République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed El Bachir El Ibrahimi - Bordj Bou Arréridj
Faculté des Mathématiques et d'informatique

Département d'informatique

MEMOIRE

Présenté en vue de l'obtention du diplôme

Master en informatique

Spécialité : **informatique Décisionnelle**

THEME

Enhanced-SPAM pour l'extraction des motifs séquentiels

Présenté par :

OUCHENE Sofiane

Devant le jury composé de :

Présidente :	*Mme. BENABIDE Sonia*
Examinateur :	*Mr. SAHA Adel*
Encadreur :	*Mr MOUSSAOUI Boubakeur*

Promotion : 2020/2021

REMERCIEMENTS

Après avoir rendu grâce à Allah le tout puissant et le miséricordieux, nous tenons à remercier :

Tout d'abord, ce travail ne serait pas aussi riche et n'aurait â pu avoir le jour sans l'aide et l'encadrement de Mr M. Boubaker, on le remercie pour la qualité de son encadrement, pour sa patience, sa rigueur et sa disponibilité durant notre préparation de ce mémoire.

Notre remerciement s'adresse également à tous nos professeurs pour leurs générosités et la grande patience dont ils ont su faire preuve malgré leurs charges académiques et professionnelles

Promotion : 2020/2021

DÉDICACE S

Je dédie entièrement ce travail à mon père et à ma mère, mes piliers, mes exemples, mes premiers supporteurs et ma plus grande force. Merci pour votre présence, votre soutien, votre aide financière, et surtout votre amour, merci de n'avoir jamais douté de moi. Tout ce que j'espère, c'est que vous soyez fiers de moi aujourd'hui.

À mon cher frère Tarek et à mes meilleures sœurs Maroua & Safa, qui font de mon univers une merveille, je leurs souhaite beaucoup de bonheur et de réussite.

A ma collègue Ichrak qui depuis des années m'encourage, me comprend et a toujours été à mes côtés, que dieu lui donne du bonheur, santé et réussite.

A tous les, les voisins et les amis que j'ai connu jusqu'à maintenant. Merci pour leurs amours et leurs encouragements.

OUCHENE Sofiane

Résumé :

Le data mining désigne le processus d'analyse de volumes massifs de données et du Big Data sous différents angles afin d'identifier des relations entre les data et de les transformer en informations exploitables. Ce mémoire présente la nouvelle approche basée sur l'algorithme SPAM qu'utilise une méthode de parcours en profondeur pour le but d'optimisé le temps d'exécution et la consommation de la mémoire.

Mots-clés : fouille de données, motifs séquentiels, extraction des motifs séquentiel, SPAM

Abstract:

Data mining refers to the process of analyzing massive volumes of data and Big Data from different angles in order to identify relationships between data and transform them into actionable information. This paper presents a new approach based on the SPAM algorithm that uses a deep traversal method to optimize execution time and memory consumption.

Keywords: data mining, sequential patterns, sequential patterns mining, SPAM

ملخص :

التنقيب عن البيانات هو عملية تحليل كميات هائلة من البيانات والبيانات الضخمة من زوايا مختلفة لتحديد العلاقات بين البيانات وتحويلها إلى رؤى قابلة للتنفيذ. تقدم هذه البحث النهج الجديد القائم على خوارزمية (SPAM) التي تستخدم طريقة المسح العميق لغرض تحسين وقت التنفيذ واستهلاك الذاكرة.

الكلمات المفتاحية: التنقيب عن البيانات ، الأنماط المتسلسلة ، الاستخراج الأنماط المتسلسلة ، الرسائل الاقتحامية (SPAM)

Table des matières :

Chapiter02 : *Extraction de motifs séquentiels*

Chapiter03 :*Contribution*

Liste des figures :

List des Tableaux :

Abbreviation

OLTP	*OnLine Transaction Processing*
WWW	*World Wide Web*
OLAP	*OnLine Analytical Processing*
CSV	*Comma-Separated Values*
API	*Application Programming Interface*
SQL	*Structured Query Language*
ROLAP	*Relational OnLine Analytical Processing*
HTML	*HyperText Markup Language*
ECBD	*Extraction des Connaissance a partire des Base des Donnée*
K-PPV	*K- Plus Proches Voisins*
K-NN	*K-Nearest Neighbors*
GRC	*Gestion de la Relation Client*
CRM	*Customer Relationship management*
SPM	*Sequential Pattern Mining*
SPADE	*Sequential PAttern Discovery using Equivalence classes*
SDB	*Sequence DataBase*
SID	*Sequence ID*
TID	Transaction ID
BFS	*Breadth-First Search*
DFS	*Depth-first search*
GSP	*Generalized Sequential Pattern*
PREFIXSPAN	*Prefix-Projected Sequential pattern Mining*
SPAM	*Sequential PAttern Mining*
CloSpan	*Closed sequential patterns in large datasets*
E-SPAM	*Enhanced Sequential PAttern Mining*
JVM	*Java virtual machine*
SPMF	*Sequential Pattern Mining Framework*

Introduction générale :

La numérisation croissante de nos activités, la capacité sans cesse accrue à stocker des données numériques, l'accumulation d'informations en tous genres qui en découle, génère un nouveau secteur d'activité qui a pour objet l'analyse de ces grandes quantités de données. Sont alors apparues de nouvelles approches, de nouvelles méthodes, de nouveaux savoirs et in fine sans doute, de nouvelles manières de penser et de travailler.

La question qui se pose est alors : Comment peut-on extraire les informations cachées au sein de grandes bases de données ? La puissance de calcul des ordinateurs actuels et la baisse des coûts de stockage laissent prédire que nous disposons des moyens physiques pour le faire. Le problème réside alors au niveau logiciel. En effet, les bases de données classiques ne sont plus de taille à faire face à l'analyse de telles informations et c'est grâce à ce besoin pressant que sont apparues les techniques d'extraction de connaissances à partir des données communément connues sous le nom de « Data Mining ». Citons parmi ces techniques : Les réseaux de neurones, Les arbres de décision, les algorithmes génétiques, Règles d'association et l'extraction des motifs séquentiel

L'extraction des motifs séquentiel est une technique qui fait pour le but de trouver toutes les séquences d'items apparaissant avec une certaine certitude dans une base de données selon une mesure d'intérêt choisie par l'utilisateur. Il s'agit donc, de construire l'ensemble de tous ces motifs intéressants appelés motifs séquentiels.

Notre travail consiste à implémenter une nouvelle approche E-SPAM une version récursive de l'algorithme SPAM, dans un deuxième temps, à évaluer et à comparer les performances de notre approche E-SPAM et SPAM en termes de temps et consommation de la mémoire selon des différent paramètres.

Introduction générale :

La structure de ce mémoire va comme suit :

Dans le chapitre 1 nous avons présenté les fondamentaux du data mining et nous expliquerons les taches de data mining, les déférents technique et les domaines d'application de data mining.

En suite dans le chapitre 2 nous expliquerons le domaine d'extraction des motifs séquentiel, et les déférent catégories des algorithmes de ce domaine et quelque algorithme dans chaque catégorie.

Et enfin dans le chapitre 3 nous avons présenté de notre nouvelle approche, les outils de développement utilisé, résultat d'exécution des deux algorithmes et discussion sur ces résultats.

Chapitre 1 :

Fouille de données

« Data Mining »

1.1. Introduction :

L'exploration de données, connue aussi sous l'expression de fouille de données, forage de données, prospection de données, data mining, ou encore extraction de connaissances à partir de données à l'aide de méthodes automatiques ou semi-automatiques utilisant un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances utiles à l'entreprise.[1]

1.2. Définition du Data Mining :

Les logiciels Data Mining font partie des outils analytiques utilisés pour l'analyse de données. Ils permettent aux utilisateurs d'analyser des données sous différents angles, de les catégoriser, et de résumer les relations identifiées. Techniquement, le Data Mining est le procédé permettant de trouver des corrélations ou des patterns entre de nombreuses bases de données relationnelles**.**

Le Data Mining repose sur des algorithmes complexes et sophistiqués permettant de segmenter les données et d'évaluer les probabilités futures**. [2]**

1.3. Motivation :

Ces dernières années, le data mining a attiré beaucoup d'attention dans l'industrie de l'information, principalement en raison de la disponibilité de grandes quantités de données et de la nécessité de les convertir en connaissances utiles, en particulier dans les applications liées à l'analyse de données, Marketing, détection des fraudes, fidélisation de la clientèle, contrôle de la production et les recherches scientifiques

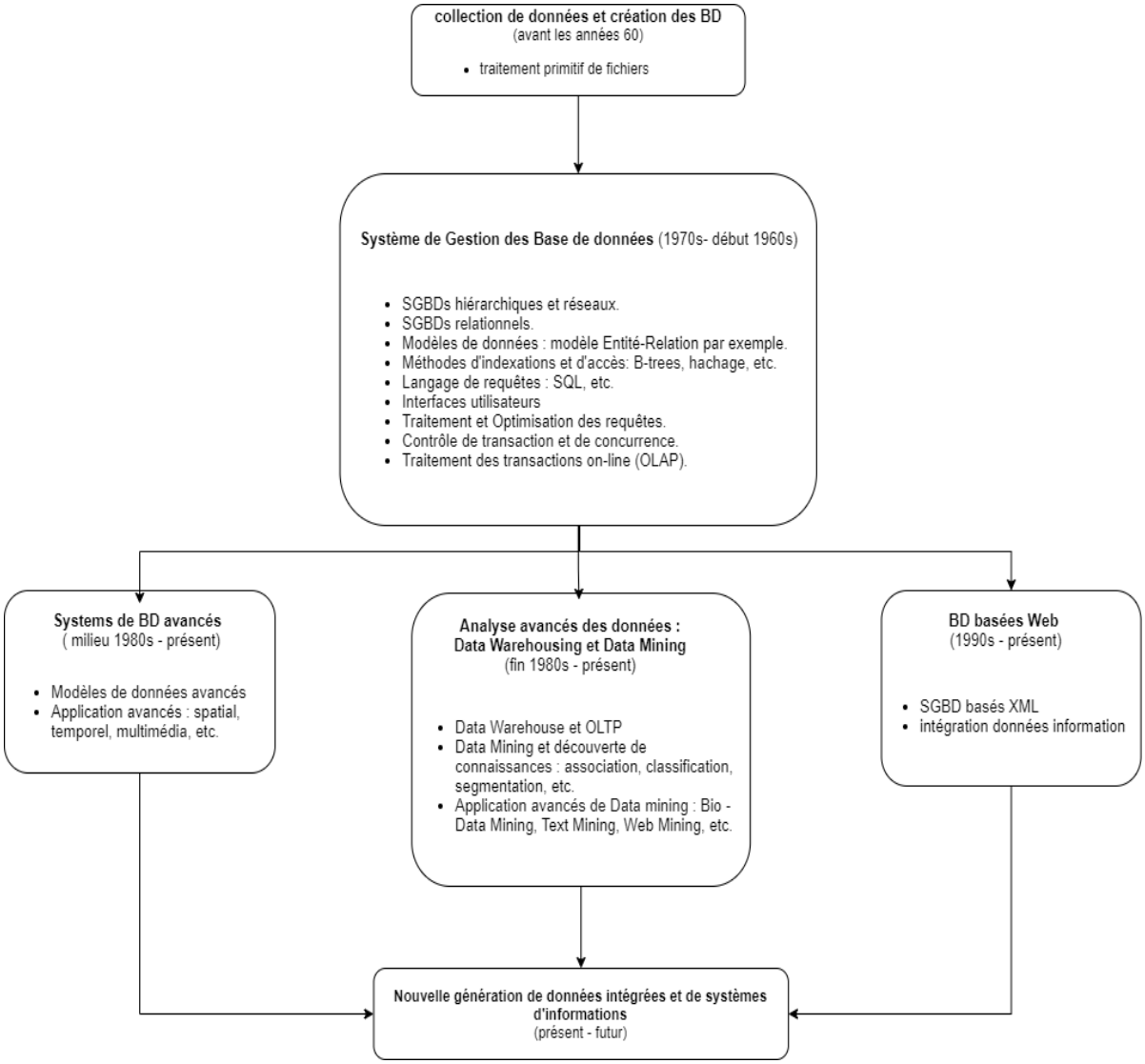


Figure 1- 1: Développement naturel de la technologie de l'information

L'énorme quantité de données collectées et stockées dans des bases de données nombreuses et volumineuses est décrite comme “Riche en données mais Pauvre en informations”, La croissance rapide et la grande quantité de données qui ont été collectées et stockées dépassent notre capacité humaine à les comprendre sans utiliser d'outils, se problème a conduit au développement d'outils d'exploration de données qui effectuent des analyses de données pour convertir des données énormes en connaissances utiles.[3] [4]

La technologie des systèmes experts modernes repose généralement sur des utilisateurs du domaine ou des experts pour saisir manuellement les connaissances dans la base de données de connaissances. Malheureusement, ce processus est incorrect et prend du temps. Les outils du data mining qui effectuent l'analyse des données peuvent identifier les modèles de données

clés qui apportent des contributions significatives à la stratégie business, à la base de connaissances et à la recherche scientifique et médicale. L'écart de plus en plus grand entre les données et les informations nécessite le développement d'une utilisation systématique des outils du data mining, transformant ainsi les Cimetière de données en pépites d'or de connaissances**. [3]** [4]

1.4. Data mining sur quels types de données :

1.4.1. Les fichiers plats :

Les fichiers plats en data mining sont définis comme des fichiers de données au format texte ou binaire, dont la structure peut être facilement extraite à l'aide d'algorithmes d'exploration de données, ce qui le rend facile à utiliser (et limité à l'utilisation). Ce type de données est principalement utilisé pour transmettre des informations. Entre plusieurs serveurs.[12]

Exemple : fichier CSV.

1.4.2. Les bases de données relationnelles :

Une base de données relationnelle est définie comme la collecte de données organisées en tableaux avec des lignes et des colonnes. Le schéma physique des bases de données relationnelles est un schéma qui définit la structure des tables. Le schéma logique dans les bases de données relationnelles est un schéma qui définit la relation entre les tables. L'API standard des bases de données relationnelles est le SQL.

Dans le data mining, des modèles structurés tels que les modèles ROLAP sont utilisés pour créer des bases de données relationnelles.

1.4.3. Les datawarehouses :

Le datawarehouse ou entrepôt de données est définies comme un endroit où les données sont intégrées et collectées à partir de plusieurs sources afin que les requêtes et les solutions puissent être exécutées. Il existe trois types d'entrepôts de données :

- Les entrepôts de données d'entreprise,
- Les data marts
- Les entrepôts virtuels.

Vous pouvez utiliser deux méthodes pour mettre à jour les données dans l'entrepôt de données : une méthode basée sur les requêtes et une méthode basée sur la mise à jour. Ces méthodes sont très utiles pour vous aider à prendre des décisions commerciales.

