République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Bordj Bou Arréridj
Faculté des Mathématiques et d'Informatique
Département d'informatique



MEMOIRE
Présenté en vue de l'obtention du diplôme
Master en informatique
Spécialités : Technologies de l'information et de la communication

THEME

Extraction des règles d'association à haute utilité basée sur les représentations concises appliquée aux données géolocalisées de Gowalla

Présenté par :
Aya Saighi Bouaouina

Soutenu publiquement le : 23/06/2025
Devant le jury composé de :
President : Moussaoui boubeuker
Examiner : nouioua mourad,benaouda nadjib
Encadreur : Attia abdelouahab
Co-Encadreur : Bensaadi Mohamed

2024/2025

Dédicace

Je dédie ce modeste travail à ceux qui illuminent ma vie de leur amour et de leur présence.

À ma chère mère **Salima**, source de tendresse et de prières silencieuses,
dont les sacrifices et l'amour inconditionnel m'ont portée jusque-là.

À mon père **Farid**, dont les mots simples et la force tranquille
ont toujours été un refuge et une inspiration.

À ma sœur **Doaa**, douce présence et complice de tous les instants.

À mes frères **Mohib Eddine** et **Salah Eddine**,
piliers solides de ma vie, et à leur manière unique de me montrer leur soutien.

À **Cherih Souheib** et **Mehiris Hichem**.

Et à moi-même, **Aya**,
pour avoir tenu bon, pour ne pas avoir abandonné,
pour avoir osé rêver et transformé ces rêves en pas concrets.

À tous ceux qui m'ont soutenue de près ou de loin,
que ces quelques lignes soient un témoignage de ma reconnaissance infinie.

Remerciements

Avant toute chose, toute louange revient à Allah, le Très-Haut, le Sage, le Miséricordieux.
Sans Sa grâce, rien n'aurait été possible.

Je tiens aussi à exprimer ma gratitude envers moi-même, moi, **Aya**,
pour avoir tenu bon malgré les tempêtes, pour n'avoir jamais abandonné,
et pour avoir cru même lorsque tout semblait incertain.

Mes remerciements les plus sincères vont à mes deux encadrants :
Monsieur **Attia Abdelouahab** et Monsieur **Bensaadi Mohamed**,
pour leur accompagnement bienveillant, leurs conseils avisés,
et leur soutien constant durant tout ce travail.

Je remercie profondément Monsieur **Benaouda Nadjib**,
pour sa disponibilité, sa patience, et sa aide précieuse tout au long de l'année.
Son accompagnement m'a été d'un immense soutien.

Et enfin, à tous mes amis et camarades,
qui m'ont offert un mot d'encouragement, un geste d'amitié,
ou une simple présence rassurante,
je vous dis du fond du cœur : **merci**.

Abstract

Our research focuses on the extraction of high-utility association rules from data originating from Location-Based Social Networks (*LBSNs*), a task that poses significant challenges due to the complexity and scale of the data. In this study, we concentrate on the **Gowalla** dataset, a social network centered around users' check-in locations.

Unlike traditional approaches that rely solely on frequency, our methodology is based on the utility value of itemsets, considering their economic or contextual importance. We propose an efficient approach based on concise representations, specifically *HUCI* (High Utility Closed Itemsets), to reduce redundancy and the volume of generated rules.

The **HUCI-Miner** algorithm is employed to extract high-utility closed itemsets, followed by the **HGB** algorithm to generate complete and non-redundant high-utility association rules. These rules are evaluated using metrics such as : **UTIL** (total utility), **AUTIL** (antecedent utility), and **UCONF** (utility-based confidence).

The experiments conducted on the **Gowalla** dataset demonstrate that this approach enables the extraction of concise, meaningful, and knowledge-rich rules, making it suitable for applications such as location recommendation and user behavior analysis in urban environments.

Résumé

Notre travail porte sur l'extraction des règles d'association à haute utilité à partir de données issues des réseaux sociaux géolocalisés (*LBSNs*), une tâche qui représente un véritable défi en raison de la complexité et du volume élevé des données. Dans ce contexte, nous nous intéressons à la base de données **Gowalla**, un réseau social centré sur les lieux visités par les utilisateurs.

Contrairement aux approches traditionnelles basées uniquement sur la fréquence, notre méthodologie repose sur la notion d'utilité des itemsets, reflétant leur importance économique ou contextuelle. Nous proposons une approche efficace s'appuyant sur les représentations concises, notamment les *HUCI* (High Utility Closed Itemsets), dans le but de réduire la redondance et le nombre de règles générées.

L'algorithme **HUCI-Miner** est utilisé pour identifier les itemsets fermés à haute utilité, suivi de l'algorithme **HGB** pour générer des règles d'association complètes, pertinentes et non redondantes. Ces règles sont évaluées à l'aide de mesures telles que : **UTIL** (utilité totale), **AUTIL** (utilité de l'antécédent), et **UCONF** (confiance utilitaire).

Les expériences menées sur la base **Gowalla** ont montré que cette approche permet d'extraire des règles concises, informatives et à forte valeur ajoutée, adaptées à des applications comme la recommandation de lieux et l'analyse du comportement des utilisateurs.

ملخص

يتناول بحثُنا موضوع استخراج قواعد الارتباط عالية المنفعة انطلاقًا من بيانات شبكات التواصل الاجتماعي المعتمدة على الموقع الجغرافي (LBSNs)، وهي مهمة تُعدّ تحديًا كبيرًا نظرًا لتعقيد هذه البيانات وتعدد أبعادها. في هذا السياق، نركز على تحليل بيانات شبكة Gowalla، وهي منصة اجتماعية تتيح للمستخدمين مشاركة أماكن تواجدهم وزياراتهم للأماكن الجغرافية.

على عكس الأساليب التقليدية التي تعتمد على التكرار فقط، تعتمد هذه المنهجية على قيمة المنفعة للمجموعات، مثل الأهمية الاقتصادية أو السياقية. نقترح مقاربة فعالة تستند إلى التمثيلات المختصرة للأنماط ذات المنفعة العالية، وبالأخص المجموعات المغلقة ذات المنفعة العالية (HUCI)، وذلك بهدف تقليص عدد الأنماط المكررة دون فقدان المعلومات الجوهرية.

يتم استخدام خوارزمية HUCI-Miner لاستخراج هذه المجموعات، تليها خوارزمية HGB لتوليد قواعد ارتباط عالية المنفعة تكون كاملة وغير مكررة. ويتم تقييم القواعد المستخرجة باستخدام مقاييس متقدمة مثل: UTIL (المنفعة الإجمالية)، AUTIL (منفعة المقدمة)، وUCONF (الثقة النفعية).

وقد أظهرت التجارب المجراة على قاعدة بيانات Gowalla أن هذا النهج يُمكّن من استخراج قواعد مختصرة، دالة، وذات قيمة معرفية عالية، مما يجعلها ملائمة لتطبيقات مثل التوصية بالأماكن وتحليل سلوك المستخدمين في البيئات الحضرية.

Table des matières

Dédicace ii

Remerciements iii

Abstract iv

Résumé v

Liste des acronymes x

Introduction Générale 1

1 Aperçu Général sur le Data Mining 3

1.1 Introduction . . . 3

1.2 Le processus de découverte de connaissances dans les bases de données . . . 4

1.2.1 Définition . . . 4

1.2.2 Objectifs . . . 4

1.2.3 Étapes du processus KDD . . . 4

1.2.4 Rôle central du Data Mining dans le processus KDD . . . 5

1.3 Fouille de données : Techniques et Domaines d'application . . . 5

1.3.1 Définition . . . 6

1.3.2 L'importance stratégique de la fouille de données . . . 6

1.3.3 Les principales tâches du Data Mining . . . 7

1.4 Conclusion . . . 8

2 Les Réseaux Sociaux Géolocalisés (LBSNs) 9

2.1 Introduction . . . 9

2.2 Intégration des dimensions spatiale et sociale : Étude des LBSNs . . . 10

2.3 Les LBSNs : définition et applications contemporaines 10
2.3.1 Exemples de réseaux LBSNs : Gowalla, Foursquare et Brightkite 11
2.4 Analyse des réseaux sociaux spatiaux : Du numérique au data mining 11
2.5 Caractéristiques et exploitations analytiques des LBSNs 11
2.6 Approches analytiques autour des LBSNs 12
2.6.1 Vers une compréhension comportementale via les LBSNs 13
2.6.2 Fondements théoriques et potentialités analytiques 13
2.6.3 LBSNs et fouille spatio-sociale : Défis méthodologiques et perspectives 13
2.7 Intégration des LBSNs dans le processus de fouille à haute utilité 14
2.8 Conclusion 14

3 High Utility Itemset Mining HUM **15**
3.1 Introduction 15
3.2 Frequent Itemset Mining(FIM) 16
3.3 Notions de base du FIM 16
3.3.1 Bases de données transactionnelles 16
3.3.2 Item (Élément) 16
3.3.3 Itemset (Ensemble d'items) 17
3.3.4 Frequent Itemset (Ensemble d'items fréquent) 17
3.3.5 Mesure de support (Support) 17
3.3.6 Défis de l'extraction des ensembles d'items fréquents 17
3.3.7 Algorithmes d'extraction des items fréquents 18
3.3.8 Stratégies de recherche dans l'extraction des itemsets fréquents 18
3.3.9 Représentation de la base de données 19
3.3.10 Génération des ensembles d'items fréquents 19
3.3.11 Limitations de l'extraction des ensembles fréquents 20
3.4 High Utility Itemset Mining HUIM 21
3.4.1 Extraction des ensembles d'items à haute utilité (HUIM) 21
3.4.2 Notions de base du HUIM 21
3.4.3 Algorithmes pour l'extraction des ensembles d'items à haute utilité 23
3.4.4 Limitations du High Utility Itemset Mining (HUIM) 23
3.5 Représentations concises des ensembles d'items à haute utilité 24
3.6 Définition des règles d'association 26
3.7 Application concrète des méthodes HUIM sur les données de Gowalla 28
3.8 Conclusion 29

4 Réalisation & Interprétation **30**

4.1 Introduction 30

4.2 Présentation des outils de développement 31

4.2.1 Eclipse 31

4.2.2 Langages de programmation : Java et Python 31

4.2.3 Python 31

4.2.4 Visual Studio 31

4.2.5 SPMF 31

4.3 L'architecture générale de la méthodologie 32

4.4 Base de Données Gowalla 34

4.4.1 Présentation de la Base de Données Gowalla 34

4.4.2 Pré-Traitement de la Base 35

4.5 Les algorithmes utilisés 37

4.5.1 l'algorithme HUCI-Miner 37

4.5.2 l'algorithme HGB 38

4.6 Exécution expérimentale des algorithmes sur la base Gowalla 40

4.6.1 Exécution de l'algorithme HUCI-Miner sur les données Gowalla 40

4.6.2 Exécution de l'algorithme HGB après HUCI-Miner 40

4.6.3 Résultats obtenus 41

4.7 Conclusion 44

Conclusion Générale **46**

References **46**

Liste des acronymes

— **KDD** : Knowledge Discovery in Databases
— **LBSNs** : Location-Based Social Networks
— **HUCI-Miner** : High Utility Closed Itemsets Miner
— **UP-Growth** : Utility Pattern Growth
— **EFIM** : Efficient High-Utility Itemset Mining
— **HUM** : High Utility Mining
— **FIM** : Frequent Itemset Mining
— **TID** : Transaction IDentifier
— **TWU** : Transaction-Weighted Utility
— **CHUIs** : Closed High Utility Itemsets
— **HGB** : High-utility Generator-Based
— **SPMF** : Sequential Pattern Mining Framework
— **CSV** : Comma-Separated Values
— **HDFS** : Hadoop Distributed File System

Table des figures

1.1 Processus de découverte de connaissances à partir de données.. 5

2.1 Modèle de réseau social géolocalisé à deux couches (utilisateurs – lieux) 10
2.2 Architecture générale d'un réseau social basé sur la localisation (LBSN) 11
2.3 Modèle géosocial multilayer : utilisateurs, contenu et emplacement 12

4.1 Présentation de l'environnement de SPMF 32
4.2 Interface graphique de la bibliothèque SPMF 32
4.3 L'architecture générale de la méthodologie 34
4.4 Un extrait du fichier SPMF généré à partir de la base de données Gowalla 36
4.5 Exécution d'algorithme UCI-Miner sur la base GOwalla 40
4.6 Exécution d'algorithme HGB avec la base GOwalla 41
4.7 Arber bipartite des 5 régles avec confiance 43
4.8 Visualisations en graphe orienté des règles d'association à haute utilité 43
4.9 Top régles d'association à haute utilité 43

Liste des tableaux

1.1 Domaines d'application du Data Mining 8

3.1 Domaines d'application du Data Mining 16
3.2 Algorithmes d'extraction des ensembles d'items fréquents 18
3.3 Résumé des algorithmes pour l'extraction des HUIs 23
3.4 Résumé des algorithmes pour l'extraction des représentations concises des HUIs 26

4.1 Caractéristiques essentielles de la base de données Gowalla 34
4.2 Exemples représentatifs de résultats générés par l'algorithme HGB à partir des motifs HUCI-Miner 42

List of Algorithms

1 HUCI-Miner($\mathscr{D}$, min_util) 38
2 HGB Construction Algorithm 39
3 HGB Construction Algorithm 39

Introduction Générale

À mesure que les technologies numériques investissent les différents domaines de l'activité humaine, les volumes de données générés quotidiennement deviennent colossaux. Ces flux informationnels, bien qu'ils constituent une richesse potentielle, ne révèlent leur véritable valeur que lorsqu'ils sont analysés, structurés et transformés en connaissances exploitables.

La fouille de données (Data Mining) constitue un levier essentiel dans cette dynamique. Elle regroupe un ensemble de techniques destinées à extraire automatiquement des modèles pertinents à partir de masses de données complexes. Parmi les approches phares de ce domaine figure l'extraction des règles d'association, qui permet de découvrir des relations significatives entre des éléments récurrents au sein de bases transactionnelles.

