UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MESURES DE SIMILARITÉ POUR L'ANALYSE DE DONNÉES DE PRODUCTION LAITIÈRES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MAITRISE EN INFORMATIQUE (INTELLIGENCE ARTIFICIELLE)

PAR

CISSE ASSAMAOU DITE SAWATOU

MARS 2025

REMERCIEMENTS

Ce mémoire est l'aboutissement de plusieurs mois de travail, de recherche, et de réflexion. Il n'aurait pu être réalisé sans le soutien et les encouragements de nombreuses personnes que je tiens à remercier sincèrement.

Avant tout, je tiens à exprimer ma profonde gratitude et mon immense respect envers mes parents. Leur amour inconditionnel, leur patience et leur soutien constant ont été les piliers de mon parcours académique et personnel. Vous m'avez inculqué des valeurs essentielles, telles que la persévérance, l'humilité et le respect, qui m'ont guidé tout au long de cette aventure. Chaque étape franchie, chaque succès remporté, je vous le dois. Merci de croire en moi, même dans les moments où je doutais de moi-même.

Je souhaite exprimer ma profonde gratitude à mon directeur de recherche, **Petko**, pour son accompagnement constant, ses conseils éclairés, et sa disponibilité tout au long de ce projet. Ses retours constructifs et sa rigueur scientifique ont grandement contribué à l'aboutissement de ce travail.

Mes remerciements s'adressent aussi à l'ensemble de mes enseignants et collègues du programme **Maîtrise en Informatique**, pour les connaissances partagées et les échanges enrichissants au fil des années.

À ma famille, merci pour votre amour inconditionnel, votre patience et vos encouragements, qui m'ont donné la force de surmonter chaque défi. Votre foi en mes capacités a été une source d'inspiration et de motivation pour atteindre mes objectifs.

À mes amis, je vous remercie pour votre présence, votre bonne humeur et vos mots réconfortants dans les moments difficiles. Votre soutien moral et votre camaraderie ont rendu ce parcours plus agréable et moins solitaire. Vous avez su être là, à la fois dans les périodes de doute et dans les moments de réussite, et pour cela, je vous suis profondément reconnaissante.

Que ce mémoire soit le reflet de ma gratitude envers vous tous, qui avez contribué, d'une manière ou d'une autre, à la réalisation de ce travail.

DÉDICACE

À ma chère mère, Pour ton amour inconditionnel,ta patience infinie et ton soutien indéfectible.

Tu as toujours cru en moi, même lorsque mes doutes prenaient le dessus, et tu as été ma lumière dans les moments les plus sombres.

Ce mémoire est le fruit de ton inspiration, de tes sacrifices et de ta foi en mes capacités.

Je te dédie ce travail avec toute ma gratitude et mon amour éternel.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

L'analyse des données liées à la production laitière joue un rôle primordial dans l'optimisation de la gestion des troupeaux et l'amélioration des performances de production. Grâce aux progrès technologiques récents, les exploitations agricoles ont accès à des dispositifs sophistiqués comme les capteurs de performance animale et les systèmes de gestion numériques, qui permettent de collecter des volumes conséquents de données. Ces informations, souvent structurées sous forme de graphes ou de tables hiérarchiques, incluent des paramètres variés, tels que les rendements en lait, les cycles de lactation et les conditions environnementales. Cette évolution technologique constitue une opportunité précieuse pour affiner les prises de décision au sein des exploitations agricoles (webagri, 2022).

Pour exploiter efficacement ces données complexes, les mesures de similarité se révèlent être des outils indispensables. Elles permettent de comparer les performances des vaches ou de groupes spécifiques en tenant compte de leurs caractéristiques individuelles ainsi que des relations structurelles présentes dans les données. Les applications de ces approches sont variées : elles incluent la sélection génétique, le diagnostic précoce des maladies et l'optimisation des pratiques d'élevage. Par exemple, la détection des similitudes dans les cycles de lactation peut faciliter le regroupement d'animaux aux performances proches, ce qui améliore les stratégies de gestion du troupeau (Alves *et al.*, 2020).

Historiquement, les mesures de similarité ont souvent reposé sur des méthodes traditionnelles comme la distance euclidienne ou le coefficient de Jaccard, appliquées sur des vecteurs d'attributs. Cependant, ces approches montrent des limites face à la complexité et aux relations hiérarchiques présentes dans les systèmes d'information modernes sur la production laitière. Ces limitations ont conduit au développement de nouvelles méthodes, notamment celles basées sur l'analyse de graphes et les modèles d'apprentissage automatique, qui permettent une meilleure prise en compte des structures relationnelles entre les données (Rifqi et Bouchon-Meunier, 2004).

En complément, des approches d'apprentissage non supervisé, utilisant des techniques avancées de calcul de similarité, offrent aujourd'hui des perspectives innovantes pour analyser des jeux de données complexes. Ces méthodes facilitent la détection des tendances et des anomalies, permettant ainsi une gestion plus proactive et efficace des troupeaux (Rifqi et Bouchon-Meunier, 2004).

Parmi les approches modernes, les **mesures de similarité structurelle** basées sur des graphes, telles que SimRank (Jeh et Widom, 2002a) et P-Rank (Zhao *et al.*, 2009), se sont révélées efficaces pour capturer les relations complexes entre les entités. Ces méthodes évaluent la similarité entre les nœuds d'un graphe en fonction de leurs voisins et des relations qui les relient, ce qui est particulièrement pertinent dans le contexte des données de production laitières, où les entités (vaches, périodes de lactation) sont fortement interconnectées. D'autre part, les approches basées sur l'embedding de graphes, comme Node2Vec (Grover et Leskovec, 2016), permettent de représenter les entités dans un espace vectoriel tout en préservant leurs propriétés structurelles. Ces représentations facilitent le calcul de similarités et l'application de techniques d'analyse traditionnelles.

D'autres méthodes, comme les réseaux de neurones de graphes (GNN), combinent les avantages des graphes et des modèles d'apprentissage profond pour apprendre des représentations complexes à partir des données brutes. Ces modèles peuvent intégrer à la fois des informations locales, telles que les caractéristiques des vaches individuelles, et des informations globales, comme les relations entre différents troupeaux. Les modèles GNN ont été utilisés avec succès dans d'autres domaines, comme la bioinformatique et l'analyse de réseaux sociaux, et commencent à être explorés pour l'analyse des données laitières.

Malgré ces avancées, l'application des mesures de similarité aux données de production laitières présente encore plusieurs défis. D'une part, les données sont souvent déséquilibrées, avec une sur-représentation de certaines catégories (par exemple, les vaches à forte production) par rapport à d'autres. D'autre part, la nature hiérarchique des données nécessite des techniques capables de capturer les relations multi-niveaux, comme celles entre les jours de test au sein d'une période de lactation ou entre différentes périodes de lactation pour une même vache. Ces spécificités appellent au développement de nouvelles approches adaptées aux caractéristiques uniques de ce domaine.

Dans ce contexte, l'analyse des données de production laitière, souvent organisées sous forme de graphes ou de bases structurées (vaches, périodes de lactation, journées de test), est devenue cruciale pour évaluer les performances et améliorer les pratiques. Cependant, ces données complexes posent plusieurs défis. Tout d'abord, leur nature hiérarchique nécessite des approches capables de traiter des niveaux d'abstraction multiples, tels que les vaches, les périodes de lactation et les journées de test. Ensuite, l'évaluation des similarités entre vaches ou groupes de vaches pour identifier des tendances, des anomalies ou des facteurs influents est compliquée par l'hétérogénéité des données. Ces dernières incluent des attributs va-

riés, comme les rendements en lait, les taux de composants (matières grasses, protéines), ainsi que des métadonnées relatives aux conditions de production. Enfin, les méthodes classiques de similarité, souvent basées sur des distances statistiques ou géométriques, se révèlent inadéquates pour capturer les relations complexes et hiérarchiques propres à ces données. Cela souligne la nécessité de développer de nouvelles approches tenant compte à la fois de la structure des données et des attributs individuels.

La capacité à mesurer avec précision la similarité entre différentes unités (vaches, lactations, journées de test) est cruciale pour de nombreuses applications. Elle permet, par exemple, de détecter des vaches aux performances similaires afin d'optimiser les pratiques d'élevage, d'identifier des schémas communs dans des situations atypiques, comme des périodes de stress ou de maladie, ou encore de prédire les performances en exploitant les relations de similarité entre individus. Pourtant, malgré l'importance de ces analyses, il n'existe pas de consensus sur une méthodologie standard adaptée au domaine de la production laitière. Les approches existantes, souvent empruntées à d'autres disciplines, n'intègrent pas suffisamment les spécificités des données laitières, notamment leur structure hiérarchique, leur hétérogénéité et les dépendances entre niveaux.

Ainsi, la problématique centrale qui émerge de cette réflexion est la suivante : *Comment développer et appliquer des mesures de similarité adaptées pour analyser les données de production laitière en tenant compte de leur structure hiérarchique, de leur hétérogénéité et de leur complexité ?* Ce mémoire se propose d'apporter une réponse à cette question en élaborant une méthodologie robuste et interprétable pour le calcul de similarité, tout en démontrant son efficacité à travers des études de cas sur des données réelles de production laitière.

Ce mémoire est structuré en cinq chapitre (5) :

- **Chapitre 1 : Notions préliminaires**
  Dans ce premier chapitre nous aborderons les essentiels concernant les vaches, les lactations et les jours de tests et aussi nous présenterons les concepts fondamentaux en lien avec les calculs de similarité dans le domaine de la production laitière.
- **Chapitre 2 : Concepts de similarité et graphes de connaissances**
  Une introduction aux concepts d'analyses statistiques, de calcul de similarité et de graphes de connaissances pour comprendre et traiter les données.

— **Chapitre 3 : Revue de la littérature et problématique**

Ce chapitre synthétise les travaux existants dans le domaine des mesures de similarité appliquées aux graphes de connaissances et formule la problématique centrale de la recherche.

— **Chapitre 4 : Méthodologie**

Une description détaillée des données utilisées, ainsi que la méthodologie adoptée pour le calcul des similarités et l'analyse des résultats sur un exemple avec une implementation, est fournie dans ce chapitre.

— **Chapitre 5 : Résultats et discussion**

Dans ce chapitre nous exposons es résultats obtenus à travers une analyse et discutons, en mettant en lumière les pertinences et les limites de notre méthodologie.

RÉSUMÉ

Ce mémoire explore les mesures de similarité dans l'analyse des données de production laitières, un enjeu crucial pour optimiser la gestion des troupeaux et améliorer les performances animales. Avec l'essor des technologies modernes de collecte de données, les fermes produisent d'importants volumes d'informations, souvent complexes et hiérarchiques, qui nécessitent des méthodologies adaptées.

Le document présente une méthodologie novatrice basée sur l'utilisation de graphes pour modéliser les relations entre les vaches, les cycles de lactation et les jours de test. Des mesures de similarité structurelle, telles que SimRank et ses variantes, ont été adaptées pour comparer efficacement ces entités tout en tenant compte de leurs attributs spécifiques et de leur organisation hiérarchique. Ces techniques permettent d'évaluer les performances, de détecter des anomalies et de prédire des tendances futures.

L'approche a été validée sur des données réelles, démontrant son efficacité dans la classification des individus et la prédiction de performances. Les résultats obtenus montrent une amélioration significative par rapport aux méthodes traditionnelles, notamment grâce à l'intégration des relations multi-niveaux et à l'utilisation de la similarité des attributs pour chaque niveau. Ces travaux ouvrent la voie à une meilleure exploitation des données laitières pour des applications variées, allant de la sélection génétique au diagnostic des maladies.

Mots-clés : Similarité structurelle, Graphes de connaissances, Production laitière, SimRank, Analyse hiérarchique, Optimisation des troupeaux.

CHAPITRE 1
NOTIONS PRÉLIMINAIRES : SIMILARITÉ ENTRE VACHES BASÉE SUR LES DONNÉES DE PRODUCTION LAITIÈRES

Dans ce chapitre nous introduisons quelques calculs de similarité appliqués dans le domaine de la production laitière, et enfin nous aborderons les essentiels concernant les vaches, lactations et jours de tests.

1.1 Introduction aux calculs de similarité dans le domaine de la production laitière

Les mesures de similarité jouent un rôle fondamental dans l'analyse des données laitières, notamment pour évaluer les performances des vaches en fonction de divers indices. Ces mesures permettent d'identifier des similitudes structurelles et attributives entre individus ou cycles de production, favorisant ainsi l'optimisation des performances et l'amélioration de la productivité des exploitations.

Un premier exemple d'application des calculs de similarité est fourni par l'étude menée par Cruz et al. (Cruz *et al.*, 2021), qui compare les chaînes de production laitière des États de Minas Gerais et Paraná au Brésil sur la période 2008-2017. Cette étude adopte une méthodologie quantitative basée sur l'utilisation du logiciel SPSS pour la collecte et l'analyse des données. Les variables analysées incluent la production de lait, le nombre de vaches laitières, la productivité et la valeur de production dans chaque municipalité des deux États. L'ANOVA a été employée pour comparer les moyennes des variables et tester les différences entre les deux régions. Par ailleurs, une analyse de clustering a permis de regrouper les municipalités selon leur productivité en quatre groupes : très basse, basse, moyenne et haute. Les résultats montrent que le Paraná présente des limites de productivité plus élevées pour chaque groupe par rapport à Minas Gerais, témoignant d'une plus grande efficacité. Cependant, certaines municipalités très productives de Minas Gerais ne figurent pas parmi les groupes les plus performants du Paraná, ce qui s'explique par des seuils de productivité différents entre les deux régions.

Un second exemple d'application des calculs de similarité est fourni par Adamczyk et al. (Adamczyk *et al.*, 2017), qui se sont intéressés à l'évaluation de l'activité physique quotidienne des vaches laitières en fonction des conditions environnementales. Deux méthodes de clustering ont été utilisées : la méthode de Ward et les réseaux de Kohonen. La méthode de Ward, un algorithme de clustering hiérarchique, vise à minimiser la somme des carrés des écarts au sein des clusters, tandis que les réseaux de Kohonen reposent sur une carte

auto-organisée (SOM) utilisant un algorithme d'apprentissage topologique pour regrouper les observations. L'utilisation combinée de ces deux approches permet une vérification croisée des résultats. L'analyse a révélé trois groupes distincts d'activités pour chaque mois d'observation. Par exemple, en juin, les périodes d'activité étaient définies comme suit : minuit à 8 h, 9 h à 17 h, et 18 h à minuit. Les clusters obtenus avec la méthode de Ward étaient cohérents avec ceux des réseaux de Kohonen, renforçant ainsi la fiabilité des résultats. Cette analyse de clustering a démontré son efficacité pour classifier les activités physiques des vaches en groupes distincts, offrant une base pour une évaluation objective de leur bien-être. Les auteurs suggèrent que des recherches futures pourraient se concentrer sur l'analyse des comportements individuels dans un groupe et l'impact des conditions environnementales, en incluant des périodes de lactation complètes pour des résultats encore plus précis.

Ces deux études illustrent le potentiel des calculs de similarité dans divers contextes de la production laitière. Elles démontrent également l'importance des outils d'analyse statistique et de clustering pour mieux comprendre les performances et les comportements des vaches, tout en ouvrant la voie à des recherches futures pour affiner ces approches.

1.2 Définitions de vache, lactation et jours de test

1.2.1 Définition de vache

La vache est la femelle adulte du bovin domestique **(Bos taurus)**, un mammifère ruminant élevé principalement pour la production de lait, de viande ou comme animal de trait. Elle joue un rôle central dans l'agriculture mondiale, fournissant des produits essentiels à l'alimentation humaine et contribuant à l'économie rurale.

1.2.1.1 Caractéristiques principales

- **Reproduction :** Une vache atteint la maturité sexuelle vers 18 mois et peut être mise à la reproduction à cet âge. La gestation dure environ neuf mois, similaire à celle des humains. Après la naissance de son premier veau, généralement vers l'âge de deux ans, la vache entre en période de lactation (de l'Agriculture et de l'Alimentation, 2023).
- **Lactation :** La période de production laitière dure environ dix mois après chaque vêlage. Pendant cette période, une vache laitière peut produire en moyenne 28 litres de lait par jour, bien que ce

chiffre puisse varier en fonction de la race et des conditions d'élevage (in World Farming (CIWF), 2023).

— **Races :** Il existe de nombreuses races de vaches et au Canada, la race bovine laitière prédominante est la **Holstein**, représentant environ 93% du cheptel laitier national (du Canada, 2024). Cette race est privilégiée pour sa *production laitière exceptionnelle*, avec une moyenne de 10 994 kg de lait par lactation standard de 305 jours. Le lait produit contient en moyenne 4,13% de matière grasse et 3,35% de protéines, répondant ainsi aux exigences de l'industrie laitière canadienne (du Canada, 2024).

  Bien que la Holstein domine largement, d'autres races sont également présentes, notamment :

  — **Ayrshire**,
  — **Suisse Brune**,
  — **Jersey**,
  — **Guernsey**,
  — **Shorthorn laitière**,
  — et la **Canadienne**.

  La race **Canadienne**, unique au Canada, est particulièrement reconnue pour sa rusticité et la richesse de son lait en matières grasses et protéines, la rendant idéale pour la production fromagère (Wikipedia, 2024).

1.2.1.2 Terminologie associée

— **Génisse :** Femelle bovine de plus de 12 mois qui n'a pas encore vêlé (Générale, 2023).
— **Taureau :** Mâle reproducteur non castré.
— **Bœuf :** Mâle castré élevé principalement pour la production de viande.

Comprendre ces distinctions est essentiel pour une gestion efficace des troupeaux et une production optimisée, que ce soit pour le lait ou la viande.

1.2.2 Définition de lactation

La lactation est le processus par lequel les femelles mammifères produisent et sécrètent du lait à travers les glandes mammaires pour nourrir leurs petits. Chez la vache laitière, ce processus est essentiel pour la production laitière et suit plusieurs phases distinctes.

1.2.2.1 Phases de la lactation chez la vache laitière

- **Mammogenèse :** Développement des tissus mammaires durant la gestation.
- **Lactogenèse :** Initiation de la production laitière autour de la parturition.
- **Galactopoïèse :** Maintien de la production de lait pendant la période de lactation.
- **Involution :** Régression des tissus mammaires après le tarissement.

Chez la vache laitière, la période de lactation commence après le vêlage et dure en moyenne 305 jours. Cette durée peut varier en fonction des pratiques d'élevage et des objectifs de production(Wikipedia,).

1.2.2.1.1 La courbe de lactation

La production laitière suit généralement une courbe caractéristique :

- **Phase ascendante :** Augmentation de la production jusqu'à un pic, généralement atteint entre la 4^{e} et la 8^{e} semaine post-partum .
- **Pic de lactation :** Période de production maximale de lait.
- **Phase descendante :** Diminution progressive de la production jusqu'au tarissement.(Agri-Mutuel,)

1.2.3 Définition des jours de test

Les jours de test sont des journées spécifiques dédiées à la collecte de données sur la production laitière des vaches, essentielles pour évaluer et optimiser les performances des troupeaux.

1.2.3.1 Collecte des données

Lors de ces journées, des informations précises sont recueillies, notamment :

- **Quantité de lait produite :** Mesurée généralement en kilogrammes.
- **Composition du lait :** Analyse des taux de matières grasses, de protéines et d'autres composants.
- **Cellules somatiques :** Indicateurs de la santé mammaire et de la qualité du lait.

1.2.3.2 Fréquence des tests

La fréquence des jours de test varie selon les protocoles adoptés :

- **Contrôle mensuel :** Un test par mois, souvent réalisé par un agent d'un organisme de contrôle des

performances laitières.

- **Contrôle bimensuel :** Deux tests par mois, alternant entre l'éleveur et un agent.

Ces protocoles sont conformes aux recommandations de l'International Committee for Animal Recording (ICAR) et sont détaillés dans le Référentiel de Contrôle des Performances pour la production de lait de vache (de l'Élevage (Idele), 2020).

1.2.3.3 Méthodes de collecte

Les méthodes de collecte des données lors des jours de test incluent :

- **Pesée du lait :** Mesure précise de la production laitière lors de chaque traite.
- **Prélèvement d'échantillons :** Collecte d'échantillons de lait pour analyse en laboratoire.

Selon le protocole, ces opérations peuvent être effectuées par l'éleveur ou par un agent qualifié (de l'Élevage (Idele), 2020).

1.2.3.4 Utilisation des données

Les informations recueillies lors des jours de test sont utilisées pour :

- **Évaluer les performances individuelles :** Identifier les vaches les plus productives et celles nécessitant une attention particulière.
- **Améliorer la gestion du troupeau :** Adapter l'alimentation, la reproduction et les soins vétérinaires en fonction des données collectées.
- **Assurer la qualité du lait :** Maintenir des standards élevés en surveillant les indicateurs de santé et de qualité.

1.3 Méthodes de collecte des données : vaches, lactations et jours de test

Dans le cadre de la gestion des troupeaux laitiers au Canada, des méthodes rigoureuses et standardisées sont employées pour collecter les données relatives aux vaches, aux lactations et aux jours de test. Ces méthodes s'appuient sur des outils technologiques avancés et des institutions reconnues, telles que Lactanet.

1.3.1 Collecte des données sur les vaches

Les données relatives aux vaches sont collectées à travers des systèmes d'identification et des dispositifs connectés.

- **Identification individuelle** : Les vaches sont équipées de boucles auriculaires RFID ou de colliers connectés, permettant un suivi précis des animaux (lactanet, 2024).
- **Données démographiques et sanitaires** : Les informations incluent la date de naissance, la race, le poids, et l'historique médical. Ces données sont enregistrées dans des bases numériques comme DairyComp 305 ou dans les bases gérées par Lactanet (VAS, 2025; lactanet, 2024).
- **Capteurs IoT** : Les dispositifs tels que SenseHub surveillent en temps réel la santé, la reproduction, et les comportements des vaches (rumination, activité physique) (Dairy, 2024).

1.3.2 Collecte des données sur les lactations

Les cycles de lactation des vaches sont surveillés de manière systématique au Canada à travers des programmes officiels de contrôle laitier.

- **Contrôles de performance** : Les données sont collectées mensuellement via des échantillons de lait analysés pour déterminer la production quotidienne et la qualité du lait (matières grasses, protéines, cellules somatiques) (lactanet, 2024).
- **Événements reproductifs** : Les dates d'insémination, de vêlage, et de tarissement sont enregistrées dans des plateformes comme Lactanet ou directement via des dispositifs automatisés (lely, 2024).
- **Machines de traite robotisées** : Les systèmes comme Lely Astronaut A5 mesurent automatiquement la production et la composition du lait en temps réel (lely, 2024).

1.3.3 Collecte des données sur les jours de test

Les jours de test constituent des moments-clés pour le suivi des performances des troupeaux et la détection d'anomalies.

- **Échantillonnage régulier** : Chaque mois, des prélèvements sont réalisés sur chaque vache pour mesurer la production, la qualité, et la santé générale des animaux.
- **Systèmes de collecte automatisée** : Les capteurs connectés (comme SCR Heatime ou Moocall) permettent d'enregistrer en continu des données comportementales et physiologiques (CIAQ, 2024; Moocall, 2024).