1.4.4. Les bases de données transactionnelles :

Les bases de données transactionnelles sont un ensemble de données organisées par horodatage, date, etc., représentant les transactions dans la base de données. Ce type de base de données à la capacité de revenir en arrière ou d'annuler son fonctionnement lorsqu'une transaction n'est pas terminée ou engagée.

Il s'agit donc d'un système très flexible où les utilisateurs peuvent modifier les informations sans changer les informations sensibles.

Les bases de données de transactions sont principalement utilisées dans les banques ou les systèmes distribués, les bases de données d'objets, etc.

1.4.5. Les de données multimédia :

La base de données multimédia est composée de supports audios, vidéo, image et texte. Ils peuvent être stockés dans une base de données orientée objet. Ils sont utilisés pour stocker des informations complexes dans un format prédéfini. Les bases de données multimédias sont utilisées dans les formats de bibliothèques numériques, les services de vidéo à la demande ou les bases de données musicales (telles que Spotify, etc.).

1.4.6. Le world wide web :

WWW fait référence au World Wide Web, qui est un ensemble de documents et de ressources (tels qu'audio, vidéo, texte, etc.) identifiés par le navigateur Web référencé par l'URL (Uniform Resource Locator) via le navigateur Web. Les pages HTML sont accessibles sur Internet. C'est le référentiel le plus hétérogène car il collecte des données à partir de plusieurs sources. Il est de nature dynamique car la quantité de données augmente et change constamment. Nous savons tous que les données d'Internet sont utilisées pour les achats en ligne, la recherche d'emploi, les recherches, les études etc.

1.5. Tâches du Data Mining :

Beaucoup de problèmes intellectuels, économiques ou même commerciaux peuvent être exprimés en termes des six tâches suivantes :

- La classification
- L'estimation
- La prédiction
- Le groupement par similitude
- L'analyse des clusters
- La description

Les trois premières tâches sont des exemples du Data Mining supervisé dont le but est d'utiliser les données disponibles pour créer un modèle décrivant une variable particulière prise comme but en termes de ces données. Le groupement par similitude et l'analyse des clusters sont des tâches non supervisées où le but est d'établir un certain rapport entre toutes les variables.

1.5.1.La classification :

La classification est la tâche la plus commune du Data Mining et qui semble être une obligation humaine. Afin de comprendre notre vie quotidienne, nous sommes constamment classifiés, catégorisés et évalués. La classification consiste à étudier les caractéristiques d'un nouvel objet pour lui attribuer une classe prédéfinie. Les objets à classifiés sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jour chaque enregistrement en déterminant un champ de classe. La tâche de classification est caractérisée par une définition de classes bien précise et un ensemble d'exemples classés auparavant. L'objectif est de créer un modèle qui peut être appliqué aux données non classifiées dans le but de les classifiées Quelques exemples de l'utilisation des tâches de classification dans les domaines de recherche et commerce sont les suivants [15] :

- Déterminer si l'utilisation d'une carte de crédit est frauduleuse.
- Diagnostiquant si une certaine maladie est présente.
- Déterminer quels numéros de téléphone corresponde aux fax.
- Déterminer quelles lignes téléphoniques est utilisées pour l'accès à Internet.

1.5.2.L'estimation :

L'estimation est similaire à la classification, sauf que la variable cible est numérique plutôt que catégorique. Les modèles sont construits en utilisant des données, qui fournissent la valeur de la variable cible, ainsi que les « prédicteurs ». Par exemple : « l'estimation de la pression artérielle d'un patient d'hôpital, basée sur son âge, son sexe, son indice de masse corporelle, et le taux de sodium. La relation entre la pression artérielle et le prédicteur variable de l'ensemble de formation nous donnerait un modèle d'estimation. Nous pouvons alors appliquer ce modèle à de nouveaux cas. [8]

1.5.3.La prédiction :

La prédiction est semblable à la classification et l'estimation, sauf que pour la prévision, les résultats se situent dans l'avenir. Exemples de tâches de prévision appliquée au marketing : « Prédire le prix d'un stock de trois mois dans le futur »

1.5.4.L'association :

La recherche de règles d'association est la tâche la plus intéressante du data mining. C'est également celle qui est la plus répandue dans le monde des affaires, notamment en marketing pour l'analyse du panier de consommation. La recherche de règles d'association cherche à découvrir les règles de quantification ou de relation entre deux ou plusieurs attributs. Les règles d'association sont de la forme « Si antécédent, puis conséquent », avec une mesure de confiance associée à la règle. La recherche de règles d'association dans une grande base de données permet de découvrir des règles cachées utiles pour la prise de décision.

Exemple de règle célèbre : lorsqu'un homme achète des couches pour bébés, il achète 2 packs d'eau dans 65% des cas. Il serait alors intéressant pour le gestionnaire d'adapter ses promotions à ces nouvelles règles.

1.5.5.Le clustering :

Le Clustering désigne le regroupement des données, des observations ou des cas dans des classes d'objets similaires. Un cluster maximise la similarité des objets de même cluster et minimise la similarité des objets de cluster différents. En effet, il n'y a pas de variable cible pour le clustering. La tâche de clustering ne cherche pas à classer, estimer, ou prédire la valeur d'une variable cible. Mais plutôt à segmenter l'ensemble des données en sous-groupes relativement homogènes à l'aide de mesures de distances.

1.5.6.La description :

Parfois, les chercheurs et les analystes essaient simplement de trouver des façons de décrire des tendances cachées dans les données. Les descriptions des modèles et des tendances servent à expliquer ou vérifier un fait. Par exemple : « ceux qui ont le plus de diplômes sont les plus susceptibles d'avoir un poste à responsabilité. ».

1.6. Extraction de connaissance à partir de données :

L'extraction de connaissances à partir de bases de données est un processus non trivial qui construit un modèle validé, nouveau, potentiellement utile et au final compréhensible, à partir de données.

Comme l'explique ce dernier auteur, l'ECBD peut se décomposer en de nombreuses étapes plus ou moins complexes mais la **figure- 1** en donne une vision synthétique. Parmi les grandes étapes de l'ECBD, on peut distinguer :

- La sélection : qui crée un ensemble de données à étudier.
- Le prétraitement : qui vise à enlever le bruit et à définir une stratégie pour traiter les données manquantes
- La transformation : où l'on recherche les meilleures structures pour représenter les données en fonction de la tâche.
- La fouille de données : la fouille proprement dite et la définition de la tâche : classification, recherche de modèles…et la définition des paramètres appropriés.
- L'interprétation et l'évaluation : pendant laquelle les patrons extraits sont analysés. La connaissance qui en est ainsi extraite est alors stockée dans la base de connaissances.

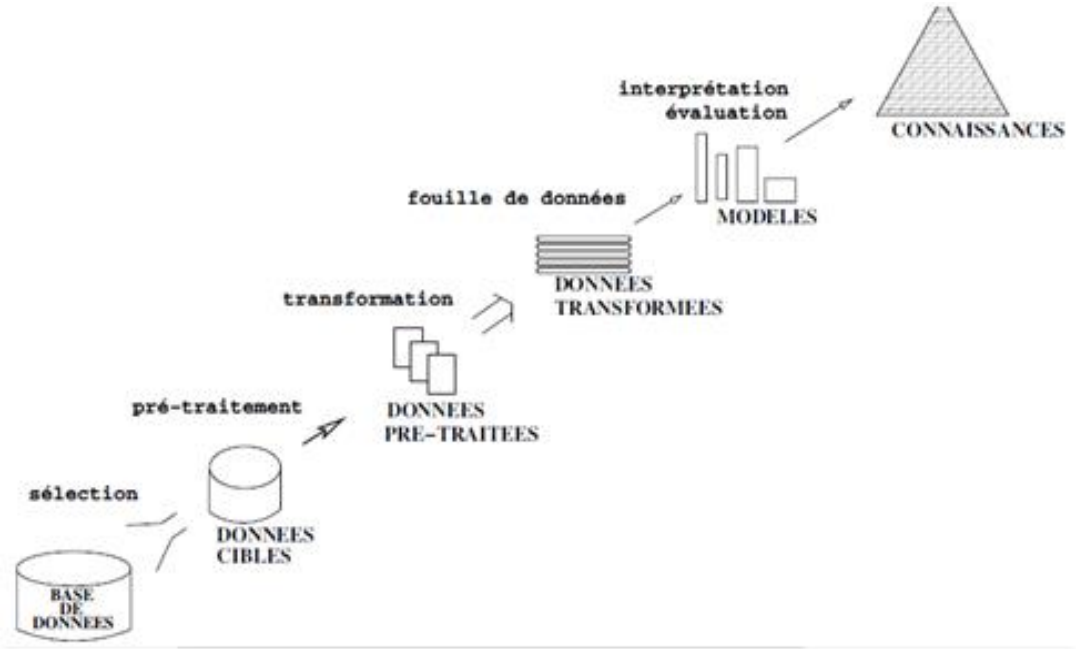


Figure 1- 2:Processus de découverte de connaissances à partir de données [4].

1.7. Quelques techniques du Data Mining :

Afin de découvrir des associations cachées dans le gisement de données et de créer des modèles à partir de ces données, diverses méthodes issues de diverses disciplines scientifiques (statistiques, intelligence artificielle, bases de données) sont utilisées pour effectuer des tâches de data mining. Dans ce titre, nous vous présentons les techniques de data mining les plus populaires.

1.7.1. Les réseaux de neurones :

Un réseau de neurones est un type spécifique de modèle de machine Learning, souvent utilisé avec l'intelligence artificielle et le deep Learning. Nommés ainsi car ils présentent différentes couches qui ressemblent à la façon dont les neurones fonctionnent dans le cerveau humain, les réseaux de neurones sont l'un des modèles de machine Learning les plus précis utilisés aujourd'hui.

1.7.1.1. Avantages et inconvénients :

Les avantages :

- Les réseaux de neurones peuvent théoriquement approximer n'importe quelle fonction continue, de sorte que les chercheurs n'ont pas à faire d'hypothèses sur le modèle sous-jacent.

Les inconvénients :

- Les réseaux de neurones ne sont généralement pas largement utilisés dans les tâches d'exploration de données, car les modèles qu'ils créent sont souvent difficiles à comprendre et prennent un long temps d'apprentissage.

1.7.2. Les arbres de décision :

Les arbres de décision (AD) sont une catégorie d'arbres utilisée dans l'exploration de données et en informatique décisionnelle. Ils emploient une représentation hiérarchique de la structure des données sous forme des séquences de décisions (tests) en vue de la prédiction d'un résultat ou d'une classe. Chaque individu (ou observation), qui doit être attribué(e) à une classe, est décrit(e) par un ensemble de variables qui sont testées dans les nœuds de l'arbre. Les tests s'effectuent dans les nœuds internes et les décisions sont prise dans les nœuds feuille.

Exemple de problème adapté à un approche par arbres de décision : comment répartir une population d'individus (e.g. clients, produits, utilisateurs, etc.) en groupes homogènes selon un ensemble de variables descriptives (e.g. âge, temps passé sur un site Web, etc.) et en fonction d'un objectif fixé (variable de sortie ; par exemple : chiffre d'affaires, probabilité de cliquer sur une publicité, etc.).**[9]**

1.7.2.1. Avantages et inconvénients :

Les avantages :

- Connaissances « intelligibles » -- validation d'expert (si arbre pas trop grand)
- Traduction directe de l'arbre vers une base de règles
- Sélection automatique des variables pertinentes
- Non paramétrique
- Traitement indifférencié selon le type des variables prédictives
- Robuste face aux données aberrantes, solutions pour les données manquantes
- Robuste face aux variables redondantes
- Rapidité et capacité à traiter des très grandes bases
- Enrichir l'interprétation des règles à l'aide des variables non sélectionnées
- Possibilité pour le praticien d'intervenir dans la construction de l'arbre **[10]**

Les inconvénients :

- Problème de stabilité sur les petites bases de données (feuilles à très petits effectifs)
- Recherche « pas-à-pas » : difficulté à trouver certaines interactions (ex. XOR)
- Peu adapté au « scoring »
- Performances moins bonnes en général par rapport aux autres méthodes (en réalité, performances fortement dépendantes de la taille de la base d'apprentissage) **[10]**

1.7.3. Algorithmes génétiques :

Les algorithmes génétiques sont mis en œuvre dans certains outils d'informatique décisionnelle ou de data mining par exemple pour rechercher une solution d'optimum à un problème par mutation des attributs (des variables) de la population étudiée.

Ils sont utilisés par exemple dans une étude d'optimisation d'un réseau de points de vente ou d'agences (banque, assurance, ...) pour tenter de répondre aux questions :

- Quelles sont les variables (superficie, effectif, ...) qui expliquent la réussite commerciale de telle ou telle agence ?
- En modifiant telle variable (mutation) de telle agence améliore-t-on son résultat ?**[17]**

1.7.3.1. Avantages et inconvénients :

Les avantages :

- Ils utilisent l'évaluation de la fonction objective sans prendre en compte sa nature ce qui lui donne plus de souplesse et un large domaine d'application.
- Ils sont dotés de parallélisme car ils travaillent sur plusieurs points en même temps il s'agit des individus de la population.
- L'utilisation de règles de transition probabilistes de croisement et de mutation permet dans certains cas d'éviter des optimums locaux et d'aller vers un optimum global

Les inconvénients :

- Temps de calcul très élever car ils nécessitent de nombreux calculs particulièrement au niveau de la fonction d'évaluation.
- Difficiles à mettre en œuvre à cause

- Des paramètres parfois difficiles à déterminer comme la taille de la population ou le taux de mutation. Ce qui implique la nécessité de plusieurs essais car le succès de l'évolution en dépend, ce qui limite encore l'efficacité de l'algorithme.
- Du choix de la fonction d'évaluation qui est critique, elle doit prendre en compte les bons paramètres du problème. Elle doit donc être choisie avec soin.

- Il est impossible d'être sûre que la solution obtenue après un nombre fini de générations soit la meilleure, on peut seulement être sûr que l'on s'est approché de la solution optimale.
- Problème des optimums locaux : lorsqu'une population évolue, il se peut que certains individus deviennent majoritaires. À ce moment, il se peut que la population converge vers cet individu et s'écarte ainsi d'individus plus intéressants mais trop éloignés de l'individu vers lequel la population converge

1.7.4. Règles d'association :

Les règles d'association sont une des méthodes de Data Mining les plus répandues dans le domaine du marketing et de la distribution. Les règles d'association générées sont de la forme "Si action1 ou condition alors action 2". Elles peuvent se situer dans le temps : "Si action1 ou condition à l'instant t1 alors action2 à l'instant t2" c'est les règles d'association séquentielles.

