RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE Dr. TAHAR MOULAY SAIDA

FACULTE : TECHNOLOGIE

DEPARTEMENT : INFORMATIQUE

MÉMOIRE DE MASTER

Option :

Réseaux informatique et systèmes répartis

Sélection semi-automatique de clé de blocage Pour le couplage d'enregistrement

---

Présenté par :

Mr BESSADAT Mohamed

Mr KHETTAB Mokhtar

Encadré par :

Mr BENYAHIA miloud

Année Universitaire 2020-2021

Dédicace

Je dédie ce modeste travail :

A toute ma famille

Spécialement ma petite belle fille yasmine hibat erahmane

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

MOKHTAR

Dédicace

Je dédie ce modeste travail :

A toute ma famille

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

MOHAMED

Résumé

Tous les ans, plusieurs organisations du monde entier subissent d'énormes pertes à cause des problèmes de qualité des données. Ainsi ces dernières sont désormais plus conscientes de l'importance de la qualité des données, et comme conséquence, beaucoup d'efforts sont investi pour améliorer la qualité des données stockées.

Parmi les principaux processus dans ce domaine est le Record Linkage (RL), également connu sous le nom de résolution d'entité.

Il s'agit d'un processus de détection des doublons qui font référence à la même entité réelle dans un ou plusieurs ensembles de données. Dans ce processus, l'étape la plus importante est celle du blocage, celle-ci vise à réduire la complexité quadratique du processus en divisant les données en un ensemble de blocs. Ainsi, la mise en correspondance n'est effectuée qu'entre les enregistrements du même bloc.

Par contre, le choix des meilleures clés de blocage pour diviser les données est une tâche difficile et, dans la plupart des cas, elle est effectuée par un expert du domaine.

Le but de notre travail est d'utilise une technique de clustering proposée par Huang comme alternative à l'analyse de clustering pour les données catégorielles uniquement, cette technique est l'algorithme k-Mode.

Les résultats que nous avons obtenus à partir des expériences sur des data sets du monde réel ont démontré l'efficacité de notre choix où le k-Mode a donné des résultats remarquables pour la sélection d'entités a renvoyé les meilleures clés de blocage.

Mots clés : Qualité des données, Record linkage, clés de blocage, technique de clustering.

ABSTRACT

Every year, many organizations around the world experience huge losses due to data quality issues. As a result, the latter are now more aware of the importance of data quality, and as a consequence, a lot of effort is invested to improve the quality of the data stored.

Among the main processes in this area is Record Linkage (RL), also known as entity resolution.

This is a process of detecting duplicates that refer to the same real entity in one or more datasets. In this process, the most important step is that of blocking, which aims to reduce the quadratic complexity of the process by dividing the data into a set of blocks. Thus, matching is only done between records in the same block.

On the other hand, choosing the best blocking keys to divide data is a difficult task and, in most cases, it is done by an expert in the field.

The aim of our work is to use a clustering technique proposed by Huang as an alternative to clustering analysis for categorical data only, this technique is the k-Mode algorithm.

The results we obtained from experiments on real-world datasets demonstrated the effectiveness of our choice where k-Mode gave remarkable results for the selection of entities that returned the best blocking keys.

Keywords : Data quality, Record linkage, blocking keys, clustering technique.

Table des matières

Introduction Général

Ces dernières années, le monde assiste à une explosion massive du volume de données. Plus précisément, après l'adoption des Smartphones et des médias sociaux qui génèrent une énorme quantité de données au quotidien. Des organisations du monde entier se sont retrouvées dans le besoin d'intégrer leurs propres données provenant de diverses sources dans différents formats. Ces données doivent être intégrées afin de faciliter le processus d'analyse des données et d'en extraire des informations utiles. Cependant, l'intégration des données peut devenir un processus très long en raison de problèmes de qualité des données, tels que des valeurs en double, des valeurs manquantes et des problèmes d'intégrité référentielle. Les parties prenantes sont désormais plus conscientes de l'importance de la qualité des données. Beaucoup d'argent est investi pour améliorer la qualité des données stockées. Le couplage d'enregistrements (Record Linkage) est l'une des tâches les plus importantes dans le domaine de la qualité des données. RL est défini comme le processus d'identification des enregistrements qui représentent la même entité du monde réel lors de la fusion de différentes sources de données. Lorsque le processus RL est exécuté sur une seule base de données, il peut être appelé processus de déduplication (Sarawagi et Bhamidipaty, 2002). Récemment, le processus RL a été exploité dans plusieurs domaines pour de multiples objectifs tels que la préservation de la vie privée, la suppression des doublons des citations bibliographiques, la comparaison des prix et la détection des fraudes.

Problématique

La meilleure façon de détecter tous les tuples qui font référence à la même entité du monde réel est de comparer chacun d'entre eux dans l'ensemble de données à tous les autres. Cependant, dans le cas d'un très grand ensemble de données, le produit cartésien pourrait aboutir à un nombre inacceptable de comparaisons. Par exemple, en appliquant le processus Record Linkage sur la base de données A et B, chacun contient 2 millions d'enregistrements finira par faire 4 milliards d'opérations d'appariement, ce qui n'est pas raisonnable. Maintenant la question qui se pose est : Y'a-t-il une technique de blocage qui facilite cette comparaison ? et comment peut-on l'implémenter ?

Organisation du mémoire

Cette étude est structurée en chapitres et organisée comme suite :

- **Chapitre 1 :** la qualité des données
- **Chapitre 2 :** Record Linkage
- **Chapitre 3 :** présente la méthode choisie et les outils utilisés pour la sélection semi-automatique de clés de blocage, plus précisément en utilisant la méthode K-modes. Enfin ce chapitre décrira l'environnement de travail et les expérimentations faites. Nous terminerons ce mémoire par une conclusion générale qui servira de base à un futur travail sur le même thème.

Chap 1 La qualité des données

1.1 Introduction

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

1.2 la qualité des données

1.2.1 Définition

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

1.2.2 Les critères de la qualité des données

1. **Les critères intrinsèques**

(a) **L'unicité :** L'unicité est le fait qu'une entité du monde réel ne soit représentée que par un seul et unique objet métier au sein de l'entreprise. Cet objet ne répond donc qu'à un identifiant unique..

L'unicité des données sert aussi à n'avoir qu'une seule description d'un produit donné. Elle contribue alors à l'amélioration de la qualité des données produit[Rgnier-Pcastaing et al., 2008].

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

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

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

(e) **L'integrité :**L'intégrité concerne les relations entre objets. Les relations importantes entre objets sont-elles toutes présentes ? Exemple : toute facture doit être associée à une commande. Si une facture n'a pas de référence vers une commande, c'est un problème d'intégrité[Rgnier-Pcastaing et al., 2008].

2. **Les critères de services**

(a) L'actualité : Une valeur de donnée est à jour si elle est correcte en dépit d'un écart possible avec la valeur exacte, due à des changements liés au temps ; une donnée est périmée à la date tsi elle est incorrecte à cette date mais était correcte aux instants précédant t. L'actualisation est le degré mesurant à quel point une donnée en

question est à jour (par exemple, l'âge ne devient obsolète qu'à la date anniversaire)[Rgnier-Pcastaing et al., 2008].

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

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

(d) La compréhensibilité : La compréhensibilité est la dimension qualité associée à la question : " cette donnée est-elle compréhensible ? ". Une donnée est compréhensible si chaque utilisateur, chaque processus, chaque application trouve facilement la bonne information parmi les attributs disponibles d'un objet. C'est le cas si celui-ci est clair et que l'alignement sémantique de l'ensemble des concepts entre tous les dépositaires (humains ou informatiques) a été réalisé et documenté [Jamm, janvier 2008].

1.2.3 l'importance de la qualité de données :

la qualité des données ne consiste pas seulement à aider les organisations à charger les bonnes données dans leurs systèmes d'information. Elle permet d'éliminer les données erronées ou les données en double. Le nettoyage des données devient une étape importante dans l'intégration des informations dans le système. La gestion de la qualité des données est la capacité de fournir des données fiables pour répondre aux besoins commerciaux et

techniques des utilisateurs. Il est mesuré en termes d'exactitude, de cohérence, d'unicité, d'exhaustivité et de disponibilité. Il s'agit d'une méthode de gestion de l'information conçue pour gérer et comparer des données entre différents systèmes d'information ou bases de données d'une entreprise. Habituellement, il s'agit de convertir des données de qualité en informations utiles essentielles à l'organisation.

1.2.4 Principaux problèmes du non qualité des données

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

1. **Création des données :**

— Entrée manuelle : absence de vérifications systématiques des formulaires de saisie
— Entrée automatique : problèmes de capture OCR, de reconnaissance de la parole Incomplétude, absence de normalisation ou inadéquation de la modélisation conceptuelle des données : attributs peu structurés, absence de contraintes d'intégrité pour maintenir la cohérence des données
— Entrée de doublons
— Approximations
— Contraintes matérielles ou logicielles
— Erreurs de mesure
— Corruption des données : faille de sécurité physique et logique des données

2. **Collecte/import des données :**

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

3. **Stockage des données :**

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

4. **Intégration des données :**

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

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

— Erreur humaine.
— Contraintes liées à la complexité de calcul.
— Contraintes logicielles, incompatibilité.
— Problèmes de passage à l'échelle, de performances et de confiance dans les résultats.