— **Validation des données** : Les informations collectées sont comparées aux analyses réalisées par les laboratoires partenaires de Lactanet pour garantir leur exactitude (lactanet, 2024).

1.4 Importance de la similarité sur les données de production laitière

L'analyse de la similarité des données de production laitière est essentielle pour optimiser la gestion des troupeaux et améliorer la rentabilité des exploitations. En identifiant des schémas communs et des divergences dans les performances des vaches, les producteurs peuvent prendre des décisions éclairées concernant l'alimentation, la santé et la reproduction.

1.4.1 Optimisation de la gestion des troupeaux

En comparant les performances individuelles des vaches, il est possible de détecter rapidement les anomalies ou les baisses de production, permettant ainsi une intervention précoce. Par exemple, l'utilisation d'outils d'analyse des données facilite l'identification des vaches nécessitant une attention particulière, contribuant ainsi à une gestion plus efficace du troupeau (Noriap, 2023).

1.4.2 Amélioration de la rentabilité

L'analyse comparative des données de production aide à déterminer les pratiques d'élevage les plus efficaces. En identifiant les facteurs qui influencent positivement la production laitière, les éleveurs peuvent ajuster leurs stratégies pour maximiser la production et, par conséquent, les revenus (de Lait, 2023).

1.4.3 Prédiction des performances futures

L'utilisation de données historiques et de modèles prédictifs permet d'anticiper les tendances de production et d'adapter les pratiques en conséquence. Cela est particulièrement utile pour planifier les périodes de reproduction, l'alimentation et la gestion sanitaire du troupeau (Analytics, 2023).

En somme, l'analyse de la similarité des données de production laitière offre aux producteurs des outils puissants pour améliorer la performance globale de leurs exploitations, assurer le bien-être animal et répondre aux exigences du marché.

CHAPITRE 2

CONCEPTS D'ANALYSES STATISTIQUES DE SIMILARITÉ ET DE GRAPHES DE CONNAISSANCES

La similarité est une notion fondamentale dans de nombreux domaines, notamment la recherche d'information, la classification, et les systèmes de recommandation. Dans ce chapitre, nous explorons le concept de similarité et son intégration dans les graphes de connaissances, une approche puissante pour modéliser des données complexes, mais avant nous mettrons en relief les concepts de distribution statistique et les méthodes d'analyse.

2.1 Analyse et méthodes statistiques des données

L'analyse statistique joue un rôle central dans la compréhension, l'exploration et l'interprétation des données. Elle permet de dégager des tendances générales, de détecter des anomalies et de modéliser des phénomènes complexes. Cette section décrit les approches générales utilisées pour analyser les distributions des données et les méthodes statistiques appliquées.

2.1.1 Analyse des distributions des données

L'analyse des distributions est une étape essentielle pour comprendre la répartition des valeurs, identifier les asymétries ou les anomalies, et garantir la qualité des données. Cette analyse repose sur des outils graphiques et des mesures statistiques pour explorer les caractéristiques principales des variables numériques.

2.1.1.1 Méthodologie générale

L'exploration des distributions des données est réalisée à l'aide de courbes de densité (Kernel Density Estimation, KDE). Cette méthode non paramétrique fournit une estimation continue de la densité de probabilité, permettant une visualisation lissée des données (Silverman, 1986). Contrairement aux histogrammes classiques, les courbes KDE facilitent l'identification des asymétries, des pics ou des valeurs aberrantes.

Étapes principales :

— **Préparation des données :** Les variables non numériques, incohérentes ou contenant des valeurs manquantes sont exclues ou traitées pour éviter les biais dans l'analyse.

— **Visualisation :** Les courbes KDE sont utilisées pour examiner la répartition des valeurs et repérer des distributions multimodales ou asymétriques.
— **Interprétation :** L'analyse graphique met en évidence les tendances générales et guide les étapes suivantes de l'analyse.

2.1.1.2 Applications et observations générales

L'analyse des distributions permet de :

— Identifier des **asymétries** ou des **distributions multimodales**, révélatrices de groupes sous-jacents ou de variabilités importantes.
— Détecter des **valeurs extrêmes** (outliers) susceptibles d'affecter les résultats des analyses statistiques.
— Vérifier la **normalité des distributions**, essentielle pour l'application de nombreuses méthodes statistiques (Montgomery, 2017).

Cette étape est cruciale pour préparer les données avant des analyses plus avancées, comme les tests statistiques ou les calculs de similarité.

2.1.2 Méthodes générales d'analyse statistique

Les méthodes statistiques permettent d'approfondir l'interprétation des données en examinant les relations, la variabilité et la structure des variables. Elles offrent des outils pour modéliser des phénomènes complexes et prendre des décisions éclairées.

2.1.2.1 Statistiques descriptives

Les statistiques descriptives constituent la base de toute analyse, fournissant une vue d'ensemble des données. Elles incluent des mesures de tendance centrale et de dispersion.

— **Moyenne :** Représente la valeur moyenne des données (Montgomery, 2017).
— **Écart-type :** Quantifie la dispersion des données autour de la moyenne (Carr *et al.*, 1996).
— **Minima et maxima :** Identifient les valeurs extrêmes, utiles pour détecter les anomalies.
— **Distribution :** Fournit des informations sur la répartition générale des données (Hastie *et al.*, 2009).

Ces mesures sont essentielles pour structurer les données avant des analyses plus complexes.

2.1.2.2 Analyse de corrélation

La corrélation examine les relations linéaires entre deux variables. Le coefficient de corrélation (r) quantifie la force et la direction de cette relation.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{2.1}$$

Applications générales :

- Identifier les relations entre variables pour orienter les modèles prédictifs (Hastie *et al.*, 2009).
- Vérifier la redondance entre les variables pour simplifier les analyses.

Un coefficient $r > 0.8$ reflète une forte corrélation positive, tandis qu'un $r < -0.8$ indique une forte corrélation négative (Montgomery, 2017).

2.1.2.3 Clustering (classification non supervisée)

Le clustering regroupe les observations en sous-ensembles homogènes appelés clusters. Cette méthode non supervisée est utile pour explorer la structure des données (Han *et al.*, 2011).

Algorithmes courants :

- **k-means :** Partitionne les données en k clusters en minimisant la variance intra-cluster.
- **Clustering hiérarchique :** Structure les données en une arborescence, facilitant leur visualisation.

Ces techniques sont largement utilisées dans les domaines de la biologie, du marketing et de l'analyse de données.

2.1.2.4 Réduction de dimensions (PCA)

L'analyse en composantes principales (PCA) réduit la dimensionnalité des données tout en conservant l'essentiel de leur variance (Jolliffe et Cadima, 2016). Elle projette les données dans un nouvel espace où chaque composante est une combinaison linéaire des variables initiales.

Applications générales :

- Identifier les dimensions les plus influentes.
- Faciliter la visualisation des données dans un espace réduit (souvent en 2D ou 3D).

2.1.3 Implications pour l'analyse statistique

L'analyse des distributions et les méthodes statistiques influencent directement les étapes ultérieures :

- Elles garantissent une meilleure préparation des données pour les analyses avancées.
- Elles fournissent une base solide pour sélectionner les tests appropriés (paramétriques ou non paramétriques).
- Elles assurent la robustesse et la fiabilité des résultats finaux.

2.2 Notion de similarité

2.2.1 Définition et fondements théoriques

la similarité quantifie à quel point deux objets sont semblables ou proches. Une mesure de similarité élevée indique que les objets partagent des caractéristiques communes selon des critères prédéfinis. Par exemple, si deux documents présentent une similarité élevée, cela peut signifier qu'ils abordent des sujets similaires ou partagent un vocabulaire similaire. Par ailleurs la dissimilarité quand à elle quantifie à quel point deux objets sont différents ce qui fait qu'elle est différente de la similarité malgré le fait que les deux quantifie la relation entre deux objets.

2.2.2 Types de données

Avant d'aborder les exemples de similarité , il est essentielle de connaître les types de données existants car le cadre de l'analyse des données et du choix de la bonne mesure de similarité. , il est essentiel de comprendre les différents types de données utilisées. Chaque type de donnée possède des propriétés uniques qui influencent les méthodes d'analyse et de calcul. Dans cette partie , nous décrivons les principales catégories de données avec des exemples.

2.2.2.1 Données nominales

Les données nominales, également appelées catégoriques, servent à classer des objets dans des catégories distinctes sans ordre ou hiérarchie. Ces données ne permettent que des comparaisons d'égalité.

Exemples :

- Les couleurs des voitures (rouge, bleu, vert).
- Le sexe d'une personne (homme, femme, autre).
- Les types de produits dans un magasin (alimentaire, électronique, vêtements).

Selon (Tan *et al.*, 2019b), ces données sont souvent codées par des valeurs numériques arbitraires pour faciliter leur traitement, bien que ces valeurs n'aient aucune signification mathématique.

2.2.2.2 Données ordinales

Les données ordinales impliquent un ordre ou un classement entre les catégories, mais les différences entre les valeurs ne sont pas nécessairement uniformes ou interprétables.

Exemples :

- Les niveaux d'éducation (primaire, secondaire, universitaire).
- Les évaluations de satisfaction (très insatisfait, insatisfait, neutre, satisfait, très satisfait).
- Les tailles de vêtements (S, M, L, XL).

Comme indiqué par (Han *et al.*, 2022), ces données permettent des analyses basées sur le classement, mais pas sur des opérations mathématiques directes comme l'addition.

2.2.2.3 Données quantitatives

Les données quantitatives représentent des valeurs numériques et peuvent être subdivisées en deux sous-catégories : les données discrètes et continues.

2.2.2.3.1 Données discrètes

Les données discrètes sont des nombres entiers représentant des objets comptables. Ces valeurs ne peuvent pas être fractionnées.

Exemples :

- Le nombre de livres empruntés dans une bibliothèque.
- Le nombre de machines dans une usine.
- Le nombre de jours dans une semaine.

2.2.2.3.2 Données continues

Les données continues peuvent prendre n'importe quelle valeur dans un intervalle donné, ce qui les rend idéales pour mesurer des quantités physiques.

Exemples :

- La température en degrés Celsius.
- La hauteur d'une personne en centimètres.
- La durée d'un trajet en heures.

D'après (James *et al.*, 2023), ces données nécessitent des outils statistiques spécifiques pour leur analyse en raison de leur précision et de leur variabilité.

2.2.2.4 Données binaires

Les données binaires sont des cas particuliers de données nominales avec seulement deux catégories possibles, souvent représentées par 0 et 1.

Exemples :

- Réponse à une question (oui ou non).
- État d'un dispositif (activé ou désactivé).
- Présence d'une condition médicale (présent ou absent).

Comme le mentionne (Aggarwal, 2015), ces données sont particulièrement utiles dans les modèles de classification binaire.

2.2.2.5 Conclusion

Chaque type de donnée possède des propriétés uniques qui influencent les méthodes d'analyse et les résultats obtenus. Une bonne compréhension des différences entre ces catégories est essentielle pour choisir

les outils appropriés et interpréter correctement les résultats.

2.2.3 Exemples de mesures de similarité

Dans cette partie nous allons aborder quelques exemples spécifique de mesures de similarité.

2.2.3.1 Coefficient de similarité simple

Le coefficient de similarité simple (Simple Matching Coefficient - SMC) est utilisé pour des attributs binaires symétriques. Il prend en compte les deux situations où les objets ont la même valeur, soit 1 ou 0. Il est calculé par la formule suivante :

$$SMC = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} \tag{2.2}$$

où f_{00} représente le nombre de paires où les attributs des deux objets sont égaux à 0, f_{11} est le nombre de paires où ceux-ci sont égaux à 1, f_{01} et f_{10} indiquent les cas où les objets ont des valeurs différentes.

Cette mesure est adaptée pour des attributs binaires équilibrés (Tan *et al.*, 2019a).

2.2.3.2 Coefficient de Jaccard

Le coefficient de Jaccard est utilisé pour des attributs binaires asymétriques, où la présence (valeur 1) est plus significative que l'absence (valeur 0).

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \tag{2.3}$$

où f_{11} est le nombre de paires où les attributs des deux objets sont égaux à 1, tandis que f_{01} et f_{10} sont les cas où les objets ont des valeurs différentes.

Cette mesure est particulièrement utile pour des données où l'absence d'une caractéristique n'est pas significative (Jaccard, 1901).Cette méthode est aussi couramment utilisée pour analyser des textes afin de détecter des duplications ou des similitudes, notamment dans des bases de données textuelles ou lors de l'alignement de séquences génomiques (Han *et al.*, 2011).

2.2.3.3 Distance inversée

La similarité numérique utilise souvent la distance inversée pour comparer deux valeurs numériques. Plus les valeurs sont proches, plus la similarité est élevée (Hastie *et al.*, 2009).

$$\text{Sim}_{\text{numérique}} = \frac{1}{1 + |x_1 - x_2|} \tag{2.4}$$

Cette approche est particulièrement utile dans les systèmes de recommandation personnalisés, où elle permet de baser les recommandations sur des préférences numériques. Elle est également largement employée pour segmenter des données quantitatives en groupes homogènes (Hastie *et al.*, 2009).

2.2.3.4 Décroissance exponentielle

Pour des données temporelles, la similarité est calculée selon une décroissance exponentielle basée sur la différence en jours (Box et Jenkins, 1970).

$$\text{Sim}_{\text{dates}} = e^{-\alpha \cdot |d_1 - d_2|} \tag{2.5}$$

Cette méthode est fréquemment utilisée dans l'analyse des séries temporelles, notamment dans des études économiques et pour modéliser les comportements utilisateurs dans des systèmes de recommandation (Box et Jenkins, 1970).

2.2.3.5 Identité

Pour des données booléennes, une simple comparaison binaire est utilisée (Russell et Norvig, 2016) :

$$\text{Sim}_{\text{booléen}} = \begin{cases} 1 & \text{si } b_1 = b_2, \\ 0 & \text{sinon.} \end{cases} \tag{2.6}$$

Cette approche est très utilisée dans les analyses logiques ou catégorielles, par exemple pour comparer des réponses oui/non dans des enquêtes ou des classifications binaires (Russell et Norvig, 2016).

2.2.3.6 Listes ou ensembles : moyenne des similarités élémentaires

Lorsque des données sont sous forme de listes ou d'ensembles, la similarité est calculée comme la moyenne des similarités entre chaque paire d'éléments correspondants (Han *et al.*, 2011).

$$\text{Sim}_{\text{listes}} = \frac{1}{n} \sum_{i=1}^{n} \text{Sim}(x_i, y_i) \tag{2.7}$$

Cette méthode est particulièrement adaptée pour analyser des ensembles complexes dans des bases de données ou comparer des structures hiérarchiques comme des arbres ou des graphes (Han *et al.*, 2011).

2.2.3.7 Distance cosinus (Cosine Similarity)

La similarité cosinus mesure la similarité entre deux vecteurs, généralement utilisée pour des données de haute dimension, telles que des représentations de documents.

$$\text{Sim}_{\cos}(A, B) = \frac{A \cdot B}{||A||\,||B||} \tag{2.8}$$

où $A \cdot B$ représente le produit scalaire des vecteurs A et B, et $||A||$ et $||B||$ sont les normes des vecteurs A et B.

Cette mesure est souvent utilisée pour évaluer la similarité entre des textes, car elle ignore la longueur des vecteurs et se concentre sur l'orientation (Singhal, 2001).

2.2.3.8 Distance Euclidienne

La distance euclidienne est une mesure classique de dissimilarité pour des attributs d'intervalle ou de ratio.

$$d(A, B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{2.9}$$

où a_i et b_i sont les valeurs des attributs des objets A et B.

Elle représente la "distance directe" entre deux points dans un espace multidimensionnel (Han *et al.*, 2011).

2.2.3.9 Distance de Manhattan

La distance de Manhattan est une mesure de dissimilarité qui utilise la somme des différences absolues entre les attributs des objets.

$$d_{\text{Manhattan}}(A, B) = \sum_{i=1}^{n} |a_i - b_i| \tag{2.10}$$

Cette mesure est souvent utilisée quand le déplacement est restreint à des chemins orthogonaux, comme dans un réseau de rues d'une ville. Elle est aussi plus robuste aux valeurs aberrantes que la distance euclidienne (Duda *et al.*, 2000).

2.2.4 La mesure de similarité mixte pour le clustering des données mixtes

Le regroupement de données ou *clustering* est une technique utilisée pour organiser des ensembles de données en groupes homogènes, appelés clusters, en fonction de leurs similarités. Cependant, le traitement des

ensembles de données mixtes, qui incluent des attributs numériques, catégoriques, binaires et ordinaux, pose des défis importants. La méthode **Mixed Similarity Measure (MSM)** représente une avancée clé pour relever ces défis en combinant différentes mesures de similarité adaptées à chaque type d'attribut, offrant une solution robuste et flexible pour le clustering des données mixtes (S. Ali *et al.*, 2017).

2.2.4.1 Contexte et limites des méthodes traditionnelles

Les approches traditionnelles comme K-means et K-modes se limitent respectivement aux données numériques et catégoriques. Par exemple, **K-means** utilise des moyennes pour déterminer les centres des clusters, ce qui est efficace pour des attributs continus mais inadapté aux données discrètes (Wu *et al.*, 2008). De son côté, **K-modes** remplace les moyennes par des modes pour traiter les données catégoriques, mais ne peut gérer des attributs numériques sans transformation préalable (Ammar *et al.*, 2012).

Pour surmonter les limitations de ces approches, deux stratégies ont été explorées :

1. **Conversion des données mixtes** en données homogènes, une solution qui entraîne une perte d'information significative (Shih *et al.*, 2010).
2. **Division des ensembles mixtes** en sous-ensembles homogènes (numériques et catégoriques), suivie de leur regroupement. Cette méthode, bien qu'efficace, est complexe et coûteuse en temps pour des ensembles volumineux (Ahmad et Dey, 2007).

Ces approches ont révélé leurs limites, motivant les chercheurs à proposer des méthodes adaptées à la nature mixte des données.

2.2.4.2 La MSM, approche et fonctionnement

La méthode **MSM**, proposée par Doaa S. Ali et ses collègues (S. Ali *et al.*, 2017), combine les avantages des algorithmes K-means et K-modes en introduisant des mesures de similarité spécifiques pour chaque type d'attribut. Les étapes principales de l'algorithme sont :

1. **Assignation initiale** : Chaque élément de données est assigné aléatoirement à un cluster.
2. **Calcul des distances** : Une mesure de similarité appropriée est appliquée pour chaque type d'attribut :
   - **Données binaires et nominales** : Distance d'appariement (Aranganayagi et Thangavel, 2009).
   - **Données ordinales** : Normalisation des écarts entre les valeurs (Mukhopadhyay et Maulik, 2007).

— **Données numériques** : Standardisation des valeurs pour une compatibilité inter-attributs.

3. **Mise à jour des centres** : Pour les attributs numériques, la règle de K-means est appliquée (moyenne). Pour les attributs catégoriques, la règle de K-modes est utilisée (valeur modale).
4. **Itérations** : Le processus est répété jusqu'à stabilisation des clusters.

2.2.4.3 Résultats expérimentaux et comparaison avec d'autres méthodes

Les performances de MSM ont été évaluées sur six ensembles de données mixtes issus de la *UCI Machine Learning Repository*. Les résultats montrent que :

— En configuration **non-évolutive**, MSM surpasse les méthodes traditionnelles (Matching, IOF, Eskin, Scaling) dans **80 % des cas**.

— En configuration **évolutive** avec optimisation par évolution différentielle (DE), MSM améliore encore ses performances, obtenant des résultats significativement meilleurs dans **90 % des cas**.

Ces résultats confirment l'efficacité de MSM, en particulier dans les situations où les ensembles de données présentent une grande diversité de types d'attributs.

2.2.4.4 Contributions Scientifiques et Perspectives

La méthode MSM marque une avancée notable dans le domaine du clustering :

— **Flexibilité** : L'intégration de mesures spécifiques pour chaque type d'attribut permet une meilleure représentation des données mixtes.

— **Efficacité** : Les résultats expérimentaux montrent une précision accrue par rapport aux méthodes existantes, tout en maintenant une complexité computationnelle raisonnable.

Pour l'avenir, les auteurs suggèrent d'explorer des extensions de MSM dans un cadre multiobjectif (Parameswari *et al.*, 2015) ou pour des données incertaines (Baghshah, 2009).

2.3 Introduction aux graphes de connaissances

Les graphes de connaissances (*Knowledge Graphs*, *KG*) constituent une structure puissante pour organiser et représenter des informations dans un format interconnecté, orienté vers les relations entre les entités. Ces graphes, formés de nœuds (représentant des entités) et d'arêtes (représentant des relations), ont été popularisés par des applications telles que le *Knowledge Graph* de Google en 2012 (Singhal, 2012), qui visait à améliorer la recherche en offrant des réponses contextuelles enrichies.

2.3.1 Définition et structure

Un graphe de connaissances est défini comme une représentation sémantique formée de triplets (s, r, o) où :

- s (sujet) et o (objet) sont des entités,
- r est une relation liant ces entités.

Ces triplets permettent de modéliser des connaissances sous forme de réseaux, reliant les entités dans un cadre explicite. Les entités et relations sont souvent enrichies par des métadonnées ou des propriétés qui décrivent leurs caractéristiques (Hogan *et al.*, 2022).

2.3.2 Applications

Les graphes de connaissances sont devenus incontournables dans plusieurs domaines :

1. **Recherche d'information** : Ils permettent des résultats contextuels, comme ceux fournis par les moteurs de recherche (Dong *et al.*, 2014).
2. **Systèmes de recommandation** : En exploitant les relations entre utilisateurs et produits pour améliorer les suggestions (Zhang *et al.*, 2016).
3. **Traitement du langage naturel (NLP)** : Ils sont utilisés pour des tâches telles que l'extraction d'entités ou l'analyse sémantique (Ji *et al.*, 2022).
4. **Sciences biomédicales** : Les graphes comme Bio2RDF structurent les connaissances sur les interactions biologiques et médicales (Belleau *et al.*, 2008).

2.3.3 Méthodes de construction

La création d'un graphe de connaissances peut être réalisée de plusieurs manières :

1. **Extraction de données structurées** : Par exemple, à partir de bases de données relationnelles ou de formats RDF (Auer *et al.*, 2007).
2. **Extraction de données non structurées** : Les textes non structurés sont une source importante de triplets via des techniques de traitement du langage naturel (Soares *et al.*, 2019).
3. **Fusion de sources multiples** : Des projets comme DBpedia ou YAGO intègrent des informations issues de bases de données multiples et hétérogènes (Suchanek *et al.*, 2007).

2.3.4 Défis et perspectives

Malgré leurs promesses, les graphes de connaissances présentent plusieurs défis :

- **Scalabilité** : La gestion des graphes de grande taille, tels que les KGs du web sémantique, nécessite des solutions adaptées (Nickel *et al.*, 2016).
- **Qualité des données** : Les graphes extraits automatiquement peuvent contenir des erreurs ou des relations ambiguës (Paulheim, 2016).
- **Interopérabilité** : La fusion de sources hétérogènes nécessite des standards communs pour garantir la cohérence.

Les recherches futures se concentrent sur l'amélioration des méthodes d'extraction, l'intégration de nouvelles sources comme les données de capteurs, et le développement de graphes dynamiques capables d'évoluer en temps réel.