Leur principale application est « l'analyse du panier de la ménagère », qui consiste, comme l'indique son nom, en la recherche d'associations entre produits sur les tickets de caisse et l'étude de ce que les clients achètent. La méthode recherche quels produits tendent à être achetés ensemble. Elles peuvent être appliquées à tout secteur d'activité pour lequel il est intéressant de rechercher des groupements potentiels de produits ou de services. Voici quelques exemples de règles: - Si un client achète du lait alors il achète du pain (90%) - Si un client achète une télévision, il achètera un récepteur satellite dans un mois (50%) - Si maladie X et traitement Y alors guérison (95%) - Si maladie X et traitement Y alors guérison dans Z années (97%) - Si présence et travail alors réussite à l'examen (99%) Ces règles sont intuitivement faciles à interpréter car elles montrent comment des produits ou des services se situent les uns par rapport aux autres. Elles sont particulièrement utiles en marketing et peuvent être facilement utilisées dans le système d'information de l'entreprise. Le but principal de cette technique est donc descriptif. Dans la mesure où les résultats peuvent être situés dans le temps, cette

technique peut être considérée comme prédictive. Cependant, il faut noter que cette méthode, si elle peut produire des règles intéressantes, peut aussi produire des règles triviales ou inutiles (provenant de particularités de l'ensemble d'apprentissage). La recherche de règles d'association est une méthode non supervisée car on ne dispose en entrée que de la description des achats. [**5**] [13]

1.7.4.1. Avantages et inconvénients :

Les avantages :

- Leur application dans plusieurs domaines de la vie quotidienne, comme l'analyse du panier de la ménagère.
- La découverte de connaissances utiles, cachées dans les grandes bases des données.
- Leur simplicité, efficacité et facilité de compréhension.
- Leur formalisme non supervisé et général.
- Leurs résultats sont clairs et faciles à interpréter**. [14]**

Inconvénients :

- Le temps énorme consacré à la recherche des itemsets fréquents.
- La grande quantité des règles d'association générées.
- La difficulté d'évaluer la qualité des règles d'associations par des indices statiques ou par l'expert du domaine.
- La production des règles triviales et inutiles qm n'apportent pas de nouvelles informations. **[14]**

1.7.5. L'algorithme des k-Plus proches voisins :

L'algorithme k-plus proche voisin (K-PPV, k nearest Neighbors en anglais ou kNN) est un algorithme de raisonnement à partir de cas qui a été développé pour la classification et peut être étendu pour inclure des problèmes d'estimation. Il vise à prendre des décisions sur la base d'un ou plusieurs cas similaires qui ont été résolus en mémoire. Comparé à d'autres méthodes de classification (arbres de décision, réseaux de neurones, algorithmes génétiques, etc.), l'algorithme KNN ne crée pas de modèles basés sur des échantillons d'apprentissage, mais c'est l'échantillon d'apprentissage, la fonction de distance et la fonction de choix de la classe en fonction des classes des voisins les plus proches, qui composent le modèle.

- Algorithme de classification par k-PPV

Paramètre : le nombre k de voisins

Donnée : un échantillon de m exemples et leurs classes

La classe d'un exemple X est c(X)

Entrée : un enregistrement Y

1. Déterminer les k plus proches exemples de Y en calculant les distances
2. Combiner les classes de ces k exemples en une classe c

Sortie : la classe de Y est c(Y)=c

- Comment cela marche-t-il ?

Nous supposons que nous avons une base de données d'apprentissage contenant N paires « entrée-sortie ». Afin d'estimer la valeur de sortie de la nouvelle entrée x, la méthode des K plus proches voisins consiste à prendre en compte (de façon identique) les k échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée x, selon une distance à définir.

Si nous prenons une base d'apprentissage de 100 éléments, Dès que nous recevons un nouvel élément que nous souhaitons classifier, l'algorithme calcule sa distance à tous les éléments de la base. Si cette base comporte 100 éléments, alors il va calculer 100 distances et donc obtenir 100 nombres réels. Si k = 25 par exemple, il cherche alors les 25 plus petits nombres parmi ces 100 nombres qui correspondent donc aux 25 éléments de la base qui sont les plus proches de l'élément que nous souhaitons classifier. La classe attribuée à l'élément à classifier est la classe majoritaire parmi ces 25 éléments

1.7.5.1. Avantages et inconvénients :

Les avantages :

- La qualité de la méthode s'améliore en introduisant de nouvelles données sans nécessiter la reconstruction d'un modèle. Ce qui représente une différence majeure avec des méthodes telles que les arbres de décision et les réseaux de neurones.
- La clarté des résultats : la classe attribuée à un objet peut être expliquée en exhibant les plus proches voisins qui ont amené à ce choix.
- La méthode peut s'appliquer à tout type de données même les données complexes tels que des informations géographiques, des textes, des images, du son. C'est parfois un critère de choix de la méthode PPV car les autres méthodes traitent difficilement les données complexes. Nous pouvons noter, également, que la méthode est robuste au bruit.

- Facile à mettre en œuvre

Les inconvénients :

- Temps de classification : la méthode ne nécessite pas d'apprentissage ce qui implique que tous les calculs sont effectués lors de la classification. Contrairement aux autres méthodes qui nécessite un apprentissage (éventuellement long) mais qui sont rapides en classification.
- Méthode donnera de mauvais résultats Si le nombre d'attributs pertinents est faible relativement au nombre total d'attributs, car la proximité sur les attributs pertinents sera noyée par les distances sur les attributs non pertinents.
- Les performances de la méthode dépendent du choix de la distance, du nombre de voisins et du mode de combinaison des réponses des voisins

1.7.6. Motifs séquentiels :

Une base de données séquentielles contient un ensemble ordonné d'éléments ou d'événements, enregistrés avec ou sans valeur concrète du temps. On retrouve de telles séquences dans de nombreuses applications comme les séquences d'achats des consommateurs, les séquences biologiques. L'extraction de motifs séquentiels est un domaine très actif de la fouille de données. Nous introduisons d'abord les concepts préliminaires relatifs aux motifs séquentiels.

Soit **I = {i1, i2, . . ., ik}** l'ensemble de tous les items. Un sous-ensemble de I est appelé un itemset. Une séquence **α = ha1, a2, . . .,** ani est une liste ordonnée d'items et (***ai ⊆ I***). Chaque itemset d'une séquence représente un ensemble d'événements qui apparaissent à la même estampille temporelle. Les différents itemset d'une séquence sont associés à des estampilles temporelles différentes. Par exemple, un consommateur peut acheter plusieurs produits lors d'un passage dans le magasin et revenir plusieurs fois faire des achats. Il peut ainsi acheter un PC et des logiciels puis revenir acheter un appareil photo numérique avec une carte mémoire puis enfin acheter une imprimante et des livres sur la photographie.

Une séquence α = ha1, a2, . . ., ami est une sous-séquence de β = hb1, b2, . . ., bni (noté α β) si et seulement si ∃i1, i2, . . ., im tels que 1 ≤ i1 < i2 < . . . < im ≤ n et a1 ⊆ bi1, . . ., am ⊆ bim. On dit également que β est une super séquence de α ou que β contient α. Étant donné un ensemble de séquences D = {s1, s2, . . ., sn}, le support d'une séquence α correspond au nombre de séquences de D qui contiennent α. Si le support d'une séquence α satisfait un seuil de support

minimum minsup, alors α est un motif séquentiel fréquent. L'objectif de la recherche de motifs séquentiels est donc d'extraire l'ensemble complet des motifs séquentiels fréquents par rapport à un seuil de support minimum minsup.[15][16]

1.8. Domaines d'application du Data Mining :

Les domaines d'application du Data Mining sont très nombreux. Parmi lesquels on peut citer :

1.8.1.Santé publique :

La science des données apparaît aujourd'hui comme une réponse majeure à de lourdes problématiques mondiales de santé et d'économie de la santé. Elle attire ainsi les plus grosses sociétés mondiales qui souhaitent investir dans ce domaine, et mobilise tous les gouvernements internationaux dans la gestion de leurs politiques de santé publique, dont les enjeux sont extrêmement élevés pour améliorer les systèmes de santé actuels et les budgets très conséquents (En France, 240 milliards d'euros sont attribués aux dépenses de santé, soit son premier budget). Les enjeux sont immenses pour améliorer le système de santé [17] :

- Détecter des signaux faibles en pharmaco-épidémiologie.
- Optimiser les parcours médico-économiques pour certaines pathologies.
- Réaliser l'évaluation comparative de l'efficacité de stratégies thérapeutiques…

1.8.2.Réseaux sociaux :

Twitter, Facebook, Instagram, Snapchat, … Pour ne citer que quelques exemples, les réseaux sociaux font partie de la vie de tous les jours de tout un chacun. Chaque mois, le monde compte plus de 2 milliards d'utilisateurs de réseaux sociaux, et 30% du temps passé on-line est consacré à l'utilisation de ces réseaux. Simple passe-temps, vecteur essentiel des liens sociaux actuels, source d'information exceptionnelle mais aussi ... outil de propagande, les réseaux sociaux sont omniprésents et leur place est devenue centrale dans la société d'aujourd'hui. [17]

Comprendre et caractériser leur structure communautaire ou leurs dynamiques temporels sous-jacentes (chaque réseau a ses particularités !), pouvoir suivre la propagation d'une information, quantifier son impact, … autant de sujets qui passionne les membres de l'Initiative.

1.8.3.Journalisme et fact-checking :

La profusion de données produites et échangées par des acteurs publics ou privés est une mine d'or pour analyser les mécanismes des entités et organisations dans des sphères telles que l'économie, la politique, la culture... La capacité d'intégrer rapidement des données hétérogènes par leur structure et leur sémantique conditionne la compréhension du monde qui nous entoure et fournit la base du débat démocratique. Ainsi, on peut croiser des sources structurées telles que les bases de données relationnelles ou des tableaux, avec des sources moins structurées, telles que des graphes sémantiques ou encore du texte. Les outils d'analyse et intégration de données ont fait leurs preuves récemment, par exemple lors de l'analyse des "Panama Papers" par un consortium international de journalistes d'investigation. Des outils d'interconnexion et d'interprétation des données permettent aussi de faciliter la tâche des journalistes spécialisés dans le "fact-checking. [17]

1.8.4.Le data mining dans le marketing :

- La gestion de la relation client (GRC ou CRM) consiste en l'ensemble des activités visant à cibler, attirer et conserver les "bons" clients.
- Détection d'associations de comportements d'achat.
- Découverte de caractéristiques de clientèle.
- Prédiction de probabilité de réponse aux campagnes de mailing.

1.9. Conclusion :

Le data mining est un processus d'aide à la décision dans lequel nous recherchons des modèles d'information dans les données, Cette technique peut être utilisée sur de nombreux types de données pour le but de prédire les tendances et les comportements futurs et prendre les bonnes décisions. Cela fait de la data mining la technologie la plus important

Chapitre 2 :

Extraction de motifs séquentiels

2.1. Introduction :

L'exploration de motifs séquentiels (SPM) est le processus qui extrait certains motifs séquentiels dont le support dépasse un seuil de support minimal prédéfini. En outre, l'exploration de motifs séquentiels permet d'extraire les séquences qui reflètent les comportements les plus fréquents dans la base de données de séquences, qui peuvent à leur tour être interprétées comme des connaissances du domaine à plusieurs fins. Pour réduire le très grand nombre de séquences en motifs séquentiels les plus intéressants et pour répondre aux différentes exigences des utilisateurs, il est important d'utiliser un support minimum qui élague les motifs séquentiels sans intérêt. Il est clair qu'un support plus élevé d'un motif séquentiel est préférable pour des motifs séquentiels plus intéressants. L'exploration de motifs séquentiels est utilisée dans plusieurs domaines [20].

Récemment, plusieurs algorithmes pour SPM ont été proposés et la plupart des algorithmes essentiels et antérieurs sont basés sur la propriété de l'algorithme Apriori proposé par Agrawal et Srikant en **1994**. Cette propriété stipule qu'un motif fréquent contient des sous-modèles qui sont à leur tour fréquents. Sur la base de cette hypothèse, une succession d'algorithmes a été proposée : en **1995**, les algorithmes AprioriAll, AprioriSome, DynamicSome ont été proposés par Agrawal et Srikant. De plus, la méthode de formatage horizontal basée sur Apriori (GSP) a été présentée en **1996** par les mêmes auteurs Agrawal et Srikant et la méthode de formatage vertical basée sur Apriori (algorithme SPADE) a été présentée par Zaki en **2001.** [19] [18]

2.2. Notations fondamentales :

2.2.1. Base de données des séquences :

Une base de données de séquences **SDB** est une liste de séquences **SDB = [s1, s2, ..., sp]** ayant des identifiants de séquence **(SID) 1, 2...p.** *Par exemple*, une base de données de séquences est présentée dans **le tableau II.1**, qui contient quatre séquences ayant les SID **1, 2, 3 et 4**. Ces séquences pourraient, par exemple, représenter des achats effectués par quatre clients. [21]

SID	Séquence
1	[{a, b}, {c}, {f, g}, {g}, {e}]
2	[{a, d}, {c}, {b}, {a, b, e, f}]
3	[{a}, {b}, {f}, {e, g}]
4	[{b}, {f, g}]

Tableau 2- 1:Base de données des séquences

2.2.2. Séquence de données :

On appelle une séquence de données est une liste ; ordonne non vide d'itemsets Si, appelé **S= <s1, s2, …, sn>** avec **i ∈ [1…n]** indique l'ordre d'apparition de si dans **S**. [21]

Exemple :

Reprenons la BDD du ***Table 2.1***. La liste ordonnée des trois transactions effectuées par le client 2 est donnée par la séquence **<{a}, {b}, {f}, {e, g}>** Avec **S1 = {a}, S2 = {b}, S3 = {f}, S4 = {e, g}**

Cette séquence se lit de la manière suivante : le client 2 a acheté l'article a, puits l'article **b**, puits l'article **f**, puits simultanément les deux articles **e** et **g**.

2.2.3. Longueur d'une séquence :

La longueur d'une séquence **S** est le nombre d'items dans cette séquence. Une séquence de longueur k est une k-séquence. [21]

Exemple :

Séquence **< {a, d}, {c}, {b}, {a, b, e, f} >** est une 8-séquence, même si cette dernière contient seulement 4 itemsets. L'item a est contenu dans deux transactions et est donc compté deux fois.

