

ACADÉMIE DE MONTPELLIER
UNIVERSITÉ MONTPELLIER II
— SCIENCES ET TECHNIQUES DU LANGUEDOC —

MÉMOIRE DE STAGE DE MASTER

SPÉCIALITÉ : Recherche en Informatique
Mention : Informatique, Mathématiques, Statistiques

Préservation de la vie privée et
Fouille de données

ALI RAMMAL

Date de soutenance : 21 Juin 2006

Effectué au sein du Laboratoire de Génie Informatique et d'Ingénierie de
Production de l'Ecole des Mines d'Alès

Sous la direction de ANNE LAURENT ET PASCAL PONCELET

Remerciements

Je tiens tout d'abord à exprimer toute ma gratitude à Monsieur YANNICK VIMONT, Directeur du Laboratoire de Génie Informatique et d'Ingénierie de Production de l'Ecole des Mines d'Alès, pour m'avoir permis de réaliser mon stage de Master dans son laboratoire.

Je remercie également Monsieur ROLAND DUCOURNAU, Responsable de la Spécialité Recherche en Informatique du Master mention IMS à l'université Montpellier II, de m'avoir accueilli au sein du département Informatique.

Je tiens à remercier mes tuteurs Madame ANNE LAURENT et Monsieur PASCAL PONCELET, pour leur encadrement, leur présence, leurs connaissances, ainsi que l'accueil qu'ils m'ont fait au sein de leur groupe de recherche. En fait c'est grâce à leur support, leurs remarques, leurs efforts, leurs conseils, leurs contrôles, leur direction et toutes les informations qu'ils m'ont fournies que ce mémoire voit le jour sous cette forme.

Je remercie également Madame SYLVIE CRUVELLIER la secrétaire du laboratoire, et Madame FRANCOISE ARMAND la responsable de la bibliothèque.

J'exprime également ma reconnaissance et mes remerciements au thésard CHEDY RAISSI, qui m'a accueilli et aidé avec beaucoup de patience tout au long de mon stage.

Enfin, je ne peux jamais oublier mes parents qui m'ont soutenu par leur support moral.

Table des matières

1	Introduction	3
1.1	Contributions	5
1.2	Organisation du mémoire	5
2	Problématique et travaux antérieurs	7
2.1	Définitions et problématiques	8
2.2	Principales méthodes d'extraction de motifs séquentiels	10
2.2.1	Méthodes basées sur Apriori (breadth-first)	10
2.2.2	Méthodes basées sur le principe depth-first	11
2.2.3	Recherche des motifs séquentiels fermés	13
2.3	Préservation de la vie privée et fouille de données	14
2.3.1	Préservation de la vie privée et motifs séquentiels dans un contexte collaboratif	15
2.3.2	Les motifs k -Anonymes	16
2.4	Discussion	17
3	Proposition	18
3.1	Définitions préliminaires	19
3.2	Inférence du support et menace de l'anonymat	22
3.3	Les canaux d'inférence pour les motifs séquentiels	23
3.4	Blocage des canaux d'inférence	30
3.5	Discussion	32
4	Conclusion	33
4.1	Perspectives	33
	Bibliographie.	35
	Annexes.	37

Chapitre 1

Introduction

Motivés par des problèmes d'aide à la décision, les chercheurs de la communauté Fouille de Données se sont intéressés, depuis une dizaine d'années, à l'élaboration de nouveaux algorithmes adaptés aux bases de données manipulées. Les connaissances extraites par ces algorithmes sont très variées puisque l'on peut retrouver par exemple des règles d'association, des motifs séquentiels, des regroupements (*clustering*), de la classification, Toutefois, les bases de données manipulés se sont avérées de plus en plus complexes au cours du temps. Ainsi, pour répondre aux nouvelles contraintes, des propositions ont permis :

- d'être plus efficaces en temps de réponses ;
- de prendre en compte de plus en plus de contraintes (temporelles, spatiales, sur les résultats attendus, ...);
- de manipuler des données de plus en plus complexes (semi-structurées, non structurées, multi-médias, ...);
- de maintenir la connaissance extraite malgré les évolutions des données
- de traiter des données disponibles sous la forme de flots ;
- ...

Aujourd'hui la communauté se retrouve confrontée à un nouveau challenge : *comment garantir que la connaissance extraite par des algorithmes de fouille ne puisse pas nuire à la vie privée ?*

Cette problématique est récente et trouve son origine dans l'adoption aux Etats-Unis de la *Health Insurance Portability and Accountability Act* (HIPAA) proposée initialement en 1996, et qui a instauré, en avril 2003, un régime de protection des renseignements personnels en matière de santé. Le département de la santé et des services sociaux américains a, de plus,

adopté un règlement couvrant les protections administratives, matérielles et techniques qui doivent être respectées pour garantir la confidentialité et l'intégrité des données personnelles de santé enregistrées ou transmises par voie électronique. Ces spécifications sont à l'heure actuelle étendues à de nombreux pays (Australie, Asie, Japon, ...) et les premières ébauches de projets apparaissent au niveau Européen. Bien entendu, ce problème de l'intégrité personnelle peut être étendu à de nombreux domaines d'applications (analyse de transactions financières, analyse de comportements sur des sites Web, ...).

Préserver la vie privée dans un contexte de fouille de données nécessite de ne pouvoir offrir de la connaissance que si celle-ci garantit de ne pas pouvoir obtenir d'information sur les individus associés. En effet, si nous savons, par exemple, que sur une base de données médicale "100% des patients du docteur X ont suivi un traitement Y après avoir eu une ordonnance du médicament Z" nous sommes capables d'obtenir de nombreuses informations sur tous les patients du docteur X. Alors que l'on pensait que les résultats de la fouille de données ne violaient pas la vie privée, l'exemple précédent montre bien, qu'à partir de la connaissance extraite, nous sommes capables d'obtenir de nombreuses informations sur les patients, i.e. dans notre exemple nous sommes capables de connaître des traitements des patients du docteur X. En fait, il semblait évident que les modèles ou les motifs extraits concernaient un nombre important d'individus et donc dissimulaient les individus. Par exemple, même s'il a plutôt été défini pour limiter l'espace de recherche, le rôle de la notion de support minimal dans le cas des algorithmes de recherche de règles d'association semblait nous garantir que les résultats obtenus ne concernaient qu'un ensemble de la population. Pour garantir que les algorithmes de fouille de données ne violent pas la vie privée des individus, de nombreuses approches ont considéré qu'elles disposaient d'une connaissance préalable sur ce qui était sensible ou non. Cependant cette approche est très subjective et surtout très difficile à mettre en œuvre.

Aussi récemment, dans le cas de la recherche de règles d'association une nouvelle approche, appelée motifs k -anonymes (*k-anonymous patterns*), a été proposée par [ABGP05] afin de ne rechercher et, ainsi de ne fournir à l'utilisateur final, que les connaissances qui ne violent pas la vie privée des individus. C'est dans ce contexte que se situe notre travail. En effet, la proposition initiale est limitée à la recherche d'itemsets et ne permet d'obtenir que des résultats par rapport aux actions menées par les individus sans se préoccuper de l'ordre de ces actions. En étendant la notion de *k-anonymous patterns* aux motifs séquentiels, nous répondons aux deux problèmes suivants :

- nous offrons une connaissance sur le comportement des individus ;
- nous généralisons la problématique des k -anonymous patterns dans la mesure où la recherche d'itemsets est un sous ensemble de la notion de motifs séquentiels, i.e. un motif est une suite ordonnée d'itemsets

1.1 Contributions

Dans ce mémoire nous proposons une nouvelle approche de préservation de la vie privée adaptée au concept de motifs séquentiels. Pour cela, nous étendons la notion de k -anonymous patterns aux motifs séquentiels et par là même nous offrons une approche plus générale.

La recherche des motifs k -anonymes nécessite de rechercher des canaux d'inférences. Dans un premier temps nous montrons comment les approches de recherche de motifs par niveaux peuvent être adaptés pour déterminer ces canaux et donc par la même nous offrons une solution qui garantit que les motifs extraits respectent notre contrainte. Cependant, de manière à optimiser les résultats, nous étendons notre approche à la prise en compte des motifs séquentiels clos i.e. fermés. En effet, ces derniers permettent de diminuer l'espace de recherche (et donc les temps de réponses) en limitant le treillis aux motifs qui sont tels qu'il ne peut pas exister de motifs plus grands, i.e. de plus grande taille, dont le support, i.e. le nombre d'occurrences, est le même. Grâce à cette approche, nous montrons qu'il est possible de facilement étendre des algorithmes de recherche de motifs clos existant pour obtenir les k -anonymes. Nous pouvons rapidement générer les "anonymes" sans régénérer le treillis dans son intégralité.

Enfin, de manière à garantir que les connaissances obtenues respectent la vie privée, nous proposons une nouvelle approche de modification des bases de données sources qui garantit que toutes les connaissances acquises respectent la contrainte et sont cohérentes par rapport aux données initiales. L'avantage de cette approche est de proposer à l'utilisateur final toutes les solutions possibles et non pas uniquement celles qui ne violent pas la vie privée.

1.2 Organisation du mémoire

Le mémoire est organisé de la manière suivante :

Dans le Chapitre 2, nous présentons en détail la problématique étudiée.

Comme cette dernière est fortement associée aux motifs séquentiels nous rappelons également les définitions préliminaires et la problématique de la recherche de motifs séquentiels. Nous présentons également les travaux antérieurs. Dans un premier temps, étant donné que nos propositions sont basées sur différents types d'algorithmes d'extraction de motifs, nous présentons les principaux algorithmes par niveaux et de recherche des motifs clos. Nous présentons ensuite les travaux existant qui concernent la préservation de la vie privée. Ce chapitre se termine par une discussion sur les limites des approches actuelles.

Le Chapitre 3 décrit notre proposition. Tout d'abord nous étendons les définitions initiales et affinons les problématiques étudiées. Nous présentons ainsi notre première approche avec ces limitations. Nous proposons d'étendre nos travaux aux motifs clos et enfin nous montrons comment modifier la base initiale pour offrir l'ensemble des connaissances à l'utilisateur. Comme précédemment ce chapitre se termine par une discussion.

Dans le Chapitre 4 nous concluons ce mémoire en revenant sur les avantages de notre proposition et nous présentons les perspectives de recherche associées.

Chapitre 2

Problématique et travaux antérieurs

Il est devenu aisé de collecter des données mais notre capacité à en extraire des informations à forte valeur ajoutée reste limitée. Pour répondre à ces opportunités, l'extraction de connaissances dans les bases de données (ECBD) est le domaine de recherche au sein duquel coopèrent statisticiens, spécialistes en bases de données et en intelligence artificielle, ou encore chercheurs en conception d'interfaces homme-machine. Parmi les techniques phares en ECBD, on trouve de nombreux travaux sur l'extraction de motifs et leurs usages. L'extraction de motifs pose des problèmes algorithmiques difficiles dans les volumes de données à traiter. Ces motifs sont souvent utilisés dans des approches descriptives (par exemple la construction de résumés des données).