2.4 La similarité dans les graphes de connaissances

La mesure de similarité dans les graphes de connaissances est essentielle pour diverses applications, telles que la recherche d'information, les systèmes de recommandation, et l'inférence logique. Étant donné la structure des graphes, les techniques de similarité combinent des approches basées sur les entités, les relations, et les représentations vectorielles des nœuds.

2.4.1 Définitions et approches

La similarité dans les graphes de connaissances peut être mesurée à plusieurs niveaux :

1. **Similarité structurelle** : Elle évalue les similitudes entre deux nœuds en fonction de leur contexte structurel, par exemple à l'aide de mesures comme SimRank (Jeh et Widom, 2002b), qui calcule la similarité sur la base de la proximité des voisins.
2. **Similarité sémantique** : Cette approche exploite la signification des entités et des relations en utilisant des méthodes telles que les embeddings, comme TransE (Bordes *et al.*, 2013), qui projette les entités et relations dans un espace vectoriel continu.
3. **Similarité contextuelle** : Elle combine des informations structurelles et sémantiques pour capturer des relations complexes entre les entités dans un graphe (Nickel *et al.*, 2016).

2.4.2 Applications

2.4.2.1 Recherche d'information :

La similarité permet d'améliorer les résultats en trouvant les nœuds pertinents par rapport à une requête donnée, même si ces nœuds ne sont pas directement reliés (Jeh et Widom, 2002b). SimRank, par exemple, est largement utilisé dans ce domaine pour établir des scores de pertinence.

2.4.2.2 Systèmes de recommandation :

Les systèmes de recommandation basés sur les graphes utilisent les similarités calculées entre utilisateurs et produits pour améliorer la personnalisation (Bordes *et al.*, 2013). TransE et ses variantes ont montré des performances élevées dans ce domaine en apprenant des représentations vectorielles qui respectent la structure du graphe.

2.4.2.3 Analyse des connaissances :

Les graphes de connaissances facilitent la déduction d'informations implicites à travers des mesures de similarité sémantique et contextuelle (Nickel *et al.*, 2016).

2.4.3 Défis et perspectives

La similarité dans les graphes de connaissances présente des défis importants :

- **Scalabilité :** Les grands graphes nécessitent des algorithmes efficaces pour calculer des similarités en temps réel.
- **Qualité des données :** Les mesures de similarité sont sensibles aux erreurs ou incohérences dans les graphes.
- **Modélisation hybride :** Les approches futures pourraient combiner des méthodes structurelles et sémantiques pour des résultats plus robustes.

Les recherches futures incluent le développement de modèles basés sur l'apprentissage profond pour calculer la similarité de manière plus efficace et précise, ainsi que l'intégration de nouvelles sources de données pour enrichir les graphes.

2.5 Conclusion

Dans ce chapitre, nous avons introduit les concepts de similarité et de graphes de connaissances, ainsi que leur synergie. Ces notions constituent une base essentielle pour les chapitres suivants, où nous examinerons leurs applications pratiques et les algorithmes spécifiques. Nous avons aborder également le concept d'analyse des distributions et les méthodes statistiques qui offrent une compréhension approfondie des données. Elles sont essentielles pour détecter des anomalies, modéliser des relations et garantir la qualité des analyses dans divers contextes scientifiques.

CHAPITRE 3
ÉTAT DE L'ART ET PROBLÉMATIQUE

Le calcul de similarité entre graphes est un sujet clé dans les domaines où les relations structurelles entre les entités jouent un rôle déterminant, comme les réseaux sociaux, la bio-informatique, et l'analyse des interactions dans les systèmes complexes. Les approches modernes exploitent des techniques d'apprentissage profond et des méthodes traditionnelles pour relever les défis liés à la diversité des graphes (taille, étiquetage, hétérogénéité) et à la complexité computationnelle. Dans ce chapitre, nous aborderons les revues de littérature sur la recherche portant sur les techniques basées sur les embeddings de graphes, les réseaux de neurones de graphes (GNN), les techniques d'appariement de sous-structures et les kernels de graphes profonds également les methodes de calcul de similarité structurelle. Enfin, nous définirons notre problématique de recherche.

3.1 Modèles d'embedding de graphes

Les modèles d'embedding de graphes visent à représenter les graphes sous forme de vecteurs dans un espace vectoriel de faible dimension tout en préservant leurs propriétés structurelles et attributives. Ces représentations permettent d'utiliser des métriques standards de similarité, comme la distance euclidienne ou le cosinus, pour comparer les graphes. Les embeddings de graphes sont particulièrement utiles pour réduire la complexité computationnelle des tâches impliquant des graphes tout en maintenant une grande précision.

3.1.1 Principe des embeddings de graphes

Un embedding de graphe peut être défini comme une fonction $f : G \rightarrow \mathbb{R}^d$, où G est un graphe et d est la dimension de l'espace d'embedding. La fonction f est apprise de manière à ce que les graphes similaires dans leur structure ou leurs attributs soient proches dans l'espace des embeddings.

Les embeddings peuvent être calculés à différents niveaux :

- **Niveau nœud :** Les représentations vectorielles sont apprises pour chaque nœud en tenant compte de ses voisins et de ses attributs.
- **Niveau graphe :** Une représentation globale est apprise pour chaque graphe en agrégeant les infor-

mations provenant de tous ses nœuds et arêtes.

3.1.2 Modèles d'embedding de niveau nœud

Les embeddings de niveau nœud se concentrent sur la création de représentations pour chaque nœud individuel en fonction de son voisinage. Ces méthodes sont adaptées lorsque l'objectif est de comparer des parties locales de graphes.

3.1.2.0.1 Node2Vec et DeepWalk

Node2Vec (Grover et Leskovec, 2016) et DeepWalk (Perozzi *et al.*, 2014) sont des méthodes pionnières pour l'apprentissage d'embeddings de nœuds. Elles s'inspirent des techniques de traitement du langage naturel (NLP) comme Word2Vec :

- Les marches aléatoires (*random walks*) sont générées à partir des graphes pour capturer les relations entre nœuds.
- Ces séquences de nœuds sont traitées comme des phrases dans un corpus de texte, et un modèle Word2Vec est utilisé pour apprendre les embeddings des nœuds.

Ces méthodes capturent efficacement les relations locales entre les nœuds, mais elles peuvent être limitées pour représenter des structures globales complexes.

3.1.2.0.2 GraphSAGE

GraphSAGE (*Graph Sample and Aggregate*) (Hamilton *et al.*, 2018) améliore les embeddings de nœuds en introduisant une approche inductive où les représentations des nœuds sont calculées en agrégeant les informations provenant de leurs voisins. Ce modèle est particulièrement adapté aux graphes évolutifs où de nouveaux nœuds ou arêtes peuvent être ajoutés.

3.1.3 Modèles d'embedding de niveau graphe

Les embeddings de niveau graphe visent à produire une représentation globale pour chaque graphe, ce qui est crucial pour des tâches comme la classification de graphes ou la comparaison globale.

3.1.3.1 Graph2Vec

Graph2Vec (Narayanan *et al.*, 2017) est une extension des techniques d'embedding de nœuds au niveau des graphes. Les étapes principales incluent :

- Extraction des sous-structures locales de chaque graphe, souvent en utilisant des sous-graphes ou des motifs fréquents.
- Utilisation de modèles comme Word2Vec pour apprendre des représentations pour ces sous-structures.
- Agrégation des embeddings des sous-structures pour produire une représentation globale du graphe.

Cette méthode est particulièrement utile pour des graphes contenant des motifs récurrents ou des structures homogènes.

3.1.3.2 DiffPool

DiffPool (*Differentiable Pooling*) (Ying *et al.*, 2019) introduit une approche hiérarchique pour générer des embeddings globaux. Ce modèle utilise des mécanismes de regroupement différentiables pour simplifier les graphes tout en conservant leurs propriétés essentielles :

- Chaque couche du modèle regroupe les nœuds similaires en un super-nœud.
- Les représentations finales des super-nœuds sont combinées pour produire un embedding global du graphe.

DiffPool est particulièrement efficace pour traiter des graphes complexes avec des structures hiérarchiques.

3.1.4 Applications des modèles d'embedding de graphes

Les modèles d'embedding de graphes sont utilisés dans une variété de domaines :

- **Bioinformatique :** Identification de structures similaires dans des réseaux biologiques ou moléculaires.
- **Vision par ordinateur :** Comparaison de graphes représentant des images ou des scènes.
- **Réseaux sociaux :** Analyse et regroupement de communautés dans des graphes sociaux.
- **Détection de fraudes :** Analyse des réseaux de transactions pour identifier des comportements suspects.

Les modèles d'embedding de graphes offrent une approche puissante et flexible pour représenter des graphes dans des espaces vectoriels tout en capturant leurs propriétés structurelles et attributives. Ils constituent une base essentielle pour de nombreuses tâches impliquant des graphes, en particulier dans

des domaines nécessitant une analyse précise et rapide des similarités.

3.2 Méthodes basées sur les réseaux de neurones de graphes

Les réseaux de neurones de graphes (graph neural networks, GNN) constituent une avancée significative dans le traitement des données structurées sous forme de graphes. Ces modèles exploitent les relations locales et globales entre les nœuds d'un graphe pour apprendre des représentations riches et adaptées à des tâches complexes telles que la classification, la prédiction de liens, et le calcul de similarité entre graphes. Les GNN se distinguent par leur capacité à combiner l'information structurelle des graphes avec des attributs propres aux nœuds ou aux arêtes.

3.2.1 Architecture et fonctionnement des GNN

L'architecture des GNN repose sur un processus itératif de propagation de messages (*message passing*) entre les nœuds d'un graphe. Ce processus est organisé en plusieurs couches, chacune permettant aux nœuds de mettre à jour leur état en fonction des informations reçues de leurs voisins.

Soit un graphe $G = (V, E)$, où V représente l'ensemble des nœuds et E l'ensemble des arêtes. Chaque nœud $v \in V$ est initialisé avec une représentation $h_v^{(0)}$, souvent dérivée de ses attributs. À chaque couche l, la représentation $h_v^{(l+1)}$ du nœud v est calculée comme suit :

$$h_v^{(l+1)} = \sigma\left(W^{(l)} \cdot \mathsf{AGG}\big(\{h_u^{(l)} : u \in \mathcal{N}(v)\}\big)\right) \tag{3.1}$$

,

où :

- $\mathcal{N}(v)$ est l'ensemble des voisins de v,
- AGG est une fonction d'agrégation, comme la somme, la moyenne, ou le maximum,
- $W^{(l)}$ est une matrice de poids apprise pour la couche l,
- σ est une fonction d'activation, comme ReLU ou Sigmoïde.

Le résultat final est une représentation pour chaque nœud ou pour l'ensemble du graphe, qui encode à la fois les informations locales et globales.

3.2.2 GNN pour le calcul de similarité de graphes

Pour évaluer la similarité entre graphes, les GNN ont été adaptés de plusieurs façons, notamment en utilisant des architectures spécifiques comme les GNN-CNN, les GNN siamois, et SimGNN.

3.2.2.1 GNN-CNN

Les modèles GNN-CNN combinent des couches GNN et CNN pour capturer les relations locales et globales entre graphes. Les représentations des nœuds produites par les couches GNN sont organisées dans une matrice d'interaction, où chaque cellule encode la similarité entre un nœud du premier graphe et un nœud du second. Cette matrice est ensuite traitée par un CNN pour détecter des motifs répétitifs et produire un score global de similarité. Ces modèles sont particulièrement efficaces dans des domaines comme la reconnaissance d'images, où les points d'intérêt des images sont représentés sous forme de graphes (Ma *et al.*, 2021).

3.2.2.2 GNN Siamois

Les architectures siamoises consistent en deux GNN identiques partageant les mêmes poids. Chaque graphe d'entrée est traité indépendamment pour produire des embeddings globaux, qui sont ensuite comparés à l'aide d'une fonction de similarité, comme la distance euclidienne ou le produit scalaire :

$$\text{Sim}(G_1, G_2) = f(h_{G_1}, h_{G_2}) \tag{3.2}$$

où h_{G_1} et h_{G_2} sont les embeddings des graphes G_1 et G_2, respectivement. Cette approche garantit une symétrie stricte dans le traitement des graphes et est adaptée à des tâches comme la détection de motifs similaires dans des bases de données de graphes (Ma *et al.*, 2021).

3.2.2.3 SimGNN

SimGNN est un modèle avancé qui combine des mécanismes globaux et locaux pour évaluer la similarité entre graphes (Bai *et al.*, 2020). Il intègre deux composants principaux :

— Un mécanisme d'attention global permettant de pondérer les nœuds en fonction de leur importance relative dans chaque graphe. Cela aide à focaliser l'analyse sur les régions les plus significatives des graphes.

— Une analyse fine des interactions locales entre les nœuds correspondants des deux graphes. Ces interactions sont ensuite agrégées pour produire un score de similarité.

SimGNN est particulièrement adapté pour des graphes où les relations structurelles jouent un rôle déterminant dans la définition de la similarité.

3.2.3 Applications des GNN pour la similarité de graphes

Les modèles basés sur les GNN sont devenus incontournables dans de nombreux domaines :

— En bioinformatique, pour identifier des molécules similaires représentées comme des graphes d'atomes.

— En vision par ordinateur, pour comparer des points clés extraits d'images et représentés sous forme de graphes.

— En analyse de réseaux sociaux, pour détecter des sous-communautés similaires dans des graphes de relations sociales.

— En détection de fraudes, pour analyser des réseaux de transactions et identifier des motifs récurrents.

Grâce à leur flexibilité et leur puissance, les GNN permettent une modélisation précise et efficace des relations complexes dans les graphes.

3.3 Méthodes basées sur l'appariement de sous-structures

Les méthodes basées sur l'appariement de sous-structures visent à simplifier la comparaison de graphes complexes en les décomposant en sous-graphes ou en groupes de sous-structures. Ces approches permettent de capturer des informations locales tout en réduisant la complexité computationnelle des comparaisons directes entre graphes entiers. Elles exploitent souvent des techniques avancées comme les hypergraphes ou des regroupements pour maximiser la précision et l'efficacité.

3.3.1 Principe général

L'idée centrale des méthodes d'appariement de sous-structures est de décomposer un graphe $G = (V, E)$ en un ensemble de sous-graphes $\{G_1, G_2, \ldots, G_k\}$, où chaque sous-graphe G_i représente une partie de la structure globale du graphe d'origine. Ces sous-graphes sont ensuite comparés en utilisant des métriques spécifiques ou des techniques d'apprentissage pour produire des scores de similarité locaux. Les scores locaux sont agrégés pour obtenir une estimation globale de la similarité entre deux graphes.

Les sous-structures peuvent inclure :

— Des sous-graphes induits basés sur les motifs récurrents.

— Des partitions ou regroupements des nœuds du graphe.

— Des hypergraphes représentant des relations complexes entre les sous-graphes.

3.3.2 H2MN : Hypergraph to Match Networks

Une méthode représentative de cette catégorie est le modèle **H2MN** (*Hypergraph to Match Networks*) (Yang *et al.*, 2024). Ce modèle décompose un graphe en un ensemble de sous-structures locales, qui sont ensuite représentées sous forme d'hypergraphes. Les hypergraphes sont construits de manière à ce que chaque hyperarête relie un groupe de nœuds ayant des interactions significatives dans le graphe d'origine.

Les étapes principales incluent :

— **Construction des hypergraphes :** Les sous-structures locales du graphe sont identifiées et représentées comme des hyperarêtes dans un hypergraphe, où chaque nœud représente une sous-structure.

— **Propagation de messages :** Un réseau neuronal d'hypergraphes est utilisé pour apprendre des représentations riches des nœuds et des arêtes.

— **Regroupement et simplification :** Une couche de regroupement réduit la complexité computationnelle en fusionnant des hypergraphes similaires.

Cette approche permet d'atteindre un compromis efficace entre précision et rapidité, ce qui en fait une méthode adaptée à des graphes de grande taille et à des applications complexes.

3.3.3 Graph Matching Networks et attention croisée

Les **Graph Matching Networks** (GMN) (Li *et al.*, 2019) sont une autre approche populaire pour l'appariement de sous-structures. Ces réseaux introduisent un mécanisme d'attention croisée pour comparer directement les sous-structures correspondantes entre deux graphes.

Le processus inclut :

— **Mise en correspondance locale :** Chaque nœud d'un graphe est comparé aux nœuds du graphe opposé en utilisant des mécanismes d'attention pondérée.

— **Agrégation des correspondances :** Les scores locaux obtenus pour chaque paire de nœuds sont combinés pour produire une estimation globale de similarité.

— **Propagation interactive :** Les représentations des nœuds sont mises à jour de manière interactive, en intégrant les informations croisées provenant de l'autre graphe.

Cette méthode est particulièrement puissante pour capturer des relations complexes et s'applique à divers contextes, notamment la reconnaissance de motifs dans les graphes.

3.3.4 Avantages de l'appariement de sous-structures

Les méthodes basées sur l'appariement de sous-structures présentent plusieurs avantages importants :

— **Réduction de la complexité :** En décomposant les graphes en sous-structures, ces approches permettent d'éviter des comparaisons coûteuses au niveau global.

— **Flexibilité :** Elles peuvent être adaptées à différents types de graphes, y compris ceux avec des attributs hétérogènes ou des relations complexes.

— **Richesse des représentations :** En utilisant des hypergraphes ou des mécanismes d'attention croisée, ces méthodes capturent des informations locales et globales de manière simultanée.

3.3.5 Applications pratiques

Les méthodes d'appariement de sous-structures trouvent des applications dans de nombreux domaines :

— **Bioinformatique :** Identification de motifs dans les réseaux biologiques ou les structures moléculaires.

— **Vision par ordinateur :** Analyse de graphes de points clés pour la reconnaissance d'objets ou la détection de similitudes dans des images.

— **Analyse de réseaux sociaux :** Comparaison de communautés ou d'interactions entre groupes dans des réseaux complexes.

— **Détection de fraudes :** Identification de transactions ou d'interactions similaires dans des bases de données représentées comme des graphes.

Ces approches permettent de traiter des graphes à grande échelle tout en maintenant une précision élevée dans l'évaluation des similarités.

3.4 Kernels de graphes profonds

Les kernels de graphes sont une classe de méthodes bien établies pour mesurer les similarités entre graphes. Ces méthodes définissent une fonction $k(G_1, G_2)$ qui quantifie la similarité entre deux graphes G_1 et G_2,

en capturant leurs propriétés structurelles et attributives. Les kernels de graphes profonds combinent les avantages des kernels traditionnels et des modèles d'apprentissage profond pour offrir une flexibilité accrue et une capacité d'adaptation aux graphes complexes et hétérogènes.

3.4.1 Concept de base des kernels de graphes

Un kernel de graphes est une fonction positive semi-définie qui compare les graphes en fonction de leurs sous-structures. Les sous-structures peuvent inclure :

- Les chemins (*walks*) ou marches aléatoires.
- Les sous-arbres extraits des graphes.
- Les graphes réduits par une partition ou un regroupement des nœuds.

Par exemple, un kernel de marches aléatoires (*random walk kernel*) calcule la similarité entre deux graphes en comptant les séquences similaires de nœuds et d'arêtes dans les deux graphes (Vishwanathan *et al.*, 2007).

3.4.2 Limites des kernels traditionnels

Bien que les kernels traditionnels soient puissants pour capturer les motifs structurels des graphes, ils souffrent de plusieurs limitations :

- Leur complexité computationnelle augmente rapidement avec la taille des graphes.
- Ils sont souvent incapables de tirer parti des attributs complexes des nœuds et des arêtes.
- Ils manquent de flexibilité pour s'adapter aux tâches spécifiques, comme la classification ou la détection de motifs.

Pour surmonter ces limitations, les kernels de graphes profonds ont été développés, combinant les avantages des kernels avec la capacité d'apprentissage des réseaux neuronaux profonds.

3.4.3 Kernels profonds

Les kernels de graphes profonds exploitent des modèles d'apprentissage profond, comme les réseaux de neurones de graphes (GNN), pour apprendre des représentations riches et adaptées des graphes avant de calculer les similarités. Le processus peut être résumé en trois étapes principales :

1. **Apprentissage de Représentations :** Les GNN ou d'autres architectures apprennent des embeddings pour les nœuds ou le graphe entier, en intégrant les attributs locaux et les relations structurelles

globales.

2. **Construction des Kernels :** Les représentations apprises sont utilisées pour définir un kernel. Par exemple, un produit scalaire ou une mesure de distance dans l'espace des embeddings peut servir de kernel.
3. **Optimisation :** Les paramètres du GNN et du kernel sont optimisés conjointement pour améliorer les performances sur des tâches spécifiques, comme la classification ou la régression.

3.4.4 Exemples

Voici quelques exemples représentatifs de kernels de graphes profonds :

— **Graph2Vec :** Inspiré des modèles Word2Vec, Graph2Vec apprend des représentations vectorielles globales pour les graphes en s'appuyant sur les caractéristiques locales des nœuds et de leurs voisinages (Narayanan *et al.*, 2017).
— **Kernels Basés sur les GNN :** Ces kernels utilisent directement les embeddings produits par les GNN comme entrée pour calculer la similarité entre graphes (Ma *et al.*, 2021).
— **Kernels Hiérarchiques :** Ces méthodes exploitent les relations hiérarchiques dans les graphes pour construire des représentations multi-échelles avant de définir un kernel (Ma *et al.*, 2021).

3.4.5 Applications

Les kernels de graphes profonds sont particulièrement adaptés pour des tâches où la structure et les attributs des graphes jouent un rôle critique :

— **Bioinformatique :** Analyse de similarité entre molécules ou réseaux biologiques.
— **Vision par ordinateur :** Comparaison de graphes représentant des images ou des scènes complexes.
— **Réseaux sociaux :** Identification de communautés similaires ou d'entités ayant des comportements similaires.
— **Détection de fraudes :** Analyse de réseaux transactionnels pour identifier des schémas frauduleux.

Les kernels de graphes profonds représentent une évolution naturelle des kernels traditionnels, combinant leur robustesse avec la puissance de l'apprentissage profond. Ces méthodes permettent de capturer des relations complexes et adaptatives entre graphes, ouvrant la voie à des applications dans des domaines variés et exigeants.

3.5 Similarité structurelle

La similarité structurelle est une méthode d'évaluation des similitudes entre entités dans des graphes ou réseaux, en se basant sur leurs contextes structurels plutôt que sur leurs attributs. Cette approche est particulièrement utile dans des domaines comme l'analyse des réseaux sociaux, les moteurs de recherche ou encore la bioinformatique, où la structure des relations entre les nœuds peut révéler des similarités significatives.

Parmi les méthodes clés utilisées pour mesurer la similarité structurelle, deux approches majeures ont émergé : **SimRank**, introduit par Jeh et Widom (Jeh et Widom, 2002a), et **P-Rank**, proposé par Zhou et al. (Zhao *et al.*, 2009). Ces deux méthodes diffèrent dans leur manière d'exploiter les relations structurelles entre nœuds, tout en partageant l'objectif commun de capturer des similarités globales en s'appuyant sur des chemins dans le graphe.

3.5.1 SimRank : similarité basée sur les voisinnages

SimRank repose sur une intuition simple mais puissante : deux nœuds sont similaires si leurs voisins sont similaires. Ce principe est exprimé mathématiquement par une équation récursive, où la similarité entre deux nœuds a et b est définie comme la moyenne pondérée des similarités entre leurs voisins respectifs (Jeh et Widom, 2002a).