— Approximations dues aux techniques de réduction des grandes dimension

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

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

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

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

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

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

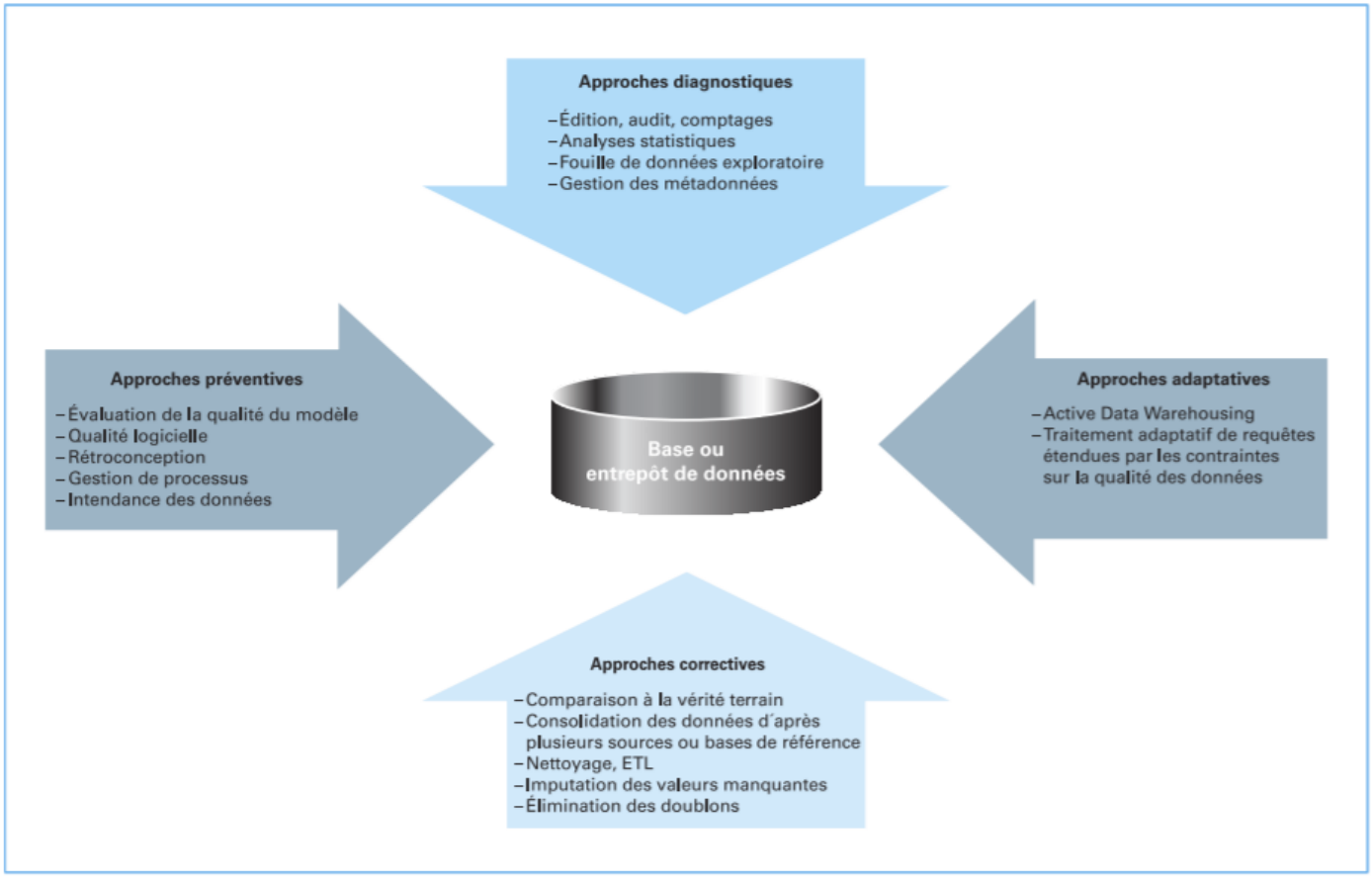

Figure 1.1 – Panorama des approches pour l'évaluation et le contrôle de la qualité des données.

1. 3 Conclusion

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

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

Dans ce chapitre, on a définir la qualité des données et leur concept, par la suite on abordera les conséquences de la non-qualité.

Chapitre 2 Record linkage

Chapitre 2 Record linkage

2.1 Définition

Record Linkage (RL), également connu sous le nom de couplage d'enregistrements, est le processus qui vise à identifier les enregistrements qui font référence à la même entité du monde réel. Les techniques de record linkage sont utilisées pour relier les enregistrements de données relatifs aux mêmes entités, telles que les patients ou les clients. Le record linkage peut être utilisé pour améliorer la qualité et l'intégrité des données, pour permettre la réutilisation des sources de données existantes pour de nouvelles études et pour réduire les coûts et les efforts d'acquisition de données pour les études de recherche.

2.2 Méthodologie du Record Linkage

Il existe deux méthodes principales de record linkage :

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

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

mortalité en utilisant des combinaisons de nom et prénom non uniques.

2.3 Les étapes de Record Linkage

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

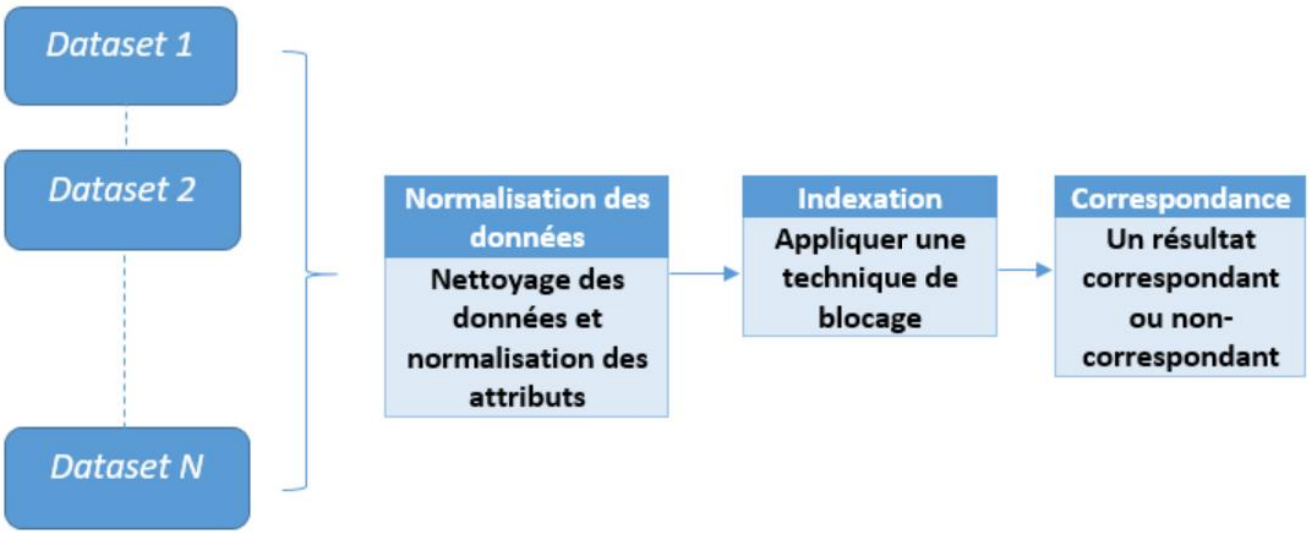

Figure 2.1 – Les étapes de Record Linkage.

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

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

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

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

2.4 Le blocage

2.4.1 Définition

Le blocage est la technique la plus utilisée dans l'étape de l'indexation.

Le blocage est le processus qui divise le dataset en un ensemble de blocs.

Tous les tuples affectés au même bloc partagent une valeur commune appelée la valeur de clé de blocage (BKV).

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

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

Figure2.2 – Exemple de clé de blocage.

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

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

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

2.4.2 L'objectif de blocage

Le blocage a deux objectifs principaux :

1. Le nombre d'appariements candidats générés doit être petit pour minimiser le nombre de comparaisons détaillées à l'étape de record linkage.

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

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

2.5 Sélection des clés de blocage

2.5.1 introduction :

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

Les deux étapes les plus importantes du processus de couplage d'enregistrements : les étapes d'indexation et La mise en correspondance.

Comme mentionné dans les séctions précédentes, «l'indexation est une étape critique du processus de RL». En fait, au cours de cette étape, le nombre de comparaisons est réduit en éliminant autant que possible les paires d'enregistrements sans correspondance. [Christen, b].

La technique d'indexation la plus utilisée est le «blocage», il est utilisé depuis les premières applications du record linkage [P.Fellegi et B.Sunter].

Le blocage consiste à créer un ensemble de blocs de manière à ce que tous les tuples du même bloc partagent la même valeur de clé de blocage[P.Fellegi et B.Sunter].

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

2.5.2 Approches du record linkage

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

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

L'idée derrière cela, est de diviser le BKV en sous-chaînes de taille Q, puis de sélectionner un nombre (fixé par l'utilisateur) de ces sous-chaînes et de les concaténer pour former les nouvelles valeurs de clés de blocage.

Dans [McCallum et al.] , les auteurs ont proposé une technique d'indexation basée sur l'algorithme de clustering Canopy. Il se compose de deux étapes principales. La première étape consiste à diviser les données en petites fractions de données appelées auvents en utilisant des méthodes peu coûteuses telles que le blocage d'index inversé. Une fois les auvents créés, la deuxième étape consiste à exécuter un algorithme de clustering classique dans chaque auvent associé à une métrique de distance coûteuse comme la distance d'édition.

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

Pour plus d'informations sur toutes les techniques d'indexation qui existent dans la littérature, une enquête a été publiée par Christen dans [Christen, b].