2.2.4. Transaction :

On appelle une transaction pour séquence **S** une tuple (SID, itemsets) formé de l'identificateur unique de la séquence, de la valeur de l'identifiant temporel pour cette transaction et l'ensemble des items de la transaction, représentant tous les items achetés par un client à la fois. [21]

2.3. Catégories des algorithmes d'extraction des motifs séquentiels :

Nous classons les algorithmes d'extraction de motifs séquentiels dans les catégories suivantes : Algorithmes de type apriori, algorithmes basés sur la recherche en largeur **(BFS**), algorithmes basés sur la recherche en profondeur **(DFS**), algorithmes basés sur des motifs séquentiels fermés et algorithmes basés sur l'incrémentation [20] :

2.3.1. Algorithmes apriori :

La première introduction des algorithmes classiques d'exploration de motifs séquentiels basés sur Apriori a été faite par **R. Agrawal et R. Srikant**. Soit une base de données de transactions comprenant des séquences de clients. Cette base de données est composée de trois attributs (identifiant du client, durée de la transaction et article acheté). Le processus de fouille a été décomposé en cinq étapes [22] :

- ***Étape de tri*** : qui trie la base de données transactionnelle selon l'identifiant client.
- ***Étape L-itemsets*** : l'objectif est d'obtenir les grands 1- itemsets à partir de la base de données triée, en fonction du seuil de support.
- ***Étape de transformation :*** cette étape remplace les séquences par les grands items qu'elles contiennent. Pour une extraction efficace, tous les grands itemsets sont mis en correspondance avec une série d'entiers. Enfin, la base de données originale sera transformée en un ensemble de séquences de clients représentées par ces grands itemsets.
- ***Étape séquentielle :*** À partir de la base de données séquentielle transformée, cette étape génère tous les motifs séquentiels fréquents.
- **Étape maximale :** Cette étape élimine les motifs séquentiels qui sont contenus dans d'autres motifs super séquentiels, car nous ne nous intéressons qu'aux motifs séquentiels maximums.

Même si l'algorithme d'**Apriori** est à la base de nombreux algorithmes efficaces développés par la suite, il n'est pas assez efficace. **R. Agrawal** et **R. Srikan** ont détecté une propriété intéressante de fermeture vers le bas, appelée **Apriori**, parmi les k-itemsets fréquents : Un k-itemsets est fréquent seulement si tous ses sous-itemsets sont fréquents. Cette propriété signifie que les itemsets fréquents peuvent être extraits en identifiant les 1-itemsets fréquents (première analyse de la base de données), puis les 1-itemsets fréquents seront utilisés pour générer des 2-itemsets fréquents candidats, ce processus sera répété à nouveau pour obtenir les 2-itemsets fréquents. Ce processus est itératif jusqu'à ce que des k-itemsets fréquents puissent être générés pour un certain k.

De nombreuses études ont été menées sur les améliorations d'Apriori, par exemple l'approche d'échantillonnage, le comptage dynamique d'items, l'exploration incrémentale, l'exploration parallèle et distribuée.

Dans certains cas, la taille des ensembles de candidats utilisant le principe d'Apriori est considérablement réduite. Cette situation peut causer deux problèmes :

- Un grand nombre d'ensembles de candidats doit être généré.
- L'utilisation de la correspondance de motifs pour scanner constamment la base de données et découvrir les candidats.

2.3.2.Algorithmes basés sur la recherche en largeur (BFS) :

Les algorithmes **Breath-first** (par niveau) décrivent les algorithmes basés sur **Apriori** car toutes les k-séquences sont construites ensemble à chaque k-ième itération de l'algorithme alors qu'elles traversent l'espace de recherche. Plusieurs algorithmes ont été développés en utilisant le principe des algorithmes **BFS**. Parmi eux, nous en citons l'algorithme **GSP** (Generalized Sequential Pattern) [20].

2.3.2.1. GSP :

L'***algorithme GSP*** proposé par ***Srikant*** et Agrawal en 1996, fait le même travail que l'algorithme AprioriAll, mais il ne nécessite pas de trouver d'abord tous les items fréquents. Cet algorithme permet de :

- Placer des limites sur la séparation temporelle entre des éléments adjacents dans un motif
- Permettre aux éléments inclus dans l'élément du motif de couvrir un ensemble de transactions dans une fenêtre temporelle spécifiée par l'utilisateur
- Permettre la découverte de motifs à différents niveaux d'une taxonomie définie par l'utilisateur.

De plus, **GSP** est conçu pour découvrir des motifs séquentiels généralisés. L'***algorithme GSP*** effectue plusieurs passages sur la base de données de séquences comme suit :

Le **SPG** utilise une représentation standard de la base de données, comme indiqué dans le tableau 1, également appelée base de données horizontale. L'algorithme **GSP** effectue une recherche par niveau pour découvrir des modèles séquentiels fréquents. Il analyse d'abord la base de données pour calculer la prise en charge de toutes les 1-séquences. Ensuite, il garde en mémoire toutes les 1-séquences fréquentes [22].

Ensuite, l'algorithme **GSP** poursuit ce processus pour générer des modèles séquentiels de longueur ***3, 4***, etc. jusqu'à ce qu'aucun modèle ne puisse être généré.

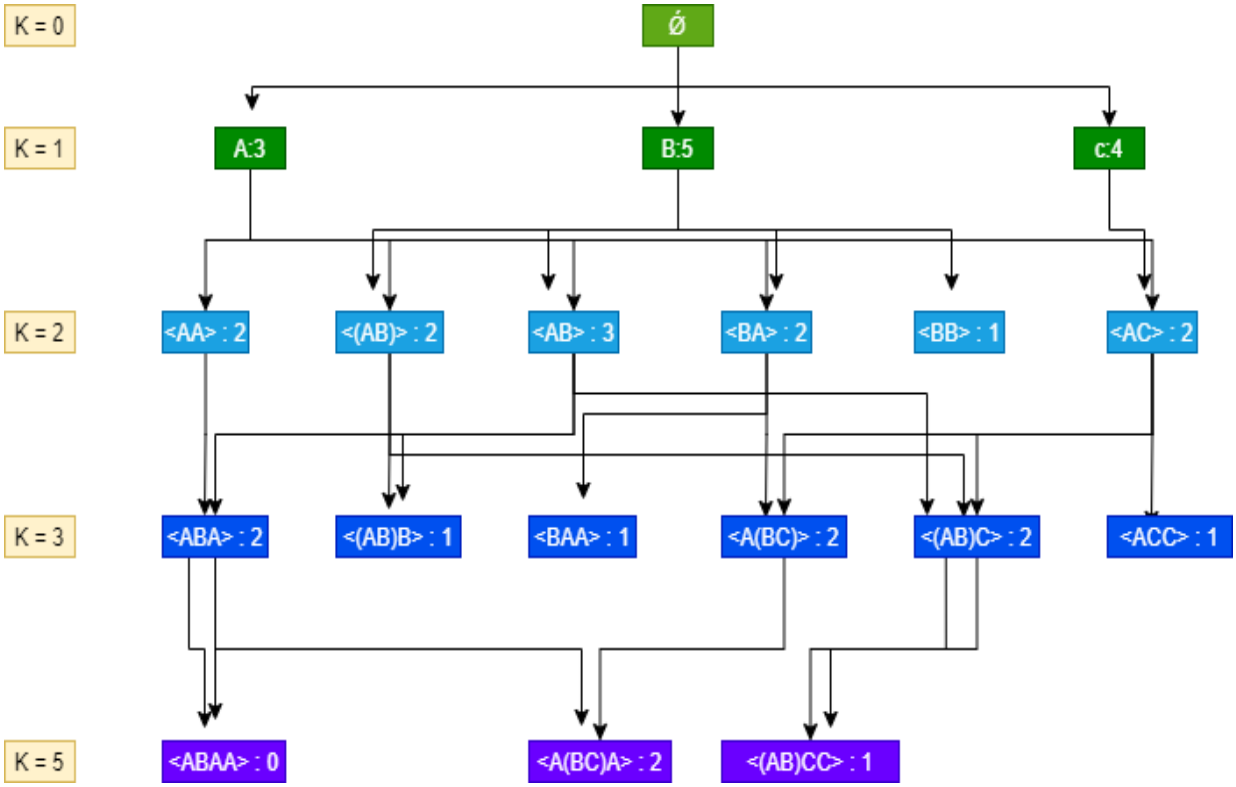


Figure 2- 1: Le processus d'extraction du SPG

aAlgorithme GSP :

```
1. Obtenier une séquence sous la forme d'une 1-longueur candidats;
2. Trouver F1 (l'ensemble des motifs séquentiels de longueur 1), après un scan unique de
3. la base de données;
4. k = 1;
5. Tant que Fk n'est pas vide faire
6.       Formez Ck+1, l'ensemble des condidats de longueur (k+1) de Fk;
7.       Si Ck+1 n'est pas vide alors
8.            Scan la base de données unique;
9.            Trouver Fk+1(l'ensemble des motifs séquentiels de longueur (k+1));
10.      Fin Si
11.      k = k+1;
12. Fin tant que
```

Figure 2- 2:algorithme GSP

Exemple :

SID	Séquence
10	< (bd) cb (ac) >
20	< (bf) (ce) b (fg) >
30	< (ah) (bf) abf >
40	< (be) (ce) d >
50	< a (bd) bcb (ade) >

Tableau 2- 2:base de données de séquence -2-

On Suppose que ***: min_sup = 2***

Les premiers candidats : Toutes les séquences simples : ***<a><b><c><d><e><f><g><h>***

Scannez la BD une fois et comptez le support pour chaque candidat :

Candidats	Sup
<a>	3
<b>	5
<c>	4
<d>	3
<e>	3
<f>	2
<g>	1
<h>	1

Tableau 2- 3:candidats 1-séquences

Les 1-séquences fréquents sont : **<a>, <b>, <c>, <d>, <e>, <f>**

Les candidats de 2-séquence :

	<a>	<b>	<c>	<d>	<e>	<f>
<a>	<aa>	<ab>	<ac>	<ad>	<ae>	<af>
<b>	<ba>	<bb>	<bc>	<bd>	<be>	<bf>
<c>	<ca>	<cb>	<cc>	<cd>	<ce>	<cf>
<d>	<da>	<db>	<dc>	<dd>	<de>	<df>
<e>	<ea>	<eb>	<ec>	<ed>	<ee>	<ef>
<f>	<fa>	<fb>	<fc>	<fd>	<fe>	<ff>

Tableau 2- 4:Les candidats de 2-séquence -1-

	<a>	<b>	<c>	<d>	<e>	<f>
<a>		<(ab)>	<(ac)>	<(ad)>	<(ae)>	<(af>
<b>			<(bc)>	<(bd)>	<(be)>	<(bf)>
<c>				<(cd)>	<(ce)>	<(cf)>
<d>					<(de)>	<(df)>
<e>						<(ef)>
<f>						

Tableau 2- 5:Les candidats de 2-séquence -2-

Candidat de 2-longeurs = ***36 + 15 = 51***

b. Les limites de GSP :

- Une grande quantité de candidats peut être générée dans les grandes bases de données. À titre indicatif, une base de données contenant 1000 1-séquences formera 1 499 500 candidats (1000 x 1000 + 1000 x (1000-1)/2). Etant donné que les candidats générés sont formés à partir de la concaténation du fichier d'amorce, plusieurs de ces candidats ne se retrouveront pas dans la base de données, ce qui représente une perte de temps.
- Une grande quantité de balayages de la base de données est requise. Étant donné que la longueur de chaque séquence candidate grandit d'un item à chaque balayage, l'identification d'une 15-séquence requiert 15 balayages de la base de données.
- Les méthodes basées sur Apriori rencontrent souvent des difficultés pour la découverte de longue séquence. Ceci est du fait que le nombre des séquences candidates est Fonction exponentielle de longueur séquences plus coutes qui les forme [22].

2.3.3. Algorithmes basés sur la recherche en profondeur (DFS) :

2.3.3.1. Algorithme SPADE :

L'algorithme **SPADE** proposé par ***Zaki*** en ***2001*** et il inclut les caractéristiques d'un partitionnement de l'espace de recherche où l'espace de recherche inclut la disposition verticale de la base de données. L'espace de recherche dans ***SPADE*** est représenté comme une structure en treillis et il utilise la notion de classes d'équivalence pour le partitionner. Il décompose le treillis original en sous-treillis plus petits, de sorte que chaque sous-treillis peut être entièrement traité en utilisant une méthode de recherche de type breadth-first ou depth-first ***(SPADE*** est également une méthode basée sur DFS). Le comptage du support **SPADE** de la méthode de la séquence candidate comprend des opérations bit à bit ou logiques. Les résultats expérimentaux

montrent que **SPADE** est environ deux fois plus rapide que **GSP.** La raison en est que **SPADE** utilise une méthode de comptage de support plus efficace basée sur la structure de la liste d'identification. De plus, **SPADE** montre une évolutivité linéaire par rapport au nombre de séquences [23] [24].

a. Algorithme SPADE :

```
1. Entrée : une base de données de séquences d'événements et un seuil de support
2. minimum MinSup;
3. Sortie : les motifs séquentiels fréquents contenus dans la base de données;
4. Utilisez la base de données pour calculer:
5.       F1 l'ensemble de tous les éléments fréquents;
6.       IdList(z) pour tout élément z de F1;
7. i = 1;
8. Tant que Fi ≠ Ø faire
9.       F(i+1) = Ø;
10.      Pour tout z1 ∈ F(i) faire
11.           Pour tout z2 ∈ F(i) faire
12.                Si Z1 et Z2 ont le même préfixe alors
13.                     Pour tout z obtenu par merge(z1, z2) faire
14.                          Calculer IdList(z) par join(IdList(z1), IdList(z2));
15.                          Utilisez IdList(z) pour déterminer si z est fréquent;
16.                          Si z est fréquent alors
17.                               F(i+1) = F(i+1) + {z};
18.                          Fin si
19.                     Fin pour
20.                Fin si
21.           Fin pour
22.      Fin pour
23.      i = i+1;
24.Fin tant que
25.Sortie F(1) U F(2) U … U F(n);
```

Figure 2- 3:Algorithme SPADE

b. Exemple :

SID	Séquence
1	**< a (abc) (ac) d (cf) >**
2	**< (ad) c (bc) (ae) >**
3	**< (ef) (ab) (df) cb >**
4	**< eg (af) cbc >**

Tableau 2- 6:base de données des séquences -2-

On Suppose que : ***min_sup = 2***

a		b		…
SID	***EID***	***SID***	***EID***	…
1	1	1	2	
1	2	2	3	
1	3	3	2	
2	1	3	5	
2	4	4	5	
3	2			
4	3			

Tableau 2- 7:Les candidats de 1-séquence -2-

ab			ba			…
SID	***EID(a)***	***EID(b)***	***SID***	***EID(b)***	***EID(a)***	…
1	*1*	2	1	2	3	
2	*1*	3	2	3	4	
3	*2*	5				
4	*3*	5				

Tableau 2- 8:Les candidats de 2-séquence -3-

aba				…
SID	***EID(a)***	***EID(b)***	***EID(a)***	….
1	1	2	3	
2	1	3	4	

Tableau 2- 9:Les candidats de 3-séquence

c. Limite de SPADE :

- La nécessité d'une très grande mémoire pour transformer et puis après stocker toute la base de données.
- Les temps de réponse au moment de compter le support des candidats générés à chaque étape est très intéressant.

