الجمهورية الجزائرية الديمقراطية الشعبية

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي و البحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة غرداية

Université de Ghardaia

كلية العلوم والتكنلوجيا

Faculté des Sciences et de Technologie

قسم الرياضيات و الاعلام الآلي

Département des Mathématiques et Informatique

MEMOIRE

Présenté pour l'obtention du **diplôme** de **MASTER**

En : Informatique

Spécialité : Systèmes Intelligents pour l'Extraction des Connaissances.

Par : Khaled BENDIDA et Abdelbasset HAKKOUMI

Sujet

Extraction des Top-k Episodes Réguliers

Soutenu publiquement devant le jury :

M.Slimane BELLAOUAR	Univ Ghardaia	Président
M.Youcef MAHDJOUB	Univ Ghardaia	Examinateur
M.Djelloul ZIADI	Univ Rouen(France)	Encadreur
M.Slimane OULAD NAOUI	Univ Ghardaia	Co-Encadreur

Année Universitaire 2015/2016

Remerciement

En premier lieu, nous tenons à remercier ""الله,notre créateur pour nous avoir la force pour accomplir ce travail.

Nous adressons nos sincères remerciements à notre encadreur monsieur Djelloul Ziadi, Professeur à l'université de Rouen qui n'a ménagé aucun effort pour que ce mémoire puisse voir le jour,sa disponibilité, ainsi pour ses conseils et ses encouragements.

Nous voudrons aussi exprime tout notre gratitude et nos remerciement à monsieur Slimane Oulad-NaouiMaitre-Assistant

à l'université de Ghardaia

pour leurs aides et leur encouragement.

Nous remercions vivement les membres du jury d'avoir accepté d'examiner ce mémoire.

Nos plus chaleureux remerciements pour tous ce qui de prés et de loin ont contribué à l'élaboration de ce travail.

Résumé

Grâce aux développements socio-économiques et aux progrès en sciences médicales, l'espérance de vie a augmenté. Aussi, l'un des défis dans les années futures est la surveillance des sujets âgés afin de les aider à vivre autonomes chez eux. L'apparition des smart-homes et les avancés dans les équipements mobiles notamment les capteurs ont permis la collecte de traces sur les activités quotidiennes des personnes à domicile. L'analyse de ces données peut informer sur l'état de santé des personnes concernées. Il est évident que l'apprentissage supervisé dans ce domaine est inapproprié étant donné qu'il exige de se disposer de données annotées dont l'acquisition est onéreuse et couteuse.

Dans ce mémoire, nous proposons de découvrir les épisodes réguliers dans ces traces. Bien entendu, l'approche se contente de l'extraction des tops k épisodes afin de réduire l'espace de recherche dont l'exploration exhaustive s'avère, outre peux intéressante, inefficace et couteuse.

Mot clé : Fouillée des Données, Fouille de Séquences, Episode Régulier, Flot de Données, Fenêtre glissante.

ABSTRACT

With the socio-economic development and progress in medical science, life expectancy has increased. Also, one of the challenges in the coming years is the monitoring of elderly patients to help them live independent at home. The emergence of smart homes and advanced in mobile devices including sensors have enabled the collection of traces about daily activities of people at home. Analysis of these data can provide information on the health status of the monitored persons. Obviously, supervised learning in this area is inappropriate since it requires to have annotated data with the acquisition is costly and expensive.

In this thesis, we propose to discover regular episodes from data stream. Of course, the approach simply extracts tops-k episodes to reduce the search space whose exhaustive exploration turns out, to be interesting, inefficient and costly.

Keywords: Data Mining, Sequence Mining, Regular Episodes, Data Stream, Sliding Window

ملخص

بفضل التطورات الاجتماعية، الاقتصادية والتقدم في العلوم الطبية، ازداد متوسط العمر المتوقع. وعليه، واحدة من بين التحديات في السنوات المقبلة هو مراقبة المرضى كبار السن لمساعدتهم على العيش المستقل في المنزل، ظهور المنازل الذكية والتطور في الأجهزة المحمولة بما في ذلك أجهزة الاستشعار سمحت بجمع المسارات و تتتبع الأنشطة اليومية للناس في داخل منازلهم. تحليل هذه البيانات يمكن أن يوفر معلومات عن الحالة الصحية للأشخاص المعنيين. ومن الواضح أنالتدريب المراقب في هذا المجال غير مناسب لأنه يتطلب الحصول على البيانات مصنفة التي هي مكلفة وباهظة .

في هذه المذكرة, نقترح اكتشاف الحلقات المنتظمة لهذه المسالك. وبطبيعة الحال, فإن النهج يكتفي ببساطة في استخراج الحلقات ك الأولى لتقليل مساحة البحث لأن الاستكشاف الشامل، إضافة إلى كونه، غير فعالة فانه مكلف.

مفاتيح :

تنقيب البيانات،تنقيب المتسلسلات ،حلقات منتظمة،تدفق البيانات،النوافذ المنزلقة.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Le domaine de l'Extraction de Connaissance dans les bases de Données (ECD ou KDD[1] en anglais) est né pour répondre aux difficultés que l'être humain prouve face à l'analyse de grands volumes de données en perpétuelle augmentation. Ce processus s'intéresse à trouver ce que l'on cherche uniquement et n'a pas besoin de tout savoir c'est-à-dire se focaliser seulement à ce que on peut qualifier d'utile seulement.

Le but est de passer un grand volume d'information par les étapes de l'ECD est de générer un ensemble de connaissances utilisables. Ce processus est similaire à celui de la transformation d'une matière première vers un produit fini.

Parmi les étapes de processus ECD nous trouvons la fouille de données qui est une étape indispensable. Cette dernière utilise plusieurs techniques tel que : la segmentation, la classification, la régression et les réseaux de neurones..etc.

Récemment, l'accent est mis sur le traitement des données se présentant sous forme de séquences, dites données séquentielles ou motifs séquentiels, telles que les séquences biologiques (ADN), les séries temporaires, les flux d'images ou de vidéo…etc. Les données séquentielles ajoutent aux algorithmes classiques une notion de temporalité et d'ordre dans les ensembles de données à traiter.

Vient s'ajouter en outre, pour les flots de données en particulier, des défis majeurs. Le premier est l'impossibilité d'effectuer plus d'une passe sur les données étant donné que la validité d'un événement dans un flot de données est attachée à son estampille temporaire qui ne peut se reproduire. Le second problème a trait à la quantité de mémoire nécessaire pour le stockage des données du flot et à l'enregistrement des résultats intermédiaires. Enfin l'efficacité des algorithmes de fouille de ce type de données est devenue aussi une des questions primordiales dans ce domaine.

L'extraction des connaissances à partir des structures plus complexes comme la vidéo-surveillance ou bien Data Stream qui est sensible à condition de stockage et insuffisant des

[1]Knowledge Discovery in Data bases.

mémoires, ce point est l'axe de notre thème qui traite le problème d'extraction des épisodes réguliers, motivée par une application d'actualité : le suivi des personnes âgées.

Dans notre travail, nous allons explorer comment contrôler les personnes âges dans leurs vie quotidienne, et ça grâce à l'amélioration de la médecine et de la qualité de la vie, avec l'utilisation d'une technologie plus développée des capteurs, les maisons intelligentes et les systèmes vivants ambiants assistés ont étalé durant la dernière décennie.

Les capteurs et appareils disséminés dans la maison enregistrent les traces de l'activité dans le cadre de la maison. Cette activité reflète la santé, et la fouille de ces traces peut révéler de l'information sur l'état de santé de la personne suivie. Il y a un intérêt croissant pour les techniques d'analyse sans surveillance, tels que les flux d'événements et de partitionnement, fréquentes et périodiques (connu sous l'appellation découverte d'épisodes réguliers ou fréquentes). Les relations entre les épisodes sont aussi étudiées.

Le sujet de ce mémoire se situe dans le domaine de la fouille des séquences, plus particulièrement les séquences d'événements reçu comme un flux de données en temps réel. Ceci ajoute des défis aux problèmes classiques connu pour ce problème. Il s'agit des questions ayant trait à la gestion efficace de la mémoire et le temps d'exécution.

Notre mémoire est organisé comme suit : On commence par l'introduction générale, puis dans le premier chapitre on parle sur la définition du Data mining en basé sur le processus KDD et les techniques et les taches du Data mining .

Le deuxième chapitre concerne le fouille des données séquentielles, il est divisé en deux grandes parties : la première c'est la définition des motifs séquentielles et leur méthodes de recherches, et la deuxième c'est l'extraction des motifs séquentielles sur les flots de données.

Le troisième chapitre se focalise sur le problème d'extraction des séquences sur le flot des données et l'étude et l'implémentation d'une approche d'extraction des tops k (les premières k épisodes intéressantes) épisodes réguliers.

Ce mémoire s'achève par une conclusion générale et des perspectives.

Chapitre 1 : Data Mining

1.1. Introduction

Durant ces dernières années, le Datamining (DM) est au cœur de toutes les préoccupations du monde car on assiste à une forte augmentation tant dans la quantité que dans la qualité des informations mémorisées dans des grandes bases de données scientifiques, économiques, financières, administratives, médicaux etc.….

Mais face à la masse exponentielle d'informations disponibles à travers des sources multiples et grâce à l'apparition de supports nouveaux, et à cause des restrictions sociales ou légales qui peuvent empêcher les analystes de recueillir des données dans un emplacement simple, les ensembles de données sont souvent physiquement distribués.

Le stockage en lui-même ne pose pas de réelles difficultés du point de vue informatique, mais le besoin de les interpréter pour trouver de nouvelles relations entre les éléments stockés dans ces bases a suscité beaucoup d'intérêt.

Ainsi, la mise au point de nouvelles techniques d'exploitation est devenu un thème important pour un grand nombre de chercheurs. L'extraction des connaissances "Knowledge Discovery" et la fouille de données "Datamining" représentent un domaine émergeant essayant de répondre à ces objectifs.

Dans ce chapitre, nous allons examiner plusieurs sections, dans la première section, nous avons parlé à la définition du "Datamining", puis à ces facteurs d'émergence, en suite qu'est-ce que ci leur objectifs. Après on a expliquons en détaillées Le processus de KDD, et dans la deuxième section les principales techniques, enfin le domaine d'application.

1.2. Historique

Au début des années 60, les ordinateurs de plus en plus utilisés pour toute sorte de calcul qu'il n'était pas envisageable de les effectuer manuellement, le progrès scientifique en matière de calcul et de qualités des données suscité des nouvelles techniques de traitement. Cette attitude opportuniste face aux données coïncida avec la diffusion de l'analyse de données.

Les promoteurs comme " Jean-Paul Benzecri " ont dû subir dans les premier temps les critiques venant des statisticiens. Malgré ça, l'analyse des données s'est développée et son intérêt grandissait en même temps que la taille des bases de donnée vers la fin des années 1980, des chercheurs en base de données tel que " Rakesh Agrawal " ont commencé à travailler sur

l’exploitation du contenu des bases de données volumineuses et ils utilisèrent le terme "data base Mining", mais celle-ci étant déposée par une entreprise "data base Workstation", ce fut "data Mining" qui s’imposa jusqu’à l’arrivée de "Shapiro Piatetski " en mars 1989 qui proposa le terme " Knowledge Discovrery" à l’occasion d’un atelier sur la découverte des connaissances dans les bases de données. Les termes Data Mining et knowledge Discovery in Data Base (KDD) sont actuellement utilisés plus ou moins indifféremment.

La communauté de data Mining a initié sa première conférence en 1995 à la suite de nombreux workshops sur le KDD entre 1989 et 1994, et en 1998 la première revue du domaine s’est créé sous le nom de " data Mining and knowledge Discovery journal ".

1.3. Définition du Datamining

Plusieurs définitions ont été proposées [1].

Le Datamining serait :

1. Le terme de Datamining signifie littéralement "forage de données".

Comme dans tout forage, son but est de pouvoir extraire un élément "la connaissance".

Ces concepts s’appuient sur le constat qu’il existe au sein de chaque entreprise des informations cachées dans le gisement de données. Ils permettent, grâce à un certain nombre de techniques spécifiques, de faire apparaître des connaissances.

2. Le Datamining est une méthodologie qui automatise la synthèse de connaissances à partir de gros volumes de données. L’essor de cette technologie est le résultat d’un accroissement dramatique de l’information numérique qui, de par son abondance, est sous-exploitée sans outil et expertise adéquats.

Cette technologie repose sur une diversité de techniques (statistiques, théorie de l’information, génie logiciel, bases de données,...) qui requièrent des compétences variés et de haut niveau.

3. la découverte de nouvelles corrélations, tendances et modèles par le tamisage d’un grand nombre de données.

4. un processus d’aide à la décision ou les utilisateurs cherche des modèles d’interprétation dans les données.

5. l’extraction d’informations originales, auparavant inconnues, potentiellement utiles à partir des données.

6. un processus de mise à jour de nouvelles corrélations, tendances et de modèles significatifs par un passage au crible des bases de données volumineuses, et par l’utilisation de modèles d’identification technique aussi bien statistiques que mathématiques.

7. Le Datamining est le processus d'exploration et de modélisation des gisements de données permettant de découvrir des informations/indicateurs inconnus pour obtenir des avantages concurrentiels.

Généralement, on s'accorde à définir le Datamining comme un ensemble de procédures de Découverte de connaissances dans les bases de données (Knowledge Discovery in Data base - KDD) .

1.4. Facteurs d'émergence du Datamining

Voici quelques facteurs qui ont contribuée l'apparition de DM :[2]

1. Production massive des données (Explosion des données .fig 1.1)

- Masse importante de données (millions de milliards d'instances) : elle double tous les 20 mois.
- Données multidimensionnelles (milliers d'attributs).
- Inexploitables par les méthodes d'analyse classiques.
- Collecte de masses importantes de données (Gbytes/heure).
- BD très larges - Very Large Databases (VLDB).
- BD denses.
- Besoin de traitement en temps réel de ces données.