3.5.1.1 Fondements mathématiques de SimRank

Soit un graphe dirigé $G = (V, E)$, où V représente l'ensemble des nœuds et E l'ensemble des arêtes. La similarité SimRank $S(a, b)$ entre deux nœuds a et b est définie comme suit :

$$S(a,b) = \begin{cases} 1 & \text{si } a = b, \\ C \cdot \frac{\sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} S(I(a)_i, I(b)_j)}{|I(a)| \cdot |I(b)|} & \text{si } a \neq b, \end{cases} \tag{3.3}$$

où :

- $I(a)$ est l'ensemble des prédécesseurs du nœud a,
- C est un facteur de pondération dans l'intervalle $[0, 1]$ qui contrôle la contribution des chemins plus longs,
- $|I(a)|$ et $|I(b)|$ sont les nombres de prédécesseurs des nœuds a et b.

La récursivité de cette équation implique que le calcul de $S(a, b)$ dépend des similarités des voisins, rendant nécessaire un processus d'itérations successives pour converger vers une solution stable (Jeh et Widom, 2002a).

3.5.1.1.1 Forces et limites

SimRank est particulièrement adapté aux graphes où la similarité structurelle est un facteur clé, comme les réseaux de citations, les systèmes de recommandation ou les bases de données relationnelles (Jeh et Widom, 2002a). Cependant, son principal inconvénient est son coût computationnel élevé, surtout pour des graphes de grande taille, en raison du besoin de calculer les similarités pour toutes les paires de voisins.

3.5.2 P-Rank : une extension de SimRank vers les graphes orientés

P-Rank (*Penetrating Rank*) est une généralisation de SimRank qui considère non seulement les relations entrantes, mais aussi les relations sortantes dans le graphe. Cette méthode vise à résoudre certaines limitations de SimRank en introduisant une mesure de similarité plus complète, adaptée aux réseaux d'information complexes où les flux directionnels jouent un rôle important (Zhao *et al.*, 2009).

3.5.2.1 Fondements mathématiques de P-Rank

P-Rank combine les contributions des relations entrantes et sortantes pour calculer la similarité structurelle. La similarité entre deux nœuds a et b est définie comme :

$$P(a, b) = \alpha \cdot S_{\text{in}}(a, b) + (1 - \alpha) \cdot S_{\text{out}}(a, b), \tag{3.4}$$

où :

- $S_{\text{in}}(a, b)$ est la similarité basée sur les relations entrantes, calculée de manière similaire à SimRank,
- $S_{\text{out}}(a, b)$ est la similarité basée sur les relations sortantes,
- $\alpha \in [0, 1]$ est un paramètre d'équilibrage entre les deux contributions.

Les calculs des similarités S_{in} et S_{out} suivent des formules analogues à SimRank, mais appliquées respectivement aux prédécesseurs et successeurs des nœuds (Zhao *et al.*, 2009).

3.5.2.2 Applications

P-Rank offre une vue plus complète de la similarité structurelle en tenant compte des flux directionnels, ce qui le rend particulièrement utile dans des réseaux tels que :

- **Réseaux sociaux :** Analyse des interactions bidirectionnelles entre utilisateurs.
- **Réseaux de transport :** Évaluation des similarités entre points de transit basées sur les flux entrants et sortants.
- **Systèmes de recommandation :** Identification des similarités dans des réseaux utilisateurs-produits (Zhao *et al.*, 2009).

3.5.3 Comparaison entre SimRank et P-Rank

SimRank et P-Rank partagent une base commune dans leur approche récursive, mais présentent des différences significatives :

- **Portée des relations :** SimRank se limite aux relations entrantes, tandis que P-Rank considère à la fois les relations entrantes et sortantes (Jeh et Widom, 2002a; Zhao *et al.*, 2009).
- **Flexibilité :** P-Rank offre un contrôle plus fin grâce au paramètre α, permettant d'ajuster l'importance relative des deux types de relations (Zhao *et al.*, 2009).
- **Complexité :** Bien que P-Rank soit plus coûteux en termes de calcul, il capture des informations structurelles plus riches, ce qui peut être crucial dans certains contextes.

Le calcul de similarité entre graphes est un domaine en pleine expansion, soutenu par des avancées en apprentissage profond. Les modèles d'embeddings, les GNN et les méthodes d'appariement offrent des solutions innovantes, mais des efforts supplémentaires sont nécessaires pour améliorer leur efficacité et leur applicabilité. Une combinaison hybride de ces approches pourrait ouvrir la voie à des solutions encore plus robustes et polyvalentes. Par ailleurs, SimRank et P-Rank sont deux mesures complémentaires pour évaluer la similarité structurelle dans des graphes. Tandis que SimRank repose sur une approche simple et élégante axée sur les voisins, P-Rank étend cette idée pour inclure des interactions directionnelles plus complexes. Ensemble, ces méthodes offrent des outils puissants pour analyser les relations structurelles dans divers types de réseaux, ouvrant la voie à des applications innovantes dans les systèmes de recommandation, les analyses de réseaux sociaux et bien plus encore (Jeh et Widom, 2002a; Zhao *et al.*, 2009).

3.6 Définition du problème

Calcul de similarité sur les données de production laitière

L'analyse de similarité entre graphes constitue un domaine de recherche actif et diversifié, soutenu par le développement des graphes de connaissances. Ces derniers permettent de représenter des entités et leurs relations de manière structurée, offrant une base solide pour modéliser et comprendre des systèmes complexes. Dans le contexte des données de production laitière, la comparaison d'entités telles que les vaches, les périodes de lactation et les jours de test, tout en prenant en compte leurs attributs spécifiques, soulève des défis liés à la nature hiérarchique et multidimensionnelle de ces données.

Les approches existantes pour mesurer la similarité entre graphes, telles que les méthodes d'embedding, les réseaux de neurones de graphes (GNN) ou les modèles de correspondance structurelle, apportent des solutions efficaces mais présentent certaines limites. Les méthodes d'embedding transforment les graphes en représentations vectorielles dans un espace réduit, ce qui peut entraîner une perte de précision lors de la réduction dimensionnelle, notamment dans les contextes où les attributs sont complexes ou fortement variés. Les réseaux de neurones de graphes, en s'appuyant sur des architectures avancées, offrent une grande précision dans la modélisation des interactions entre nœuds, mais cela se fait souvent au prix d'une complexité computationnelle élevée, limitant leur applicabilité dans des environnements où les ressources sont contraintes. Les modèles de correspondance structurelle, quant à eux, permettent une comparaison fine en alignant dynamiquement les éléments des graphes, mais peinent à gérer efficacement les relations hiérarchiques et multi-niveaux.

Dans le cadre des données laitières, ces limites se manifestent par une difficulté à capturer les relations complexes entre les différents niveaux d'organisation : vaches, périodes de lactation et jours de test par exemple. La plupart des approches actuelles ne permettent pas d'intégrer simultanément des attributs spécifiques à chaque niveau et les relations structurelles globales, limitant leur capacité à détecter des similitudes riches et détaillées nécessaires pour des applications telles que la sélection animale ou l'amélioration des performances de production.

Ainsi, il est nécessaire de développer une méthodologie qui :

1. Prend en compte la structure hiérarchique des données laitières, organisée en plusieurs niveaux d'abstraction.

2. Intègre à la fois les similarités structurelles et attributives, permettant une comparaison précise des entités et de leurs relations.
3. Réduit la complexité computationnelle pour garantir une applicabilité dans des contextes opérationnels.
4. Reste flexible et généralisable pour d'autres domaines nécessitant l'analyse de graphes complexes.

Cette problématique, au carrefour de l'analyse de graphes et des sciences animales, appelle des solutions innovantes capables de capturer les interactions hiérarchiques et multi-dimensionnelles propres aux données de production laitière.

CHAPITRE 4
MÉTHODOLOGIE

Ce chapitre est divisé en trois blocs principaux. Nous commençons par présenter les données et les étapes de pré-traitement associées à celles-ci. Ensuite, nous exposons l'approche proposée pour faire ces calculs de similarité. Enfin, nous décrivons comment utiliser cette méthode en l'appliquant sur un exemple et nous exposerons notre implémentation. Pour atteindre nos objectifs, nous avons utilisé plusieurs outils et bibliothèques, présentés ci-dessous :

- **Langage de programmation :** Python a été employé pour toutes les étapes de traitement et d'analyse des données.
- **Bibliothèques principales :**
  - Pandas : Pour la manipulation et l'analyse des données.
  - Matplotlib : Pour la création de visualisations graphiques.
  - Seaborn : Pour des visualisations statistiques attrayantes et détaillées.
  - NumPy : Pour les calculs numériques avancés.
  - Scikit-learn : Pour les modèles d'apprentissage automatique, notamment le clustering avec KMeans.
- **Environnement de développement :** Jupyter Notebook a été utilisé pour faciliter le développement, l'expérimentation et la documentation du projet.

4.1 Description des données

Notre étude consiste à calculer la similarité sur les données de production laitière. Un vaste ensemble de données sur la production laitière a été fourni par Lactanet, le Réseau canadien pour l'excellence laitière. Ces données sont généralement hétérogènes, c'est-à-dire qu'elles couvrent partiellement ou totalement la santé, la nutrition, le rendement et la génétique. Elles possèdent également une structure complexe, avec une grande variété de données pour un animal unique, dispersées dans de nombreux enregistrements et de multiples tables. Lactanet est le centre d'expertise en production laitière couvrant la province de Québec et les régions atlantiques du Canada. Les données accumulées par Lactanet sur la production laitière et le contrôle laitier décrivent 6 670 troupeaux et 1,5 million de vaches.

Les concepts clés reflétés dans les données comprennent les échantillonnages du contrôle laitier et les analyses de laboratoire associées qui estiment les principaux composants du lait : matières grasses, protéines,

azote uréique du lait, etc. Dans l'ensemble, les enregistrements fournis par Lactanet représentent plus de 3 milliards de données. Cet énorme ensemble de données cache des concepts potentiellement significatifs, par exemple les vaches improductives qui admettent une amélioration par rapport à celles qui doivent être vendues rapidement, ainsi que les modèles de comportement des vaches ou des agriculteurs, qui doivent être découverts.

Notre dataset est structuré en trois sous-ensembles, chacun correspondant à un niveau clé de données relatives aux animaux, leurs lactations et les tests réalisés sur leur production laitière. Dans notre cas nous avons extrait un dataset de 1000 vaches de la base de données initiales et pour l'ensemble des 1000 vaches nous comptabilisons 2081 lactations et 15062 tests, Voici une description détaillée des attributs incluses dans chaque sous-ensemble :

4.1.1 Jeu de données Animaux

Ce sous-ensemble regroupe des informations essentielles sur chaque animal dans l'étude. Les attributs sélectionnés permettent d'identifier, de caractériser et de suivre les mouvements des animaux dans les troupeaux. Voici les détails des attributs :

- **Identifiant de troupeau** : Un identifiant unique pour relier les animaux à leur troupeau.
- **Identifiant de l'animal** : Un identifiant individuel pour chaque animal.
- **Race de l'animal** : L'indication de la race à laquelle l'animal appartient.
- **Date de naissance** : La date de naissance de chaque animal.
- **Date d'entrée dans le troupeau** : La date à laquelle l'animal a été intégré au troupeau.
- **Date de sortie du troupeau** : La date où l'animal a quitté le troupeau.
- **Raison de sortie du troupeau 1 & 2** : Les raisons principales et secondaires expliquant la sortie de l'animal (par exemple, vente, décès, réforme, etc.).

4.1.2 Jeu de données Lactations

Ce sous-ensemble décrit les informations relatives aux différentes périodes de lactation des animaux, en incluant des caractéristiques liées au vêlage et à l'élevage des veaux. Les manipulations et filtres appliqués garantissent la qualité et la cohérence des données :

- **id_vache** : Identifiant unique de la vache.
- **lactation_num** : Numéro de la lactation.

— **end_date** : Date de fin de la lactation.

— **start_date** : Date de début de la lactation.

— **lact_start_reasn** : Raison du début de la lactation.

— **ler_cd** : Code LER (Local Event Record).

— **day_305_milk** : Production de lait standardisée sur 305 jours (kg).

— **lact_date_yld_milk** : Production totale de lait pour la période de lactation (kg).

— **bca_milk** : Valeur BCA (Breed Class Average) pour la production de lait.

— **lact_prsstncy_milk** : Persistante de la lactation pour la production de lait.

— **day_305_fat** : Production de matières grasses standardisée sur 305 jours (kg).

— **lact_date_yld_fat** : Production totale de matières grasses pour la période de lactation (kg).

— **bca_fat** : Valeur BCA pour la production de matières grasses.

— **lact_prsstcy_fat** : Persistante de la lactation pour les matières grasses.

— **day_305_prot** : Production de protéines standardisée sur 305 jours (kg).

— **lact_date_yld_prot** : Production totale de protéines pour la période de lactation (kg).

— **bca_prot** : Valeur BCA pour la production de protéines.

— **lact_prsstncy_prot** : Persistante de la lactation pour les protéines.

— **scc_linear_score_avg** : Score linéaire moyen des cellules somatiques.

— **scc_milk_wgt_tot** : Poids total du lait associé au comptage des cellules somatiques.

— **clvng_ease_1** : Facilité de vêlage pour le premier vêlage.

— **calf_sex_1** : Sexe du veau pour le premier vêlage.

— **calf_size_1** : Taille du veau pour le premier vêlage.

— **survival_ind_1** : Indicateur de survie du veau pour le premier vêlage.

— **clvng_ease_2** : Facilité de vêlage pour le deuxième vêlage.

— **calf_sex_2** : Sexe du veau pour le deuxième vêlage.

— **calf_size_2** : Taille du veau pour le deuxième vêlage.

— **survival_ind_2** : Indicateur de survie du veau pour le deuxième vêlage.

— **cumul_milk_value** : Valeur cumulative de la production de lait.

— **cumul_feed_cost** : Coût cumulatif de l'alimentation.

— **peak_dim** : Jours au pic de lactation.

— **peak_milk_yld** : Production maximale de lait (kg).

— **lid** : Identifiant unique de lactation.

4.1.3 Jeu de données Tests Animaux

Ce sous-ensemble contient les mesures réalisées sur la production laitière et les performances des animaux lors des tests. Les données sélectionnées permettent d'évaluer la production et la qualité du lait, ainsi que certains indicateurs économiques :

- **Date du test** : Le jour où le test a été effectué.
- **Jours en lait** : Le nombre de jours depuis le début de la lactation au moment du test.
- **Production laitière sur 24h** : La quantité de lait produite en une journée.
- **Pourcentage de matières grasses et de protéines** : Indicateurs de la qualité du lait.
- **Comptage des cellules somatiques** : Un indicateur de la santé mammaire.
- **Azote uréique dans le lait (MUN)** : Un indicateur de l'efficacité alimentaire.
- **Lactose et BHB (*Beta-Hydroxybutyrate*)** : Paramètres supplémentaires pour évaluer la composition du lait et l'état métabolique de l'animal.
- **Valeur quotidienne du lait et coût quotidien de l'alimentation** : Des indicateurs économiques pour évaluer les marges de production.

4.1.4 Analyse statistique des données

L'analyse des valeurs manquantes dans les fichiers qui contiennent nos données révèle les points suivants :

- **Fichier** `vache_data` :
  - Une seule colonne (`lhr_cd_2`) contient des valeurs manquantes (960 valeurs manquantes).
  - Les données relatives aux dates (`birth_date`, `enter_herd_date`, et `left_herd_date`) sont complètes.

Tableau 4.1 – Analyse des données manquantes pour les données sur les vaches

Colonne	valeurs manquantes
anm_id	0
enter_herd_date	0
birth_date	0
left_herd_date	0
lhr_cd	0

Suite à la page suivante

Tableau 4.1 – Analyse des données manquantes pour les données sur les vaches

Colonne	valeurs manquantes
lhr_cd_2	960
anb_cd	0

— **Fichier** `lac_data` :

— Plusieurs colonnes, comme `day_305_milk` (620 valeurs manquantes), présentent des données incomplètes.

— Les colonnes relatives aux pics de production (`peak_milk_yld`) et aux valeurs cumulées (`cumul_milk_value`) ont également des valeurs manquantes.

Tableau 4.2 – Analyse des données manquantes pour les données sur les lactations

Colonne	valeurs manquantes
anm_id	0
lact_no	0
end_date	0
start_date	0
lact_start_reasn	0
ler_cd	0
day_305_milk	620
lact_date_yld_milk	12
bca_milk	621
lact_prsstncy_milk	12
day_305_fat	622
lact_date_yld_fat	14
bca_fat	623
lact_prsstcy_fat	14
day_305_prot	622
lact_date_yld_prot	14

Suite à la page suivante

Tableau 4.2 – Analyse des données manquantes pour les données sur les lactations

Colonne	valeurs manquantes
bca_prot	623
lact_prsstncy_prot	14
scc_linear_score_avg	5
scc_milk_wgt_tot	5
clvng_ease_1	7
calf_sex_1	93
calf_size_1	181
survival_ind_1	7
clvng_ease_2	2035
calf_sex_2	2035
calf_size_2	2042
survival_ind_2	2035
cumul_milk_value	4
cumul_feed_cost	1594
peak_dim	12
peak_milk_yld	12
lid	0

— **Fichier** `tes_data` :
  — Les colonnes liées à la composition du lait (`protein`, `fat`) sont complètes.
  — Certaines colonnes, comme `scc` (`1` valeur manquante) et `mun` (`6443` valeurs manquantes), présentent des données incomplètes.
  — Les colonnes économiques, comme `daily_feed_cost` et `profit`, contiennent un nombre significatif de valeurs manquantes.

Tableau 4.3 – Analyse des données manquantes pour les données sur les jours de tests

Colonne	valeurs manquantes
anm_id	0
lact_no	0
test_date	0
ans_cd	0
dim	0
hr_24_milk	1
fat	0
protein	0
scc	1
mun	6443
lactose	0
abnrml_status	1
lact_start_date	0
milkng_fqcy	91
daily_milk_value	83
daily_feed_cost	10872
bhb	0
profit	10895
lid	0

Pour traiter les valeurs manquantes présentes dans nos données. Nous avons défini une fonction qui fait deux tâches principales : la suppression des colonnes ayant un pourcentage élevé de valeurs manquantes et l'imputation des valeurs manquantes restantes en fonction du type de données.

- **Suppression des colonnes avec trop de valeurs manquantes** : Si une colonne contient un pourcentage de valeurs manquantes supérieur à un seuil prédéfini, elle est supprimée. Ce seuil est configurable (par défaut, 50%).
- **Imputation des valeurs manquantes restantes** :
  - Les colonnes numériques voient leurs valeurs manquantes remplacées par la médiane des valeurs non manquantes.

— Les colonnes catégoriques voient leurs valeurs manquantes remplacées par la valeur la plus fréquente (mode) dans chaque colonne.

4.1.5 Statistiques descriptives des données numériques

Les statistiques descriptives suivantes ont été calculées pour les colonnes numériques :

— **Fichier** `vache_data` :
  — Les identifiants des animaux (`anm_id`) varient de $10,245,383$ à $10,471,127$.
— **Fichier** `lac_data` :
  — La production laitière moyenne sur 305 jours (`day_305_milk`) est d'environ $9,500\,\mathrm{kg}$, avec une variabilité significative.
  — Le pic de production moyenne (`peak_milk_yld`) est d'environ $35\,\mathrm{kg/j}$.
— **Fichier** `tes_data` :
  — La production moyenne journalière (`hr_24_milk`) est de $30,86\,\mathrm{kg}$.
  — La teneur moyenne en protéines (`protein`) est de $3,39\%$, avec une distribution homogène.

4.1.6 Évaluation de la distribution des données

L'analyse des distributions des données a été effectuée pour examiner les variables numériques de chaque ensemble de données et identifier des tendances ou anomalies éventuelles. Cette étape est essentielle pour garantir la pertinence des calculs de similarité et la puissance des résultats. Voici les observations pour chaque ensemble de données.

4.1.6.1 Méthodologie

Les distributions des variables numériques ont été visualisées à l'aide de courbes de densité (*kernel density estimation*, KDE). Cette approche permet d'explorer la répartition des valeurs tout en identifiant les éventuelles asymétries, pics ou valeurs aberrantes. Les données non numériques ou celles présentant des incompatibilités ont été exclues de l'analyse graphique.

4.1.6.2 Observations générales

- Pour les ensembles de données analysés (*vache_data*, *lac_data*, et *tes_data*), les distributions des variables numériques ont révélé une variabilité importante dans certaines mesures clés.
- Certaines variables, comme les identifiants (`anm_id`) ou les classifications, n'ont pas pu être visualisées en raison de formats non numériques ou de valeurs incohérentes. Ces colonnes serviront d'identifiant et ne seront pas intégrées dans les calculs de similarité.
- Les courbes de densité montrent que certaines variables, notamment celles relatives à la production laitière ou aux scores cellulaires (SCC), présentent des distributions asymétriques ou des valeurs extrêmes.

4.1.6.3 Problèmes rencontrés

- Plusieurs colonnes contiennent des valeurs non numériques ou des formats incompatibles, empêchant leur inclusion dans l'analyse graphique. Ces problèmes sont principalement dus à des erreurs de typage ou à des valeurs manquantes.
- Les variables avec des distributions fortement asymétriques ou une présence significative de valeurs aberrantes peuvent biaiser les analyses subséquentes si elles ne sont pas traitées correctement.

Cette analyse fournit une base solide pour le prétraitement des données et leur intégration dans un modèle de similarité adapté, garantissant une évaluation robuste des relations entre les entités étudiées.

4.1.7 Visualisations des distributions et relations clés

Nous avons produit quelques statistiques sur les variables numériques. Une partie d'entre elles sont visualisées par les Figures 4.1, 4.2 et 4.3.

4.1.8 Conclusions

L'analyse descriptive met en évidence la variabilité significative dans les performances des vaches (production, composition du lait). Les valeurs manquantes doivent être traitées avant d'entreprendre des analyses de similarité. La relation entre le SCC et la production suggère des possibilités d'exploration supplémentaires pour évaluer l'impact des indicateurs de santé sur la performance.

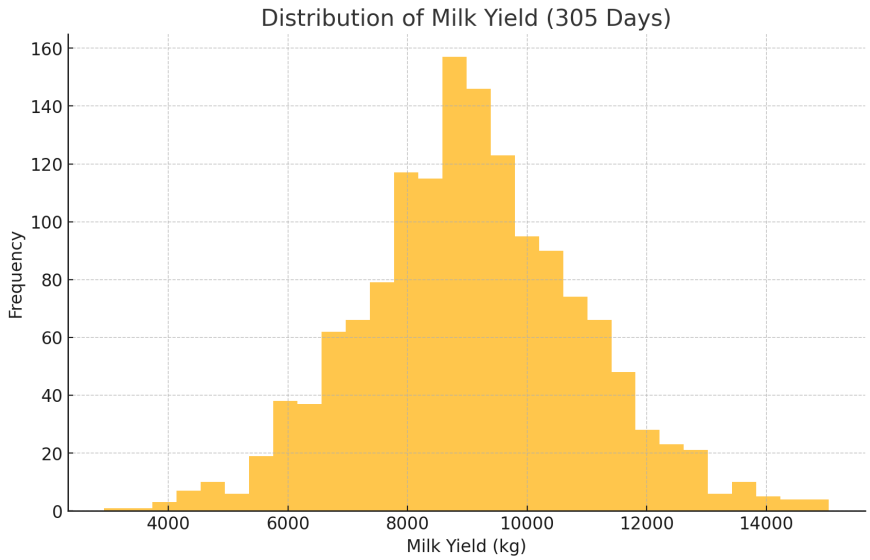

Figure 4.1 – Distribution de la production laitière sur 305 jours (`day_305_milk`).