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

Plusieurs algorithmes de correspondance de chaînes existent dans les littératures certains d'entre eux appartiennent à la famille des encodages phonétiques comme (Soundex et phonex [Holmes et McCabe], phoenix [Gadd], NYSIIS et Double-Metaphone [Philips]. D'autres appartiennent à la famille de recherche de motifs comme l'algorithme Edit-distance qui est défini dans [lev] comme le nombre d'insertions, de suppressions et de substitutions pour transformer une chaîne en une autre. Il est généralement mis en oeuvre à l'aide d'un programme dynamique . Plus d'informations sur les techniques d'appariement quittées peuvent être trouvées dans[Elmagarmid et al.]. Une autre approche a été proposée pour détecter les doublons en utilisant des techniques d'apprentissage automatique comme dans [OUHAB et al.] où les auteurs ont utilisé l'algorithme SVM pour classer la paire d'enregistrements comme correspondances ou non.

2.5.3 Approches de blocage

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

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

— L'indexation Q-gram L et al. [a] est également une approche de blocage très populaire qui dépend du choix initial de la clé de

blocage. Dans cette approche, la valeur de la clé de blocage est divisée en un ensemble de Q-gams. Ensuite, ces Q-grammes obtenus sont utilisés comme nouvelles clés de blocage pour former les nouveaux blocs.

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

— L'approche des quartiers triés (sorted neighborhood)[M.A et S.J] commence également par la génération des clés de blocage. Une fois que cela est fait, les enregistrements sont triés en fonction de leurs BK générés et une fenêtre glissante, de taille W, se déplace sur les enregistrements. Tous les enregistrements qui sont dans la même plage de fenêtres sont comparés les uns aux autres. Cette approche a été améliorée plus tard dans [Yan et al.].

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

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

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

2.5.4 Sélection de Clé de blocage

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

— [Vogel et Naumann] ont proposé une approche automatique pour la sélection des clés de blocage basée sur les combinaisons d'uni gramme en tant que clés de blocage générées [Vogel et Naumann].

La première étape de cette approche consiste à générer toutes les combinaisons d'uni-gramme possibles. Pour chaque combinaison, un algorithme de détection de doublons est exécuté sur un dataset de référence. Après chaque exécution, si le nombre de doublons détectés est acceptable, la qualité globale de la clé de blocage est calculée.

Toutes les clés de blocage sélectionnées à partir de cette étape seront stockées et triées en fonction de leur qualité. Une fois la première étape effectuée, le processus de record linkage est exécuté sur un dataset de test à l'aide de la liste des clés de blocage triées de l'étape précédente et le meilleur BK satisfaisant aux critères d'arrêt choisis est sélectionné.

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

paires positives et négatives, tandis que la ligne du milieu représente tous les prédicats de blocage générés.

— [**?**] ont proposé un algorithme de couverture séquentielle (SCA) modifié pour générer les meilleurs schémas de blocage. Ils ont utilisé les correspondances du dataset comme exemple positif de l'algorithme SCA. Toutes les approches précédentes sont basées sur l'existence d'un dataset standard, ce qui n'est pas le cas des datasets du monde réel.

2.6 Conclusion

On peut déduire à partir des approches discutées que :

_ La plupart des approches de blocage dépendent de la sélection initiale de la clé de blocage.

_ proposer une solution de sélection de clé de blocage est l'une des priorités les plus importantes de la communauté RL.

_ La communauté Record Linkage a proposé plusieurs approches de sélection des clés de blocage.

_ Le problème de la sélection des clés de blocage peut être considéré comme un problème de sélection de fonctionnalités où toutes les clés de blocage possibles sont supposées être générées et l'objectif est de sélectionner le meilleur sous-ensemble de ces clés qui peut accélérer les performances de l'approche RL .

_ Plusieurs algorithmes ont été utilisés pour résoudre le problème de la sélection des attributs.

Chap 3
Méthode utilisée et Expérimentations

Chap 3
Méthode utilisée et Expérimentations

3.1 INRODUCTION

Après avoir exposé la théorie nécessaire à la compréhension de notre projet à travers les deux chapitres, nous allons présenter dans notre dernier chapitre la méthodologie suivie pour répondre à la problématique posée dans l'introduction générale, nous proposons une nouvelle approche de sélection des clés de blocage semi-automatique dans le processus de record linkage. Cette approche proposée est basée sur l'algorithme K-Modes pour l'étape d'indexation .Nous l'expliquons dans cette section en détails et comment nous avons adapté cet algorithme pour résoudre le blocage de la sélection des touches problème.

3.2. K- Modes

3.2. 1.Présentation de la méthode

K-Modes a été proposé pour la première fois en 1998 par HUANG [15]. Cet algorithme est considéré comme une extension de l'algorithme de cluster classique K-Means, il a été proposé dans afin de regrouper des données catégorielles, ce qui n'est pas le cas de la Algorithme K-Means qui accepte uniquement les attributs numériques. Bien sûr, il existe l'algorithme de cluster hiérarchique classique qui traite à la fois des données catégorielles et numériques, mais son la complexité quadratique le rend inadapté au regroupement de grands ensembles de données. L'utilisation des modes K dans notre approche nous a permis de éliminer l'étape de conversion des données numériques qui était un beaucoup de temps et une étape nécessaire à faire avec les k-means algorithme. Les auteurs ont basé leur algorithme sur trois principaux points : (1) Mesure de dissimila rite simple (démontrée dans équation 1) afin de faire correspondre les objets. (2) Utilisation modes à la place des moyens et (3) Approche basée sur la fréquence pour mesurer le mode d'un ensemble.

$$d_1(X,Y) = \sum_{j=1}^{m} \delta(x_i, y_i), \quad \alpha = \begin{cases} 0 & (x_i = y_i) \\ 1 & (x_i \neq y_i) \end{cases} \qquad (1)$$

Pour la sélection initiale des modes, deux techniques ont été proposées dans le journal. La première consiste à affecter le premier K enregistrements distincts comme modes initiaux. La seconde approche consiste de mesurer les fréquences pour toutes les catégories de chaque attribut et les trier par ordre décroissant selon leurs fréquences. Une fois cela fait, nous attribuons la première fréquence catégories aux premiers k-modes initiaux.

Dans notre travail, nous utilisons l'algorithme K-Modes pour regrouper données en blocs. Chaque bloc contiendra des correspondances possibles. Le cluster utilisera uniquement les clés de blocage générées dans la étape précédente en tant qu'attributs de cluster, au lieu d'utiliser tous les attributs du jeu de données, afin de gagner du temps et parce que les clés de blocage contient les informations les plus importantes sur les enregistrements.

3.2.2. Filtrage adaptatif

Une fois le cluster terminé et avant de passer au l'étape d'appariement, nous exécutons le filtrage adaptatif présenté dans [12]. Les auteurs ont proposé une nouvelle approche afin de réduire la comparaison du nombre de paires d'enregistrements une fois le blocage terminé et c'est en ignorant la comparaison entre les paires qui sont dans le même cluster mais sont considérés comme des correspondances improbables.

Cette approche a été proposée après avoir observé que tous les les méthodes de blocage peuvent générer de très gros blocs. La première technique de filtrage utilisée dans cette approche est "filtrage de longueur", où deux enregistrements sont déclarés comme une correspondance peu sympathiques si la différence entre la longueur de leur variable de filtrage est supérieure à une valeur prédéfinie K. La deuxième technique est le filtrage de comptage où chacun des les variables de filtrage des deux enregistrements comparés sont divisées en un ensemble de bigrammes, puis le nombre du commun bi-grammes entre les deux chaînes est comparé à Cmin où **Cmin = max(s1; s2) - 2k + 1** avec k représentent l'édition

distance entre les deux variables de filtrage. Si le nombre de le bi-gramme commun est plus petit que Cmin que les enregistrements sont déclarés comme des correspondances peu sympathiques.

Les expériences sur des jeux de données du monde réel et sur des données synthétiques ensembles de données [12], ont montré que les techniques de filtrage ont réduit le nombre de comparaisons jusqu'à 80 % par rapport au blocage traditionnel même lorsqu'il s'agit de petits blocs.

3.2.3. Correspondance

La dernière étape de notre approche est l'appariement entre l'enregistrement pair dans chaque groupe. Nous avons choisi d'utiliser un ensemble de métriques de similarité pour le fait qu'ils sont conçus pour traiter avec des erreurs typographiques [8] ce qui est le cas des plus réels ensembles de données du monde. Plusieurs métriques pour la correspondance de chaînes existent dans le la littérature a été évaluée (distance d'édition, similarité de Jaro Winkler, Similitude de Jaccard, distance de Smith-Waterman, Q-Grams et Suite).

3.3. La sélection semi-automatique des clés de blocage avec K-modes

Dans cette approche, K-Modes est utilisé comme étape d'indexation en regroupant les données en utilisant uniquement les clés de blocage qui sont les sous-ensembles de fonctionnalités actuellement sélectionné. Les meilleures clés de blocage en termes de fitness sont celles dans lesquelles K-Modes regroupe les enregistrements les plus dupliqués lorsqu'ils sont utilisés comme attributs de clustering.

En conséquence, la fonction de fitness est le paramètre de complétude de la paire (PC). Le PC mesure le nombre de doublons détectés par une approche RL en utilisant les clés de blocage sélectionnées.