Cependant, les méthodes traditionnelles d'extraction de règles reposent principalement sur des mesures statistiques telles que le support et la confiance. Ces critères, bien qu'efficaces dans certains contextes, ne tiennent pas compte de l'utilité économique ou stratégique des items impliqués. Ainsi, des motifs fréquemment observés mais faiblement rentables peuvent être privilégiés au détriment d'ensembles moins fréquents mais à forte valeur.

Pour remédier à cette limitation, l'approche dite du *High Utility Itemset Mining (HUIM)* a été développée. Elle vise à identifier les ensembles d'items dont la contribution globale en termes d'utilité dépasse un seuil prédéfini. Cette orientation permet une lecture plus fine des données, en lien avec les objectifs économiques réels. Néanmoins, elle entraîne une complexité computationnelle accrue, notamment en raison du nombre élevé de motifs candidats générés.

Dans ce contexte, les recherches récentes se sont tournées vers des stratégies de réduction, notamment à travers l'adoption de représentations concises telles que les itemsets fermés, maximaux ou générateurs. Ces représentations permettent de synthétiser l'information sans perte de pertinence, tout en allégeant le coût computationnel.

Parallèlement à ces avancées méthodologiques, l'émergence des *réseaux sociaux géolocalisés (LBSNs)* constitue une source précieuse de données mixtes, combinant dimensions sociale et spatiale. Ces réseaux, tels que *Gowalla* ou *Foursquare*, génèrent des données transactionnelles implicites exploitables dans les systèmes de fouille à haute utilité.

Le présent travail s'inscrit dans cette perspective et vise à extraire des règles d'association à haute utilité en s'appuyant sur des algorithmes exploitant ces représentations concises, tout en intégrant une étude appliquée à des données issues de LBSNs. L'étude porte sur plusieurs aspects complémentaires :

— L'analyse du comportement et de la performance de certaines approches algorithmiques récentes en matière d'extraction d'itemsets et de règles à haute utilité ;
— L'étude de l'impact de différents paramètres de contrôle (tels que le seuil de l'utilité minimale ou celui de la confiance minimale) sur la quantité, la qualité et la compacité des règles générées ;
— L'exploration d'un réseau social géolocalisé réel (*Gowalla*) et sa transformation en base transactionnelle ;
— La mise en œuvre expérimentale de ces algorithmes via des outils reconnus, notamment la bibliothèque *SPMF*, afin de comparer objectivement leurs performances.

Afin de répondre à ces objectifs, le mémoire est structuré selon quatre volets complémentaires :

— Une première partie introductive sur les bases de la fouille de données et des règles d'association à haute utilité ;
— Une seconde partie consacrée aux réseaux sociaux géolocalisés et à leur intégration analytique dans les processus de fouille ;
— Une troisième partie technique dédiée à la description détaillée des algorithmes retenus, ainsi qu'aux modèles de représentation concises ;
— Une quatrième partie expérimentale, consacrée à la mise en œuvre, l'analyse et la comparaison des résultats obtenus à travers des cas d'application concrets.

À travers cette étude, nous cherchons à démontrer que l'intégration conjointe de la notion d'utilité, des représentations concises et des données issues de LBSNs constitue une voie pertinente vers une fouille de données plus efficace, plus informative et mieux adaptée aux exigences réelles de valorisation de l'information.

Chapitre 1

Aperçu Général sur le Data Mining

1.1 Introduction

La croissance rapide des volumes de données disponibles dans les environnements informatiques modernes a transformé les besoins d'analyse classique. Face à cette abondance, de simples méthodes statistiques ne suffisent plus pour dégager des connaissances utiles. C'est dans ce contexte que la fouille de données, ou *Data Mining*, est apparue comme une solution efficace et incontournable.

Ce chapitre présente un aperçu général sur le domaine de la fouille de données. Il s'attarde dans un premier temps sur la définition du processus d'extraction de connaissances à partir des données, plus connu sous l'appellation **KDD (Knowledge Discovery in Databases)**, en détaillant ses différentes étapes.

Par la suite, nous abordons les principales techniques de fouille de données, telles que la classification, l'estimation, la description, la recherche de motifs fréquents, et la prédiction. Chaque technique est illustrée par des exemples d'applications réels dans des secteurs variés comme la santé, l'éducation, le marketing ou encore la finance.

Enfin, le chapitre se termine par une discussion sur les défis liés à la mise en œuvre de la fouille de données, ainsi que sur ses perspectives stratégiques. Cette base théorique est essentielle pour appréhender les enjeux actuels de l'analyse des données et pour mieux comprendre le rôle stratégique de la fouille de données dans divers contextes décisionnels..

1.2 Le processus de découverte de connaissances dans les bases de données

1.2.1 Définition

"Knowledge Discovery in Databases (KDD) est défini comme un processus non trivial visant à extraire des motifs valides, nouveaux, utiles et compréhensibles à partir des données. Ce processus comprend plusieurs étapes interconnectées, notamment la sélection des données, leur nettoyage, leur transformation, leur analyse à l'aide de techniques de fouille de données, ainsi que l'interprétation des résultats obtenus."[1]

1.2.2 Objectifs

Le processus d'extraction des connaissances à partir des bases de données vise à identifier des motifs et des connaissances utiles à partir des données, ce qui permet de :

- **Améliorer la prise de décision** : en révélant des schémas cachés pouvant aider à comprendre les tendances générales.
- **Optimiser la performance** : en fournissant de nouvelles perspectives susceptibles d'améliorer les processus ou les produits.
- **Analyser de grandes quantités de données** : en facilitant le traitement d'importants volumes d'informations pour en extraire des résultats pertinents.

Ces objectifs jouent un rôle essentiel dans le soutien de nombreux secteurs comme les affaires, l'éducation, la santé et bien d'autres domaines [**?**].

1.2.3 Étapes du processus KDD

Sélection des données : Choisir les données pertinentes.

Prétraitement : Nettoyer les données, gérer les valeurs manquantes.

Transformation : Formater les données pour l'analyse.

Fouille de données (Data Mining) : Appliquer des algorithmes pour découvrir des motifs.

Évaluation : Vérifier la qualité et la pertinence des motifs extraits.

Utilisation des connaissances : Intégrer les résultats dans le système décisionnel.

Comme illustré dans la figure 1.1, le processus KDD suit plusieurs étapes clés.

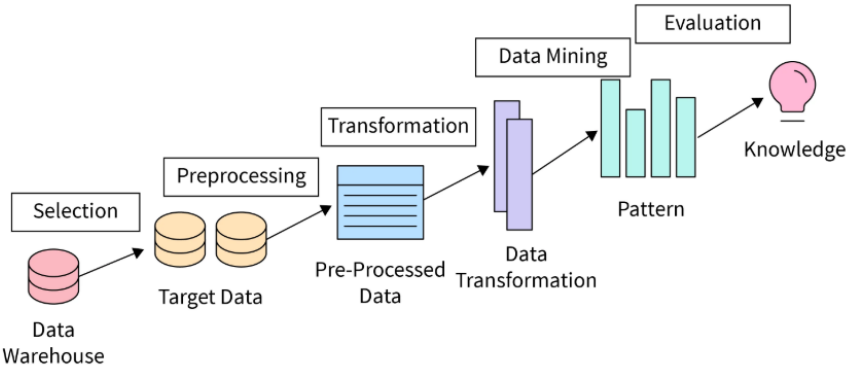


FIGURE 1.1 – Processus de découverte de connaissances à partir de données..

1.2.4 Rôle central du Data Mining dans le processus KDD

Le Data Mining représente l'étape clé et centrale du processus de découverte de connaissances dans les bases de données (KDD). Il consiste à appliquer des techniques d'analyse telles que la classification, la régression, le clustering ou encore l'analyse d'associations, afin d'extraire des motifs pertinents, nouveaux et potentiellement utiles à partir de grandes quantités de données.

Ce processus ne se limite pas à la simple extraction d'informations, mais inclut également l'évaluation et l'interprétation des résultats, dans le but de produire une connaissance compréhensible et applicable dans un contexte décisionnel.[2]

1.3 Fouille de données : Techniques et Domaines d'application

Au cours de la décennie 1996–2005, le domaine de la fouille de données et de la découverte de connaissances a connu une transformation significative. Initialement perçu comme un sujet académique prometteur, il a évolué pour devenir un pilier essentiel dans le monde des affaires et de l'analyse. Cette transition a été marquée par une maturation des techniques, une adoption accrue dans divers secteurs industriels et une reconnaissance de sa valeur stratégique.

Cette évolution a été influencée par plusieurs facteurs externes majeurs, notamment la croissance exponentielle du web et du commerce électronique, les avancées significatives en biologie (telles que les microarrays d'ADN) et l'utilisation controversée de la fouille de données pour la sécurité intérieure. De

nouveaux domaines de recherche ont émergé, tels que l'analyse des réseaux sociaux, la fouille du web et la fouille multimédia. Des progrès considérables ont été réalisés dans le raffinement des algorithmes de règles d'association, et de nouveaux algorithmes importants, tels que les machines à vecteurs de support (SVM), ont été largement acceptés.[3]

1.3.1 Définition

Selon David Hand, Heikki Mannila et Padhraic Smyth dans leur ouvrage Principles of Data Mining (2001), la fouille de données est définie comme suit :

« L'analyse de grands ensembles de données observationnelles afin de découvrir des relations inattendues et de résumer les données de manière nouvelle, compréhensible et utile pour le propriétaire des données. »[4]

Cette discipline repose sur des techniques issues de l'intelligence artificielle, des statistiques et de l'apprentissage automatique.

Un autre auteur, Frawley, Piatetsky-Shapiro et Matheus (1992), propose une définition complémentaire :

« L'extraction non triviale d'informations implicites, inconnues auparavant et potentiellement utiles à partir de données. »[5]

Cette définition met l'accent sur l'aspect exploratoire de la fouille de données et son rôle dans la découverte de connaissances cachées.

1.3.2 L'importance stratégique de la fouille de données

La fouille de données occupe aujourd'hui une place centrale dans les stratégies des entreprises modernes. Elle ne se limite plus à une tâche technique, mais constitue un levier stratégique pour guider la prise de décision, optimiser les performances et anticiper les tendances.

Selon Talend **(entreprise spécialisée dans les solutions de gestion des données) :** «Lorsqu'elle est employée efficacement, la fouille de données peut fournir aux organisations un avantage considérable par rapport à leurs concurrents. Elle permet en effet de mieux comprendre les clients, de développer des stratégies marketing efficaces, d'augmenter les revenus et de réduire les coûts. »

InsightSoftware souligne également que : « Le data mining soutient la prise de décision stratégique, améliore l'efficacité opérationnelle et favorise une meilleure compréhension des données. »

Enfin, selon **ia-data-analytics.fr :** « La fouille de données permet de transformer des données massives en connaissances exploitables pour orienter les décisions stratégiques. »

1.3.3 Les principales tâches du Data Mining

De nombreux défis dans les domaines intellectuel, économique ou commercial peuvent être abordés à travers différentes tâches fondamentales du Data Mining. Ces tâches sont généralement regroupées selon la nature de l'apprentissage utilisé.

Les tâches supervisées, telles que la classification, l'estimation et la prédiction, reposent sur l'utilisation de données historiques étiquetées pour construire un modèle capable de généraliser ou de prévoir une variable cible à partir des données disponibles [6].

En revanche, les tâches non supervisées telles que le regroupement par similarité (clustering) et l'analyse des groupes (clusters) ne nécessitent pas de variable cible, et visent plutôt à découvrir des structures naturelles ou des corrélations implicites entre les variables présentes dans les données[6] .

Il existe aussi des tâches hybrides comme la description, qui peut intervenir à la fois dans un contexte supervisé et non supervisé, selon la finalité de l'analyse et le type de données manipulées[7].

1.3.3.1 La classification

La classification consiste à affecter chaque élément à une catégorie prédéfinie en fonction de ses attributs. Par exemple, classer des e-mails comme spam ou non-spam.[7]

1.3.3.2 L'estimation

Cette tâche permet d'estimer une valeur numérique à partir des données existantes. Par exemple, estimer le revenu annuel d'un client en se basant sur ses transactions passées.[6]

1.3.3.3 La prédiction

Elle a pour objectif de prévoir des valeurs ou des événements futurs à partir des données historiques. Par exemple, prédire le volume des ventes pour le trimestre suivant.[6]

1.3.3.4 Le regroupement (Clustering)

Il s'agit de regrouper des objets similaires sans catégories prédéfinies. Cela permet d'identifier des structures cachées dans les données, comme des groupes de clients aux comportements d'achat similaires.[6]

1.3.3.5 L'analyse des cluster

Une fois les clusters formés, cette tâche consiste à analyser et interpréter les caractéristiques communes de chaque groupe afin de mieux comprendre les profils représentés.[7]

1.3.3.6 La description

Elle vise à résumer et interpréter les modèles extraits pour les rendre compréhensibles. Cela inclut par exemple la description des variables clés dans un segment de clientèle.[7]

1.3.3.7 L'extraction de motifs (Pattern Mining)

Cette tâche cherche à découvrir des régularités, des associations fréquentes ou des séquences significatives dans les données. Elle est particulièrement utile dans les systèmes de recommandation ou l'analyse de panier d'achats.[6]

Secteur	**Exemple d'application**
Banque	Détection de fraude, analyse du risque de crédit
Assurance	Modélisation des risques, fausses déclarations
Télécommunication	Prévision du churn, gestion du trafic réseau
Commerce	Recommandation de produits, analyse d'achats
Santé	Aide au diagnostic, suivi des traitements
Éducation	Suivi des étudiants, prédiction de l'échec scolaire

TABLE 1.1 – Domaines d'application du Data Mining

1.4 Conclusion

Dans ce chapitre, nous avons posé les bases conceptuelles de la fouille de données, en mettant en lumière ses origines, ses principales étapes à travers le processus KDD, ainsi que ses différentes approches et domaines d'application. Ce tour d'horizon nous a permis de mieux comprendre l'importance stratégique de l'analyse des données dans divers contextes décisionnels. Nous avons également distingué les types de techniques utilisées en fouille, qu'elles soient descriptives ou prédictives, et nous avons souligné les enjeux liés à l'extraction de connaissances pertinentes à partir de données massives. Ces éléments constituent une introduction indispensable pour appréhender les limitations des approches traditionnelles et pour justifier l'émergence de méthodes plus orientées vers la valeur. Cette réflexion théorique met ainsi en évidence les fondements nécessaires à une compréhension approfondie des méthodes modernes en fouille de données.

