Structuration de contenus visuels à travers les méthodes basées sur les voisins partagés

1. Introduction générale :

Avec la croissance de l'Internet et la baisse des prix des périphériques de stockage, la quantité de données ne cesse d'augmenter. Trouver des outils pour manipuler les grands dépôts de données numériques devient une nécessité. Qui n'est jamais revenu de vacances et se retrouve avec un grand nombre de photos prises pendant le séjour à différents endroits et avec des personnes différentes? Garder les photos dans un répertoire unique rend la navigation très longue et particulièrement quand nous sommes à la recherche de photos d'un lieu déterminé ou d'une personne spécifique. La structuration des images similaires dans un ensemble de groupes rend la navigation dans la collection très sympathique et efficace. Mais, les collections de photos personnelles n'est pas la seule application qui nécessite une structuration des données: les collections d'images scientifiques (les images de satellites, les images de plantes, des images médicales) ont aussi un grand besoin de modèles de découverte, de regroupement, de résumé et même de recommandation.

Malgré les années de recherche sur le regroupement des données, il n'existe toujours pas un regroupement qui soit la meilleure solution. Chaque méthode de classification représente certains avantages et certaines limitations et est appliquée à des problèmes et des données restreintes. Pour cette raison, il y a autant de solutions aux problèmes que de méthodes de clustering. Les utilisateurs ont plusieurs algorithmes de classification, mais ne savent pas lequel est le plus approprié. Il n'existe pas d'outil générique qui peut être appliqué à n'importe quelle application, sur n'importe quelles données, en utilisant n'importe quelle modalité. L'hétérogénéité des données d'une application à une autre est l'une des causes de ce problème. Même dans une seule application, nous pouvons trouver des informations de sources hétérogènes qui peuvent être explorées pour profiter de chacune des sources. Dans certains cas, l'hétérogénéité des données et l'utilisation de différentes modalités conduisent à l'utilisation de différentes fonctions de similarité, ce qui n'est pas très pratique.

Cependant, certains algorithmes de clustering ont besoin de paramètres adhoc pour produire des résultats pertinents qui peuvent être très difficiles à régler par l'utilisateur plus particulièrement quand il utilise des bases de données différentes. D'autres algorithmes de clustering produisent des formes spécifiques de clusters et ne peuvent pas traiter, par exemple, des groupes de densités différentes. Toutes ces contraintes font que l'utilisateur est perdu dans le choix de la méthode de classification qui soit la plus appropriée.
Le regroupement basé sur la stratégie des voisins partagés semble être prometteur et faire face

à ces problèmes. L'information des voisins partagés semble être approprié pour faire face aux différentes natures des données, des différentes fonctions de similarité, et des diverses modalités.

Cette thèse s'appuie sur cette idée. Nous proposons une nouvelle méthode théorique de clustering basée sur l'information des plus proches voisins partagés en utilisant l'approche a contrario. Grâce à la sélection du voisinage optimal de chaque élément, nous montrons à l'aide de données synthétiques comment notre méthode est robuste contre les données aberrantes (des voisins bruités).

Pour accélérer le calcul des voisins partagés, nous proposons un nouveau algorithme de factorisation basé sur la récursivité du calcul. La méthode de regroupement proposée est comparée à la classification spectrale et nous montrons par l'expérience que notre méthode est plus robuste et moins sensible à la taille du graphe.

Le prochain défi abordé dans ce travail est le prolongement de notre méthode de clustering à un cadre multi-source de voisins partagés. Dans ce cas, chaque objet n'est pas associé à une seule liste ordonnée des plus proches voisins mais un ensemble de listes provenant de différentes sources d'information.
Nous proposons une méthode générique multi-source basée sur les voisins partagés qui peut être appliquée à des sources multimédias, y compris le texte, les images, les vidéos et les documents audio. Le fait que seules les listes des plus proches voisins soient utilisées comme entrée de notre méthode de clustering rend cela possible. Alors que les méthodes SNN (« Shared Nearest Neighbours ») semblent être parfaitement adaptées à des ensembles de sources d'information hétérogènes, ce problème multi-source n'a pas été étonnamment abordé dans la littérature.

