquel les données sont disponibles et à jour, reflétant l'état actuel de l'objet ou de l'événement qu'elles décrivent. Les données opportunes sont récentes, pertinentes et accessibles, et elles soutiennent la prise de décision et l'action en temps réel. La rapidité est essentielle à la qualité des données, car elle garantit que les décisions et les actions basées sur les données sont réactives et efficaces.

5-Unicité :L'unicité fait référence au degré dans lequel les données sont exemptes de doublons, de redondances et d'ambiguïtés, et garantit que chaque élément de données est distinct et identifiable. Des données uniques sont essentielles à la qualité des données car elles garantissent que les décisions et les actions basées sur les données sont précises et sans ambiguïté.

6-Validité :Degré auquel les données sont conformes à un ensemble de règles ou de contraintes prédéfinies. Cela peut être mesuré en évaluant la proportion de données qui enfreignent ces règles ou contraintes.

1-3-2-Les Critères de Sécurité à L'actualité:

1-L'accessibilité:L'accessibilité concerne la facilité d'accès aux données. D'une part, il est nécessaire de savoir où se trouve réellement l'information et quel est lepoint de vérité de de la donnée au sein du SI. D'autre part, il faut adapter cette accessibilité en fonction de la volumétrie des bases et de l'usage qu'il est fait de la donnée par le calibrage d'un mode événement (à chaque mise à jour) et d'un mode requête (à la demande de l'utilisateur) ou encore en mode batch pour les synchronisations en masse. **[2]**

2-La pertinence:C'est le critère plébiscité par les équipes métier.Une donnée est de qualité que si elle est pertinente c'est-à-dire utile. Une donnée trop détaillée peut être inutile par les processus ou les briques applicatives qui la consomment.Afin d'être pertinente la donnée doit être adaptée et en adéquation avec sonusage.

3-La compréhensibilité:La donnée doit en effet être compréhensible pour chaque utilisateur. Ce critère pré suppose un alignement dans la signification de l'attribut ou del'objet.Afin d'atteindre ce niveau de compréhension partagé,il est préconisé que les univers,concepts et termes métier soient documentés par l'existence d'un glossaire métier,d'un dictionnaire de données,d'un inventaire des traitements et des usages de cette donnée.

1-4-Principaux problèmes du non qualité des données :

I-4-1-Les problèmes de la qualité des données :

1-Données en double :Des enregistrements en double peuvent survenir pour diverses raisons, notamment une erreur humaine lors de la saisie des données ou des problèmes système.Ils gonflent la taille de vos bases de données,gaspillant des ressources de stockage et faussant les résultats d'analyse.

2- Données inexactes :Les inexactitudes des données peuvent provenir de fautes de frappe, de fausses informations ou d'entrées obsolètes.

Ces inexactitudes peuvent conduire à des informations erronées et à des prises de décisions erronées.

3- Informations obsolètes :Les données peuvent devenir rapidement obsolètes, en particulier dans les secteurs en évolution rapide. S'appuyer sur des données obsolètes peut fausser les décisions stratégiques et l'allocation des ressources.

4-Valeurs manquantes :Les lacunes dans vos données peuvent avoir de graves conséquences sur les analyses et donner lieu à des informations trompeuses. Parfois, les données manquantes peuvent être plus nuisibles que les données incorrectes.

5- Données non standardisées :Les différents formats, unités ou terminologies selon les sources de données peuvent entraver l'efficacité de l'analyse des données, rendant difficile la comparaison ou l'agrégation des données.

6- Sécurité et confidentialité des données :Des protocoles de sécurité des données inadéquats peuvent conduire à un accès non autorisé aux données, ce qui met non seulement en danger les données elles-mêmes, mais peut également avoir des conséquences juridiques.

1-4-2-Les problèmes de rencontres:

1-Création des données:

1-**Entrée manuelle** :absence de vérifications systématiquedes formulaires de saisie.
2-**Entrée automatique** : problémesde capture OCR,dereconnaissance de la parole incomplétude,absence de normalisation ou inadéquation de la modélisation conceptuelle des données:attributs peu structures,absence de contraintes d'intégrité pourmaintenir la coherence des données.
3- **Entrée de doubloons.**
4**- Approximations.**
5**-Erreur demesure**.
6-**Corruption des données:faille de sécurité physique et logique des données.**

7-Contraintes matérielles ou logicielles.

2-Collecte/importdesdonnées:

1) Destructionou mutilation d'information par des prétraitements inappropriés.
2) Perte de données : buffer over flows, problèmes de transmission.
3) Absence de vérification dans les procédures d'importmassif.
4) -Introduction d'erreurs par les programmes de conversion de données.

3-Stockagedes données :

1) Absence de méta-données.
2) Absence de mise à jour et de rafraîchissement des données obsolètes ou répliquées.
3) Modifications ad-hoc.
4) Modèles et structures de données inappropriés,spécifications incomplètes ou évolution des besoins dans l'analyse et conception dusystème.
5) Contraintes matérielles ou logicielles.

4-Intégration desdonnées:

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

5-Rechercheetanalysedesdonnées:

1) Erreur humaine.
2) Contraintes liées à la complexité de calcul.
3) Contrainteslogicielles, incompatibilité.
4) Problèmes de passage à l'échelle,de performances et de confiance dans les résultats.
5) Approximations dues aux techniques de réduction des grandes dimensions.

1-5-Approches générales et cas pratique pour détecter et corriger lesproblèmesdequalitédes données :

Comme le représente la Figure 1, on peut classer la plupart des travaux abordant la problématique de la qualité des données selon quatre grands types d'approches complémentaire. [3]

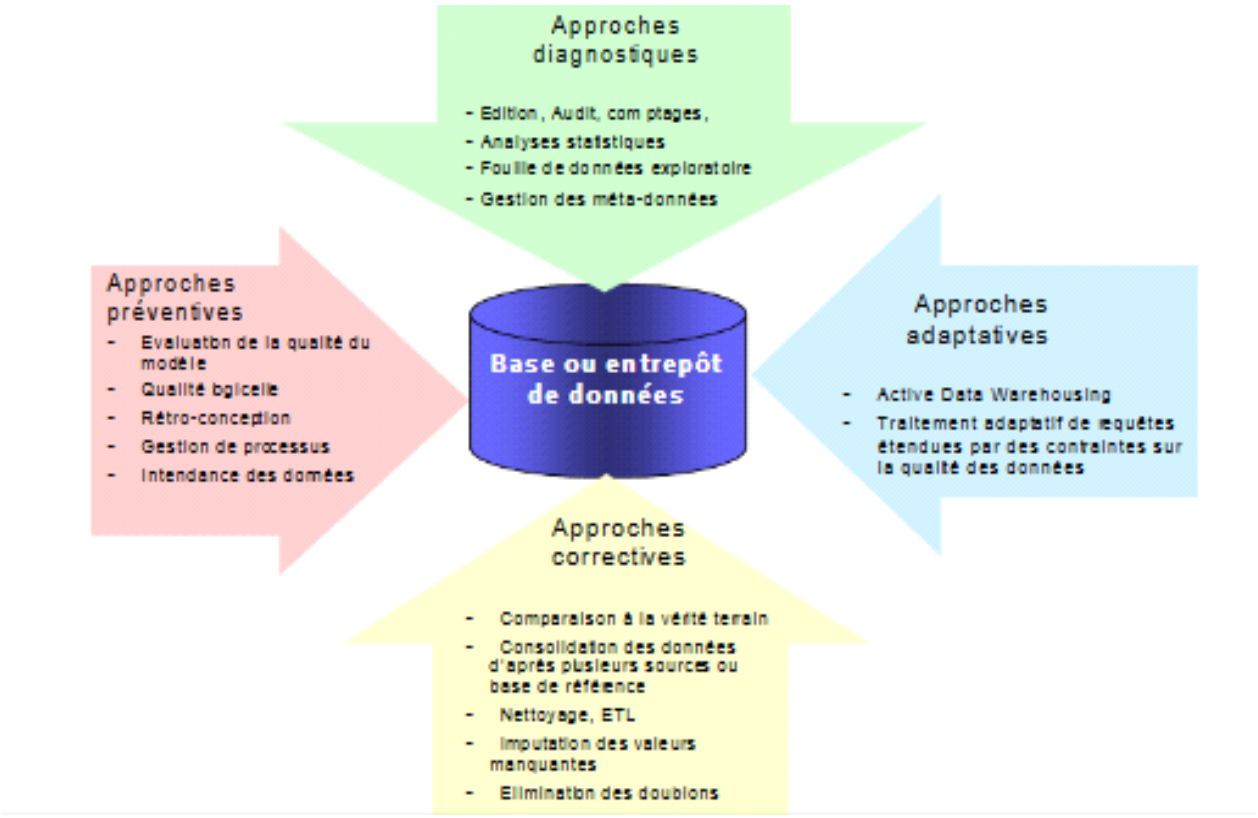


Figure 1 : Panorama des approches pour l'évaluation et le contrôle de laqualitédes données.

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

2-Les approches diagnostiques :centrées sur des méthodes Statistiques et d'analyse et de fouille de données exploratoire permettant de détecter des anomalies sur les données.

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

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

1-6-L'importance de la qualité de données:

La qualité des données joue un rôle crucial dans le couplage d'enregistrements,car elle a un impact direct sur l'exactitude, la fiabilité et l'efficacité du processus de couplage. Voici quelques raisons pour lesquelles la qualité des données est importante dans le couplage d'enregistrements :

- Correspondance précise.
- Réduction des faux positifs et des faux négatifs.
- Intégration des données améliorée.
- Analyse de données améliorée.
- Efficacité accrue.
- Amélioration de la confidentialité et de la sécurité des données.

1-7-L'intégration des données :

L'intégration des données dans le couplage d'enregistrements est complexe,mais essentielle dans de nombreux domaines, tels que les soins de santé et la finance. les données intégrées sont analysées pour obtenir des informations et prendre des décisions éclairées.

1-8-Conclusion :

La conclusion concernant la qualité des données dans le couplage d'enregistrements est que les techniques de nettoyage des données ont un effet minime sur la qualité du couplage. En fait, un nettoyage approfondi peut même entraîner une diminution de la qualité. En effet, les techniques de nettoyage réduisent généralement la variabilité, ce qui rend les enregistrements corrects plus susceptibles de correspondre, mais rend également les enregistrements incorrects plus susceptibles de correspondre, ce qui l'emporte sur les correspondances correctes et réduit la qualité globale.

Par conséquent, des précautions doivent être prises lors du processus de nettoyage des données.

Chapitre02:Le couplage d’enregistrement

2-1-Introduction:

Dans le monde actuel axé sur les données, les organisations sont confrontées à une quantité considérable de données provenant de diverses sources, notamment des bases de données et des feuilles de calcul et des fichiers.Cependant,ces données sont souvent fragmentées et incomplètes et incohérentes, ce qui rend difficile une compréhension globale des entités qu'elles représentent. Par exemple, une même personne peut détenir plusieurs enregistrements dans différents systèmes, chacun contenant différents niveaux d'informations et d'exactitude. Ce phénomène est communément appelé problème du « silo de données ».

Pour surmonter ce défi, les organisations ont besoin d'un moyen d'identifier et de relier ces enregistrements disparates, créant ainsi une vue unifiée de leurs données. C'est là qu'intervient le couplage d'enregistrements : une technique puissante qui permet d'intégrer des données provenant de plusieurs sources, d'identifier les enregistrements en double et de les relier entre eux pour former une représentation unique et précise d'une entité. Ce faisant, le couplage des enregistrements ouvre de nouvelles perspectives, améliore la qualité des données et renforce les capacités de prise de décision.

2-2-Définition de couplage d'enregistrement:

Est la tâche consistant à rechercher des enregistrements dans un ensemble de données qui font référence à la même entité dans différentes sources de données. Cela est nécessaire lors de la jonction de différents ensembles de données basés sur des entités qui peuvent ou non partager un identifiant commun, ce qui peut être dû à des différences dans la forme des enregistrements, l'emplacement de stockage ou le style ou les préférences du conservateur.

En informatique, le couplage d'enregistrements est également connu sous le nom de couplage de données ou de déduplication (dans le cas de la recherché d'enregistrements en double dans un seul fichier). [4]

Alors, 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 une seule. RL aide à améliorer considérablement la qualité des données en supprimant tous les enregistrements qui font référence à la même entité réelle.

2-3-Types de couplage :

Il existe deux types de couplage d'enregistrements : l'appariement exact et l'appariement statistique. L'appariement statistique se divise en deux sous-types : le couplage d'enregistrements déterministe et le couplage d'enregistrements probabiliste. [5]

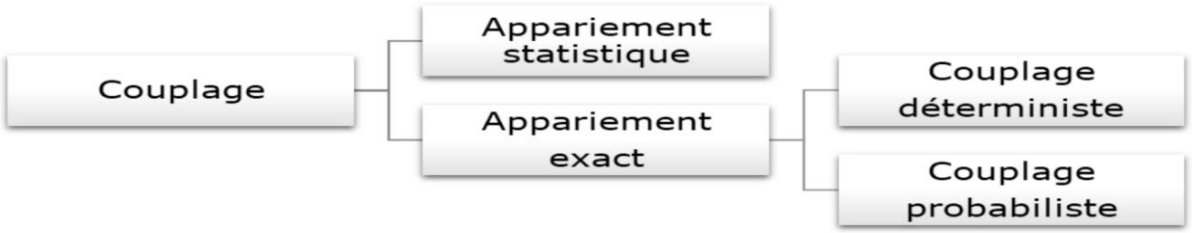


Figure 2 : Les types de couplage d'enregistrement

2-3-1-Appariement exact :

L'objectif de l'appariement exact est de relier les informations relatives à un enregistrement particulier dans un fichier aux informations d'un fichier secondaire afin de créer un seul fichier avec des informations correctes pour chaque enregistrement. Le couplage est effectué au niveau de l'enregistrement, par exemple un lien entre les enregistrements de mortalité et le recensement de la population.