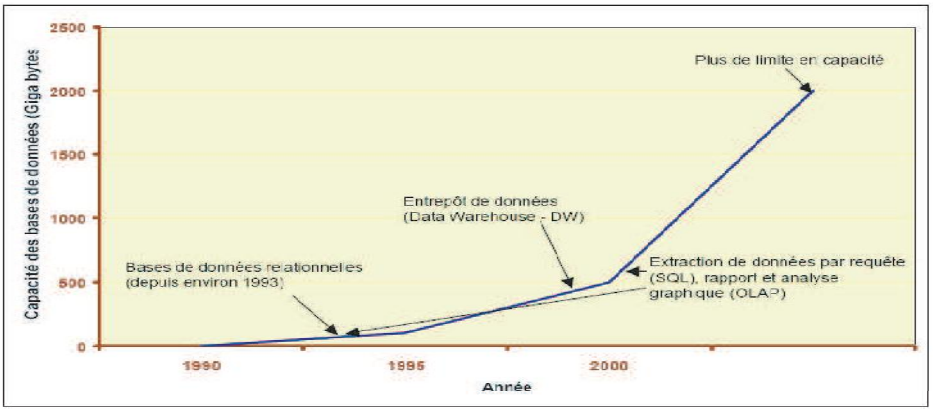


Fig 1.1: Evolution de la taille des bases de données.

2. Grandes capacité de stockage.
3. Puissants processeurs.
4. Améliorer la productivité Forte pression due à la concurrence du marché.

- Brièveté du cycle de vie des produits.
- Besoin de prendre des décisions stratégiques efficaces.

5. Contexte très concurrentiel.
6. Disponibilité de logiciels de DM.

7. Croissance en puissance/coût des machines capables

- de supporter de gros volumes de données.
- d'exécuter le processus intensif d'exploration.
- hétérogénéité des supports de stockage.

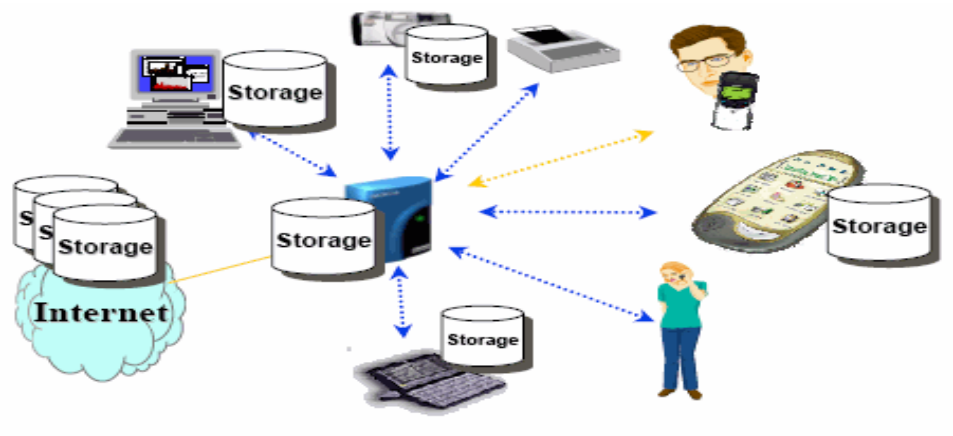


Fig 1.2 : Masse importante de données –supports hétérogènes.

1.5. Objectifs du Datamining

Le Datamining est un processus s'appuyant sur les données. Il consiste à extraire des informations des bases de données de l'entreprise (informations prévisibles mais non évidentes) et à les transformer en connaissances originales ou en modèles par des techniques à base d'algorithmes. Il exploite la base de données (qui peut concerner plusieurs millions d'acteurs) pour sélectionner les données les plus pertinentes, en faire l'analyse et en tirer des conclusions qui orienteront les décisions marketing .En autre, il permet une optimisation du ciblage, la découverte de nouveaux segments et la création de nouveaux indicateurs. Cette connaissance du client permet une grande créativité marketing [3].Mettre au point des programmes marketing produit, prix, distribution, publicité, promotion, plus adaptés à leurs besoins.

On peut dire que l'émergence du data mining (DM) devenu pour réaliser les trios objectif suivant :

- Le data mining pourra tenter d'expliquer un événement ou un incident indiscernable(Expliqation**)**. Par la consultation des informations contenues dans l'entrepôt de données de l'entreprise, on peut être en mesure de formuler la question suivante : pour qu'elle raison perd-t-on des clients pour tel produit spécifique dans telle

région? tout en se basant sur des données collectées ou des mises en signification de paramètres liés, le data mining va essayer de trouver un certain nombre d'explication à cette question. Le Data Mining va aider à trouver des hypothèses d'explications.

- Le data Mining aidera à confirmer un comportement ou une hypothèse(Confirmation**)**. Dans le cas où le décideur aurait un doute concernant une hypothèse, le data Mining pourra tenter de confirmer cette hypothèse en la vérifiant en appliquant des méthodes statistiques ou d'intelligence artificielle.
- enfin, le data mining peut explorer les données pour découvrir un lien "inconnu" jusque-là (exploration). Quand le décideur n'as pas d'hypothèse ou d'idée sur un fait précis, il peut demander au système de proposer des associations ou des corrélations qui pourront aboutir à une explication. Il est utopique de croire que le data mining pourrait remplacer la réflexion humaine. Le data mining ne doit être vu et utiliser uniquement en tant qu'aide à la prise de décision. Par contre, l'informatique décisionnelle dans son ensemble, et plus particulièrement le data mining permet de suggérer des hypothèses. La décision finale appartiendra toujours au décideur.

1.6. Découverte de la connaissance dans les bases de données (Knowledge Discovery in Data bases)

L’extraction de connaissance à partir de bases de données, traduction littérale de" Knowledge Discovery in data base ".

Le KDD adopte la définition suivante :

"Knowledge Discovery in data bases is the non-trivial process of identifying valid, novel, potentially useful and understandable patterns in data".

Dans cette définition, chaque mot a son importance et ses auteurs Fayyad, PIATETSKI-SHAPIRO et SMYTH, ont pris le soin de les décortiquer :

Le terme data fait référence aux données, c’est-á-dire aux tuples des bases de données et le mot patterns désigne une expression dans un langage décrivant un sous-ensemble ou un modèle applicable à ce sous-ensemble. Ainsi, dans notre contexte, l’extraction de "patterns" signifie ajuster un modèle aux données, trouver une structure dans les données ou d’une façon générale trouver une description élaborée de l’ensemble des données.

1.7. Le processus de KDD

Le processus d'Extraction de Connaissances à partir de Données (ECD ou KDD) est découpé en six parties : préparation, nettoyage, enrichissement, codage, fouille et validation. L'enchaînement des différentes étapes est présenté dans la fig 1.3. [4].

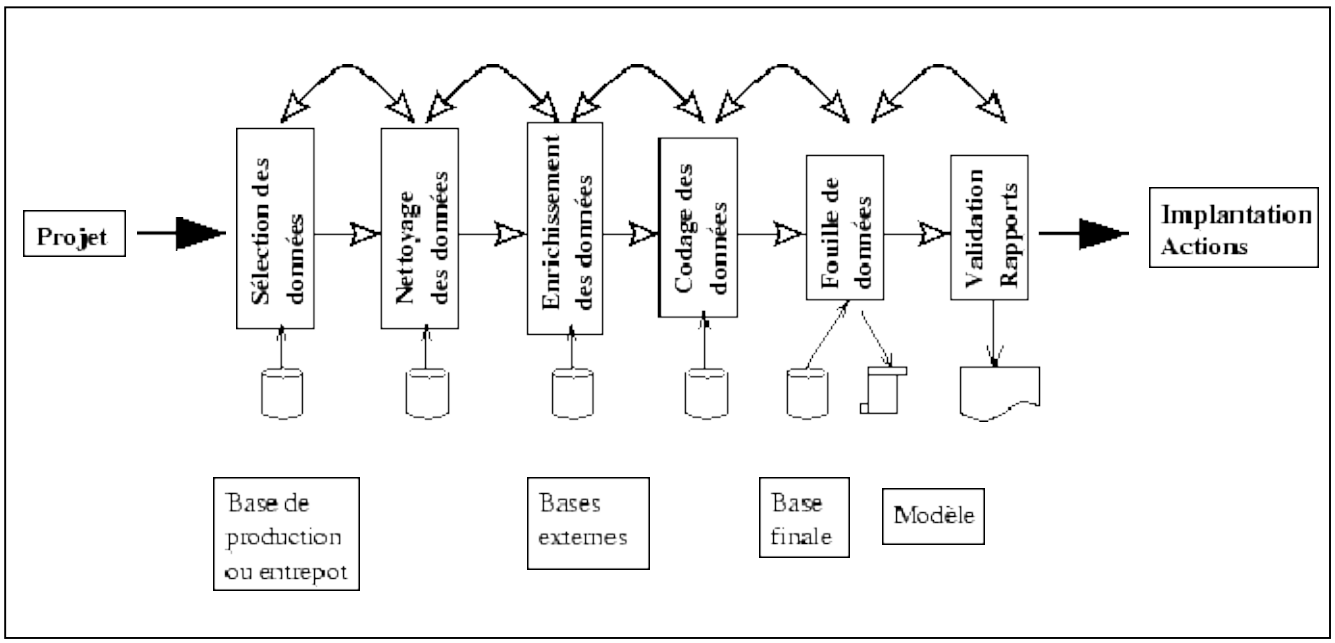


Fig. 1.3 : Les étapes du processus d'extraction de connaissances à partir de données

1.7.1 Préparation des données

Cette étape consiste dans un premier temps à obtenir des données en accord avec les objectifs que l'on s'impose. Ces données sont tout d'abord copiées sur une machine adéquate, pour des questions de performance. L'obtention des données est souvent réalisée à l'aide outils de requêtage (OLAP[2], SQL[3],……).

1.7.2 Nettoyage :

Corriger certain incohérence dans les donnes tel que :

-Doublons, erreurs de saisie :

Les doublons vont donner plus d'importance aux valeurs répétées. Une erreur de saisie pourra à l'inverse cacher une répétition.

-Intégrité de domaine :

Un contrôle sur les domaines des valeurs permet de retrouver des valeurs aberrantes.

-Informations manquantes :

[2]OnLine Analytical Processing.

[3]Structured Query Language.

C'est le terme utilisé pour désigner le cas où des champs ne contiennent aucune donnée. Mais parfois, l'absence d'information peut être une information.

1.7.3 Enrichissement :

On peut avoir recours à d'autres bases pour enrichir nos données. L'opération va se traduire par l'ajout de nouveaux champs en conservant souvent le même nombre d'enregistrements. Les difficultés ici sont de pouvoir relier les données hétérogènes et l'introduction de nouvelles valeurs donc la phase de nettoyage sera certainement de nouveau utile.

1.7.4 Codage, normalisation :

À ce stade du processus, les choix sont particulièrement guidés par l'algorithme de fouille utilisée et des ajustements des choix de codage sont souvent nécessaires.

- Regroupements :

Lorsqu'il est important de considérer les attributs pour la fouille de données, il est obligatoire d'opérer des regroupements et ainsi obtenir un nombre de valeurs raisonnable.

- Attributs discrets :

Les attributs discrets prennent leurs valeurs (souvent textuelles) dans un ensemble fini donné. Par exemple, le même codage en deux valeurs 0 et 1 sera réalisé avec les champs Oui/Non issus de l'enrichissement.

- Changements de type :

Pour certaines manipulations, il est préférable de modifier les types de certains attributs. Par exemple, la date de naissance et la date d'abonnement nous pouvons les convertir en âge ou en durée.

1.7.5 Fouille :

La fouille de données est le cœur du processus car elle permet d'extraire de l'information des données.

1.7.6 Validation :

Les méthodes de validation vont dépendre de la nature de la tâche et du problème considéré. Nous distinguerons deux modes de validation : statistique et par expertise. Pour certains domaines d'application (le diagnostic médical, par exemple), il est essentiel que le modèle produit soit compréhensible. Il y a donc une première validation du modèle produit par l'expert, celle-ci peut être complétée par une validation statistique sur des bases de cas existantes.

1.8. Les taches du Datamining

Contrairement aux idées reçues, le Datamining n'est pas le remède miracle capable de résoudre toutes les difficultés ou besoins de l'entreprise. Cependant, une multitude de problèmes d'ordre intellectuel, économique ou commercial peuvent être regroupé, dans leur formalisation, dans l'une des tâches suivantes :

1.8.1. Classification :

La classification se fait naturellement depuis déjà bien longtemps pour comprendre et communiquer notre vision du monde (par exemple les espèces animales, minérales ou végétales)."Elle consiste à examiner des caractéristiques d'un élément nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini".

Dans le cadre informatique, les éléments sont représentés par un enregistrement et le résultat de la classification viendra alimenter un champ supplémentaire.

La classification permet de créer des classes d'individus (terme à prendre dans son acception statistique). Celles-ci sont discrètes : homme / femme, oui / non, rouge / vert / bleu.

La démarche en générale est la suivante :

(a) On identifie la variable objectif, ou variable de classe (par exemple client/ non client), et d'autres variables qui pourraient permettre de la prévoir (par exemple des caractéristiques individuelles, ou le comportement sur le site).

(b) On regroupe un certain nombre de cas pour lesquels à la fois la variable objective et les variables potentiellement explicatives sont disponibles.

(c) Sur ces cas, on essaie de construire un modèle qui relie de façon satisfaisante les variables explicatives à la variable objective ;

(d) On vérifie que le modèle est suffisamment robuste (cette vérification peut s'effectuer soit par tests statistique adapté soit simplement en appliquant le modèle mis au point à un autre jeu de caser).