Notre contribution porte sur deux points. Tout d'abord, nous introduisons une étape de sélection des sources d'informations dans le calcul des scores des clusters candidats.
Tout élément de la base à structurer est ainsi associé à son propre sous-ensemble optimal de modalité maximisant une mesure normalisée multi-source. Comme le montrent les expériences, cette étape de sélection de source rend notre approche largement robuste à la présence des sources locales aberrantes, à savoir des sources non pertinentes qui produisent des listes bruitées des plus proches voisins.

Deuxièmement, dans ce cas, multi-source, nous proposons une étape de remodelage des groupes lors de la construction des clusters candidats. Les éléments manquants ne sont pas récupérés seulement lors de l'élimination des clusters redondants, comme nous le faisons dans le cas mono-source, mais aussi à partir des autres K listes disponibles des plus proches voisins appartenant à l'ensemble optimal des sources sélectionnées pour chaque groupe.

Grâce aux données synthétiques, nous démontrons l'efficacité et la robustesse de notre méthode face aux sources d'information bruitées. Notre méthode multi-source basée sur l'information des voisins partagés est appliquée au regroupement multimodal des résultats de recherche, l'exploitation d'objets visuels et le regroupement d'images basé sur plusieurs sous-espaces aléatoires.

Enfin, nous étudions le cas où les plus proches voisins d'un élément appartiennent à un autre ensemble. Dans ce cas bipartite, la similitude des deux éléments est évaluée par leurs plus proches voisins partagés appartenant à un ensemble disjoint. Nous proposons une nouvelle méthode de classification bipartite basée sur les plus proches voisins partagés et nous l'appliquons à la suggestion d'objets visuels. Nous cherchons à résoudre les problèmes de perception des utilisateurs en appliquant notre méthode de clustering bipartite.

Nous abordons également le problème de suggestion d'objets visuels répondant à des requêtes utilisateurs. Les expériences montrent que cette nouvelle méthode dépasse les résultats des méthodes de l'état de l'art.

2. Contribution 1: Revisiter la méthode de clustering basée sur les plus proches voisins partagés

Le regroupement non supervisé de données reste une étape cruciale dans la plupart des approches de recherche multimédia par exemple la fouille d'événements ou la suggestion de requêtes visuelles.
Cependant, la performance et l'applicabilité d'un grand nombre d'approches de clustering classiques obligent souvent des choix particuliers de représentation des données et des mesures de similitude.
Certaines méthodes, telles que k-means et ses variantes nécessitent l'utilisation de paramètres Lp ou d'autres mesures spécifiques de similitude de données, d'autres, comme BIRCH et CURE rajoute un calcul prohibitif qui augmente quand la dimension de la représentation des données est élevée, en raison de leur dépendance à l'égard des structures de données. Ces hypothèses sont particulièrement problématiques dans un contexte multimédia qui implique généralement des données et des mesures de similarité hétérogènes.
Une approche alternative intéressante au clustering qui nécessite des tests comparatifs seulement des valeurs de similarité est l'utilisation de ce que nous appelons l'information des voisins partagés.
Ici, nous vous présentons notre première contribution au paradigme de regroupement: nous introduisons un nouveau formalisme SNN basé sur la théorie de la décision a contrario. Cela nous permet de tirer des scores de connectivité plus fiable des clusters candidats et une interprétation plus intuitive des voisinages locaux optimaux.

Principes de base, notations et définitions :

Le principe de base des méthodes de clustering SNN est de considérer qu'un cluster parfait est composé d'éléments qui ont tous leurs voisins dans le même cluster. Deux éléments sont considérés comme similaires, non par rapport à leur valeur de similarité, mais plutôt par le degré de similarité de leurs voisins respectifs. Cela signifie que si deux éléments ont une forte proportion de voisins en commun, il est raisonnable de les mettre dans le même groupe. L'avantage est qu'aucune hypothèse n'est faite sur la forme, la densité ou la métrique utilisée.