2-3-2-Appariement statistique :

L'objectif de l'appariement statistique est de créer un fichier reflétant la distribution de la population sous-jacente. Les enregistrements qui sont combinés ne correspondent pas nécessairement à la même entité, telle qu'une personne ou une entreprise. Les fichiers qui sont appariés peuvent avoir des unités différentes, mais se référer à la même population. On suppose que la relation des variables dans la population sera similaire à la relation dans les fichiers. Cette méthode est principalement utilisée dans les études de marché et rarement par les agences statistiques officielles.

2-3-2-1-Couplage d'enregistrements déterministe :

Est une méthode utilisée pour identifier et lier les enregistrements faisant référence à la même entité dans différentes sources de données, sur la base de correspondances exactes de variables d'identification spécifiques, telles que le nom, le numéro de sécurité sociale ou la date de naissance. Cette approche est moins flexible que le couplage probabiliste d'enregistrements, car elle repose sur des correspondances exactes, mais elle peut être plus précise lorsque la qualité des données est élevée et que le nombre de correspondances possibles est limité. Le couplage déterministe d'enregistrements est souvent utilisé dans des applications telles que l'intégration de données, le nettoyage des données et la gestion des données de référence.

2-3-2-2Couplage 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. Le but de l'appariement statistique est de créer un fichier qui refléte la répartition sous-jacente de la population. 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.

2-4-Les étapes de Record Linkage :

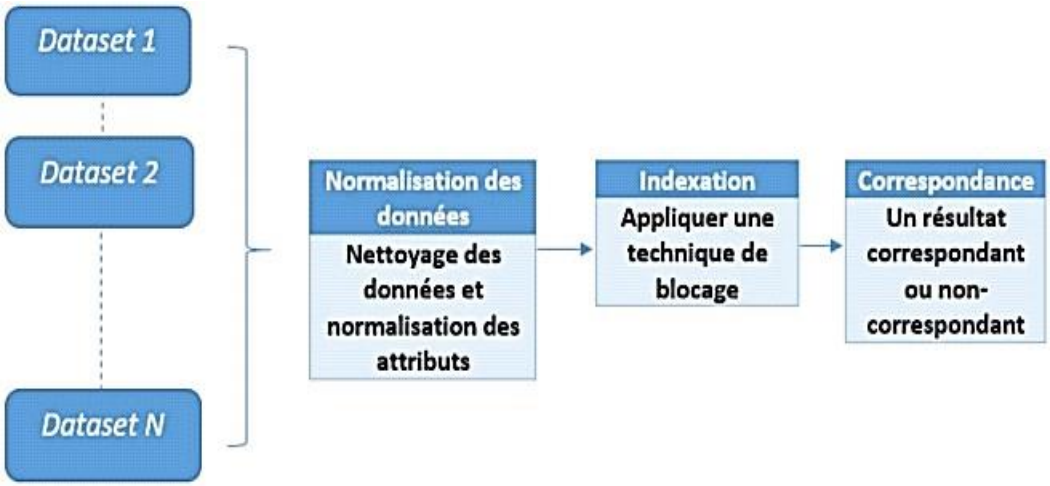


Figure 3 : Les étapes de couplage d'enregistrement

2-4-1-Nettoyage et normalisation:

Le nettoyage et la normalisation sont des étapes essentielles du couplage d'enregistrements, un processus de liaison d'enregistrements sur différents ensembles de données. Le nettoyage implique la correction, la suppression ou la modification des valeurs de données incorrectes ou incohérentes pour améliorer la qualité et l'exactitude des données.

La normalisation transforme les données dans un format cohérent, permettant une comparaison et une mise en correspondance efficaces des enregistrements. Cela inclut la suppression des jetons, des espaces et des crochets indésirables, ainsi que la conversion des chaînes en codes phonétiques à l'aide d'algorithmes tels que Soundex pour identifier les mots à consonance similaire malgré des orthographes différentes.

2-4-2-L'indexation:

L'indexation dans le couplage d'enregistrements est un processus de sélection d'un sous-ensemble d'enregistrements à partir d'un ensemble de données susceptibles de correspondre, réduisant ainsi le nombre de comparaisons nécessaires pour relier les enregistrements.

2-4-2-1-K-Modes:

Définition :

Les K-Modes sont un type d'algorithme de clustering utilisé pour les données catégorielles. Il s'agit d'une extension de l'algorithme K-Means, conçu pour les données numériques. K-Modes est utilisé pour partitionner les données en K clusters en fonction des modes des variables catégorielles.

Dans les K-Modes, chaque cluster est représenté par un mode, qui est la valeur la plus fréquente dans chaque catégorie. L'algorithme met à jour les modes de manière itérative et réaffecte les objets au cluster avec le mode le plus similaire jusqu'à ce que la convergence ou un critère d'arrêt soit atteint.

K-Modes est particulièrement utile pour regrouper des données catégorielles, telles que des données textuelles, des réponses à des enquêtes ou des données génomiques.

Les auteur sont basé leur algorithme sur trois principaux points:

(1) Une simple mesure de dissemblance de qui correspond à un objet.
(2) Utilisation modes à la place des moyens.
(3) Approche basée sur la fréquence pour mesurer le monde d'un ensemble.
(4) Deux méthodes ont été proposes dans le magazine pour la selection initiale desmodes.

Les modes K sont un algorithme de clustering qui peut être utilisé dans le couplage d'enregistrements pour regrouper des enregistrements similaires en fonction de leurs attributs catégoriels. Il est particulièrement utile lorsqu'il s'agit de données catégorielles, car il peut gérer la rareté inhérente et la nature non numérique de ces données. Dans le couplage d'enregistrements, les modes K peuvent être utilisés pour identifier les enregistrements en double dans différents ensembles de données en regroupant les enregistrements avec des attributs similaires. [6]

2-4-2-2-Le blocage:

1-Définition:

Le blocage est la technique la plus utilisée dans l'étape de l'indexation.
Le blocage fait référence au processus de division d'un grand ensemble de données en morceaux plus petits et plus faciles à gérer, appelés blocs, pour faciliter une indexation et des requêtes efficaces.

Chaque bloc contient généralement un sous-ensemble d'enregistrements partageant des caractéristiques similaires, telles que des valeurs similaires dans un attribut spécifique ou une plage de valeurs. Le blocage est utilisé pour réduire le nombre de comparaisons requises lors de l'indexation et de l'interrogation, rendant ainsi le processus plus rapide et plus efficace.

Exemple:
Blocage des clés générées à partir du jeu de données du restaurant.
Deux clés bloquante sont été générées.

Le premier(BK1)est l'encodage phonétique Soundex du nom du restaurant concaténé avec le numéro det éléphone.
Le deuxième (BK2) est l'encodage phonétique NYSIIS du restaurant Nom concaténé avec le numéro d'adresse.

BK1	BK2	Name	Address	City	Phone	Type
A6553102461501	LASANG435	arnie morton's of Chicago	435 s. la cienega blv.	Los Angeles	310/246-1501	American
H3413104721211	STADAC12224	art's deli	12224 ventura bold	Studio city	818-762-1221	delis

Tableau1:clés de blocage dans l'ensemble des données du restaurant.

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

a-la valeur de la clé de blocage:La valeur de la clé de blocage détermine la manière dont les enregistrements similaires sont regroupés en blocs. Une bonne clé de blocage doit avoir une forte correlationavec la similarité des enregistrements, de sorte que des enregistrements similaires aient des valeurs de clé de blocage similaires. Le choix de la valeur de la clé de blocage peut avoir un impact significatif sur la qualité des blocs et sur les performances globales de la technique de blocage.

b-Le nombre de clés de blocage :Le nombre de clés de blocage détermine le nombre d'attributs ou de fonctionnalités différents utilisés pour créer la clé de blocage. L'utilisation de plusieurs clés de blocage peut améliorer la précision de la technique de blocage en capturant davantage de nuances dans les données. Cependant, l'augmentation du nombre de clés de blocage peut également augmenter la charge de calcul et entraîner un ralentissement des performances. Le nombre optimal de clés de blocage dépend de la complexité des données et du cas d'utilisation spécifique. [7]

3-L'objective de blockage:

L'objectif bloquant du couplage d'enregistrements est de regrouper les enregistrements potentiellement concordants, réduisant ainsi le nombre considérable de comparaisons possibles. Ceci est réalisé en créant un ou plusieurs index de blocage, qui visent à réduire autant que possible le nombre de comparaisons effectuées tout en garantissant qu'aucune correspondance potentielle n'est négligée en raison du processus d'indexation.

2-4-2-3-La sélection automatique des clés :

1-Définition de GWO:

Le Grey Wolf Optimizer (GWO) est un algorithme d'optimisation métaheuristique inspiré de la hiérarchie sociale et du comportement de chasse des loups gris. Il a été proposé en 2014 par Mirjalili et al. comme une nouvelle technique d'optimisation pour résoudre des problèmes d'optimisation complexes.

2-Comportement des loups gris lors de la chasse :

Tout d'abord, nous expliquons le comportement des loups lors de lachasse qui utilise quatre types de loups gris tels que : alpha, bêta,delta et oméga pour simuler la hiérarchie duleadership, et qui peutse résumer en trois étapes principales :de la recherche de proies, de l'encerclement des proies et de l'attaque des proies.

Les loups gris ont la capacité de reconnaître l'emplacement desproies et de les encercler,puis ils recherchent principalement enfonction de la position de l'alpha, du bêta et du delta. Ils divergentles uns des autres pour rechercher des proies et convergent pour attaquerles proies.

La chasse est généralement guidée par l'alpha. Le bêta et le deltapeuvent également participer à la chasse à l'occasion. Cependant,dans un espace de recherche abstrait, nous n'avons aucune idée del'emplacement de l'optimum (proie). Ils terminent la chasse en attaquantla proie lorsqu'elle cesse de bouger.

3-Inspiration :

Le loup gris (Canis lupus) appartient à la famille des Canidés. Les loups gris sont considérés comme des prédateurs au sommet, ce qui signifie qu'ils se situent au sommet de la chaîne alimentaire. Les loups gris préfèrent généralement vivre en meute. La taille du groupe est de 5 à 12 personnes en moyenne. Ce qui est particulièrement intéressant, c'est qu'ils ont une hiérarchie sociale dominante très stricte.

Les dirigeants sont un homme et une femme, appelés alphas. L'alpha est principalement responsable de la prise de décisions concernant la chasse, le lieu de sommeil,l'heure de réveil, etc. Les décisions de l'alpha sont dictées à la meute. Cependant,une sorte de comportement démocratique a également été observé, dans lequel un alpha suit les autres loups de la meute. Lors des rassemblements, la meute entière reconnaît l'alpha en gardant la queue baissée. Le loup alpha est aussi appelé loup dominant puisque ses ordres doivent être suivis par la meute. Les loups alpha ne sont autorisés à s'accoupler qu'en meute. Fait intéressant, l'alpha n'est pas nécessairement le membre le plus fort de la meute mais le meilleur en termes de gestion de la meute. Cela montre que l'organisation et la discipline d'une meute sont bien plus importantes que sa force.

Le deuxième niveau de la hiérarchie des loups gris est bêta. Les bêtas sont des loups subordonnés qui aident l'alpha dans la prise de décision ou dans d'autres activités de la meute. Le loup bêta peut être un mâle ou une femelle, et il est probablement le meilleur candidat pour devenir l'alpha au cas où l'un des loups alpha décèderait ou deviendrait très vieux. Le loup bêta doit respecter l'alpha, mais commande également les autres loups de niveau inférieur. Il joue le rôle de conseiller de l'alpha et de disciplineur de la meute. La version bêta renforce les commandes de l'alpha tout au long du pack et donne un feedback à l'alpha.

Le loup gris le moins bien classé est oméga. L'oméga joue le rôle de bouc émissaire. Les loups Omega doivent toujours se soumettre à tous les autres loups dominants. Ce sont les derniers loups autorisés à manger. Il peut sembler que l'oméga n'est pas un individu important dans la meute, mais il a été observé que l'ensemble de la meute est confronté à des combats internes et à des problèmes en cas de perte de l'oméga. Cela est dû à l'évacuation de la violence et de la frustration de tous les loups par le(s) oméga(s). Cela aide à satisfaire l'ensemble de la meute et à maintenir la structure de domination. Dans certains cas, les oméga sont aussi les baby-sitters de la meute.

Si un loup n'est pas un alpha, un bêta ou un oméga, il est appelé subordonné (ou delta dans certaines références). Les loups Delta doivent se soumettre aux alphas et aux bêtas, mais ils

dominent l'oméga. Les éclaireurs, les sentinelles, les anciens, les chasseurs et les gardiens appartiennent à cette catégorie.

Les scouts sont chargés de surveiller les limites du territoire et d'avertir la meute en cas de danger. Les sentinelles protègent et garantissent la sécurité de la meute. Les aînés sont des loups expérimentés qui étaient autrefois alpha ou bêta. Les chasseurs aident les alphas et les bêtas lorsqu'ils chassent des proies et fournissent de la nourriture à la meute. Enfin, les soigneurs sont chargés de soigner les loups faibles, malades et blessés de la meute.