Notre approche proposée peut être résumée par les points suivants :

- créez toutes les listes de clés de blocage possibles.
- Initialisez la première tuple qui est un sous-ensemble d'entités aléatoires de la liste des clés de blocage précédemment crée.
- Le meilleur membre de la dernière tuple est sélectionné comme meilleur sous-ensemble de fonctionnalités à utiliser comme clés de blocage.

3 .4 Créations des clés candidates :

Avant de créer la liste des clés candidates, une étape essentielle de prétrai-tement ne peut être négligée. Il s'agit, en fait, de nettoyer l'ensemble A. en d'autres termes, il faut éliminer les attributs de mauvaise qualité de l'ensemble A. Deux paramètres sont utilisés pour calculer la qualité globale d'un attribut. Premièrement, l'exhaustivité représente le pourcentage de valeur nulle concernant les attributs spécifiés [L.L et al.]. Nous avons utilisé la mesure NBC (Null-based Com-pleteness) où l'exhaustivité est mesurée à l'aide de l'équation 3.En utilisant cette méthode, la valeur 1 représente le meilleur résultat et 0 le pire. Tous les attributs qui ont une valeur d'exhaustivité inférieure au seuil prédéfini sont éliminés de la génération de liste de clés de blocage candidates.

$$\text{Completness(Attj)} = 1 - \frac{\text{number of Null values in Attj}}{Number\ of\ instances} \quad (3)$$

Le deuxième paramètre est la cardinalité d'un attribut. La cardinalité représente le nombre de valeurs distinctes pour un attribut spécifié. Dans le processus RL, les attributs à très faible cardinalité ne conviennent pas pour être utilisés comme clés de blocage. Par exemple, l'utilisation de l'attribut sex comme clé de

blocage divise les données en seulement 2 blocs (M / F). Par conséquent, dans notre approche, les attributs à très faible cardinalité sont éliminés de la génération de liste de clés de blocage candidates.

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

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

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

3.5. Le couplage d'enregistrements

Comme mentionné précédemment, afin de tester les performances de chaque sous-ensemble à partir des fonctionnalités sélectionnées, nous avons utilisé l'approche de liaison d'enregistrements basée sur les modes K [H.N et al.]. Dans cette approche, les données sont regroupées en blocs en utilisant

uniquement les clés de blocage comme attributs de clustering, qui sont dans notre cas les fonctionnalités sélectionnées. Une fois le clustering terminé, la correspondance entre les enregistrements du même cluster est effectuée à l'aide d'une métrique de similarité de chaîne comme la similarité Jaro-Winkler. Le nombre de valeurs dupliquées détectées est exprimé à l'aide du paramètre de complétude de pair (PC) (Équation 4.1).

$$\mathrm{PC} = \frac{\textbf{nomber de paires d'enregitrements détectées}}{\textbf{nombre de paires d'enregistements en double dans le dataset}} \quad \text{Équation (4.1)}$$

D'autres paramètres sont utilisés pour mesurer la performance d'une approche de couplage d'enregistrements. La ration de réduction (RR) (équation 4.2) est utilisée pour mesurer dans quelle mesure la technique de blocage a réussi à réduire le nombre de comparaisons. La mesure F (équation 4.3) est utilisée pour contrôler le compromis entre RR et PC.

$$\mathbf{RR} = \mathbf{1} - \frac{\textbf{nomber de paires d'enregitrements détectées}}{\textbf{nombre de paires d'enregistements en double dans le dataset}} \quad (4.2)$$

$$\boldsymbol{fMesure} = \mathbf{2} * \frac{\mathbf{RR} * \mathbf{PC}}{\mathbf{RR} + \mathbf{PC}} \quad (4.3)$$

3.6. Implémentation et expérimentation

3.6.1. Environnement de travail

- Environnement matériel

• Processeur : Intel i5-5400U CPU @ 2,40 GHz 2,50 GHz

• Mémoire installée (RAM) Mémoire installée (RAM) : 8,00 Go

• Type du système : Type du système : Système d'exploitation

64 bits, processeur x64

- **Langage de programmation**

Parmi les différents langages de programmation existant dans le monde de développement, nous avons choisi le JAVA. C'est un langage de programmation orienté objet, développé par Sun Microsys-tems. Il permet de créer des logiciels compatibles avec de nombreux systèmes d'exploitations (Windows, Linux, Macintosh, Solaris). Java donne aussi la possibilité de développer des programmes pour téléphones portables et assistants personnels. Enfin, ce langage peut être utilisé sur internet pour des petites applications intégrées à la page web (applet) ou encore comme langage serveur (jsp) 1.
La technologie Java est indissociable du domaine de l'informatique

2.**Environnement logiciel :**

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

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

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

3.6.2. Présentation de l’ interface de notre application :

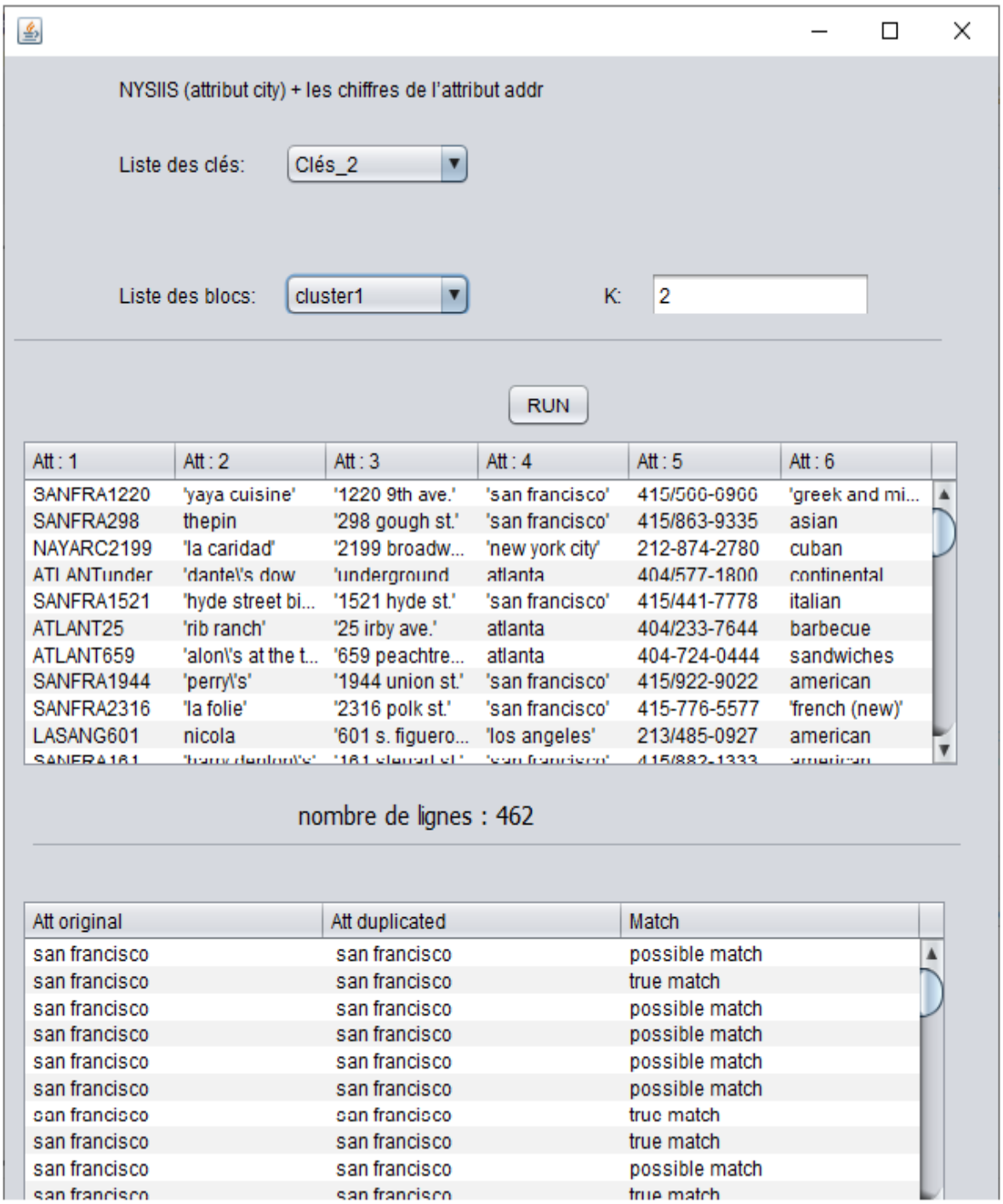

Fig. 3.1 l’ interface de notre application

L’exécution en (K=8 et Les clés=Clés_1)