Soit un ensemble X de N éléments d'un certain domaine D. Le descripteur de ces éléments peut être de n'importe quel type. Nous supposons l'existence d'une fonction Fk qui associe à tout élément x de X son K plus proches voisins selon une certaine mesure de similarité ou de distance. Notez que cette indication peut être totalement inconnue. Seul le classement qu'elle produit est connu.

Notons que nnk (x) $\in$ X retourne le k-ième plus proche voisin de x $\in$ X par rapport à une certaine métrique ou distance. La fonction Fk est définie comme suit:

Définition 1. La fonction Fk des K plus proches voisins:

$$Fk\ (x) = \{nnk\ (x) \mid 0 < k \leq K\}$$

Pour tout i < j, l'élément nni (x) est plus pertinent ou similaire à x que nnj (x).
Nous rappelons que la fonction Fk joue le rôle d'un oracle ou d'une source d'information qui renvoie une liste ordonnée des éléments pertinents à chaque requête. Comme aucune hypothèse n'est faite sur la nature des objets ou la mesure de similarité utilisée pour comparer des éléments entre eux, nous définissons la similarité entre une paire de points x1 et x2 en fonction du nombre des plus proches voisins partagés.

$$SIM\ (x1, x2) = |\ Fk\ (x1) \cap Fk\ (x2)\ |$$

Mais si la métrique initiale et la similitude des voisins partagés sont toutes les deux utilisées pour évaluer la similitude, elles ne partagent pas les mêmes propriétés fondamentales. La métrique (fonction de la distance parfois appelée tout simplement distance) satisfait quatre propriétés fondamentales:
-La non-négativité: d (x, y) $\geq$ 0
- L'identité: d (x, y) = 0 si et seulement si x = y
- La symétrie: d (x, y) = d (y, x)
- L'inégalité triangulaire: d (x, z) = d (x, y) + d (y, z)

La similitude des voisins partagés est non négative et symétrique, mais elle ne nécessite pas de satisfaire l'inégalité du triangle. En outre, le fait que deux éléments x1 et x2 ont des éléments en commun, ne signifie pas nécessairement que x1 appartient à la K-NN
x2 et vice-versa.

Nous proposons d'utiliser le principe a contrario. Une telle normalisation a contrario a déjà été proposé pour le clustering, mais pas pour SNN clustering. L'approche a contrario est une formalisation mathématique d'un principe de regroupement perceptuel. Plus un regroupement d'éléments est très peu probable, plus l'apparition d'une telle disposition est importante et plus les éléments devraient être regroupés en un seul groupe. Les clusters sont détectés a contrario à une hypothèse nulle ou modèle de fond. La signifiance d'un groupe d'éléments est mesurée par le nombre de fausses alarmes (Nfa) qui ont été produites sous l'hypothèse nulle. Plus le Nfa est moins élevé et plus le groupe est considéré pertinent. Cette mesure est utilisée par exemple pour classer les groupes et pour décider si un cluster est un groupe naturel dans lequel les valeurs aberrantes ont été éliminées.

L’algorithme de regroupement proposé :

Après avoir défini nos nouvelles mesures décrivant la qualité de chaque cluster sur la base de l’approche a contrario, nous pouvons décrire notre regroupement. Le but est de trouver les groupes optimaux qui maximisent les scores a contrario.
Ce problème combinatoire ne peut pas être résolu exactement. En pratique, nous utilisons une heuristique pour réduire le nombre de solution en considérant d'abord que chaque élément est le centre d’un cluster candidat.
Notre algorithme de classification est basé sur deux étapes principales, la construction de cluster candidat et la sélection finale des clusters.
Notre proposition de regroupement basé sur les voisins partagés est la suivante:

- **La construction de cluster candidat:** Chaque élément $x \in X$ est considéré comme un centre du cluster candidat et un voisinage optimal doit être calculé pour celui-ci en faisant varier la taille k de 1 à K et en sélectionnant la taille du voisinage qui maximise le score contrario. Nous obtenons N clusters candidats de différentes qualités.