Outre la hiérarchie sociale des loups, la chasse en groupe est un autre comportement social intéressant des loups gris. Selon Muro et al[8].les principales phases de la chasse au loup gris sont les suivantes :

•Traquer, pour suivre et approcher la proie.
•Poursuivre, encercler et harceler la proie jusqu'à ce qu'elle s'arrête de bouger.
•Attaque vers la proie.

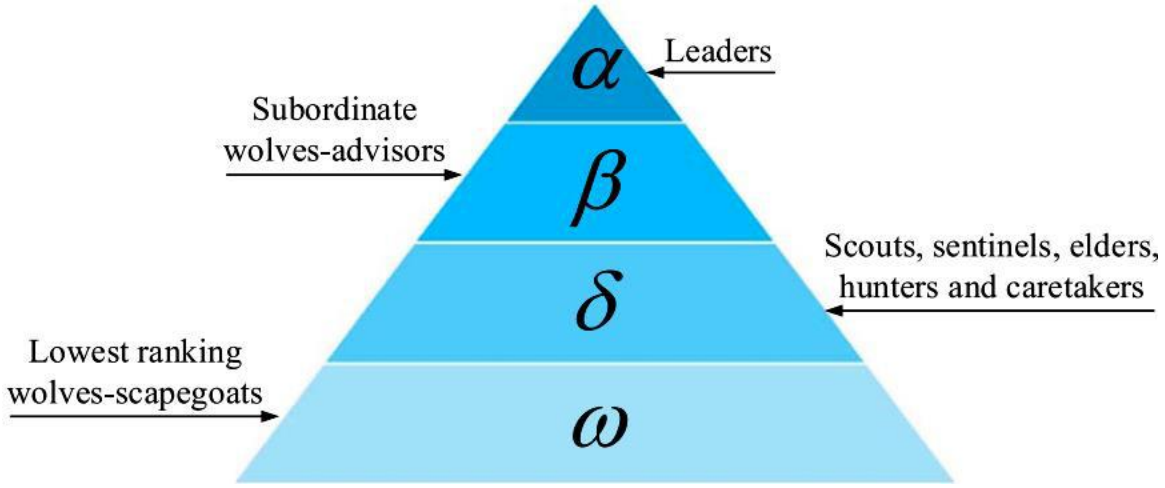


Figure 4 : Hiérarchie du loup gris (la dominance diminue de haut en bas)

4-Modèle mathématique :

La technique de chasse et la hiérarchie sociale des loups gris sont modélisées mathématiquement afin de concevoir GWO et d'en réaliser l'optimisation. Les modèles mathématiques proposés pour la hiérarchie sociale, le suivi, l'encerclement et l'attaque des proies sont les suivants :

a-Hierarchie sociale :

Afin de modéliser mathématiquement la hiérarchie sociale des loups lors de la conception de GWO, nous considérons la solution la plus adaptée comme l'alpha (α). Par conséquent, les deuxième et troisième meilleures solutions sont respectivement appelées bêta (β) et delta (δ). Le reste des solutions candidates sont supposées être des oméga (ω). Dans l'algorithme GWO, la recherche (optimisation) est guidée par α, β et δ. Les loups ω suivent ces trois loups.

b-Proie encerclée :

Comme mentionné ci-dessus, les loups gris encerclent leurs proies pendant la chasse. Afin de modéliser mathématiquement le comportement encerclant, les équations suivantes sont proposées :

$$\vec{D} = \left|\vec{C}.\vec{X}_p(t) - \vec{X}(t)\right| \qquad (3.1)$$

$$\vec{X}(t+1) = \vec{X}_p(t) - \vec{A}.\vec{D} \qquad (3.2)$$

où t indique l'itération en cours,$\vec{A}$ et $\vec{C}$ sont des vecteurs de coefficients ,$\vec{X}_p$est le vecteur position de la proie ,$\vec{X}$ indique le vecteur de position d'un loup gris.

Les vecteurs $\vec{A}$ et $\vec{C}$ sont calculés comme suit :

$$\vec{A} = 2\vec{a}.\vec{r}_1 - \vec{a} \qquad (3.3)$$

$$\vec{C} = 2.\vec{r}_2 \qquad (3.4)$$

Où les composants de $\vec{a}$ sont linéairement diminués de 2 à 0 au cours des itérations et r_1, r_2

Avec les équations ci-dessus, un loup gris en position (X,Y) peut mettre à jour sa position en fonction de la position de la proie (X*,Y*). Différents endroits autour du meilleur agent peuvent être atteints par rapport à la position actuelle en ajustant la valeur de $\vec{A}et\vec{C}$.Par exemple, (X*-X,Y*) peut être atteint en définissant $\vec{A} = (1,0)$ et $\vec{C} = (1,1)$.
Notez que les vecteurs aléatoires r_1 et r_2 permettent aux loups d'atteindre n'importe quelle position entre les deux points particuliers. Ainsi, un loup gris peut mettre à jour sa position à l'intérieur de l'espace autour de la proie dans n'importe quel endroit aléatoire grâce aux équations mentionnées ci-dessus.

c-Chasse :

Les loups gris ont la capacité de reconnaître l'emplacement de leurs proies et de les encercler. La chasse est généralement guidée par l'alpha. Les bêta et les delta peuvent également participer occasionnellement à la chasse. Cependant, dans un espace de recherche abstrait, nous n'avons aucune idée de l'emplacement de l'optimum (proie).

Afin de simuler mathématiquement le comportement de chasse des loups gris, nous supposons que les alpha (meilleure solution candidate) bêta et delta ont une meilleure connaissance de la localisation potentielle des proies.

Par conséquent, nous sauvegardons les trois premières meilleures solutions obtenues jusqu'à présent et obligeons les autres agents de recherche (y compris les omégas) à mettre à jour leurspositions en fonction de la position du meilleur agent de recherche. Les formules suivantes sont proposées à cet égard :

$$\vec{D}_\alpha = |\vec{C}_1.\vec{X}_\alpha - \vec{X}|, \vec{D}_\beta = |\vec{C}_2.\vec{X}_\beta - \vec{X}| \ , \vec{D}_\delta = |\vec{C}_3.\vec{X}_\delta - \vec{X}| \qquad (3.5)$$

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1.(\vec{D}_\alpha), \vec{X}_2 = \vec{X}_\beta - \vec{A}_2.(\vec{D}_\beta), \vec{X}_3 = \vec{X}_\delta - \vec{A}_3.(\vec{D}_\delta) \qquad (3.6)$$

$$\vec{X}(t+1) = (\vec{X}_1 + \vec{X}_2 + \vec{X}_3)/3 \qquad (3.7)$$

Avec ces équations, un agent de recherche met à jour sa position en fonction de alpha, bêta et delta dans un espace de recherche à n dimensions. De plus, la position finale serait à un endroit aléatoire dans un cercle défini par les positions alpha, bêta et delta dans l'espace de recherche. En d'autres termes, alpha, bêta et delta estiment la position de la proie, et les autres loups mettent à jour leurs positions de manière aléatoire autour de la proie.

d-Attaquer une proie (exploitation) :

Comme mentionné ci-dessus, les loups gris terminent la chasse en attaquant la proie lorsqu'elle arrête de bouger.Afin de modéliser mathématiquement l'approche de la proie,nous diminuons la valeur de $\vec{a}$.Notez que la plage de fluctuation de $\vec{A}$est également diminué de $\vec{a}$.Autrement dit $\vec{A}$est une valeur aléatoire dans l'intervalle [-2a,2a] où a diminue de 2 à 0 au cours des itérations. Lorsque des valeurs aléatoires de $\vec{A}$ sont dans [-1,1], la position suivante d'un agent de recherche peut être dans n'importe quelle position entre sa position actuelle et la position de la proie.

Avec les opérateurs proposés jusqu'à présent, l'algorithme GWO permet à ses agents de recherche de mettre à jour leur position en fonction de l'emplacement de l'alpha, du bêta et du delta ; et attaque vers la proie. Cependant, l'algorithme GWO est sujet à une stagnation des solutions locales avec ces opérateurs. Il est vrai que le mécanisme d'encerclement

proposé montre dans une certaine mesure l'exploration, mais GWO a besoin de plus d'opérateurs pour mettre l'accent sur l'exploration.

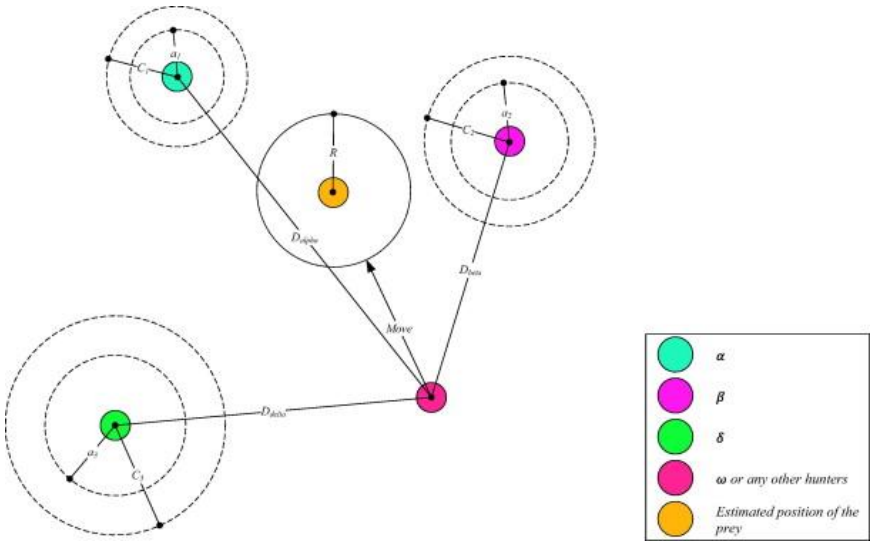


Figure 5 : Mise à jour de la position des loups gris dans GWO

e-Recherche de proies (exploration) :

Les loups gris recherchent principalement en fonction de la position de l'alpha, du bêta et du delta. Ils s'écartent les uns des autres pour rechercher des proies et convergent pour attaquer des proies. Afin de modéliser mathématiquement la divergence, nous utilisons $\overrightarrow{A}$ avec des valeurs aléatoires supérieures à 1 ou inférieures à -1 pour obliger l'agent de recherche à s'écarter de la proie. Cela met l'accent sur l'exploration et permet à l'algorithme GWO d'effectuer une recherche globale. |A|>1 force les loups gris à s'écarter de la proie dans l'espoir de trouver une proie plus en forme. Une autre composante de GWO qui favorise l'exploration est $\vec{C}$,qui contient des valeurs aléatoires dans [0, 2]. Cette composante fournit des poids aléatoires pour les proies afin d'accentuer de manière stochastique (C>1) ou d'atténuer (C<1) l'effet de la proie dans la définition de la distance dans l'équation (3.1). Cela aide GWO à montrer un comportement plus aléatoire tout au long de l'optimisation,favorisant l'exploration et l'évitement des optima locaux. Il convient de mentionner ici que C n'est pas linéairement diminué contrairement à A.

Nous exigeons délibérément que C fournisse des valeurs aléatoires à tout moment afin de mettre l'accent sur l'exploration non seulement lors des itérations initiales mais également lors des itérations finales. Ce composant est très utile en cas de stagnation des optima

locaux, notamment dans les itérations finales.Le vecteur C peut également être considéré comme l'effet des obstacles à l'approche des proies dans la nature. D'une manière générale,les obstacles naturels apparaissent sur les parcours de chasse des loups et les empêchent en fait de s'approcher rapidement et facilement de leurs proies. C'est exactement ce que fait le vecteur C.

En fonction de la position du loup, il peut donner un poids aléatoire à la proie et la rendre plus difficile et plus éloignée pour atteindre les loups ou vice versa.

Pour résumer, le processus de recherche commence par la création d'une population aléatoire de loups gris (solutions candidates) dans l'algorithme GWO. Au fil des itérations, les loups alpha, bêta et delta estiment la position probable de la proie. Chaque solution candidate met à jour sa distance par rapport à la proie. Le paramètre est diminué de 2 à 0 afin de mettre respectivement l'accent sur l'exploration et l'exploitation. Les solutions candidates ont tendance à s'écarter de la proie lorsque $|\vec{A}| > 1$et converge vers la proie lorsque $|\vec{A}| < 1$. Enfin, l'algorithme GWO se termine par la satisfaction d'un critère de fin.

5-Algorithme d'optimisation de Grey Wolf Optimizer pour la sélection automatique des clés de blocage :

L'algorithme d'optimisation Grey Wolf Optimizer a été récemmentproposé pour résoudre les problèmes d'optimisation.Comme mentionné ci-dessus, notre objectif dans ce chapitre estd'adapter l'algorithme GWO au problème de sélection de clé de blocage.Ce dernier peut en effet être modélisé comme un problèmede sélection de caractéristiques. La population initiale est un groupede sous-ensembles de fonctionnalités, c'est-à-dire des clés de blocage.

Ainsi, chaque membre d'une population représente un sousensembleconcurrent de clés de blocage. De plus, les sous-ensemblesn'ont pas nécessairement les mêmes tailles ou éléments. Les populationssont régénérées à l'aide de l'algorithme GWO.L'aptitude de chaque membre d'une population est calculée en utilisantl'approche Record Linkage proposée dans [H.N et al.] de manièreglobale. Dans cette approche, K-Modes est utilisé comme étape d'indexation en regroupant les données en utilisant uniquement les clés de blocage qui sont le sous-ensemble de fonctionnalités actuellement sélectionné.