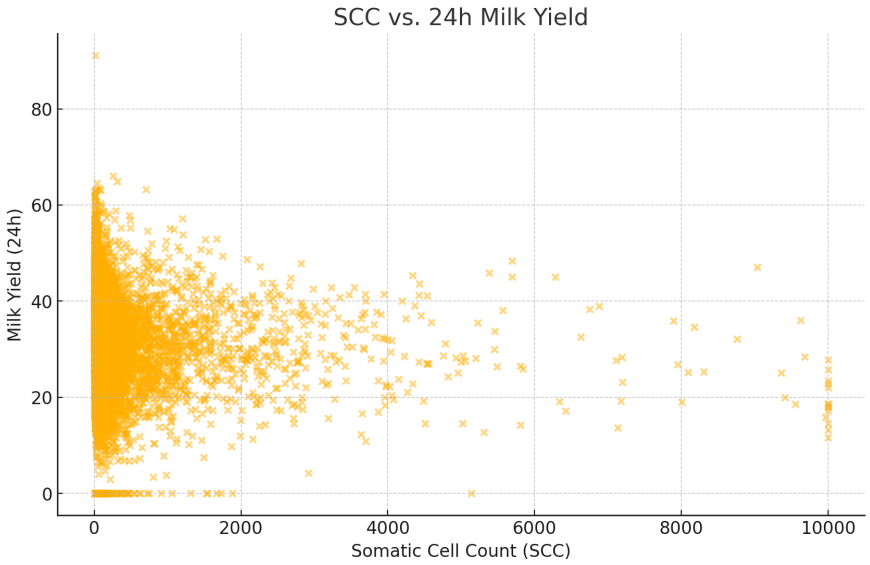

Figure 4.2 – Relation entre le SCC (`scc`) et la production journalière (`hr_24_milk`).

4.2 Principes de conception

4.2.1 Intuition

Afin de valider notre mesure de similarité elle sera utilisée pour évaluer la ressemblance entre les vaches en tenant compte d'un ou plusieurs indices de performance. Ceux-ci seront notre base d'évaluation : un tel choix est dicté par l'objectif global du projet laitier, soit d'améliorer ces performances. Ainsi, la bonne cor-

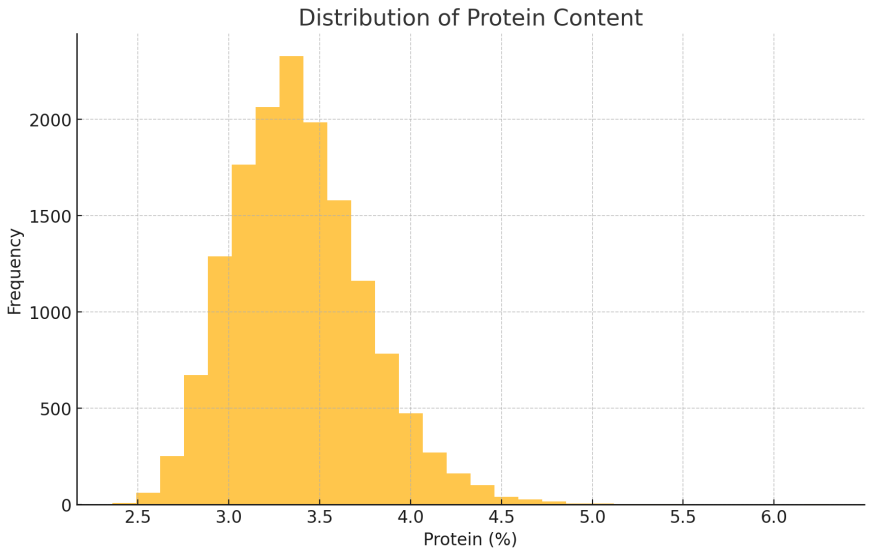

Figure 4.3 – Distribution de la teneur en protéines (`protein`).

respondance entre l'indice et la similarité indiquera, de manière indirecte, la pertinence de notre mesure.

Plus spécifiquement, nous allons comparer les deux quantité de manière indirecte. En d'autres mots, plutôt que de chercher à calculer une mesure statistique entre les deux, nous allons comparer leur effet sur l'ordonnancement des vaches par rapport à une vache donnée. L'intuition derrière est que la similarité sera utilisée, en grande partie, dans le choix des vaches le plus semblables à une vache donnée (par exemple, pour décider si une vache peut être remplacée suite à une maladie ou pour décider d'isoler une groupe homogène de vaches au sein du troupeau entier). Cela revient à établir un ordre entre vaches basé sur cette similarité. Du coup, bien que cela peut entraîner une certaine perte de précision, nous proposons, à des fin de comparaison, de ce focaliser non pas sur les valeurs exactes des deux mesures, mais plutôt sur l'ordre induit sur un ensemble de vaches.

4.2.2 Approche

Il existe une multitude de mesures de corrélation entre deux ordonnancement sur le même ensemble d'objets de données.

4.2.2.1 Nature des graphes :

Les graphes utilisés dans cette conception sont organisés de manière hiérarchique par niveau. Chaque niveau correspond à un type de nœud, et ces nœuds sont associés à des attributs spécifiques qui doivent être évalués. En raison de cette hiérarchie :

- Les nœuds de chaque niveau sont typés, c'est-à-dire que seuls des nœuds du même type peuvent être comparés entre eux.
- Pour cette conception, les graphes sont considérés non orientés, ce qui signifie que les relations entre les nœuds (les arêtes) ne portent pas de direction particulière.

4.2.2.2 SimRank appliqué sur des graphes multiples

Dans l'article original de SimRank (Jeh et Widom, 2002a), la méthode est définie pour un seul graphe. L'objectif de SimRank est de calculer la similarité entre des nœuds au sein d'un même graphe en se basant sur l'idée que "deux objets sont similaires si leurs voisins sont similaires". La similarité est donc propagée à travers le graphe en évaluant les relations entre les voisins de chaque nœud.

L'application de SimRank à deux graphes distincts, telle que nous la proposons ici, n'est cependant pas directement couverte dans l'article initial. Cela représente une extension non triviale de la méthode, nécessitant des adaptations spécifiques.

Dans cette adaptation, SimRank est appliqué sur deux graphes distincts les comparaisons ne se font donc qu'entre des nœuds de même type dans les deux graphes.

4.2.2.3 Restriction de la comparaison

SimRank ne peut comparer que des nœuds du même niveau structurel (par exemple, des paires *"Vache vs Vache"*, *"Lactation vs Lactation"*). Il est impossible de comparer des nœuds de types différents (comme une Vache avec un Test). Par exemple :

- Lors de la comparaison entre deux périodes de lactation, on forme des couples de tests respectifs, ainsi qu'un couple de vaches respectives pour représenter les comparaisons.

4.3 Particularités méthodologiques

4.3.1 Fonction de similarité générique

La fonction de similarité générique repose sur les équations de SimRank et est notée $\mathrm{Sim}(v_1, v_2)$. Elle s'applique à tous les couples de nœuds homogènes entre les deux graphes, quel que soit leur type.

La fonction $\mathrm{Sim}(v_1, v_2)$ est calculée de manière itérative, c'est-à-dire, à l'aide d'une suite $\{\mathrm{Sim}^i(v_1, v_2)\}$ dont les membres sont déterminés selon le processus suivant :

- **Étape initiale (itération 0)** : On commence avec une valeur initiale que peut être calculée indépendement de la structure du graphe.
- **Itérations suivantes** : À chaque itération $i+1$, la nouvelle valeur de la similarité est calculée comme une somme pondérée des similarités de tous les couples de voisins homogènes des nœuds v_1 et v_2 à l'itération précédente i.

4.3.2 Valeur d'initialisation

La valeur d'initialisation, à l'étape 0, est donnée par la similarité d'attributs normalisée $\mathrm{sim}_a(v_1, v_2)$. Cette similarité est définie comme une somme pondérée des similarités individuelles sur l'ensemble des attributs communs aux nœuds v_1 et v_2.

4.4 Définitions mathématiques

Soit (v_1, v_2) un couple de nœuds homogène (c'est-à-dire composé de nœuds de même type), et soit $\mathrm{sim}_a(v_1, v_2)$ leur similarité d'attributs normalisée. La similarité SimRank adaptée, notée $\mathrm{Sim}(v_1, v_2)$, est définie comme suit :

$$\mathrm{Sim}(v_1, v_2) = c \cdot \frac{1}{|N_h(v_1, v_2)|} \sum_{(v_1', v_2') \in N_h(v_1, v_2)} \mathrm{Sim}(v_1', v_2') \tag{4.1}$$

Dans une interprétation non orienté des liens du graphe, la défition ci-dessus mène à des dépendances circulaires entre les fonctions mathématiques. Ainsi, la similarité entre deux vaches dépend de la similarité des couples de lactations, mais l'inverse est aussi vrai. En conséquence, les définitions respectives ne peuvent

pas être exploitées par un calcul directe comme pour les mesures classiques dans les données tabulaires. À la place, les définitons respectives sont interprétées comme des équations qu'il faudrait résoudre afin d'obtenir les valeurs de similarité en tant que solutions du système d'équations résultant.

Techniquement, une résoution approximative et itérative est appliquée : les valeurs sont obtenues comme les limites la limite d'une suite définie comme suit :

4.4.1 Initialisation

$$Sim^0(v_1, v_2) = \sigma(v_1, v_2) \tag{4.2}$$

où $\sigma(v_1, v_2)$ est une fonction statique de similarité, c'est-à-dire une constante ou toute autre fonction qui peut être calculée *a priori* (sans égard pour les voisinages des noeuds en argument). Divers choix sont possibles, nous avons opté pour une substitution par la fonction $\mathrm{sim_a}(v_1, v_2)$ qui représente la similarité d'attributs entre v_1 et v_2.

4.4.2 Itération

$$\mathrm{Sim}^{i+1}(v_1, v_2) = c \cdot \frac{1}{|N_h(v_1, v_2)|} \sum_{(v_1', v_2') \in N_h(v_1, v_2)} \mathrm{Sim}^i(v_1', v_2') \tag{4.3}$$

où :

- c est un facteur d'atténuation (généralement compris entre 0 et 1) qui contrôle l'influence des itérations précédentes.
- $N_h(v_1, v_2)$ est l'ensemble des couples de voisins homogènes de v_1 et v_2, c'est-à-dire les couples (v_1', v_2') tels que :
  - $v_1' \in N(v_1)$ (voisins de v_1 dans le premier graphe),
  - $v_2' \in N(v_2)$ (voisins de v_2 dans le second graphe),
  - v_1' et v_2' sont du même type (Vache, Lactation, etc.).

4.4.3 Valeurs finales

Pour tout couple de nœuds homogène (v_1, v_2)

$$\mathrm{Sim}(v_1, v_2) = \mathrm{Sim}^{\infty}(v_1, v_2)$$

où $\mathrm{Sim}^{\infty}(v_1, v_2)$ est la limite de la suite $\{\mathrm{Sim}^{i}(v_1, v_2)\}$ (son membre de rang infini).

4.5 Explication des choix

- **Hiérarchie et typage des nœuds** : Ces contraintes permettent d'assurer que les comparaisons sont significatives en respectant la structure logique des graphes. Comparer des entités du même niveau (ex : des vaches avec des vaches) garantit que les similarités sont cohérentes.
- **Utilisation d'une somme pondérée** : Cette approche permet de prendre en compte l'ensemble des voisins tout en atténuant l'impact des contributions à chaque itération, afin d'éviter une propagation non contrôlée des similarités au fil des itérations.
- **Dépendance itérative** : La structure itérative est un principe central de SimRank, où les similarités des nœuds dépendent des similarités de leurs voisins, ce qui introduit une notion de proximité structurelle.

4.5.1 Validation

4.5.1.1 Principes de la conception

- **Particularité** : Dans notre cas SimRank s'applique sur 2 graphes et non sur un seul comme défini dans l'article initial (Jeh et Widom, 2002a)
  1. Les graphes sont organisés hiérarchiquement (par niveau), les noeuds sont typés et ont leur propres attributs évalués.
  2. Pour ce premier exercice de conception, ils sont considérés comme non orientés.
- **Restriction** : SimRank s'applique uniquement sur les couple de noeud de même niveau structurel (Vache vs Vache), donc pas sur de couples hétérogènes (pas de Vache vs Test).
  1. En particulier, en comparant deux lactations, on forme **tous les couples** de leurs Tests respectifs et **le couple** de leurs Vaches respectives.
- **Particularité** :
  1. On définit **une seule** fonction de similarité **générique** —celle tirée des équations de SimRank, $Sim(v_1, v_2)$— pour tous les couples de noeuds homogènes entre graphes (v_1, v_2) (quelque soit leur type).

2. Elle est calculée, pour un couple de noeuds de manière itérative :
   (a) En commençant par la valeur d'**initialisation** (étape 0)
   (b) En recalculant à chaque itération i+1 la nouvelle valeur de la similarité comme une somme pondérée des valeurs pour tous les couples de voisins à l'étape i.
3. pour ce 1er exercice, on utilise comme valeur d'initialisation celle calculée par la similarité d'attributs normalisée $sim_a(v_1, v_2)$,
4. elle aussi est définie de manière générique comme une somme pondérée des similarités individuelles sur l'ensemble des attributs communs.

4.5.1.2 Définitions mathématiques

Soit (v_1, v_2) un couple de noeuds homogène (même type) et soit $sim_a(v_1, v_2)$ sa similarité d'attributs normalisée,

La similarité SimRank **adaptée** $Sim(v_1, v_2)$ est définie comme la valeur limite (membre $Sim^{\infty}(v_1, v_2)$) de la suite définie itérativement comme suit :

$$Sim^0(v_1, v_2) = sim_a(v_1, v_2) \tag{4.4}$$

$$Sim^{i+1}(v_1, v_2) = \frac{c}{|N_h(v_1, v_2)|} \sum_{(v'_1, v'_2) \in N_h(v_1, v_2)} Sim^i(v'_1, v'_2) \tag{4.5}$$

où $N_h(v_1, v_2)$ est le voisinage homogène du couple (v_1, v_2) soit l'ensemble de tous les couples (v'_1, v'_2) tels que $v'_1 \in N(v_1)$, $v'_2 \in N(v_2)$ et v'_1, v'_2 ont le même type (Vache, Lactation, Test).

4.5.1.3 Principes de la conception (variante avec facteur constant)

— **Particularité 1** : Dans notre cas, le calcul de similarité (SimRank+) s'applique sur 2 graphes et non sur un seul comme défini dans l'article d'origine.

1. Les graphes sont organisés hiérarchiquement (par niveau), les noeuds sont typés et ont leur propres attributs que la similarité doit évaluer aussi.
2. Les deux graphes sont considérés comme non orientés.

— **Restriction** : SimRank+ s'applique uniquement sur les couples de noeud de même niveau structurel (Vache vs Vache, Lactation vs Lactation), donc pas sur de couples hétérogènes (par exemple, pas de Vache vs Test).

1. En particulier, en comparant deux lactations, on forme **tous les couples** de leurs Tests respectifs et **le couple** de leurs Vaches respectives.

— **Particularité 2** :

1. On définit **une seule** fonction de similarité **générique** —celle adaptée des équations de SimRank, $Sim(v_1, v_2)$— pour tous les couples de noeuds homogènes entre graphes (v_1, v_2) (quelque soit leur type).
2. Elle est calculée, pour un couple de noeuds de manière itérative :
   (a) En commençant par la valeur d'**initialisation** (étape 0)
   (b) En recalculant à chaque itération $i + 1$ la nouvelle valeur de la similarité comme une somme pondérée (α, $1 - \alpha$) des deux facteurs :
      — la similarité de l'étape 0 - (*ceci est une différence majeure de SimRank+ avec le modèle de SimRank qui n'a pas de facteur constant*)
      — la combinaison linéaire (somme pondérée) des valeurs de la similarité pour tous les couples de voisins à l'étape i.
3. On utilise comme valeur d'initialisation la similarité d'attributs normalisée $sim_a(v_1, v_2)$,
4. $sim_a(v_1, v_2)$ est aussi définie de manière générique, comme une **somme pondérée** des similarités individuelles sur l'ensemble des attributs communs.

4.5.1.4 Définitions mathématiques

Soit (v_1, v_2) un couple de noeuds homogène (même type) et soit $sim_a(v_1, v_2)$ sa similarité d'attributs normalisée,

La similarité SimRank **adaptée**, SimRank+, notée $Sim(v_1, v_2)$, est définie comme la valeur limite (membre $Sim^{\infty}(v_1, v_2)$) de la suite définie itérativement comme suit :

$$Sim^0(v_1, v_2) = sim_a(v_1, v_2) \tag{4.6}$$

$$Sim^{i+1}(v_1, v_2) = \alpha Sim^0(v_1, v_2) + \frac{1-\alpha}{|N_h(v_1, v_2)|} \sum_{(v_1', v_2') \in N_h(v_1, v_2)} Sim^i(v_1', v_2') \tag{4.7}$$

où $N_h(v_1, v_2)$ est le voisinage homogène du couple (v_1, v_2) soit l'ensemble de tous les couples (v_1', v_2') tels que $v_1' \in N(v_1)$, $v_2' \in N(v_2)$ et v_1', v_2' ont le même type (Vache, Lactation, Test).

4.6 Calcul de similarité, illustration

Pour le calcul de similarité, nous allons illustrer un exemple entre deux vaches. Veuillez trouver ci-dessous deux diagrammes représentant les détails pour chaque vache, suivis d'une description de la méthode utilisée pour calculer la similarité. Nous avons appliqué notre modèle SimRank adapté. Le calcul a été effectué manuellement pour la similarité des attributs, tandis que les itérations suivantes ont été calculées de manière itérative à l'aide d'un fichier Excel, afin de démontrer la convergence de nos résultats.

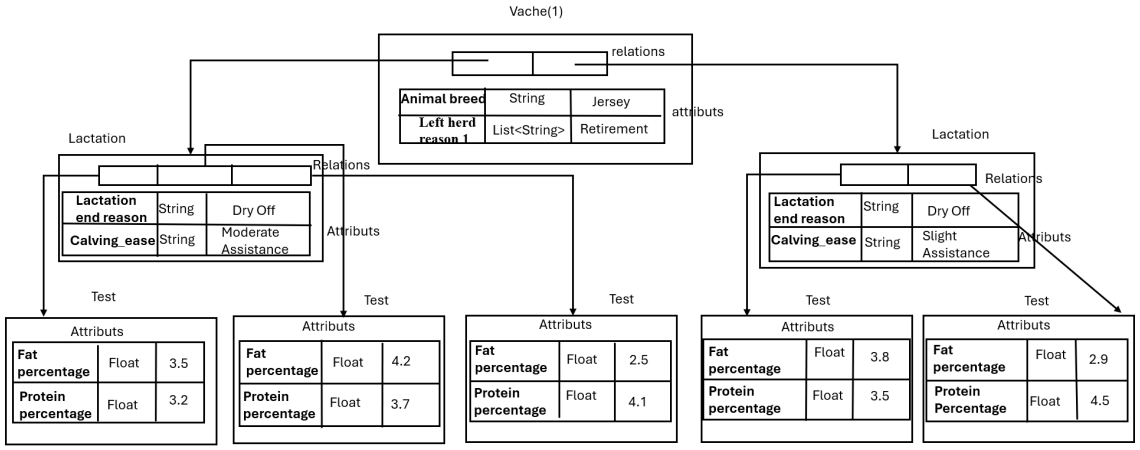

Figure 4.4 – Diagramme pour la vache 1

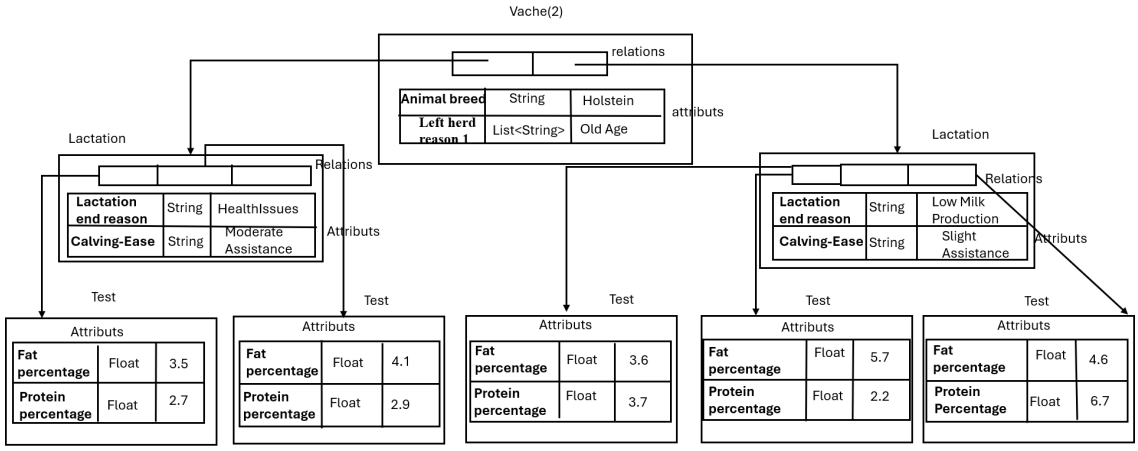

Figure 4.5 – Diagramme pour la vache 2

4.6.1 Description des attributs choisis pour chaque niveau

1. **Données sur les vaches** :
   - race animale (Animal breed)
   - Raison de sortie du troupeau (Left herd reason)
2. **Informations sur la lactation** :
   - Raison de fin de lactation (Lactation end reason)
   - facilité de vêlage (Calving ease)
3. **Détails des jours de test** :
   - Pourcentage de matière grasse (Fat percentage)
   - Pourcentage de protéine (Protein percentage)

4.6.2 Première formule de conception

La première conception repose sur la définition mathématique suivante :

Définitions mathématiques : Soit (v_1, v_2) un couple de nœuds homogènes (de même type) et soit $sima(v_1, v_2)$ leur similarité d'attributs normalisée.

La similarité SimRank adaptée, notée $Sim(v_1, v_2)$, est définie comme la valeur limite $Sim^{\infty}(v_1, v_2)$ de la suite suivante, définie itérativement :

1. $Sim^0(v_1, v_2) = sima(v_1, v_2)$
2. $Sim^{i+1}(v_1, v_2) = c \cdot \frac{1}{|N_h(v_1,v_2)|} \sum_{(v'_1,v'_2)\in N_h(v_1,v_2)} Sim^i(v'_1, v'_2)$

Explications des termes :

— $N_h(v_1, v_2)$ est le voisinage homogène du couple (v_1, v_2), c'est-à-dire l'ensemble de tous les couples (v'_1, v'_2) tels que $v'_1 \in N(v_1)$, $v'_2 \in N(v_2)$ et v'_1, v'_2 sont de même type (par exemple, Vache, Lactation, Test).