Chapitre 2

Les Réseaux Sociaux Géolocalisés (LBSNs)

2.1 Introduction

Avec l'évolution rapide des technologies de l'information et de la communication, une nouvelle génération de réseaux sociaux a vu le jour : les réseaux sociaux géolocalisés, ou *Location-Based Social Networks (LBSNs)*. Ces réseaux permettent aux utilisateurs de partager leur position géographique en temps réel et d'interagir avec des contenus et des personnes à proximité, générant ainsi des données riches à double dimension : sociale et spatiale.

La diffusion croissante de ces réseaux, tels que *Foursquare*, *Gowalla* et *Brightkite*, a engendré d'importants volumes de données portant sur la mobilité et les comportements des utilisateurs. Une fois transformées en format transactionnel, ces données deviennent exploitables dans des tâches d'analyse approfondie, notamment dans le domaine de l'extraction de connaissances.

Dans ce contexte, ce chapitre propose une étude structurée autour des LBSNs, abordant leur définition, leurs caractéristiques et leurs usages analytiques. Une attention particulière est accordée à leur intégration dans les systèmes d'extraction des règles d'association à haute utilité, en lien avec les algorithmes de représentation concise des données, contribuant ainsi à l'atteinte des objectifs analytiques de cette étude.

2.2 Intégration des dimensions spatiale et sociale : Étude des LBSNs

Les réseaux sociaux basés sur la localisation ont transformé les dynamiques d'interaction numérique. Les relations sociales ne se limitent plus à des connexions abstraites, mais sont désormais fortement influencées par la position géographique. Scellato et al. [8] ont démontré que les lieux communs fréquentés par les utilisateurs favorisent la création de liens sociaux.

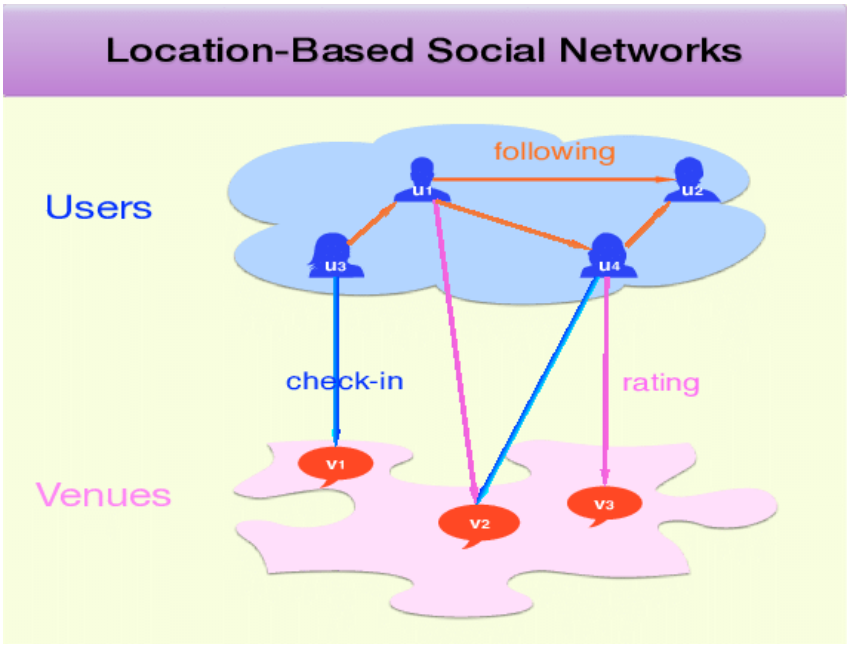


FIGURE 2.1 – Modèle de réseau social géolocalisé à deux couches (utilisateurs – lieux)

2.3 Les LBSNs : définition et applications contemporaines

Les LBSNs permettent des interactions sociales contextuelles. Grâce aux fonctionnalités de géolocalisation comme le "check-in", ils sont utilisés dans la publicité ciblée, les recommandations de lieux et la gestion des foules. Zheng et al. [9] ont démontré l'impact des données GPS historiques sur l'efficacité des systèmes de recommandation.

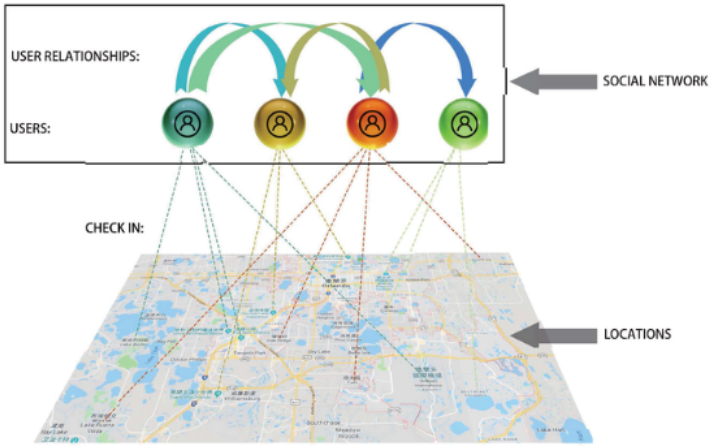


FIGURE 2.2 – Architecture générale d'un réseau social basé sur la localisation (LBSN)

2.3.1 Exemples de réseaux LBSNs : Gowalla, Foursquare et Brightkite

Foursquare a introduit le concept de check-ins et de recommandations sociales. *Gowalla*, avec une approche ludique, proposait une collecte d'objets virtuels selon les lieux visités. *Brightkite* permettait le partage de position en temps réel. Les données issues de ces réseaux ont été utilisées par Cho et al. [10] pour analyser la relation entre mobilité et formation des liens sociaux.

2.4 Analyse des réseaux sociaux spatiaux : Du numérique au data mining

Chaque interaction dans un LBSN peut être modélisée comme une transaction. Bao et al. [11] ont combiné préférences utilisateur et localisation dans un modèle de recommandation, illustrant l'intérêt des LBSNs comme sources de données transactionnelles complexes.

2.5 Caractéristiques et exploitations analytiques des LBSNs

La centralité du lieu influence la structure des interactions sociales, dans la mesure où certains lieux — appelés points d'intérêt (POI) — agissent comme des nœuds de haute centralité dans les réseaux de mobilité. Ces lieux, en raison de leur accessibilité ou de leur popularité, facilitent les interactions et peuvent être considérés comme des hubs comportementaux.

Les types de lieux (restaurants, écoles, transports, etc.), leur fréquence de visite et leur contexte fonctionnel (travail, loisirs, santé) permettent une analyse comportementale fine, révélant des régularités dans les habitudes individuelles et collectives. Ces éléments sont exploités pour segmenter les profils utilisateurs, comprendre les dynamiques sociales, et anticiper les flux de déplacement.

Cho et al. [10] ont confirmé le rôle fondamental de la mobilité dans la dynamique des réseaux sociaux, en mettant en lumière l'effet de la proximité géographique sur la formation et l'intensité des liens sociaux.

La figure 2.3 illustre un modèle géosocial multilayer combinant les utilisateurs, le contenu et la localisation, soulignant ainsi la complexité structurelle des LBSNs.

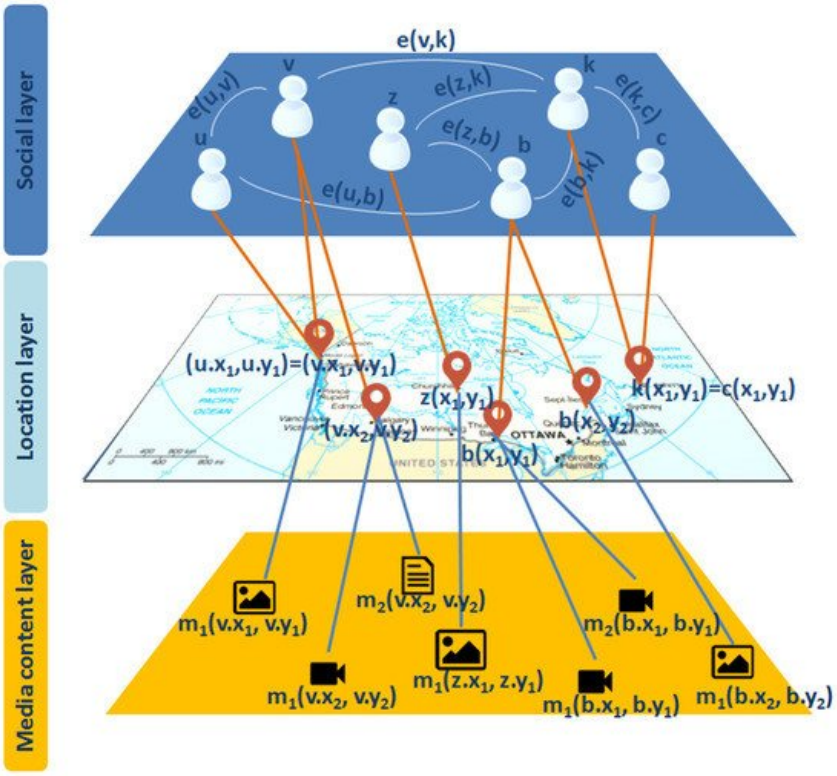


FIGURE 2.3 – Modèle géosocial multilayer : utilisateurs, contenu et emplacement

2.6 Approches analytiques autour des LBSNs

L'analyse des réseaux sociaux géolocalisés (LBSNs) ne se limite pas à une simple observation des interactions sociales dans l'espace. Elle implique une variété d'approches analytiques visant à modéliser, prédire et comprendre les comportements des utilisateurs en fonction de leurs déplacements et de leurs connexions sociales. Les sous-sections qui suivent présentent les principales perspectives développées dans ce cadre.

2.6.1 Vers une compréhension comportementale via les LBSNs

L'étude des routines de déplacement permet de prédire les comportements futurs en exploitant les séquences temporelles et spatiales des check-ins. Ces routines peuvent être modélisées via des chaînes de Markov, des modèles de prédiction séquentielle (RNN, LSTM), ou encore des modèles probabilistes bayésiens.

Gao et al. [12] ont proposé un modèle corrélant la position géographique et les relations sociales afin d'anticiper les comportements futurs, tels que la probabilité qu'un utilisateur visite un lieu en fonction de ceux fréquentés par ses amis.

2.6.2 Fondements théoriques et potentialités analytiques

Les LBSNs s'appuient sur les concepts de graphes hétérogènes, où les sommets représentent des entités diverses (utilisateurs, lieux, contenus) et les arêtes, des relations spatio-sociales. Cette modélisation permet d'appliquer des techniques de classification, de clustering (ex. : DBSCAN spatio-temporel), ou d'analyse communautaire.

Liu et Lee [13] ont développé un modèle hybride alliant classification statistique et analyse de graphes pour explorer la structure multidimensionnelle des interactions dans un LBSN, montrant que la prise en compte simultanée des dimensions spatiales et sociales améliore la précision des analyses.

2.6.3 LBSNs et fouille spatio-sociale : Défis méthodologiques et perspectives

L'exploitation massive des données issues des LBSNs soulève plusieurs défis :

— **Scalabilité** : Le volume élevé de données géolocalisées requiert l'utilisation de frameworks distribués (ex. : Spark, Hadoop).
— **Redondance et bruit** : Les check-ins multiples dans un même lieu peuvent biaiser les analyses.
— **Vie privée** : Les données de localisation soulèvent des enjeux éthiques (traçabilité, anonymisation).

Zaki et al. [14] ont proposé des algorithmes de représentation concise, comme les itemsets fermés fréquents, facilitant l'extraction de motifs pertinents à partir de transactions complexes tout en réduisant le volume de sortie.

2.7 Intégration des LBSNs dans le processus de fouille à haute utilité

Les réseaux sociaux géolocalisés (LBSNs) ne représentent pas uniquement des plateformes d'interaction sociale, mais également des sources riches de données exploitables dans des contextes analytiques avancés. Les check-ins, les itinéraires d'utilisateurs et les relations spatio-temporelles peuvent être convertis en transactions formelles : chaque transaction représente un utilisateur ou une session temporelle, et chaque item représente un lieu visité ou une catégorie de lieu.

L'attribution d'une **valeur d'utilité** à chaque item — comme la fréquence de visite, la durée moyenne de séjour ou la dépense estimée — permet d'appliquer des algorithmes de fouille à haute utilité (HUIM), tels que HUCI-Miner, UP-Growth ou EFIM.

Ce type d'analyse permet de faire émerger des motifs à forte valeur ajoutée (par exemple, la combinaison de lieux régulièrement visités par un segment spécifique d'utilisateurs), qui peuvent ensuite être utilisés dans des systèmes de recommandation, des campagnes de géomarketing ou la planification urbaine.

2.8 Conclusion

Dans ce chapitre, nous avons exploré les fondements des réseaux sociaux géolocalisés (LBSNs), en définissant leur concept, leurs caractéristiques structurelles, ainsi que leurs principales applications analytiques. Nous avons mis en évidence la double dimension – sociale et spatiale – qui distingue ces réseaux des plateformes sociales traditionnelles, et qui en fait une source précieuse d'information contextuelle.

L'analyse des LBSNs permet de modéliser des comportements humains complexes, en tenant compte à la fois des interactions sociales et des dynamiques géographiques. Cette structure hybride offre des opportunités prometteuses en matière d'extraction de connaissances, notamment dans les contextes où la localisation joue un rôle stratégique.

Ainsi, les LBSNs se positionnent comme des environnements riches et pertinents pour l'analyse avancée des données, grâce à leur capacité à refléter des réalités comportementales ancrées dans l'espace et la société.