Les meilleures clés de blocage en termes de fitnessplus dupliqués lorsqu'ils sont utilisés comme attributs de clustering.En conséquence, la fonction de fitness est le paramètre de complétudede la paire (PC). Le PC mesure le nombre de doublons détectés par une approche RL en utilisant les ckés de blocage sélectionnées.

Notre approche proposée peut être résumée par les points suivants :

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

6-Génération des clés candidates :

La liste des clés candidates est celle à partir de laquelle la populationinitiale de l'algorithme GWO sera sélectionnée au hasard. Avant degénérer la liste des clés candidates, une étape essentielle de prétraitementne peut être négligée. Il s'agit, en fait, de nettoyer l'ensembleA. en d'autres termes, il faut éliminer les attributs de mauvaise qualitéde l'ensemble A. Deux paramètres sont utilisés pour calculerla qualité globale d'un attribut.
Premièrement,l'exhaustivité représentele pourcentage de valeur nulle concernant les attributs spécifiés[L.L et al.]. Nous avons utilisé la mesure NBC (Null-basedCompleteness)où l'exhaustivité est mesurée à l'aide de l'équation 3.8 .En utilisant cette méthode, la valeur 1 représente le meilleur résultatet 0 le pire. Tous les attributs qui ont une valeur d'exhaustivité inférieureau seuil prédéfini sont éliminés de la génération de liste declés de blocage candidates.

$$Completness(Attj) = 1 - \frac{number\ of\ Null\ values\ in\ Attj}{Number\ of\ instances} \qquad (3.8)$$

Le deuxième paramètre est la cardinalité d'un attribut. La cardinalitéreprésente le nombre de valeurs distinctes pour un attributspécifié. Dans le processus RL, les attributs à très faible cardinaliténe conviennent pas pour être utilisés comme clés de blocage.

Par exemple, l'utilisation de l'attribut sex comme clé de blocage diviseles données en seulement 2 blocs (M / F). Par conséquent, dansnotre approche, les attributs à très faible cardinalité sont éliminés dela génération de liste de clés de blocage candidates.Une fois que les attributs de mauvaise qualité sont éliminés ; pourchaque dataset D, différentes clés de blocage peuvent être généréesen fonction du domaine de dataset et du type d'attributs.

7-L'algorithme GWO :

- Initialize the grey wolf population Xi (i = 1, 2, ..., n)
- Initialize a, A, and C
- Calculate the fitness of each search agent
- X_α =the best search agent
- X_β =the second best search agent
- X_δ =the third best search agent
    - while (t < Max number of iterations)
        - for eachsearch agent
            - Update the position of the current search agent by above equations
        - end for
        - Update a, A, and C
        - Calculate the fitness of all search agents
        - Update X_α , X_β, and X_δ
        - t=t+1
    - end while

return X_α

Pour voir comment GWO est théoriquement capable de résoudre des problèmes d'optimisation, quelques points peuvent être notés :

- La hiérarchie sociale proposée aide GWO à sauvegarder les meilleures solutions obtenues jusqu'à présent au cours de l'itération.
- Le mécanisme d'encerclement proposé définit un voisinage en forme de cercle autour des solutions qui peut être étendu à des dimensions supérieures sous la forme d'une hypersphère.
- Les paramètres aléatoires A et C aident les solutions candidates à avoir des hyper-sphères avec différents rayons aléatoires.
- La méthode de chasse proposée permet aux solutions candidates de localiser la position probable de la proie.
- L'exploration et l'exploitation sont garanties par les valeurs adaptatives de a et A
- Les valeurs adaptatives des paramètres a et A permettent à GWO de passer en douceur entre l'exploration et l'exploitation.
- Avec A décroissant, la moitié des itérations est consacrée à l'exploration ($|A| \geq 1$) et l'autre moitié est dédiée à l'exploitation ($|A| < 1$).
- Le GWO n'a que deux paramètres principaux à régler (a et C).

8-La structure de GWO :est détaillée dans la figure (6) :

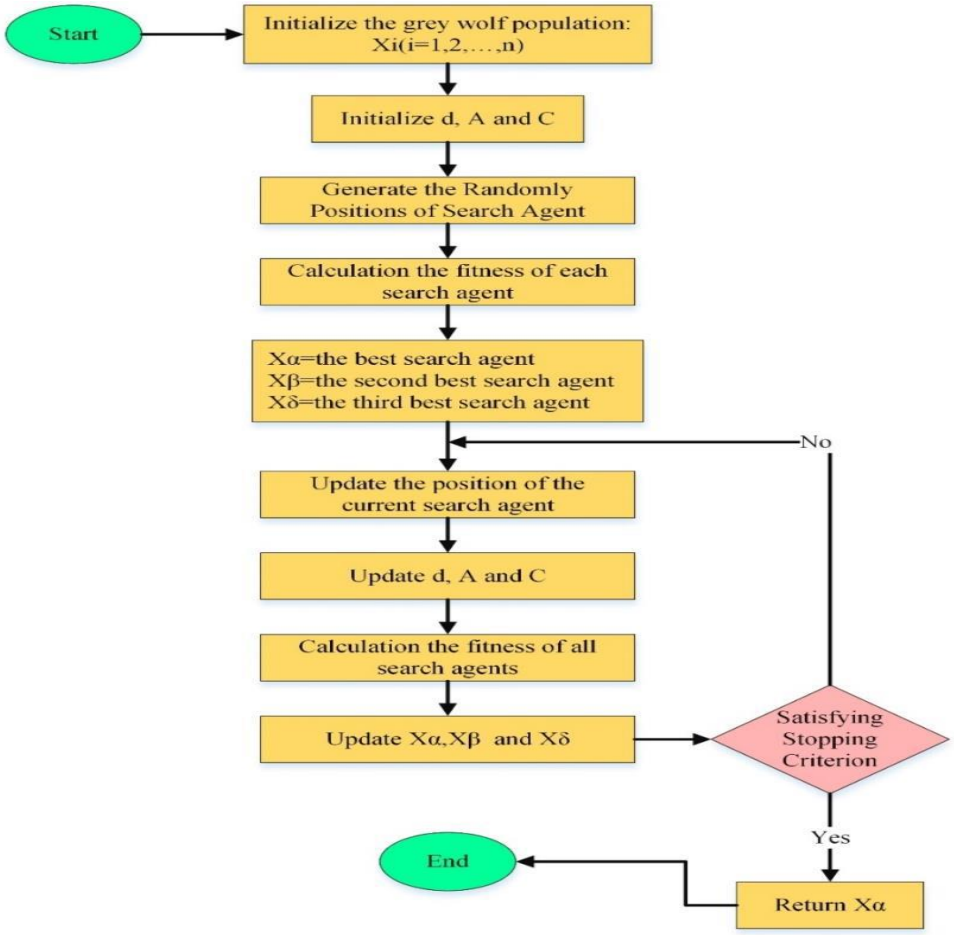


Figure 6 :La structure de GWO

2-4-3-La mise en correspondance des paires d'enregistrements indexes:

Dans le couplage d'enregistrements, la mise en correspondance de paires d'enregistrements indexés fait référence au processus d'identification et de liaison d'enregistrements similaires entre deux ou plusieurs ensembles de données sur la base d'un ensemble de clés de blocage ou de critères d'indexation prédéfinis. L'objectif est de trouver des enregistrements identiques ou similaires qui correspondent à la même entité, comme une personne, une organisation ou un produit, en comparant les valeurs des clés de blocage. Les paires correspondantes résultantes sont ensuite utilisées pour fusionner ou lier les enregistrements, permettant ainsi l'intégration des données, la déduplication et l'amélioration de la qualité des données.

2-4-3-1Bald eagle search:

1-Définition de l'algorithme :

L'algorithme Bald EagleSearch (BES), proposé par H.A. Alsattar [9] en 2020. Il s'inspire du comportement de recherche de nourriture des pygargues à tête blanche et est connu pour sa robuste capacité de recherche mondiale. L'algorithme est divisé en trois parties : sélection de l'espace de recherche, recherche dans l'espace de recherche sélectionné et balayage. Les trois stades de prédation du BES sont illustrés à la figure 7.

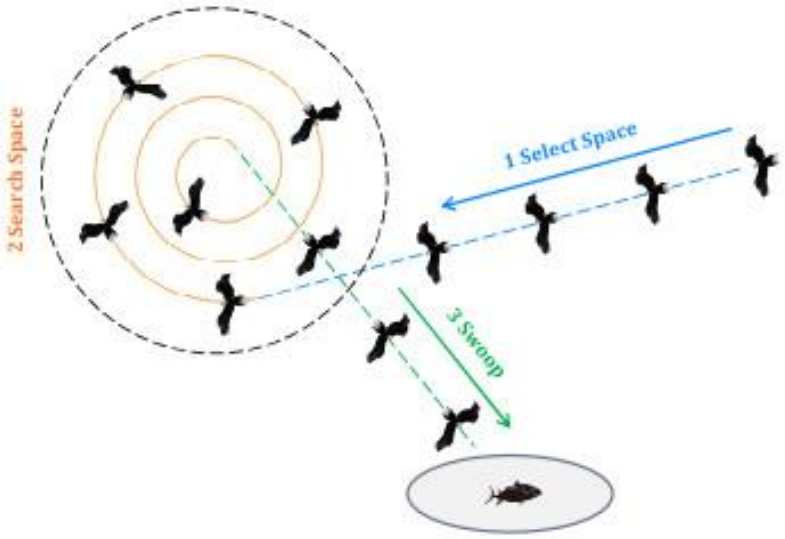


Figure 7 :les étapes de BES [10]

2- Les étapes de BES :

a-Sélectionner un espace :

Lors de l'étape de sélection, le pygargue à tête blanche commence par identifier et Sélectionner l'emplacement le plus prometteur dans l'espace de recherche défini.
Cette sélection est basée sur la quantité de nourriture disponible dans chaque zone.
Le modèle mathématique de cette étape peut être exprimé comme suit :

$$P_{i,new} = P_{best} + \alpha * r(P_{mean} - P_i) \qquad (1)$$

où $P_{i,new}$ est la position mise à jour du i – thPygargue à tête blanche. P_{best} Indique la meilleure position actuelle du pygargue à tête blanche. Le paramètre de changement de position $\alpha \in (1.5,2)$ et r est un nombre aléatoire appartenant à (0, 1).

De plus, P_{mean} est la position moyenne du pygargue à tête blanche. P_i désigne le i –th position du pygargue à tête blanche.

b-Espace de recherche :

Une fois l'étape de sélection terminée, le BES entre dans l'étape de recherche. Au cours de cette étape, l'algorithme imite le comportement de chasse du pygargue à tête blanche en recherchant systématiquement des proies dans la zone préalablement identifiée.L'aigle se déplace dans un mouvement circulaire, élargissanprogressivement sa recherche en forme de spirale. La représentation mathématique de cette étape peut être formulée comme suit :

$$P_{i,new} = P_i + y(i) * (P_i - P_{i+1}) + x(i) * (P_i - p_{mean}) \quad (2)$$

$$x(i) = \frac{xr(i)}{max(|xr|)} \qquad y(i) = \frac{yr(i)}{max(|yr|)} \quad (3)$$

$$xr(i) = r(i) * \sin(\theta(i)) \qquad yr(i) = r(i) * \cos(\theta(i)) \quad (4)$$

$$\theta(i) = a * \pi * rand \text{ et } \quad r(i) = \theta(i) + R * rand \quad (5)$$

Où x(i) et y(i) sont la position du pygargue à tête blanche dans l'espace polaire, ils ont tous des valeurs comprises entre −1 et 1. P_{i+1}est la prochaine position mise à jour du i - th Pygargue à tête blanche.$\theta(i)$et r(i) sont l'angle polaire et le diamètre polaire. a et R sont des paramètres de spirale, allant de : (5, 10), (0,5, 2). rand est un nombre aléatoire, sa valeur est comprise entre 0 et 1.

c-Coup :

Enfin, pendant la phase de plongée, le pygargue à tête blanche plonge pours'attaquer à la proie enfermée dans l'espace de recherche. L'équation (6) décrit le comportement de chasse du pygargue à tête blanche au cours de cette étape.

$$P_{i,new} = rand * P_{best} + x1(i) * (P_i - c_1 * P_{mean}) + y1(i) * (P_i - c_2 * P_{best}) \quad (6)$$

$$X1(i) = \frac{xr(i)}{max(|xr|)} \qquad y1(i) = \frac{yr(i)}{max(|yr|)} \quad (7)$$

$$xr(i) = r(i) * \sin(\theta(i)) \qquad yr(i) = r(i) * \cos(\theta(i)) \quad (8)$$

$$\theta(i) = a * \pi * rand \text{ et } \quad r(i) = \theta(i) \quad (9)$$

Le BES comprend des coefficients de rehaussement c1 et c2 , qui ont chacun une valeurcomprise entre 1 et 2. Pour une compréhension plus complète de l'algorithme BES,l'algorithme 1 fournit le pseudo-code.