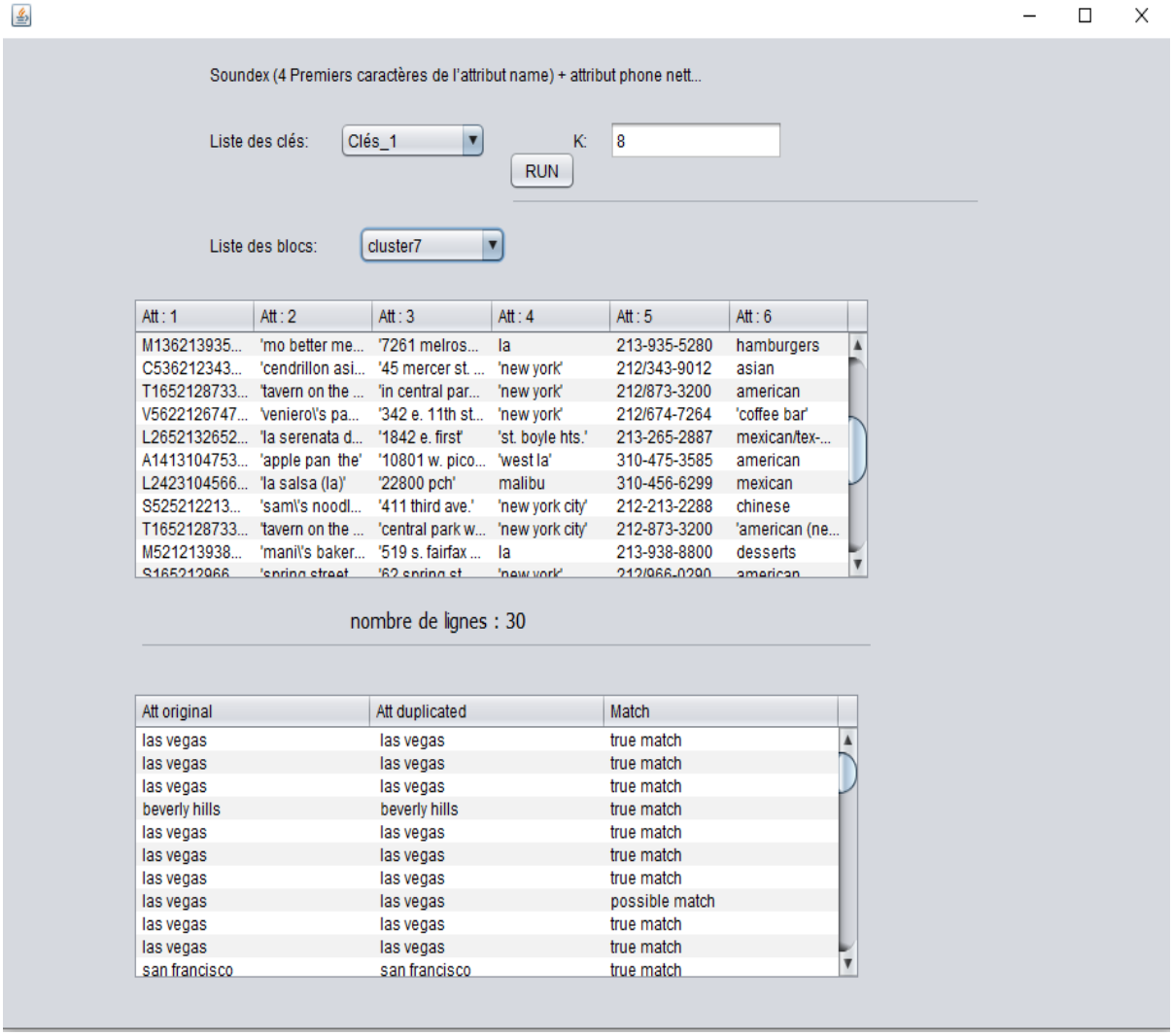

Fig. 3.2

L’exécution en (K=8 et Les clés=Clés_2)

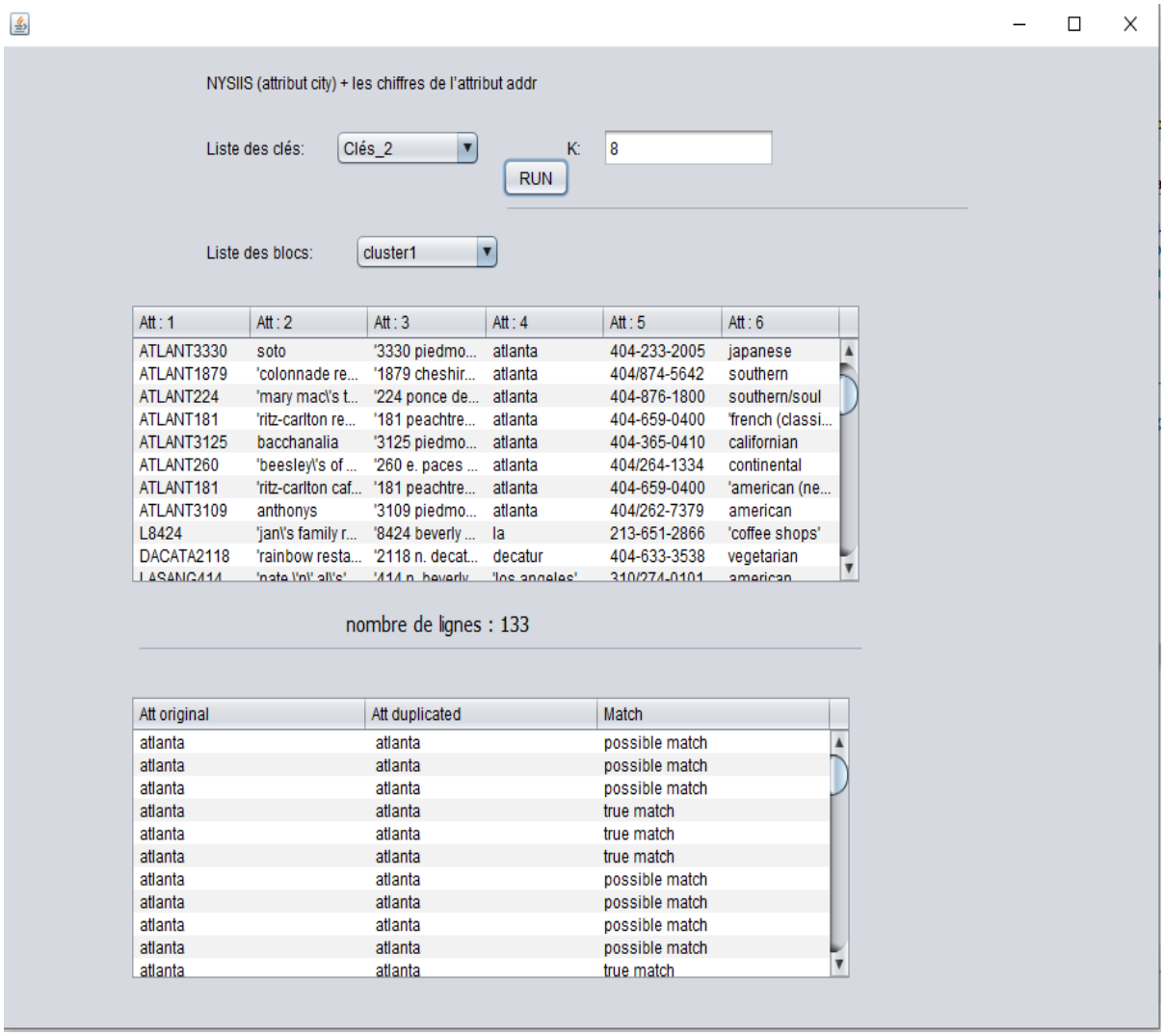

Fig. 3.3

L’exécution en (K=8 et Les clés=Clés_3)

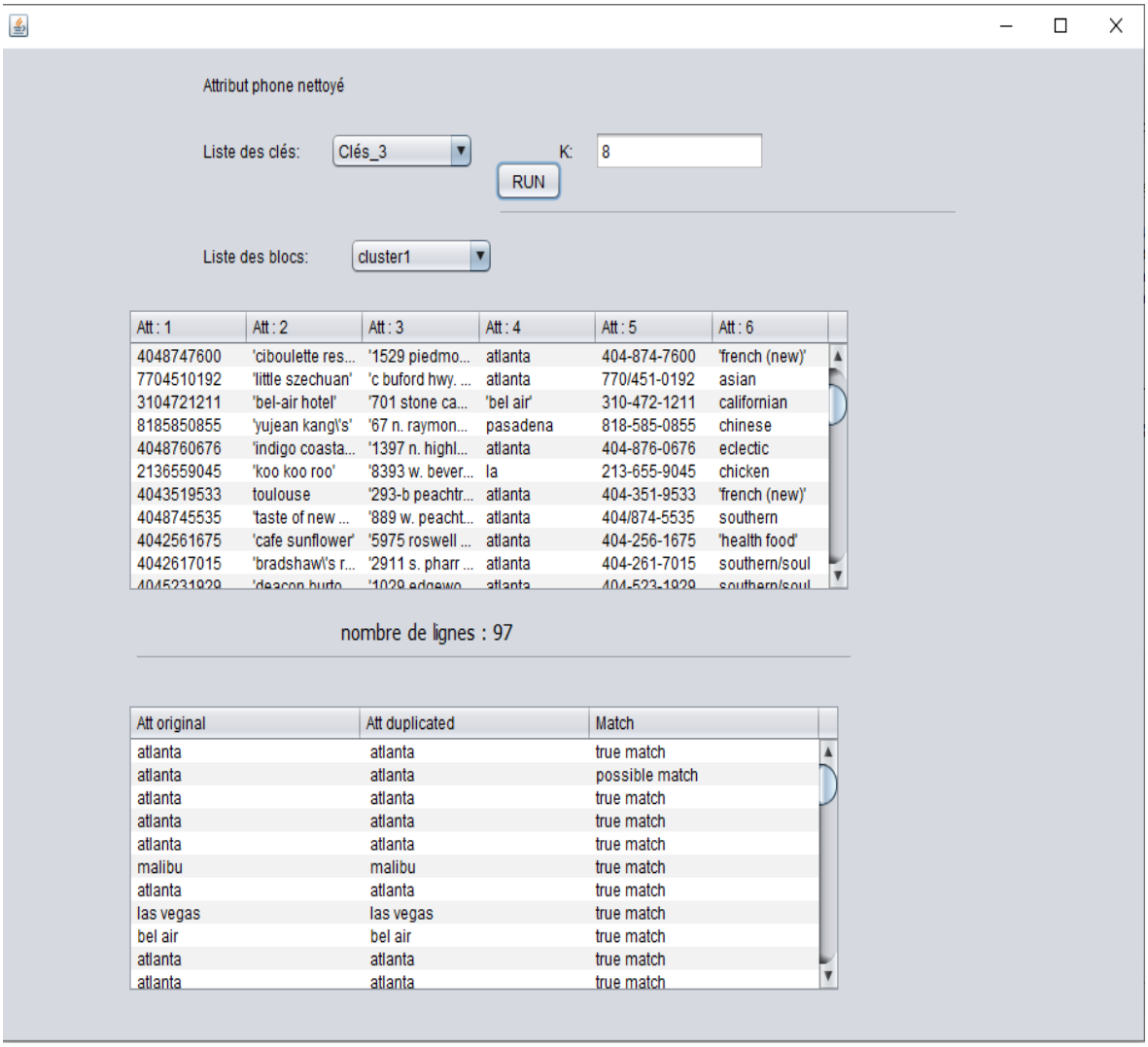

Fig. 3.4

L’exécution en (K=16 et Les clés=Clés_1)