Chapitre 3

High Utility Itemset Mining HUM

3.1 Introduction

Les approches classiques d'extraction de motifs fréquents, bien qu'utiles dans de nombreux contextes, présentent des limites importantes lorsqu'il s'agit d'évaluer la valeur réelle des items impliqués dans les transactions. En effet, le critère de fréquence ne suffit pas à capturer l'importance stratégique ou économique d'un item dans des environnements où la rentabilité est une priorité.

C'est dans cette optique qu'est apparue la fouille des motifs à haute utilité (*High Utility Itemset Mining - HUIM*). Cette approche vise à identifier les ensembles d'items dont l'impact global est significatif selon une métrique d'utilité définie, souvent en fonction de la quantité et du profit unitaire. Contrairement aux techniques traditionnelles, HUIM permet donc une analyse plus fine, orientée valeur, en ligne avec les exigences des applications réelles.

Ce chapitre est consacré à la présentation des fondements du HUIM. Nous y abordons d'abord les concepts préliminaires liés à la fouille des itemsets fréquents, avant de mettre en lumière les insuffisances qui ont conduit à l'émergence du paradigme HUIM. Ensuite, nous détaillons les notions clés associées à cette nouvelle approche, telles que l'utilité transactionnelle, le seuil *minUtil*, ou encore le TWU (Transaction Weighted Utility).

Enfin, une attention particulière est portée aux représentations concises des motifs à haute utilité, qui permettent de réduire la redondance des résultats extraits tout en conservant leur valeur analytique. L'ensemble de ces éléments constitue une base théorique solide pour la suite du mémoire, qui portera sur la mise en œuvre expérimentale et l'analyse comparative de différentes stratégies de fouille.

3.2 Frequent Itemset Mining(FIM)

Selon Agrawal et Srikant (1994), l'extraction d'ensembles d'items fréquents (Frequent Itemset Mining) est une tâche visant à identifier les groupes d'éléments apparaissant fréquemment ensemble dans une base de données transactionnelle, avec un taux de support supérieur à un seuil prédéfini. Cette technique permet de découvrir des relations cachées et des motifs récurrents entre les éléments au sein de grands ensembles de données. Elle est applicable dans divers domaines tels que l'analyse du comportement, la détection de fraudes, les systèmes de recommandation, ou encore l'aide à la décision. L'objectif est d'améliorer la compréhension des structures internes des données et de faciliter l'extraction de connaissances. FIM constitue une étape essentielle dans la génération de règles d'association et dans de nombreuses applications du data mining.[15]

3.3 Notions de base du FIM

Pour aborder le FIM, il est essentiel de définir ses concepts fondamentaux :

3.3.1 Bases de données transactionnelles

Une base de données transactionnelle regroupe un ensemble de transactions, chacune représentant une collection d'items achetés ou associés simultanément. Chaque transaction peut être vue comme une instance indépendante permettant d'analyser la co-occurrence des items. Ces bases sont fondamentales pour les méthodes d'extraction des motifs fréquents dans les données. [16]

TID	Transaction
T1	a, c, d
T2	b, c, e
T3	a, b, c, e
T4	b, e
T5	a, b, c, e

TABLE 3.1 – Domaines d'application du Data Mining

3.3.2 Item (Élément)

Un item est une entité indivisible représentée dans une base de données transactionnelle.Par Exemple : Dans une base de données d'achats, « lait », « pain » ou « fromage » sont des items.[17]

3.3.3 Itemset (Ensemble d'items)

Un itemset est un ensemble non vide d'items. Il représente un groupe d'éléments co-présents dans une transaction.par Exemple : pain, fromage, beurre est un itemset contenant trois éléments.[17]

3.3.4 Frequent Itemset (Ensemble d'items fréquent)

Un ensemble d'items fréquent est un groupe d'éléments apparaissant conjointement dans un nombre important de transactions d'une base de données. Il est qualifié de fréquent lorsque sa fréquence (support) dépasse un seuil minimal prédéfini. Ce concept est central dans les tâches de fouille d'associations pour révéler des corrélations significatives entre les données.[16]

3.3.5 Mesure de support (Support)

Le support est une mesure qui quantifie la fréquence d'apparition d'un itemset dans une base de données transactionnelle. Il permet de déterminer si un motif est suffisamment représentatif pour être considéré comme pertinent dans l'analyse.[17]

3.3.6 Défis de l'extraction des ensembles d'items fréquents

3.3.6.1 Taille de l'espace de recherche

Le nombre possible d'ensembles d'items dans une base contenant m items est $2^m - 1$, ce qui engendre une complexité exponentielle même pour de petites bases.[17]

3.3.6.2 Surabondance des motifs

Les algorithmes de FIM génèrent souvent trop de motifs, dont plusieurs sont redondants, rendant l'analyse difficile.[17]

3.3.6.3 Limites du support

Le support ne reflète pas toujours l'intérêt réel des items, ce qui peut entraîner la perte de motifs significatifs mais rares.[17]

3.3.6.4 Données complexes et évolutives

Les données incertaines ou en évolution posent des défis que les méthodes classiques peinent à gérer efficacement.[17]

3.3.7 Algorithmes d'extraction des items fréquents

Afin de réaliser une extraction efficace des ensembles d'items fréquents, il est primordial de concevoir des approches capables d'éviter l'énumération exhaustive de toutes les combinaisons possibles. Pour cela, les algorithmes doivent manipuler les ensembles d'items de manière optimisée. Plusieurs méthodes performantes ont ainsi vu le jour, notamment Apriori, FP-Growth, Eclat et LCM. Ces algorithmes, bien qu'ils produisent les mêmes résultats attendus à partir des mêmes données en entrée, se distinguent principalement par les techniques employées ainsi que les structures de données adoptées afin d'améliorer l'efficacité du processus de découverte [17] .

Algorithme	**Type de recherche**	**Représentation de la BD**
Apriori	Recherche en largeur	Horizontal (Transactions brutes)
AprioriTID	Recherche en largeur	Vertical (TID-lists)
FP-Growth	Recherche en profondeur	Horizontal (FP-Tree)
Eclat	Recherche en profondeur	Vertical (TID-lists)
dEclat	Recherche en profondeur	Vertical (Diffsets)
Relim	Recherche en profondeur	Horizontal (Base réduite dynamiquement)
H-Mine	Recherche en profondeur	Horizontal (H-Struct)
LCMFreq	Recherche en profondeur	Horizontal (Transactions compressées)
PrePost	Recherche en profondeur	Vertical (N-lists)
PrePost+	Recherche en profondeur	Vertical (N-lists optimisées)
FIN	Recherche en profondeur	Vertical (N-lists)
DFIN	Recherche en profondeur	Vertical (Diff-N-lists)
NegFIN	Recherche en profondeur	Vertical (N-lists avec négation)

TABLE 3.2 – Algorithmes d'extraction des ensembles d'items fréquents

3.3.8 Stratégies de recherche dans l'extraction des itemsets fréquents

Pour extraire efficacement les ensembles d'items fréquents à partir d'une base de données transactionnelle, les algorithmes utilisent principalement deux stratégies de recherche dans l'espace des solutions :

Recherche en largeur (Breadth-First Search) La recherche en largeur consiste à explorer l'espace des itemsets par **niveau de taille croissante**. L'algorithme commence par examiner tous les itemsets de taille 1, puis ceux de taille 2, et ainsi de suite jusqu'à atteindre les ensembles de taille maximale. Cette approche permet de **réduire le nombre de candidats** grâce à des propriétés anti-monotones, comme la propriété Apriori. Généralement, elle utilise des structures de type arbre de hachage (Hash Tree) pour accélérer le comptage de support.[18]

Exemple : L'algorithme **Apriori** est un exemple classique de cette stratégie. Il génère tous les candidats de taille k à partir de ceux de taille $k-1$, puis les évalue un à un.

Recherche en profondeur (Depth-First Search) La recherche en profondeur explore l'espace des itemsets en suivant une **stratégie récursive** : elle commence par un itemset de taille 1, puis y ajoute progressivement d'autres items pour former des ensembles plus grands, avant de revenir en arrière (backtracking). Cette stratégie est plus **économe en mémoire** car elle n'a pas besoin de stocker tous les candidats simultanément.[19] *Exemples* :

- **FP-Growth** : construit un arbre compact (FP-Tree) et applique une exploration récursive basée sur les sous-bases conditionnelles.[19]
- **Eclat** : utilise l'intersection des identifiants de transactions (TID) et explore les itemsets selon une approche DFS.[19]
- **LCM** : méthode efficace pour extraire les itemsets fermés fréquents, également fondée sur la recherche en profondeur.[19]

3.3.9 Représentation de la base de données

La représentation de la base de données joue un rôle crucial dans l'efficacité des algorithmes d'extraction des itemsets fréquents. Deux approches principales sont utilisées :

- **Représentation horizontale :** dans cette approche, la base de données est considérée comme une liste de transactions, où chaque transaction contient un ensemble d'items. Cette méthode est simple et intuitive, mais elle peut devenir inefficace lorsque le nombre de transactions est très élevé.[17]
- **Représentation verticale :** chaque item est associé à une liste des identifiants des transactions (TID) dans lesquelles il apparaît, appelée *tidset*. Cette représentation facilite le calcul des intersections nécessaires pour l'extraction des motifs fréquents, ce qui peut améliorer la performance de certains algorithmes.[17]

Le choix entre ces deux représentations dépend du contexte d'application et des caractéristiques des données.

3.3.10 Génération des ensembles d'items fréquents

La génération des ensembles d'items fréquents est une étape essentielle dans l'extraction de motifs à partir de bases de données transactionnelles. Elle consiste à identifier les combinaisons d'items qui

apparaissent fréquemment ensemble dans les transactions. Cette tâche repose généralement sur une mesure de support, qui indique la fréquence d'apparition d'un ensemble d'items.Deux stratégies principales sont utilisées :

— **Approche ascendante (bottom-up)** : utilisée dans des algorithmes comme Apriori, elle commence par les itemsets de taille 1 et génère progressivement des itemsets plus grands en combinant les plus petits.[17]
— **Approche descendante (top-down)** : elle explore les itemsets les plus grands en premier, souvent utilisée dans des contextes spécifiques où les grands itemsets sont connus ou ciblés.[17]

L'efficacité de cette génération dépend largement des structures de données et des heuristiques utilisées pour éviter la génération de candidats inutiles.

3.3.11 Limitations de l'extraction des ensembles fréquents

L'extraction des ensembles fréquents (Frequent Itemset Mining - FIM) présente plusieurs limitations importantes :

1. **Explosion combinatoire** : Lorsque le nombre d'items augmente ou que le seuil de support minimal est réduit, le nombre d'ensembles candidats croît de manière exponentielle, ce qui entraîne une consommation importante de temps et de mémoire.[17]
2. **Résultats excessifs ou peu utiles** : Des milliers voire des millions d'ensembles fréquents peuvent être extraits, rendant leur interprétation et leur exploitation difficiles dans la pratique.[17]
3. **Ignorance de l'utilité** : La plupart des algorithmes FIM se basent uniquement sur la fréquence d'apparition, sans prendre en compte l'utilité, le profit ou l'importance relative des items.[17]
4. **Incapacité à détecter les relations négatives** : Les algorithmes classiques ne révèlent généralement que les corrélations positives, négligeant les relations négatives où la présence d'un item implique l'absence d'un autre.[17]
5. **Seuil de support arbitraire** : Le choix d'un seuil de support adéquat est difficile. Un seuil trop élevé peut ignorer des motifs intéressants, tandis qu'un seuil trop bas génère un grand nombre de motifs inutiles.[17]
6. **Manque de prise en compte du contexte** : Les algorithmes traditionnels ne considèrent pas des facteurs contextuels tels que le temps, le lieu ou le profil de l'utilisateur, ce qui limite la pertinence des motifs extraits.[17]

3.4 High Utility Itemset Mining HUIM

3.4.1 Extraction des ensembles d'items à haute utilité (HUIM)

L'extraction des ensembles d'items à haute utilité (High Utility Itemset Mining - HUIM) est un domaine de recherche émergent en fouille de données qui vise à identifier des ensembles d'items ayant une grande importance, souvent mesurée en termes de profit, quantité ou autre valeur utilisateur. Contrairement à l'extraction traditionnelle d'ensembles fréquents, HUIM ne se limite pas à la fréquence d'apparition mais prend en compte l'utilité associée à chaque item, ce qui permet d'extraire des motifs plus significatifs pour la prise de décision.

Selon Fournier-Viger et al. (2019), HUIM a été introduit pour répondre aux limites de la fouille d'ensembles fréquents, notamment l'incapacité à détecter les motifs peu fréquents mais à forte valeur ajoutée. Ce paradigme est largement applicable dans les domaines commerciaux, comme l'analyse des paniers d'achat, la recommandation de produits ou la gestion des stocks.[17]

3.4.2 Notions de base du HUIM

3.4.2.1 Base de données transactionnelle

Une base de données transactionnelle est un ensemble de transactions, où chaque transaction est un ensemble d'items associés à des quantités. Chaque transaction possède un identifiant unique appelé **TID (Transaction ID)**.

TID	Items de la transaction
T1	A(2), B(1), C(1)
T2	A(1), C(3)
T3	B(2), D(1)

3.4.2.2 Identifiant de transaction (TID)

Le TID est utilisé pour identifier chaque transaction de manière unique. Il est essentiel dans les structures comme *utility-list* pour suivre les contributions des transactions à l'utilité des itemsets.[17]

3.4.2.3 Profit unitaire

Chaque item est associé à un **profit unitaire** (unit profit), c'est-à-dire la valeur gagnée pour chaque unité vendue. Cette valeur est fixe et donnée en amont.[17]

Item	Profit unitaire
A	5
B	2
C	1
D	3

3.4.2.4 Utilité d'un item dans une transaction

Elle est calculée par :

$$u(i,T) = \text{profit unitaire}(i) \times \text{quantité}(i,T)$$

[17]

Par exemple : $u(A,T1) = 5 \times 2 = 10$.