Dans un premier temps, le problème de l'extraction de motifs séquentiels peut sembler proche de celui de l'extraction de règles d'association. Ce rapprochement s'avère cependant très fragile en raison d'un élément clé qui est propre à l'extraction de motifs séquentiels : la temporalité. Cette notion permet à la fois de distinguer à l'intérieur des enregistrements un ordre d'apparition mais aussi de regrouper certains éléments. En effet si les règles d'association s'appliquent à des données de type itemsets (et permettent l'extraction de règles intra-transaction), la recherche de motifs séquentiels s'applique à des données de type séquences d'itemsets (et permet donc l'extraction de règles inter-transactions).

Nous proposons dans ce chapitre de décrire la problématique de l'extraction de motifs séquentiels ainsi que des principaux travaux existants dans ce domaine. Nous proposons également une description des travaux liés aux

motifs séquentiels fermés dans la mesure où nos travaux tireront partis de ces approches. Nous avons vu dans le chapitre d'introduction que des premières approches existent pour préserver la vie privée dans le contexte des règles d'association ou plus particulièrement de la recherche d'itemsets. Au cours de ce chapitre nous reviendrons sur ces principales approches. Etant donné qu'il n'existe pas à l'heure actuelle d'approche de recherche de motifs permettant de garantir l'anonymat des utilisateurs, nous définirons la problématique associée.

Le Chapitre est organisé de la manière suivante. Dans la Section 2.1 nous proposons les définitions associées à la recherche de motifs séquentiels et présentons brièvement la problématique étudiée (une présentation plus précise est proposée dans le Chapitre suivant). Nous présentons dans la Section 2.2, les principales méthodes d'extraction de motifs en nous focalisant sur celles par niveaux et celles qui s'intéressent aux motifs clos. La Section 2.3 décrit les principaux travaux dans le domaine de la fouille de données qui permettent de garantir la préservation de la vie privée. Enfin nous concluons ce Chapitre par une discussion.

2.1 Définitions et problématiques

La problématique de l'extraction de motifs séquentiels peut être perçue comme une extension de celle de l'extraction de règles d'association. En effet la prise en compte de la temporalité dans les enregistrements à étudier permet une plus grande précision dans les résultats, mais implique aussi un plus grand nombre de calculs et de contraintes. Le problème de la recherche de séquences dans une base de données de transactions est présentée dans [AS95] de la façon suivante (nous gardons, au niveau des définitions, les concepts de clients et d'achats) :

Définition 1 *Une transaction constitue, pour un client C , l'ensemble des items achetés par C à une même date. Dans une base de données client, une transaction s'écrit sous la forme d'un ensemble : $id\text{-client}, id\text{-date}, itemset$. un itemset est un ensemble non vide d'items évalué à vrai noté $(i_1 i_2 \dots i_k)$. Une séquence est une liste ordonnée, non vide, d'itemsets notée $\langle s_1 s_2 \dots s_n \rangle$ où s_j est un itemset. une séquence de données est une séquences représentant les achats d'un client. soit T_1, T_2, \dots, T_n les transactions d'un client, ordonnées par dates d'achat croissantes et soit $itemset(T_i)$ l'ensemble des items correspondants à T_i , alors la séquence de données de ce client est $\langle itemset(T_1) itemset(T_2) \dots itemset(T_n) \rangle$*

Exemple 1 Soit C un client et $S = \langle (A) (D E) (H) \rangle$, la séquence de données représentant les achats de ce client. S peut être interprétée par "C a acheté l'item A, puis en même temps les items D et E et enfin l'item H".

Définition 2 Soit $s_1 = \langle a_1 a_2 \dots a_n \rangle$ et $s_2 = \langle b_1 b_2 \dots b_m \rangle$ deux séquences de données. s_1 est incluse dans s_2 ($s_1 \subseteq s_2$) si et seulement si il existe $i_1 < i_2 < \dots < i_n$ des entiers tels que $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \dots, a_n \subseteq b_{i_n}$.

Exemple 2 La séquence $s_1 = \langle (C) (D E) (H) \rangle$ est incluse dans la séquence $s_2 = \langle (G) (C H) (I) (D E F)(H) \rangle$ (i.e. $s_1 \subseteq s_2$) car $(C) \subseteq (C H)$, $(D E) \subseteq (D E F)$ et $(H) \subseteq (H)$. En revanche $\langle (C) (E) \rangle \not\subseteq \langle (C E) \rangle$ (et vice versa).

Définition 3 Un client supporte une séquence s si s est incluse dans la séquence de données de ce client. Le support d'une séquence s est calculé comme étant le pourcentage des clients qui supportent s . Soit minSupp le support minimum fixé par l'utilisateur. Une séquence qui vérifie le support minimum (i.e. dont le support est supérieur à minSupp) est une séquence fréquente.

Remarque 1 Une séquence de données n'est prise en compte qu'une seule fois pour calculer le support d'une séquence fréquente, i.e. il peut présenter plusieurs fois le même comportement, le processus de recherche de séquences considère qu'il produit ce comportement sans tenir compte du nombre de ses apparitions dans la séquence de données.

Les deux propriétés suivantes considèrent le cas des sous-ensembles par rapport aux calculs du support et de l'inclusion.

Propriété 1 Soit s_1 et s_2 deux séquences. Si $s_1 \subseteq s_2$ alors $\text{support}(s_1) \geq \text{support}(s_2)$.

Propriété 2 Soit s_1 une séquence non fréquente. Quelle que soit s_2 telle que $s_1 \subseteq s_2$, s_2 est une séquence non fréquente.

La propriété 1 se justifie par le fait que toute séquence de données d dans DB supportant s_2 supporte obligatoirement s_1 (l'inverse ne se vérifie pas). La propriété 2 est une conséquence de la propriété 1. En effet, d'après cette propriété, $\text{support}(B) \leq \text{support}(A) < \text{minSupp}$, donc B n'est pas fréquent.

A partir des définitions et des propriétés précédentes nous pouvons définir la problématique de la recherche de motifs séquentiels de la manière suivante :

Définition 4 Soit une base de données DB , l'ensemble L^{DB} des séquences fréquentes maximales (également notées motifs séquentiels) est constitué de toutes les séquences fréquentes telles que pour chaque s dans L^{DB} , s n'est incluse dans aucune autre séquence de L^{DB} . Le problème de la recherche de séquences maximales (sequential patterns dans [AS95]) consiste à trouver l'ensemble L^{DB} .

Dans le cadre de ce mémoire, nous étendons la problématique de la recherche de motifs à la prise en compte de la vie privée :

Définition 5 Soit une base de données DB . Soit L^{DB} l'ensemble des séquences fréquentes maximales. Le problème de la recherche de séquences maximales anonymes consiste à trouver l'ensemble L^{DB} tel qu'il n'est pas possible d'obtenir d'informations sur les clients de DB à partir de L^{DB} .

Dans les sections suivantes, nous présentons tout d'abord les approches de recherche de motifs et nous abordons les travaux actuels pour la préservation de la vie privée.

2.2 Principales méthodes d'extraction de motifs séquentiels

2.2.1 Méthodes basées sur Apriori (breadth-first)

La méthode GSP (Generalized Sequential Patterns) [SA96] a été l'une des premières propositions pour résoudre la problématique des motifs séquentiels (ce travail fait suite à [AS95]). Les auteurs, en définissant la problématique de l'extraction de motifs séquentiels, ont également proposé un algorithme reprenant les principes d'Apriori, conçu pour l'extraction de règles d'association. Cependant, les difficultés relatives à la prise en compte de la temporalité ont rapidement conduit à la mise en place d'une méthode de génération de candidats adaptée à ce contexte. Celle-ci maintient cependant les principes d'une recherche "en largeur d'abord" puisque les candidats sont générés en fonction de leur longueur et non de leur préfixe.

Algorithme pionnier : GSP et sa structure

Dans [AS95] nous trouvons un résumé des techniques mises en œuvre depuis le début du projet Quest d'IBM. Ce projet est à l'origine de l'algorithme GSP [SA96], extension de Apriori, lui-même destiné à reprendre l'algorithme AIS présenté dans [AIS93].

GSP est un algorithme basé sur la méthode générer-élaguer mise en place depuis Apriori et destinée à effectuer un nombre de passes raisonnable sur la base de données. La technique généralement utilisée par les algorithmes de recherche de séquences est basée sur une création de candidats, suivie du test de ces candidats pour confirmer leur fréquence dans la base. Bénéficiant de propriétés relatives aux séquences et à leur fréquence d'apparition, ces techniques sont tout de même contraintes "d'essayer" des séquences avant de les déterminer fréquentes (ou non).

Pour évaluer le support de chaque candidat en fonction d'une séquence de données, GSP utilise 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 candidates, 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.

2.2.2 Méthodes basées sur le principe depth-first

La prise en compte de la temporalité dans les transactions a conduit de nombreux auteurs à privilégier des méthodes de recherche dites "en profondeur d'abord" pour extraire les motifs séquentiels. C'est le cas de PSP [MCP98], qui met en place et exploite un arbre de préfixes pour gérer les candidats. C'est aussi le principe adopté par [HPMa⁺00] avec FREESPAN et amélioré par [PHMa⁺01] avec l'algorithme PREFIXSPAN. PREFIXSPAN implémente de plus un principe de re-écriture de la base de données en fonction des préfixes des motifs séquentiels fréquents découverts (ou d'une indexation en fonction de la mémoire disponible). Nous présentons dans cette section une sélection de méthodes basées sur ce principe de la recherche "en profondeur d'abord".

PSP

Les auteurs de [MCP98] estiment que l'arbre de hachage utilisé dans [AS95, SA96] présente un défaut qu'il est facile de constater. En effet lors de la recherche des feuilles susceptibles de contenir des candidats inclus dans la séquence analysée, la structure utilisée ne tient pas compte des changements de date entre les items de la séquence qui servent à la navigation. Par exemple, avec la séquence $\langle (A\ C) (B\ D) \rangle$, l'algorithme va atteindre la feuille

du sommet C (fils de A), alors que cette feuille peut contenir deux types de candidats :

- ceux qui commencent par $\langle (A) (C) \dots$ d'un côté
- et ceux qui commencent par $\langle (A C) \dots$ de l'autre.

Le but est alors de mettre en place une structure d'arbre de préfixes, pour gérer les candidats. L'algorithme PSP (Prefix Tree for Sequential Pattern), destiné à exploiter cette structure, est basé sur la méthode générer-élaguer. Le principe de base de cette structure consiste à factoriser les séquences candidates en fonction de leur préfixe. Cette factorisation, inspirée de celle mise en place dans [AS95], pousse un peu plus loin l'exploitation des préfixes communs que présentent les candidats. En effet les auteurs proposent de prendre en compte les changements d'itemsets dans cette factorisation. L'arbre de préfixes ainsi proposé ne stocke plus les candidats dans les feuilles, mais permet de retrouver les candidats de la façon suivante : tout chemin de la racine à une feuille représente un candidat et tout candidat est représenté par un chemin de la racine à une feuille. De plus, pour prendre en compte le changement d'itemset, l'arbre est doté de deux types de branches. Le premier type, entre deux items, signifie que les items sont dans le même itemset alors que le second signifie qu'il y a un changement d'itemset entre ces deux items.