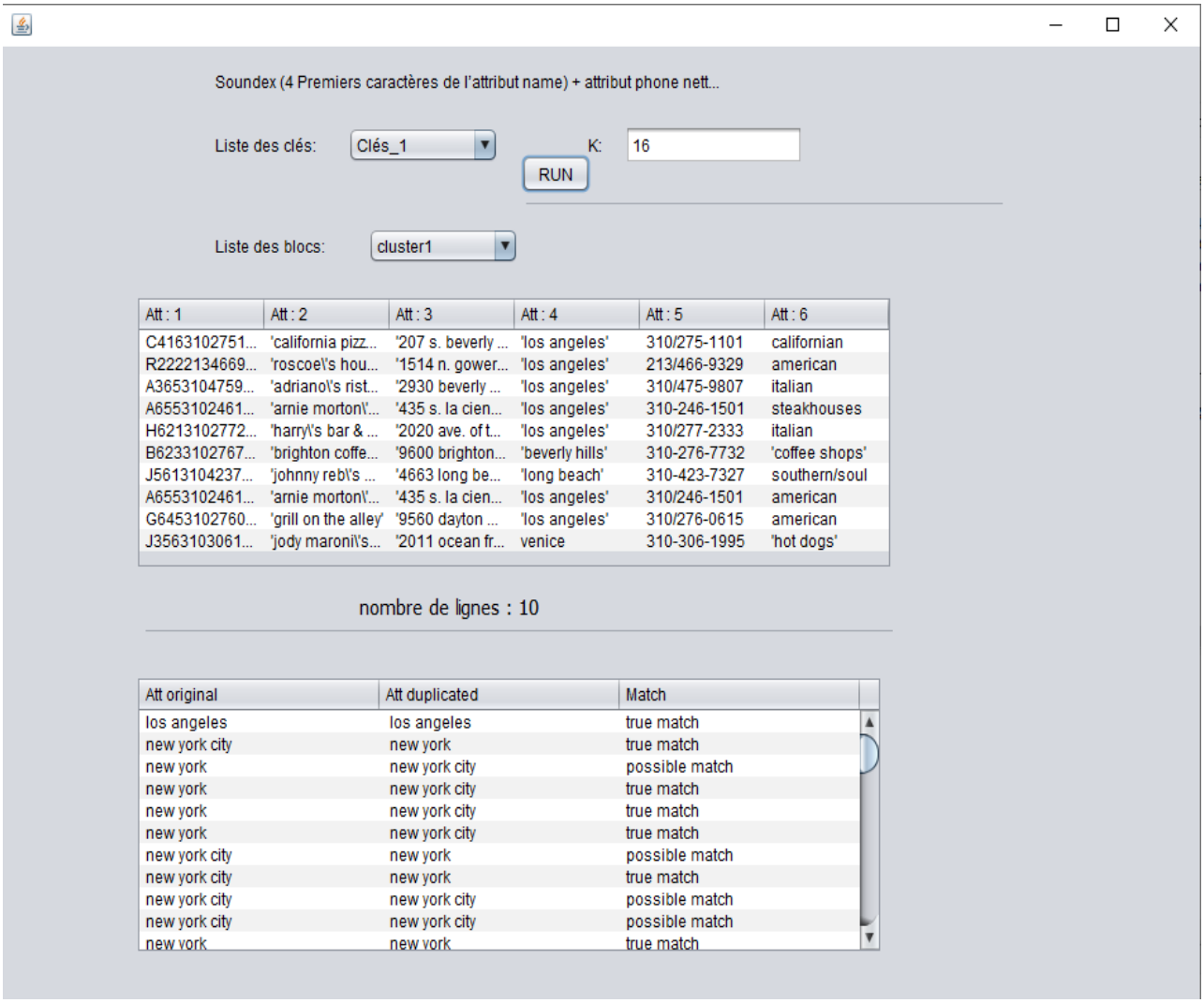

Fig. 3.5

L’exécution en (K=16 et Les clés=Clés_2)

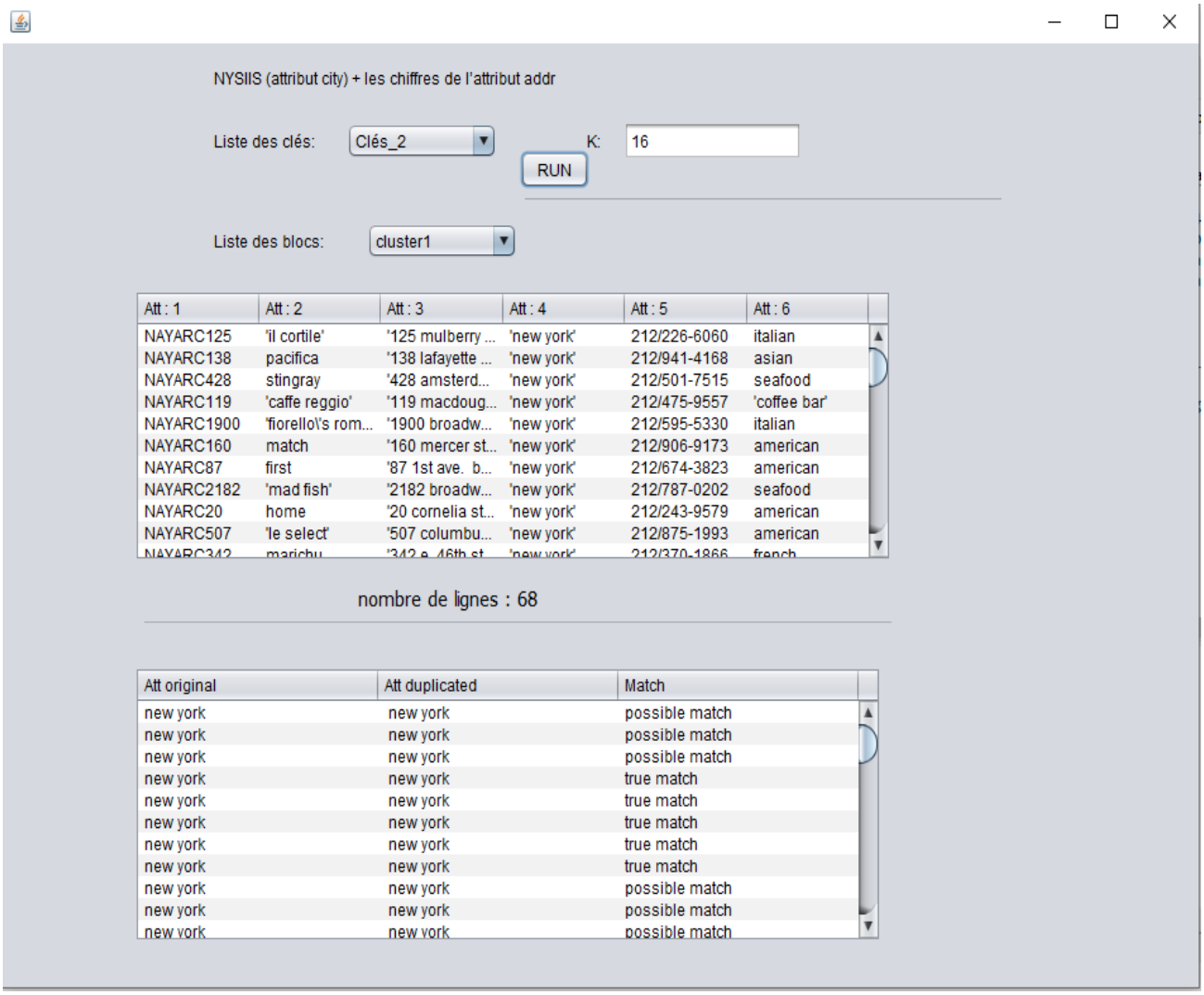

Fig. 3.5

L’exécution en (K=16 et Les clés=Clés_3)

