Chapitre 1

Introduction

Introduite par (Srikant et Agrawal, 1996), l'extraction de motifs séquentiels fréquents permet de découvrir des corrélations entre des événements selon une relation d'ordre (e.g. le temps). Ce problème est devenu au fil des années un domaine actif de la fouille de données avec de nombreux algorithmes à la clé.

La problématique de l'extraction de motifs séquentiels peut être vue comme une extension de la problématique de l'extraction d'itemsets et de règles d'associations. En effet, dans ce cadre, la dimension temporelle n'est pas prise en compte alors que pour les motifs séquentiels elle occupe une place centrale. La recherche de tels motifs consiste ainsi à extraire des enchaînements d'ensembles d'items, couramment associés sur une période de temps bien spécifiée. En fait, cette recherche met en évidence des associations inter-transactions, contrairement à celle des règles d'association qui extrait des combinaisons intra-transactions. Par exemple, des motifs séquentiels peuvent montrer que "60% des gens qui achètent une télévision, achètent un magnétoscope dans les deux ans qui suivent". Ce problème, posé à l'origine dans un contexte de marketing, intéresse à présent des domaines aussi variés que les télécommunications (détection de fraudes), la finance, ou encore la médecine (identification des symptômes précédant les maladies).

1.1 Domaines d'application

La découverte des relations séquentielles dans les ensembles de données est actuellement largement utilisée dans des domaines très variés tel que le marketing, l'aide à la décision, le management, la bioinformatique, l'analyse des performances des systèmes, l'analyse des réseaux de communication etc.

Les données séquentielles considérées sont des suites ordonnées de symboles (lettres,

signaux, états, événements, ...) et sont au cœur de domaines aussi divers que la fouille de texte, l'examen de séquences ADN, le monitoring de l'activation d'appareils, l'étude des comportements dans le temps d'acheteurs ou d'utilisateurs, conséquences de catastrophes naturelles ou encore l'étude de carrières.

Dans le Web, plusieurs applications réelles peuvent être envisagés. Un système de fouille de données séquentielle peut découvrir que :

- 30% des internautes qui ont visité le site de l'université de Biskra, on fait dans un délai de 15 jours une recherche google du mot "dattes"
- 15% des personnes qui ont acheté le livre de Kamber "Data mining : conpcepts and techniques" ont acheté dans le mois d'après le livre "Mining sequential patterns from large data sets"
- 40 % des personnes qui ont changé de PC pendant plus de deux fois dans une année sont des informaticiens,
- 50% des personnes qui achètent dans une bourse dans la semaine qui suit 3 achats gagnants perd plus de 60% de leurs gains.
- ...

1.2 Types de séquences

Selon le type des données qui les composent, les séquences de données peuvent être différenciées en deux types.

- Séquences symboliques : codes, caractères, articles, ...
- Séquences numériques : appelées aussi "séries temporelles".

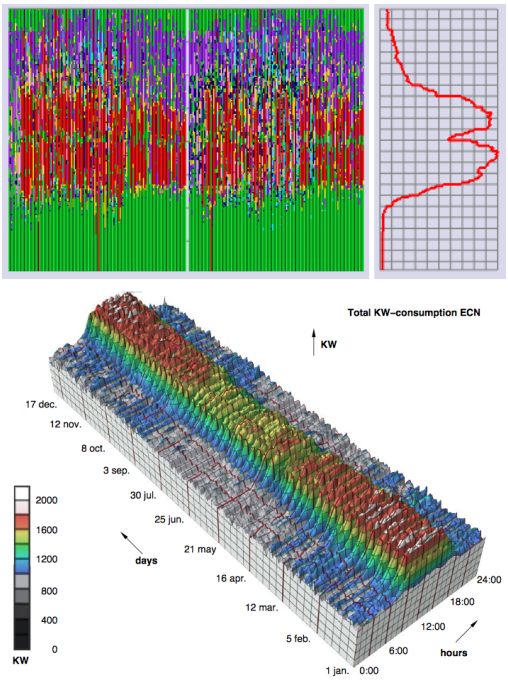
Selon l'ordre d'apparition dans les données, les séquences peuvent être de trois types :

- Motifs périodiques : ensembles d'items qui se répètent périodiquement : chaque semaine, chaque fin de semaine, le dernier jeudi de chaque mois, dans les cinq derniers jours du mois,...

  Ils peuvent être non fréquents mais périodiques.
- Motifs statistiques : ensemble d'items qui apparaissent fréquemment dans l'ensemble de données
- Motifs approximatif : A cause de certains bruits dans les données, la périodicité de certains items peut être perturbée. C'est des motifs approximatifs.

1.3 Visualisation de séquences

Les bases de données transactionnelles contenant des données séquentielles sont souvent très volumineuses et des visualisations adéquates peuvent contribuer efficacement à leur compréhension et analyse. représenter chaque item par une couleur différente et afficher les données à une grande résolution permet d'avoir un aperçu général des données. Les figures suivantes en représentent des exemples :



1.4 Représentation des séquences

Les données séquentielles sont généralement extraites à partir de bases de données transactionnelles contenant par exemples, le numéro de la transaction et éventuellement, sa date et heure, le code du client et l'ensemble des items achetés :

TID	Date	CID	Itemset
1	12/09/2017	1	{a,b,d }
2	17/09/2017	2	{b}
3	19/09/2017	1	{b,c,d }
4	28/09/2017	2	{a,b,c }
5	03/10/2017	3	{a,b }
6	06/10/2017	1	{b,c,d }
7	08/10/2017	3	{b,c,d }

Pour une analyse efficace, les bases transactionnelles sont souvent transformées et représentées sous une forme plus aisée pour l'analyse, on trouve différentes représentations :

1.4.1 Base de séquences

Dans cette représentation, les items des transactions sont regroupés par un sujet commun (client, jour, mois, ...)

CID	Séquence
1	{ {a,b,d }, {b,c,d }, {b,c,d } }
2	{ {b}, {a,b,c } }
3	{ {a,b },{b,c,d } }

1.4.2 Bitmap

La représentation Bitmap, représente les items en colonnes et les transactions par des vecteurs binaires représentant la présence ou non des items dans la transaction :