2.3.3.2. FreeSpan :

FreeSpan est un algorithme proposé par **Pei** et **al**. en **2001** dans le but de réduire la génération de sous-séquences candidats. Il utilise des bases de données projetées pour générer des annotations de bases de données afin de guider le processus d'exploration pour trouver rapidement des motifs fréquents. L'idée générale de **FreeSpan** est d'utiliser des éléments fréquents pour projeter des bases de données de séquences dans un ensemble de bases de données projetées plus petites en utilisant récursivement les ensembles fréquents actuellement minés, et des fragments de sous-séquences dans chaque base de données projetée sont générés, respectivement. Deux alternatives de projections de bases de données peuvent être utilisées : la projection niveau par niveau ou la projection niveau alternatif. La méthode utilisée par **FreeSpan** divise les données et l'ensemble des motifs fréquents à tester, et limite chaque test effectué à la plus petite base de données projetée correspondante. **FreeSpan** ne scanne la base de données originale que trois fois, quelle que soit la longueur maximale de la séquence. Les résultats expérimentaux montrent que **FreeSpan** est efficace, qu'il extrait l'ensemble complet de motifs et qu'il est considérablement plus rapide que l'algorithme **GSP**. Le coût principal de **FreeSpan** est de traiter les bases de données projetées[24].

2.3.3.3. PrefixSpan :

Une méthode de croissance des motifs basée sur la projection est utilisée dans l'algorithme PrefixSpan pour l'extraction de motifs séquentiels. L'idée de base de cette méthode est qu'au lieu de projeter les bases de données de séquences en évaluant les occurrences fréquentes des sous-séquences, la projection est faite sur le préfixe fréquent. Cela permet de réduire le temps de traitement, ce qui augmente finalement l'efficacité de l'algorithme**. Jian Pei** et **al** ont proposé un nouvel algorithme appelé **PrefixSpan** (*Prefix-projected Sequential Pattern Mining)* qui fonctionne sur la projection de la base de données et la croissance des motifs séquentiels. La technique de division et de recherche d'espace est mise en œuvre par ***PrefixSpan***. L'algorithme extrait des motifs séquentiels en suivant les étapes suivantes [25] :

- Trouver des motifs séquentiels de **longueur 1**. La séquence **S** donnée est analysée pour obtenir l'élément *(préfixe*) qui apparaît fréquemment dans **S**. Le nombre de fois où cet élément apparaît est égal à la longueur-l de cet élément. La **longueur-l** est donnée par la notation ***<pattern> : <count>.***
- Diviser l'espace de recherche. Sur la base du préfixe dérivé de la première étape, l'ensemble des motifs séquentiels est divisé dans cette phase.

➢ Trouver des sous-ensembles de motifs séquentiels. Les bases de données projetées sont construites et les motifs séquentiels sont extraits de ces bases de données. Seules les séquences fréquentes locales sont explorées dans les bases de données projetées afin d'étendre les motifs séquentiels. Le coût de la construction des bases de données projetées est assez élevé. Les méthodes de projection à deux niveaux et de pseudo-projection sont utilisées pour réduire ce coût, ce qui augmente finalement l'efficacité de l'algorithme.

a. Algorithme PrefixSpan :

```
L'entrée de PrefixSpan est une base de données de séquences et un seuil spécifié par
l'utilisateur appelé support minimum (minsup);
Sortie : L'ensemble complet de motifs séquentiels;
En analysant S| α une fois, on trouve l'ensemble des éléments fréquents b tels que
      b peut être assemblé au dernier élément de α pour former un motif séquentiel;
      Où
      On peut ajouter <b> à α pour former un motif séquentiel;
Pour chaque élément fréquent b
      l'annexer à α pour former un motif séquentiel α', et sortir α' ;
Fin pour

Pour chaque α'
      construire la base de données α'-projetée S|α';
      et appeler PrefixSpan (α', i+1, S|α');
Fin pour
```

Figure 2- 4:Algorithme PrefixSpan

b. Exemple :

Préfix	Suffixe (projection)
<a>	**< (abc) (ac) d (c f) >**
<aa>	**< (_bc) (ac) d (cf) >**
<ab>	**< (_c) (ac) d (cf) >**

Tableau 2- 10:Exemple de suffixe et préfix

***1er étape* :** trouver les séquences 1-langeurs : **<a>, <b>, <c>, <d>, <e>, <f>**

2eme étape : Diviser l'espace de recherche et exploiter chaque BD projetée

- <a>- BD projetée
- <b>- BD projetée
- ….
- <f>- BD projetée

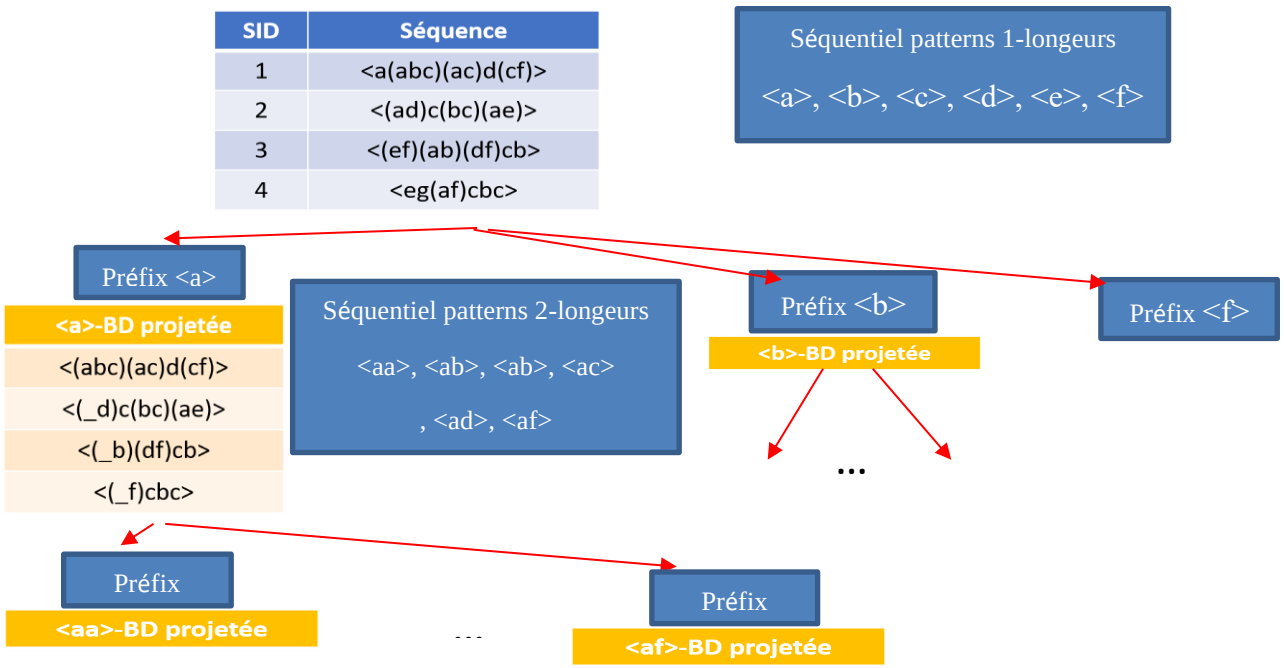


Figure 2- 5:PrefixSpan : Extraction de bases de données projetées de préfixes.

c. Limite de PrefixSpan :

Algorithme ***PREFIXSPAN*** : il s'agit de l'algorithme de base utilisé pour écrire des séquences et représenter des données en projection horizontale. Pendant le processus d'analyse, l'emplacement des données dans la mémoire principale est projeté sur les données structurées de la base de données. Données basées sur des préfixes qui existent habituellement. La cohérence permet de réaliser des bénéfices substantiels en dirigeant les travaux d'extraction vers la partie utile de la base. Un autre avantage de cette idée est que l'extraction ne peut se faire que de deux manières. De plus, cette méthode ne génère pas de séquences candidates. Le prix payé est une énorme limite de mémoire, toutes les projections doivent tenir en mémoire, ce qui n'est pas garanti dans toutes les situations.

2.3.3.4. SPAM :

Proposé par J. Ayres, J. Flannick, J. Gehrke et T. Yiu, cet algorithme utilise une méthode de parcours en profondeur combinée à une représentation bitmap verticale pour stocker chaque séquence, permettant une compression bitmap significative ainsi qu'un comptage de support efficace. SPAM utilise une représentation bitmap verticale des données qui sont créées pour chaque élément du jeu de données. Chaque bitmap contient un bit représentant chaque transaction dans le jeu de données, si l'élément i apparaît dans la transaction j, alors le bit relatif à la transaction j du bitmap pour l'élément i est mis à 1, sinon il est mis à 0. Un comptage et une génération de candidats efficaces peuvent être activés si le bitmap doit être partitionné afin de

s'assurer que toutes les séquences de transaction dans la base de données apparaissent ensemble dans le bitmap. L'idée de représentation bitmap de SPAM nécessite beaucoup de mémoire, elle est donc très efficace pour les bases de données qui ont des modèles séquentiels très longs. En outre, une caractéristique importante de cet algorithme est la production de nouveaux ensembles fréquents en ligne et de manière incrémentielle. Les résultats expérimentaux montrent que cet algorithme est plus efficace que SPADE et PrefixSpan sur les grands ensembles de données, mais il consomme plus d'espace que SPADE et PrefixSpan [26].

*a. **Algorithme SPAM** :*

```
1. DFS-Pruning (node n = (S1, ...., Sk), Sn , In)
2. Début
3.      Stemp = φ;
4.      Itemp = φ;
5.      Pour chaque (i ∈ Sn)
6.           si ((s1, ....., sk , {i}) est fréquent)
7.                Stemp = Stemp U{i};
8.           Fin si
9.      Fin pour
10.     Pour chaque (i ∈ Stemp)
11.          DFS-Pruning((s1,.............,sk,{i}),Stemp,
12.          tous les éléments de Stemp supérieurs à i );
13.     Fin pour
14.     Pour chaque (i∈In)
15.          si ((s1, ..... , sk " {i}) est fréquent)
16.               Itemp = Itemp U {i};
17.          Fin si
18.     Fin pour
19.     Pour chaque (i∈Itemp)
20.          DFS-Pruning ((S1, ......, Sk U {i}), Stemp,
21.          tous les éléments de Itemp supérieurs à i);
22.     Fin pour
23.Fin
```

Figure 2- 6: Algorithme SPAM

Vous allez trouver plus de détails sur cet algorithme dans le chapitre suivant, expliquant la phase d'élagage et la représentation des données avec un exemple

2.3.4. Algorithmes basés sur des motifs séquentiels fermés :

Les algorithmes d'exploration de motifs séquentiels présentés précédemment explorent l'ensemble complet de sous-séquences fréquentes satisfaisant à un seuil de support minimal. Néanmoins, comme une longue séquence fréquente contient un nombre combiné de sous-séquences fréquentes, le processus d'extraction génère un grand nombre de sous-séquences fréquentes pour les motifs longs, ce qui est coûteux en temps et en espace. L'extraction de motifs fréquents (itemsets et séquences) n'a pas besoin d'extraire tous les motifs fréquents, mais

seulement les motifs fermés, car cela permet d'obtenir une meilleure efficacité, ce qui peut réellement réduire le nombre de sous-séquences fréquentes. Nous présentons, dans la section suivante l'algorithme reconnu CloSpan[20] :

2.3.4.1. CloSpan :

Proposé par X.Yan et R.Afshar en 2003 pour réduire le coût en temps et en espace lors de la génération d'un nombre explosif de motifs de séquences fréquentes. CloSpan n'exploite que les sous-séquences fréquentes fermées (les séquences ne contenant aucune super séquence avec le même support), au lieu d'exploiter l'ensemble complet des sous-séquences fréquentes. Le processus d'extraction utilisé par CloSpan est divisé en deux étapes. La première étape génère un ensemble de candidats qui est plus grand que l'ensemble final de séquences fermées. Cet ensemble est appelé ensemble de séquences fermées suspectes (le sur-ensemble de l'ensemble de séquences fermées). Une méthode d'élagage est appelée dans la deuxième étape pour éliminer les séquences non fermées. La principale différence entre CloSpan et PrefixSpan réside dans l'implémentation de CloSpan qui est un mécanisme de terminaison précoce qui évite de parcourir inutilement l'espace de recherche. L'utilisation des méthodes de sous-modèle et de super-modèle à rebours permet d'absorber ou de fusionner certains modèles, ce qui réduit considérablement la croissance de l'espace de recherche [27].

a. Algorithme CloSpan :

```
1. CloSpan (s, Ds, min_supp, L)
2. Entrées : séquence s, base de données projeté Ds, et le support minimal
3. Sortie :  le treillis de recherche de préfixes L.
4. Debut
5.      Vérifier s'il existe une séquence découverte s' telle que soit s ⊆ s',
6.      soit s'⊆ s, et la taille de la base de données L(Ds)=L(Ds').
7.      si  super-modèle ou sous-modèle existe, alors
8.           Modifier le lien en L ;
9.           Retour
10.     sinon, insérer s dans L ;
11.          Scanner Ds une fois, trouver l'ensemble d'items fréquents α tel que
12.               α peut être annexé pour former un motif séquentiel s ◊ α.
13.          Si aucun α valide n'est disponible, alors
14.               Return
15.          Fin si
16.     Fin si
17.     Pour chaque α valide, faire
18.          CloSpan(s ◊ α, Ds ◊ α, minsupp, L)
19.          Return
20.     Fin pour
21.fin
```

Figure 2- 7:Algorithme CloSpan

2.3.5. Algorithmes basés sur l'incrémentation :

Dans le cadre de l'exploration de motifs séquentiels, l'algorithme incrémental peut être utilisé pour l'exploration des mises à jour fréquentes et incrémentielles des bases de données (insertions et suppressions). Nous distinguons deux cas pour développer un algorithme incrémental [20] :

- ***Les séquences complètes (modèle de séquence)*** sont insérées dans et/ou supprimées de la base de données originale ;
- ***La base de données originale*** contient une séquence qui est mise à jour en ajoutant de nouvelles transactions à la fin.