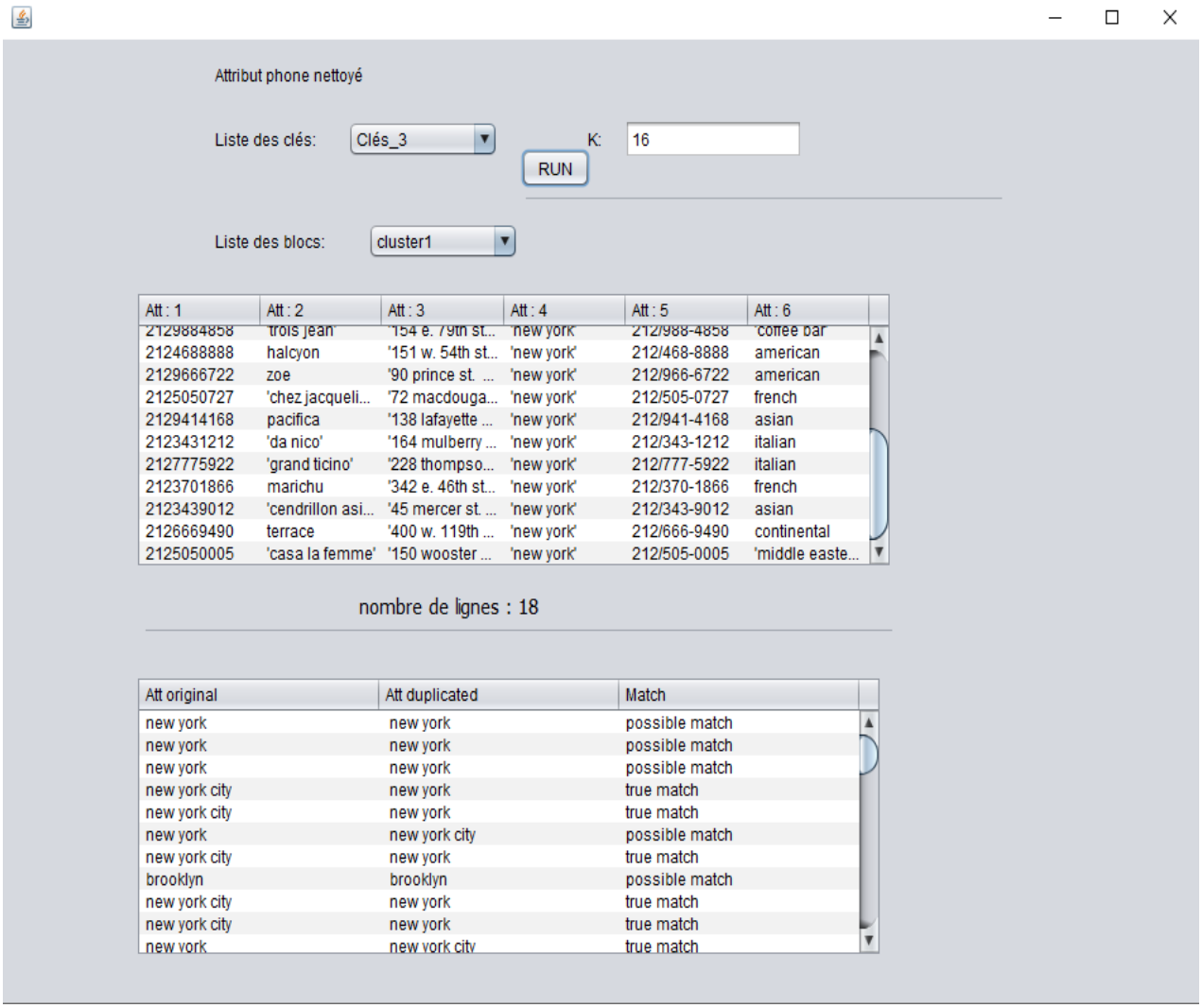

Fig. 3.6

L’exécution en (K=50 et Les clés=Clés_1)

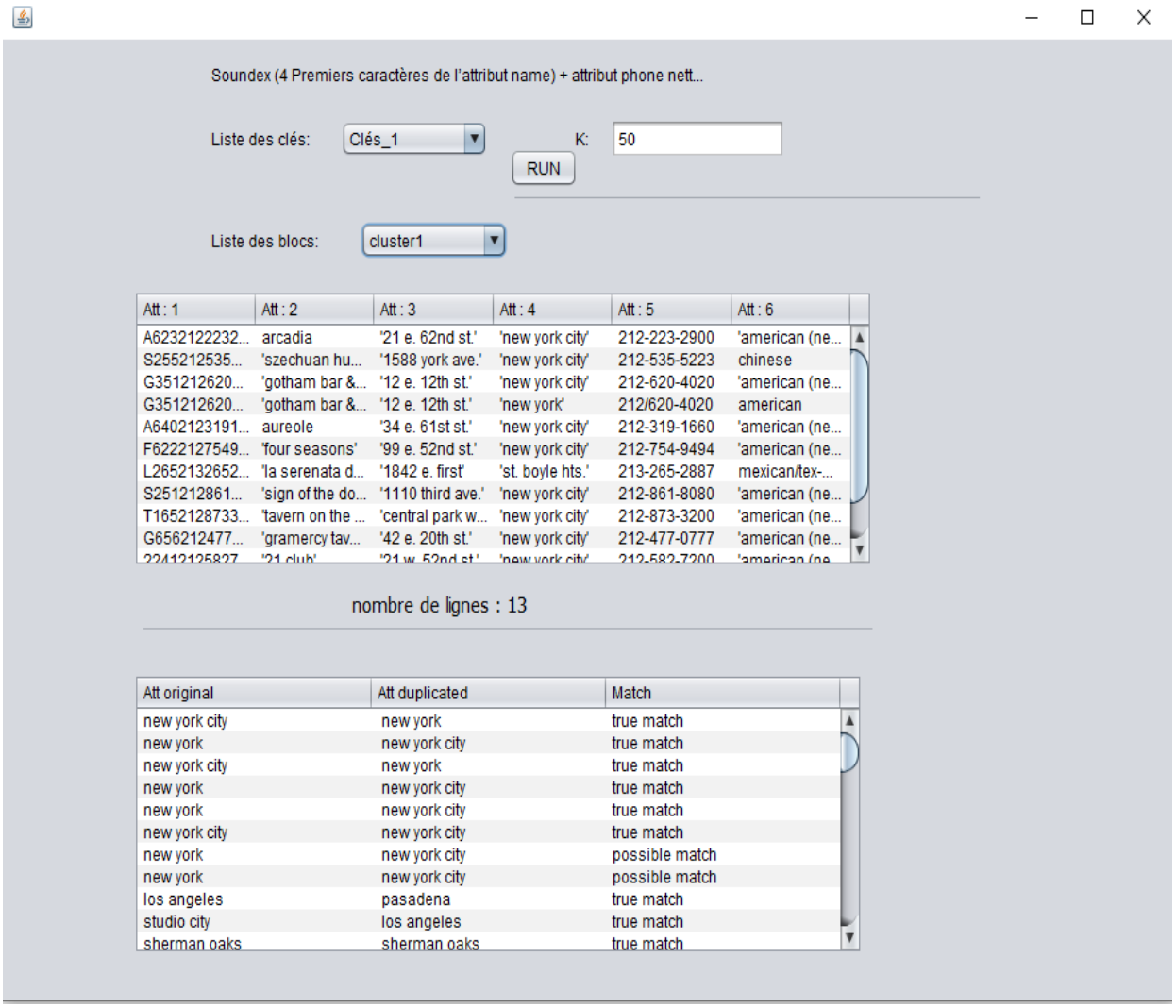

Fig. 3.7

L’exécution en (K=50 et Les clés=Clés_2)