— c est un facteur d'atténuation (généralement $c \in [0, 1]$).

4.6.2.1 Observations

Lors de nos calculs avec la première conception, nous avons constaté que les valeurs ne convergeaient pas. Cela était dû au fait que certains tests appartenaient à la même lactation et Cette situation a engendré une tendance vers zéro pour nos valeurs de similarité des jours de tests mais les lactations et les vaches atteignent la valeur zéro. Cette observation met en évidence

4.6.2.2 Tendance vers zéro avec les itérations

La structure du graphe est hautement uniforme c'est à dire que beaucoup de nœuds ont des voisins similaires ou répétitifs, les valeurs de similarité $Simi(v'_1, v'_2)$ deviennent de plus en plus petites au fur et à mesure des itérations. Cela est dû à :

— Une dilution des valeurs causée par la moyenne $\frac{1}{|Nh(v_1,v_2)|}$.

— Une perte de "diversité" dans les connexions du graphe, où les similarités entre les nœuds voisins sont toutes proches ou égales.

— Le facteur c amplifie cette atténuation à chaque itération.

4.6.2.3 Problème d'uniformité (ou uniformise)

Plusieurs nœuds sont connectés systématiquement aux mêmes nœuds clés que nous allons (ici appelés "nœuds étoiles"), cela entraîne une redondance dans le calcul de $Nh(v_1, v_2)$. Cette redondance fait que la somme $\sum Simi(v'_1, v'_2)$ est dominée par les mêmes valeurs, conduisant à une disparition progressive des différences entre les similarités.

4.6.2.4 Solutions possibles pour interpréter et résoudre ce problème

4.6.2.4.1 Comprendre le phénomène

- Le fait que les valeurs tendent vers zéro n'indique pas nécessairement une absence de convergence. Cela peut être dû à la structure du graphe et à la manière dont les voisinages sont définis.
- Ce comportement peut signifier que les similarités entre les nœuds sont uniformisées, ce qui est un biais introduit par le graphe ou la formule.

4.6.2.4.2 Adapter la normalisation

- Introduire un terme qui empêche la normalisation excessive : par exemple, ajouter un biais constant ou utiliser une normalisation moins stricte.
- Par exemple, Introduire un poids inversement proportionnel au degré des nœuds voisins pour limiter leur influence excessive sur $Simi + 1(v_1, v_2)$.

4.6.3 Explication de la deuxième formule

La deuxième formule propose une amélioration de la prémière formule pour résoudre le problème d'atténuation des similarités à chaque itération en combinant deux composantes :

1. **Initialisation avec une similarité de base (Sim_0) :**
   - $Sim_0(v_1, v_2)$ est la similarité initiale entre v_1 et v_2, calculée à l'aide d'une fonction de similarité directe ($sim_a(v_1, v_2)$) basée sur les attributs ou une autre métrique de base.
   - Cette composante garantit que même si les itérations futures réduisent les valeurs, la similarité de base reste présente dans le calcul.
2. **Mise à jour itérative ajustée :**
   - La deuxième partie de la formule, $\frac{1}{|Nh(v_1,v_2)|} \sum_{(v_1',v_2') \in Nh(v_1,v_2)} Sim_i(v_1', v_2')$, calcule la similarité moyenne basée sur les voisins des nœuds v_1 et v_2, comme dans la formule initiale.
   - Ce terme incorpore l'influence structurelle des voisinages.
3. **Combinaison pondérée :**
   - La nouvelle formule introduit un facteur de pondération, α, pour combiner les deux composantes :
     - $\alpha Sim_0(v_1, v_2)$ donne plus ou moins de poids à la similarité initiale, selon la valeur de α.

— $(1 - \alpha)$ pondère l'importance de la similarité calculée à partir des voisins.

— Cette approche équilibre l'influence de la similarité initiale et celle obtenue par les itérations, évitant que les valeurs ne tendent trop rapidement vers zéro ou soient dominées par les voisins.

4.6.4 Avantages de cette formule

— **Préservation de l'information initiale :** La similarité initiale (Sim_0) garantit que des informations importantes sur v_1 et v_2 ne sont pas perdues au fil des itérations.

— **Flexibilité :** Le paramètre α permet de régler l'importance relative des deux composantes, selon le contexte ou les besoins spécifiques.

— **Stabilité accrue :** En conservant une partie fixe (Sim_0), cette formule limite l'effet de dilution observé dans la première méthode.

En résumé, cette formule constitue une amélioration robuste qui combine l'information initiale avec des mises à jour itératives, tout en évitant les problèmes de disparition des similarités observés avec la première formule. Ces conclusions résultent des observations réalisées à partir des illustrations effectuées sur nos graphes d'exemple dans Excel. Vous trouverez ci-dessous une explication détaillée de la méthode utilisée pour effectuer les calculs avec la deuxième formule, qui aboutit à une convergence.

4.6.5 Hypothèses

Le facteur d'atténuation utilisé dans les calculs est :

$$\alpha = 0.7$$

Les voisinages sont définis comme suit :

— Les voisins d'une vache sont ses lactations.

— Les voisins d'une lactation sont la vache correspondante et ses tests.

— Les voisins d'un test sont les lactations auxquelles il appartient.

4.6.6 Notations

— Vc_1 et Vc_2 : les deux vaches.

— $Lt_{1,1}$ et $Lt_{1,2}$: les lactations de Vc_1.

— $Lt_{2,1}$ et $Lt_{2,2}$: les lactations de Vc_2.

— $Ts_{x,y,z}$: le test z lié à la lactation (x, y).

4.6.7 Similarités d'attributs

Le Tableau 4.4 résume notre méthode de calcul des similarités initiales en fonction des différents types d'attributs, ainsi que les méthodes et exemples associés pour calculer la similarité.

Type d'attributs	Méthode de similarité	Exemple de sortie
Valeurs manquantes	Retourne 0.0	`None` vs 3 $\rightarrow$ 0.0
Chaînes de caractères	Jaccard (intersection / union des caractères)	`"chat"` vs `"chaton"` $\rightarrow$ 0.67
Nombres	Distance absolue, inversée par une formule simple	3 vs 4 $\rightarrow$ 0.5
Dates	Décroissance exponentielle selon la différence en jours	`"2023-01-01"` vs `"2023-01-11"` $\rightarrow$ 0.905
Booléens	Identité stricte (1.0 ou 0.0)	`True` vs `False` $\rightarrow$ 0.0
Collections	Moyenne des similarités entre éléments correspondants	`[1, 2]` vs `[1, 3]` $\rightarrow$ 0.75

Tableau 4.4 – Résumé de notre méthode de calcul de la similarité des attributs

4.6.7.1 Similarité d'attributs des vaches

Les attributs considérés pour les vaches sont les suivants :

— Vc_1 : { $animal\ breed = Jersey, rsn = [Retirement]$ }

— Vc_2 : { $animal\ breed = Holstein, rsn = [Old\ age]$ }

Exemple de calculs :

$$sim_a^{animal\ breed}(Vc_1, Vc_2) = 0.1818$$

$$sim_a^{rsn}(Vc_1, Vc_2) = 0.083$$

La similarité initiale entre les vaches est donc :

$$Sim^0(Vc_1, Vc_2) = \frac{0.1818 + 0.083}{2} = 0.13$$

4.6.7.2 Similarité d'attributs des lactations

Les lactations sont comparées à l'aide des attributs $ersn$ et cv. Exemple pour $Lt_{1,1}$ et $Lt_{2,1}$:

$$sim_a^{l\_e\_r}(Lt_{1,1}, Lt_{2,1}) = 0$$

$$sim_a^{cv}(Lt_{1,1}, Lt_{2,1}) = 1$$

La moyenne donne :

$$Sim^0(Lt_{1,1}, Lt_{2,1}) = \frac{0+1}{2} = 0.5$$

Autres exemples :

$$Sim^0(Lt_{1,2}, Lt_{2,2}) = 0.29, \quad Sim^0(Lt_{1,1}, Lt_{2,2}) = 0.26, \quad Sim^0(Lt_{1,2}, Lt_{2,1}) = 0.52$$

4.6.7.3 Similarité d'attributs des tests

Les tests sont comparés à l'aide des attributs numériques mg (% de matière grasse) et pr (% de protéines). Exemple pour $Ts_{1,1,1}$ et $Ts_{2,1,1}$:

$$sim_a^{mg}(Ts_{1,1,1}, Ts_{2,1,1}) = 1, \quad sim_a^{pr}(Ts_{1,1,1}, Ts_{2,1,1}) = 0.66$$

La moyenne donne :

$$Sim^0(Ts_{1,1,1}, Ts_{2,1,1}) = \frac{1 + 0.66}{2} = 0.83$$

Le Tableau 4.5 indique les valeurs de Sim^0 pour les tests.

4.6.8 Première itération (i = 1)

On applique la formule itérative :

$$Sim^{i+1}(v_1, v_2) = \alpha Sim^0(v_1, v_2) + \frac{1-\alpha}{|N_h(v_1, v_2)|} \sum_{(v_1', v_2') \in N_h(v_1, v_2)} Sim^i(v_1', v_2')$$

Paires de nœuds (T_{s1}, T_{s2})	**Sim**^{0}
sima$(T_{s1,1,1}, T_{s2,1,1})$	0.83
sima$(T_{s1,1,1}, T_{s2,1,2})$	0.69
sima$(T_{s1,1,1}, T_{s2,2,1})$	0.78
sima$(T_{s1,1,1}, T_{s2,2,2})$	0.40
sima$(T_{s1,1,1}, T_{s2,2,3})$	0.35
sima$(T_{s1,1,2}, T_{s2,1,1})$	0.54
sima$(T_{s1,1,2}, T_{s2,1,2})$	0.73
sima$(T_{s1,1,2}, T_{s2,2,1})$	0.81
sima$(T_{s1,1,2}, T_{s2,2,2})$	0.4
sima$(T_{s1,1,2}, T_{s2,2,3})$	0.48
sima$(T_{s1,1,3}, T_{s2,1,1})$	0.46
sima$(T_{s1,1,3}, T_{s2,1,2})$	0.42
sima$(T_{s1,1,3}, T_{s2,2,1})$	0.60
sima$(T_{s1,1,3}, T_{s2,2,2})$	0.30
sima$(T_{s1,1,3}, T_{s2,2,3})$	0.30
sima$(T_{s1,2,1}, T_{s2,1,1})$	0.66
sima$(T_{s1,2,1}, T_{s2,1,2})$	0.69
sima$(T_{s1,2,1}, T_{s2,2,1})$	0.83
sima$(T_{s1,2,1}, T_{s2,2,2})$	0.39
sima$(T_{s1,2,1}, T_{s2,2,3})$	0.39
sima$(T_{s1,2,2}, T_{s2,1,1})$	0.49
sima$(T_{s1,2,2}, T_{s2,1,2})$	0.42
sima$(T_{s1,2,2}, T_{s2,2,1})$	0.57
sima$(T_{s1,2,2}, T_{s2,2,2})$	0.28
sima$(T_{s1,2,2}, T_{s2,2,3})$	0.34

Tableau 4.5 – Valeurs de Sim^0 pour différentes paires de nœuds (T_{s1}, T_{s2}).

Pour les Vaches :

$$\begin{aligned} Sim^1(Vc_1, Vc_2) = 0.7 \cdot Sim^0(Vc_1, Vc_2) \\ + \frac{1-0.7}{4}\Big(Sim^0(Lt_{1,1}, Lt_{2,1}) + Sim^0(Lt_{1,1}, Lt_{2,2}) \\ + Sim^0(Lt_{1,2}, Lt_{2,1}) + Sim^0(Lt_{1,2}, Lt_{2,2})\Big) \end{aligned}$$

En substituant les valeurs de Sim^0 :

$$Sim^1(Vc_1, Vc_2) = 0.7 \cdot 0.13 + \frac{1-0.7}{4}(0.5 + 0.29 + 0.26 + 0.52) = 0.21$$

Pour les Lactations : De manière similaire, pour $Lt_{1,1}$ et $Lt_{2,1}$, nous avons :

$$\begin{aligned} Sim^1(Vc_1, Vc_2) = 0.7 \cdot Sim^0(Vc_1, Vc_2) \\ + \frac{1-0.7}{4}\left(Sim^0(Lt_{1,1}, Lt_{2,1}) + Sim^0(Lt_{1,1}, Lt_{2,2})\right. \\ \left. + Sim^0(Lt_{1,2}, Lt_{2,1}) + Sim^0(Lt_{1,2}, Lt_{2,2})\right) \end{aligned}$$

En substituant les valeurs :

$$Sim^1(Lt_{1,1}, Lt_{2,1}) = 0.7 \cdot 0.5 + \frac{1-0.7}{7}(0.13 + 0.83 + 0.69 + 0.545 + 0.73 + 0.45 + 0.41) = 0.51$$

De manière similaire, nous avons calculé pour les autres paires de lactations (Tableau 4.6).

Paire de nœuds	**Calcul de Sim^1**
$Sim^1(Lt_{1,1}, Lt_{2,2})$	$0.7 \cdot 0.29 + \frac{1-0.7}{10}(0.13 + 0.78 + 0.40 + 0.34 + 0.81 + 0.40$ $+0.48 + 0.59 + 0.29 + 30) = 0.34$
$Sim^1(Lt_{1,2}, Lt_{2,1})$	$0.7 \cdot 0.26 + \frac{1-0.7}{5}(0.13 + 0.66 + 0.69 + 0.49 + 0.41) = 0.32$
$Sim^1(Lt_{1,2}, Lt_{2,2})$	$0.7 \cdot 0.52 + \frac{1-0.7}{7}(0.13 + 0.83 + 0.38 + 0.39 + 0.57$ $+0.28 + 0.34) = 0.49$

Tableau 4.6 – Calculs de Sim^1 pour différentes paires de lactations

Finalement, nous calculons Sim^1 pour les tests en suivant le même schéma. Les valeurs sont indiquées dans la Tableau 4.7.

Après la première itération, le même calcul a été répété de manière itérative jusqu'à atteindre une convergence, c'est-à-dire un point où les valeurs obtenues restent stables et ne varient plus.

Paires de nœuds (T_{s1}, T_{s2})	**Sim1**
sima$(T_{s1,1,1}, T_{s2,1,1})$	0.73
sima$(T_{s1,1,1}, T_{s2,1,2})$	0.63
sima$(T_{s1,1,1}, T_{s2,2,1})$	0.63
sima$(T_{s1,1,1}, T_{s2,2,2})$	0.37
sima$(T_{s1,1,1}, T_{s2,2,3})$	0.33
sima$(T_{s1,1,2}, T_{s2,1,1})$	0.53
sima$(T_{s1,1,2}, T_{s2,1,2})$	0.66
sima$(T_{s1,1,2}, T_{s2,2,1})$	0.65
sima$(T_{s1,1,2}, T_{s2,2,2})$	0.36
sima$(T_{s1,1,2}, T_{s2,2,3})$	0.42
sima$(T_{s1,1,3}, T_{s2,1,1})$	0.47
sima$(T_{s1,1,3}, T_{s2,1,2})$	0.44
sima$(T_{s1,1,3}, T_{s2,2,1})$	0.50
sima$(T_{s1,1,3}, T_{s2,2,2})$	0.29
sima$(T_{s1,1,3}, T_{s2,2,3})$	0.29
sima$(T_{s1,2,1}, T_{s2,1,1})$	0.54
sima$(T_{s1,2,1}, T_{s2,1,2})$	0.56
sima$(T_{s1,2,1}, T_{s2,2,1})$	0.74
sima$(T_{s1,2,1}, T_{s2,2,2})$	0.43
sima$(T_{s1,2,1}, T_{s2,2,3})$	0.43
sima$(T_{s1,2,2}, T_{s2,1,1})$	0.42
sima$(T_{s1,2,2}, T_{s2,1,2})$	0.37
sima$(T_{s1,2,2}, T_{s2,2,1})$	0.55
sima$(T_{s1,2,2}, T_{s2,2,2})$	0.35
sima$(T_{s1,2,2}, T_{s2,2,3})$	0.39

Tableau 4.7 – Valeurs de Sim1 pour différentes paires de nœuds (T_{s1}, T_{s2})

4.7 Implémentation de l'algorithme SimRank+

L'algorithme **SimRank+** est une extension de l'algorithme SimRank, qui nous avons conçu pour évaluer la similarité entre des paires de nœuds dans un graphe. Il combine des aspects structurels (voisinage des nœuds) et des caractéristiques des nœuds (attributs) pour produire une mesure de similarité robuste, tout en assurant une convergence rapide grâce à une approche itérative.

4.7.1 Objectif de l'algorithme

Le but principal de SimRank+ est de calculer la similarité entre des paires de nœuds homogènes (nœuds ayant le même type) dans un graphe hétérogène. Il s'agit d'une tâche essentielle dans des domaines tels que :

- La recherche d'entités similaires dans des graphes de connaissances.
- La classification de nœuds en fonction de leurs propriétés structurelles et attributaires.
- L'analyse de réseaux sociaux ou biologiques pour identifier des entités ayant des fonctions similaires.

SimRank+ est particulièrement utile dans les contextes où les graphes contiennent des types variés de nœuds et de relations, nécessitant une approche spécifique pour des paires de nœuds homogènes.

4.7.2 Paramètres et initialisation

L'algorithme commence par l'initialisation des paramètres suivants :

- **Coefficient d'atténuation** (α) : Ce paramètre contrôle l'importance relative des similarités initiales des nœuds par rapport aux similarités structurelles calculées à chaque itération. Une valeur typique est $\alpha = 0.7$, ce qui donne un équilibre entre ces deux composantes.
- **Seuil de convergence** (ϵ) : Utilisé pour déterminer si les différences entre les valeurs de similarité calculées à deux itérations consécutives sont suffisamment petites pour arrêter l'algorithme. Une valeur courante est $\epsilon = 10^{-5}$.
- **Nombre maximum d'itérations** (*max_iterations*) : Un plafond pour éviter une boucle infinie dans le cas où la convergence n'est pas atteinte. Typiquement, *max_iterations* est fixé à 100.
- **Dictionnaires de stockage** :
  - *sim_current* : Contient les similarités calculées à l'itération en cours.
  - *sim_next* : Contient les similarités prévues pour l'itération suivante.
  - *sim_0* : Contient les similarités initiales calculées à partir des attributs des nœuds.

4.7.3 Principales étapes de l'algorithme

4.7.3.1 Calcul des similarités initiales (Sim_0)

Pour chaque paire de nœuds homogènes (v_1, v_2), la similarité initiale est calculée en utilisant une fonction appelée *sima*, qui mesure la similarité entre les attributs des nœuds. Cette étape garantit que l'algorithme prend en compte les propriétés intrinsèques des nœuds dès le départ. La fonction *sima* peut être définie comme suit :

$$sima(v_1, v_2) = \textit{attribut\_similarity}(v_1, v_2)$$

où *attribut_similarity* est une fonction spécifique, par exemple basée sur une distance euclidienne ou cosinus entre les vecteurs d'attributs des nœuds.

4.7.3.2 Détection des voisins homogènes (Nh)

Pour chaque paire de nœuds (v_1, v_2), l'ensemble des voisins homogènes est identifié. Cet ensemble contient toutes les paires (v_1', v_2') où v_1' est un voisin de v_1 et v_2' est un voisin de v_2, et où v_1' et v_2' sont de même type. Cette étape exploite les relations structurelles dans le graphe pour guider le calcul des similarités.

4.7.3.3 Mise à jour des similarités (Sim_{t+1})

L'algorithme met à jour les valeurs de similarité pour chaque paire de nœuds en combinant :

- La similarité initiale (Sim_0).
- La moyenne des similarités des voisins homogènes.

La formule de mise à jour est donnée par :

$$Sim_{t+1}(v_1, v_2) = \alpha \cdot Sim_0(v_1, v_2) + (1 - \alpha) \cdot \frac{\sum_{(v_1', v_2') \in Nh(v_1, v_2)} Sim_t(v_1', v_2')}{|Nh(v_1, v_2)|}$$

Si l'ensemble des voisins homogènes est vide, la contribution de cette partie est nulle.

4.7.3.4 Critère de convergence

Après chaque itération, la différence absolue entre $Sim_{t+1}(v_1, v_2)$ et $Sim_t(v_1, v_2)$ est comparée au seuil ϵ. Si toutes les différences sont inférieures à ϵ, l'algorithme converge. Sinon, il passe à l'itération suivante.

4.7.4 Pseudo-code détaillé

Le pseudo-code de notre méthode est présenté dans l'algorithme 1.

Algorithm 1 Calcul de SimRank+ pour des paires de nœuds homogènes

Initialisation :
$\alpha \leftarrow 0.7$ ▷ Coefficient d'atténuation
$\epsilon \leftarrow 1e-5$ ▷ Seuil de convergence
$max\_iterations \leftarrow 100$ ▷ Nombre maximum d'itérations
$sim^0, sim\_current, sim\_next \leftarrow \{\}$ ▷ Dictionnaires pour la similarité initiale, courante et réactualisée
for all $(v_1, v_2) \in homogeneous\_node\_pairs$ **do**
    $sim^0[(v_1, v_2)] \leftarrow$ sima(v_1, v_2) ▷ Fonction calculant attribut_similarity(v_1, v_2)
    $sim\_current[(v_1, v_2)] \leftarrow$ sima(v_1, v_2)
end for
$iteration \leftarrow 1$
while $\neg converged$ **and** $iteration < max\_iterations$ **do**
    $converged \leftarrow True$
    for all $(v_1, v_2) \in homogeneous\_node\_pairs$ **do**
        $nh\_v1\_v2 \leftarrow$ Nh(v_1, v_2) ▷ Les couples $(v_1', v_2') \in N(v_1) \times N(v_2)$ de même type
        $sum\_neighbors\_sim \leftarrow 0$
        if $|nh\_v1\_v2| > 0$ **then**
            for all $(v_1', v_2') \in nh\_v1\_v2$ **do**
                $sum\_neighbors\_sim \leftarrow sum\_neighbors\_sim + sim\_current[(v_1', v_2')]$
            end for
            $average\_neighbor\_sim \leftarrow sum\_neighbors\_sim / |nh\_v1\_v2|$
        else
            $average\_neighbor\_sim \leftarrow 0$
        end if
        $sim\_next[(v_1, v_2)] \leftarrow \alpha \cdot sim^0[(v_1, v_2)] + (1 - \alpha) \cdot average\_neighbor\_sim$
        if $|sim\_next[(v_1, v_2)] - sim\_current[(v_1, v_2)]| > \epsilon$ **then**
            $converged \leftarrow False$
        end if
    end for
    $sim\_current \leftarrow sim\_next$; $iteration++$
end while
Résultat : $sim\_current$ contient la similarité Sim$\infty(v_1, v_2)$ pour toutes les paires homogènes.

CHAPITRE 5
RÉSULTATS ET DISCUSSION

Dans ce chapitre, nous présenterons l'analyse des résultats. Ensuite, nous exposerons nos résultats concernant la valeur de **alpha**. Enfin, nous présenterons une discussion par rapport à notre méthodologie et les résultats obtenus.