- **La sélection finale des clusters:** Après la sélection des groupes candidats, nous obtenons des groupes susceptibles d'être pertinents pour chaque élément, mais bon nombre de ces groupes sont encore très semblables parce que les éléments proches pourraient générer environ le même cluster candidat.
  Par conséquent, les clusters redondants doivent être éliminés et seulement les clusters différents doivent être sélectionnés.
  Pour cette étape, nous utilisons une heuristique simple basée sur le chevauchement entre les clusters candidats. Un paramètre de chevauchement est généralement utilisé pour cette étape. Pour cela, tout d'abord, nous trions tous les clusters candidats dans l'ordre décroissant de leurs scores a contrario et ensuite nous itérons entre eux.
  Si un cluster rencontré a moins de chevauchement avec tous les groupes précédemment retenus, il est ajouté à la liste finale des clusters. Sinon, le groupe rencontré est considéré comme similaire au groupe retenu et doit être utilisé pour le remodeler.
  Pour cela, la contribution de tous les éléments des deux groupes est calculée et triée dans l'ordre décroissant.
  Finalement, le groupe conservé sera construit à partir des éléments qui améliorent la qualité du cluster original. Cette étape de remodelage est utile parce que seuls les éléments pertinents du cluster original sont conservés et remplacent les éléments qui sont peu associés dans le but de donner un nouveau cluster de meilleure qualité.
  Dans l'ensemble, on remarque ici que notre méthode de regroupement ne nécessite que le paramètre de chevauchement pour décider si deux groupes sont similaires ou non. Il peut être choisi de façon naturelle, sans aucune connaissance de la nature de l'ensemble de données ou de sa distribution.

Conclusion:
Dans cette partie, nous avons revisité une méthode basée sur les voisins partagés sur deux points. Nous avons introduit un nouveau formalisme SNN basé sur la théorie de la décision a contrario. Cela nous a permis de calculer des scores de connectivité plus fiables des groupes candidats et une interprétation plus intuitive des voisinages locaux optimaux. Nous avons

également proposé un nouveau algorithme de factorisation pour accélérer le calcul intensif nécessaire pour déterminer la matrice des voisins partagés. Nous avons comparé notre méthode de regroupement à la classification spectrale. Nous avons montré que notre méthode est globalement plus robuste face aux bruits (voisins non pertinents).
Cette partie est la base pour la deuxième et la troisième contribution de cette thèse.