PREFIXSPAN

Dans [HPMa⁺00], les auteurs proposent l'algorithme FREESPAN (*Frequent pattern projected Sequential pattern mining*). L'idée générale est de proposer des projections récursives de la base de données en fonction des items fréquents. La base est alors projetée en plusieurs bases plus petites et les séquences fréquentes grandissent avec le nombre de projections. Les temps de réponses sont alors améliorés car chaque base projetée est plus petite et facile à traiter. Ce travail est le point de départ d'autres études sur la projection de bases de données en recherche de motifs séquentiels. FREESPAN présente tout de même un défaut selon ses auteurs : une sous-séquence peut être générée par n'importe quelle combinaison dans une séquence, donc FREESPAN doit conserver la totalité de la séquence dans la base d'origine sans réduire sa taille.

La méthode PREFIXSPAN, présentée dans [PHMa⁺01], se base sur une étude du nombre de candidats qu'un algorithme de recherche de motifs séquentiels peut avoir à produire afin de déterminer les séquences fréquentes. L'objectif des auteurs est alors de réduire le nombre de candidats générés. Pour parvenir à cet objectif, PREFIXSPAN propose (à l'instar de PSP avec

les candidats) d'analyser les préfixes communs que présentent les séquences de données de la base à traiter. À 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 cherche à faire croître la taille des motifs séquentiels découverts en appliquant la même méthode de manière récursive.

Deux sortes de projections sont alors mises en place pour réaliser cette méthode : la projection dite "niveau par niveau" et la "bi-projection". Au final, les auteurs proposent une méthode d'indexation permettant de considérer plusieurs bases virtuelles à partir d'une seule, dans le cas où les bases générées ne pourraient être maintenues en mémoire en raison de leurs tailles.

2.2.3 Recherche des motifs séquentiels fermés

L'extraction de motifs séquentiels devient problématique selon la longueur des motifs séquentiels extraits. Les auteurs de [YHA03] illustrent ce problème avec l'exemple d'une base de données ne contenant qu'un seul motif : $\langle (a_1)(a_2)\dots(a_{100}) \rangle$. Dans ce cas, il faudra générer $2^{100} - 1$ sous-séquences fréquentes avec un support minimum de 1. Ces sous-séquences seront redondantes car elles auront toutes le même support que $\langle (a_1)(a_2)\dots(a_{100}) \rangle$. Dans [YHA03], les auteurs définissent donc la problématique de la recherche des motifs séquentiels fermés (closed sequential patterns), inspirée de la recherche d'itemsets fermés. Ils proposent CLOSPAN, le premier algorithme capable de résoudre ce problème et optimisé pour cela. Dans [WH04], l'approche BIDE utilise une nouvelle manière d'étendre les séquences et optimise l'espace de recherche en analysant à l'avance les motifs à étendre.

Définition 6 Soit $minSup$, le support minimum et FS l'ensemble des motifs séquentiels fréquents correspondants. L'ensemble des motifs séquentiels fermés CS est défini comme :

$$CS = \{s/s \in FS \text{ et } \nexists s' \text{ telle que } s' \subset s \text{ et } support(s') = support(s)\}.$$

CLOSPAN

CLOSPAN [YHA03] est une méthode basée sur le principe depth-first et implémente l'algorithme PREFIXSPAN. En fait, il s'agit d'une optimisation de ce dernier, destinée à élaguer l'espace de recherche en évitant de parcourir certaines branches dans le processus de divisions récursives (en détectant par avance les motifs séquentiels non fermés). Le principe de CLOSPAN repose sur deux éléments essentiels : l'ordre lexicographique des séquences et la

détection de liens systématiques entre deux items (i.e. " β apparaît toujours avant γ dans la base de données").

BIDE

Etant donné que CLOSPAN conserve l'historique des séquences candidates, il ne s'avère pas efficace dans le cas de bases contenant de trop nombreuses séquences fermées. Pour pallier ce problème, une nouvelle approche, BIDE (BI-Directional Extension) est proposée dans [WH04]. L'idée générale est d'étendre les séquences dans les deux directions, i.e. en avant (forward extension) et en arrière (backward extension). En effet, considérons une séquence $S = i_1 i_2 \dots i_n$, celle-ci peut être étendue de trois manières possibles : ajout d'un item après i_n , ajout d'un item entre $i_1 i_2 \dots i_n$, ajout d'un item avant i_1 . La première correspond à une extension en avant et les deux dernières à une extension en arrière. Ainsi, les auteurs montrent que pour une séquence S , s'il n'y a pas d'extension avant ni d'extension arrière alors S est une séquence fermée. Comme dans CLOSPAN, une base projetée est constituée. Pour une séquence S , son ensemble d'items extensibles en avant, i.e. les items qui peuvent être ajoutés à la fin de S , est constitué par les items locaux dont le support est égal à celui de la séquence. Ces items locaux sont simplement trouvés en parcourant la base projetée pour ce préfixe et en comptant le nombre d'items. Pour effectuer rapidement cette opération, la projection utilisée est une pseudo projection comme dans [PHW02]. De manière à définir les extensions possibles en arrière, il faut dans un premier temps rechercher, pour les items d'une séquence, quelles sont les extensions en arrière possibles. Pour cela, il est nécessaire de remonter dans la séquence pour examiner avec quel item il est possible de l'étendre [WH04].

2.3 Préservation de la vie privée et fouille de données

L'une des préoccupations actuelle dans le domaine de la fouille de données est d'assurer la préservation de la vie privée, i.e. de garantir quel que soit le traitement effectué sur les données qu'il est impossible de retrouver des informations concernant un individu si celui-ci en a fait la demande. Concernant directement l'étape de fouille de données, cela consiste en fait à considérer les deux problèmes suivants :

- Comment réunir différentes informations tout en garantissant l'anonymat lors des différentes étapes de mise en commun de plusieurs sources

d'information ?

- Comment garantir que les motifs obtenus ne permettent pas d'identifier un comportement individuel ?

En ce qui concerne le premier point, de récents travaux ont été proposés dans un contexte collaboratif. Ces derniers redéfinissent des algorithmes d'extraction de motifs en considérant par exemple des approches d'anonymisation lors de la manipulation des données. Par contre, en ce qui concerne les motifs, il n'existe aucune proposition concernant le second point, i.e. garantir, à partir des résultats obtenus d'algorithmes traditionnels, que les connaissances acquises respectent la vie privée. Les seuls travaux, proches de notre problématique, ont été développés par le laboratoire KDD de l'Université de Pise, en définissant la notion de k -Anonymous Patterns [ABGP05] mais sont limités aux itemsets, i.e. à un sous ensemble des motifs séquentiels .

2.3.1 Préservation de la vie privée et motifs séquentiels dans un contexte collaboratif

Dans [ZMC04], les auteurs considèrent le problème de l'extraction de motifs séquentiels de différentes base de données stockées dans plusieurs lieux et nécessitant une collaboration entre les différentes bases (Privacy-Preserving Collaborative Sequential Pattern Mining). Ils proposent une approche pour extraire les motifs à l'aide de jointures et garantissent que les données privées des différentes bases sont préservées.

L'idée générale est d'effectuer un "mapping" (transformation) : à chaque item de la base on associe un entier puis, pour chaque base, on crée une matrice binaire où les colonnes sont les items transformés et les lignes sont les clients ($C_{ij} = 1$ correspond au fait que le client i supporte l'item j). Ensuite l'étape de fouille débute à l'aide de ces données transformées.

L'algorithme utilise une approche assez classique d'extraction de motifs et fonctionne de la manière suivante :

- L'ensemble des 1-séquences fréquentes est tout d'abord extrait ;
- Les séquences candidates sont générés et testés sur les bases de données ;
- Pour garantir l'anonymat, un protocole sécurisé est appliqué lors de

chaque vérification de séquences sur les bases de données.

Dans [KPTT06], les auteurs généralisent l'approche en considérant qu'il n'existe aucune contrainte entre les différents items des différentes bases et utilisent un protocole basé sur des opérateurs logiques. L'originalité de l'approche réside dans le fait que trois parties sont considérées : 2 parties permettent d'anonymiser les données et 1 partie effectue les différents calculs (support, fréquence, ...). A chaque étape du processus d'extraction, les données sont anonymisées et aucune des parties ne peut connaître les informations contenues par les autres. En outre, le protocole défini garantit que même si deux parties communiquent, elles ne peuvent pas obtenir le résultat final.

2.3.2 Les motifs k -Anonymes

Dans [ABGP05], les auteurs considèrent le problème de la manière suivante : ils recherchent dans les connaissances extraites quelles sont celles qui peuvent violer l'anonymat. Leur proposition se résume à des itemsets fréquents. Pour cela, la notion de k -anonymous patterns est introduite et définie de la manière suivante : on analyse l'ensemble d'itemsets fréquents pour savoir lesquels forment une menace pour l'anonymat de l'individu, l'idée principale est de tester tous les supports des motifs inférés d'un itemset fréquent, si le support d'un tel motif est inférieur à un seuil d'anonymat k alors ce motif est dit non k -anonyme. Un motif non k -anonyme est donc un itemset qui ne respecte pas l'anonymat. De manière à illustrer cette notion, considérons l'exemple suivant. Soit l'itemset fréquent $I = \{abc\}$ il est possible d'inférer de cet itemset le motif $p = a \wedge b \wedge \neg c$ si $Sup_D(p) \geq k$ ou $Sup_D(p) = 0$ alors ce motif est dit k -anonyme.

Les auteurs de [ABGP05] proposent un algorithme pour détecter les motifs non k -anonymes en se basant sur la notion des canaux d'inférence. Un canal d'inférence correspond à n'importe quelle sous structure de la collection d'itemsets (avec leurs supports respectifs) à partir de laquelle il est possible d'inférer des motifs non k -anonymes. Par exemple, soit un itemset X et un item 'a' qui n'appartient pas à X . Soit le super-set $X \cup \{a\}$ avec $0 < Sup_D(X) - Sup_D(X \cup \{a\}) < k$
 Dans ce cas la paire $\langle X, Sup_D(X); X \cup \{a\}, Sup_D(X \cup \{a\}) \rangle$ est un canal d'inférence pour le motif non k -anonyme $X \wedge \neg a$, dont le support est donné directement par $Sup_D(X) - Sup_D(X \cup \{a\})$

De manière à optimiser l'approche, les auteurs de [ABGP05] diminuent l'espace de recherche des canaux d'inférence en éliminant les canaux d'infé-

rence redondants. Deux canaux d'inférence sont dits redondants s'ils spécifient les mêmes transactions. L'optimisation est basée sur la notion d'itemsets fermés : un itemset est dit fermé s'il n'existe pas un sur-ensemble de cet itemset ayant le même support.

2.4 Discussion

Après l'engouement des travaux de recherche pour les règles d'association, les motifs séquentiels ont été très étudiés ces dernières années. Nous avons pu constater que de nombreux travaux de recherche ont été menés. Initialement, les premiers travaux ont consisté à améliorer les performances des propositions. Dans ce cadre, de nouvelles structures de données ou de nouvelles représentations des données sources ont été mises au point.