2.3.5.1. IncSpan :

IncSpan est un algorithme proposé dans [27] utilisé pour l'exploration incrémentale sur de multiples incréments de base de données. Le développement de l'algorithme ***IncSpan*** est basé sur deux idées nouvelles. La première idée qui présente plusieurs bonnes propriétés et conduit à des pratiques efficaces est l'utilisation d'un ensemble de séquences "presque fréquentes" comme candidats dans la base de données mise à jour. La seconde idée est constituée de deux techniques d'optimisation conçues pour améliorer les performances, à savoir la correspondance inverse de motifs et la projection partagée. La première technique est utilisée pour faire correspondre un motif séquentiel dans une séquence. La mise en correspondance de motifs inversés peut élaguer l'espace de recherche supplémentaire, alors que les transactions ajoutées se trouvent à la fin d'une séquence. La projection partagée est destinée à réduire le nombre de projections de la base de données pour certaines séquences ayant un préfixe commun [28].

b. Algorithme IncSpan :

```
1. IncSpan (D', min_sup, μ, FS, SFS)
2. Début
3.      Entrée: Une base de données ajoutée D' , min_sup, μ, des séquences fréquentes
4.           FS dans D, des séquences semi-fréquentes SFS dans D.
5.      Sortie: FS' et SFS'
6.      FS' = ø
7.      SFS' = ø
8.      Scanner la LDB pour les articles individuels
9.      Ajouter un nouvel item fréquent dans FS' ;
10.     Ajouter un nouvel item semi-fréquent dans SFS' ;
11.     pour chaque nouvel élément i dans FS', faire
12.          PrefixSpan(i, D'|i, μ * min sup, FS', SFS') ;
13.     Fin pour
14.     pour chaque motif p dans FS ou SFS, faire
15.          vérifier Δsup(p) ;
16.          si sup(p) = supD(p) + Δsup(p) ≥ min sup
17.               insérer(FS', p) ;
18.               si supLDB(p) ≥ (1 - μ)min sup
19.                    PrefixSpan(p, D'|p, μ * min sup, FS', SFS') ;
20.          Sinon
21.               insert(SFS', p) ;
22.          Fin si
23.     Fin pour
24.     retour ;
25.Fin
```

Figure 2- 8:Algorithme IncSpan

Caractéristiques comparatives de différents algorithmes d'extraction de motifs séquentiels

2.4. Conclusion :

Bien que le concept de l'exploration de données séquentielles soit nouveau, il a fait des progrès considérables en quelques temps. Plusieurs approches concernées par l'extraction de motifs séquentiels ont été proposées pour améliorer l'efficacité des algorithmes, soit avec de nouvelles structures, de nouvelles approches ou par la gestion de la base de données dans la mémoire de l'ordinateur.

Par conséquent, dans ce chapitre on a classifié l'exploration de motifs séquentiels en cinq classes principales (parmi d'autres classes), Apriori, DFS, BFS, les motifs séquentiels fermés et les algorithmes basés sur les motifs incrémentaux.

Chapitre 3 : *Contribution*

3.1. Introduction :

Dans ce chapitre nous détaillant la nouvelle approche nommé E-SPAM (**E**nhanced **S**equential **PA**ttern **M**ining) en appliquant la récursivité sur la forme générale de SPAM, les simulations faites sur les deux algorithmes déjà mentionné ont montrées la supériorité de notre approche E-SPAM en termes de temps et consommation de la mémoire.

3.2. SPAM :

Comme on a vu dans le chapitre 2, ***SPAM*** est l'un des algorithmes d'exploration de motifs séquentiels de type ***DFS***. Il suppose que toute la base de données utilisée pour l'algorithme tient entièrement dans la mémoire. Pour les séquences données, il génère un arbre lexicographique. La racine de l'arbre commence toujours par une chaîne vide. Les nœuds enfants de l'arbre sont formés par une séquence étendue ou une séquence étendue d'éléments. Pour une séquence contenante ***a, b, c, d***, nous devons générer une analyse lexicographique de l'arbre, qui commence alors par le nœud racine {}[29].

Soit la base de données suivante :

SID	Séquence
1	({a, b, d}, {b, c, d}, {b, c, d})
2	({b}, {a, b, c})
3	({a, b}, {b, c, d})

Tableau 3- 1:Base de données de séquence

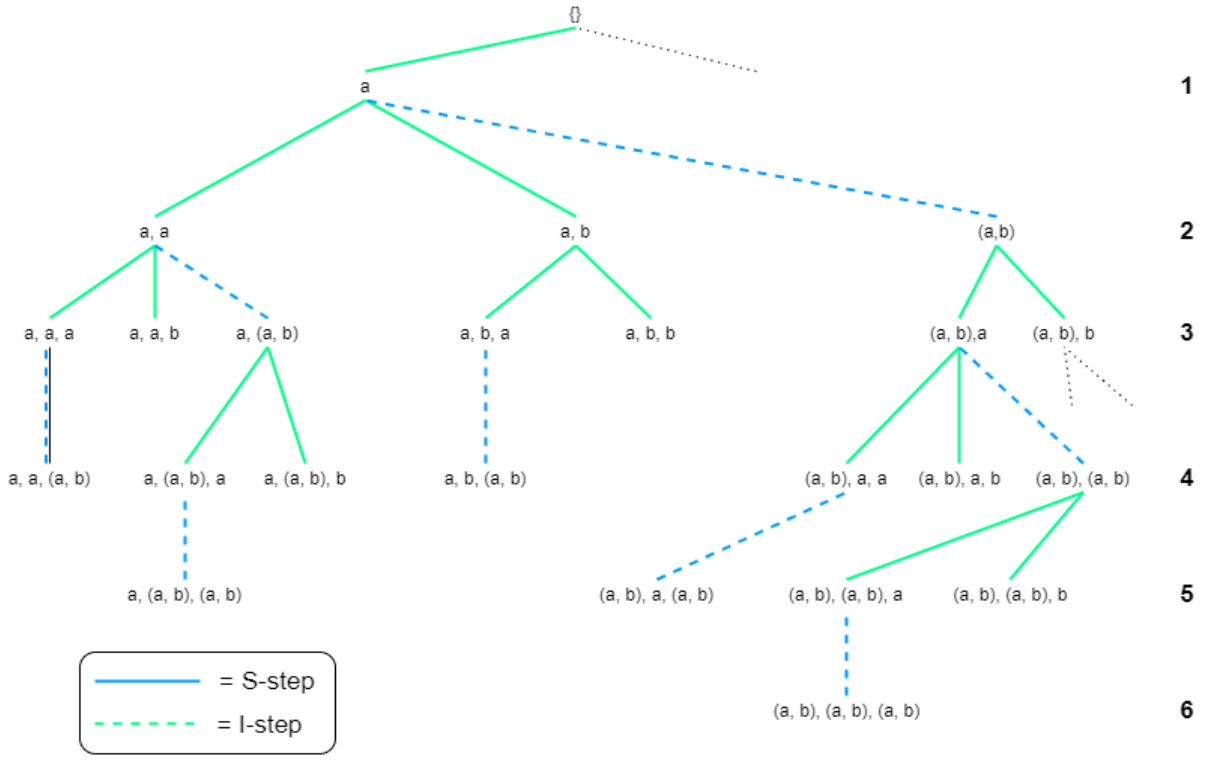


Figure 3- 1:L'arbre séquentiel lexicographique

Au niveau 1, les éléments ***a, b, c*** et d sont considérés séparément. Considérons d'abord l'élément a, pour la génération de candidats, il prendra l'étape d'extension de séquence (s-step) et générera une séquence différente comme ***{a, a}, {a, b}.*** De même, il génère une séquence élargie de l'ensemble d'éléments par l'étape d'extension de l'ensemble d'éléments (ou I-step) comme ***(a, a)*** au niveau 2. De même, avec différents éléments au niveau 3, on génère une séquence de longueur 3.

Une fois que l'arbre complet est généré pour découvrir ou rechercher un ensemble de sous-séquences spécifié par l'utilisateur, l'algorithme traverse la stratégie de recherche en profondeur. À chaque nœud, le support de la séquence est testé. Les noyaux ayant une valeur de support supérieure ou égale au support minimal sont stockés et répète DFS récursivement sur ces noyaux. Sinon, le noyau n'est pas considéré par le principe d'Apriori.

Pour l'algorithme de comptage de support, il utilise une structure bitmap verticale. Pour le tableau 1(tableau12), la représentation bitmap des données peut être donnée comme ci-dessous. Comme la représentation bitmap de la longueur maximale de la séquence est de 3, le bitmap vertical est composé de 3 bits et comme le nombre de séquences est de 3 dans le tableau 1, les bitmaps ont 3 emplacements.

SID	TID	{a}	{b}	{c}	{d}
1	1	1	1	0	1
1	3	0	1	1	1
1	6	0	1	1	1
-	-	0	0	0	0
2	2	0	1	0	0
2	4	1	1	1	0
-	-	0	0	0	0
-	-	0	0	0	0
3	5	1	1	0	0
3	7	0	1	1	1
-	-	0	0	0	0
		0	0	0	0

Tableau 3- 2:Représentation bitmap de l'ensemble des données dans le tableau -1-

Pour la présentation bitmap **de S-Step,** chaque bit après le premier indice de 1 est mis à zéro et chaque bit après cette position d'indice est mis à 1.

({a})		({a})s		{b}		({a}, {b})
1		0		1		0
0		1		1		1
0		1		1		1
0		1		0		0
0	S-step → Processus	0	&	1	Résultat →	0
1		1		1		0
0		1		1		0
0		1		0		0
0		1		0		0
1		0		1		0
0		1		1		1
0		1		0		0
0		1		0		0

Tableau 3- 3:Traitement de S-step sur le bitmap de la séquence ({a}) représentée sur la tableau -2-

Pour la représentation bitmap de la I-Step, les bitmaps des ensembles d'éléments nouvellement ajoutés sont logiquement combinés par ET avec la séquence générée par la S-Step.

({a}, {b})		{d}		({a}, {b, d})
0		1		0
1		1		1
1		1		1
0		0		0
0	&	0	Résultat →	0
0		0		0
0		0		0
0		0		0
0		0		0
1		1		1
0		0		0
0		0		0

Tableau 3- 4:Traitement de I-step sur la séquence bitmap ({a}, {b}) représentée sur le tableau -3-.

Pour améliorer les performances de l'algorithme, des techniques d'élagage sont utilisées avec la **S-extension** et **l'I-extension** d'un nœud. La technique d'élagage **S-step** élague ***les enfants S-step.*** Pour l'élagage, elle applique le principe d'Apriori, c'est-à-dire que la séquence **({a}{a})** et **({a}, {b}))** est donnée et si **({a}{b}** n'est pas fréquente, alors **({a}{a}{b}), ({a}{c}{b})** ou **({a}, {a, b})** ou **({a}, {c, b})** sont ignorés. De même, la technique d'élagage I-step élague les I-step enfants. Pour l'élagage à **I-step**, elle applique le même principe d'Apriori pour l'ensemble d'éléments, c'est-à-dire que pour les séquences d'ensembles d'éléments **({a, b})** et **({a, c})**, si **({a, c})** n'est pas fréquent, alors **({a, b, c})** n'est pas fréquent.

Des contraintes peuvent être ajoutées comme l'écart minimum et maximum entre les deux éléments. Avec les contraintes mingap et maxgap, l'étape de transformation est modifiée pour restreindre le nombre de positions dans lesquelles l'élément suivant peut apparaître après le

premier élément. Si le premier élément est **{a}** et le suivant **{b}** et que la contrainte **mingap=1** et **maxgap=1**, alors toutes les séquences **{a, c, b}, {a, b, b}** etc. sont des séquences de l'ensemble de données mentionné précédemment [26].

Les expressions régulières peuvent également être utilisées pour limiter le nombre de motifs intéressants. Si **a + b** est donné, on considère que toute la séquence contenante a ou a peuvent être obtenues.

3.3. E-SPAM :

Notre proposition repose sur la modification de l'algorithme SPAM, cette amélioration réside dans la manière d'exécution de cet algorithme.

En fait on a pensé à remplacer la fonction DFS-puring par quatre fonction 'sSteps', 'sPatternRecorded', 'iStep', 'iPatternRecorded'.

A l'origine la fonction 'DFS-puring' fonctionne d'une manière itérative ce qui peut alourdir l'exécution, demande plus d'espace en mémoire (tableau 5), la communication entre les quatre fonctions proposées ainsi que leur implémentation sont présentées dans la figure 12, il est important de tirer l'attention que les fonctions sont récursives et qu'elles sont complémentaire d'une manière à gagner plus d'espace mémoire et de garantir un temps de repense acceptable. Les résultats de simulation confirment ce gain (section 3, 5).

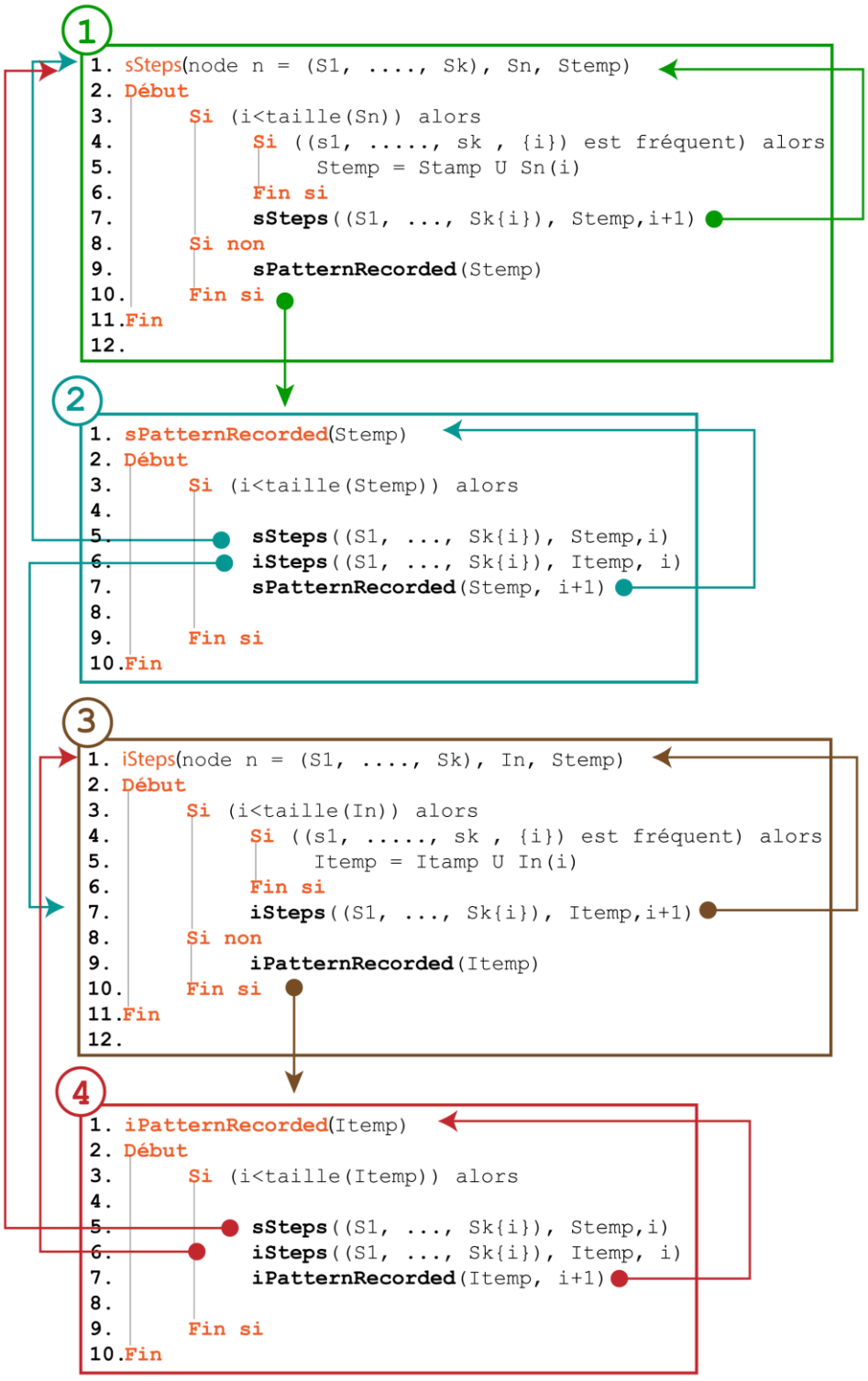


Figure 3- 2:Algorithme E-SPAM

3.4. Les outils de développement

3.4.1. IntelliJ IDEA :

IntelliJ IDEA également appelé « IntelliJ », « IDEA » ou « IDJ » est un environnement de développement intégré (en anglais Integrated Development Environment - IDE) destiné au développement de logiciels informatiques reposant sur la technologie Java. Il est développé par JetBrains (anciennement « IntelliJ ») et disponible en deux versions, l'une communautaire, open source, sous licence Apache 2 et l'autre propriétaire, protégée par une licence commerciale. Tous deux supportent les langages de programmation Java, Kotlin, Groovy et Scala.[30]