3. Contribution 2: Regroupement multi-source basé sur les plus proches voisins partagés

La disponibilité croissante d'appareils (ordinateurs portables, téléphones intelligents, appareils photo) a conduit à une explosion de la quantité d'information que l'utilisateur doit faire face afin de les utiliser efficacement.
Nous utilisons souvent des données sans surveillance provenant de sources différentes. Par exemple, les images ont de nombreuses propriétés (couleur, texture, etc) et les métadonnées (annotations textuelles, EXIF, etc) qui sont très différentes d'une source d'information à une autre. L'objectif est de profiter de toutes les sources d'information disponibles afin d'être en mesure d'avoir des informations plus significatives.
Pour exploiter cela, le traitement multimodal est devenu une stratégie attrayante.
Il est connu que le traitement des données provenant d'une source d'information unique non pertinente contribuera à la création d'un mauvais résultat. Est-ce que l'utilisation de multiples sources de données multiples a plus de chance d'améliorer le résultat? Comment pouvons-nous traiter des sources d'information lorsque certaines d'entre elles sont caractérisées d'incertaines? Nous avons besoin d'une combinaison efficace des sources et pas seulement une fusion triviale. Une combinaison efficace est cruciale.
Pour faire face à l'incertitude de certaines sources de données disponibles, nous proposons un regroupement qui n'est pas appliqué à toutes les sources disponibles mais au sous-ensemble optimal de sources. Pour chaque groupe, nous avons besoin de sélectionner les sources d'information pertinentes et ignorer celles qui sont bruitées. De cette façon, le bruit peut être réduit en combinant des sources de manière efficace et nous pouvons alors surmonter la mauvaise qualité des sources en corrigeant les erreurs produites par chaque source individuelle.
Le second problème est de savoir comment regrouper des données décrites dans des espaces physiques différents avec des dimensionnalités différentes, comme par exemple, le regroupement de données visuelles avec des données textuelles?
Pour ce faire, le regroupement basé sur les voisins partagés semble approprié car même dans des contextes hétérogènes dans lesquels les caractéristiques sous-jacentes et les valeurs de similarité n'ont pas une simple interprétation unique, deux éléments ayant une forte proportion de voisins en commun doivent être attribués au même groupe. Deux éléments sont considérés comme fortement associés non pas en raison de leur valeur de similarité, mais par le nombre de voisins en commun selon leurs sources respectives. Les méthodes de regroupement basées sur les plus proches voisins partagés (SNN) semblent donc être parfaitement adaptées au contexte multimodal.
Ces méthode SNN, comme indiqué précédemment, sont en mesure de combler les lacunes des approches de regroupement classiques: elles ne souffrent pas de la malédiction de la dimensionnalité, elles sont robustes aux données bruitées, elles ne nécessitent pas de fixer à l'avance le nombre de clusters et finalement elles ne nécessitent aucune connaissance

explicite de la nature ou de la représentation des données. Ces propriétés les rendent largement génériques à des fins de fouille de données ou de structuration multimédia, quelle que soit la nature des données ou les mesures de similarité utilisées.
La principale originalité de notre approche est que nous introduisons une étape de sélection de source d'information dans le calcul des mesures décrivant la qualité des clusters grâce à une standardisation a contrario de la somme des scores individuels SNN. En plus, chaque cluster est associé à son propre sous-ensemble optimal de sources qui maximise le score a contrario multi-source. Par conséquence, tous les clusters résultants n'ont pas nécessairement les mêmes sources d'information, contrairement aux autres travaux antérieurs dans la littérature.

Conclusion:

Les techniques de regroupement basées sur les plus proches voisins partagés (SNN) sont bien connues pour surmonter plusieurs insuffisances des approches de clustering traditionnelles, notamment la forte dimensionnalité et les limitations des métriques. Cependant, les méthodes SNN précédentes étaient limitées à une source d'information unique alors qu'elles semblent être très bien adaptées pour des données hétérogènes, généralement dans des contextes multimodaux.
Dans cette partie, nous avons introduit une nouvelle méthode de regroupement multi-source basée sur l'information des voisins partagés. Nous avons d'abord étendu la méthode existante SNN mono-source au cas multi-source et nous avons introduit une étape initiale de sélection automatique de sources lors de la construction des clusters candidats. Le point clé est que chaque groupe résultant est construit avec son propre sous-ensemble optimal de modalités qui améliore la robustesse aux sources d'information bruyantes ou aberrantes. Nous avons expérimenté notre méthode avec les données synthétiques impliquant différentes sources d'information. Nous avons démontré l'efficacité et la robustesse de notre méthode face à la présence de sources bruyantes.