Les techniques les plus appropriées à la classification sont : Les arbres de décision, le raisonnement basé sur la mémoire, Eventuellement l'analyse des liens, Réseaux bayésiens, présente dans la section suivante.

1.8.2 Estimation :

Contrairement à la classification, le résultat d'une estimation permet d'obtenir une variable continue. Celle-ci est obtenue par une ou plusieurs fonctions combinant les données en entrée. Le résultat d'une estimation permet de procéder aux classifications grâce â un barème. Par exemple, on peut estimer le revenu d'un ménage selon divers critères (type de véhicule et

nombre, profession ou catégorie socioprofessionnelle, type d'habitation, etc.). Il sera ensuite possible de définir des tranches de revenus pour classifier les individus. Un des intérêts de l'estimation est de pouvoir ordonner les résultats pour ne retenir si on le désire que les n meilleures valeurs.

Cette technique est souvent utilisée en marketing, combinée aux d'autres, pour proposer des offres aux meilleurs clients potentiels. Enfin, il est facile de mesurer la position d'un élément dans sa classe si celui-ci a été estimé ce qui peut être particulièrement important pour les cas limitrophes.

L'estimation comporte les étapes suivantes :

(a) On identifier la variable objective, et d'autres variables qui pourraient permettre de la prévoir.

(b) On regroupe un certain nombre de cas pour lesquels à la fois la variable objective et les variables potentiellement explicatives sont disponibles.

(c) Sur ces cas, on essaie de construire un modèle qui relie de façon satisfassent les variables explicatives à la variable objective.

(d) On vérifie que le modèle est suffisamment robuste.

La technique la plus approprié à l'estimation est :

– les réseaux de neurones.

1.8.3 Prédiction :

La prédiction ressemble la classification et à l'estimation mais dans une échelle temporelle différente. Tout comme les tâches précédentes, elle s'appuie sur le passé et le présent mais son résultat se situe dans un futur généralement précisé La seule méthode pour mesurer la qualité de la prédiction est d'attendre !

Les techniques les plus appropriées à la prédiction sont :

(a) Les règles d'associations.

(b) Les arbres de décision.

(c) les réseaux de neurones.

1.8.4 L'analyse d'association :

Le regroupement par similitudes consiste à grouper les éléments qui vont naturellement ensembles. La technique la plus appropriée au regroupement par similitudes est l'analyse du panier de la ménagère.

1.8.5 Description : C'est souvent l'une des premières tâches demandées à un outil de Data Mining. On lui demande de décrire les données d'une base complexe. Cela engendre souvent

une exploitation supplémentaire en vue de fournir des explications. La technique la plus appropriée à la description est l'analyse du panier de la ménagère.

1.8.6 Optimisation :

Pour résoudre de nombreux problèmes, il est courant pour chaque solution potentielle d'y associer une fonction d'évaluation. Le but de l'optimisation est de maximiser ou minimiser cette fonction. Quelques spécialistes considèrent que ce type de problème ne relève pas du DataMining. La technique la plus appropriée à l'optimisation est les réseaux de neurones.

1.9. Les Techniques de Datamining

Les techniques de Datamining se distinguent selon les données à traiter et selon le caractère opératoire. Dans le cas de données textuelles, on parle de Texte Mining. Dans le cas de données numériques quantitatives ou qualitatives, on parlera de Datamining.

Le caractère opératoire d'une technique désigne son objectif ; deux objectifs sont retenus pour le Data Mining : d'une part, un objectif de description ; d'autre part, un objectif de modélisation/prédiction.

Cette simplification permet de positionner les techniques de Data Mining les unes par rapport aux autres et même leur façon d'utilisation. Elles se répartissent selon trois grandes familles :[5]

- *Les techniques statistiques multidimensionnelles* : (Factorisation, segmentation, classification, régression)
- *Les techniques de visualisation des données* : (Règles d'association, réseaux bayésiens).
- *Les techniques s'appuyant sur l'intelligence artificielle* : (réseaux de neurones).

Ces techniques sont utilisées dans deux contextes différents :

- La modélisation descriptive, on retrouve les techniques de classification (centre mobile, classification ascendante hiérarchique), factorielles (ACP) et d'association (Règle d'association, réseaux bayésiens).
- La modélisation prédictive, on retrouve les techniques de segmentation (d'arbre de décision et de réseaux de neurones) et d'estimation (régression, réseaux de neurones).

1.9.1. Les techniques statistiques multidimensionnelles :

Elles permettent, de réduire l'espace, de fournir les représentations graphiques, d'exploiter, de fouiller, de représenter de grands ensembles de données.

1.9.1.1. *Techniques de régression* :

Ce sont les techniques les plus anciennes et les plus connues.

Il y a :

La régression linéaire simple : modélisation d'une variable quantitative par une autre variable quantitative.

La régression linéaire multiple : modélisation d'une variable quantitative par une somme de variables quantitatives.

La régression non linéaire polynomiale : modélisation d'une variable quantitative par un polynôme de variables.

La régression logistique : modélisation d'une variable qualitative binaire par un polynôme de variables quantitatives ou qualitatives.

La régression par voisinage : modélisation d'une variable quantitative en ne prenant que les observations les plus proches.

1.9.1.2. *Techniques factorielles* :

Les techniques factorielles procèdent de la même manière que les techniques de régression. Elles s'en différencient par le fait que ses variables du modèle élaboré sont les axes factoriels, combinaison linéaire des variables d'origine, non corrélés deux à deux au sens de la métrique utilisée pour la construction de l'analyse, et les variables du modèle sont rangées par ordre d'importance liées à la variance des axes factoriels ; ce qui permet de modéliser la situation avec un petit nombre de variables (axes factoriels) significatif en terme de variance expliquée.

La prédiction est effectuée soit directement par affectation d'un élément supplémentaire dans l'analyse factorielle, soit au moyen d'une régression sur facteurs. Il y a plusieurs types d'analyse factorielle :

-L'analyse en composantes principales.

-L'analyse discriminante.

-L'analyse des correspondances binaires.

-L'analyse des correspondances multiples.

-L'analyse factorielle canonique.

1.9.1.3. *La segmentation* :

L'analyse des clusters consiste à segmenter une population hétérogène en sous populations homogènes. Contrairement à la classification, les sous populations ne sont pas préétablis. La technique la plus appropriée à la clusterisation est l'analyse des clusters.

1.9.1.4. Construction d'arbre de décision :

Il s'agit de diviser la population successivement en sous-groupes selon les valeurs prises par les variables, qui à chaque stade, on discrimine mieux la variable modélisé. La variable à expliquer est par définition le sommet de l'arbre.

L'objectif est de prédire la probabilité d'appartenance de chaque individu à l'une ou l'autre des catégories de la variable à expliquer.

Il existe plusieurs méthodes de construction des arbres, la plus ancienne est la méthode arbre *C5*, la plus récente et la plus robuste est la méthode *CART*. L'arbre de décision est particulièrement efficace sur de gros volumes de données qualitatives et il est facilement interprétable par des non statisticiens mais il se révèle en revanche moins efficace que les régressions sur les données quantitatives.

– Avantages :

1. Utilisation naturelle dans les domaines basés sur des règles.
2. Calcul simplifié rapide et peu coûteux sur la classification.
3. Adapté à la manipulation des variables continues ou énumératives.
4. Indication claire des champs.

– Inconvénients :

1. Peu performants si beaucoup de classes.
2. Apprentissage des arbres coûteux en calculs.
3. Problèmes avec certaines boites de classification rectangulaire.

1.9.1.5. *La classification* :

Ces techniques sont destinées à produire des groupements de lignes ou de colonnes dans un tableau de manière à apparaître des classes d'individus. Les présentations se font sous forme de partitions simples des ensembles étudiés (méthode d'agrégation par les centres mobiles ou K-means) ou de partitions hiérarchisées (algorithmes ascendants ou descendants) qui ont un aspect d'arbre (au sens de la théorie des graphes).

1. 9.2 Les techniques de visualisation des données :

1.9.2.1. *L'association* :

Les associations les plus simples sont calculées en fonction du nombre d'occurrences simultanées de couples de modalités.

1.9.2.2. *Règle d'association* :

Les associations se feront ici sur des n-tuples de modalités, c'est à dire les occurrences simultanées de plusieurs modalités, la force de l'association proprement dite est mesurée par la probabilité conditionnelle de retrouver une modalité y en co-occurrence avec un ensemble de *n* autres a, b, c, d qui le sont déjà cette probabilité s'appelle la confiance. La règle correspondante est " si a, b, c, d sont en co-occurrence alors y l'est aussi avec eux " et s'exprime comme y<=a, b, c, d. Une règle est considéré comme bonne et est sélectionnée si son seuil de confiance est élevé et si elle est présente souvent, c'est à dire si elle a un support élevé parmi les différents algorithmes, celui qui est le plus souvent utilisé est *APRIORI* et a développé par IBM. Cette technique est utilisée dans le cas de personnalisation d'une offre ou de vente croisé.

1.9.2.3. *Les réseaux Bayesiens* :

Le réseau Bayesiens est un modèle graphique probabiliste de connaissance. C'est une distribution de probabilité (appelée quantification de la connaissance) factorisée suivant un graphe (appelé structure de la connaissance). Ce graphe représente en un certain sens les relations entre variables en terme causal, relations déterminées par des lois de probabilités conditionnelles, elles-mêmes détruites des fréquences conditionnelles associées à un tableau de données.

1.9.3 Modèle de dépendance basée sur l'intelligence artificielle :

1.9.3.1.*Réseaux de neurones* :

Les réseaux de neurones sont utilisés pour prédire les valeurs prises par une ou plusieurs variables. Fondée au départ sur des analogies biologiques, le réseau de neurones est utile pour mettre en évidence des relations non linéaires entre les variables. Il sert aussi à faire une bonne approximation des relations si l'on sait qu'il existe des relations non linaires mais non identifiées entre plus de trois variables.

1.10. Types de données

En principe, le data mining peut s'appliquer à tous les types de données. Toutefois, selon chaque type de données, les algorithmes de data mining diffèrent. Quelques exemples de types de données auxquels peut s'appliquer la fouille de données sont : [6]

_LesFichiers plat"flat file" : Ce sont des fichiers en format texte ou binaire, contenant un enregistrement par ligne, avec des champs séparés par des délimiteurs, tels que les virgules ou les tabulations. Dans ce type de fichiers, les données peuvent être des transactions, des séries temporelles, des mesures scientifiques .....etc.

_ *Base de données relationnelle* : Une base de données relationnelle est une base de données consistant dans des tableaux séparés, avec des liaisons explicitement définies et dont les éléments peuvent être combinés sélectivement comme des résultats à des interrogations. Chaque tableau contient des colonnes (correspondantes à des tuples) et des lignes (correspondantes à des attributs) ;

_ *Entrepôt de données* : Ce terme désigne une base de données utilisée pour collecter et stocker des informations provenant de multiples bases de données (souvent hétérogènes) et qui les traite comme un tout-unitaire (comme une seule base de données) ;

_*Base de données transactionnelle* : Une base de données transactionnelle est un ensemble d'enregistrements représentant des transactions, chaque enregistrement ayant une marque temporelle, un identificateur et un ensemble d'articles ;

_ B*ase de données multimédia* : Ce type de base de données inclut des vidéos, des images, des audio et des textes média ;

_ B*ase de données temporelles* : Elles contiennent des données organisées dans le temps, comme par exemple des activités log ;

_ B*ases de données orientées objet et relationnelle objet* : Il s'agit d'un type spécial de base de données (ou base de données relationnelle) où les données sont des objets ;

_ Bases de données spatiales : Une telle base de données est optimisée afin de garder des données spatiales et de pouvoir en être interroger ;

_ W*orld wide web* : Le WWW est le dépôt le plus hétérogène et dynamique possible.

1.11. Les domaines d'application du Datamining

Les applications du datamining sont multiples, on cite quelques domaines d'application :

Activités commerciales : grande distribution, vente par correspondance, banque, assurances.

- Segmentation de la clientèle.
- Détermination du profil du consommateur.
- Analyse du panier de la ménagère.
- Mise au point de stratégies de rétention de la clientèle.
- Prédiction des ventes.
- Détection de fraudes.
- Identification de clients à risques .

Activités financières

- Recherche de corrélations entre les indicateurs financiers.
- Maximiser le retour sur investissement de portefeuilles d'action.

Activités de gestion des ressources humaines

- Prévision du plan de carrière.
- Aide au recrutement.

Activités industrielles

- Détection et diagnostic de pannes et de défauts.
- Analyse des flux dans les réseaux de distribution.

Activités scientifiques

- Diagnostic médical.
- Santé publique.
- Etude du génome.
- Analyse chimique et pharmaceutique.
- Exploitation de données astronomique.

1.12. Les logiciels de Datamining

Plusieurs logiciels commerciales et expérimentales de datamining ou bien dit certains sont gratuits et d'autre sont payants sont développées jusqu'à aujourd'hui. Il existe des plateformes fonctionnent monoposte (exécution et expérimentation locale) et d'autre de type d'accès à distance base sur l'architecture client-serveur.[7]

Nous avons citons quelque exemple de telle logiciels qui utilise souvent dans DM :

- WEKA : c'est une bibliothèque gratuite implémentée en Java qui comporte des algorithmes d'apprentissage et une plateforme de programmation visuelle. Elle donne à l'utilisateur la possibilité d'ajouter facilement ses propres algorithmes.
- TANGRA: c'est une plateforme gratuite d'expérimentation pour la fouille de données .Implémentée en Delphi, elle contient des méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique, mais ne comporte pas beaucoup de méthode de visualisation et de techniques interactives. Elle est désignée aux étudiants pour l'enseignement et aux chercheurs.
- ORANGE :c'est une plateforme d'expérimentation pour la DM, elle est implémentée en *C++/Qt/Python* et englobe des méthodes d'apprentissage automatique et de visualisation.
- Oracle DataMining (ODM) :une option d'*Oracle Databse 11g Enterprise* qui permet de produire des informations prédictives exploitables et d'élaborer des applications de business intelligence intégrées. Avec la fonction d'extraction des données intégrée à

Oracle Database 11g ,les clients peuvent identifier des méthodes et des informations cachées dans leurs données, les développeurs d'application peuvent rapidement automatiser la découverte et la distribution de nouvelles prévisions , de méthodes et de découvertes dans l'ensemble de l'entreprise.