3.4.2.Java :

Java est un langage de programmation orienté objet créé par James Gosling et Patrick Naughton, employés de Sun Microsystems, avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982), présenté officiellement le 23 mai 1995 au SunWorld.

La société Sun a été ensuite rachetée en 2009 par la société Oracle qui détient et maintient désormais Java.

Une particularité de Java est que les logiciels écrits dans ce langage sont compilés vers une représentation binaire intermédiaire qui peut être exécutée dans une machine virtuelle Java (JVM) en faisant abstraction du système d'exploitation. [31]

3.4.3.SPMF

SPMF est une bibliothèque de fouille de données à code source ouvert, spécialisée dans la fouille de motifs, offrant des implémentations de plus de 120 algorithmes de fouille de données. Elle a été utilisée dans plus de 310 articles de recherche pour résoudre des problèmes appliqués dans un large éventail de domaines allant de l'attribution d'auteurs à la recommandation de restaurants. Ses implémentations sont également couramment utilisées comme points de référence dans les articles de recherche, et il a également été intégré dans plusieurs logiciels d'analyse de données. Après trois ans de développement, cet article présente la deuxième révision majeure de la bibliothèque, appelée SPMF 2, qui fournit (1) plus de 60 nouvelles implémentations d'algorithmes (y compris de nouveaux algorithmes pour la prédiction de séquences), (2) une interface utilisateur améliorée avec visualisation de modèles (3) un nouveau système de plug-in, (4) des performances améliorées, et (5) un support pour l'exploration de textes. [32]

3.5. Résultats de l'exécution des algorithmes SPAM et E-SPAM :

Dans cette section on va voire les résultats d'exécution des deux algorithmes sur deux déférentes bases de données et minsup, les résultats sont représentés sur les graphes et les tableaux ci-déçus

3.5.1. La base de données de séquence :

Le format de fichier d'entrée est un fichier texte où chaque ligne représente une séquence. Chaque élément d'une séquence est un entier positif (> 0) et il est séparé par la valeur "-1". La valeur "-2" indique la fin d'une séquence (elle apparaît à la fin de chaque ligne). Par exemple, le fichier d'entrée peut contenir les deux lignes suivantes (deux séquences).

3.5.1.1. ***Base de données 1*** *:*

Cette base de données contient plus de 700 séquences et la taille moyenne des séquences et 100

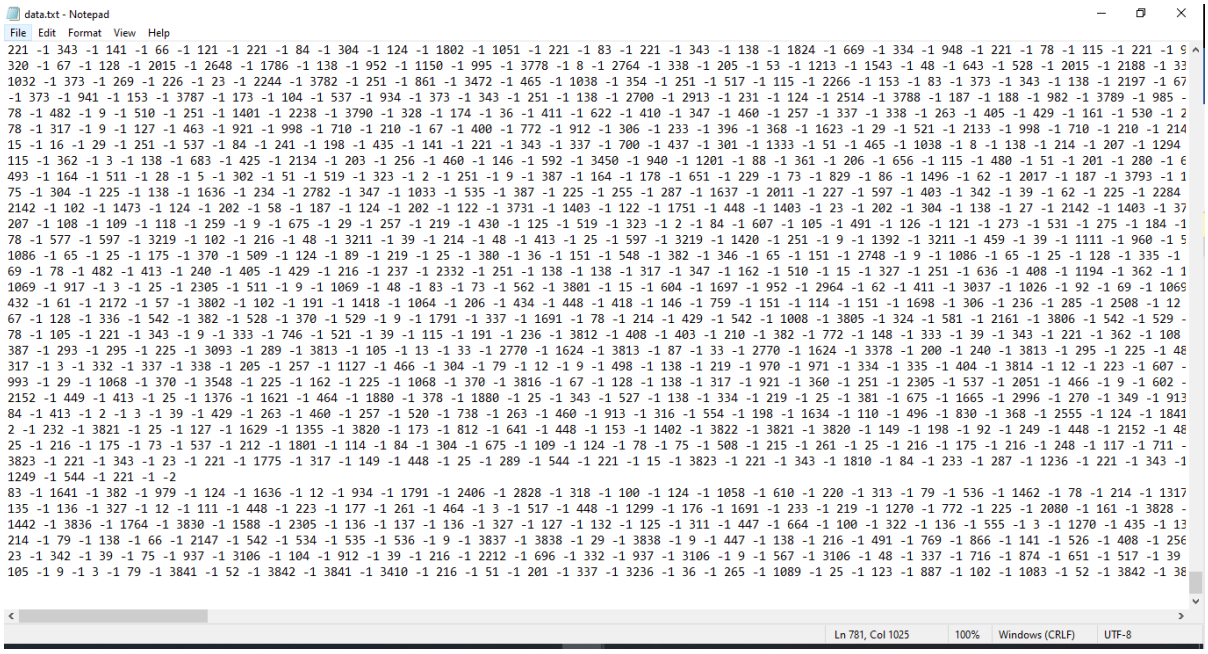


Figure 3- 3:Base de données de séquence 1

Algorithme	Critère Minsup	5	10	15	20	25	50
E-SPAM	Mémoire (MB)	460.24	305.07	184.25	149.07	136.94	132.01
	Temps (S)	409.54	65.85	26.47	15.97	8.71	3.2
SPAM	Mémoire (MB)	456.70	298.14	180	173.3	152.4	144.67
	Temps (S)	425.53	71.87	30.45	18.33	10.03	4.2

Tableau 3- 5:Résultats de l'exécution des algorithmes SPAM et E-SPAM avec les minsup (5, 10, 15, 20, 25,50)

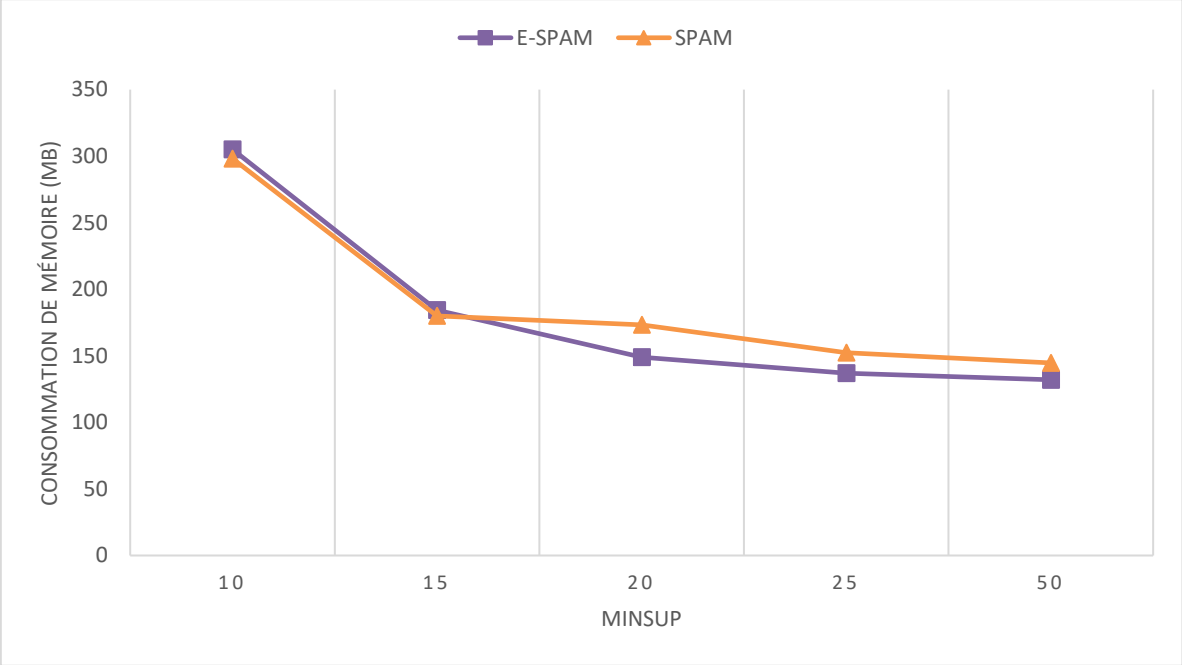


Figure 3- 4: Résultats obtenus de minsup (10, 15, 20, 25, 50) par rapport à la consommation de mémoire

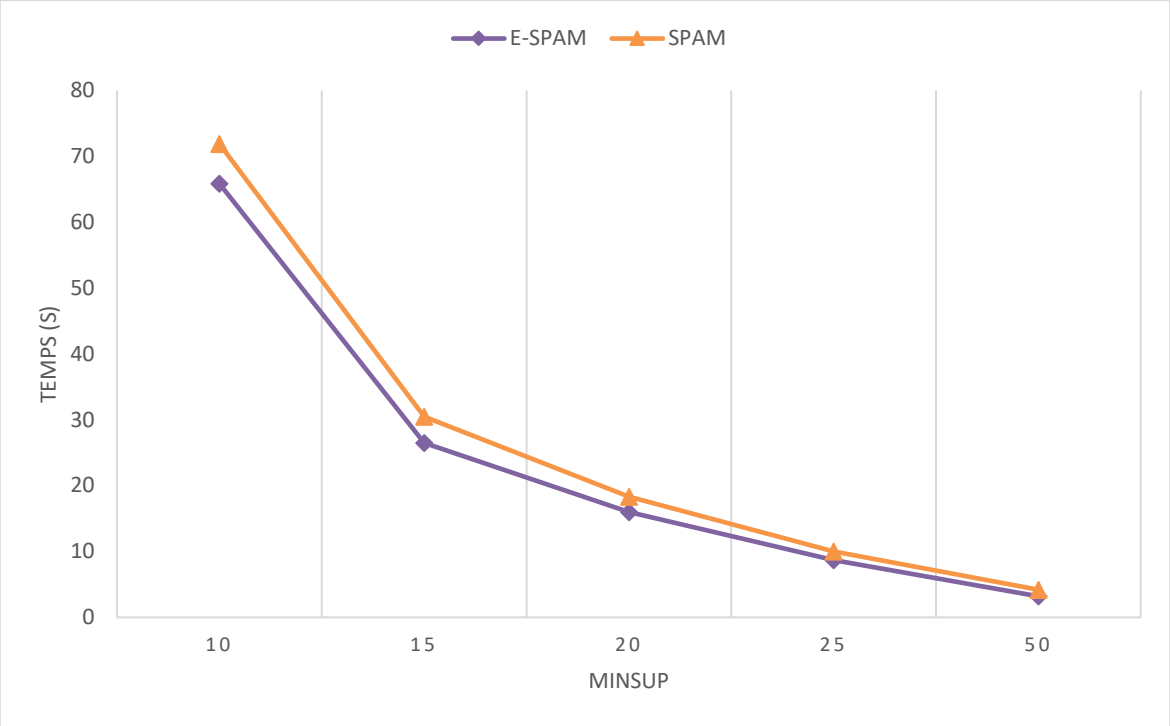
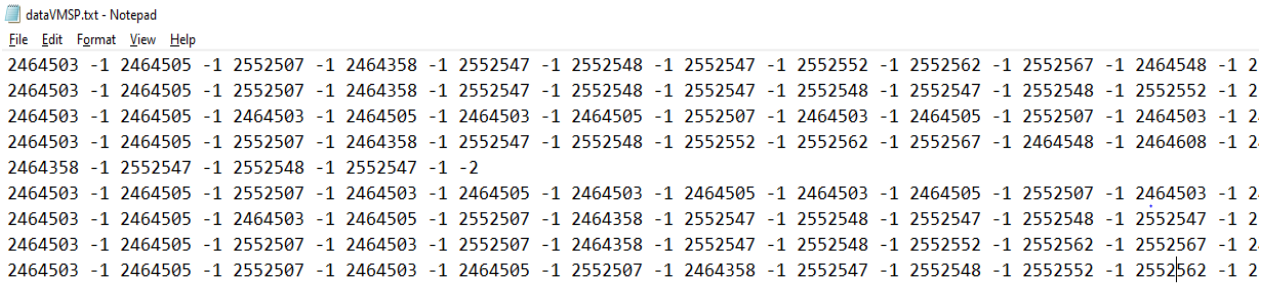


Figure 3- 5:Résultats obtenus de minsup (10, 15, 20, 25, 50) par rapport au temps