3-Pseudo-code de l'algorithme BES :

Input: population size N,dimension dim,maximum iteration number Maxiter, upper and lower bounds up, lb

Output: the optimal solution

1. Random initialization Point P_i
2. Calculate the fitness values
3. **While** (t <Max_iter)

Select stage

4. **For** (point i)
5. Update individual position using Eq (6)
6. **If** $f(P_{new})<f(P_i)$
7. $Pi=Pnew$
8. **End If**
9. **If** $f(P_{new})<f(P_{best})$
10. $Pbest=Pnew$
11. **End If**
12. **End For**

Search stage

13. **For** (point i)
14. Update individual position using Eq (7)
15. **If** $f(P_{new})<f(P_i)$
16. $P_i=P_{new}$
17. **End If**
18. **If** $f(P_{new})<f(P_{best})$
19. $P_{best}=P_{new}$
20. **End If**
21. **End For**

Swoop stage

22. **For** (point i)
23. Update individual position
24. **If** $f(P_{new})<f(Pi)$
25. $P_i=P_{new}$
26. **End If**
27. **If** $f(P_{new})<f(P_{best})$
28. $P_{best}=P_{new}$
29. **End If**
30. **End For**
31. **End While**

4-L'Organigramme de l'algorithme BES :

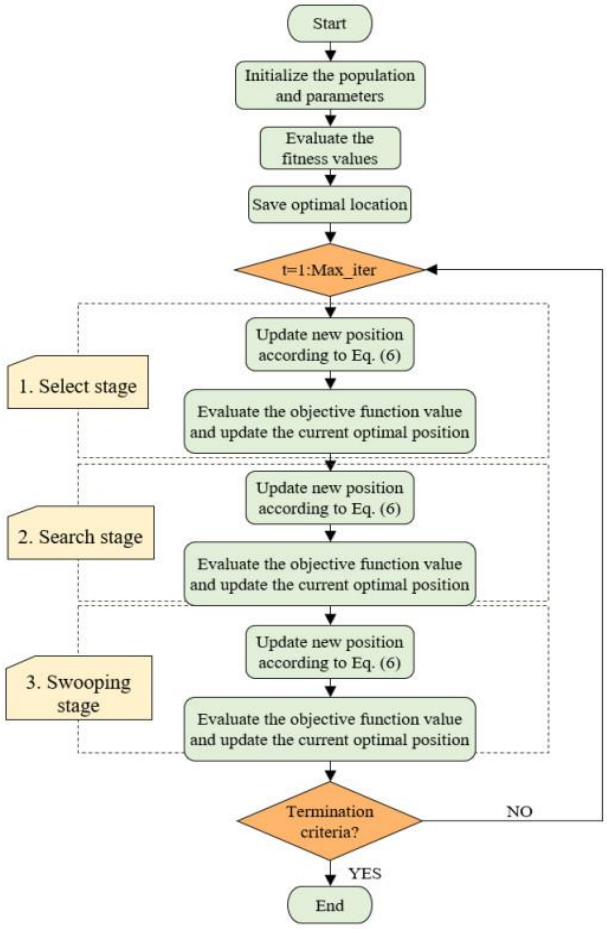


Figure 8 :L'organigramme de l'algorithme BES

2-5- Codage phonétique:

a-Soundex:Les principales étapes du Soundex sont:

- Conservez la première lettre de la chaîne.
- Remplacez toutes les consonnes en utilisant les règles suivantes

0 pour les caractères A,E,H,I,O,U,W,Y.
1 pour les caractères B, F, P, V.
2 pourC,G,J,K,Q,S,X,Z.
3 pourD,T.
4 pourL
5 remplace M,N
6 remplace
le caractére
R.

Dans le cas où la chaîne est trop courte, l'algorithme complète les trois chiffres après le premier caractère par des zéros. .[11]

b-NYSIIS(Système d'identification et de renseignement de l'État de NewYork):

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 en D).

2-6-Recherche de motifs:

a-Distance d'édition:

La distance ou distance de Levenshtein est une mesure de la similitude entre deux cordes, proposée en 1965 par Vladimir Levenshtein. C'est l'une des métriquesles plus courantes pour comparer deux séquences de caractères.

En général, il est défini comme le nombre d'insertions, de suppressions et de mises à journécessaires pour transformer une chaîne en une autre. L'exemple suivant montrecommentcalculerlescoûts depassaged'un mot àun 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

Tableau2:les calculs des coûts de passage d'un mot à un autre

Donc la distance d'édition entre le deux mots (Intention, Exécution) est la somme des coûts pour transformer la première chaîne en la seconde qui est égale à Cinq.Les étapes qui sont représentés dans l'exemple ne sont pas la seule solution pour transformer la première chaîne dans la seconde mais sont celles qui ont le moins de coût.

b-Jaro-Winkler:

$$Jaro\_Sim(s_1, s_2) = \begin{cases} 0, & m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), & otherwise \end{cases}$$

- 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 échellede préfixes P afin de privilégier les chaînes de caractères qui commencent par lemême préfixe L pour une longueur maximale de quatre. La similarité de Jaro-Winklerest définie comme suit : [12]

jaroWinlker_Sim (s1 , s2) = Jaro_Sim(s1 , s2) + LP(1 - Jaro_Sim(s1 , s2))

Où:

- Jaro_Sim(s1 ,s2)estlasimilaritéJaroentreleschaînes decaractères.
- Lestlalongueurdupréfixe.
- 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 similitude entredeux jeux d'échantillons qui peuvent être le cas de Strings.
Afin de mesurer le Distance de Jaccard,il faut d'abord calculer le coefficient de Jaccard qu'est défini comme suit : [13]

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

Sil' on fait,la distance de Jaccard n'est obtenue que par la soustraction du coefficient de Jaccard de 1.

$$Jaccard_{distance}(A, B) = 1 - Jaccard(A, B)$$

2-7-Conclusion:

Le couplage d'enregistrements continue de prendre de l'importance entant qu'activité fondamentale dans les organismes statistiques.Le nombre de listes administratives et de fichiers commerciaux disponibles a augmenté de façon exponentielle et offre aux organismes statistiques la possibilité d'accumuler des informations grâce au couplage d'enregistrements pour appuyer la production de statistiques officielles.

Chapitre03:Implémentation et Expérimentation

3-1-Introduction:

Dans ce chapitre, nous décrivons l'environement et les technologies utilisées pour l'implémentation de mon application, puis je présente mon contributiondans le domaine du couplage d'enregistrements. Une solution est proposée pourchacun des défis abordés dans le chapitre précédent. Enfin je présentequelquescapturesd'écran.

3-2-Environnement de travail:

L'environnement de développement est unfacteur important qui doit être détaillé pour connaîtredans 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 mon application sur une machine hp avec un processeur Intel(R) Core TM i5-7200U CPU, une vitesse de 2.71 Ghz et une capacité mémoire de 8 GB.
Système d'exploitation Windows 10 professionnel de64bits.

Langage utilisé :Java.
IDE utilisée : NetBeans

Outil de development:

3-2-1NetBeansIDE:

NetBeans IDE est un environnement de développement intégré gratuit et à code source ouvert destiné au développement d'applications sous Windows, Mac,Linuxet Solaris.L'environnement IDE simplifie le développement d'applications Web,d'entreprise, de bureau et mobiles utilisant les plates-formes Java et HTML5. Iloffre également une assistance pour le développement d'applications PHP etC/C++.
NetBeans constitue par ailleurs une plateforme qui permet le développement d'applications spécifiques (bibliothèqueSwing(Java)). L'IDE NetBeans s'appuie sur cette plateforme.

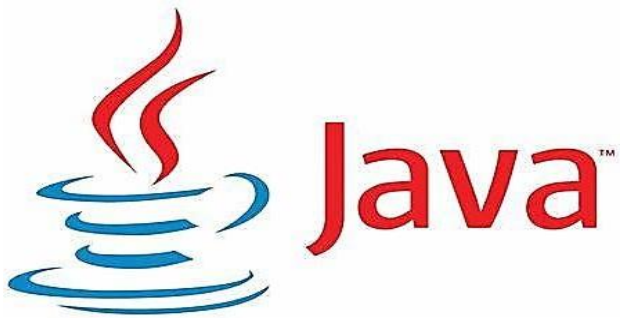
3-2-2-Langage de programmation:

3-2-2-1-Définition Java:

Java est un langage de programmation et une plateforme informatique qui ontété 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 cessede 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 don't 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. 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.



3-2-2-2-JavaFXIntégration:

JavaFX est une bibliothèque graphique intégrée dans le JRE et le JDK de Java.Oracleladécritcomme«TheRichClientPlatform»,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 desarchitectures avec des paradigmes objet, et aussi de pouvoir utiliser le typagestatique.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'utilisationdu «scenegraphe».

Cette présentation ne fait pas dans le bling-bling, même si JavaFX est doué pour cela,enpréférant 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.

3-2-2-3-JavaFXSceneBuilder(SceneBuilder):

Vous permet de concevoir rapidement des interfaces utilisateur d'application JavaFX enfaisant 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 arrière plan.SceneBuilder peut être utilisé comme un outil de conception autonome, mais ilpeut 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'EDI NetBeans,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 SceneBuilder, d'exécuter les exemples SceneBuilder et de générer un modèle pour le fichier source du contrôleur.

3-2-3Les avantages:

- ✔Basé sur Java(JavaSEetME).
- ✔Utilisable sur tous les écrans:navigateurs,mobile,TV,etc.
- ✔Open Source.
- ✔Déploiementsur navigateur et ordinateur de bureau "Desktop" sans modification.

- ✔Collaboration designers et développeurs. Possibilité d'intégrer des codes en Java et JavaFX
- ✔Moins de code pour générer une interface et des composants Graphiques(NSY).

3-3-Présentationdel'application:

->L'interface principale de l'application:

La figure 9 montre la page d'accueil de notre application qui nous permet de charger l'ensemble des données utilisé dans notre système.

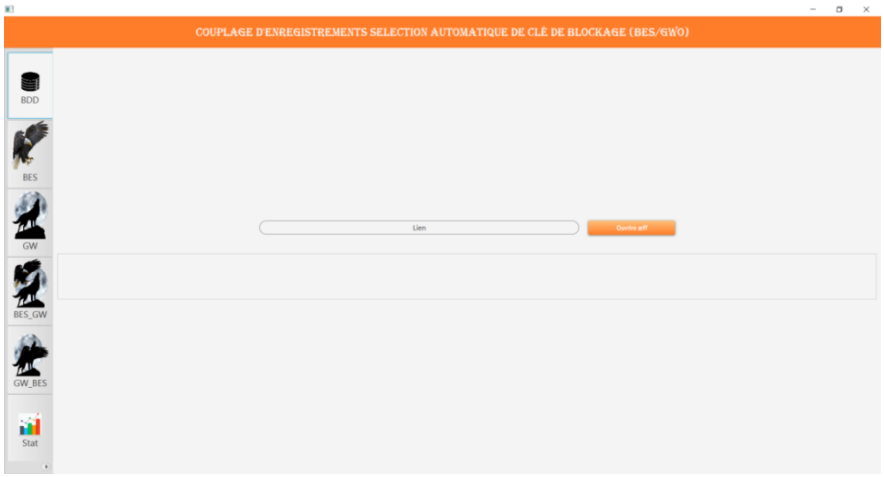


Figure 9:L'interface principale de l'application

->selectionné un data set:

L'ensemble des données restaurant est chargé dans une fichier.arf.

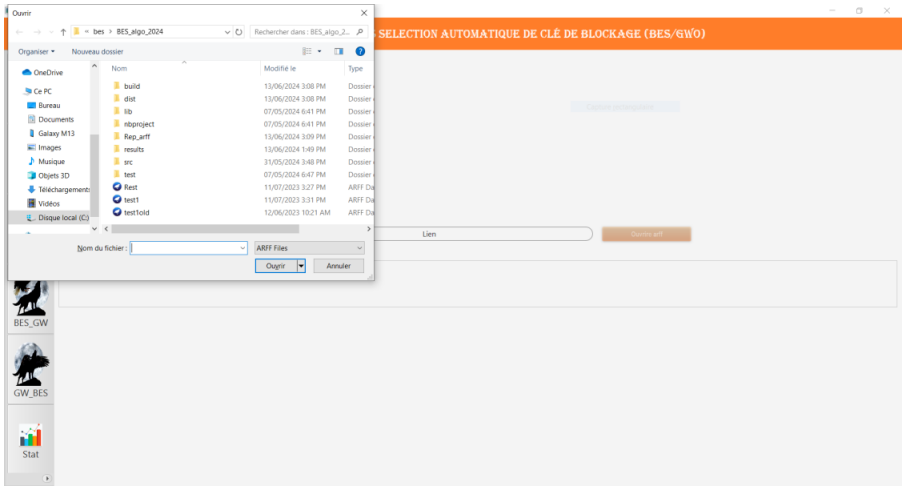


Figure10:Sélection de fichier data test1.arff

-> Chargement de l'ensemble de donnees dans l'application :
La figure 11 montre le chargement de Data set Restaurant dans notre application afin qu'on puisse travailler avec :