1.13. Les barrières majeures aux Datamining

Même si le Datamining peut apporter beaucoup de valeur ajoutée à un système d'information, il reste encore plusieurs difficultés auxquelles les différentes offres essaient de répondre.

Le coût est un facteur bloquant majeur pour les entreprises qui veulent s'équiper d'un SIAD. Une solution monoposte coûte environ 20000 dollars, ce qui empêche toute implémentation d'envergure. Même si certains éditeurs proposent des solutions à moins de 5000 dollars, le prix reste une barrière psychologique à franchir [8].

La quantité monstrueuse de données qu'il faut stocker sur des serveurs back office géants. Il est nécessaire d'avoir un nombre suffisant de données pour que les résultats obtenus soient statistiquement viables. Il est aussi évident que les utilisateurs doivent anticiper les conclusions, donc délimiter clairement le domaine de données sur lequel il souhaite mener le Datamining.

La plupart des logiciels de Datamining ne restent compréhensibles que par un petit nombre d'utilisateurs avertis. D'ailleurs, beaucoup d'outils exécutent le plus gros du travail dans une boîte noire. Heureusement, les arbres de décision permettent de palier à ce problème en proposant un éclatement du processus en plusieurs sous branches.

La préparation des données prend largement 80% du temps du processus complet. Les outils d'aujourd'hui permettent à l'utilisateur de mieux sélectionner les données à traiter, en proposant de vues externes qui constituent une couche sémantique supplémentaire plus compréhensible. Ce processus d'épuration et de sélection de données est appelé "sampling".

La décision d'adopter un SIAD est risquée pour une entreprise: en effet, il lui est très difficile d'estimer le retour sur investissement de la chose, étant donné que l'on ne connaît pas a priori les résultats que l'on obtiendra éventuellement. Il devient de plus en plus nécessaire de proposer des solutions incrémentielles, qui s'adaptent progressivement aux besoins grandissants des entreprises.

1.14. Conclusion

Dans le contexte d'une entreprise intégrée numériquement, des quantités très importantes de données sont générées chaque jour. Ces données permettent de décrire tout aussi bien les produits et les processus de l'entreprise. Bien que le volume de données produit augmente sans cesse à un rythme qui s'accélère, ces données sont le plus souvent archivées dans des entrepôts de données sans être aux préalables exploitées. Nous supposons que ces données contiennent de l'information cachée qu'il peut être pertinent d'extraire et d'utiliser pour la prise de décision et l'amélioration des connaissances de l'entreprise sur elle-même, sur ses produits et sur ses clients.

Le datamining est une activité en pleine expansion, qui préfigure ce que seront le système d'information de demain. On abandonne peu à peu les bases de données infocentres qui n'étaient que des délocalisations de données brutes et non exploitables pour s'intéresser à des bases plus intelligentes, s'adaptant au besoin à son environnement et rendant transparents les accès à l'information.

Le mot d'ordre est trop de données trop d'information …moins de données plus d'information. Les entreprises peuvent maintenant s'équiper de systèmes d'aide à la décision intègres à leur métier et à leur activité économique.

Chapitre 2 : Fouille de données séquentielles

2.1. Introduction

La notion de recherche des motifs séquentiels a été introduite pour la première fois par *Rakesh Agrawal en 1995*.Elle a été intégré dans plusieurs domaines tels que : la biologie, la fouille d'usage du Web, la détection d'anomalie, la fouille de flux de données, l'extraction de motifs spatio-temporels...etc.

Dans ce chapitre nous présenterons la notion de fouille de donnes séquentielles, puis nous discuterons les algorithmes d'extraction de motifs séquentiels, en suite nous expliquerons l'extraction de motifs séquentiels sur les flots de données (Data Stream).

2.2. Recherche des motifs séquentiels

2.2.1. Définition et problématique

Suite au développement de nouvelles applications, une extension de la fouille de données est apparue : la fouille de motifs séquentiels qui consiste en la découverte des données se présentant sous forme de séquences.

Le concept de séquence a été défini premièrement dans [09, 21]. Les définitions suivantes sont tirées de[23].

Définition1 : (Itemset). Un itemset est un ensemble non vide d'items noté ($i_1 i_2 \ldots i_k$).

Définition2 : (Séquence). Une séquence est une liste ordonnée, non vide, d'itemsets notée<(it_1) . . . (it_n)> où (it_j) est un itemset. T(I) représente l'ensemble de toutes les séquences possibles à partir des items présents dans l'ensemble I.

ID Transactions	Itemset
T_1	{Cola,Pain}
T_2	{Chips,Pain}
T_3	{Pain,Yaourt}
T_4	{Chpis, Pain,Chocolat}
T_5	{Chocolat, Cola,Chips,Yaourt,Pain}

Tab. 2.1: Exemple de base de données contenant 5 transactions.

Définition 3 : (Base de données transactionnelles) . Soit I un ensemble fini d'items. Une base de données transactionnelles *D* basée sur les items de *I* est un ensemble fini de paires (*SID*, *T*), appelées transactions, avec *SID* ∈ {1, 2, . . .} un identifiant et $T \in T(I)$ une séquence construite sur *I*. Pour deux transactions quelconques $(SID_1 , T_1) \neq (SID_2 , T_2) \in D$ implique que $SID_1 \neq SID_2$.

Définition4 (Inclusion).Une séquence $S' = <(it'_1)\ldots. (it'_n)>$ est une sous-séquence de$S =<(it_1)\ldots.(it_m)>$, notée $S' \leq S$, si $\exists\ i_1<i_2< \ldots i_j \ldots <i_n$ tels que $it'_1 \subseteq it_{i1}$, $it'_2 \subseteq it_{i2}$,. . ….. $it'_n \subseteq it_{in}$. Si $S \leq S'$ et $S' \leq S$, les séquences sont dites incomparables et sont notées $S <> S'$.
De plus, une séquence est dite régulière si chaque itemset it_j contient le même unique item *i*.
Exemple 2 La séquence S´={(a)(b, c)(d)} est incluse dans la séquence S ={(a, d, e)(g, h)(f)(b, c, e)(d, e, f)}(i.e. $S' \leq S$) car (a) ⊆ (a, d, e), (b, c) ⊆ (b, c, e) et (d) ⊆ (d, e, f). En revanche,{(a)(b)}⊈{(a, b)}(et vice versa). Les deux séquences {(a)(b)} et {(a, b)} sont dites incomparables.

Définition5 (Support). Le support d'une séquence *S* dans une base de données transactionnelles *D*, noté Support(*S*, *D*) ou Support(*S*) quand le contexte est clair, est défini tel que :
Support(*S*, *D*) = |{(SID, *T*) ∈*D*|*S* ≤ *T*}|

Définition 6 (Fréquence). La fréquence de *S* dans *D*, noté f_S , est par :

$$f_S = \frac{\text{Support}(S,D)}{|D|}$$

Le problème d'extraction des motifs séquentiels peut être défini formellement comme suit.
Définition 7 (Extraction des motifs séquentiels fréquents). Soit *D* une base de données transactionnelles. Étant donné un seuil de support minimal (ou un seuil de fréquence minimale) σ, le problème d'extraction de motifs séquentiels est l'énumération de toutes les séquences *S* dans *D* telles que $f_S \geq \sigma$. L'ensemble des motifs séquentiels pour la valeur σ dans la base de données *D* est noté $F_{Seqs}(D, \sigma)$.
$F_{Seqs}(D, \sigma) := \{S \mid \text{Support}(S, D) \geq \sigma\}$
Avec $0 < \sigma \leq 1$ dans le cas où σ est un seuil de fréquence et $0 < \sigma \leq |D|$ dans le cas où σ est un

seuil de support.[23]

La propriété Apriori étendue aux séquences est la suivante. Elle permet d'élaguer l'espace de recherche.

Propriété 1[11].

Soit S' et S deux séquences. Si $S' \leq S$ alors Support(S') ≥Support(S) (ou $f_{S'} \geq f_S$).

Le dual de la propriété précédente est donné ci-après.

Propriété 2.

Soit S' une séquence non fréquente. Quelle que soit S telle que $S' \leq S$, S est une séquence non fréquente.

Exemple 3 Considérons la base de données D représentée dans la Tab 2.2.

Id.Sequence	Sequences
1	(a,b)(a,b)(b,d,e)
2	(a,b,c,d,e)(b ,e)
3	(b,c,e)
4	(a,c)(b,c,e)
5	(c) (c) (d)(b,c,e)

Tab 2.2: La base de données transactionnelle exemple 3

Avec un seuil de fréquence minimum σ de $\frac{3}{5}$ soit 60% (i.e. pour qu'une séquence s soit retenue, il faut qu'au moins trois séquences dans la base de données la supportent), les séquences fréquentes sont alors les suivantes :

$$FSeqs(\mathcal{D}, \sigma) = \left\{ \begin{array}{lll} \langle a \rangle : 3 & \langle (a)(b) \rangle : 3 & \langle (c)(b,e) \rangle : 3 \\ \langle b \rangle : 5 & \langle (a)(e) \rangle : 3 & \langle (a)(b,e) \rangle : 3 \\ \langle c \rangle : 4 & \langle (b,c) \rangle : 4 & \\ \langle d \rangle : 3 & \langle (b,e) \rangle : 5 & \\ \langle e \rangle : 5 & \langle (c)(b) \rangle : 3 & \\ & \langle (c,e) \rangle : 4 & \\ & \langle (c)(e) \rangle : 3 & \end{array} \right\}$$

2.2.2. Les méthodes d'extraction des motifs séquentiels [23]

Les méthodes existantes différentes essentiellement sur la manière de parcourir l'espace de recherche (largeur d'abord ou profondeur d'abord) et sur les structures de données utilisées pour indexer la base de données et faciliter une énumération rapide. Les algorithmes d'extraction de motifs séquentiels peuvent être classés en trois grandes catégories :

1. Méthodes horizontales
2. Méthodes verticales
3. Méthodes par projection

2.2.2.1. Méthodes horizontales : (par niveau)

Parmi les premiers algorithmes, on trouve l'algorithme GSP (Generalized Sequential Patterns) qui donne un processus traitant l'ensemble des séquences par niveau.[11]

L'algorithme GSP :

Cet algorithme est fondé sur le principe de son antécédent pour les itemsets fréquents Apriori. Il utilise la technique de recherche en BFS pour parcourir l'ensemble de la base une fois, il suit toujours le principe de générer-tester, c'est-à-dire de la création de candidats, puis du test de ces candidats pour vérifier leur support.

La génération de candidats se fait dans le cadre général par auto-jointure sur le dernier ensemble extrait de motifs fréquents.

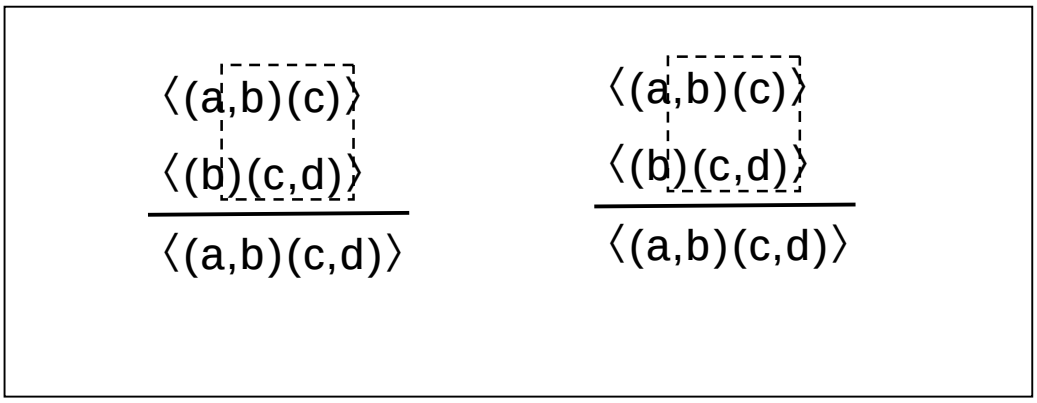


Fig2.1.Exemple de jointure entre candidats dans GSP

Ainsi, pour générer les séquences candidates de taille k, l'algorithme GSP fait une opération de jointure sur l'ensemble de motifs fréquents de taille K-1. Nous illustrons cette opération de génération dans la fig 2.1 en étendant la séquence ‹(a, b)(c). Il possède la propriété anti-monotone où toutes les séquences d'une séquence fréquente doivent être aussi fréquentes.