3.4.2.5 Utilité d'une transaction (TU)

L'utilité totale d'une transaction est la somme des utilités de tous les items qu'elle contient :

$$TU(T) = \sum_{i \in T} u(i,T)$$

[17]

Exemple pour T1 : $TU(T1) = u(A,T1) + u(B,T1) + u(C,T1) = 10 + 2 + 1 = 13$.

3.4.2.6 Transaction-Weighted Utilization (TWU)

Le TWU d'un ensemble d'items X est la somme des utilités des transactions contenant X :

$$TWU(X) = \sum_{T \supseteq X} TU(T)$$

Le TWU est utilisé pour estimer rapidement si un ensemble peut atteindre le seuil d'utilité minimal.[17]

3.4.2.7 Ensemble à haute utilité (High Utility Itemset)

Un ensemble d'items X est dit à haute utilité s'il vérifie :

$$u(X) \geq \text{minUtil}$$

où $u(X)$ est la somme de ses utilités dans toutes les transactions.[17]

3.4.2.8 Propriétés importantes

— **Propriété descendante du TWU** : si TWU(X) < minUtil, alors aucun super-ensemble de X ne peut être à haute utilité.[17]

— **Pas de propriété anti-monotone stricte pour l'utilité réelle** : contrairement au support, l'utilité d'un itemset peut être supérieure à celle de ses sous-ensembles.[17]

3.4.3 Algorithmes pour l'extraction des ensembles d'items à haute utilité

L'extraction des ensembles d'items à haute utilité (HUIs) vise à identifier les motifs les plus rentables dans les bases de données transactionnelles contenant des informations de profit. Plusieurs algorithmes ont été proposés dans la littérature pour cette tâche, en fonction de la stratégie de recherche adoptée (profondeur ou largeur), de la représentation des données (horizontale ou verticale), et du nombre de phases utilisées.

Algorithme	Type de recherche	Nb de phases	Représentation de la base	Extensions
EFIM (Zida et al., 2015)	Depth-first	1	Verticale (Utility-list)	Optimisations par projection et tri
FHM (Fournier-Viger et al., 2014)	Depth-first	1	Verticale (Utility-list)	Pruning avec EUCP
HUI-Miner (Liu & Qu, 2012)	Depth-first	1	Verticale (Utility-list)	Aucune
HUP-Miner (Krishnamoorthy, 2014)	Breadth-first	2	Horizontale	Extension d'IHUP
mHUIMiner (Peng et al., 2017)	Depth-first	1	Verticale	Optimisation mémoire
UFH (Dawar et al., 2017)	Breadth-first	2	Horizontale	Gestion efficace de TWU
HMiner (Krishnamoorthy, 2017)	Depth-first	1	Verticale	Amélioration de HUI-Miner
ULB-Miner (Duong et al., 2018)	Breadth-first	2	Horizontale	Utilisation des listes binaires
IHUP (Ahmed et al., 2009)	Breadth-first	2	Horizontale (IHUP-tree)	Base des autres algorithmes
Two-Phase (Liu et al., 2005)	Breadth-first	2	Horizontale	TWU-based pruning
UP-Growth (Tseng et al., 20a)	Breadth-first	2	Horizontale (UP-tree)	Stratégies de réduction
UP-Growth+ (Tseng et al., 2013)	Breadth-first	2	Horizontale (UP-tree)	Version optimisée
UP-Hist (Dawar et al., 2015)	Breadth-first	2	Horizontale	Pruning par histogrammes
d2HUP (Liu et al., 2012)	Depth-first	1	Horizontale	Double pruning heuristique
FHIM (Sahoo et al., 2015)	Depth-first	1	Verticale	Optimisation par compression

TABLE 3.3 – Résumé des algorithmes pour l'extraction des HUIs

3.4.4 Limitations du High Utility Itemset Mining (HUIM)

Le *High Utility Itemset Mining (HUIM)* est une tâche puissante mais extrêmement complexe, en raison principalement de la taille massive de l'espace de recherche. Selon **Fournier-Viger et al. (2017)** :

> *"The number of possible itemsets can be very large. For example, if a retail store has 10,000 items on its shelves (m = 10,000), the utilities of 2^10,000 1 possible itemsets should be calculated, which is unmanageable using the naive approach."*
> [17]

Cela montre que la taille exponentielle de l'espace de recherche constitue l'un des plus grands défis du HUIM, même pour des bases de données relativement petites. Cette caractéristique entraîne une consommation mémoire élevée et un temps de traitement conséquent.

Pour faire face à ces limitations, plusieurs algorithmes ont été développés en s'appuyant sur des techniques de réduction de l'espace de recherche. Parmi celles-ci, l'utilisation de **Représentations concises** s'est révélée particulièrement efficace. Ces représentations permettent de diminuer le nombre de motifs extraits, tout en conservant leur pertinence, ce qui améliore considérablement les performances globales des algorithmes.

3.5 Représentations concises des ensembles d'items à haute utilité

Les **représentations concises** visent à réduire le nombre d'ensembles d'items à haute utilité (HUIs) extraits, tout en conservant l'information essentielle. Elles permettent une analyse plus efficace et une meilleure compréhension des motifs pertinents. Les principales représentations concises sont les suivantes :

— **Ensembles fermés à haute utilité (CHUIs)** : Un ensemble d'items X est un CHUI s'il n'existe aucun ensemble Y tel que $X \subset Y$, $Y \in HUIs$ et $sup(X) = sup(Y)$. Formellement :

$$CHUIs = \{X \in HUIs \mid Y \in HUIs : X \subset Y \wedge sup(X) = sup(Y)\}$$

Les CHUIs permettent de réduire la redondance des motifs extraits. De plus, en stockant des informations supplémentaires sur la structure des CHUIs, il est possible de reconstruire l'ensemble complet des HUIs sans re-scanner la base de données.[17]

— **Ensembles maximaux à haute utilité (MHUIs)** : Un ensemble d'items X est un MHUI s'il n'existe aucun ensemble Y tel que $X \subset Y$ et $Y \in HUIs$. Formellement :

$$MHUIs = \{X \in HUIs \mid Y \in HUIs : X \subset Y\}$$

Les MHUIs représentent les ensembles les plus larges à haute utilité. Bien qu'ils réduisent considérablement le nombre de motifs, ils ne permettent pas de reconstruire l'ensemble complet des HUIs, car certaines informations peuvent être perdues.[17]

— **Générateurs à haute utilité (GHUIs)** : Un ensemble d'items X est un GHUI s'il satisfait deux conditions :

1. Il n'existe aucun ensemble $Y \subset X$ tel que $sup(X) = sup(Y)$.
2. Il existe un ensemble $Z \supseteq X$ tel que $u(Z) \geq minutil$.

Formellement :

$$GHUIs = \{X \mid Y \subset X : sup(X) = sup(Y) \wedge \exists Z \supseteq X : u(Z) \geq minutil\}$$

Les GHUIs représentent les plus petits ensembles d'items communs à des transactions contenant des ensembles à haute utilité. Ils sont particulièrement utiles pour des applications telles que la classification ou la compression des données.[17]

— **Ensembles minimaux à haute utilité (MinHUIs)** : Un ensemble d'items X est un MinHUI si $u(X) \geq minutil$ et qu'aucun de ses sous-ensembles $Y \subset X$ ne satisfait $u(Y) \geq minutil$. Formellement :

$$MinHUIs = \{X \mid u(X) \geq minutil \wedge Y \subset X : u(Y) \geq minutil\}$$

Les MinHUIs identifient les plus petits ensembles d'items générant une utilité élevée. Ils sont particulièrement intéressants pour des applications marketing, car ils permettent de cibler des combinaisons de produits simples mais rentables.[17]

Algorithmes pour l'extraction des représentations concises :

L'extraction des représentations concises des ensembles d'items à haute utilité permet de résumer l'information utile tout en évitant la redondance. Plusieurs types de représentations ont été présentés précédemment, tels que les ensembles fermés, maximaux, minimaux et les générateurs.

Afin d'identifier ces formes compactes de manière efficace, divers algorithmes ont été proposés dans la littérature. **Le tableau suivant** offre un aperçu organisé de ces algorithmes, en les classant selon le type de représentation qu'ils visent. Ces méthodes permettent généralement une extraction plus rapide et plus concise que la découverte de l'ensemble complet des itemsets à haute utilité, tout en conservant une richesse informative suffisante pour de nombreuses applications analytiques.

Algorithme	Type de recherche	Nb de phases	Représentation de la base	Extensions
CHUD (20a/2015)	HUIs fermés	2	Utility-lists	CHUD
HUG-Miner (2014)	Générateurs HUIs	2	Transactions transformées	HUCI-Miner
GHUI-Miner (2014)	Générateurs HUIs	1	Base projetée	-
CHUI-Miner (2015)	HUIs fermés	2	Utility-lists	CHUI-Miner
HUCI-Miner (2015)	HUIs fermés + générateurs	2	Utility-lists croisées	HUCI-Miner
MinFHM (2016)	HUIs minimaux	2	Utility-lists améliorées	Pruning
EFIM-Closed (2016)	HUIs fermés	2	Utility-lists	EFIM
CLS-Miner (2019)	HUIs fermés	2	Utility-lists compressées	CLS-Miner
HMiner_Closed (2019)	HUIs fermés	1	Utility-lists	HMiner
CHUI-Miner Max (2019)	HUIs maximaux	2	Utility-lists	CHUI-Miner Max

TABLE 3.4 – Résumé des algorithmes pour l'extraction des représentations concises des HUIs

3.6 Définition des règles d'association

Les règles d'association désignent des implications de la forme $X \Rightarrow Y$ entre deux ensembles d'items dans une base de données transactionnelle. Introduites par **Agrawal et Srikant (1994)**, elles permettent d'identifier des cooccurrences fréquentes entre les items, dans le but de révéler des relations intéressantes ou significatives. Le processus de génération repose principalement sur deux étapes :

— l'extraction des ensembles d'items fréquents à l'aide d'un seuil minimal de support,

— puis la construction de règles satisfaisant un seuil de confiance donné.

Ces règles ont été largement utilisées dans des domaines comme le commerce électronique ou les systèmes de recommandation [15].

3.6.0.1 Règles d'association basées sur la fréquence

Les approches basées sur la fréquence s'appuient sur l'identification des cooccurrences fréquentes entre les items dans une base de données transactionnelle. Elles reposent sur des mesures telles que le support et la confiance pour extraire des règles significatives [15].

Concepts fondamentaux

— **Support** : Le support désigne la fréquence à laquelle un ensemble d'éléments $X \cup Y$ apparaît dans la base de données. Il se calcule par la formule :

$$\text{Support}(X \Rightarrow Y) = \frac{|\{t \in D : X \cup Y \subseteq t\}|}{|D|}$$

[15]

— **Confiance (Confidence)** : La confiance mesure la probabilité que Y soit présent dans une transaction contenant déjà X. Elle est exprimée par :

$$\text{Confiance}(X \Rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

[15]

— **Lift** : Le lift évalue l'indépendance statistique entre X et Y, en comparant la confiance avec le support de Y. Il est donné par :

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Confiance}(X \Rightarrow Y)}{\text{Support}(Y)} = \frac{P(X \cup Y)}{P(X) \cdot P(Y)}$$

[15]

Algorithmes d'extraction des motifs fréquents

— **Apriori** : Il s'agit d'un algorithme proposé par Agrawal et Srikant en 1994 pour découvrir des itemsets fréquents et générer des règles d'association dans de grandes bases de données. Il fonctionne de manière itérative en explorant les ensembles d'items de taille croissante, tout en éliminant les combinaisons non fréquentes à l'aide de la propriété Apriori (tout sous-ensemble d'un itemset fréquent est lui-même fréquent) [15].

— **FP-Growth** : Proposé par Han, Pei et Yin en 2000, cet algorithme vise à extraire des motifs fréquents sans générer de candidats. Il repose sur une structure compacte appelée *FP-tree*, construite à partir de la base de données compressée. En explorant cette structure, FP-Growth permet une recherche plus rapide et plus efficace, notamment dans les bases volumineuses [20].

3.6.0.2 Règles d'association à utilité élevée

Contrairement aux règles d'association traditionnelles qui s'appuient uniquement sur la fréquence des items, les règles d'association à utilité élevée prennent en compte la valeur ou le profit associé à chaque item. Cette approche permet de découvrir des règles plus pertinentes dans les contextes où les items n'ont pas tous la même importance. Selon Fournier-Viger et al. (2019), l'extraction de tels motifs vise à identifier des ensembles d'items générant une utilité globale élevée, tout en respectant des contraintes comme le support minimal ou la confiance [17].

Concepts fondamentaux

— **Utilité d'un item** : dans le High Utility Itemset Mining (HUIM), l'utilité représente une mesure numérique indiquant la valeur ou l'importance d'un item dans une transaction. Elle est calculée

comme le produit entre la quantité de l'item dans la transaction et sa valeur unitaire (ex. profit) :

$$u(i,t) = q(i,t) \times p(i)$$

où $q(i,t)$ est la quantité de i dans la transaction t, et $p(i)$ est sa valeur unitaire.

— **Utilité d'un itemset** : l'utilité d'un itemset X dans une transaction t est la somme des utilités des items qui le composent :

$$u(X,t) = \sum_{i \in X} u(i,t)$$

L'utilité totale d'un itemset X dans la base de données D est la somme de ses utilités dans toutes les transactions contenant X :

$$u(X) = \sum_{t \in D \wedge X \subseteq t} u(X,t)$$

[17]

— **Seuil d'utilité minimale (minUtil)** : il s'agit d'une valeur fixée par l'utilisateur qui permet de déterminer si un itemset ou une règle possède une utilité suffisante pour être considéré comme intéressant. Un itemset est dit à utilité élevée si son utilité totale dépasse ce seuil.[17]

— **Règle d'association à utilité élevée (HUAR)** : une règle de la forme $X \Rightarrow Y$ est considérée comme à utilité élevée si elle satisfait simultanément trois critères : un support minimal, une confiance minimale, et une utilité totale supérieure à un seuil fixé ($minUtil$). Ce type de règle permet d'identifier des relations à la fois fréquentes et économiquement significatives . [17]

3.7 Application concrète des méthodes HUIM sur les données de Gowalla

Les concepts et techniques présentés dans ce chapitre ne sont pas uniquement d'ordre théorique. Ils constituent les fondements méthodologiques de l'application concrète développée dans le cadre de ce travail de recherche, portant sur les données issues du réseau social géolocalisé *Gowalla*.