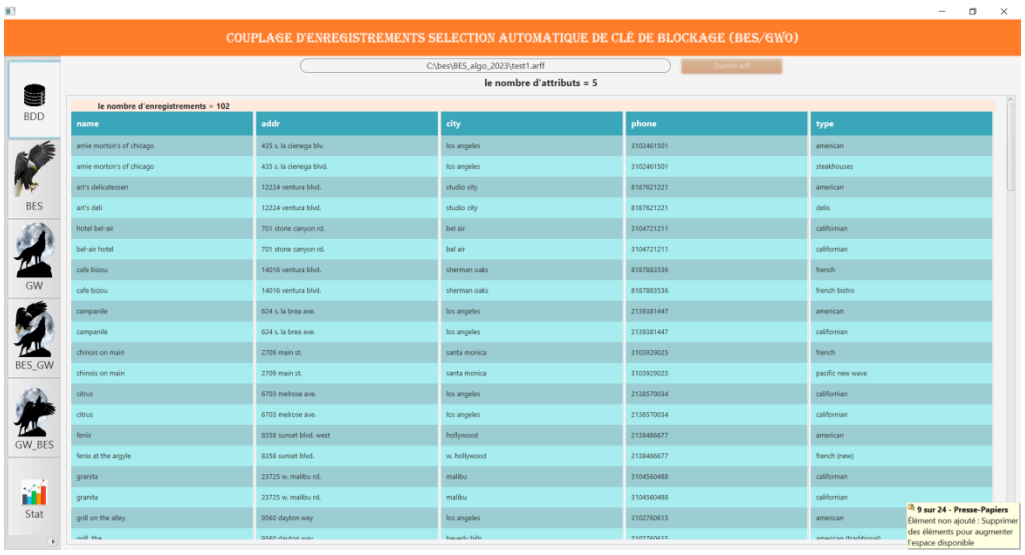


Figure11:Affichage des données

3-3-1-BES:

->selection des paramétres de l'éxecution:

a-les paramétres de K_mode:nombre des blocks et nombre d'itération

b-les paramétres de phonetique:soundex ou NYSIIS

c-les paramétres de distance:EditDistance ou jaccardSimilarity ou javaWinklerDistance

d-les paramétres de BES: max des attributs et min des attributs at max des itérations et le nombre de population

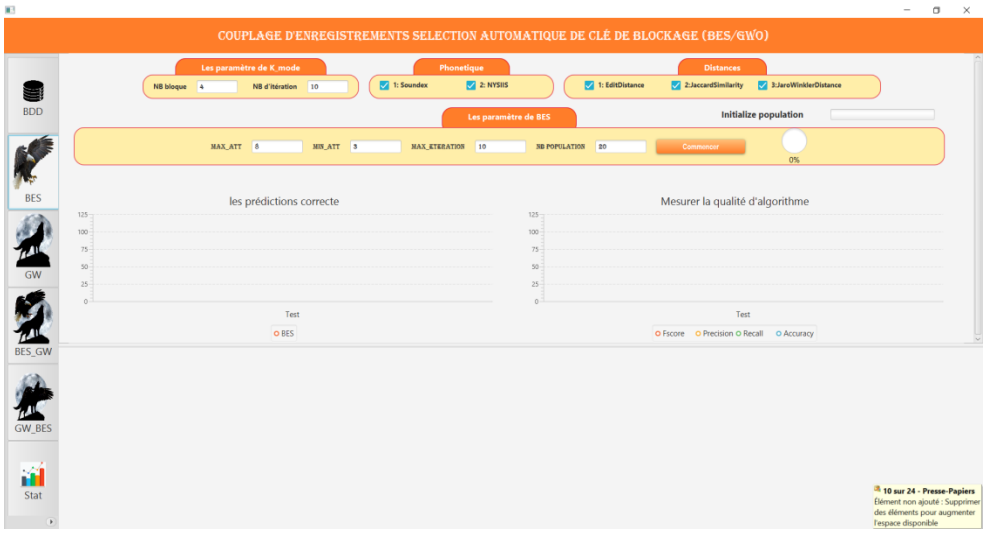
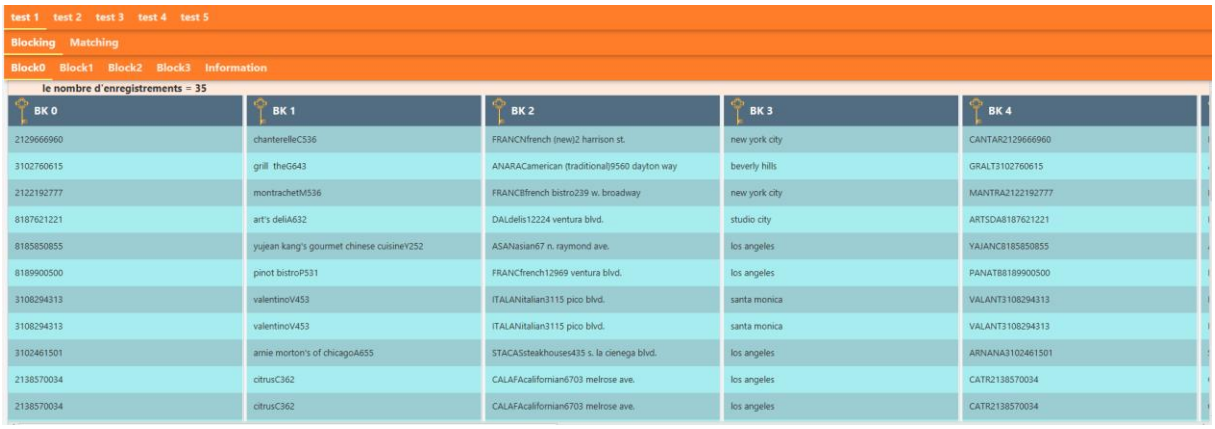


Figure12:Sélection les paramétres(K_mode,phonetique,distance,BES)

->Lancement de l'éxecution:

1-création des blocks:



BK 0	BK 1	BK 2	BK 3	BK 4
2129666960	chanterelleC536	FRANCNfrench (new)2 harrison st.	new york city	CANTAR2129666960
3102760615	grill theG643	ANARACamerican (traditional)9560 dayton way	beverly hills	GRALT3102760615
2122192777	montrachetM536	FRANCBfrench bistro239 w. broadway	new york city	MANTRA2122192777
8187621221	art's deliA632	DALdelis12224 ventura blvd.	studio city	ARTSDA8187621221
8185850855	yujean kang's gourmet chinese cuisineY252	ASANasian67 n. raymond ave.	los angeles	YAJANC8185850855
8189900500	pinot bistroP531	FRANCfrench12969 ventura blvd.	los angeles	PANATB8189900500
3108294313	valentinoV453	ITALANitalian3115 pico blvd.	santa monica	VALANT3108294313
3108294313	valentinoV453	ITALANitalian3115 pico blvd.	santa monica	VALANT3108294313
3102461501	arnie morton's of chicagoA655	STACASsteakhouses435 s. la cienega blvd.	los angeles	ARNANA3102461501
2138570034	citrusC362	CALAFAcalifornian6703 melrose ave.	los angeles	CATR2138570034
2138570034	citrusC362	CALAFAcalifornian6703 melrose ave.	los angeles	CATR2138570034

Figure13:Création des block

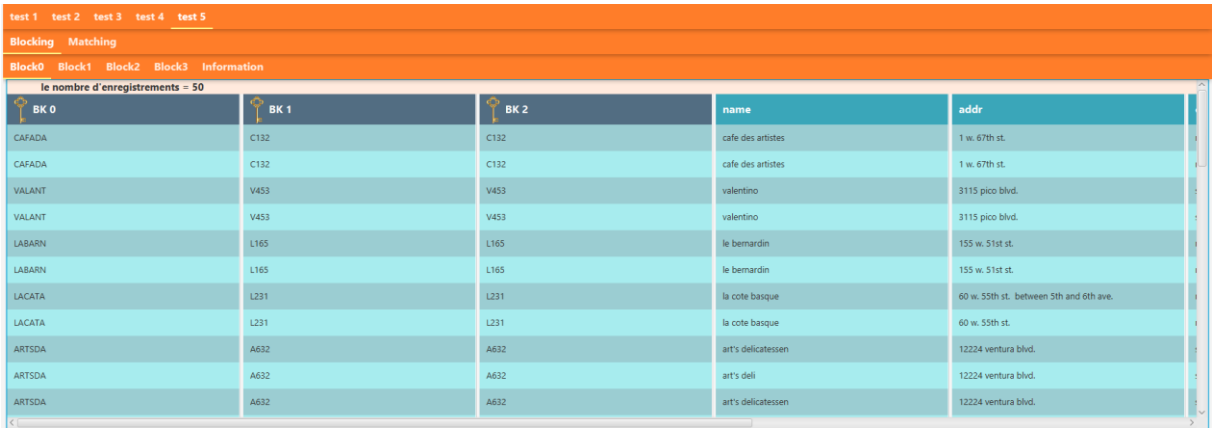


Figure14:Block 0 de test 5

->Information des blocs

La figure 15 montre les informations principale utilisé pour chaque bloc et matchMetric représente la quantité de changement pour diviser les enregistrements en cluster, où la valeur zéro représente la répétition de les même enregistrements dans cluster a chaque itération.

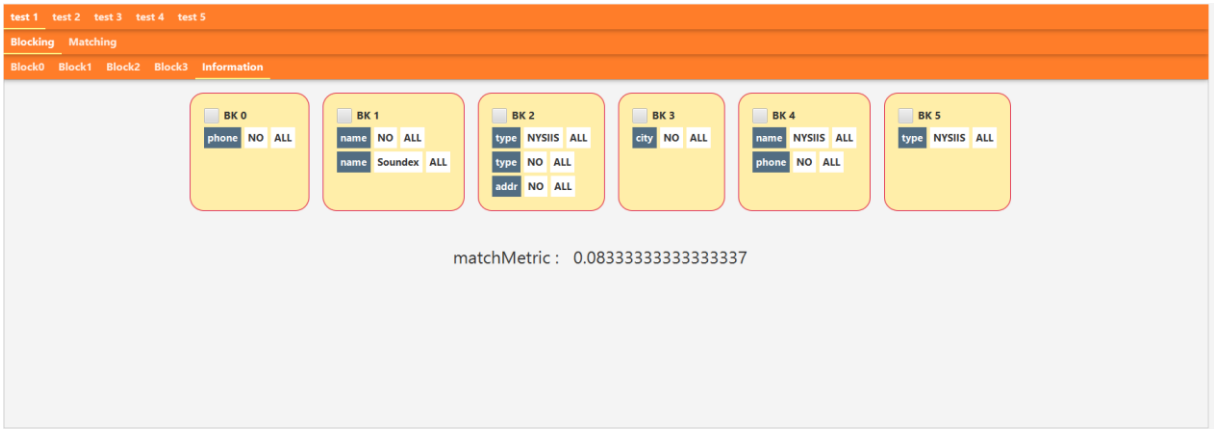


Figure 15: Les clés de blocage de test1

->Résultat de matching:

La figure 16 montre les résultat obtenues aprés le Matching. Comme on peut voir 4 résultats déférentes obtenues dans la matrice de confusion :

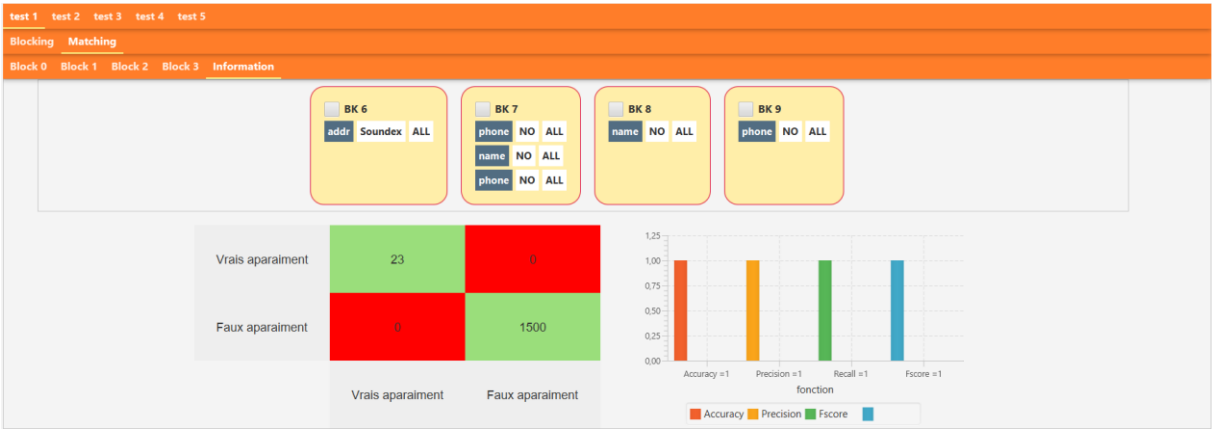


Figure16:Les clés de matching avec la matrice de confusion

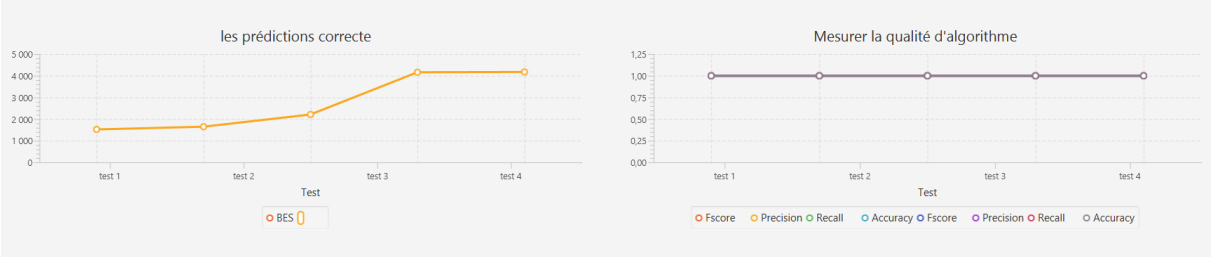


Figure17:Les résultats graphique du record

3-3-2GWO:

->sélection des paramétres de l'éxecution:
a-les paramétres de K_mode, phonetique et distance
b-les paramétres de GWO(max attribute, minattribute, maxitération et nombre population) .