A notre connaissance, à part les deux approches que nous avons présenté, il n'existe pas de propositions d'algorithmes de motifs séquentiels qui préservent la vie privée. Ces deux approches partent de l'hypothèse que de nouveaux algorithmes d'extraction sont définis. Même s'ils répondent en partie à notre problème, ces travaux ne sont pas capables de répondre à la question suivante : à partir d'un ensemble de motifs extraits à partir d'un algorithme de recherche de motifs quelconque, est-il possible de garantir que les connaissances préservent la vie privée. Nous avons vu que l'approche proposée par [ABGP05], via les k -anonymous patterns, pouvait répondre en partie à notre problème. Cependant cette dernière souffre des limitations des itemsets par rapport aux motifs séquentiels, i.e. elle ne peut s'appliquer qu'à des séquences réduites à un seul itemset. Dans le cas d'applications réelles, cette contrainte est cependant trop restrictive et il devient indispensable de proposer une nouvelle approche généralisant la problématique. Dans le chapitre suivant, nous montrerons comment généraliser la notion d'itemsets k -anonymes à celle de motifs séquentiels k -anonymes.

Chapitre 3

Proposition

Notre objectif est de produire un modèle d'extraction valide (i.e. qui respecte la problématique des motifs séquentiels) sans révéler les données "privées". Pour cela, nous allons étudier quelles sont les caractéristiques de résultats d'une extraction qui menacent l'anonymat des individus. Dans notre contexte, nous considérons qu'aucune transformation ou "anonymisation" des données initiales n'est faite. Notre objectif est de considérer que nous partons de jeux de données réels sur lesquelles nous garantissons que les résultats respectent la vie privée, i.e. conservent l'anonymat des clients de la base.

Nous formalisons l'idée de k -anonyme pour les motifs séquentiels et décrivons les inférences qu'un adversaire peut exploiter pour rechercher les motifs séquentiels non k -anonymes.

Nous étudions les propriétés qui permettent d'identifier des motifs séquentiels dangereux cachés dans un ensemble des séquences fréquentes et nous proposons une manière simple et très efficace pour éliminer ces menaces.

Le chapitre est organisé de la manière suivante. Dans la section 3.1, nous étendons les définitions du chapitre précédent. La section 3.2 précise les problématiques étudiées en caractérisant les attaques considérées. Dans la section 3.3, nous étendons la notion de canal d'inférence aux motifs séquentiels. Nous décrivons tout d'abord un algorithme naïf d'extraction et nous l'étendons en utilisant la notion de motifs fermés. Ce dernier possède l'avantage de réduire considérablement l'espace de recherche et de minimiser les canaux redondants. Au cours de la section 3.4, nous ré-examinons les bases de données sources de manière à y insérer les données qui peuvent garantir que tous les résultats obtenus par une extraction préservent la vie privée. Enfin, nous

concluons ce chapitre dans la section 3.5 par une discussion.

3.1 Définitions préliminaires

Définition 7 Soit $I = \{i_1, \dots, i_p\}$ un ensemble d'items distincts. Nous avons vu précédemment qu'une transaction constitue, pour un client C , l'ensemble des items achetés par C à une même date. Dans une base de données client, une transaction peut également s'écrire sous la forme d'un ensemble : id -client, id -date, et un vecteur binaire p -dimensionnel enregistrant les valeurs d'items.

Plus exactement, une base de données peut être considérée de la manière suivante :

Définition 8 Une base de données binaire D se compose d'un ensemble de littéraux $I = \{i_1, \dots, i_p\}$ connus sous le nom d'items (articles) et multi-ensembles de transactions $T = \{T_1, \dots, T_m\}$.

D

T_{time}	C_{id}	a	b	c	d	e	f	g	h
10	1	0	0	1	1	0	0	0	0
10	3	1	1	0	0	0	1	0	0
10	4	0	0	0	1	0	0	1	1
15	1	1	1	1	0	0	0	0	0
15	2	1	1	0	0	0	1	0	0
15	5	0	0	0	1	0	0	0	0
15	6	0	1	0	0	0	1	0	0
15	7	0	0	0	1	0	0	0	0
20	1	1	1	0	0	0	1	0	0
20	2	0	0	0	0	1	0	0	0
20	4	0	1	0	0	0	1	0	0
20	5	0	1	0	0	0	0	0	0
20	6	1	0	0	0	0	0	0	0
20	7	0	1	0	0	0	1	0	0
25	1	1	0	1	1	0	1	0	0
25	4	1	0	0	0	0	0	1	1
30	5	1	0	0	0	0	0	0	0
30	7	1	0	0	0	0	0	0	0

TAB. 3.1 – Exemple d’une base de données binaire

C_1	$(cd)(abc)(abf)(acdf)$
C_2	$(abf)(e)$
C_3	(abf)
C_4	$(dgh)(bf)(agh)$
C_5	$(d)(b)(a)$
C_6	$(bf)(a)$
C_7	$(d)(bf)(a)$

TAB. 3.2 – Transformation en séquences des données des clients

Nous étendons à présent la notion de motifs à celle de motifs généralisés.

Définition 9 *Un motif séquentiel généralisé pour les littéraux de I (ensemble des items) est une liste de formules logiques. Chaque formule est obtenue en reliant des conditions sur la valeur de certains variables en utilisant les connecteurs logiques ET (\wedge) et OU (\vee). Le domaine de tous les motifs séquentiels généralisés extraits est appelé $Pat(I)$.*

Exemple 3 *Considérons l'exemple de la table 3.2 où $(dg)(bf)$ appartient à la séquence de données du client C_4 , $(d \vee g)(b \wedge \neg f)$ est un motif séquentiel généralisé.*

Dans ce contexte, la notion de support correspond au nombre de clients pour lesquels un motif est vrai :

Définition 10 *Soit une base de données D et un motif séquentiel généralisé P et un client $C \in D$. On dit $P(C)$ si C rend P vrai donc*
 $Sup_D(P) = |\{C \in D \mid P(C)\}|$

Si pour un motif séquentiel généralisé donné ce nombre est très petit (i.e. plus petit qu'un seuil k) mais pas nul alors ce motif représente une menace pour l'anonymat des individus qui rendent ce motif vrai. Le seuil k correspond au seuil d'anonymat :

Définition 11 *Soit une base de données D . Soit un seuil d'anonymat k . Un motif séquentiel généralisé P est dit k -anonyme si $Sup_D(P) \succ k$ ou $Sup_D(P) = 0$.*

Exemple 4 *Soit la base de données D , et $k = 2$, le motif séquentiel généralisé $(\neg d)(b \wedge f)$ est 2-anonyme car son support vaut 3, en fait ce motif est dans les séquences des données des clients C_2 , C_3 et C_6 . (Pour voir d'autres exemples des motifs séquentiels généralisés k -anonymes cf Annexe).*

La classe la plus étudiée pour les motifs séquentiels généralisé est la séquence d'itemsets.

Définition 12 *L'ensemble de toutes les séquences d'itemsets est la classe des motifs séquentiels qui se compose de toutes les séquences possibles des conjonctions de la forme : $s_1 \dots s_N$ avec $s_k = (i_m \wedge \dots \wedge i_n)$ k allant de 1 à N et $N \leq$ longueur de la plus longue séquence client.*

En donnant une base de données D et un seuil de support σ , le problème de fouille des séquences fréquentes demande de calculer $F(D, \sigma) = \{ \langle X, Sup_D(X) \rangle \mid X \in 2^{|Y_1 + \dots + Y_N| - 1} \ Y_i \in 2^I \wedge Sup_D(X) \geq \sigma \}$ ($N \leq$ longueur de la plus long séquence client). La séquence d'itemsets est habituellement notée sous la forme d'une liste d'ensemble de conjonction des items.

Exemple 5 $F(D, 3) = \{ \langle (a), 7 \rangle, \langle (b), 7 \rangle, \langle (d), 4 \rangle, \langle (f), 6 \rangle, \langle (ab), 3 \rangle, \langle (af), 3 \rangle, \langle (b)(a), 5 \rangle, \langle (bf), 6 \rangle, \langle (d)(a), 4 \rangle, \langle (d)(b), 4 \rangle, \langle (d)(f), 3 \rangle, \langle (f)(a), 4 \rangle, \langle (abf), 3 \rangle, \langle (bf)(a), 4 \rangle, \langle (d)(bf), 3 \rangle, \langle (d)(b)(a), 4 \rangle, \langle (d)(f)(a), 3 \rangle, \langle (d)(bf)(a), 3 \rangle \}$

3.2 Inférence du support et menace de l'anonymat

Notre objectif est d'extraire l'ensemble des séquences fréquentes mais au préalable, nous devons déterminer d'après les éléments de cet ensemble lesquels forment une menace pour l'anonymat des individus. Nous définissons formellement les genres d'attaques que nous considérons. Puisque nous proposons un ensemble de séquences fréquentes, les seules attaques possibles concernent celles dont les motifs sont non k -anonymes, i.e. tout motif p tel que $0 \prec sup_D(p) \prec k$.

Définition 13 *L'ensemble des séquences σ -fréquentes correspond à l'ensemble des séquences qui ont des supports $\geq \sigma$. L'ensemble des séquences σ -fréquentes S est compatible avec une base de données D si $S = F(D, \sigma)$.*

Nous étudions comment "assainir" un ensemble de motifs de manière à ce que le résultat produit soit toujours compatible avec au moins une base de données.

Le premier problème que nous considérons est la détection de la menace de l'anonymat dans le résultat d'une extraction des séquences fréquentes. Même s'il existe des travaux pour les itemsets (*cf* Chapitre précédent), le fait de travailler sur des motifs plutôt que sur des itemsets fait que les résultats préalables ne sont pas adaptés.

Exemple 6 *Dans cet exemple nous illustrons le fait que même si une séquence fréquente est composée d'itemsets fréquents et k -anonymes, la séquence n'est pas forcément k -anonyme.*

C_1	$(abc)(de)$
C_2	$(abc)(de)$
C_3	$(abc)(de)$
C_4	$(d)(abc)$
C_5	$(abc)(d)$

Soit $\sigma = 3$ et $k = 2$, nous avons les itemsets (abc) et (de) fréquents et 2-anonymes (cf Annexe). Cependant, la séquence fréquente $\langle (abc)(de) \rangle$ n'est pas 2-anonyme car nous avons $(abc)(d) \subseteq (abc)(de)$ et $\text{sup}((a \wedge b \wedge c)(d \wedge \neg e)) = 1 < 2$ (seulement le client C_5 supporte ce motif).

Problème 1 Soit une collection des séquences fréquentes $F(D, \sigma)$. Soit un seuil d'anonymat k . Le premier problème consiste à détecter tous les canaux d'inférence possibles qui existent dans $F(D, \sigma)$:

$$S \subseteq F(D, \sigma) : \exists P \in \text{Pat}(I) : S \models 0 < \text{Sup}_D(P) < k$$

Il s'agit de vérifier qu'il n'existe pas de canaux d'inférence dans cette collection fréquente. Dans ce cas le résultat est sans risque autrement les deux cas suivants sont à considérer :

1. Fouiller une autre collection des séquences fréquentes avec des conditions différentes. Par exemple un seuil du support plus élevé pour chercher une collection admissible.
2. Transformer la collection pour enlever les canaux d'inférence.