Ces données, une fois prétraitées et transformées en un format transactionnel, seront exploitées par des algorithmes d'extraction de motifs à haute utilité, notamment *HUCI-Miner* et *HGB*. L'objectif est d'extraire des règles d'association présentant une forte valeur contextuelle, en tirant parti de la structure spatiale et sociale des interactions au sein de Gowalla.

Cette étude de cas servira ainsi à illustrer la pertinence pratique des méthodes abordées, tout en démontrant leur efficacité dans l'analyse de données géolocalisées à forte valeur ajoutée.

3.8 Conclusion

Dans ce chapitre, nous avons exploré les fondements de la fouille des ensembles d'items à haute utilité (HUIM), en présentant ses concepts clés, ainsi que les mesures associées telles que l'utilité transactionnelle et le seuil minimal d'utilité (minUtil). Nous avons également comparé cette approche aux méthodes classiques de fouille d'itemsets fréquents (FIM), en mettant en évidence leurs limites dans des contextes où la fréquence seule ne suffit pas à évaluer l'importance des éléments.

Ce cadre théorique met en lumière l'intérêt des techniques HUIM dans l'extraction de connaissances à forte valeur ajoutée, et prépare la transition vers l'expérimentation sur des données réelles issues du réseau géolocalisé Gowalla. .

Nous avons également mis en évidence les limites des approches traditionnelles et la nécessité d'intégrer des paramètres de valeur ou d'utilité dans les processus d'extraction de connaissances. Les méthodes HUIM permettent ainsi de découvrir des règles plus pertinentes dans les contextes où la simple fréquence des items ne suffit pas à refléter leur importance.

Cette base conceptuelle, articulée autour des principes du HUIM, prépare ainsi la transition vers l'expérimentation. Dans ce qui suit, nous mettrons en œuvre ces approches sur les données issues du réseau Gowalla, afin de valider leur pertinence dans un contexte réel d'analyse géolocalisée.

Chapitre 4

Réalisation & Interprétation

4.1 Introduction

Après avoir défini les fondements théoriques de l'extraction des règles d'association à haute utilité, et examiné les différentes représentations concises ainsi que l'intérêt analytique des réseaux sociaux géolocalisés, ce chapitre marque le passage à la phase pratique du travail.

Il présente la mise en œuvre expérimentale des algorithmes sélectionnés, à savoir **HUCI-Miner** pour l'extraction des itemsets fermés à haute utilité, et **HGB** pour la génération de règles d'association pertinentes et non redondantes. L'expérimentation est réalisée sur un jeu de données réel, issu de la plateforme géo-sociale *Gowalla*, transformé préalablement en un format transactionnel adapté à l'outil *SPMF*.

Ce chapitre décrit en détail les étapes suivies : depuis la phase de prétraitement des données jusqu'à l'exécution des algorithmes, en passant par la construction des structures d'utilité et la configuration des paramètres d'entrée. Il comprend également l'analyse des résultats obtenus, à travers des représentations graphiques et des tableaux comparatifs, afin d'évaluer la qualité, la compacité et la pertinence des motifs extraits.

L'objectif est d'apprécier concrètement l'efficacité des méthodes proposées dans un environnement réel, et de mettre en évidence l'apport des approches fondées sur l'utilité et la représentation concise dans les tâches de fouille de données transactionnelles.

4.2 Présentation des outils de développement

4.2.1 Eclipse 

Eclipse est une plateforme de développement libre utilisée dans divers projets open source, notamment pour le développement en Java. Une étude a analysé les contributions au code source de plusieurs projets au sein de la fondation Eclipse, montrant l'importance de sa communauté et son architecture modulaire. [21]

4.2.2 Langages de programmation : Java et Python

4.2.2.1 Java 

Java est un langage de programmation orienté objet, robuste et indépendant de la plateforme grâce à la machine virtuelle Java (JVM). Il est largement utilisé pour les applications d'entreprise, le développement mobile (Android), et les systèmes embarqués.[22]

4.2.3 Python 

Python est un langage de programmation interprété, de haut niveau et à la syntaxe claire. Il est très utilisé dans le domaine scientifique, l'analyse de données, l'intelligence artificielle, le développement web, et l'automatisation.[23]

4.2.4 Visual Studio 

Visual Studio est un environnement de développement intégré (IDE) développé par Microsoft. Il prend en charge plusieurs langages comme C++, C#, Python, et JavaScript. Il est apprécié pour son éditeur intelligent, ses outils de débogage, et son intégration avec Git.[24]

4.2.5 SPMF (Sequential Pattern Mining Framework) 

La bibliothèque SPMF (Sequential Pattern Mining Framework) est une plateforme open source développée en Java, spécialisée dans l'extraction de motifs à partir de divers types de données. Elle offre un large éventail d'algorithmes pour la fouille de motifs séquentiels, les itemsets fréquents, les motifs à haute utilité, les règles d'association, les séries temporelles, le clustering et la classification.

Avec plus de 250 algorithmes implémentés, SPMF constitue un outil incontournable pour les chercheurs et praticiens dans le domaine du Data Mining. Grâce à son architecture modulaire, elle peut être intégrée facilement dans des projets Java ou utilisée indépendamment via une interface graphique conviviale ou en ligne de commande.[25]

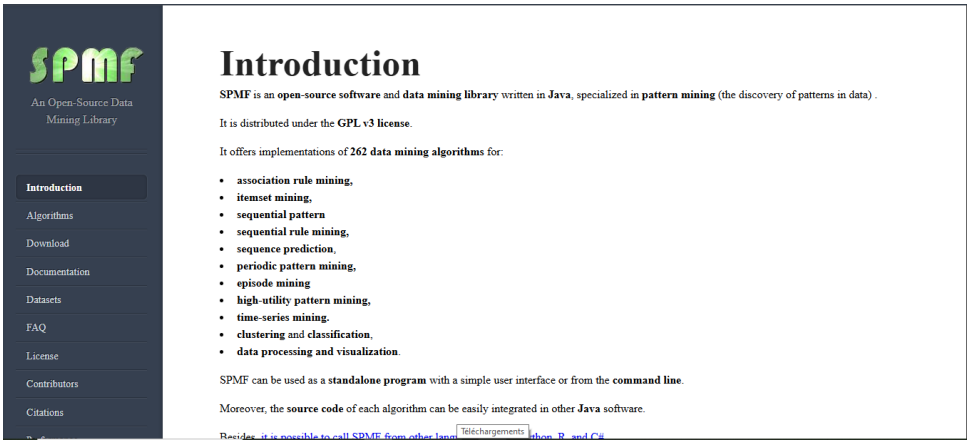


FIGURE 4.1 – Présentation de l'environnement de SPMF

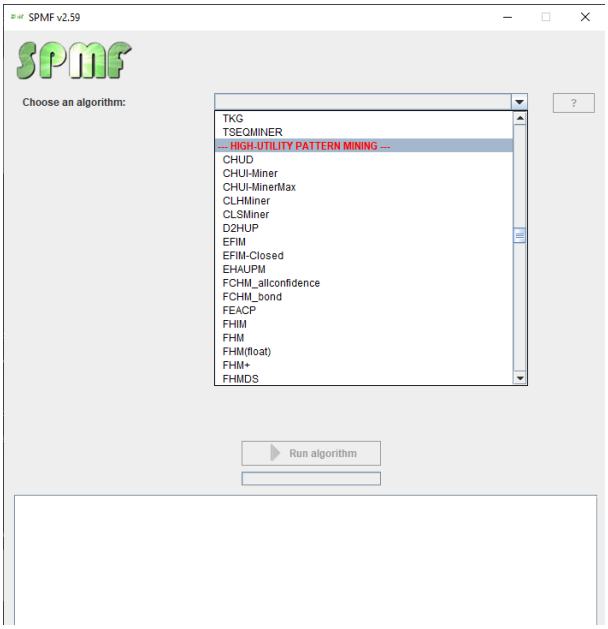


FIGURE 4.2 – Interface graphique de la bibliothèque SPMF

4.3 L'architecture générale de la méthodologie

La méthodologie adoptée dans ce travail repose sur une approche progressive et structurée, visant l'extraction de règles d'association à haute utilité à partir d'une base de données transactionnelle réelle. Cette approche intègre l'utilisation successive de deux algorithmes complémentaires : **HUCI-Miner**,

pour l'identification des itemsets fermés à haute utilité, et **HGB**, pour la génération de règles d'association non redondantes.

L'architecture générale de la démarche se décline en six étapes principales :

beginenumerate

Prétraitement des données : Cette étape commence par l'importation de la base de données *Gowalla*. Elle consiste à nettoyer et réorganiser les données, puis à leur ajouter un nouvel attribut. Ensuite, les données sont exportées sous forme d'un fichier *CSV* compatible avec le format *SPMF*. Un script de transformation est appliqué pour convertir la base en un format transactionnel, prêt pour l'analyse.

Attribution d'utilité : Une fois la base transactionnelle générée, une valeur d'utilité est attribuée à chaque item dans chaque transaction. Cette utilité quantitative constitue une étape préparatoire essentielle pour la suite du processus.

Extraction des itemsets à haute utilité (HUIs) : L'algorithme **HUCI-Miner**, implémenté en Java, est appliqué sur la base transactionnelle enrichie. Il permet d'extraire les ensembles d'items dont l'utilité dépasse un seuil minimal (*minUtil*).

Identification des itemsets fermés à haute utilité (HUCI) : L'algorithme **HUCI-Miner** effectue également une étape de filtrage permettant de ne conserver que les itemsets fermés, c'est-à-dire ceux qui ne peuvent être étendus sans perte de spécificité. Cette opération permet de réduire la redondance dans les résultats.

Génération de règles d'association à haute utilité : L'algorithme **HGB**, également implémenté en Java, est utilisé pour générer des règles d'association à partir des HUCI extraits. Ces règles se distinguent par leur pertinence, leur utilité, et leur non-redondance.

Analyse et évaluation des résultats : Les règles générées sont ensuite analysées selon plusieurs critères : quantité, compacité, et qualité informative. Cette étape permet également d'évaluer les performances globales des algorithmes **HUCI-Miner** et **HGB** appliqués sur des données massives.

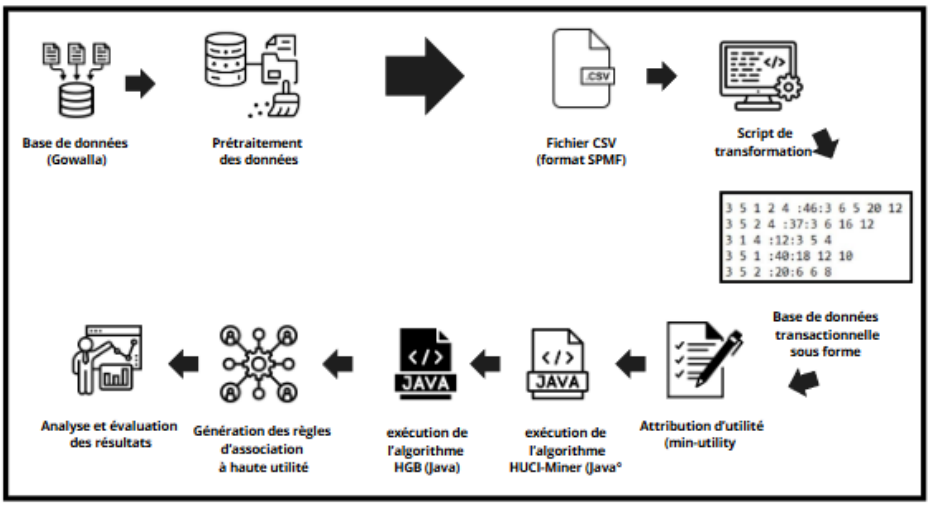


FIGURE 4.3 – L'architecture générale de la méthodologie

4.4 Base de Données Gowalla

4.4.1 Présentation de la Base de Données Gowalla

La base de données **Gowalla** provient d'un réseau social géolocalisé, largement utilisé dans la recherche pour l'analyse des comportements des utilisateurs dans l'espace et le temps. Ces données ont été publiées par le *Stanford Network Analysis Project (SNAP)* et décrites en détail dans l'étude de Cho et al. [26], qui constitue une référence dans le domaine des réseaux sociaux géographiques.

Les caractéristiques principales de cette base sont résumées dans le tableau suivant [27] :

TABLE 4.1 – Caractéristiques essentielles de la base de données Gowalla

Propriété	**Valeur**
Nombre total d'utilisateurs	196 591
Nombre de relations d'amitié	950 327
Nombre de check-ins	6 442 890
Nombre estimé de lieux	~1,2 million
Type de graphe	Non orienté
Période de collecte	Février 2009 – Octobre 2010

Ces données sont disponibles publiquement sur le site officiel de SNAP [27].

Modélisation transactionnelle Pour adapter cette base au contexte de l'extraction à haute utilité, chaque utilisateur est modélisé comme une transaction, et les sites visités (identifiés par

un identifiant unique) sont représentés comme des items. Ainsi, chaque transaction regroupe l'ensemble des check-ins effectués par l'utilisateur.[26]

Utilité externe Chaque item peut se voir attribuer une *utilité externe* (comme la fréquence de visite, la popularité ou la durée moyenne de présence), ce qui rend possible l'application d'algorithmes de fouille des motifs à haute utilité.

Ce cas d'étude constitue un socle concret pour les expérimentations développées dans ce mémoire, notamment dans le cadre de l'extraction de règles d'association à haute utilité avec représentations concises.[26]

4.4.2 Pré-Traitement de la Base

Les fichiers d'origine de la base de données **Gowalla** ne sont pas directement exploitables par la bibliothèque *SPMF*. Par conséquent, un processus de prétraitement s'est avéré nécessaire afin de transformer cette base en une base de données transactionnelle conforme au format accepté par l'outil SPMF.