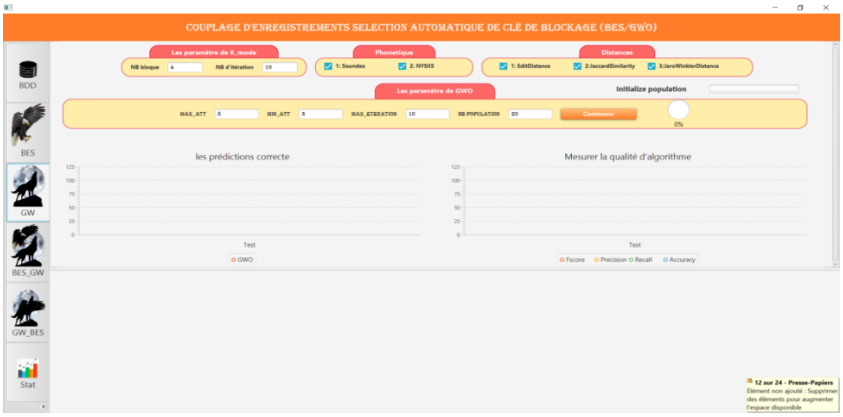


Figure 18:Séléction des paramétres de(K_mode,phonetique,distance et GWO)

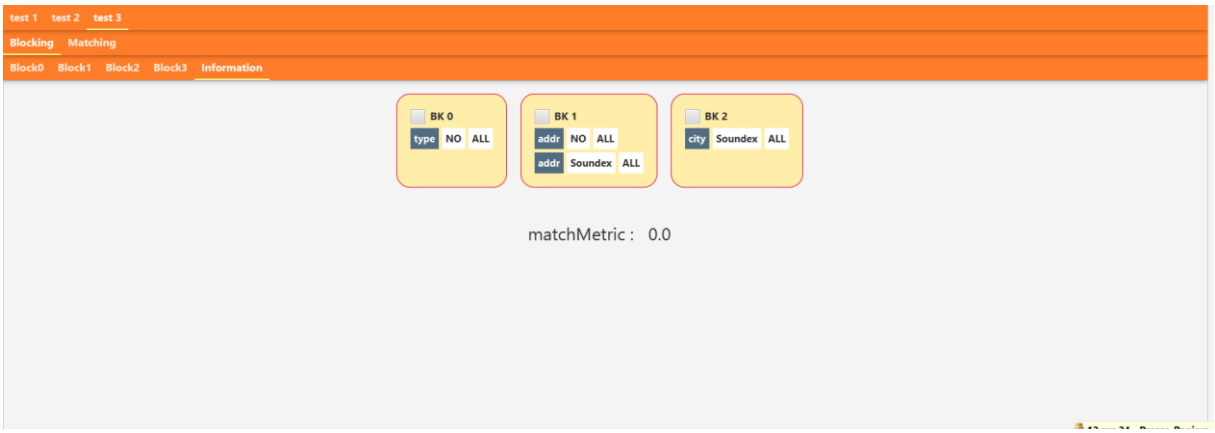


Figure19:Les clés de blocage de test3

Figure 20:Les clés de matching avec la matrice de confusion

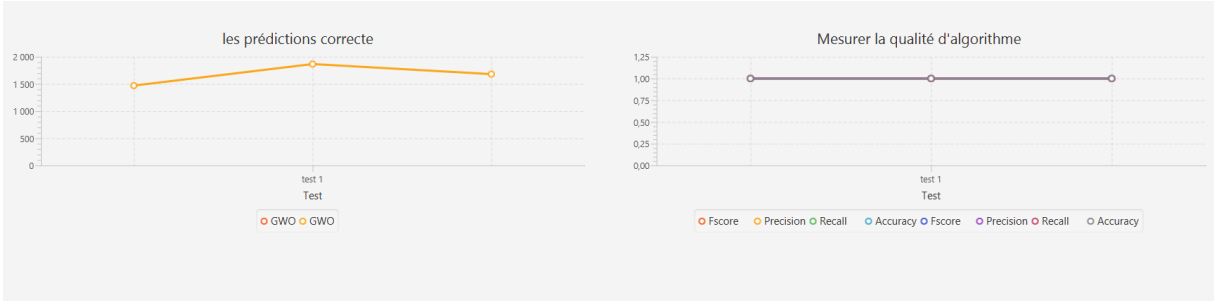


Figure 21:Résultats du record

3-3-3BES_GWO

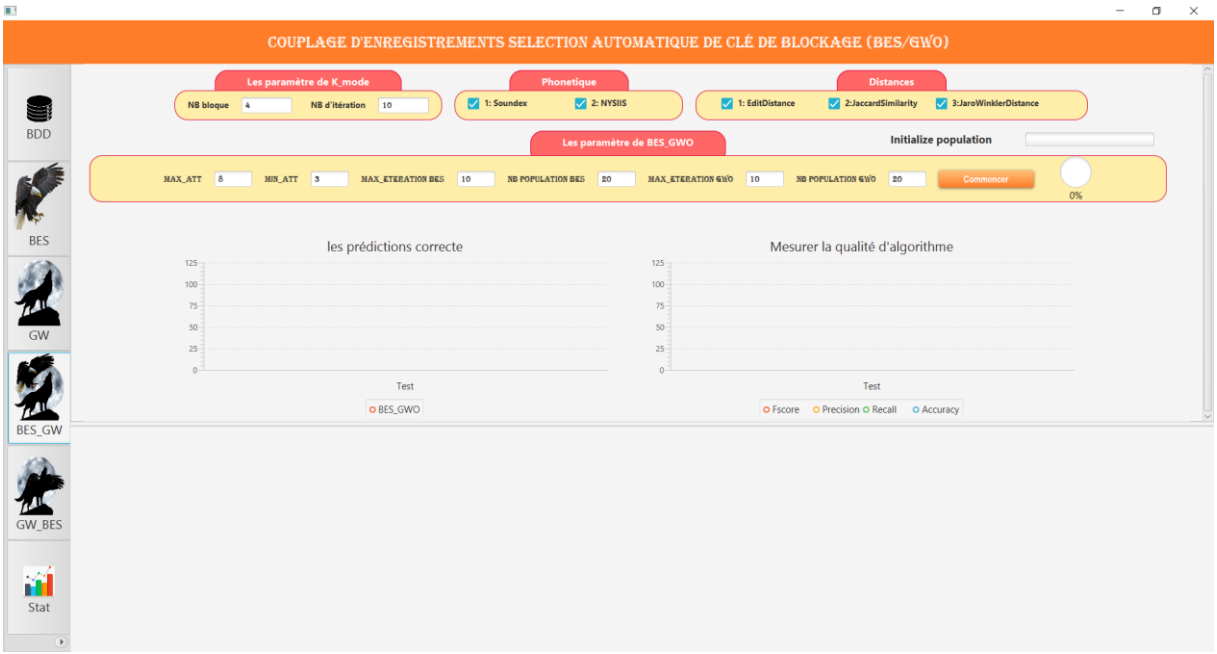


Figure 22:Sélection des paramétres de BES etGWO

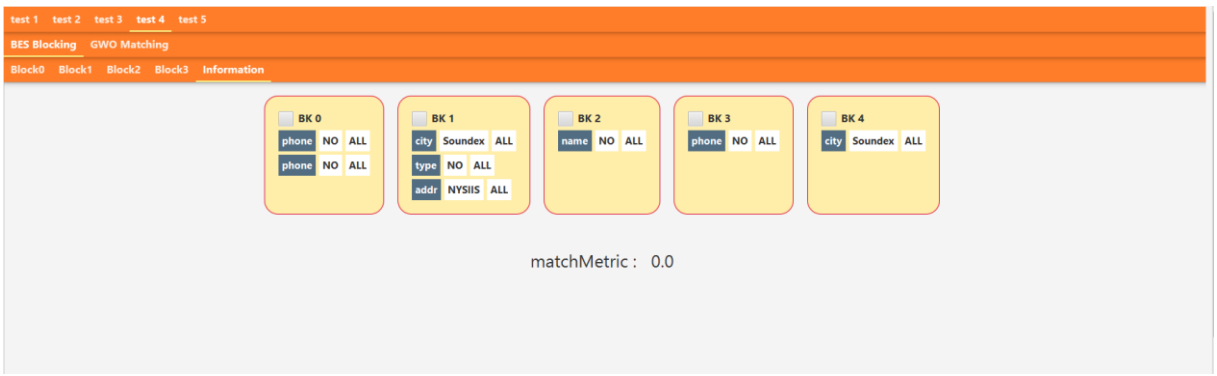


Figure 23:Les clés de blocage de test4

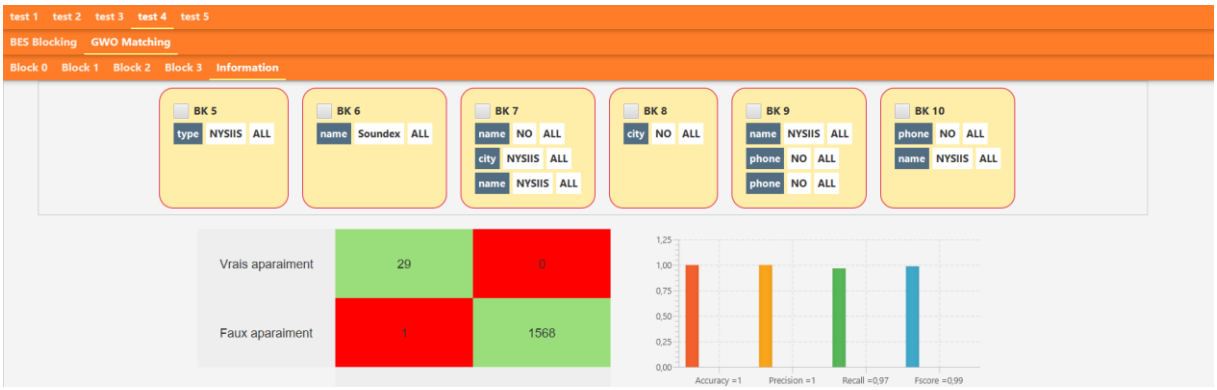


Figure24:Les clés de matching avec la matrice de confusion

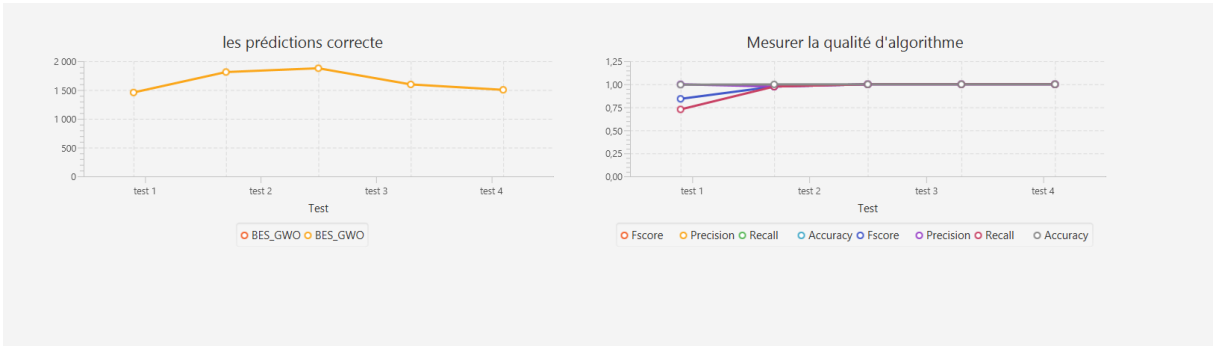


Figure 25:Résultats du record

3-2-4-GWO-BES:

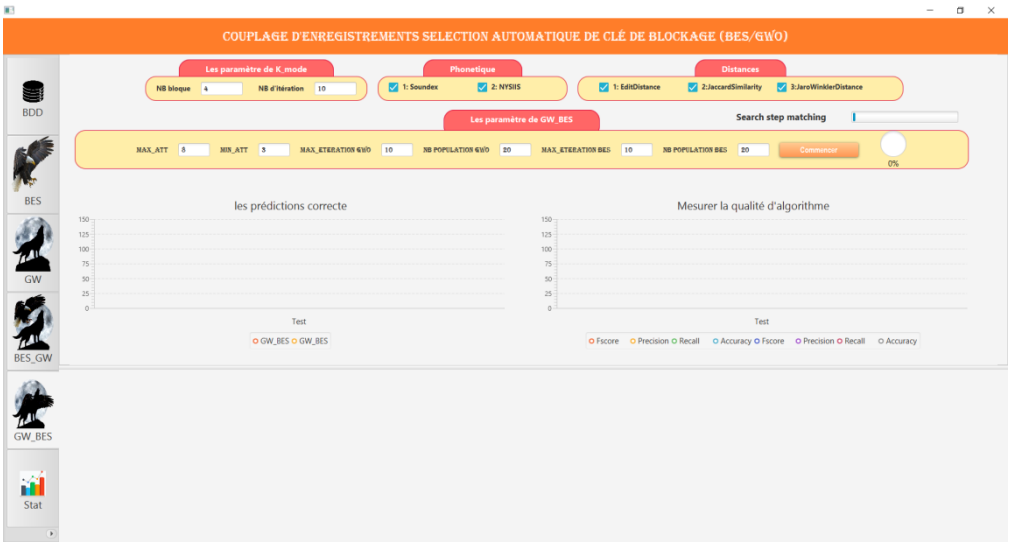


Figure 26:Sélection des parémétres de GWO et BES

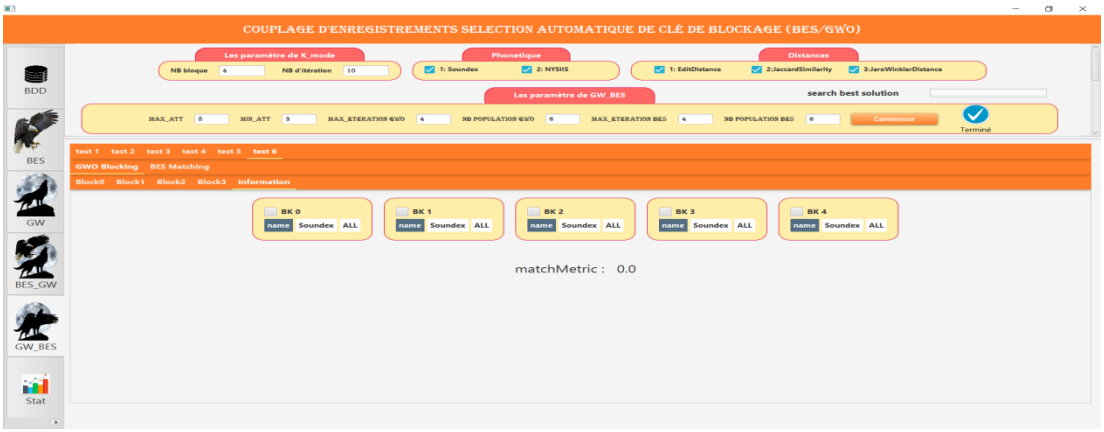


Figure 27:Les clés de blocage de test6

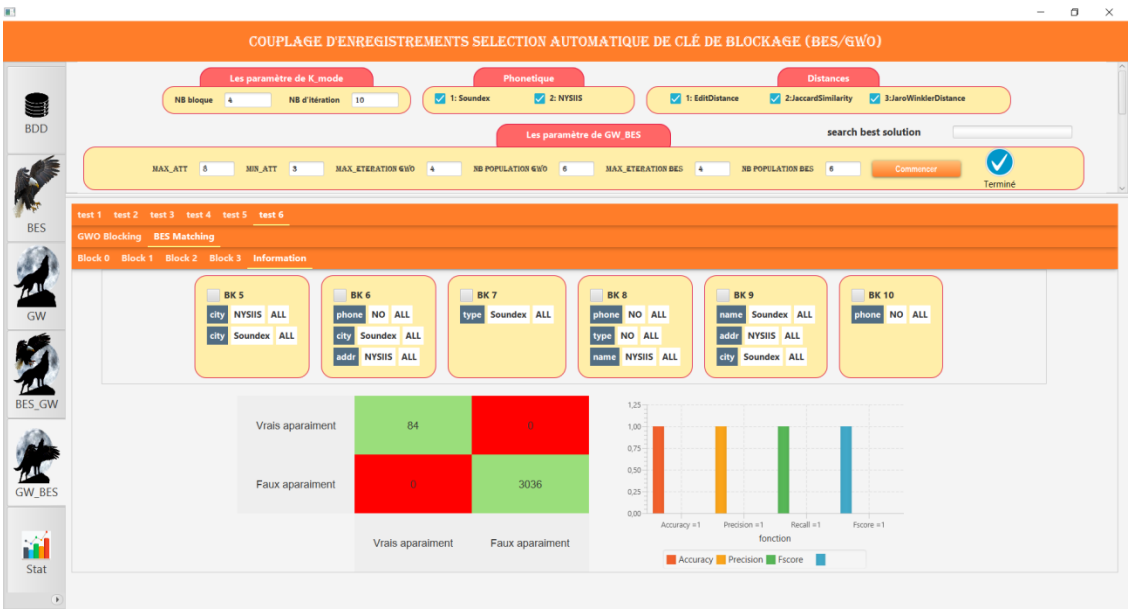


Figure28:Les clés de matchinget matrice de confusion

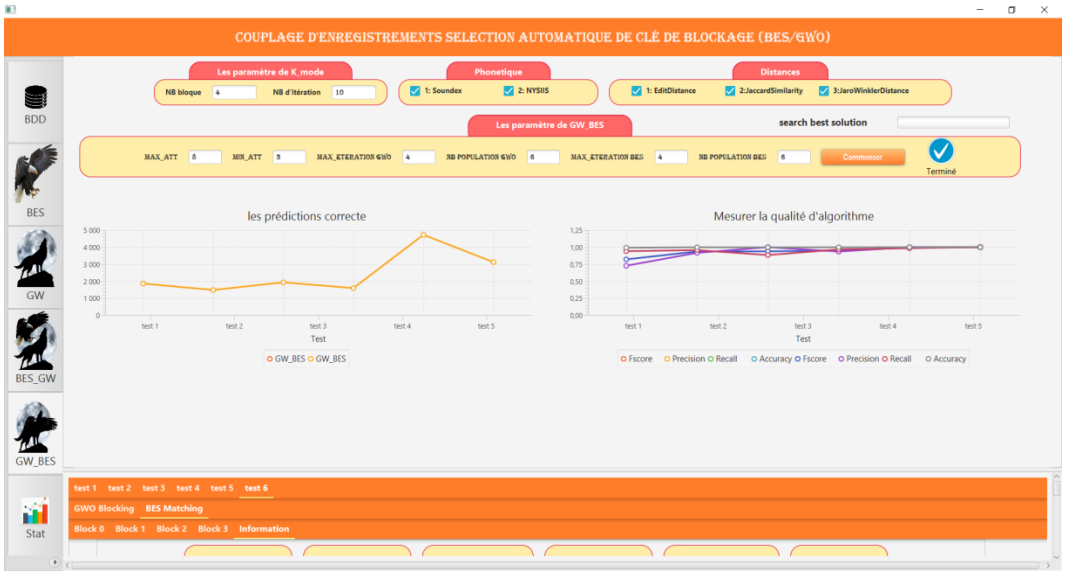


Figure29:Résultats du record

->Comparaison:

La figure 29 montre la comparaison entre les quatre algorithme :

On Remarque dans l'algorithme 4 que nous obtenu d'excellents résultats

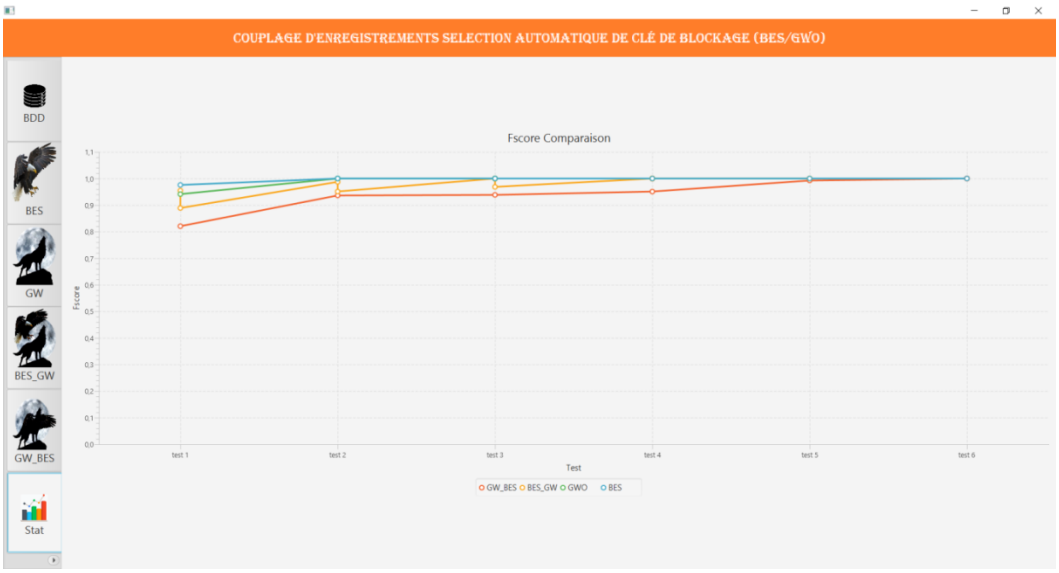


Figure 30:Comparaison entre les 4 algorithme

3-4- Evaluation :

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

1-Accuracy :

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

$$A = T_p + T_n \setminus T_p + F_n + F_p + T_n$$

2- Précision :

Cette métrique est utilisée pour mesurer la précision des comparaisons.

$P = T_p \setminus T_p + F_p$

3-Recall :

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

$R = T_p \setminus T_p + F_n$

4-F-Score

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

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

Avec :

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

Tn (Vrais negatifs) : paires qui apparaissent dans différents clusters a 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 vérite terrain. Connu sous le nom de fausses correspondances.

Fn (Faux negatifs) : 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 manques.

3-5- Conclusion:

Dans ce chapitre, nous allons présenter une approche de méthode hybride pour un couplage d'enregistrements efficace.Les expériences sur deux ensembles de données du monde réel ont montré l'efficacité de l'algorithme k-modes, BES et GWO pour sélectionner le meilleur sous-ensemble de clés de blocage et de clés de matching.J'ai consacré à la réalisation et implémentation des différentes procédures, ainsi que l'évaluation du résultat obtenu.

Coclusion générale

« En conclusion, la méthode hybride de couplage d'enregistrements présentée dans cette étude démontre une amélioration significative en termes de précision et d'efficacité par rapport aux méthodes de couplage traditionnelles. En combinant les atouts de différentes techniques de couplage, telles que les approches déterministes, probabilistes et basées sur l'apprentissage automatique, le La méthode hybride est capable de gérer efficacement des ensembles de données complexes avec différents niveaux de qualité et de complexité.

Les résultats de cette étude montrent que la méthode hybride est capable d'atteindre des niveaux élevés de précision et de rappel, surpassant dans de nombreux cas les méthodes de liaison individuelles. La capacité de la méthode à s'adapter à différents ensembles de données et scénarios en fait un outil polyvalent et fiable pour les tâches de couplage d'enregistrements.L'utilisation d'une approche hybride permet également de tirer parti des connaissances et de l'expertise spécifiques à un domaine, permettant ainsi l'incorporation de sources de données et de fonctionnalités supplémentaires susceptibles d'améliorer la précision des liens. En outre, la flexibilité et l'évolutivité de la méthode la rendent bien adaptée aux tâches d'intégration de données à grande échelle, pour lesquelles les méthodes de liaison traditionnelles peuvent s'avérer prohibitives en termes de calcul.

Dans l'ensemble, la méthode hybride de couplage d'enregistrements présentée dans cette étude a le potentiel d'améliorer considérablement la précision et l'efficacité des tâches d'intégration des données, permettant ainsi aux chercheurs et aux praticiens de prendre des décisions plus éclairées et d'acquérir de nouvelles connaissances à partir des données liées. Alors que les données continuent de croître en taille et en complexité, le développement de méthodes de liaison innovantes et efficaces, comme cette approche hybride,sera crucial pour libérer tout le potentiel de la recherche et des applications basées sur les données. »

Bibliographie

Bibliographie:

[1] Franck Rgnier-Pcastaing, Michel Gabassi, Jacques Finet,Enjeux et mthodes dela gestion des donnes,livre,(2008).

[2] JEMM research, DES DONNES QUALIT :Exploitez lecapital de votre organisa-tion ,livre blanc,(janvier 2008).

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

[4] https://recordlinkage.readthedocs.io/en/latest/about.html

[5] https://www.ncbi.nlm.nih.gov/sites/books/NBK253312/

[6] Emary E, Zawbaa H.M, Ghany K.K.A., et A.E.and Pârv B Hassanien. Firefly optimization algorithm for feature selection. Dans Actes de la 7e Conférence desBalkans sur l'informatique, page 26. ACM.

[7] An Automatic Blocking Keys Selection For Efficient Record Linkage , International Journal of Organizational and Collective Intelligence ,Volume 11 •Issue 1 • January-March 2021

[8] Muro C, Escobedo R, Spector L, Coppinger R. Wolf-pack (Canis lupus) hunting strategies emerge from simple rules in computational simulations. Behav Process 2011;88:192–7.

[9] U. Goel, S. Varshney, A. Jain, S. Maheshwari, A. Shukla, Three dimensional path planning for UAVs in dynamic environment using glow-worm swarm optimization, *Proc. Comput. Sci.*, **133** (2018), 230–239. https://doi.org/10.1016/j.procs.2018.07.028 doi: 10.1016/j.procs.2018.07.028

[10] Y. Zhang, Y. Zhou, G. Zhou, Q. Luo, B. Zhu, A curve approximation approach using bio-inspired polar coordinate bald eagle search

[11] algorithm, *Int. J. Comput. Intell. Sys.*, **15** (2022), 30. https://doi.org/10.1007/s44196-022-00084-7 doi: 10.1007/s44196-022-00084-7

[12] TN Gadd. Phonix : The algorithm. Program, 24(4) :363– 366,. Köpcke H, Thor A, et Rahm E. Evaluation of entity resolution approaches on real-world

[13] David Holmes et M.Catherine McCabe. Improving precision and recall forsoundex retrieval. Dans Information Technology : Coding and Computing, 2002.Proceedings. International Conference on, page 22–26. IEEE.

[14] Benkhaled hamid naceur.Data Quality in the Big Data context. Technologiede l'information 15/02/2021, page 42