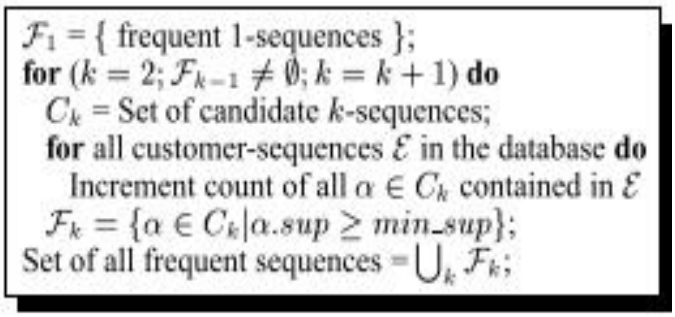
GSP fonctionne en plusieurs passages, il effectue plusieurs passes sur l'ensemble de données. Dans la première passe, un ensemble de candidats 1-séquences sont identifiées. A partir des éléments fréquents identifiés, un ensemble de candidats 2 séquences est généré et un autre passage est fait pour calculer leur fréquence d'apparition. Les fréquentes 2 séquences sont ensuite utilisées pour générer les candidats 3-séquences et ce processus est répété jusqu'à ce qu'aucune des séquences plus fréquentes n'est trouvée. Alors L'algorithme peut être résumé à deux étapes importantes :

1. Initialement, chaque item dans la base est un candidat de longueur 1.
2. Pour chaque niveau (séquences de longueur k) faire :

– Scanner la base pour déterminer le support de chaque séquence dans la base

– Garder les séquences fréquentes

– Générer les séquences candidates de longueur k+1 à partir des séquences fréquentes de longueur k générées, en utilisant Apriori. Répéter jusqu'à ce qu'aucune séquence candidate ne peut être trouvée.



-Algorithme GSP-

Pour le calcul de support de chaque candidat en fonction d'une séquence de données, cet algorithme exploite une structure d'arbre de hachage destinée à organiser les candidats. Les candidats sont stockés en fonction de leur préfixe.

Pour ajouter un candidat dans l'arbre des séquences candidate, GSP parcourt ce candidat et effectue la descente correspondante dans l'arbre. Pour trouver quelles séquences candidates sont incluses dans une séquence de données, GSP parcourt l'arbre en appliquant une fonction de hachage sur chaque item de la séquence de données. Quand une feuille est atteinte, elle contient des candidats potentiels pour la séquence de données.

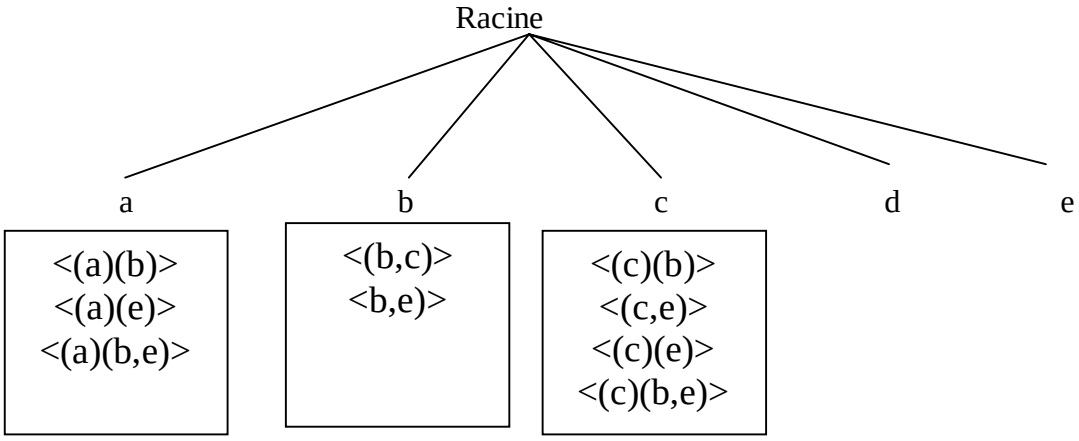


Fig. 2.2.La structure d'arbre de hachage utilisée par l'algorithme GSP

2.2.2.2. Méthodes verticales

Pour une amélioration de l'extraction des motifs séquentiels, d'autres algorithmes ont été conçu et qui ont pour objectif la résolution du problème de gestion de la mémoire. Dans cette orientation l'algorithme SPADE (Sequential PAttern Discovery using Equivalent Class) a été conçu [12].

Algorithme SPADE :

Cette approche a pour objectif de réduire l'espace de recherche par regroupant des candidats en catégorie. Elle utilise une représentation verticale qui consiste à inverser la méthode d'indexation de la base de données pour extraire de manière plus efficace des motifs séquentiels.

L'algorithme SPADE introduit une nouvelle représentation des motifs séquentiels en les regroupant selon des classes d'équivalences construites sur une relation d'équivalence ayant des préfixes commun qui permettant de décomposer le problème en sous-problèmes qui seront traités séparément en mémoire.

Il gère les candidats et les séquences fréquentes à l'aide de classes d'équivalence comme suit :

Deux k-séquences appartiennent à la même classe si elles présentent un préfixe commun de taille $(k - 1)$.

Plus formellement, soit $P_{k-1}(s)$ la séquence de taille k-1 qui préfixe la séquence s. Comme s est fréquente, $P_{k-1}(s) \in F_{k-1}$, avec F_{k-1} les fréquents de taille $k - 1$. Une classe d'équivalence est définie de la manière suivante :

$[\rho \in F_{k-1}] = \{s \in F_k \mid P_{k-1}(s) = \rho\}$

Le principal avantage de SPADE par rapport à GSP est l'utilisation de listes verticales temporaires pour les jointures. Ces listes, bornées, sont beaucoup plus petites et rapides à générer que celles utilisées par GSP ce qui permet d'améliorer grandement le temps de calcul

nécessaire ainsi que de réduire les besoins de mémoire. En plus, l'utilisation d'ensembles de catégorie permet de restreindre le domaine de recherche autour des éléments fréquents ce qui conduit à un important gain de temps.

ENTRÉES: minSupp,D /* Support minimal et base de données */
F1 ={ éléments fréquents ou 1-séquences fréquentes }
F2 ={2-séquences fréquentes}
ε ={classes équivalentes $[X]_{\theta 1}$}}
pour∀[X] ∈ε**faire**
EnumérateurFréquentSeq([X])
fin pour

-Algorithme SPDE-.

***Exemple* :** La fig 2.3 représente la base de données exemple dans le format vertical pour les séquences ‹a›, ‹b› et ‹(a)(b)› . En effet, la transformation consiste à associer à chaque k-séquence l'ensemble des couples (Id Séquence, transaction) qui lui correspondent dans la base. Le support d'une séquence est le cardinal de l'ensemble constitué par les identifiants de séquences. Dans le cas de la séquence ‹a›, l'ensemble des identifiants est : {1, 2, 4} de cardinal 3, le support de cette séquence est donc 3.

Séquence 〈a〉	
Id.Séq.	Transactions
1	1
1	2
2	1
4	1

Séquence〈b〉	
Id.Séq.	Transactions
1	1
1	2
1	3
2	1
2	2
3	1
4	2
5	4

Séquence〈(a)(b)〉	
Id.Séq.	Transactions
1	1
1	2
2	1
2	2
4	1

Fig.2.3: La base de données au format vertical pour les séquences 〈 a〉 , 〈 b〉 et 〈 (a)(b)〉

2.2.2.3. Méthodes par projection

Avec les contraintes des temps de réponse, des algorithmes comme FreeSpan (Frequent pattern projected Sequential pattern mining) est apparu. Il utilise des méthodes de recherche dites “en profondeur d’abord” (DFS) pour extraire les motifs séquentiels. Ilexploite un arbre de préfixes pour gérer les candidats. FreeSpan amélioré par d’autres études sur la projection de bases de données en recherche de motifs séquentiels, ce principe est utilisé par l’algorithme PrefixSpan.[13]

L’algorithme PrefixSpan

L’approche PrefixSpan diffère de GSP et SPADE par la méthode de génération des candidats. Elle possède les caractéristiques suivantes :

- Basée sur la projection de base de donne en recherche de motifs séquentiels.
- Les motifs sont générés de façon séquentielle dans les bases de données projetées par examen des segments localement fréquentes.

L’objectif de l’algorithme PrefixSpan est de réduire le nombre de candidats générés par l’analyse des préfixes communs que présentent les séquences de données de la base à traiter. A partir de cette analyse, l’algorithme construit des bases de données intermédiaires qui sont des projections de la base d’origine déduites à partir des préfixes identifiés , ensuite dans chaque base obtenue, PrefixSpan applique un comptage du support des différents items afin de faire croître la taille des motifs séquentiels découverts.

Le processus d’extraction de motifs dans l’approche PrefixSpanse déroulé en différents étapes de la manière suivante :

Premièrement PrefixSpan recherche et compte le support des séquences de taille 1, en plus il divise la base en plusieurs partitions selon le nombre d’items fréquents. Chaque partition est en fait une base de données qui prend un des items fréquents comme préfixe. En procédant à un comptage de tous les items fréquents sur les bases de données projetées, PrefixSpan extrait ainsi toutes les séquences fréquentes de taille 2. Ce processus est répété récursivement jusqu’à ce que les bases de données projetées soient vides, ou qu’il n’y ait plus de séquences fréquentes possibles.

PrefixSpan a surclassé les autres méthodes principalement de trois façons:

- Il produit les modèles sans génération de candidats.
- La réduction des données peut être effectuée de manière efficace par les projections.
- L'utilisation de l'espace mémoire est presque stable.

2.3Les flots des données

Grace à de nombreuses applications qui sont apparues, une grande quantité de donnes doit être générée dans un temps raisonnable avec une capacité de stockage limite, est né le concept de flot de donnes.

2.3.1. Définition d'un flot des données

Un flot de données est une séquence de données structurées que l'on peut considérer comme infinie d'éléments générés de façon continue à un rythme rapide et parfois variable.

2.3.2. Modèles des flots des données

Il est reconnu qu'il existe trois modèles de flots de données [14].Le choix du modèle est fonction des besoins de l'utilisateur et de l'application.

Notons un flot de donnée F, les élément arrivant sur ce flot sont notés a_1, a_2, . . .(où a_1 est l'élément le plus ancien du flot) et décrivent un signal sous-jacent à p variables A: [0, . . . , p−1] à valeurs dans R. Les trois modèles suivants diffèrent uniquement sur la manière dont les éléments ai décrivent le signal A:[23]

– *Le modèle des séries temporelles* : dans ce modèle chaque élément $a_i = A[i]$, il existe donc une adéquation complète entre le flot de données et le signal qu'il décrit.

– *Le modèle de la caisse enregistreuse* : dans ce modèle chaque élément a_i vient s'ajouter (ou s'incrémenter) à $A[j]$. En fait, on peut définir chaque élément $a_i = (j, I_i)$ avec I_i positif et $A_i[j] = A_{i-1}[j] + I_i$. Sémantiquement parlant, ce modèle tient compte de l'augmentation du signal pour une valeur $j \in [0, . . . , p-1]$ à chaque instant i. Ce modèle est le modèle le plus utilisé généralement dans les applications de flots de données.

– *Le modèle du tourniquet* : dans ce modèle, et contrairement au modèle précédent de la caisse enregistreuse, chaque élément a_i peut soit s'ajouter soit se soustraire à la valeur précédente dans $A[j]$. Ainsi, pour chaque élément $a_i = (j, I_i)$ avec I_i positif ou négatif, $A_i[j] = A_{i-1}[j] + I_i$. Bien que ce modèle soit le plus général possible, il est très rarement utilisé par les applications. Ceci est dû généralement à la difficulté de calculer des bornes sur la variation du signal A.

2.3.3. Gestion du temps sur le flot des donnes

Parmi les caractéristiques les plus intéressantes du flot de donnes est le temps d'arrivée de chaque sur le flot appelée« estampille temporelle ». Il est toujours possible de générer cette estampille en s'appuyant sur la date de réception de l'élément par le système traitant ce flot. Dans la mesure où la taille d'un flot de données est non bornée et potentiellement infinie, il n'est pas réaliste, de stocker et traiter l'intégralité d'un flot, il est donc essentiel de trouve une

solution, comme par exemple se limiter à une partie du flot. Dans la littérature, cette partie est généralement appelée fenêtre [14]. Nous décrivons ci-dessous les principaux types de fenêtres existants : [23]

– **La fenêtre fixe**. Cette fenêtre est une portion stricte du flot de données. Par exemple la fenêtre stricte entre les données du début du mois d'Avril jusqu'à la fin du mois de Mai.

–**La fenêtre Point de Repère**. Si en revanche l'une ou l'autre des dates est relative on parle de fenêtre point de repère, une fenêtre entre le 1er mars et le temps courant en est un exemple.

– **La fenêtre glissante**. Cette fenêtre contient deux bornes dynamiques mises à jour selon deux conditions possibles : soit par réévaluation après chaque nouvelle arrivée d'un élément sur le flot (réévaluation avide), soit la réévaluation se fait après une période de temps bien précise choisie par l'utilisateur.

2.3.4. Extraction de séquences à partir du flot des données

Très peu d'algorithmes ont été proposés pour l'extraction de séquences fréquentes sur les flots de données. On va détailler trois approches de fouilles de données:

Dans [15], Marascu et al. Proposent l'algorithme SMDS (Sequence Mining in Data Streams) qui permet l'extraction de motifs séquentiels sur les flots. L'algorithme SMDS se base sur une approche par segmentation (clustering) de séquences. Les séquences sont regroupées selon une méthode d'alignement et permet ainsi de construire au fur et à mesure de l'avancée du flot des ensembles de séquences similaires.

Dans [16] un algorithme d'extraction de motifs séquentiels maximaux sur les flots de données appelé SPEED a été proposé. L'idée principale est de toujours maintenir une bordure qui contiendrait les séquences maximales sous-fréquentes potentielles et qui pourraient devenir fréquentes au fur et à mesure de l'évolution du flot. L'approche se base sur l'utilisation de tables de fenêtres temporelles logarithmiques afin de pouvoir garder un historique étendu des motifs séquentiels maximaux. L'algorithme SPEED procède à une transformation des itemsets afin de représenter chacun des itemsets au moyen d'un seul entier. Pour ce faire, nous associons à chaque item un nombre premier, cette association permet ainsi d'utiliser des propriétés de congruences et d'arithmétiques simples pour les tests d'inclusions de séquences.