La 2^{ème} solution engendre un nouveau problème

Problème 2 Soit une collection de séquences fréquentes $F(D, \sigma)$ et soit l'ensemble de tous ses canaux d'inférence. Notre problème consiste à transformer $F(D, \sigma)$ en une collection de séquences fréquentes O_k qui peuvent être révélée sans risque.

O_k doit satisfaire les conditions suivantes :

1. $\nexists P \in \text{Pat}(I) : O_k \models 0 < \text{Sup}_D(P) < k$
2. $\exists D' : O_k = F(D', \sigma)$

Nous répondons à ces différentes problématiques dans les sections suivantes.

3.3 Les canaux d'inférence pour les motifs séquentiels

De manière intuitive, un simple canal d'inférence est donné par n'importe quelle séquence S avec sa super-séquence $\langle (S)(a) \rangle$ (S suivi par 'a' où 'a' est

un item de la base) tel que : $0 \prec \text{Sup}_D(S) - \text{Sup}_D((S)(a)) \prec k$
 Dans ce cas la paire $\langle S, \text{Sup}_D(S); (S)(a), \text{Sup}_D((S)(a)) \rangle$ est un canal d'inférence pour le motif séquentiel non kanonyme $(S)(\neg a)$ tel que son support est directement donné par .

De manière à illustrer cette notion, considérons l'exemple suivant.

Exemple 7 Soit $k= 2$

$$(d)(f) \in F(D, 3) \quad \text{Sup}_D((d)(f)) = 3$$

$$(d) \subseteq (d)(f) \quad \text{Sup}_D((d)) = 4$$

$$\text{Sup}_D((d)) - \text{Sup}_D((d)(f)) = 1 \prec 2$$

donc $\langle (d), 4; (d)(f), 3 \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non 2-anonyme $(d)(\neg f)$

En général si nous connaissons le support de la séquence $J = \langle a_1, a_2, \dots, a_n \rangle$, le support de la séquence $I = \langle b_1, b_2, \dots, b_m \rangle$ avec $(m \leq n)$ et les supports de toutes les séquences X telles que $I \subseteq X \subseteq J$ nous pouvons calculer le support du motif séquentiel généralisé $P = \langle p_1, p_2, \dots, p_n \rangle$

$$P_i = \begin{cases} \overline{b_j a_i / b_j} & \text{si } b_j \subseteq a_i \text{ avec } j \in \{1, \dots, m\} \\ \overline{a_i} & \text{sinon} \end{cases}$$

Exemple 8 Soit $J = \underbrace{(abe)}_{a_1} \underbrace{(abc)}_{a_2} \underbrace{(dce)}_{a_3} \underbrace{(bd)}_{a_4} \underbrace{(a)}_{a_5} \quad n = 5$

et $I = \underbrace{(ac)}_{b_1} \underbrace{(b)}_{b_2} \underbrace{(a)}_{b_3} \quad m = 3 \quad I \subseteq J$

donc le motif séquentiel associé à I et J est

$$P = \underbrace{(\neg a \wedge \neg b \wedge \neg e)}_{\overline{a_1}} \underbrace{(a \wedge \neg b \wedge c)}_{\overline{b_1 a_2 / b_1}} \underbrace{(\neg d \wedge \neg c \wedge \neg e)}_{\overline{a_3}} \underbrace{(b \wedge \neg d)}_{\overline{b_2 a_4 / b_2}} \underbrace{(a)}_{\overline{b_3 a_5 / b_3}}$$

En considérant le treillis associé à toutes les possibilités d'extension des séquences, nous obtenons le lemme suivant qui permet de caractériser le support de P :

Lemme 1 En donnant un motif séquentiel généralisé associé à deux séquence I et J ($I \subseteq J$) (P, I et J définis ci-dessus) la formule de calcul du support de P devient

$$\text{Sup}_D(P) = \sum_{I \subseteq X \subseteq J} (-1)^{|X/I|} \Delta_X \text{Sup}_D(X)$$

avec

$$\Delta_X = \begin{cases} 1 & \text{si } |X/I| = 1 \text{ ou } 0 \\ |\text{nombre des fils de } X \text{ contenant } I| - 1 & \text{sinon} \end{cases}$$

De manière à illustrer le calcul du support. Considérons l'exemple suivant :

Exemple 9 Soit $J = (ab)(ab)$ et $I = (a)(b)$. P est tel que $P = (a \wedge \neg b)(\neg a \wedge b)$
 $Sup_D(P) = \Delta Sup_D((a)(b)) - \Delta Sup_D((a)(ab)) - \Delta Sup_D((ab)(a)) +$
 $\Delta Sup_D((ab)(ab))$
 $= 1 * Sup_D((a)(b)) - 1 * Sup_D((a)(ab)) - 1 * Sup_D((ab)(a)) + (2 -$
 $1) Sup_D((ab)(ab)).$ (cf Annexe)

Nous pouvons maintenant définir de manière formelle les canaux d'inférences pour les motifs séquentiels :

Définition 14 Soient I et J deux séquences $I \subseteq J$ et P le motif séquentiel généralisé associé. Nous notons $P = C_I^J$ et nous avons $Sup_D(C_I^J) = f_I^J$. Si $0 \prec f_I^J \prec k$ alors C_I^J est appelé canal d'inférence.

Théorème 1 $\forall P \in Pat(I) : 0 \prec Sup_D(P) \prec k . \exists C_I^J : 0 \prec f_I^J(D) \prec k .$

Preuve Dans le cas d'un motif séquentiel conjonctif la preuve est la conséquence directe du Lemme 1.

Autrement, P peut s'écrire sous la forme de $P = P_1 \vee P_2 \vee \dots \vee P_n$ avec P_i est un motif séquentiel conjonctif, nous avons donc $Sup_D(P) \geq \max(Sup_D(P_i))$. $1 \leq i \leq n$. Si nous avons $Sup_D(P) \prec k$ alors $\forall i$ nous avons $Sup_D(P_i) \prec k$ et si $Sup_D(P) \succ 0$ alors $\exists i$ tel que $Sup_D(P_i) \succ 0$ alors $0 \prec Sup_D(P_i) \prec k$.

D'après le théorème 1, nous pouvons conclure que toutes les menaces de l'anonymat sont liées à un canal d'inférence de la forme C_I^J .

Algorithm 1: Naïve Inference Channel Detector

Data: $F(D, \sigma), k$

Result: $Ch(k, F(D, \sigma))$ ensemble des canaux d'inférence détectés

$Ch(k, F(D, \sigma)) = \phi;$

forall $\langle J, \text{sup}(J) \rangle \in F(D, \sigma)$ **do**

forall $I \subseteq J$ **do**

compute $f_I^J;$

if $0 \prec f_I^J \prec k$ **then**

insert $\langle C_I^J, f_I^J \rangle$ **in** $Ch(k, F(D, \sigma))$

L'exemple suivant illustre les canaux d'inférence obtenus pour la base de données :

Exemple 10 $Ch(2, F(D, 3)) = \left\{ \langle C_\phi^{(f)}, 1 \rangle, \langle C_{(b)}^{(bf)}, 1 \rangle, \langle C_{(d)}^{(d)(f)}, 1 \rangle, \langle C_{(b)}^{(abf)}, 1 \rangle, \langle C_{(b)(a)}^{(bf)(a)} \rangle, \langle C_{(d)(b)}^{(d)(bf)} \rangle, \langle C_{(b)(a)}^{(d)(b)(a)} \rangle, \langle C_{(d)(a)}^{(d)(f)(a)} \rangle, \langle C_{(f)(a)}^{(d)(f)(a)} \rangle, \langle C_{(bf)(a)}^{(d)(bf)(a)} \rangle, \langle C_{(d)(b)(a)}^{(d)(bf)(a)} \rangle \right\}.$
(Pour le calcul des supports cf Annexe).

Dans la partie suivante, nous considérons la propriété d'anti-monotonie. Cette dernière sera non seulement utile pour déterminer les canaux d'inférence mais sera également utilisée pour optimiser la proposition.

Propriété d'anti-monotonie

La propriété d'anti-monotonie, définie initialement pour les règles d'association correspond dans le cas des motifs séquentiels à :

Définition 15 Soient S_i et S_j deux séquences $S_i = \langle a_1, a_2, \dots, a_m \rangle$ et $S_j = \langle b_1, b_2, \dots, b_n \rangle$ avec $(m \leq n)$, on dit que $S_i \subseteq S_j$ s'il existe $i_1 \prec i_2 \prec \dots \prec i_m$ tel que $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \dots, a_m \subseteq b_{i_m}$

En d'autres termes, cette propriété permet de savoir que si un motif est fréquent, tous les sous motifs associés sont également fréquents. En considérant l'extension proposée pour les motifs séquentiels, nous obtenons :

Définition 16 Soient les deux motifs séquentiels conjonctifs C_I^J et C_H^L on dit que $C_I^J \leq C_H^L$ si $I \subseteq H$ et $(J/I) \subseteq L$.

Nous pouvons ainsi déduire directement la propriété suivante :

Proposition 1 $C_I^J \leq C_H^L \Rightarrow \forall D. f_I^J(D) \geq f_H^L(D)$.

Lors de la détection des canaux d'inférence si nous trouvons un motif séquentiel conjonctif C_I^J tel que $f_I^J \geq k$ (i.e. k -anonyme) alors la propriété d'anti-monotonie nous permet d'éviter de tester les supports de tous les motifs qui sont inclus dans C_I^J (dans l'ordre de \leq) car ils sont $\geq k$ (i.e. ces motifs sont k -anonymes).

A partir de l'algorithme 1, nous pouvons obtenir tous les canaux d'inférences. Cependant même si cette proposition est complète elle souffre du fait que les canaux d'inférences trouvés peuvent être redondants. Nous proposons donc dans le reste de cette section d'optimiser l'approche en n'extrayant que les canaux d'inférences non redondants.

Canaux d'inférence redondants

Dans l'exemple 10 nous pouvons constater qu'il existe des canaux d'inférence qui sont redondants. Considérons, par exemple les deux canaux d'inférence $\langle C_{(b)(a)}^{(d)(b)(a)}, 1 \rangle \leq \langle C_{(bf)(a)}^{(d)(bf)(a)}, 1 \rangle$. Ces deux motifs séquentiels non k -anonymes associés identifient seulement le client C_6 même si l'un est plus spécifique que l'autre et sont donc redondants. Il est facile de voir qu'il existe d'autres canaux redondants dans $Ch(3, F(D, 2))$ (cf Annexe).

Notre problème consiste maintenant non seulement à rechercher les canaux d'inférence mais également de n'obtenir que ceux qui sont non redondants. Dans la suite, nous considérons la théorie des séquences fermées pour nous aider à résoudre ce problème.