4. Contribution 3: Regroupement bipartite basé sur les voisins partagés

Dans cette partie, nous nous sommes intéressés au graphe bipartite dans lequel les nœuds ne sont pas dans un seul ensemble, mais sont répartis en deux ensembles disjoints. Les k plus proches voisins d'un ensemble A d'éléments à structurer appartiennent à un deuxième ensemble disjoint B. De plus, les éléments d'un même ensemble ne sont pas reliés entre eux, seules les paires de nœuds appartenant à des ensembles différents sont connectés. La relation entre les nœuds d'un même ensemble est exprimée par le nombre de nœuds en commun dans le second ensemble.
Les graphes bipartis sont utilisés par exemple pour saisir la relation entre les utilisateurs et leurs intérêts, les utilisateurs et leurs requêtes, une page web et ses annonces et une photo et ses tags.
Dans cette thèse, nous nous intéressons à la structuration de tels graphes. Nous étudions le regroupement bipartite basé sur l'information des plus proches voisins partagés (SNN).
Les méthodes de classification SNN n'ont pas encore été étudiées dans le cas des graphes bipartites auparavant. De tels graphes sont naturels pour de nombreuses applications telles que les documents et les mots. Un mot appartient à un ensemble de documents et dans le même

temps, un document contient un ensemble de mots. La motivation peut être de regrouper les documents ayant des mots communs.

Conclusion
Dans cette partie, nous avons étendu les méthodes SNN dans le contexte des graphes bipartites c.-à-d. les voisins de chaque élément d'un cluster se trouvent dans un ensemble disjoint. Nous avons introduit une nouvelle mesure de pertinence SNN dédiée pour ce contexte asymétrique et nous avons montré comment elle peut être utilisée pour sélectionner localement des groupes optimaux bipartites.
En utilisant des données synthétiques, nous avons comparé notre méthode à un algorithme de classification spectrale bipartite. Nous avons démontré que notre méthode est plus robuste face l'instabilité des K-NN. Cette contribution est encore prospective et nous pensons pouvoir encore l'améliorer par des algorithmes d'optimisation du temps de calcul.

5. Conclusions générales

Cette thèse traite le problème de structuration de contenu en utilisant le regroupement basé sur l'information des plus proches voisins. Notre motivation était que le regroupement classique des données oblige souvent de faire des choix particuliers de représentation des données et des mesures de similarité. De telles hypothèses sont particulièrement problématiques dans le contexte du multimédia qui implique généralement des données hétérogènes et des mesures de similarité diverses.
Pour cette raison, nous étudions de nouveaux paradigmes de regroupement basés sur le principe des plus proches voisins partagés (SNN) qui sont aptes à surmonter la complexité et l'hétérogénéité des données et la haute dimensionnalité.
Tout d'abord, nous avons proposé de revoir l'état de l'art des méthodes existantes basées sur l'information des voisins partagés. Nous avons présenté un nouveau formalisme SNN basé sur la théorie de la décision a contrario.
Cela nous a permis de proposer un score de connectivité plus fiable qui est utilisé pour sélectionner des voisinages optimaux. L'avantage d'utiliser cette mesure a contrario est qu'elle n'est pas biaisée par rapport aux tailles des clusters et qu'elle est interprétable.
L'idée est de sélectionner le meilleur voisinage pour chaque élément, puis d'éliminer les clusters candidats redondants en utilisant une stratégie qui privilégie les meilleurs d'entre eux et qui permet d'ajouter des éléments pertinents aux clusters finaux.
Nous avons également proposé un nouveau algorithme de factorisation pour permettre une accélération du calcul intensif des matrices des voisins partagés.
En utilisant des données synthétiques, nous avons démontré que notre méthode est capable de sélectionner le meilleur voisinage en présence de voisins bruités.
Par rapport à la classification spectrale de Ng, Jordan et Weiss, nous avons montré que notre méthode est plus robuste et moins sensible à la taille du graphe d'entrée.
La deuxième contribution de cette thèse est une généralisation de la proposition SNN au cas multi-source. La principale originalité de notre approche est que nous introduisons une étape de sélection des sources d'information pour le calcul des scores des clusters candidats.
Tout groupe arbitraire est ainsi associé à son propre sous-ensemble optimal de modalités qui maximise une mesure normalisée multi-source. Comme le montre les expériences sur les données synthétique, cette étape de sélection de sources rend notre approche non seulement largement robuste à la présence de sources locales aberrantes, mais améliore également la qualité des clusters.