La troisième approche d'extraction de motifs séquentiels est quelque peu différente des deux précédentes puisqu'elle prend en compte l'extraction sur un ensemble de flots de données. Les auteurs Chen et al[17] considèrent ainsi que les motifs extraits de cet ensemble de flots de données sont des motifs multidimensionnels. Les auteurs choisissent un modèle de fenêtre glissante avec un système d'alignement des transactions des différents flots de données afin

d’extraire ces motifs séquentiels multidimensionnels. Leur algorithme, nommé MILE, utilise l’algorithme PrefixSpan à chaque étape de glissement afin d’extraire ces motifs.

2.3.5. Méthodes de traitement des flots des données

Il existe trois types d'approches de traitement des flots de données :[18]

_ *agglomératif (ou itératif)* : les données sont traitées par ordre d'apparition, une après l'autre ;

_ *par paquets* : les données sont groupées dans des paquets d'une même taille fixe et ensuite les paquets seront traités l'un après l'autre.

_*à l'aide des fenêtres glissantes* : une fenêtre de taille fixe dont le premier point glisse dans le temps ou à l'arrivée de nouvelles données.

Selon la manière dont la taille de la fenêtre (ou du paquet) est établie, les deux derniers types se divisent à leur tour dans deux sous-types :

_ *taille établie selon le nombre de données* ;

_ *taille établie selon l'intervalle de temps.*

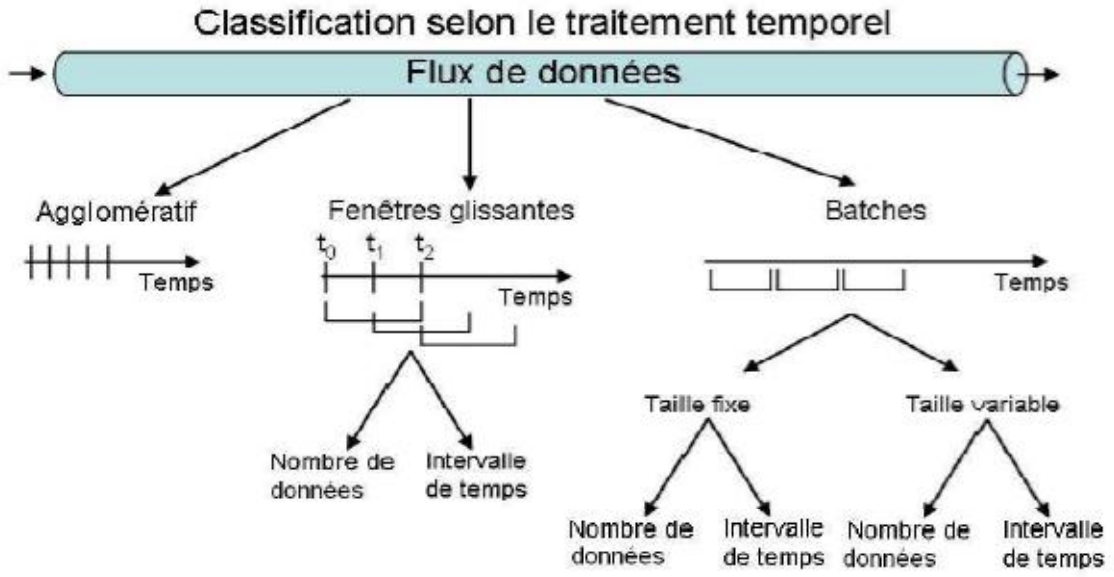


Fig. 2.4: Classification des méthodes de traitement des flots de données.

2.3.6. Domaines d'applications

Il apparaît que de nombreuses applications génèrent des données sous la forme de flux de données, ces dernières se retrouvent dans des domaines extrêmement variés allant des technologies de l’information à la physique prévisionniste. Et selon [19] les flots de données se divisent en deux grandes catégories :

_ *Flots de données transactionnels* ;

_ *Flots de données de mesure*.

La première catégorie s'intéresse aux interactions entre des entités, alors que la deuxième surveille leurs évolutions dans le temps.

- **Flot de données transactionnelles**
  On trouve par exemple : les transactions bancaires, les télécommunications, le Web…
- **Flot de données de mesure**

Les flots de données de mesure surveillent l'évolution des états des entités.

On peut les trouver dans : la surveillance des réseaux IP, La détection des spams, La détection d'anomalies, Réseaux de Capteurs

2.4. Conclusion

En fin de ce chapitre, on peut dire que la problématique de la recherche de motifs séquentiels est apparue pour permettre l'extraction des comportements typiques au cours du temps et répondre ainsi aux besoins de nouvelles applications (gestion d'alarmes dans les réseaux de télécommunication, analyse de comportements de clients, analyse des accès dans un serveur Web, ….etc). On peut dire que la fouille de flots de données est la technique qui consiste à explorer les données qui arrivent en un flot continu, illimité, avec une grande rapidité, et dont certains paramètres fondamentaux se modifient avec le temps, comme par exemple :l'analyse des flots de données émis par des capteurs domotiques. Parmi les exemples d'applications qui peuvent apparaitre dans les domaines d'activités de la vie de tous les jours citons :la surveillance des activités les personnes âgées pour assurer un suivi médical et permettent d'évaluer leurs autonomie.

Le chapitre suivant discute l'implémentation d'un algorithme qui fait extraire la liste de top-k épisodes réguliers concernent les différentes activités des personnes âgées dans environnement intelligente.

Chapitre 3 : Top-K épisodes réguliers

3.1. Introduction

Dans les dernières années est apparue une problématique sociale marquée par le vieillissement de la population. Du plus en plus l'importance que les individus accordent à leur autonomie, autrement dit, de rester indépendants dans leurs propres maisons. Pour cette raison, des systèmes de surveillance à distance qui assure un suivi médical et une assistance sociale pour l'aide les personnes âgé ont été mises en place.

Un système de cette famille assure par l'émergence des nouvelles technologies telles que les maisons intelligents et les capteurs des donnes. Les capteurs permettent de collecter les traces des activités quotidiennes des personnes à contrôler.

Dans ce chapitre, nous allons présenter les motivations qui portent sur l'émergence des systèmes d'assistance des personnes âgée à domicile, puis on définit les différentes technologies utilisées par ce système comme les maisons intelligentes et les systèmes des capteurs.

Enfin, nous illustrerons le travail d'implémentation pour l'extraction des épisodes réguliers. On s'intéressera exactement sur la liste des top-k épisodes les intéressantes.

3.2 Motivation

Le système de supervision à domicile a été arrivé pour résoudre le problème d'assistance des personnes âgées pour assure leur indépendances et leur autonomies. Ce système joue un rôle plus important par suivi des activités de la vie quotidienne des pulpations fragile pour objectif de maintien à domicile, avec un renforcement de leur sécurité et une amélioration de leur qualité de vie. L'assistance à domicile est une nouvelle discipline née grâce aux technologies d'assistance et l'intelligence ambiante, dans ce contexte on trouver deux notions essentielles représentant les différents types de technologies utilisées pour mesurer l'activité, à savoir les maisons intelligentes et le capteur des données.

- **Maisons intelligentes :** on peut définir une maison intelligentes comme un environnement intelligent équipe par différents types des technologies tel que les

capteurs et les actionneurs pour assure une meilleur qualité de vie avec la prise en compte des besoins des habitants comme l'indépendante particulièrement dans le cas des personnes âgées.

- **Les capteurs :** sont des dispositifs mobiles qui peuvent être utilisé pour mesure les activistes des personnes qui habitent dans un espaces intelligent et détecter l'interaction entre ces personnes et leur environnement. Ils sont aussi exploite pour assure un meilleur suivi du comportement des habitants qui permet de déduire par exemple l'état sante ou bien les besoins d'assistance.

Il existe plusieurs type de système de capteur destinés au maintien à domicile (Fig 3.1) tel que les capteur de donnes physiologiques, les capteurs d'activités et le capteur environnementaux, en plus on peut trouver des capteur fixe sur différentes parties du corps et d'autres installes dans la maison intelligente par exemple : vêtement intelligents, bracelet aphycare, système TRIDENT, LOCADYN 3D...

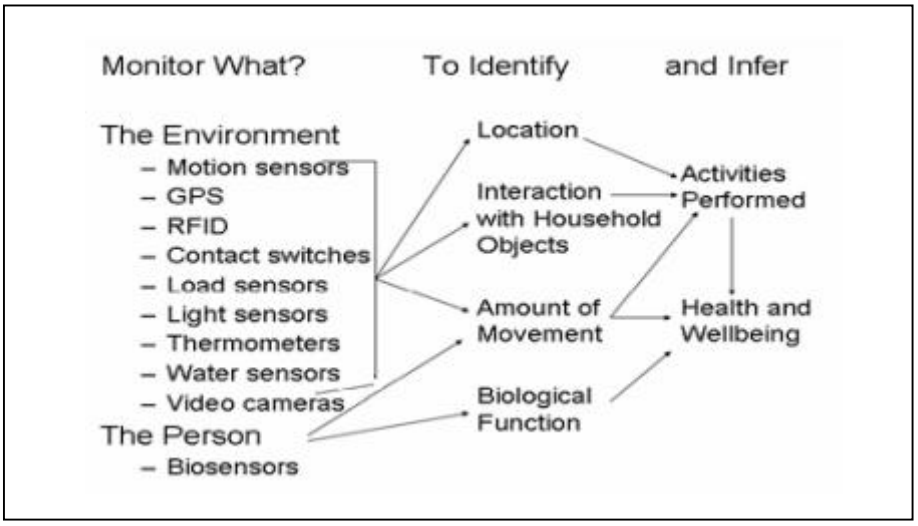


Fig 3.1 : Capteurs pour la surveillance de l'activité. [24]

3.3. Extraction des épisodes réguliers

Le problème d'extraction des épisodes a été introduit par Mannila,H. et al. dans [21], celui des épisodes réguliers a été récemment étudié par Komate. A et al. dans [20].

Dans ce qui suit nous présentons les notions et les définitions nécessaires à la compréhension des extraction des épisodes réguliers, cette présentation est issue de [20, 21,22], et notre but est d'implémente l'algorithme proposé par (Komate.A , Julie.S, Philippe.L) [20] .

3.3.1. Définition du problème

Nous présentons ici les concepts utilisés pour la découverte d'épisodes à partir de flux de capteurs. Nous introduisons aussi le problème de l'extraction du top- k épisodes réguliers.
Un flux d'événements est une séquence potentiellement infinie des événements classés:

$$DS = \langle (e_1,t_1), (e_2,t_2),\ldots.. \ldots (e_i,t_i),\ldots\ldots\rangle$$

Où (e_i,t_i) i^{éme} événement dans la séquence, avec e_i l'étiquette de l'événement prenant des valeurs dans un alphabet fini ξ; et t_i l'estampille temporel (timestamp) de l'événement. Les événements sont triés par leurs estampille temporel (pour tout i, $t_i \leq t_{i+1}$)).

Une épisode $E = \{e_1,e_2 ,\ldots e_n\}$ est un ensemble (non ordonnées, sans étiquettes d'événement en double)*de* n étiquettes d'événements sur ξ. Les étiquettes de groupe d'épisodes qui se produisent ensemble, mettant ainsi en évidence les relations entre les événements. Ils constituent une abstraction caractérisant les activités qui se déroulent dans la maison sans connaissance préalable ou expert.
Dans les définitions suivantes, E fait référence à une épisode avec n étiquettes d'événement $\{e_1,e_2,\ldots,e_n\}$.
L'objectif dans ce problème est de découvrir des épisodes réguliers dans le passé récent, et d'actualiser (mettre à jour) ces connaissances lorsque de nouveaux événements se produisent. Nous considérons donc un modèle de fenêtre coulissante, où une fenêtre W est composé de m lots consécutifs, à savoir $W = (B_i, B_{i+1}, \ldots ,B_{i+m-1})$, où chaque lot B_i est une séquence d'événements. Quand un nouveau lot de données B_{i+m} arrive, B_i devient obsolète et sera retiré de la fenêtre W. Les lots couvrent une période ou une unité de temps fixée par l'utilisateur (par exemple, un jour, une semaine, etc.)

Définition1 : (occurrence d'une épisode E). Il y a un occurrence o de E entre les temps t_1 et t_n, s'il existe une permutation p (1, ... n) et n times tamps (t_1, ... t_n), de sorte que o =$\langle(e_{p\,(1)}, t_1)$, ... $(e_{p(n)},t_n)\rangle$ est une sous-séquence de la fenêtre W. o est appelé une T_{ep}-*occurrence* si : $t_n - t_1 < T_{ep}$. Cette condition constitue une contrainte en plus dans ce problème, c'est-à-dire, que seuls les T_{ep}-occurrences sont pris en compte dans les mesures de régularité.

Définition2 :(occurrence minimal d'un épisode *E*). Soit *o* une occurrence de *E, s'étalant entre* $t_1$et t_n. *o* est une occurrence minimale s'il y a pas d'occurrence *o′* de *E*, à partir de t_1' à t_n'tel que :$t_1 \leq t_1'$, $t_n' \leq t_n$, et $t_n' - t_1' < t_n - t_1$.

Définition3 : (Les occurrences nonchevauchantes d'un épisode*E*). Soient *o* et *o′* des occurrences minimales de *E* allant respectivement, à partir de t_1 à t_n et t_1'à t_n' . *o* et *o′* sont non-chevauchantes si min (t_n, t_n')< max (t_1, t_1'). La liste des occurrences (Tep) minimale nonchevauchantes de *E* est notée par NMO^E**.**

Définition4 :(Régularité d'un épisode *E*). Soit NMO^E la séquence ordonnée des occurrences minimales non chevauchantes de *E*. La régularité r^E de l'épisode *E* est définie comme la valeur maximale entre :

- La régularité entre l'heure de début de la fenêtre (t_{sw}) et l'heure de début de la première occurrence minimale dans la $NMO^E(t_1)$:$r_{sw} = t_1 - t_{sw}$
- La régularité entre chaque paire d'occurrences consécutives de la $NMO^E o_u$, allant de t_1 à $t_n$et o_{u+1}allant de t_1' à t_n' : $r_u = t_1' - t_n$
- La régularité entre la dernière occurrence $o = \langle (e_{p(1)}, t_1), \ldots\ldots\ldots (e_{p(n)}, t_n) \rangle$ en NMO^Eet le dernier de la fenêtre (t_{ew}): $r_{ew} = t_{ew} - t_n$

On peut remarquer qu'un épisode *E* est plus régulier qu'un autre épisode *E′* si sa valeur de régularité est plus faible.