La base **Gowalla** est constituée de données géo-sociales représentant les enregistrements de visites (check-ins) effectuées par des utilisateurs dans différents lieux. Toutefois, ces données ne suivent pas le modèle transactionnel classique requis pour l'extraction des motifs à haute utilité. Il a donc fallu les restructurer selon les exigences du format *SPMF*.

Le format *SPMF* repose sur un fichier texte, dans lequel chaque ligne représente une transaction effectuée par un utilisateur unique. Chaque ligne est structurée de la manière suivante :

```
{item1 item2 item3 ...} : total utility : {profit1 profit2 profit3
...}
```

Où `{item1 item2 ...}` correspond aux identifiants des lieux visités par un utilisateur donné, `{profit1 profit2 ...}` représente les valeurs d'utilité (par exemple, la fréquence ou la pondération attribuée à chaque lieu), et la `total utility` désigne l'utilité globale de la transaction.

Pour construire cette structure, nous avons développé un script en *Python* permettant de regrouper les check-ins par identifiant utilisateur. Ainsi, chaque ensemble de lieux visités par

un même utilisateur est interprété comme une transaction unique. Les valeurs d'utilité associées à chaque lieu ont été calculées en fonction de leur fréquence d'apparition, ou d'autres critères pertinents selon le contexte d'analyse.

Ce processus a permis de générer un fichier transactionnel conforme au format *SPMF*, prêt à être exploité par les algorithmes d'extraction de motifs à haute utilité. Voici un aperçu de la base après la transformation en format SPMF :

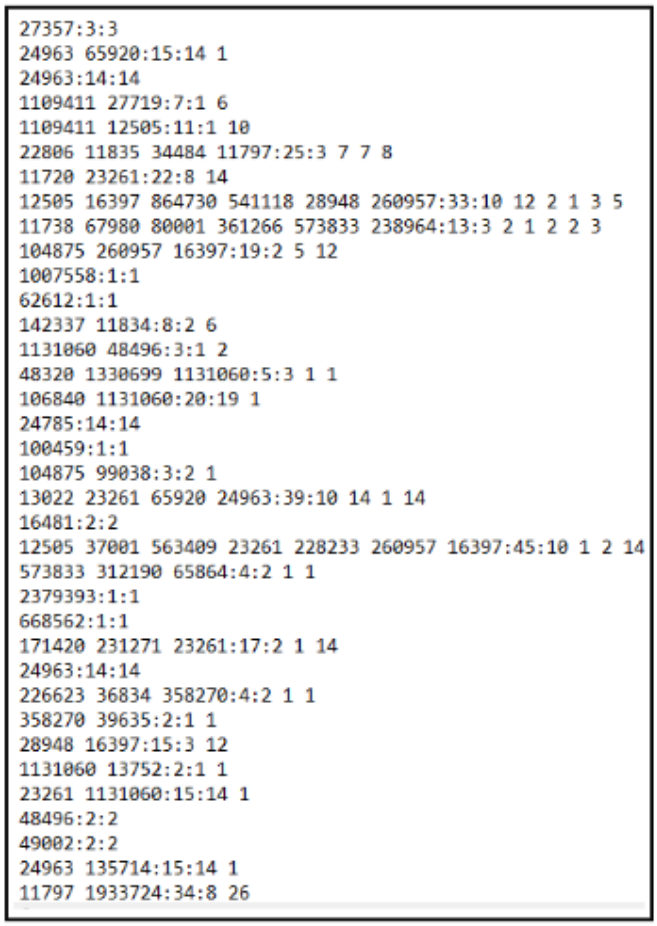


FIGURE 4.4 – Un extrait du fichier SPMF généré à partir de la base de données Gowalla

À ce stade, nous allons appliquer les algorithmes **HUCI-Miner** et **HGB** sur la base de données **Gowalla**, puis procéder à une analyse approfondie des résultats obtenus afin d'en dégager des règles d'association à haute utilité pertinentes et non redondantes.

4.5 Les algorithmes utilisés

4.5.1 l'algorithme HUCI-Miner

L'algorithme **HUCI-Miner** a été conçu pour extraire efficacement les ensembles d'items fermés à haute utilité (HUCI) à partir de bases de données transactionnelles. Contrairement aux approches traditionnelles qui génèrent d'abord tous les motifs utiles avant d'identifier les fermés, HUCI-Miner intègre ces étapes en une seule procédure optimisée.

Cette méthode repose sur une structure de données appelée *EU-list* (Extended Utility List), qui enregistre les informations d'utilité de chaque item dans les transactions. Grâce à cette structure, l'algorithme peut réduire le nombre de parcours de la base et éviter la génération de motifs redondants.

En appliquant des stratégies de combinaison contrôlée des listes d'utilité et de vérification de la fermeture, HUCI-Miner identifie uniquement les motifs significatifs qui dépassent un seuil d'utilité défini (*minUtil*). Cela permet de produire un ensemble de motifs compacts, sans perte d'information, facilitant ainsi l'extraction de règles d'association pertinentes.[28] L'algorithme complet est présenté ci-dessous à travers un pseudo-code illustrant les différentes étapes de traitement :

Algorithm 1 HUCI-Miner($\mathscr{D}$, min_util)

Input: Transaction dataset $\mathscr{D}$, minimum utility threshold *min_util*

Output: High utility closed itemsets with their generators

Procedure HUCI_Miner($\mathscr{D}$, min_util)

$CH \leftarrow \emptyset, HG \leftarrow \emptyset$ $H \leftarrow FHIM(\mathscr{D}, \text{min_util}) \leftarrow \{H_1, H_2, \ldots, H_{max}\}$

foreach $h \in H_1$ **do**

    $h.closed \leftarrow$ true $h.key \leftarrow$ true

for $k = 2$ *to max* **do**

    if $H_k \neq \emptyset$ **then**

        foreach $h \in H_k$ **do**

            $h.key \leftarrow$ true, $h.closed \leftarrow$ true **foreach** *subset* $h' \in H_{k-1}$ *of* h **do**

                if $supp(h') = supp(h)$ **then**

                    $h.key \leftarrow$ false $h'.closed \leftarrow$ false

        $CH_{k-1} \leftarrow \{h \in H_{k-1} \mid h.closed = \text{true}\}$ **Call** Get_generators(CH_{k-1}, H_k)

$CH_{max} \leftarrow H_{max}$ **Call** Get_generators(CH_{max})

$CH \leftarrow \bigcup_k CH_k$ with generators Compute utility unit array for each closed itemset

return *CH with associated generators*

4.5.2 l'algorithme HGB

L'algorithme **HGB** (High Utility Generators Based) a été conçu pour extraire des règles d'association à haute utilité tout en réduisant la redondance des résultats. Contrairement aux méthodes classiques qui produisent un grand volume de règles, HGB s'appuie sur l'utilisation conjointe des itemsets fermés à haute utilité (HUCI) et de leurs générateurs minimaux afin de générer uniquement les règles pertinentes et non redondantes.

Le processus consiste à identifier, pour chaque HUCI, l'ensemble des générateurs associés. Une règle potentielle est construite en comparant un générateur avec son HUCI correspondant. Cette règle est conservée uniquement si elle satisfait deux conditions : dépasser un seuil d'utilité minimale (*minUtil*) et atteindre un niveau suffisant de confiance utilitaire (*utility-confidence*).

Cette stratégie permet de produire un ensemble concis et interprétable de règles d'associa-

tion, tout en maintenant une qualité informative élevée, ce qui est particulièrement utile pour l'analyse de grandes bases de données transactionnelles.

Algorithm 2 HGB Construction Algorithm

Input: $HUCI$: Set of high utility closed itemsets with utility unit arrays ;

min_uconf : Minimum utility-confidence ;

min_util : Minimum utility threshold

Output: HGB : High Utility Generic Basis

foreach *itemset* $h \in HUCI$ **do**

  $L_{ma} \leftarrow \emptyset$; `// Minimal antecedents`

  foreach *subset* $h' \subseteq h$ *in increasing order of size* **do**

    $L_{temp} \leftarrow \emptyset$ **foreach** $g \in HG_{h'}$ *and* $g \not\subseteq h$ **do**

      if $\frac{lu_V(g,h)}{u(g)} \geq min_uconf$ ***and*** $g_s \in L_{ma}$ *where* $g_s \subseteq g$ **then**

        $L_{ma} \leftarrow L_{ma} \cup \{g\}$

      else

        $L_{temp} \leftarrow L_{temp} \cup \{g\}$

    foreach $g \in L_{temp}$ **do**

      $A_1 \leftarrow \{i_1, i_2, \ldots, i_k\}$ where each $i_j \in h' \setminus g$

      for $j = 1$ *to* k *and* $A_j \neq \emptyset$ **do**

        foreach $l \in A_j$ **do**

          if $\frac{lu_V(g,h)}{u(g|h)} \geq min_uconf$ ***and*** $g_s \in L_{ma}$ *where* $g_s \subseteq \{g \cup l\}$ **then**

            Remove all $l' \subseteq \{g \cup l\}$ from L_{ma} $L_{ma} \leftarrow L_{ma} \cup \{g \cup l\}$

        $A_{j+1} \leftarrow \text{Apriorigen}(A_j, min_util)$; `// No pruning applied`

Algorithm 3 HGB Construction Algorithm

foreach $g_s \in L_{ma}$ **do**

  Create rule $R : g_s \rightarrow h \setminus g_s$ Set $R.utility \leftarrow h.utility$ Set $R.uconf \leftarrow \frac{lu_V(g,h)}{u(g_s)}$ $HGB \leftarrow HGB \cup \{R\}$

return HGB

4.6 Exécution expérimentale des algorithmes sur la base Gowalla

4.6.1 Exécution de l'algorithme HUCI-Miner sur les données Gowalla

L'algorithme **HUCI-Miner** (High Utility Closed Itemset Miner) a été appliqué à la base de données géo-sociale **Gowalla**, après sa conversion au format transactionnel compatible avec l'outil **SPMF v2.59**. Cet algorithme vise à extraire uniquement les *itemsets fermés à haute utilité*, c'est-à-dire des ensembles d'items qui ne sont contenus dans aucun autre ensemble avec la même utilité. Ce filtrage permet d'obtenir une représentation concise et sans redondance, facilitant ainsi la génération de règles pertinentes.

L'exécution a été réalisée via l'interface graphique de SPMF en sélectionnant le fichier de données transformé, un fichier de sortie, et en fixant une valeur pour le *minUtil*. À la fin du traitement, les résultats ont été enregistrés automatiquement dans un fichier texte contenant les HUCI extraits.

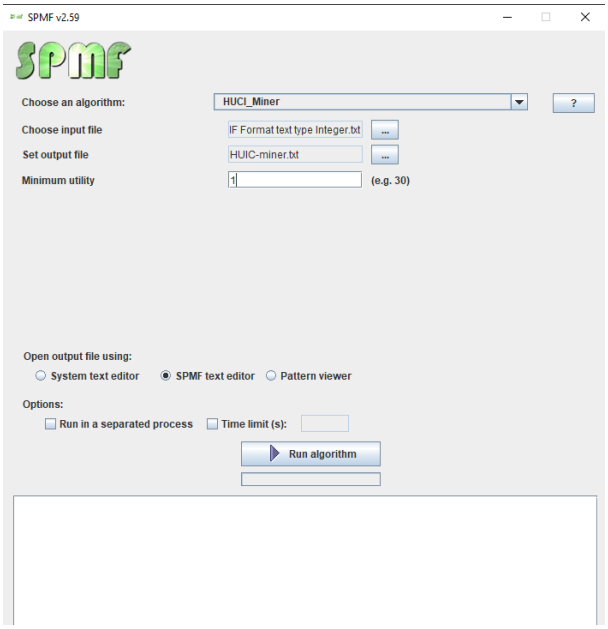


FIGURE 4.5 – Exécution d'algorithme UCI-Miner sur la base GOwalla

4.6.2 Exécution de l'algorithme HGB après HUCI-Miner

Après l'extraction des itemsets fermés à haute utilité, l'algorithme **HGB** (High-utility Generator-based) a été appliqué afin de générer un ensemble complet, concis et non redondant de règles

d'association à haute utilité.

L'algorithme utilise comme entrée le fichier HUCI obtenu précédemment, auquel s'ajoute la définition d'un *seuil de confiance utilitaire minimale (minUConf)*. SPMF permet alors de générer des règles sous forme `antécédent  conséquent`, accompagnées de leurs mesures d'utilité (`UTIL`), d'utilité de l'antécédent (`AUTIL`) et de confiance utilitaire (`UCONF`).

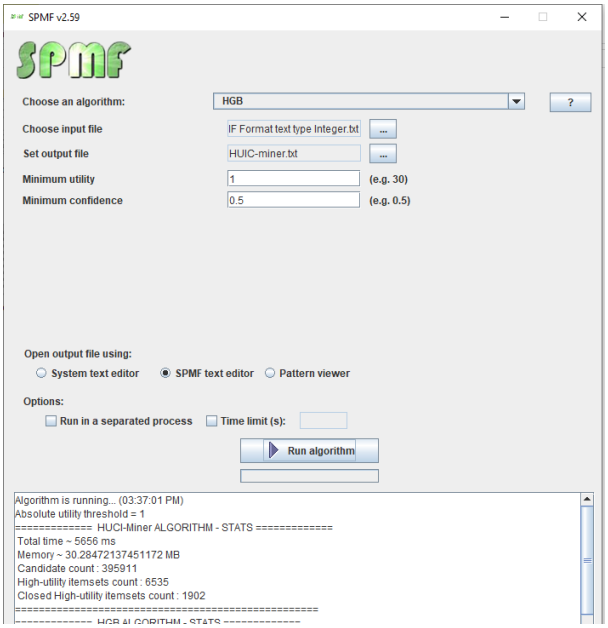


FIGURE 4.6 – Exécution d'algorithme HGB avec la base GOwalla

4.6.3 Résultats obtenus

Après l'extraction des itemsets fermés à haute utilité à l'aide de l'algorithme **HUCI-Miner**, l'algorithme **HGB** a été appliqué afin de générer un ensemble complet, concis et non redondant de règles d'association à haute utilité.