5.1 Analyse des résultats

L'analyse des similarités calculées pour chaque paire de vaches repose sur trois dimensions principales : les attributs des vaches, les attributs sur les performances au cours des périodes de lactation, et les attributs des tests journaliers. Ces similarités sont ensuite agrégées pour produire une similarité globale. Cette approche permet d'évaluer les relations entre les entités de manière hiérarchique tout en tenant compte des attributs spécifiques à chaque niveau. Pour l'analyse une sélection des 5000 premières paires de vaches à été faites.

5.1.1 Analyse des similarités par dimension

Les résultats montrent que la similarité globale est influencée par la contribution des trois dimensions :

- **Similarité des vaches** : Moyenne de **0.231** avec un écart type de **0.0668**, ce qui reflète une hétérogénéité significative entre les attributs propres aux vaches.
- **Similarité des lactations** : Moyenne légèrement supérieure de **0.2426** avec une faible dispersion (**écart type de 0.027**), indiquant une plus grande homogénéité des performances des lactations.
- **Similarité des tests** : Moyenne élevée de **0.4687** avec un écart type très faible (**0.0247**), faisant des jours de tests la dimension la plus stable et discriminante.
- **Similarité globale** : Moyenne de **0.3296** avec une dispersion modérée (**écart type de 0.0253**), combinant les trois dimensions.

Les résultats montrent que les jours de tests apportent une granularité importante pour évaluer les similarités, tandis que les attributs propres aux vaches et les lactations ajoutent des informations complémentaires mais plus dispersées, et aussi les similarités globales, comprises entre 0.297 et 0.327, traduisent une agrégation équilibrée des trois dimensions, montrant que la méthodologie multi-niveaux intègre efficacement

les relations structurelles et attributives.

Le Tableau 5.1 présente les valeurs calculées pour un échantillon de paires de vaches :

Tableau 5.1 – Similarités calculées pour chaque paire de vaches

Paire de Vaches	Similarité Vaches	Similarité Lactations	Similarité Tests	Similarité Globale
(10245383, 10245790)	0.219831	0.230205	0.441960	0.311795
(10245383, 10245964)	0.232809	0.242697	0.449193	0.322329
(10245383, 10246408)	0.220064	0.220214	0.413062	0.297308
(10245383, 10247127)	0.257130	0.234880	0.448030	0.326815
(10245383, 10247623)	0.224055	0.241295	0.457065	0.322431
(10245383, 10248855)	0.222391	0.227969	0.438985	0.310702
(10245383, 10249684)	0.220000	0.224242	0.411371	0.297821
(10245383, 10249900)	0.346922	0.223075	0.412487	0.335994
(10245383, 10251039)	0.225377	0.242165	0.449646	0.320121
(10245383, 10251186)	0.214324	0.220296	0.447292	0.309303
(10245383, 10251188)	0.227701	0.234005	0.453150	0.319772
(10245383, 10252198)	0.215899	0.227070	0.428638	0.304346
(10245383, 10253155)	0.251293	0.215422	0.454573	0.321844
(10245383, 10253594)	0.233300	0.244332	0.440481	0.319482
(10245383, 10253625)	0.238931	0.236438	0.442819	0.319738
(10245383, 10253650)	0.081421	0.213071	0.395780	0.246660
(10245383, 10253954)	0.217892	0.232190	0.443948	0.312604
(10245383, 10254626)	0.227504	0.233345	0.449515	0.318061
(10245383, 10254656)	0.224466	0.234886	0.435117	0.311852
(10245383, 10254667)	0.219340	0.228568	0.433729	0.307864

5.1.2 Implications des résultats

Les résultats obtenus permettent de :

— **Identifier des groupes homogènes :** Les paires avec des similarités globales proches (0.322329 et 0.322431) peuvent être regroupées pour des analyses plus approfondies.
— **Évaluer les relations multi-niveaux :** La méthodologie démontre que les trois dimensions contribuent de manière significative à la similarité globale, tout en capturant des nuances entre les paires de vaches.
— **Optimiser la gestion des troupeaux :** Les similarités calculées peuvent aider à regrouper les vaches en fonction de leurs performances ou attributs, facilitant ainsi les décisions en matière de sélection ou d'alimentation.

Ces analyses confirment que la méthodologie développée est robuste et adaptée pour traiter des données hiérarchiques et multidimensionnelles telles que celles de la production laitière.

5.2 Analyse de la formule

La formule proposée :

$$\text{Sim}(v_1, v_2) = \alpha \cdot \text{Sim}_0(v_1, v_2) + (1-\alpha) \cdot \frac{1}{|N_h(v_1, v_2)|} \sum_{(v_1', v_2') \in N_h(v_1, v_2)} \text{Sim}_i(v_1', v_2') \tag{5.1}$$

est une **fonction itérative** qui mesure la similarité entre deux nœuds homogènes v_1 et v_2. Cette formule combine deux types de similarité :

1. $\text{Sim}_0(v_1, v_2)$: La *similarité des attributs*, représentant la ressemblance directe entre les propriétés (ou caractéristiques) des nœuds v_1 et v_2. Cette contribution est pondérée par le paramètre α.
2. $\text{Sim}_i(v_1', v_2')$: La *similarité des liens*, qui considère la ressemblance entre les voisins respectifs des nœuds v_1 et v_2. La moyenne de ces similarités pour les paires de voisins est pondérée par $(1-\alpha)$.

5.2.1 Rôle du paramètre

Le paramètre $\alpha \in [0, 1]$ joue un rôle fondamental dans l'équilibre entre les deux types de similarités :

— **Si α est grand (proche de 1)** : La formule accorde une importance prédominante à la *similarité des attributs* (Sim_0). Dans ce cas, la similarité est principalement déterminée par les caractéristiques internes des nœuds v_1 et v_2. Cela conduit à une **convergence plus rapide** du processus itératif, car

la similarité des attributs est directement calculée et ne dépend pas des voisins. En d'autres termes, moins d'itérations seront nécessaires pour stabiliser la valeur de la similarité globale.

— **Si α est petit (proche de 0)** : La formule donne davantage de poids à la *similarité des liens* (Sim_i), c'est-à-dire aux relations entre les nœuds et leurs voisins. Cette approche repose sur une propagation des similarités à travers les voisins, ce qui augmente la **dépendance structurelle** et nécessite plusieurs itérations pour que les valeurs convergent. Par conséquent, **la convergence est plus lente** car la propagation de la similarité dans le graphe devient plus complexe et progressive.

5.2.2 Interprétation de la convergence

— **Lorsque** $\alpha \to 1$: La similarité finale entre v_1 et v_2 est essentiellement déterminée par leurs attributs internes. Le calcul devient rapide car il ne dépend presque pas de la structure environnante (les voisins). Cela est particulièrement avantageux si les attributs des nœuds sont fortement discriminants ou informatifs.

— **Lorsque** $\alpha \to 0$: La similarité est dominée par les liens structurels. Le processus itératif prend alors plus de temps pour converger, car la similarité doit se propager à travers les voisins ($N_h(v_1, v_2)$) jusqu'à atteindre un équilibre stable.

— **Cas intermédiaires ($0 < \alpha < 1$)** : La formule équilibre les deux types de similarités. La vitesse de convergence dépend alors de l'importance relative des attributs par rapport aux relations structurelles. Une valeur intermédiaire de α permet de combiner efficacement les informations des attributs et des liens, au prix d'une convergence modérée.

5.2.3 Conclusion

Le paramètre α contrôle directement la rapidité de convergence de la formule en ajustant l'importance relative des attributs et des liens. Une valeur élevée de α favorise la similarité basée sur les attributs, conduisant à une **convergence plus rapide**. À l'inverse, une valeur faible de α favorise la similarité structurelle, ce qui **ralentit la convergence** en raison de la propagation nécessaire à travers les voisins. Le choix optimal de α dépend donc du contexte d'application et de la nature des données (attributs discriminants vs structure informative).

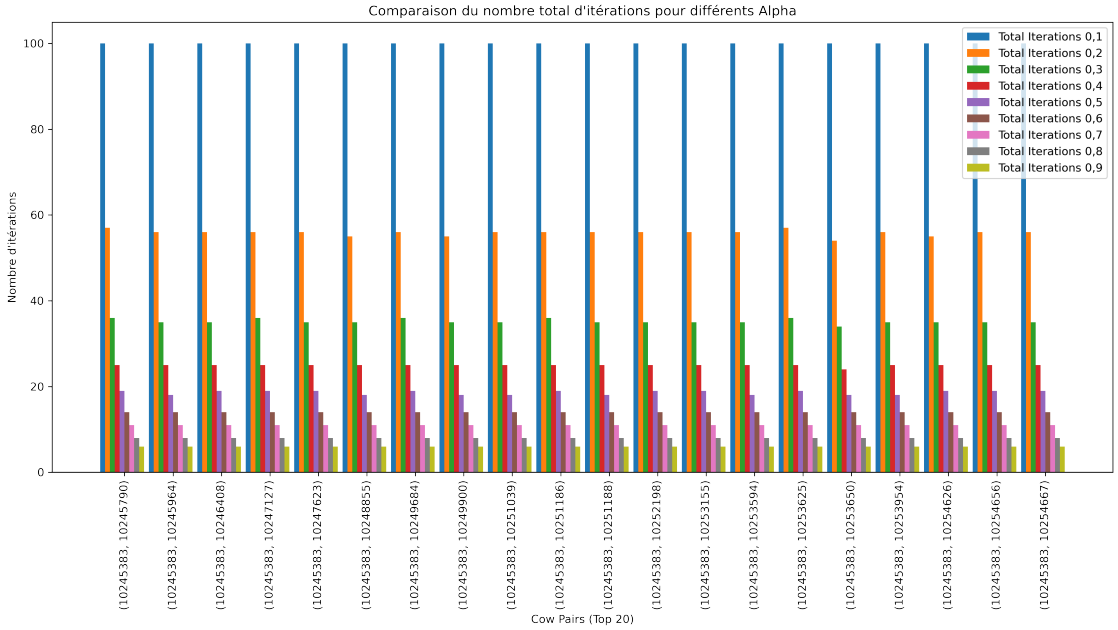

Figure 5.1 – Comparaison des resultats pour différents α

5.3 Analyse statique des similarités entre Vaches

Cette section présente une analyse approfondie des similarités entre les paires de vaches, en se basant sur quatre dimensions principales : la similarité des attributs des vaches, celle des performances en lactation, celle des données des jours de tests, et une similarité globale combinant ces trois aspects. Les résultats incluent des corrélations entre dimensions,des observations sur la distribution des similarités, de l'analyse des clusters et aussi le clustering des paires de vaches.

5.3.1 Analyse des corrélations

La matrice de corrélation (voir la Figure 5.2) révèle des relations intéressantes entre les dimensions :

- **Corrélation élevée entre Total Similarity Vaches et Similarité Globale** ($r = 0.85$) : Les attributs physiques ou génétiques des vaches jouent un rôle prédominant dans le calcul de la similarité globale. Cette forte corrélation reflète l'importance de ces caractéristiques structurelles.
- **Corrélation modérée entre Total Similarity Lactations et Similarité Globale** ($r = 0.45$) : Les performances des lactations ont une influence moyenne sur la similarité globale, suggérant que les facteurs environnementaux et alimentaires jouent un rôle complémentaire aux attributs physiques.
- **Corrélation faible entre Total Similarity Tests et Similarité Globale** ($r = 0.46$) : Bien que les simi-

larités basées sur les tests journaliers soient importantes, leur influence sur la similarité globale est relativement moindre. Cependant, leur faible variance en fait un indicateur fiable pour détecter des tendances spécifiques.

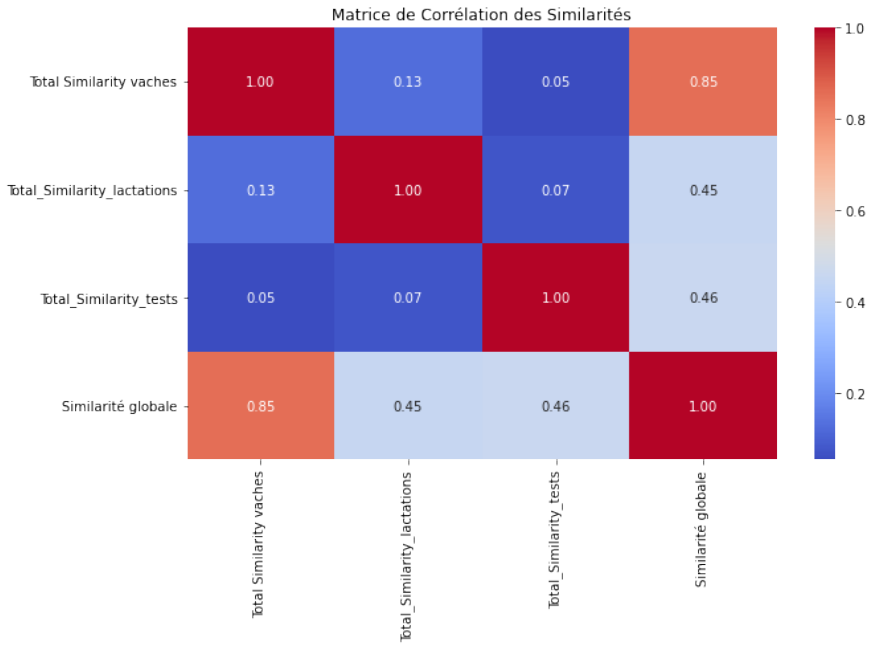

Figure 5.2 – Matrice de corrélation des dimensions de similarité.

5.3.2 Distribution des similarités

Le boxplot (Figure 5.3) révèle des observations importantes sur la variabilité et la fiabilité des différentes mesures de similarité :

- **Total Similarity Tests** : Cette mesure présente une médiane élevée et un intervalle interquartile (IQR) étroit, indiquant une grande cohérence entre les paires de vaches. Cela en fait une mesure fiable pour évaluer les performances quotidiennes.
- **Total Similarity Lactations** : Avec un IQR légèrement plus large et une médiane inférieure, cette mesure reflète une variabilité accrue, probablement liée aux interactions entre génétique et environnement.
- **Total Similarity Vaches** : Cette mesure affiche la plus grande variabilité et une médiane plus basse, reflétant la diversité génétique ou physique importante entre les vaches.

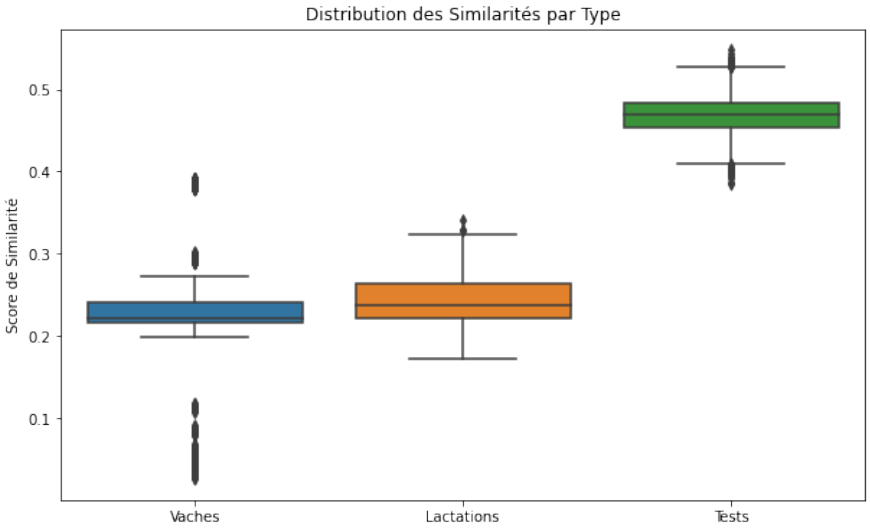

Figure 5.3 – Distribution des similarités par dimension.

5.3.3 Analyse des clusters entre la similarité globale et la similarité des tests

Le graphique Similarité globale vs Similarité des Tests (Figure 5.4) met en lumière des relations complexes entre les dimensions :

- **Présence de clusters** : Des regroupements distincts sont visibles, suggérant que certaines sous-populations de vaches partagent des profils de similarité similaires. Ces clusters pourraient refléter des liens génétiques ou environnementaux communs.
- **Impact des lactations** : Les similarités élevées en lactation (échelle de couleur) se regroupent dans les zones de haute similarité globale, indiquant que les performances lactées contribuent fortement à ces scores.
- **Taille des points** : Les grandes tailles de points (correspondant à des similarités élevées des vaches) sont associées à des scores globaux élevés, confirmant le rôle central des attributs physiques.

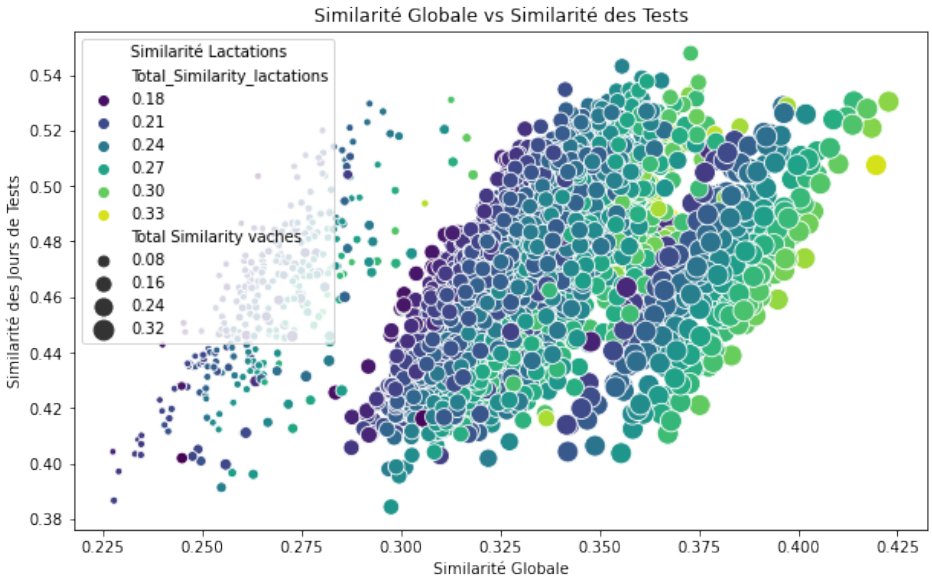

Figure 5.4 – Similarité globale vs similarité des tests

5.3.4 Recommandations et implications pratiques

Les résultats de cette analyse offrent plusieurs pistes pour la gestion du troupeau et l'optimisation des performances :

1. **Priorisation des jours de tests** : Les tests quotidiens étant la dimension la plus discriminante, il est recommandé de les prioriser pour identifier les performances sous-optimales ou détecter des anomalies.
2. **Segmentation des vaches** : La similarité globale peut être utilisée pour regrouper les vaches en clusters homogènes, facilitant une gestion ciblée et efficace.
3. **Approfondissements** :
   — Étudier l'impact de seuils de similarité spécifiques sur les décisions opérationnelles.
   — Analyser les relations entre clusters et autres variables (génétiques, environnementales, sanitaires).

Ces recommandations visent à améliorer la productivité et le bien-être animal, tout en optimisant la prise de décision dans la gestion du troupeau.

5.4 Clustering des paires de Vaches

L'algorithme k-means a permis de diviser les paires de vaches en trois clusters basés sur leurs similarités globales. La Figure 5.5 illustre cette répartition, où chaque point représente une paire de vaches et sa similarité globale est indiquée sur l'axe x.

Figure 5.5 – Clustering des paires de vaches basé sur la Similarité Globale.

Les clusters identifiés révèlent des tendances distinctes dans les similarités globales :

- **Cluster 0 :** Représente les paires avec des similarités globales faibles (< 0.30), indiquant des relations plus éloignées entre ces paires.
- **Cluster 1 :** Contient les paires avec les plus fortes similarités globales (> 0.33), représentant des relations étroites et homogènes.
- **Cluster 2 :** Inclut les paires avec des similarités globales moyennes ($0.30 \leq$ similarité globale ≤ 0.33), montrant des relations modérément homogènes.

Le Tableau 5.2 présente un échantillon des résultats, indiquant les similarités globales et les clusters correspondants.

5.4.1 Interprétation des résultats

Les résultats révèlent des informations importantes pour la gestion des troupeaux :

- **Groupes homogènes :** Les paires du Cluster 1, ayant les plus fortes similarités globales, représentent

Tableau 5.2 – Similarité globale et cluster pour un échantillon de paires de Vaches.

Paire de Vaches	Similarité Globale	Cluster
(10245383, 10245790)	0.311795	2
(10245383, 10245964)	0.322329	1
(10245383, 10246408)	0.297308	2
(10245383, 10247127)	0.326815	1
(10245383, 10247623)	0.322431	1
...	...	...
(10247623, 10253650)	0.250766	0
(10247623, 10253954)	0.345379	1
(10247623, 10254626)	0.341071	1
(10247623, 10254656)	0.320735	2
(10247623, 10254667)	0.306259	2

des vaches avec des performances ou des caractéristiques très proches. Ces groupes peuvent être utilisés pour des stratégies spécifiques, comme la sélection animale.

— **Relations multi-niveaux :** Les clusters montrent que les relations entre les paires varient selon la similarité globale. Le Cluster 0 peut indiquer des vaches présentant des performances ou comportements atypiques nécessitant une attention particulière.

— **Optimisation :** Les clusters homogènes permettent une gestion ciblée, notamment en termes d'alimentation et de diagnostic, afin de maximiser les performances et minimiser les inefficacités.

5.5 Discussion

L'objectif principal de notre recherche était de développer une méthodologie multi-niveau permettant d'analyser les similarités entre des paires de vaches laitières. En s'appuyant sur trois niveaux complémentaires - les attributs individuels des vaches, les attributs sur leurs performances en lactation et les attributs pour les tests journaliers - cette approche vise à fournir une analyse globale et détaillée des relations présentes dans les données. Une telle analyse est essentielle pour comprendre les dynamiques des troupeaux, identifier des groupes homogènes et détecter des anomalies ou des comportements atypiques.

5.5.1 Lien avec l'état de l'art

Le calcul de la similarité entre entités hiérarchiques et multi-niveaux s'inscrit dans un cadre exploré par des approches comme SimRank (Jeh et Widom, 2002a) et les Graph Neural Networks (GNN) (Ma *et al.*, 2021). Ces méthodes exploitent les relations structurelles entre les nœuds d'un graphe et leurs attributs pour calculer la similarité. Cependant, notre approche hybride se distingue en :

- Combinant de manière équilibre les similarités attributives et structurelles, a l'instar des kernels de graphes profonds (Narayanan *et al.*, 2017), tout en restant interprétable.
- Exploitant la granularité des données journalières pour détecter des tendances fines, une dimension souvent sous-exploitée dans les approches traditionnelles.
- Intégrant une structure hiérarchique explicite (vaches, lactations, tests), ce qui la différencie des approches globales comme SimRank, plus adaptées aux relations simples ou plates.

5.5.2 Impact computationnel

Par rapport aux méthodes comme SimRank ou P-Rank (Zhao *et al.*, 2009), notre méthode présente une complexité computationnelle réduite grâce a l'utilisation d'une agrégation pondérée des niveaux. Cependant, il est important de noter que le temps d'exécution reste très long, en raison du processus d'agrégation multi-niveaux et des itérations nécessaires pour assurer la convergence des calculs. En revanche, une optimisation du paramètre α reste cruciale pour équilibrer les similarités structurelles et attributives et également réduire le temps d'exécution.