Definition5 : (Top-*k* épisodes réguliers). On s'intéresse à la liste des épisodes classée par ordre croissant de leurs régularités. On peut dit qu'un épisode *E* appartient une liste des top-*k* épisode réguliers s'il n'y a pas plus de *k* - 1 épisodes ayant des valeurs inférieures de régularité à celle de *E*.

Avec l'ensemble des paramètres donnés par l'utilisateur: un certain nombre d'épisodes *k* intéressant voulu, une durée de traitement par lots, un nombre *m* de lots dans la fenêtre, la durée maximale d'épisodes *Tep*-occurrences; nous abordons le problème de l'extraction du top-*k* épisodes réguliers. Autrement dit, nous découvrons les *k* épisodes avec les valeurs de régularité plus bas dans la fenêtre coulissante sur flots de données DS.

3.3.2. Algorithme d'extraction des top-k épisodes réguliers (*TKRES*[4])

Dans cette section, nous présentons *TKRES [20]*, un algorithme à une passe efficace pour l'extraction du top-k épisodes réguliers dans un flux de données de capteur. *TKRES* recherche

[4]Top –K Regular Episodes.

les épisodes dans une fenêtre glissante contenant *m* lots consécutifs d'événements, et peut être divisé en deux étapes principales: l'initialisation, c'est-à-dire l'extraction du top- k épisodes réguliers de la première fenêtre (les premiers *m* lots du flux d'entrée B_1 a B_m), Et la mise à jour avec un lot entrant, la mise à jour des connaissances sur les *top-k* épisodes réguliers présents dans la nouvelle fenêtre (les lots précédents, sauf le plus ancien lot, plus le nouveau lot entrant).

Algorithme de l'extraction initiale

Algorithm 1*TKRES* –initial mining

Input: k : number of episodes to be discovered,m batches of sensor data (B_1,... B_m)
Output: top- k list containing in the top-k most regular episodes,
k -tree,containing the occurrence information for the episodes on(B1,... Bm)

1: Initialize the tree, and create entries for all single events
2: **for each** batch B_i**do**
3: **for each** event (e_j, t_j)in B_i**do**
4: update e_j 's occurrence information in the tree with timestamp t_j
5: Compute theregularity of each event label
6: Collect the k labels with the lowest regularity into the sorted top-k list
7: Compute the depth dot of the k-tree to be created
8: **for** depth d = 1to dot-1**do**
9: **for each** episodes X and Y at depthd having d-1 common labels **do**
10: Merge episodes X and Y to be Z (it contains thus d+1labels)
11: Create a node for Z in the k-tree, setto be a child of X
12: Get the occurrence times for Z from X and Y. Infer its regularity r^Z
13: **if** $r^Z < r^{kth}$**then**
14: Remove the kth episode from the top-k list, insert Z
15: **for** depth d = dot to $|\xi|$ **do**
16: **for each** pair of episodes X and Y at depthd with d-1 common labels where $r^X \leq r^{kth}$ and $r^Y \leq r^{kth}$**do**
17: Merge episodes X and Y to be Z
18: Get the occurrence times for Z from X and Y. Infer its regularity r^Z
19: **if** $r^Z < r^{kth}$**then**
20: Remove the k^{th} episode from the top-k list, insert Z
21: Create a node for Z in the k-tree, setto be a child of X on depth d+1

Algorithme de l'extraction à l'arrivée du nouveau lot (batch)

Algorithm2*TKRES* –mining a new incoming batch of sensor data

Input: k : number of episodes to be discovered, B_{i+m}: the new batch,
k -tree,with the occurrence information for the episodeson(B_i,... B_{i+m-1})
Output: top- k list containing in the top-k most regular episodes,
k -tree,with the occurrence information for the episodeson(B_{i+1},... B_{i+m})

1: Empty the top-k list
2: Remove all the nodes at depth higher than dot
3: For each node, remove the occurrence times occurring during Bi
4: **for each** event (e_j, t_j) in the new batch B_{i+m}**do**
5: Collect t_j in the node for e_j in the k -tree
6: Recompute the regularity for the episodes at depth 1
7: Collect the k labels with the lowest regularity into the sorted top-k list
8: **for** depth d = 1 to dot **do**
9: **for each** episodes X and Y at depth d having d − 1 common labels **do**
10: Merge episodes X and Y to be Z
11: Get the occurrence times for Z from X and Y (Only for what is occurring during B_{i+m}).Infer its regularity r^Z
12: If it is not already in the tree, create a node for Z, set to be a child of X
13: **if** $r^{X \cup Y} < r^{kth}$**then**
14: Remove the k^{th} episode from the top-k list, insert Z instead
15: **for** depth d = dot to $|\xi|$ **do**
16: **for each** pair of episodes X and Y at depth d with d − 1 common labels
where r^X and $r^Y \leq r^{kth}$**do**
17: Merge episodes X and Y to be Z
18: Get the occurrence times for Z from X and Y. Infer its regularity r^Z
19: **if** $r^Z < r^{kth}$ **then**
20: Remove the k^{th} episode from the top-k list, insert Z instead
21: Create a node for Z in the k-tree,setto be a child of X on depth d + 1

3.4. Implémentation

3.4.1. Environnement d'exécution de l'algorithme *TKRES*

Nous avons essayé d'implémenté l'algorithme *TKRES* dans un environnement du Java comme un langage de développement. Nous examiner cette algorithme a une base de données real appelle *Aruba* de projet CASAS[5] du Université Washington "WSU"(on donne une brève définition ci-dessus) appartient les traces de mouvement à travers des capteurs d'une personne âgée habite dans un environnement intelligent (Smart Home) indiqué par des étiquettes avec date et heure.

- **Projet CASAS :**

Dans le projet CASAS, l'habitat est vu comme un agent intelligent qui perçoit son environnement à partir d'un ensemble de capteurs et réagit à travers des dispositifs capables de piloter le chauffage ou la lumière. Le but principal est de mettre en œuvre un contrôleur intelligent capable de raisonner pour minimiser le coût de fonctionnement de l'habitat été n même temps d'améliorer le confort des personnes. Les expérimentations sont effectuées dans un appartement situé sur le campus de l'université de Washington, appartement qui comporte une salle de séjour, une cuisine, des toilettes et trois chambres (Fig 3.2).

Fig3.2 :Plan d'une maison intelligent (Smart Home) avec la position des capteurs.

[5]Center for Advanced Studies in Adaptive Systems

3.4.2. Les étapes d'exécution

L'exécution de l'algorithme *TKRES* avec base de données *Aruba* pour extraction les tops K épisodes réguliers selon les étapes suivant :

- L'utilisateur entre des paramètres prédéfinis :

  -k : nombre des épisodes réguliers pour être extrait.

  -m : nombre des lots avec leur durée qui consiste le fenêtre.

  -T_{ep} : temps d'occurrence maximale de l'épisode.
- Construire l'arbre préfix avec les étiquettes représentant les épisodes de taille 1.
- Pour chaque lot arrive en affecte les temps des événements a les nœuds d'arbre.
- Calcule la régularité pour chaque épisode de taille 1.
- Extraire les k épisodes les plus réguliers.
- Calculer la profondeur de l'arbre.
- Génère les épisodes plus long a travers de chaque pair des épisodes de taille 1jesuqe atteindre la profondeur qui calculer.
- Mise à jour la liste de top k épisodes réguliers.

3.4.3. Interface d'application

- Utilisateur donne les paramètres prédéfini : [K-Durée de lot-Nombre de lot-T_ep]

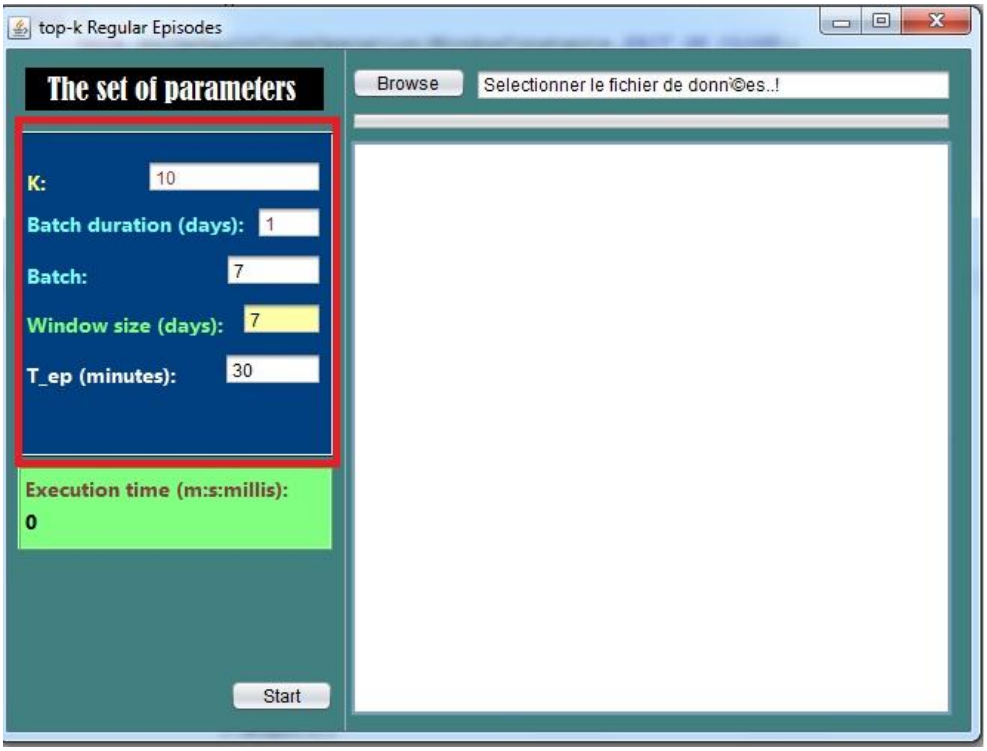


Fig3.3. Saisie des paramètres par utilisateur.

- Charger la base de donne *Aruba* qui disponible a site de centre CASAS du université de Washington(http://ailab.wsu.edu/casas/datasets/):

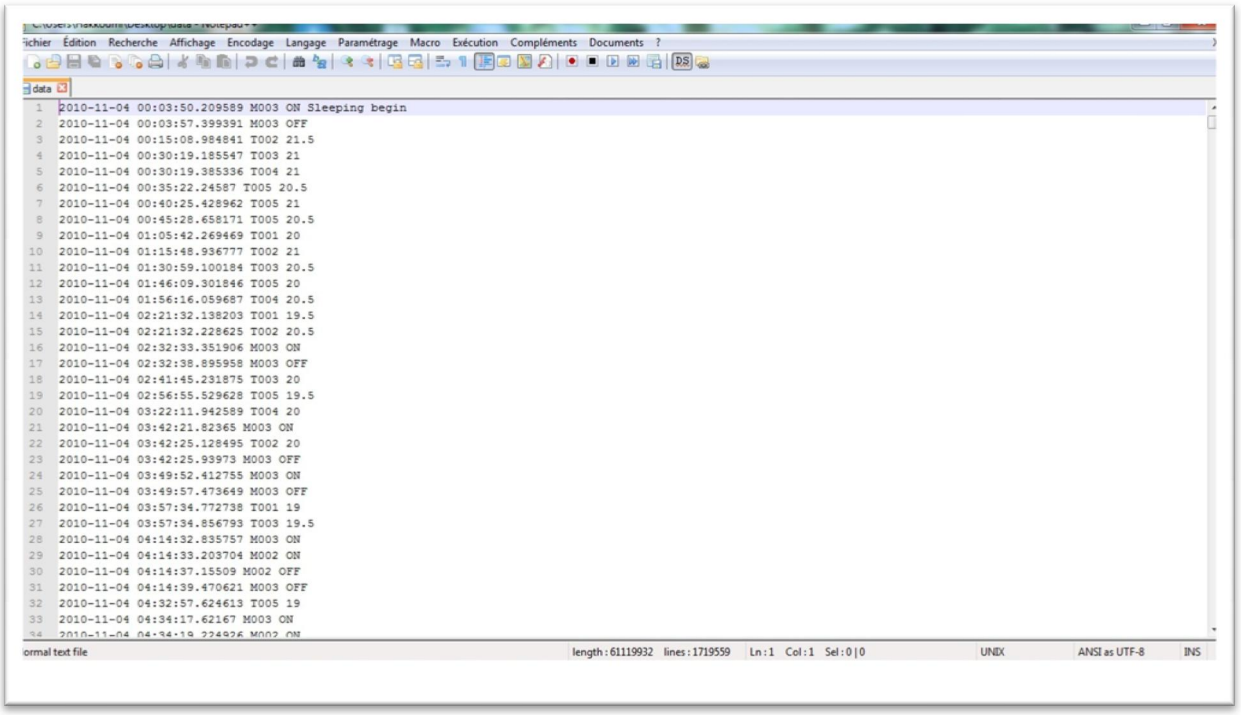


Fig3.4. Base de données Aruba de centre CASAS.