Définition 17 Soit σ , le support minimum et $F(D, \sigma)$ l'ensemble des séquences fréquentes correspondants. L'ensemble des séquences fermées $Cl(D, \sigma)$ est défini comme :

$$Cl(D, \sigma) = \{s/s \in F(D, \sigma) \text{ et } \nexists s' \text{ telle que } s' \subset s \text{ et } support(s') = support(s)\}.$$

En fait, les séquences fermées sont une représentation concise de toutes les séquences fréquentes, i.e. elles contiennent les mêmes informations sans redondance et groupent ensemble les séquences qui identifient le même groupe de clients (cf Chapitre précédent).

Définition 18 Soit C un ensemble des clients appartenant à une base de données D et s une séquence d'itemsets. On donne la fonction $f(C)$ qui retourne la plus large séquence d'itemsets commune entre tous les éléments de

C , la fonction $g(s)$ qui retourne l'ensemble des clients qui supportent s , et la fonction composée $f \circ g$ appelée opérateur de Galois. Ainsi, une séquence s est dite fermée par rapport à l'ensemble des clients de D si et seulement si $T(s) = f \circ g(s) = f(g(s)) = s$.

Définition 19 Soit une base de données D et un seuil minimum du support σ , le problème de fouille des séquences fréquentes fermées exige de calculer : $Cl(D, \sigma) = \{ \langle X, Sup_D(x) \rangle \in F(D, \sigma) / X = T(X) \}$.

une séquence s est dite fréquente maximal si et seulement si elle est fréquente fermée et $\nexists s' \supset s$ avec $s' \in Cl(D, \sigma)$.

Exemple 11 Dans la base de données D , nous avons

$Cl(D, 3) = \{ \langle (a), 7 \rangle, \langle (b), 7 \rangle, \langle (bf), 6 \rangle, \langle (b)(a), 5 \rangle, \langle (abf), 3 \rangle, \langle (bf)(a), 4 \rangle, \langle (d)(b)(a), 4 \rangle, \langle (d)(bf)(a), 3 \rangle \}$.

dans cette situation (abf) et $(d)(bf)(a)$ sont les séquences fréquentes maximales.

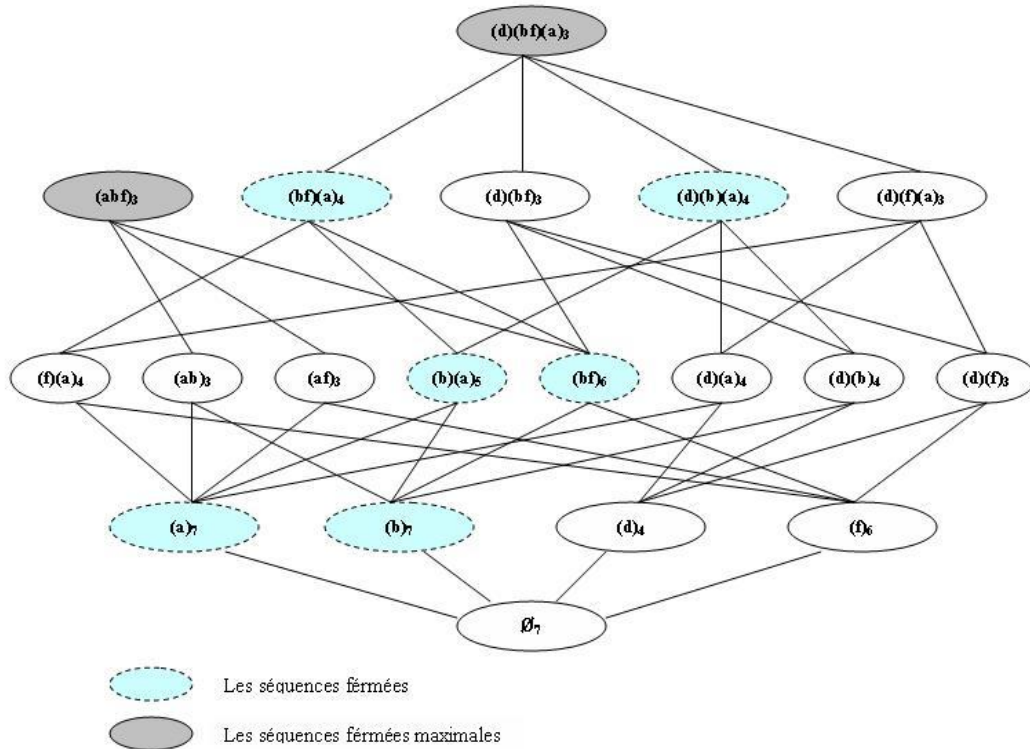


FIG. 3.1 – Le treillis de séquences 3-fréquentes

L'approche que nous proposons est d'utiliser le même concept pour éliminer les canaux d'inférence redondants et ne retenir ainsi que les canaux pertinents. En ne retenant que ces canaux pertinents, notre objectif est de réduire le nombre de contrôle requis pour vérifier les violations d'anonymat.

Définition 20 Soit l'ensemble de toutes les séquences fréquentes fermées $Cl(D, \sigma)$, nous définissons :

$$MCh(k, Cl(D, \sigma)) = \{ \langle C_I^J, f_I^J(D) \rangle \mid I \in Cl(D, \sigma), J \text{ maximal} \}$$

Comme l'ensemble de toutes les séquences fréquentes fermées $Cl(D, \sigma)$ contient toutes les informations de $F(D, \sigma)$ dans une représentation plus compacte nous savons que l'ensemble $MCh(k, Cl(D, \sigma))$ représente, sans redondance, tous les canaux d'inférence de $F(D, \sigma)$. Ceci signifie que pour détecter tous les canaux d'inférence de $F(D, \sigma)$, nous pouvons uniquement considérer les canaux d'inférence dans $MCh(k, Cl(D, \sigma))$ et ainsi exécuter un nombre réduit d'opérations.

L'exemple suivant illustre le nombre de canaux obtenus en ne considérant que les fermés. En fait à partir de tous les fermés, nous sommes capables d'extrapoler la valeur de support minimale de toutes les sous séquences incluses et donc nous évitons de régénérer tout le treillis.

Exemple 12 Dans la base de données D nous avons :

$$MCh(2, Cl(D, 3)) = \left\{ \left\langle C_{(b)}^{(abf)}, 1 \right\rangle, \left\langle C_{(bf)(a)}^{(d)(bf)(a)}, 1 \right\rangle, \left\langle C_{(d)(b)(a)}^{(d)(bf)(a)}, 1 \right\rangle \right\}$$

L'algorithme 2 représente une manière optimisée d'identifier toutes les menaces à l'anonymat. A partir de tous les fermés, nous retenons les plus grands, i.e. ceux qui sont maximaux. Il suffit alors de calculer les f_I^J correspondant comme nous l'avons vu dans l'approche initiale. L'avantage de l'approche est que les J correspondent aux fermés maximaux et les I correspondent aux fermés. Il est aisé de voir que le nombre d'opérations effectuées est nettement inférieur au nombre d'opérations à réaliser dans le cas d'un treillis complet.

Algorithm 2: Optimized Inference Channel Detector

Data: $Cl(D, \sigma), k$

Result: $MCh(k, Cl(D, \sigma))$

$M = \{I \in Cl(D, \sigma) / I \text{ est maximale}\};$

$MCh(k, F(D, \sigma)) = \phi;$

forall $J \in M$ **do**

forall $I \in Cl(D, \sigma)$ **such that** $I \subseteq J$ **do**

compute $f_I^J;$

if $0 \prec f_I^J \prec k$ **then**

insert $\langle C_I^J, f_I^J \rangle$ **in** $MCh(k, Cl(D, \sigma))$

3.4 Blocage des canaux d'inférence

Dans cette section nous étudions comment bloquer les menaces de violation d'anonymat décrites dans la section précédente. La première approche, assez naïve, consiste simplement à éliminer des résultats les paires de séquences I, J tels que C_I^J est un canal d'inférence. Malheureusement, cette approche produit un résultat qui n'est pas (en général) compatible avec les bases de données manipulées et surtout qui n'offre pas tous les résultats.

De manière à éviter ce problème, nous proposons d'"aseptiser" les séquences fréquentes par rapport aux canaux d'inférence détectés par l'algorithme 2. L'idée que nous proposons peut être résumée de la manière suivante : pour tous les canaux d'inférence C_I^J , nous incrémentons le support de la séquences I par k pour imposer $f_I^J \succ k$. Bien entendu, pour maintenir la compatibilité par rapport aux données de la base, nous incrémentons les supports de tous les sous-séquences de I en conséquence.

Proposition 2 *Soit S un ensemble de séquences σ -fréquentes compatible avec au moins une base de données D donc en incrémentant par k le support d'une séquence $I \in S$ et de chaque sous-séquence de I , nous obtenons un autre ensemble de séquences σ -fréquentes qui est compatible avec une base de données D' , obtenue en ajoutant à D k clients supportant seulement I .*

Minimiser le nombre d'insertions de tuples

Bien que notre but principal soit de cacher chaque canal d'inférence, nous voulons également minimiser le nombre de tuples à insérer. L'idée principale consiste en fait à exploiter la propriété d'anti-monotonie des motifs.

Définition 21 Soit $I = s_1, s_2, \dots, s_n$ et $J = t_1, t_2, \dots, t_m$ sont deux séquences d'itemsets avec $m \leq n$ on définit la séquence

$$I \cup J = s_1 \cup t_1, s_2 \cup t_2, \dots, s_m \cup t_m, s_{m+1}, \dots, s_n .$$

Remarque : Soient deux motifs différents $C_{(a)(b)(f)}^{(ac)(bd)(f)}$ et $C_{(b)(ae)}^{(be)(ace)}$. Nous obtenons en effectuant une jointure $C_{(ab)(abe)(f)}^{(abce)(abcde)(f)}$. Généralement deux motifs C_I^J et $C_{I'}^{J'}$ peuvent être regroupés dans $C_{I''}^{J''}$ si et seulement s'il existe un $C_{I''}^{J''}$ tel que $C_{I''}^{J''} \supseteq C_I^J$ et $C_{I''}^{J''} \supseteq C_{I'}^{J'}$, avec $I \cap (J'/I') = \phi$ et $I' \cap (J/I) = \phi$. Dans ce cas-ci nous pouvons avoir $J'' = J \cup J'$ et $I'' = I \cup I'$. Dans l'algorithme 3, $smax$ dénote l'ensemble de telles super-séquences maximales obtenues par jointures.