5.5.3 Interprétation des résultats

Les résultats obtenus démontrent que chaque niveau joue un rôle distinct et complémentaire dans l'évaluation des similarités globales. Les attributs individuels des vaches, par exemple, reflètent les caractéristiques génétiques ou physiologiques des vaches. Leur dispersion notable indique une forte variabilité au sein du troupeau, ce qui est attendu dans des populations animales présentant des différences en termes de lignées, d'âge ou de conditions de vie. Ce niveau est particulièrement utile pour évaluer les différences fondamentales entre les individus.

Les attributs pour les performances en lactation montrent une plus grande homogénéité, probablement

en raison de la standardisation des pratiques d'élevage et de gestion des troupeaux. Les faibles écarts dans cette dimension suggèrent que les conditions environnementales et les interventions humaines jouent un rôle important dans la réduction des variations, rendant cette dimension plus stable et prévisible.

Enfin, les tests journaliers, avec leur moyenne élevée et leur faible variance, se révèlent être l'indicateur le plus stable et le plus précis pour capturer les variations subtiles de performance ou de comportement. Ces résultats mettent en avant l'importance de collecter des données régulières et granuleuses pour évaluer efficacement la santé et la productivité des vaches.

La similarité globale, résultant de l'intégration des trois niveaux, présente une dispersion modérée, traduisant un équilibre dans la contribution de chaque dimension. Cette mesure synthétique est particulièrement utile pour identifier des tendances générales tout en capturant des relations spécifiques.

5.5.4 Implications pratiques

Les implications pratiques de cette étude sont nombreuses et touchent divers aspects de la gestion des troupeaux laitiers. Premièrement, la capacité à identifier des groupes homogènes de vaches, basés sur leurs similarités globales, offre une opportunité précieuse pour appliquer des stratégies de gestion différenciées. Par exemple, les vaches ayant des similarités élevées peuvent être regroupées pour des programmes de reproduction spécifiques ou des régimes alimentaires optimisés.

Deuxièmement, la détection des anomalies est facilitée par cette méthodologie. Les vaches présentant des similarités faibles avec le reste du troupeau peuvent être identifiées comme des cas atypiques, nécessitant une attention particulière. Ces anomalies pourraient indiquer des problèmes de santé sous-jacents, des conditions environnementales défavorables ou des besoins spécifiques en termes de soins.

Troisièmement, l'analyse des données journalières permet une réactivité accrue dans la gestion quotidienne des troupeaux. Les baisses de performance ou les changements comportementaux peuvent être détectés rapidement, permettant des interventions précoces et ciblées. Cette approche proactive est essentielle pour maintenir un niveau de productivité élevé tout en assurant le bien-être des animaux.

5.5.5 Analyse des limites

Malgré ses avantages, cette recherche présente certaines limitations qui doivent être prises en compte. Tout d'abord, la méthodologie repose sur des poids fixes pour chaque dimension dans le calcul de la similarité globale. Cette simplification pourrait limiter la flexibilité de l'analyse, car elle ne tient pas compte des variations possibles dans l'importance relative des dimensions en fonction des contextes.

De plus, l'analyse est basée sur des données statiques, ce qui signifie qu'elle ne capture pas les variations temporelles. Les performances des vaches peuvent évoluer au fil du temps en réponse à des facteurs tels que l'âge, les saisons ou les changements dans les pratiques d'élevage. Une analyse longitudinale permettrait de mieux comprendre ces dynamiques et d'affiner les recommandations.

Enfin, l'étude est limitée par la taille et l'homogénéité de l'échantillon. Les conclusions tirées pourraient ne pas être entièrement généralisables à d'autres troupeaux ou contextes d'élevage. Des études supplémentaires impliquant des échantillons plus diversifiés seraient nécessaires pour valider et étendre ces résultats.

5.5.6 Perspectives futures

Les perspectives futures de cette recherche sont nombreuses et prometteuses. L'une des principales améliorations possibles serait d'intégrer une pondération dynamique des dimensions, permettant d'ajuster automatiquement leur influence en fonction des données spécifiques ou des objectifs de l'analyse. Cela pourrait être réalisé à l'aide de techniques d'apprentissage automatique, qui offrent des outils puissants pour optimiser les modèles.

Une autre direction intéressante serait de passer d'une analyse statique à une analyse dynamique en intégrant des séries temporelles. Cela permettrait non seulement de capturer les tendances à long terme, mais aussi d'identifier des événements ponctuels significatifs, tels que des périodes de stress ou des changements dans les performances.

Enfin, cette méthodologie pourrait être appliquée à d'autres espèces animales ou à d'autres domaines où les relations hiérarchiques jouent un rôle clé. Par exemple, elle pourrait être utilisée pour analyser les similarités dans les systèmes biologiques complexes ou pour optimiser la gestion dans d'autres types d'élevage.

CONCLUSION

La méthodologie développée a permis d'acquérir et de consolider des compétences variées en analyse de données, modélisation de graphes et production laitière. Cette combinaison de connaissances nous a permis de concevoir une approche pour calculer les similarités structurelles et attributives dans des graphes hiérarchiques, visant à évaluer les relations entre les vaches, les lactations et les jours de test. Notre étude visait à élaborer une méthode pour identifier des similarités spécifiques aux performances des vaches dans des conditions variées. Nous avons démontré que ces similarités peuvent être discriminantes pour analyser les performances et détecter des anomalies, offrant ainsi des perspectives intéressantes pour l'optimisation des troupeaux et la productivité des exploitations.

Afin d'évaluer de manière exhaustive notre approche, nous avons appliqué nos méthodes à plusieurs ensembles de données. Ces ensembles ont été soumis à diverses évaluations et comparaisons, permettant ainsi de valider les mesures de similarité grâce à des tests sur les relations multi-niveaux. Les résultats de notre étude ont mis en évidence l'efficacité de notre approche dans l'analyse des performances animales. Nos résultats ont révélé l'existence de similarités structurales cohérentes entre différentes vaches et lactations. Parmi ces similarités, une proportion significative a montré une homogénéité de plus de 50% entre les groupes étudiés. Une analyse approfondie des similarités communes a également souligné l'importance d'une approche multifactorielle pour optimiser l'identification des patterns de performance.

Cependant, notre recherche a également révélé des défis méthodologiques significatifs, notamment en ce qui concerne la gestion des données déséquilibrées et la complexité des dépendances hiérarchiques. En conclusion, notre étude a contribué à l'amélioration de l'analyse des données complexes en utilisant les avancées technologiques pour renforcer la gestion des troupeaux laitiers. En envisageant l'avenir, une extension prometteuse de notre méthodologie serait le développement d'une approche intégrant les données environnementales et comportementales. Cette méthode plus complète pourrait permettre une gestion encore plus précise et proactive des troupeaux, contribuant ainsi à améliorer la productivité et le bien-être global des animaux. Toutefois, cela nécessiterait une infrastructure accrue pour garantir la robustesse et la précision des analyses. En somme, notre étude marque une avancée significative dans l'analyse des données laitières et leur intégration dans les outils d'agriculture de précision. En continuant à explorer ces domaines, nous aspirons à favoriser une agriculture durable et technologiquement avancée.

BIBLIOGRAPHIE

Adamczyk, K., Cywicka, D., Herbut, P. et Trześniowska, E. (2017). The application of cluster analysis methods in assessment of daily physical activity of dairy cows milked in the Voluntary Milking System. *Computers and Electronics in Agriculture*, *141*, 65–72. http://dx.doi.org/10.1016/j.compag.2017.07.007. Récupéré le 2024-11-22 de https://linkinghub.elsevier.com/retrieve/pii/S0168169917300868

Aggarwal, C. C. (2015). *Data Mining : The Textbook*. Springer.

Agri-Mutuel. La production laitière des vaches allaitantes : un enjeu qui pèse. Récupéré de https://www.agri-mutuel.com/elevage/la-production-laitiere-des-vaches-allaitantes-un-enjeu-qui-pese/

Ahmad, A. et Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, *63*(2), 503–527. http://dx.doi.org/10.1016/j.datak.2007.03.016. Récupéré le 2025-01-05 de https://linkinghub.elsevier.com/retrieve/pii/S0169023X0700050X

Alves, G., Couceiro, M. et Napoli, A. (2020). Sélection de mesures de similarité pour les données catégorielles. Dans *EGC 2020 - 20^{e} édition de la conférence Extraction et Gestion des Connaissances*, Bruxelles, Belgique. Récupéré le 2025-03-10 de https://hal.science/hal-02410221/document

Ammar, A., Elouedi, Z. et Lingras, P. (2012). K-Modes Clustering Using Possibilistic Membership. In S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo, et R. R. Yager (dir.), *Advances in Computational Intelligence*, volume 299 596–605. Berlin, Heidelberg : Springer Berlin Heidelberg. Series Title : Communications in Computer and Information Science

Analytics, F. (2023). Harnessing data for sustainable dairy farming. Récupéré de https://www.fossanalytics.com/fr-fr/news-articles/rmt/harnessing-data-for-sustainable

Aranganayagi, S. et Thangavel, K. (2009). Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure. *3*(3).

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. et Ives, Z. (2007). DBpedia : A Nucleus for a Web of Open Data. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, et P. Cudré-Mauroux (dir.), *The Semantic Web*, volume 4825 722–735. Berlin, Heidelberg : Springer Berlin Heidelberg. Series Title : Lecture Notes in Computer Science

Baghshah, M. S. (2009). Semi-Supervised Metric Learning Using Pairwise Constraints. Récupéré de https://www.ijcai.org/proceedings/2009

Bai, Y., Ding, H., Bian, S., Chen, T., Sun, Y. et Wang, W. (2020). SimGNN : A Neural Network Approach to Fast Graph Similarity Computation. arXiv :1808.05689 [cs],

http://dx.doi.org/10.48550/arXiv.1808.05689. Récupéré le 2025-01-05 de http://arxiv.org/abs/1808.05689

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. et Morissette, J. (2008). Bio2RDF : Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, *41*(5), 706–716. http://dx.doi.org/10.1016/j.jbi.2008.03.004. Récupéré le 2025-01-05 de https://linkinghub.elsevier.com/retrieve/pii/S1532046408000415

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. et Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-relational Data.

Box, G. E. et Jenkins, G. M. (1970). *Time Series Analysis : Forecasting and Control*. Holden-Day.

Carr, J. E., Austin, J., Hatfield, D. B. et Bailey, J. S. (1996). The standard deviation as an informative measure of variability in reporting interobserver agreement means. *Journal of Behavior Therapy and Experimental Psychiatry*, *27*(3), 263–267. http://dx.doi.org/10.1016/S0005-7916(96)00024-9. Récupéré le 2025-01-05 de https://linkinghub.elsevier.com/retrieve/pii/S0005791696000249

CIAQ (2024). heatime- surveillance de la reproduction et de la santé des vaches. Récupéré de https://ciaq.com/heatime/

Cruz, A. O., Silva, J. K. L., Silva, E. M. D., Santos, A. C. D. et Antonialli, L. M. (2021). Differences and similarities in the milk production chain : a comparative analysis with the states of Minas Gerais and Paraná. *Independent Journal of Management & Production*, *12*(4), 1034–1051. http://dx.doi.org/10.14807/ijmp.v12i4.1309. Récupéré le 2024-11-22 de http://www.ijmp.jor.br/index.php/ijmp/article/view/1309

Dairy, S. (2024). Sensehub - moniteur de santé et reproduction bovine. Récupéré de https://fr.sensehub.global/monitoring/

de l'Agriculture et de l'Alimentation, M. (2023). Le bien-être et la protection des vaches laitières. Récupéré de https://agriculture.gouv.fr/le-bien-etre-et-la-protection-des-vaches-laitieres

de Lait, F. T. (2023). Valoriser les données de la filière laitière : défis et opportunités. Récupéré de https://franceterredelait.fr/valoriser-les-donnees-de-la-filiere-laitiere-defis-et-opportunites

de l'Élevage (Idele), I. (2020). Référentiel de contrôle des performances pour la production de lait de vache : Protocoles et méthodes de qualification. Récupéré de https://idele.fr/fileadmin/medias/Documents/3_Controle_des_Performances_Lait_Protocoles_et_methodes_de_qualification_20200402.pdf

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S. et Zhang, W. (2014). Knowledge vault : a web-scale approach to probabilistic knowledge fusion. Dans *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 601–610., New York New York USA. ACM. http://dx.doi.org/10.1145/2623330.2623623. Récupéré le 2025-01-05 de https://dl.acm.org/doi/10.1145/2623330.2623623

du Canada, G. (2024). Centre canadien d'information laitière. Récupéré de https://agriculture.canada.ca/fr/secteur/production-animale/centre-canadien-information-laitiere/industrie-laitiere

Duda, R. O., Hart, P. E. et Stork, D. G. (2000). *Pattern Classification* (2nd éd.). John Wiley & Sons.

Grover, A. et Leskovec, J. (2016). node2vec : Scalable Feature Learning for Networks. arXiv :1607.00653 [cs], http://dx.doi.org/10.48550/arXiv.1607.00653. Récupéré le 2025-01-05 de http://arxiv.org/abs/1607.00653

Générale, L. C. (2023). Différence entre bœuf, taureau, vache, génisse et veau. Récupéré de https://www.laculturegenerale.com/difference-boeuf-taureau-vache-genisse-veau/

Hamilton, W. L., Ying, R. et Leskovec, J. (2018). Inductive Representation Learning on Large Graphs. arXiv :1706.02216 [cs], http://dx.doi.org/10.48550/arXiv.1706.02216. Récupéré le 2025-01-05 de http://arxiv.org/abs/1706.02216

Han, J., Kamber, M. et Pei, J. (2011). *Data Mining : Concepts and Techniques*. Morgan Kaufmann.

Han, J., Pei, J. et Tong, H. (2022). *Data Mining : Concepts and Techniques* (4th éd.). Morgan Kaufmann.

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S. et Zimmermann, A. (2022). Knowledge Graphs. *ACM Computing Surveys*, *54*(4), 1–37. http://dx.doi.org/10.1145/3447772. Récupéré le 2025-01-05 de https://dl.acm.org/doi/10.1145/3447772

in World Farming (CIWF), C. (2023). Vaches laitières : comprendre leur mode de vie. Récupéré de https://www.ciwf.fr/animaux-delevage/vaches-laitieres/

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et du Jura. Medium : text/html,application/pdf,text/html Publisher : Imprimerie Corbaz & Comp., http://dx.doi.org/10.5169/SEALS-266450. Récupéré le 2025-01-05 de https://www.e-periodica.ch/digbib/view?pid=bsv-002:1901:37::790

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2023). *An Introduction to Statistical Learning* (2nd éd.). Springer.

Jeh, G. et Widom, J. (2002a). SimRank : A Measure of Structural-Context Similarity. Dans *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 538–543. ACM.

Jeh, G. et Widom, J. (2002b). SimRank : A Measure of Structural-Context Similarity. Dans *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 538–543. ACM. http://dx.doi.org/10.1145/775047.775126

Ji, S., Pan, S., Cambria, E., Marttinen, P. et Yu, P. S. (2022). A Survey on Knowledge Graphs : Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(2), 494–514. http://dx.doi.org/10.1109/TNNLS.2021.3070843. Récupéré le 2024-11-22 de https://ieeexplore.ieee.org/document/9416312/

Jolliffe, I. T. et Cadima, J. (2016). *Principal Component Analysis*. Springer.

lactanet (2024). Lactanet - contrôle laitier et gestion des données au canada. Récupéré de https://lactanet.ca/

lely (2024). Lely astronaut a5 - robot de traite. Récupéré de https://www.lely.com/fr/solutions/traite/astronaut-a5/

Li, Y., Gu, C., Dullien, T., Vinyals, O. et Kohli, P. (2019). Graph Matching Networks for Learning the Similarity of Graph Structured Objects. arXiv :1904.12787 [cs], http://dx.doi.org/10.48550/arXiv.1904.12787. Récupéré le 2025-01-05 de http://arxiv.org/abs/1904.12787

Ma, G., Ahmed, N. K., Willke, T. L. et Yu, P. S. (2021). Deep graph similarity learning : a survey. *Data Mining and Knowledge Discovery*, *35*(3), 688–725. http://dx.doi.org/10.1007/s10618-020-00733-5. Récupéré le 2024-11-22 de https://link.springer.com/10.1007/s10618-020-00733-5

Montgomery, D. C. (2017). *Introduction to Statistical Quality Control*. Wiley.

Moocall (2024). Moocall - capteur de vêlage. Récupéré de https://moocall.com/

Mukhopadhyay, A. et Maulik, U. (2007). Multiobjective approach to categorical data clustering. Dans *2007 IEEE Congress on Evolutionary Computation*, 1296–1303., Singapore. IEEE. http://dx.doi.org/10.1109/CEC.2007.4424620. Récupéré le 2025-01-05 de http://ieeexplore.ieee.org/document/4424620/

Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y. et Jaiswal, S. (2017). graph2vec : Learning Distributed Representations of Graphs. arXiv :1707.05005 [cs], http://dx.doi.org/10.48550/arXiv.1707.05005. Récupéré le 2025-01-05 de http://arxiv.org/abs/1707.05005

Nickel, M., Murphy, K., Tresp, V. et Gabrilovich, E. (2016). A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, *104*(1), 11–33. arXiv :1503.00759 [stat], http://dx.doi.org/10.1109/JPROC.2015.2483592. Récupéré le 2025-01-05 de http://arxiv.org/abs/1503.00759

Noriap (2023). Améliorer les performances en production laitière : les outils disponibles. Récupéré de https://www.noriap.com/blog/ameliorer-performances-production-laitiere

Parameswari, P., Samath, J. A. et Saranya, S. (2015). Scalable Clustering Using Rank Based Preprocessing Technique for Mixed Data Sets Using Enhanced Rock Algorithm. *African Journal of Basic & Applied Sciences*, *7*(3), 129–136.

Paulheim, H. (2016). Knowledge graph refinement : A survey of approaches and evaluation methods. *Semantic Web*, *8*(3), 489–508. http://dx.doi.org/10.3233/SW-160218. Récupéré le 2025-01-05 de https://journals.sagepub.com/doi/full/10.3233/SW-160218

Perozzi, B., Al-Rfou, R. et Skiena, S. (2014). DeepWalk : Online Learning of Social Representations. Dans *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710. arXiv :1403.6652 [cs], http://dx.doi.org/10.1145/2623330.2623732. Récupéré le 2025-01-05 de http://arxiv.org/abs/1403.6652

Rifqi, M. et Bouchon-Meunier, B. (2004). Set-theoretic similarity measures. Dans *Proceedings of the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*. Récupéré de https://webia.lip6.fr/~rifqi/RifqiBouchon.pdf

Russell, S. et Norvig, P. (2016). *Artificial Intelligence : A Modern Approach*. Pearson Education.

S. Ali, D., Ghoneim, A. et Saleh, M. (2017). Data Clustering Method based on Mixed Similarity Measures :. Dans *Proceedings of the 6th International Conference on Operations Research and Enterprise Systems*, 192–199., Porto, Portugal. SCITEPRESS - Science and Technology Publications. http://dx.doi.org/10.5220/0006245601920199. Récupéré le 2025-01-05 de http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006245601920199

Shih, M.-Y., Jheng, J.-W. et Lai, L.-F. (2010). A Two-Step Method for Clustering Mixed Categroical and Numeric Data. *Tamkang Journal of Science and Engineering*, *13*(1), 11–19.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. CRC Press.

Singhal, A. (2001). Modern information retrieval : A brief overview. *IEEE Data Engineering Bulletin*, *24*(4), 35–43.

Singhal, A. (2012). Introducing the knowledge graph : Things, not strings. *Google Official Blog*. Récupéré de https://blog.google/products/search/introducing-knowledge-graph-things-not/

Soares, L. B., FitzGerald, N., Ling, J. et Kwiatkowski, T. (2019). Matching the Blanks : Distributional Similarity for Relation Learning. arXiv :1906.03158 [cs], http://dx.doi.org/10.48550/arXiv.1906.03158. Récupéré le 2025-01-05 de http://arxiv.org/abs/1906.03158

Suchanek, F. M., Kasneci, G. et Weikum, G. (2007). Yago : A core of semantic knowledge unifying wordnet and wikipedia. Dans *Proceedings of the 16th International Conference on World Wide Web (WWW)*, 697–706. ACM. http://dx.doi.org/10.1145/1242572.1242667

Tan, P.-N., Steinbach, M., Karpatne, A. et Kumar, V. (2019a). *Introduction to Data Mining* (2nd éd.). Pearson.

Tan, P.-N., Steinbach, M., Kumar, V. et Karpatne, A. (2019b). *Introduction to Data Mining* (2nd éd.). Pearson.

VAS (2025). Welcome to dc305. Récupéré de https://dc-help.vas.com/GetStarted/WelcomeToDC305.htm

Vishwanathan, S., Borgwardt, K. M. et Schraudolph, N. N. (2007). Fast Computation of Graph Kernels. In B. Schölkopf, J. Platt, et T. Hofmann (dir.), *Advances in Neural Information Processing Systems 19* 1449–1456. The MIT Press

webagri (2022). Des capteurs au service de l'élevage de précision. Consulté sur Web-Agri. Récupéré de https://www.web-agri.fr/dossiers-eleveur-laitier/article/829448/des-capteurs-au-service-de-l-elevage-de-precision

Wikipedia. Lactation. Récupéré de https://fr.wikipedia.org/wiki/Lactation

Wikipedia (2024). Race bovine canadienne. Récupéré de https://fr.wikipedia.org/wiki/Race_bovine_canadienne

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J. et Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. http://dx.doi.org/10.1007/s10115-007-0114-2. Récupéré le 2025-01-05 de http://link.springer.com/10.1007/s10115-007-0114-2

Yang, P., Wang, H., Yang, J., Qian, Z., Zhang, Y. et Lin, X. (2024). Deep Learning Approaches for Similarity Computation : A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *36*(12), 7893–7912. http://dx.doi.org/10.1109/TKDE.2024.3422484. Récupéré le 2024-11-22 de https://ieeexplore.ieee.org/document/10584318/

Ying, R., You, J., Morris, C., Ren, X., Hamilton, W. L. et Leskovec, J. (2019). Hierarchical Graph Representation Learning with Differentiable Pooling. arXiv :1806.08804 [cs], http://dx.doi.org/10.48550/arXiv.1806.08804. Récupéré le 2025-01-05 de http://arxiv.org/abs/1806.08804

Zhang, F., Yuan, N. J., Lian, D., Xie, X. et Ma, W.-Y. (2016). Collaborative Knowledge Base Embedding for Recommender Systems. Dans *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 353–362., San Francisco California USA. ACM. http://dx.doi.org/10.1145/2939672.2939673. Récupéré le 2025-01-05 de https://dl.acm.org/doi/10.1145/2939672.2939673

Zhao, P., Han, J. et Sun, Y. (2009). P-Rank : a comprehensive structural similarity measure over information networks. Dans *Proceedings of the 18th ACM conference on Information and knowledge management*, 553–562., Hong Kong China. ACM. http://dx.doi.org/10.1145/1645953.1646025. Récupéré le 2024-11-22 de https://dl.acm.org/doi/10.1145/1645953.1646025