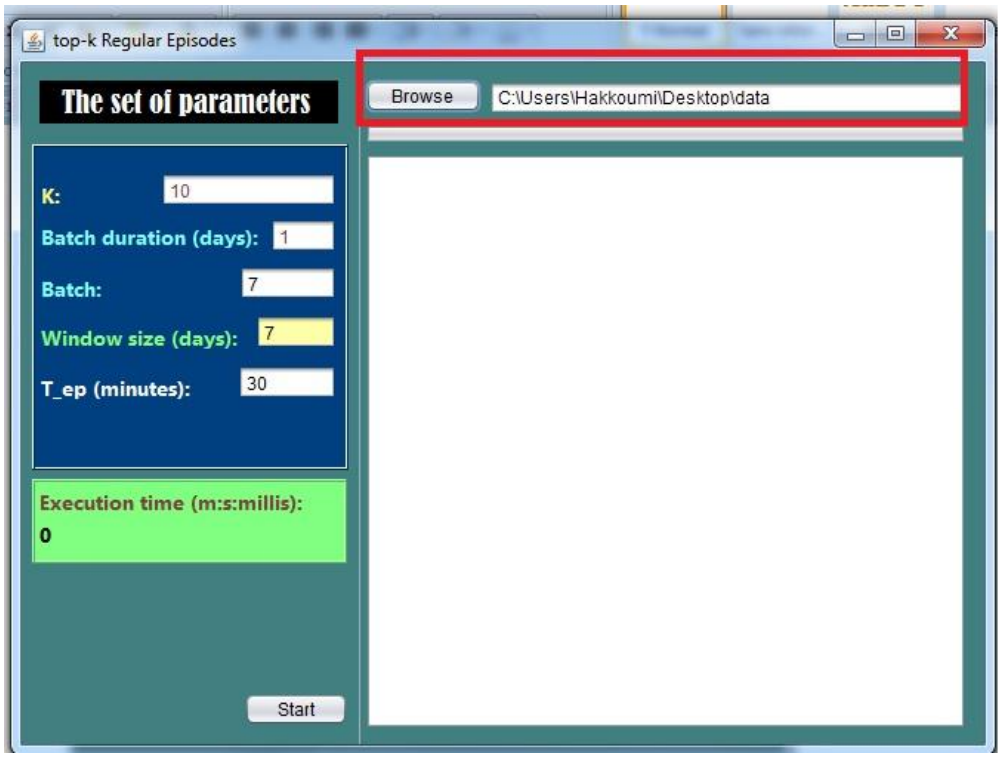


Fig3.5. Parcours et chargement de la base de donné *Aruba*.

- Afficher la liste de top-K épisodes réguliers :

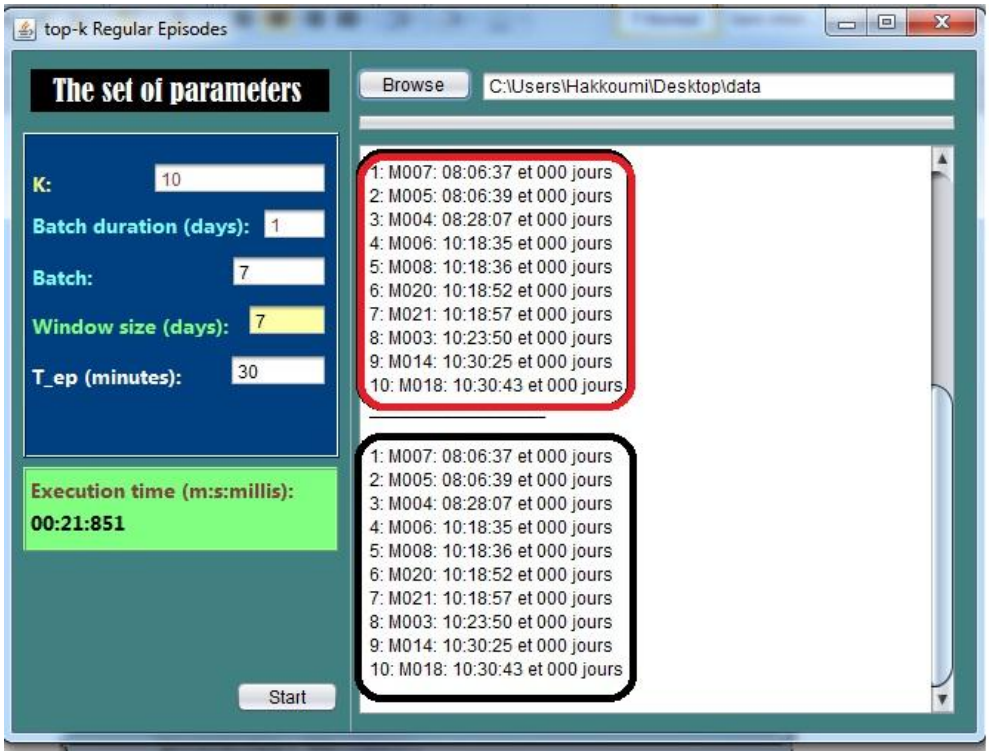


Fig3.6.La liste des top-k épisodes réguliers

- Les temps d'exécution sont différents selon les paramètres qui donnent par l'utilisateur :

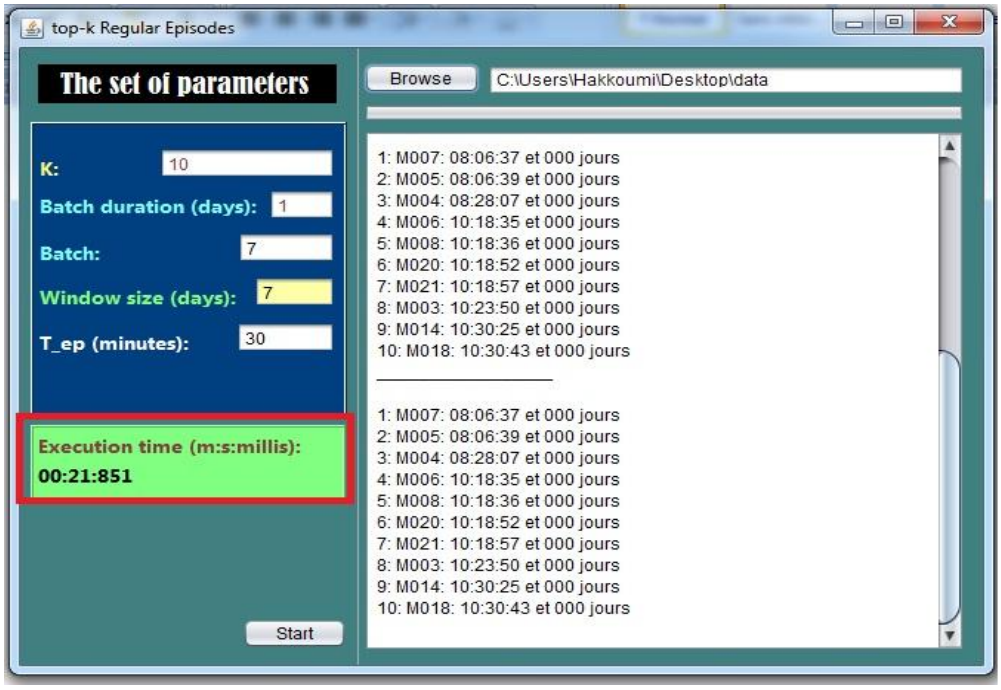


Fig3.7.Temps d'exécution pour chaque ensemble des paramètres.

On peut dit que l'extraction des épisodes le plus réguliers donne la trajectoire régulière dans l'appartement, et peut aider les médecins et les assistants sociaux pour améliore et adapter disposition de la maison où les personnes âgée a habité.

3.5. Conclusion :

Dans ce chapitre, premièrement nous avons présenté un système de monitorage des activités d'une personne à domicile en utilisant des capteurs vidéo et des capteurs d'activité. L'analyse de la qualité de la méthode et des résultats a mis en évidence à la fois les potentialités et la complexité de mise en pratique de la combinaison d'informations issues de caméras et des capteurs de d'activité. Le manque de données expérimentales issues d'un système réel ou des personnes âgées simulent des scénarios qui relatent leur activités quotidiennes à domicile n'a cependant pas permis une validation complète des résultats obtenus.

Puis nous avons présenté l'implémentation d'un algorithme appelle TKRES qui devise en deux étapes l'initialisation et la mise à jours de la fenêtre par un entrant d'un nouvelle lot. Comme un résultat de cette algorithme nous extraction un liste de top k épisodes réguliers.

Conclusion générale et perspectives

Dans ce travail notre objectif est d'étudier une modélisation pour l'extraction des connaissances nécessaire à partir d'une grande quantité des donnes homogène, grâce à ça nous avons passé par la définition de data minig (DM) en générale qui base sur le processus KDD (processus d'extraction de connaissance) puis et selon les besoin de notre travail nous avons expliqué une partie de fouille de donne à savoir la fouille de donnes séquentielle, et en dernier partie en vue l'algorithme qui fait l'extraction des séquences sur le flot de donnée pour trouver top- k épisodes réguliers .

Dans ce mémoire, nous avons étudié le problème de la fouille des séquences d'événements appliquée à la surveillance des personnes âgées à domicile à partir de flot de données enregistrées par un ensemble de capteurs (vidéo et activité),Tout écart par rapport à un profil comportemental est susceptible de correspondre à une situation inquiétante ou critique. Ce problème présente des difficultés lies à la gestion efficace à la fois de la mémoire et le temps d'exécution. Pour cela, les algorithmes dans ce champ utilise le principe des fenêtres glissante, aussi nous avons implémentes une structure d'arbre de préfixe dont la profondeur est bornée et fixe à l'avance.

Comme extension à ce travail, on peut envisager :

1. D'exploiter la philosophie de flots de données dans d'autre application comme vidéosurveillance,
2. Le passage à l'échelle : explorer des algorithmes et structure des données plus efficaces.
3. Parallélisation et distribution de ces algorithmes.

Nous espérons à notre travail d'apporte une validation pratique pour fouille des donnes séquentielles comme Data Stream, et donne une bonne plate-forme d'autre pour réaliser des applications dédiées à ce domaine.

Bibliographie

[1] EDWIN DIDAY. Data mining et analyse des ventes d'une chaine de magasins, PROMOTION 2002.

[2]Zighed & Rakotomalala, « Extraction des Connaissances à partir des Données (ECD) », in Techniques de l'Ingénieur, 2002.

[3]PHILIPPE BESSE. Data mining 2 modélisations statistiques et apprentissage. LABORATOIRE DE STATISTIQUE ET PROBABILITES, janvier 2003.

[4] Processus de découverte d'information.
http://www.grappa.univ-lille3.fr/polys/fouille/sortie004.html

[5] Mémento techniques n 14 base de donnée et système d'information3
http : //www.rd.francetelecom.com/fr/conseil/mento14/chap6c.html.

[6] Zaïane, O.R. cours cmpt690 principles of knowledge discovery in databases, 1999.

[7] Http://chirouble.univ-lyon2.fr/~ricco/data-mining/logiciels : liste de logiciels libres de data mining.

[8] STEPHANE TYFFERY. Data mining et scoring (bases de données et gestion de la Relation client), 1996.

[9] R. Agrawal R. Srikant : Mining sequential patterns. In Proceedings of the 11th International Conference on Data Engineering (ICDE 95), pages 3–14, Tapei, Taiwan, 1995.

[10] R. Agrawal, T. Imielinski et A. Swami : Mining association rules between sets of items in large database. In Proceedings of the International Conference on Management of Data (ACM SIGMOD 93), pages 207–216, 1993.

[11] R. Srikant et R. Agrawal : Mining sequential patterns : Generalizations and performance improvements. In Proceedings of the 5th International Conference on Extending Database Technology (EDBT 96), pages 3–17, Avignon, France, 1996.

[12] M.J. Zaki : Spade : An efficient algorithm for mining frequent sequences. Machine Learning, 42(1/2):31–60, 2001.

[13] J. Pei, J. Han, B. Mortazavi-asl, H. Pinto, Q. Chen et U. Dayal : Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth. In Proceedings of 17th International Conference on Data Engineering (ICDE 01), pages 215–224, Heidelberg, Germany, 2001.

[14] C. Aggarwal, éditeur. Data Streams : Models and Algorithms. Springer, 2007.

[15] A. Marascu et F. Masseglia : Extraction de motifs séquentiels dans les flots de données d'usage du web. In EGC, pages 627–638, 2006.

[16] C. Raissi, P. Poncelet et M. Teisseire : Speed : Mining maximal sequential patterns over data streams. In Proceedings of the 3rd IEEE International Conference On Intelligent Systems (IS2006), Westminster, UK, 2006.

[17] G. Chen, X. Wu et X. Zhu : Mining sequential patterns accross data streams. Rapport technique CS-05-04, University of Vermont, march 2005.

[18] A.Marascu.: "Extraction de motifs séquentiels dans les flux de données ".Thèse de doctorat de l' université Nice - Sophia Antipolis, Septembre 2009.Page 75.

[19] Koudas, N. & Srivastava, D. (2005). Data stream query processing. In ICDE '05 : Proceedings of the 21st International Conference on Data Engineering, 1145, IEEE Computer Society, Washington, DC, USA.

[20] Komate. A, Julie. S, Philippe .L. « Mining top-k Regular Episodes fromSensor Streams ».7th International Conference on Advances in Information Technology.2015

[21] Mannila,H., Toivonen,H.,Verkamo.I. « Discovery of Frequent Episodes in Event Sequences ».Data Mining and Knowledge Discovery 1, 259–289 .1997.

[22] Avinash. A,Srivatsan.L , Sastry P. S. . « A unified view of the apriori-based algorithms for frequent episode discovery ». Springer 2011.

[23] Chedy RAISSI, « Extractions des Séquences Fréquentes : des bases de données statiques aux flots de données».Thèse de doctorat de l'université de Montpellier 2, 2008.

[24] Nadia Zouba,« Analyse Multicapteur pour la Reconnaissance d'Activités Humaines».Rapport de Stage de l'Universités Paris 8, 2006.