Cette approche repose sur les représentations fermées extraites précédemment et nécessite la définition d'un seuil de confiance utilitaire minimale (*minUConf*), ajusté en fonction des caractéristiques de la base de données.

Les résultats obtenus sont présentés dans le tableau suivant, qui illustre plusieurs exemples de règles générées, accompagnées de leurs mesures d'utilité et de confiance utilitaire :

Règle	**Utilité**	**Confiance utilitaire**
Rockefeller Center ⇒ Times Square	2352	0.72
Hard Rock Café ⇒ Times Square	948	0.65
Disney Store ⇒ Rockefeller Center	516	0.48
Trailer Park Lounge ⇒ Amy Ruth's	2	1.0
DuMont Restaurant ⇒ Roberta's	2	1.0
Starbucks ⇒ Jamaica LIRR Station	2	1.0
Sahadi Importing Co. ⇒ The Bell House	2	1.0
Crosby Street Hotel ⇒ Dean & Deluca	2	1.0
Press 195 ⇒ Neil Simon Theatre	2	1.0
Gordon Ramsay at The London ⇒ Bouchon Bakery	2	1.0
Del Posto ⇒ Google	2	1.0
Ed's Lobster Bar ⇒ Calexico Cart	2	1.0
Sweet Afton ⇒ Brooklyn Label	2	1.0
Ozzie's Coffee & Tea ⇒ IKEA	2	1.0
Mug's Ale House ⇒ T.B.D.	2	1.0
Hell's Kitchen ⇒ Mesa Grill	2	1.0
Local 138 ⇒ The Rusty Knot	2	1.0
Yola's Cafe ⇒ Soho House New York	2	1.0
T.B.D. ⇒ Mug's Ale House	2	1.0
Robin des Bois ⇒ Bar Tabac	2	1.0
Esca ⇒ St James Theater	2	1.0
675 Bar ⇒ Bill's Bar & Burger	2	1.0
Salon De Ning ⇒ Upstairs at the Kimberly Hotel	2	1.0
7 Eleven ⇒ Friend's Tavern	2	1.0

TABLE 4.2 – Exemples représentatifs de résultats générés par l'algorithme HGB à partir des motifs HUCI-Miner

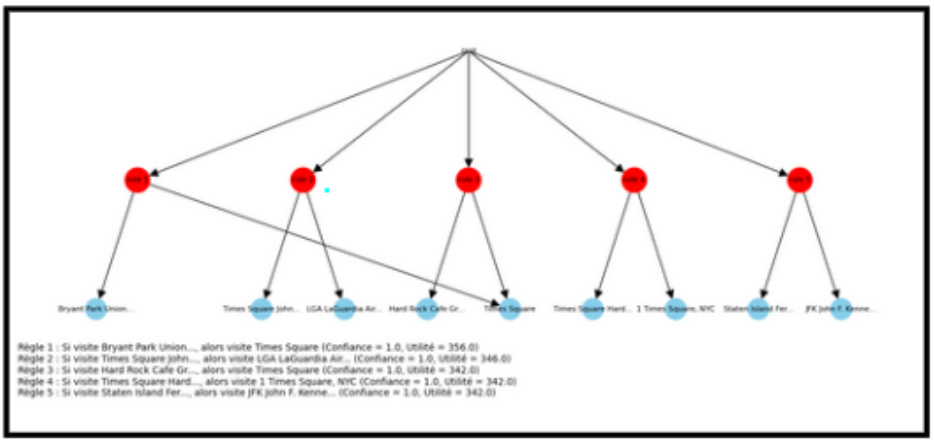


FIGURE 4.7 – Arber bipartite des 5 régles avec confiance

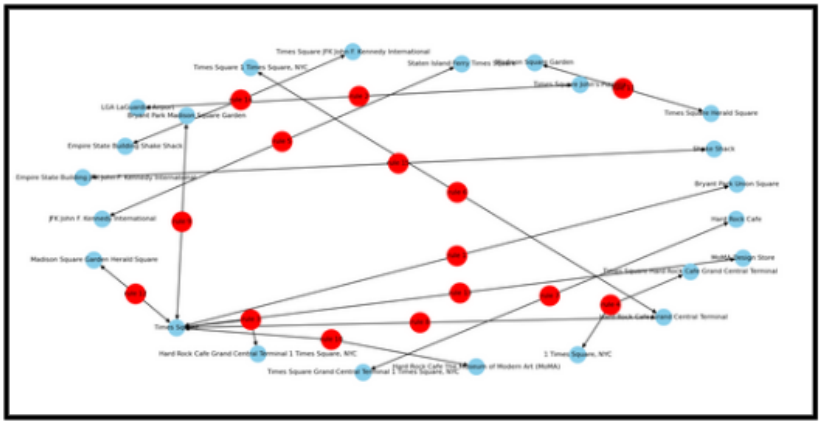


FIGURE 4.8 – Visualisations en graphe orienté des règles d'association à haute utilité

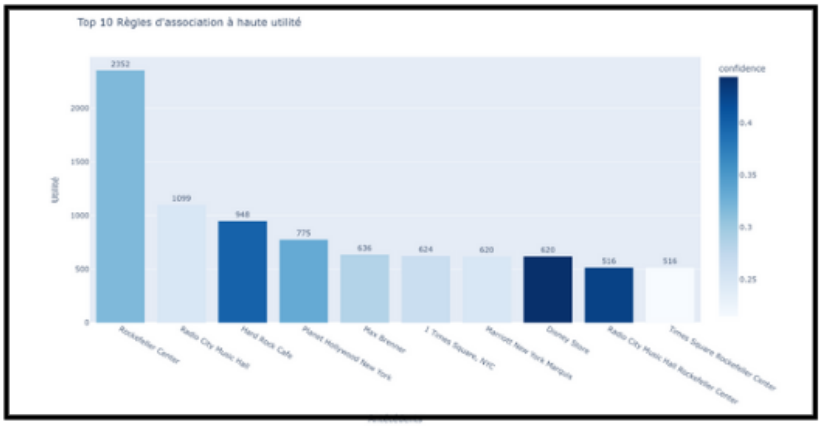


FIGURE 4.9 – Top régles d'association à haute utilité

4.7 Conclusion

Ce chapitre a constitué le cœur pratique de notre étude, en concrétisant les concepts théoriques abordés dans les chapitres précédents à travers une mise en œuvre expérimentale sur des données réelles.

Nous avons présenté les différentes étapes nécessaires pour transformer la base de données **Gowalla** en un format transactionnel exploitable par l'outil *SPMF*, ce qui a permis l'application efficace des algorithmes **HUCI-Miner** et **HGB**. Bien que complexe, la phase de prétraitement a joué un rôle fondamental dans la structuration des données et leur adéquation aux objectifs analytiques de ce travail.

L'exécution des algorithmes a permis d'extraire des itemsets fermés à haute utilité ainsi que des règles d'association pertinentes, démontrant la performance de l'approche combinée HUCI-Miner / HGB pour produire des motifs informatifs, compacts et non redondants.

Ainsi, ce travail expérimental apporte une validation concrète des choix méthodologiques adoptés, et illustre la faisabilité et la pertinence de l'extraction de règles à haute utilité à partir de données complexes issues de réseaux sociaux géolocalisés.

Conclusion générale

À l'ère de l'explosion des données, la capacité à extraire des connaissances utiles et exploitables constitue un enjeu stratégique pour de nombreux domaines. Ce mémoire s'est inscrit dans cette problématique en explorant l'extraction des règles d'association à haute utilité, une approche qui vise à dépasser les limites des techniques classiques en intégrant la notion d'importance économique ou stratégique des items.

L'étude a d'abord établi les fondements théoriques de cette problématique, en présentant les concepts clés de la fouille de données et des itemsets à haute utilité. L'accent a ensuite été mis sur les représentations concises telles que les itemsets fermés, qui permettent une réduction efficace du volume des motifs tout en conservant leur valeur informative.

L'une des originalités de ce travail réside dans l'intégration des *réseaux sociaux géolocalisés (LBSNs)* en tant que source de données complexes. En exploitant les données issues de la base *Gowalla*, nous avons pu démontrer la pertinence de ces environnements pour la fouille spatio-sociale à haute utilité. Une étape de prétraitement a permis de transformer cette base en un format transactionnel compatible avec l'outil *SPMF*, facilitant ainsi l'application des algorithmes.

Deux algorithmes ont été sélectionnés et implémentés pour mener l'expérimentation : **HUCI-Miner**, pour l'extraction des itemsets fermés à haute utilité, et **HGB**, pour la génération de règles d'association non redondantes. L'analyse des résultats a montré que cette combinaison permet d'obtenir un ensemble de règles pertinentes, compactes et informatives, tout en maîtrisant la complexité computationnelle.

Les résultats obtenus confirment que l'intégration conjointe de la notion d'utilité, des représentations concises et des données issues des LBSNs constitue une approche prometteuse

pour améliorer l'efficacité et la qualité des systèmes de découverte de connaissances. Ce travail ouvre ainsi la voie à des perspectives futures intéressantes, notamment l'extension vers des données temporelles, l'intégration de critères multicritères ou encore l'optimisation de la performance à grande échelle.

Bibliographie

[1] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, “Knowledge discovery in databases : An overview,” *AI magazine*, vol. 13, no. 3, pp. 57–57, 1992.

[2] J. Han, J. Pei, and H. Tong, *Data mining : concepts and techniques*. Morgan kaufmann, 2022.

[3] G. Piatetsky-Shapiro, “Data mining and knowledge discovery 1996 to 2005 : overcoming the hype and moving from “university” to “business” and “analytics”,” *Data Mining and Knowledge Discovery*, vol. 15, no. 1, pp. 99–105, 2007.

[4] D. J. Hand, “Principles of data mining,” *Drug safety*, vol. 30, pp. 621–622, 2007.

[5] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, “Knowledge discovery in databases : An overview,” *AI magazine*, vol. 13, no. 3, pp. 57–57, 1992.

[6] D. T. Larose and C. D. Larose, *Discovering knowledge in data : an introduction to data mining*. John Wiley & Sons, 2014.

[7] M. J. Berry and G. S. Linoff, *Data mining techniques*. John Wiley & Sons, 2009.

[8] S. Scellato, A. Noulas, and C. Mascolo, “Exploiting place features in link prediction on location-based social networks,” in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2011, pp. 1046–1054.

[9] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, “Collaborative location and activity recommendations with gps history data,” in *Proceedings of the 19th International Conference on World Wide Web (WWW)*. ACM, 2010, pp. 1029–1038.

[10] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility : User movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2011, pp. 1082–1090.

[11] J. Bao, Y. Zheng, and M. F. Mokbel, "Location-based and preference-aware recommendation using sparse geo-social networking data," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*. ACM, 2012, pp. 199–208.

[12] H. Gao, J. Tang, and H. Liu, "gscorr : Modeling geo-social correlations for new check-ins on location-based social networks," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2012, pp. 1582–1586.

[13] B. Liu and W.-C. Lee, "A hybrid classification model for social network analysis," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1391–1397, 2010.

[14] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD)*. AAAI Press, 1997, pp. 283–286.

[15] R. S. Agrawal and R. Srikant, "R. fast algorithms for mining association rules," in *Proceedings of the 20th international conference on very large data bases, vldb*, 1994, pp. 487–499.

[16] J. M. Luna, P. Fournier-Viger, and S. Ventura, "Frequent itemset mining : A 25 years review," *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, vol. 9, no. 6, p. e1329, 2019.

[17] P. Fournier-Viger, J. C.-W. Lin, B. Vo, T. T. Chi, J. Zhang, and H. B. Le, "A survey of itemset mining," *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, vol. 7, no. 4, p. e1207, 2017.

[18] B. Goethals, "A priori property and breadth-first search algorithms," in *Encyclopedia of database systems/Liu, Ling [edit.]; et al.[edit.]*, 2009, pp. 124–127.

[19] M. J. Zaki, “Scalable algorithms for association mining,” *IEEE transactions on knowledge and data engineering*, vol. 12, no. 3, pp. 372–390, 2000.

[20] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” *ACM sigmod record*, vol. 29, no. 2, pp. 1–12, 2000.

[21] Q. C. Taylor, J. L. Krein, A. C. MacLean, and C. D. Knutson, “An analysis of author contribution patterns in Eclipse Foundation project source code,” in *IFIP International Conference on Open Source Systems*. Springer, 2011, pp. 269–281.

[22] Oracle Corporation, “Java Technologies – Oracle,” https://www.oracle.com/java/technologies/, 2025, accessed : 2025-05-27.

[23] Python Software Foundation, “About Python – Official Website,” https://www.python.org/about/, 2025, accessed : 2025-05-27.

[24] Microsoft Corporation, “Visual Studio – Code Editor and IDE,” https://visualstudio.microsoft.com/, 2025, accessed : 2025-05-27.

[25] Philippe Fournier-Viger, “SPMF : An Open-Source Data Mining Library,” http://www.philippe-fournier-viger.com/spmf/, 2025, accessed : 2025-05-27.

[26] E. Cho, S. A. Myers, and J. Leskovec, “Friendship and mobility : user movement in location-based social networks,” in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2011, pp. 1082–1090. [Online]. Available : https://snap.stanford.edu/data/loc-gowalla.html

[27] J. Leskovec and A. Krevl, “Snap datasets : Gowalla social network,” https://snap.stanford.edu/data/loc-gowalla.html, 2014, stanford Network Analysis Project (SNAP).

[28] L. K. Sahoo, M. R. Patra, and A. K. Rath, “Huci-miner : An efficient algorithm for mining high utility closed itemsets,” *Expert Systems with Applications*, vol. 42, no. 21, pp. 8033–8045, 2015.