Algorithm 3: Inference Channel Sanitization

Data: $F(D, \sigma), k$

Result: O^k

- 1 : compute $Cl(D, \sigma)$ from $F(D, \sigma)$;
 - 2 : compute $MCh(k, Cl(D, \sigma))$ from $Cl(D, \sigma)$;
 - 3 : $smax = \phi$;
 - 4 : **forall** $\langle C_I^M, f_I^M \rangle \in MCh(k, Cl(D, \sigma))$ **do**
 - 5 : **if** $\exists \langle C_A^B, f_A^B \rangle \in smax$ **such that** $A \cap (M/I) = \phi$ **et**
 $I \cap (B/A) = \phi$ **then**
 - 6 : $smax = smax / \{ \langle C_A^B, f_A^B \rangle \}$;
 - 7 : $smax = smax \cup \{ \langle C_{I \cup A}^{M \cup B}, f_{I \cup A}^{M \cup B} \rangle \}$;
 - 8 : **else**
 - 9 : $smax = smax \cup \{ \langle C_I^M, f_I^M \rangle \}$;
 - 10 : **forall** $\langle X, sup(X) \rangle \in F(D, \sigma)$ **such that** $X \subseteq I$ **do**
 - 11 : $sup^k(I) = sup(I)$;
 - 12 : **forall** $\langle C_I^J, f_I^J \rangle \in smax$ **do**
 - 13 : $sup^k(I) = sup^k(I) + k$;
 - 14 : $O^k = O^k \cup \{ \langle I, sup^k(I) \rangle \}$;
-

De la ligne 3 à 9 l'algorithme 3 calcule l'ensemble des canaux d'inférence $smax$ de $MCh(k, Cl(D, \sigma))$ en exploitant la remarque précédente. Comme $|smax| \leq |MCh(k, Cl(D, \sigma))|$, nous pouvons ainsi réduire le nombre total d'insertions. De la ligne 10 à 14 l'algorithme incrémente le support des séquences qui sont dans des canaux d'inférence de $smax$.

Exemple 13 Dans la base de données D en incrémentant les supports de (b) , $(bf)(a)$ et $(d)(b)(a)$ et ses sous-séquences par 2 nous enlevons tous les canaux d'inférence qui sont dans l'ensemble des séquences fréquentes de D

3.5 Discussion

Dans ce chapitre nous avons étendu la notion de motifs k -anonymes aux motifs séquentiels. Pour cela, nous avons tout d'abord présenté une approche naïve qui utilise le treillis dans son intégralité. Cette approche est étendue en considérant les motifs séquentiels fermés. L'avantage dans ce cas est limiter l'espace de recherche et ainsi de n'obtenir que des canaux d'inférences non redondants.

L'approche proposée généralise les premières propositions sur les itemsets. En effet, les itemsets représentant un sous ensemble des motifs séquentiels, en montrant qu'il était possible de rechercher des canaux sur des motifs nous sommes également à même de considérer des motifs réduits à un seul itemset. En étendant les canaux, nous avons également été confrontés à la redondance. Si nous examinons plus attentivement le premier algorithme, il est clair que nous sommes obligés de considérer tout le treillis qui peut être généré (de la même manière que les algorithmes par niveaux). La propriété d'anti-monotonie a été utilisée dans des travaux précédents pour optimiser l'extraction, notamment en minimisant l'espace de recherche. Dans le contexte des canaux, nous souffrons des mêmes problèmes : comment optimiser l'extraction ? Nous avons pu montrer que cette propriété pouvait être généralisée également dans notre contexte. En couplant cette approche à celle des motifs fermés nous sommes alors à même de produire des résultats sans aucune redondance. Le fait d'utiliser la notion de fermé offre également l'avantage de pouvoir utiliser des approches existantes (*cf* Chapitre précédent) pour extraire les canaux sans beaucoup de modification (nous reviendrons sur cet aspect lors de la conclusion générale).

En proposant d'augmenter les supports (et encore à l'aide de la propriété d'anti-monotonie), nous avons également montré que nous étions capables d'extraire tous les canaux d'inférence pertinents.

Chapitre 4

Conclusion

Dans ce mémoire, nous avons considéré la préservation de la vie privée et la fouille de données. Les nouveaux projets comme HIPAA ont eu des conséquences très fortes dans le cadre de la modélisation, l'accès, la manipulation des données. Même si, initialement, les conséquences sur les approches d'extraction n'étaient pas évidentes, il faut tout de même constater qu'elles existent et sont de plus en plus importantes. Tout au long de ce mémoire nous avons vu qu'il était aisé à partir des connaissances extraites d'obtenir des informations sur les individus. Toutefois, les propositions initiales souffrent du fait que pour préserver la vie privée, il est indispensable de reconcevoir de nouveaux algorithmes. Nous avons vu que des approches commençaient à exister et que surtout la notion de k -anonymous patterns offrait l'avantage de ne pas directement reconsidérer les algorithmes d'extraction mais plutôt utilisaient les résultats de l'extraction. Cependant les motifs k -anonymes proposés se limitent aux itemsets et donc sont peu adaptés à de nombreux domaines d'applications. En étendant ce problème aux motifs séquentiels, nous offrons de nouveaux types de connaissances. Le fait d'utiliser les résultats des algorithmes existants pour extraire à partir des motifs séquentiels les canaux d'inférence offre la possibilité de rapidement adapter les travaux existant à la préservation de la vie privée.

4.1 Perspectives

La première perspective de ce travail est de tester sur des jeux de données réelles notre approche. En effet, il existe, comme nous l'avons décrit, des algorithmes d'extraction de motifs fermés et il est maintenant indispensable de les coupler à notre approche afin de proposer une approche uniforme. Au cours de ce stage, ce travail n'a pas pu être effectué dans la mesure où

les travaux sur les motifs séquentiels fermés sont assez limités et surtout les implémentations existantes ne considèrent pas les séquences comme une succession d'itemsets. En effet parmi les travaux que nous avons eu l'occasion d'examiner, les séquences sont réduites à des successions d'items plutôt que d'itemsets. En étendant les propositions actuelles aux itemsets, nous disposerons d'algorithmes sur lesquels il sera aisé de coupler notre approche.

Au cours de nos travaux, nous avons vu que pour optimiser l'approche, la notion de fermé était intéressante. Toutefois, à l'heure actuelle, de nouvelles approches ont été définies pour les itemsets et il serait intéressant d'examiner l'impact sur les motifs séquentiel. Par exemple, les nouvelles approches basées sur les free ou sur les dérivables offrent l'opportunité de réduire considérablement le treillis manipulés et donc minimise l'espace de recherche. L'un des objectif serait de considérer comment notre approche peut tirer profit de telles propositions.

Bibliographie

- [ABGP05] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi. K-anonymous patterns. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 05)*, Porto, Portugal, 2005.
- [AIS93] R. Agrawal, T. Imielinski, and 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.
- [AS95] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proc. of ICDE'95*, 1995.
- [HPMa⁺00] J. Han, J. Pei, B. Mortazavi-asl, Q. Chen, U. Dayal, and M. Hsu. Freespan : Frequent pattern-projected sequential pattern mining. In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (KDD 00)*, pages 355–359, Boston, USA, 2000.
- [KPTT06] V. Kapoor, P. Poncelet, F. Trouset, and M. Teisseire. Privacy-preserving sequential pattern mining in distributed databases. In *Technical Report LGI2P Research Center*, 2006.
- [MCP98] F. Maseglia, F. Cathala, and P. Poncelet. The PSP approach for mining sequential patterns. In *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD 98)*, pages 176–184, Nantes, France, 1998.
- [PHMa⁺01] J. Pei, J. Han, B. Mortazavi-asl, H. Pinto, Q. Chen, and 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.
- [PHW02] J. Pei, J. Han, and W. Wang. Mining sequential patterns with constraints in large databases. In *Proceedings of the 10th In-*

- ternational Conference on Information and Knowledge Management (CIKM 02)*, pages 18–25, MCLean, USA, 2002.
- [SA96] R. Srikant and R. Agrawal. Mining sequential patterns : Generalizations and performance improvements. In *Proc. of EDBT'96*, 1996.
- [WH04] J. Wang and J. Han. Bide : Efficient mining of frequent closed sequences. In *Proceedings of the International Conference on Data Engineering (ICEDE 04)*, Boston, M.A., 2004.
- [YHA03] X. Yan, J. Han, and R. Afshar. Clospan : Mining closed sequential patterns in large databases. In *Proceedings of the SDM 03 Conference*, San Francisco, CA, 2003.
- [ZMC04] J. Z. Zhan, S. Matwin, and L. Chang. Privacy-preserving collaborative sequential pattern mining. In *Proceedings of Workshop on Link Analysis, Counter-terrorism and Privacy*, 2004.

Annexe

Itemsets k -anonymes

C_1	$(abc)(de)$
C_2	$(abc)(de)$
C_3	$(abc)(de)$
C_4	$(d)(abc)$
C_5	$(abc)(d)$

Dans la base de données ci-dessus (avec $\sigma = 3$ et $k = 2$) nous avons les itemsets (abc) et (de) sont fréquents mais encore sont \mathcal{L} -anonymes car tous les motifs inférés de ces itemsets ont un support nul ou bien ≥ 2 .

En fait pour l'itemset (abc) les motifs inférés de lui sont :
 $(\neg a \wedge \neg b \wedge \neg c); (a \wedge \neg b \wedge \neg c); (\neg a \wedge b \wedge \neg c); (\neg a \wedge \neg b \wedge c); (a \wedge b \wedge \neg c); (a \wedge \neg b \wedge c); (\neg a \wedge b \wedge c); (a \wedge b \wedge c)$
le motif $(a \wedge b \wedge c)$ a un support 5, et tous les autres ont un support nul.

Pour l'itemset (de) les motifs inférés de lui sont :
 $(\neg d \wedge \neg e); (d \wedge \neg e); (\neg d \wedge e); (d \wedge e)$
le motif $(d \wedge e)$ a un support 3 (i.e. C_1, C_2 et C_3), et $(d \wedge \neg e)$ a un support 2 (i.e. C_4 et C_5), et les autres ont un support nul.

Calcul des supports des canaux d'inférence

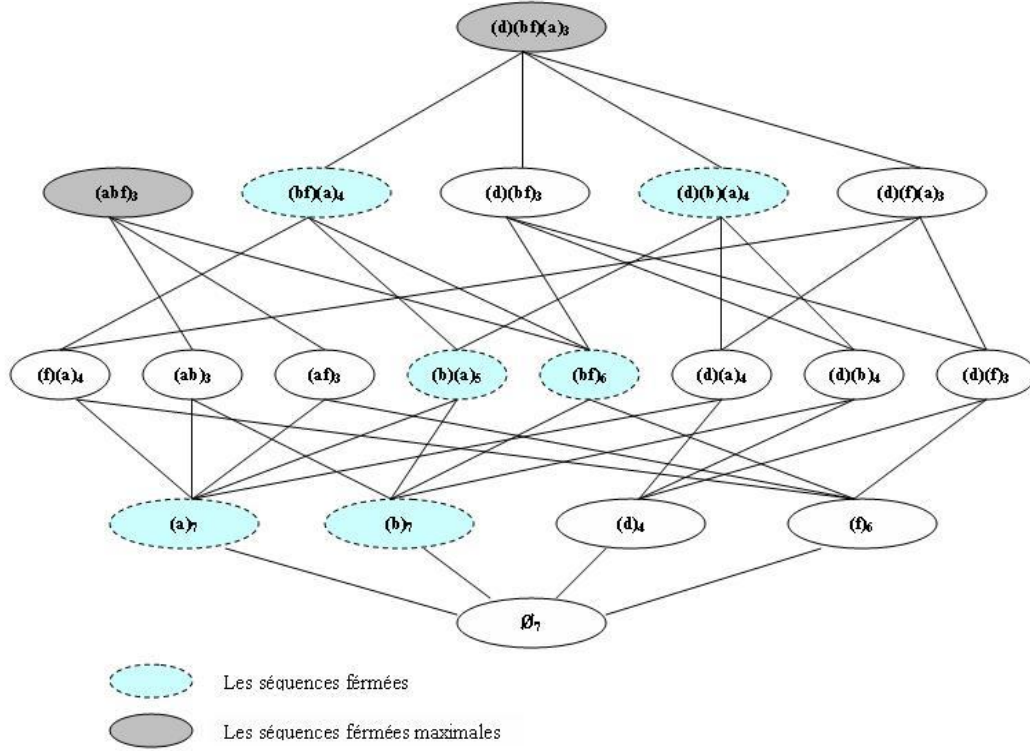


FIG. 4.1 – Le treillis de séquences 3-fréquentes

Ensemble des séquences 3-fréquentes :