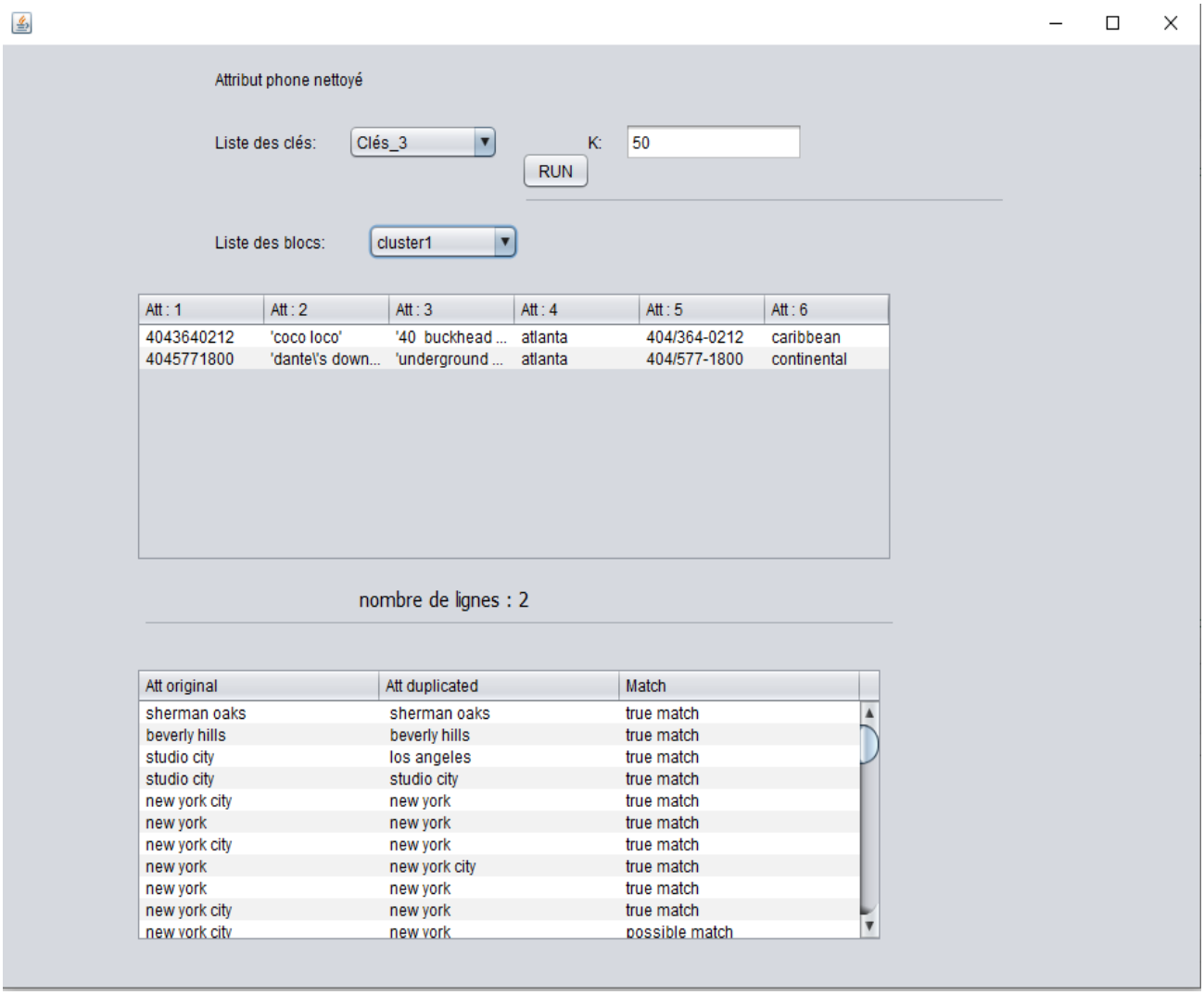

Fig. 3.8

L’exécution en (K=50 et Les clés=Clés_3)

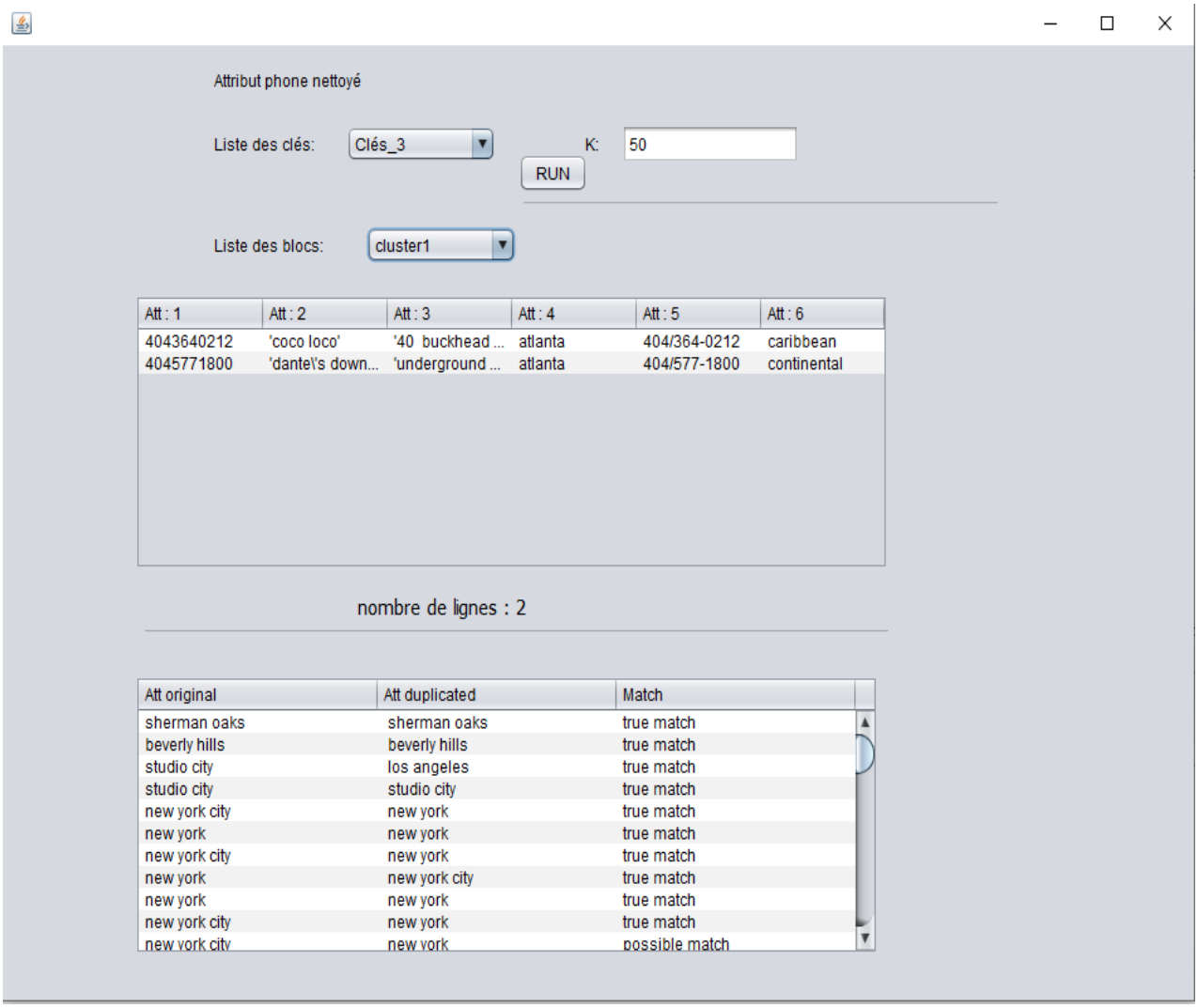

Fig. 3.9

TABLE I
SELECTED DATASETS FOR EXPERIMENTS

Data set	Nombre des instances	True matches
Restaurant	864	100

TABLE II

RÉSULTATS POUR L'ENSEMBLE DE DONNÉES DU RESTAURANT

	K = 8							
CLE	ATT 1	ATT 2	ATT 3	ATT 4	ATT 5	ATT 6	ATT 7	ATT 8
1	46	51	126	178	80	317	28	38
2	79	108	26	213	48	242	03	145
3	91	59	199	123	128	32	74	158

3.7 . CONCLUSION

Dans ce chapitre, nous avons présenté une approche de sélection semi-automatique des clés de blocage pour un couplage d'enregistrements efficace. L'approche était basée sur l'algorithme Optimiser récemment introduit. Notre choix d'utiliser une méta-heuristique s'est fait après avoir remarqué leur capacité à résoudre les problèmes NP-Hard, ce qui est le cas de la sélection de fonctionnalités. Nous avons généré un ensemble de clés de blocage aléatoires à partir desquelles la population initiale de k-modes est sélectionnée au hasard. Nous avons choisi d'utiliser la sélection des fonctionnalités de la paire du blocage basé sur les K-Modes comme fonction de fitness. Les expériences sur deux ensembles de données du monde réel ont montré l'efficacité de l'algorithme k-modes pour sélectionner le meilleur sous-ensemble de clés de blocage.

Conclusion Générale

Conclusion Générale

Conclusion Générale

A l'ère du « big data », les sources d'informations disponibles se multiplient et les volumes de données potentiellement accessibles augmentent de façon exponentielle. La qualité des informations et leur véracité ont de fait pris une importance majeure, pour cela de nombreuses méthodes pour identifier, mesurer et résoudre certains problèmes de qualité des données existent.

Notre étude est concentrée sur la sélection des attributs, plus particulièrement, la sélection des clés de blocage qui est un élément essentiel dans ce domaine. Parmi des nombreuses approches décrivant ce problème, nous avons fixé notre étude sur la sélection automatique des clés de blocage à l'aide d'une méta-heuristique. Pour notre expérimentation nous avons choisi k-modes comme algorithme en tant qu'étape d'indexation, et l'algorithme de k-modes comme une méthode d'indexation et de clsutering.

L'avantage majeur de l'algorithme k-modes réside dans le fait qu'il traite directement avec les données catégorielles, nous n'avons donc pas à les convertir en données numériques. L'étape de correspondance (matching) consiste à faire correspondre les paires d'enregistrements dans le même bloc et décider si elles représentent une correspondance vraie ou possible, cette décision a été faites en comparant les clés de blocage (BK) au lieu de comparer tous les attributs des data sets.

Enfin, l'analyse de convergence de k-modes a confirmé la convergence de cet algorithme. De plus, les résultats des problèmes de conception technique ont également montré que l'algorithme k-modes a des performances élevées dans des espaces de recherche inconnus et difficiles.

L'algorithme k-modes a finalement été appliqué à deux data sets du monde réel. Les résultats ont montré une amélioration par rapporta la sélection manuelle, montrant l'applicabilité de l'algorithme proposé dans la résolution de problèmes réels.

Pour répondre aux questions soulevées lors de la problématique est sachant que le volume de donnés ne cesse d'augmenter dans le temps. Ce qui rend la sélection semi-automatique des BK nécessaire et indispensable.

Bibliographie

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 match problems. Proceedings of the VLDB Endowment, 3(1-2) :484–493, a.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Dif N, Attaoui M, et Elberrichi Z. Gene selection for microarray data classification using hybrid meta-heuristics. Dans International Symposium on Modelling and Implementation of Complex Systems, page 119–132. Springer.

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

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

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

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

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

Kalsi S, Kaur H, et Chang V. Dna cryptography and deep learning using genetic algorithm with nw algorithm for key generation. Journal of medical systems, 42(1) :17.

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

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

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

Public License :

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

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

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

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