Nous avons conclu que la combinaison efficace de sources peut compenser la faible qualité des sources indépendantes très bruyantes.
Nous avons appliqué notre méthode de regroupement multi-source basée sur les voisins partagés à la structuration du contenu visuel dans différentes applications.
Dans une expérience sur les feuilles, nous avons voulu aider les botanistes à identifier et formaliser des catégories morphologiques utiles qui relient des espèces entre elles en utilisant notre regroupement multi-source.
L'objectif de notre deuxième application multi-source est de structurer le contenu d'un résultat de recherche. En considérant l'information visuelle et textuelle des images, nous avons obtenu des clusters qui étaient sémantiquement et visuellement cohérents et d'autres clusters qui étaient soit que visuellement cohérents, soit que sémantiquement pertinents. Une telle réorganisation est très efficace pour explorer les résultats d'une recherche et très significative pour les utilisateurs qui ont besoin de connaitre l'information pourquoi des images ont été regroupées ensemble.
Enfin, nous avons proposé d'étendre le regroupement SNN dans le contexte des graphes bipartites ç-à-dire lorsque les voisins de chaque élément se trouvent dans un ensemble disjoint. Pour cela, nous avons introduit une nouvelle mesure de pertinence SNN que nous avons revisité pour ce contexte asymétrique et nous l'avons utilisé pour sélectionner localement des clusters bipartites optimaux.
En utilisant des données synthétiques, nous avons démontré que notre regroupement SNN bipartite était plus pertinent que la classification spectrale et ceci grâce à la sélection des éléments pertinents lors de la création des clusters candidats.
En se basant sur la découverte d'objets, nous avons introduit un nouveau paradigme de recherche visuelle, c'est-à-dire basé sur la suggestion de requêtes visuelles. L'idée est de proposer à l'utilisateur des requêtes visuelles qui représentent les objets les plus fréquents dans la base de données. Plutôt que de laisser l'utilisateur choisir n'importe quelle région d'intérêt dans une image, le système vous proposera seulement les requêtes visuelles (régions de l'image) qui sont réellement pertinentes.
Tout au long de toutes ces différentes applications dans cette thèse, nous avons montré comment l'information des voisins partagés est susceptible d'être utilisée sur des représentations différentes de données et des modalités diverses. Nous avons également montré comment nos méthodes proposées peuvent être appliquées même en présence de sources fortement bruitées.

6. Perspectives

Comme les méthodes de regroupement basées sur l'information des voisins partagés proposées dans cette thèse ont démontré leur potentiel pour sélectionner un voisinage optimal pour chaque élément, nous proposons d'utiliser ces voisinages optimaux pour construire le graphe d'entrée pour la classification spectrale. Cette dernière est très sensible à la qualité des graphes d'entrée.
En utilisant ces voisinages optimaux, nous produisons un graphe facile à regrouper et nous améliorons la robustesse de la méthode spectrale face aux voisinages bruités. De tels voisinages optimaux pourraient être utilisés dans de nombreuses applications telles que les systèmes de recommandation.
Nous allons aussi approfondir notre travail pour qu'il soit capable de traiter de grandes bases de données ce qui est la principale limite de notre méthode SNN proposée.

Pour cela, nous prévoyons de tester deux idées: la première consiste à utiliser une approche hiérarchique qui traite ce problème et le second est de proposer un hachage sensible aux voisins partagés. L'idée principale de cette seconde proposition est d'utiliser une fonction de hachage tel que les points qui ont des voisins en commun dans l'espace d'origine ont une forte probabilité d'avoir la même valeur de hachage.
Enfin, nous voulons étendre notre méthode à la classification supervisée SNN. Utiliser la similitude SNN pourrait offrir de meilleures performances en particulier dans le contexte multi-source.