$$F(D, 3) = \{ \langle (a), 7 \rangle, \langle (b), 7 \rangle, \langle (d), 4 \rangle, \langle (f), 6 \rangle, \langle (ab), 3 \rangle, \langle (af), 3 \rangle, \langle (b)(a), 5 \rangle, \langle (bf), 6 \rangle, \langle (d)(a), 4 \rangle, \langle (d)(b), 4 \rangle, \langle (d)(f), 3 \rangle, \langle (f)(a), 4 \rangle, \langle (abf), 3 \rangle, \langle (bf)(a), 4 \rangle, \langle (d)(bf), 3 \rangle, \langle (d)(b)(a), 4 \rangle, \langle (d)(f)(a), 3 \rangle, \langle (d)(bf)(a), 3 \rangle \}$$

pour $k = 2$:

$f_{\phi}^{(f)} = \text{Sup}_D(\phi) - \text{Sup}_D((f)) = 7 - 6 = 1 < 2$. Donc $\langle C_{\phi}^{(f)}, 1 \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non 2-anonyme ($\neg f$)

$f_{(b)}^{(bf)} = \text{Sup}_D((b)) - \text{Sup}_D((bf)) = 7 - 6 = 1$. Donc $\langle C_{(b)}^{(bf)}, 1 \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non 2-anonyme ($b \wedge \neg f$)

$f_{(d)}^{(d)(f)} = Sup_D((d)) - Sup_D((d)(f)) = 4 - 3 = 1$. Donc $\langle C_{(d)}^{(d)(f)}, 1 \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non \mathcal{Q} -anonyme $(\neg d)(f)$

$f_{(b)}^{(abf)} = Sup_D((b)) - Sup_D((ab)) - Sup_D((bf)) + \Delta Sup_D((abf)) = 7 - 3 - 6 + (2 - 1)3 = 1$. Donc $\langle C_{(b)}^{(abf)}, 1 \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non \mathcal{Q} -anonyme $(\neg a \wedge b \wedge \neg f)$

$f_{(b)(a)}^{(bf)(a)} = Sup_D((b)(a)) - Sup_D((bf)(a)) = 5 - 4 = 1$. Donc $\langle C_{(b)(a)}^{(bf)(a)} \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non \mathcal{Q} -anonyme $(b \wedge \neg f)(a)$

$f_{(d)(b)}^{(d)(bf)} = Sup_D((d)(b)) - Sup_D((d)(bf)) = 4 - 3 = 1$. Donc $\langle C_{(d)(b)}^{(d)(bf)}, 1 \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non \mathcal{Q} -anonyme $(d)(b \wedge \neg f)$

$f_{(b)(a)}^{(d)(b)(a)} = Sup_D((b)(a)) - Sup_D((d)(b)(a)) = 5 - 4 = 1$. Donc $\langle C_{(b)(a)}^{(d)(b)(a)}, 1 \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non \mathcal{Q} -anonyme $(\neg d)(b)(a)$

$f_{(d)(a)}^{(d)(f)(a)} = Sup_D((d)(a)) - Sup_D((d)(f)(a)) = 4 - 3 = 1$. Donc $\langle C_{(d)(a)}^{(d)(f)(a)}, 1 \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non \mathcal{Q} -anonyme $(d)(\neg f)(a)$

$f_{(f)(a)}^{(d)(f)(a)} = Sup_D((f)(a)) - Sup_D((d)(f)(a)) = 4 - 3 = 1$. Donc $\langle C_{(f)(a)}^{(d)(f)(a)}, 1 \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non \mathcal{Q} -anonyme $(\neg d)(f)(a)$

$f_{(bf)(a)}^{(d)(bf)(a)} = Sup_D((bf)(a)) - Sup_D((d)(bf)(a)) = 4 - 3 = 1$. Donc $\langle C_{(bf)(a)}^{(d)(bf)(a)}, 1 \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non \mathcal{Q} -anonyme $(\neg d)(b \wedge f)(a)$

$f_{(d)(b)(a)}^{(d)(bf)(a)} = Sup_D((d)(b)(a)) - Sup_D((d)(bf)(a)) = 4 - 3 = 1$. Donc $\langle C_{(d)(b)(a)}^{(d)(bf)(a)}, 1 \rangle$ est un canal d'inférence pour le motif séquentiel généralisé non \mathcal{Q} -anonyme $(d)(b \wedge \neg f)(a)$

Exemples des motifs séquentiels k -anonymes

Pour $k = 2$, le motif séquentiel généralisé $(\neg d)(\neg b \wedge f)(a)$ est 2-anonyme

car

$$f_{(f)(a)}^{(d)(bf)(a)} = \text{Sup}_D((f)(a)) - \text{Sup}_D((bf)(a)) - \text{Sup}_D((d)(f)(a)) + \Delta - \text{Sup}_D((d)(bf)(a)) = 4 - 4 - 3 + (2 - 1)3 = 0.$$

De même le motif séquentiel généralisé $(a \wedge \neg b \wedge \neg f)$ est 2-anonyme car

$$f_{(a)}^{(abf)} = \text{Sup}_D((a)) - \text{Sup}_D((ab)) - \text{Sup}_D((af)) + \Delta \text{Sup}_D((abf)) = 7 - 3 - 3 + (2 - 1)3 = 4.$$

$(d)(\neg b \wedge \neg f)$ est 2-anonyme

$$f_{(d)}^{(d)(bf)} = \text{Sup}_D((d)) - \text{Sup}_D((d)(b)) - \text{Sup}_D((d)(f)) + \Delta \text{Sup}_D((d)(bf)) = 4 - 4 - 3 + (2 - 1)3 = 0.$$

$(\neg d)(\neg a)$ est 2-anonyme

$$f_{\phi}^{(d)(a)} = \text{Sup}_D(\phi) - \text{Sup}_D((a)) - \text{Sup}_D((d)) + \Delta \text{Sup}_D((d)(a)) = 7 - 7 - 4 + (2 - 1)4 = 0.$$

Nous ne décrivons pas toutes les autres séquences k -anonymes.

Canaux d'inférence redondants

Dans l'ensemble de tous les canaux d'inférence qui sont dans la base de données D ($Ch(2, F(D, 3))$) nous avons beaucoup des canaux d'inférence redondants.

$$Ch(2, F(D, 3)) = \left\{ \left\langle C_{\phi}^{(f)}, 1 \right\rangle, \left\langle C_{(b)}^{(bf)}, 1 \right\rangle, \left\langle C_{(d)}^{(d)(f)}, 1 \right\rangle, \left\langle C_{(b)}^{(abf)}, 1 \right\rangle, \right. \\ \left. \left\langle C_{(b)(a)}^{(bf)(a)}, 1 \right\rangle, \left\langle C_{(d)(b)}^{(d)(bf)}, 1 \right\rangle, \left\langle C_{(b)(a)}^{(d)(b)(a)}, 1 \right\rangle, \left\langle C_{(d)(a)}^{(d)(f)(a)}, 1 \right\rangle, \left\langle C_{(f)(a)}^{(d)(f)(a)}, 1 \right\rangle, \right. \\ \left. \left\langle C_{(bf)(a)}^{(d)(bf)(a)}, 1 \right\rangle, \left\langle C_{(d)(b)(a)}^{(d)(bf)(a)}, 1 \right\rangle \right\}$$

Dans cet ensemble nous avons :

$$\left\langle C_{\phi}^{(f)}, 1 \right\rangle \leq \left\langle C_{(b)}^{(abf)}, 1 \right\rangle \text{ et } \left\langle C_{(b)}^{(bf)}, 1 \right\rangle \leq \left\langle C_{(b)}^{(abf)}, 1 \right\rangle$$

Ces trois motifs séquentiels non k -anonymes associés identifient seulement le client C_5 , et sont donc redondants.

Donc cette famille peut être représentée par le canal d'inférence $\left\langle C_{(b)}^{(abf)}, 1 \right\rangle$ avec (b) est fermée, et (abf) est maximale.

$$\left\langle C_{(b)(a)}^{(d)(b)(a)}, 1 \right\rangle \leq \left\langle C_{(bf)(a)}^{(d)(bf)(a)}, 1 \right\rangle \text{ et } \left\langle C_{(f)(a)}^{(d)(f)(a)}, 1 \right\rangle \leq \left\langle C_{(bf)(a)}^{(d)(bf)(a)}, 1 \right\rangle$$

Ces trois motifs séquentiels non k -anonymes associés identifient seulement le client C_6 , et sont donc redondants.

Donc cette famille peut être représentée par le canal d'inférence $\left\langle C_{(bf)(a)}^{(d)(bf)(a)}, 1 \right\rangle$ avec $(bf)(a)$ est fermée, et $(d)(bf)(a)$ est maximale.

$$\left\langle C_{\phi}^{(f)}, 1 \right\rangle \leq \left\langle C_{(d)(b)(a)}^{(d)(bf)(a)}, 1 \right\rangle, \left\langle C_{(b)}^{(bf)}, 1 \right\rangle \leq \left\langle C_{(d)(b)(a)}^{(d)(bf)(a)}, 1 \right\rangle, \left\langle C_{(d)}^{(d)(f)}, 1 \right\rangle \leq \\ \left\langle C_{(d)(b)(a)}^{(d)(bf)(a)}, 1 \right\rangle, \left\langle C_{(b)(a)}^{(bf)(a)}, 1 \right\rangle \leq \left\langle C_{(d)(b)(a)}^{(d)(bf)(a)}, 1 \right\rangle, \left\langle C_{(d)(b)}^{(d)(bf)}, 1 \right\rangle \leq \left\langle C_{(d)(b)(a)}^{(d)(bf)(a)}, 1 \right\rangle \\ \text{et } \left\langle C_{(d)(a)}^{(d)(f)(a)}, 1 \right\rangle \leq \left\langle C_{(d)(b)(a)}^{(d)(bf)(a)}, 1 \right\rangle$$

Ces sept motifs séquentiels non k -anonymes associés identifient seulement le client C_5 , et sont donc redondants.

Donc cette famille peut être représentée par le canal d'inférence $\left\langle C_{(d)(b)(a)}^{(d)(bf)(a)}, 1 \right\rangle$ avec $(d)(b)(a)$ est fermée, et $(d)(bf)(a)$ est maximale.