3.5.1.2. Base de données 2 :

Cette base de données contient 9 séquences et la taille moyenne des séquences et 30


dataVMSP.txt - Notepad

File Edit Format View Help

```
2464503 -1 2464505 -1 2552507 -1 2464358 -1 2552547 -1 2552548 -1 2552547 -1 2552552 -1 2552562 -1 2552567 -1 2464548 -1 2
2464503 -1 2464505 -1 2552507 -1 2464358 -1 2552547 -1 2552548 -1 2552547 -1 2552548 -1 2552547 -1 2552548 -1 2552552 -1 2
2464503 -1 2464505 -1 2464503 -1 2464505 -1 2464503 -1 2464505 -1 2552507 -1 2464503 -1 2464505 -1 2552507 -1 2464503 -1 2
2464503 -1 2464505 -1 2552507 -1 2464358 -1 2552547 -1 2552548 -1 2552552 -1 2552562 -1 2552567 -1 2464548 -1 2464608 -1 2
2464358 -1 2552547 -1 2552548 -1 2552547 -1 -2
2464503 -1 2464505 -1 2552507 -1 2464503 -1 2464505 -1 2464503 -1 2464505 -1 2464503 -1 2464505 -1 2552507 -1 2464503 -1 2
2464503 -1 2464505 -1 2464503 -1 2464505 -1 2552507 -1 2464358 -1 2552547 -1 2552548 -1 2552547 -1 2552548 -1 2552547 -1 2
2464503 -1 2464505 -1 2552507 -1 2464503 -1 2552507 -1 2464358 -1 2552547 -1 2552548 -1 2552552 -1 2552562 -1 2552567 -1 2
2464503 -1 2464505 -1 2552507 -1 2464503 -1 2464505 -1 2552507 -1 2464358 -1 2552547 -1 2552548 -1 2552552 -1 2552562 -1 2
```

Figure 3- 6:Base de données de séquence 2

Algorithme	Critère Minsup	2	3	4	5	6	9
E-SPAM	Mémoire (MB)	107	100.46	73.86	61.35	14.08	2.28
	Temps (S)	65.83	3.95	1.94	0.691	0.2	0.15
SPAM	Mémoire (MB)	129	107.46	73.86	61.38	14.08	2.28
	Temps (S)	69.54	4.28	2.08	0.718	0.28	0.15

Tableau 3- 6:Résultats de l'exécution des algorithmes SPAM et E-SPAM avec les minsup (2, 3, 4, 5, 6,9)

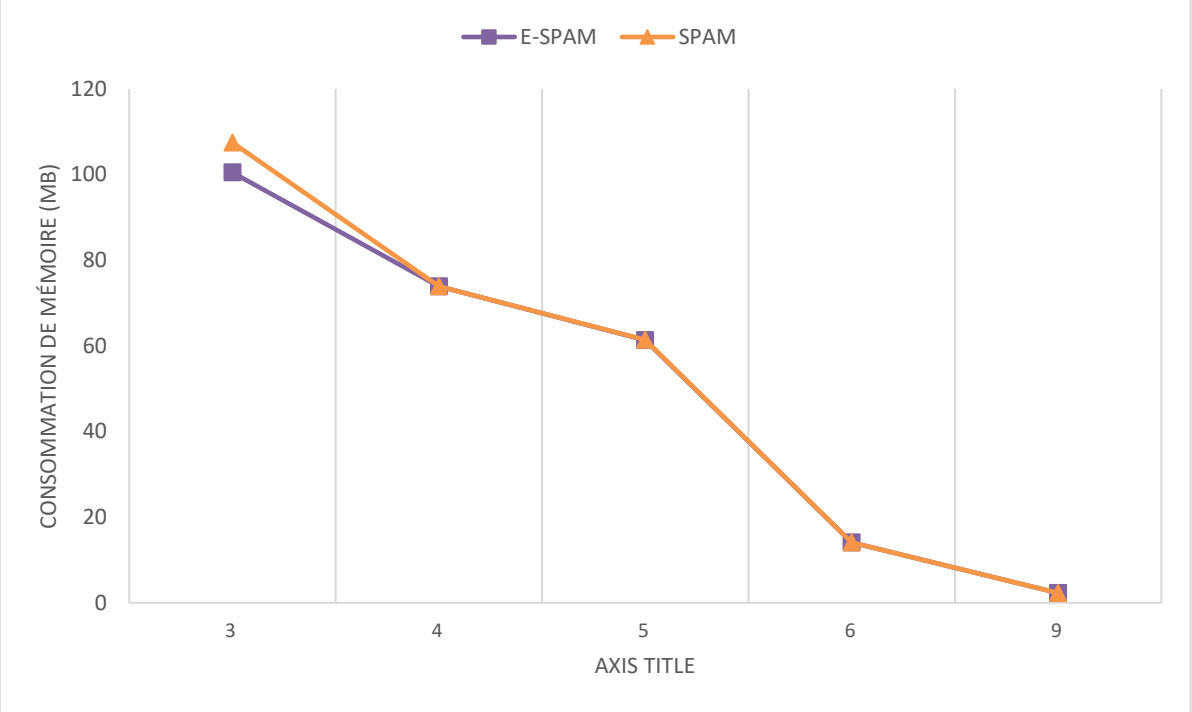


Figure 3- 7:Résultats obtenus de minsup (3, 4, 5, 6, 9) par rapport à la consommation de mémoire

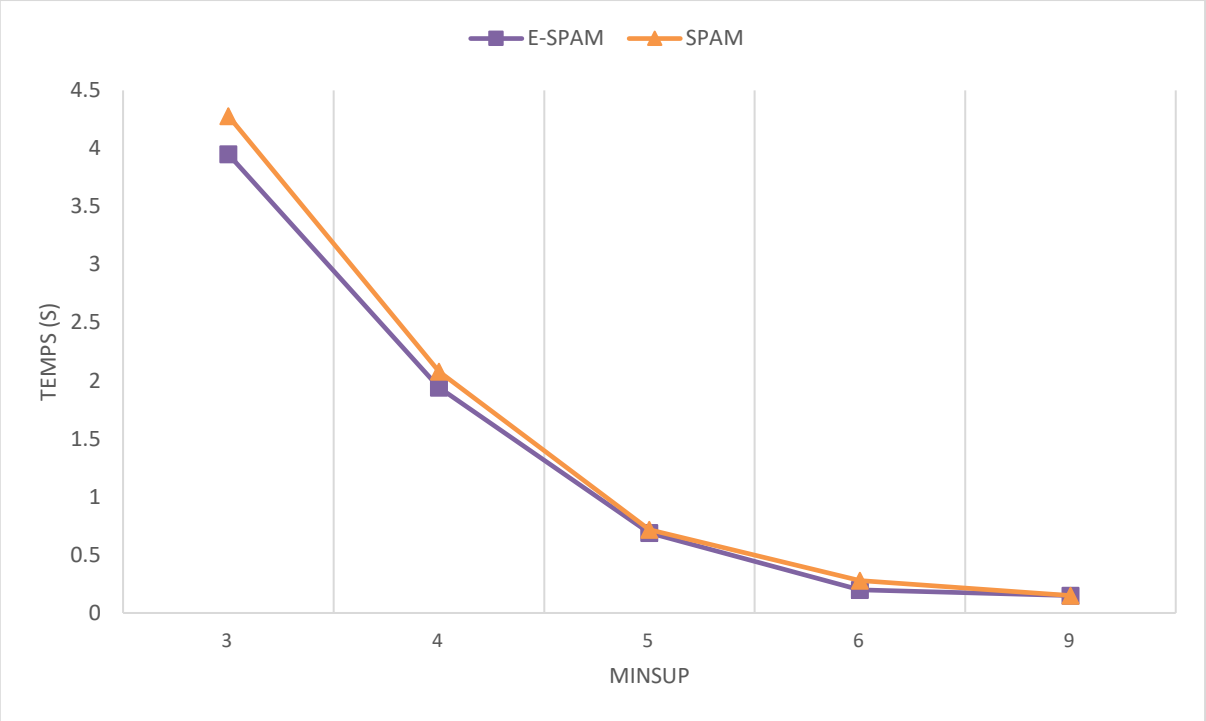


Figure 3- 8:Résultats obtenus de minsup (3, 4, 5, 6, 9) par rapport au temps

3.6. Comparaison

Après plusieurs tests et recensements des résultats de simulation comme le montre les figures (14, 15, 17 et 18), notre amélioration a donnée des résultats meilleurs d'un point de vue espace on temps d'exécution, et ce dans les valeurs minsup, plus que la valeur minsup augmente les deux courbes converge, mais ça n'empêche qu'il reste toujours une amélioration par rapport à l'algorithme originale. Cette diminution revient à la taille de la base des séquences qui va augmenter de plus en plus et par conséquent nécessite de ressources de stockage et de temps de traiter

3.7. Conclusion :

Dans ce chapitre nous avant montré les différents outils utilisés pour l'implémentation de l'algorithme E-SPAM et son algorithme avec l'explication de son fonctionnement, on a présenté les résultats de la simulation des deux algorithmes sur plusieurs base de données des séquences, ces résultats ont montré la supériorité de notre approche E-SPAM en termes de temps.

Conclusion générale :

Dans le cadre de ce mémoire de fin d'étude on a implémenté une nouvelle approche nommé E-SPAM basé sur l'algorithme SPAM utilisant la récursivité pour le but d'optimisé le temps d'exécution et la complexité mémoire et nous avant montré la supériorité de notre approche par rapport à l'algorithme SPAM au terme de temps et la consommation de la mémoire.

Lors de ce travail nous avant présenté la notion de data mining qui représente le processus d'analyse de volumes massifs de données et du Big Data sous différents angles afin d'identifier des relations entre les data et de les transformer en informations exploitables, les déférents tâches de data mining et les techniques qui peuvent être appliqué pour accomplir ces tâches.

Par la suite nous avant présenter le domaine extraction des motifs séquentiels et les déférent catégories des algorithmes de ce domaine et quelque algorithme dans chaque catégorie des exemples explicatif et leurs pseudo algorithme.

Et présentation de notre nouvelle approche E-SPAM avec une explication de la modification qu'on a fait sur l'algorithme SPAM avec son pseudo Algorithme et une comparaison entre les deux algorithmes avec déférant base de données de séquence qui montre la supériorité de notre algorithme sur le terme de temps et la consommation de la mémoire.

A partir des points que nous avons étudier, nous pouvons dire que le domaine du data mining est capable d'évoluer et riche en sujets de recherche qui attend quelqu'un qui vient relever ses défis.

Référence :

[1] https://www.coursehero.com/file/84733162/Cours-DataMining-R-seance1pdf/

[2] https://www.lebigdata.fr/data-mining-definition-exemples%2018/04/2021%2023 :00

[3] Jaiwei Han, Micheline Kamber Jian Pei : data mining concepts et technique

[4] https://www.kmit.in/department/LectureNotes/DM%20UNIT-I%20Notes.pdf

[5] http://www.univ-usto.dz/theses_en_ligne/doc_num.php?explnum_id=669

[6] https://www.mediafinances.net/informatique/donnees-data-mining/

[7] https://jafwin.com/2019/07/05/lessentiel-a-savoir-sur-le-data-mining/

[8] https://www.petite-entreprise.net/P-2595-83-G1-principales-taches-du-data-mining.html

[9] http://cedric.cnam.fr/vertigo/cours/ml2/coursArbresDecision.html

[10] http://eric.univ-lyon2.fr/~ricco/cours/slides/Arbres_de_decision_Introduction.pdf

[11] https://www.techno-science.net/glossaire-definition/Algorithme-genetique.html

[12] Jaiwei Han, Micheline Kamber Jian Pei : data mining concepts et technique

[13] Oualid Ouarem, Farid Nouioua, and Philippe Fournier-Viger; Mining Episode Rules From Event Sequences Under Non-Overlapping Frequency. 2021

[14] http://depot-e.uqtr.ca/id/eprint/1201/1/030110265.pdf

[15] Marc Plantevit. Extraction De Motifs Séquentiels Dans Des Données Multidimensionelles. Informatique [cs]. Université Montpellier II - Sciences et Techniques du Languedoc, 2008. Français. Fftel00319242f. 7 September 2008

[16] Florent Masseguelame, Teisseire Pascal Poncelet. Extraction de motifs séquentiels. Problème et méthode revue des sciences et technologies de l'information série T+ISI : Ingénierie des Systèmes d'Information Lavoisier 2004, 9 (3/4), pp.183-210.10.3166/isi 9.3-4.183-210 lirnum 00108563. 3 novembre 2018

[17]https://portail.polytechnique.edu/datascience/fr/recherche/quelques-domaines-dapplication%2019/04/2021%2010%20:45

[18] Dr. Shalini Bhaskar Bajaj, et Deepika Garg survey on Sequence mining Algorithms. November 2016

[19] Jawahar. S. A comparative study of Sequential patterns mining Algorithms. December 2015

[20]Thabet Slimani, and Amor Lazzez SEQUENTIAL MINING: PATTERNS AND ALGORITHMS ANALYSIS Computer Science, Taif University & LARODEC Lab, Saudia Arabia, Computer Science, Taif University, Saudia Arabia. Janvier 2013

[21] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, Yun Sing Koh, Rincy Thomas A survey of Sequential pattern mining. Février 2017

[22] Allia Mohamed Rachid Bouadi Tassadit, El Motaouakil Sami, Keira Mamadou

Fouille de données : règles séquentielles Université Montpellier II

[23] Marion Leleu, Christophe Rigotte, Jean François Boulinant, et Guillonne Enviand. GSpade : mining Sequential patterns over datasets with consecutive Repetitions 2003

[24] Jian Pei Jiawei Han Behzad Mortazavi-Asl Helen Pinto Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth The work was supported in part by the Natural Sciences and Engineering Research Council of Canada (grant NSERC-A3723), the Networks of Centres of Excellence of Canada (grant NCE/IRIS-3), and the Hewlett-Packard Lab, U.S.A.

[25] Patrik saraf, R. R sedamker, sheettal rathi. PrefixSpan Algorithm for finding sequential patterns with various constraints. 3 Janvier 2015

[26] Jay Ayres, Jason Flannick, Johannes Gehrke, Tomi Yiu sequential pattern mining using a bitmap representation. 2002

[27] Gregories S. Budhi, Yulin, hery Gurnawan. Clospan sequential pattern mining for books recommendation university library. Juin 2013

[28] Hang Cheng, Xifen Yan, Jiawei Han. IncSpan: Incremental mining od sequential patterns in large database. May 2005

[29] Joshua Ho, Lior Lukov, Sanjay Chawla International Sequential Pattern Mining with Constraints on Large Protein Databases Conference on Management of Data COMAD 2005b, Hyderabad, India, December 20–22, 2005 °c Computer Society

[30]https://www.google.com/search?q=intellij+idea+%3F&rlz=1C1SQJL_frDZ928DZ928&oq=intelli&aqs=chrome.1.69i57j69i59j35i39j0i67j0i67i433j69i60j69i61l2.5615j0j7&sourceid=chrome&ie=UTF-8

[31] https://fr.wikipedia.org/wiki/Java_(langage)

[32]https://www.researchgate.net/publication/307587257_The_SPMF_Open-Source_Data_Mining_Library_Version_2