

Cours sur l'Analyse de Données Séquentielles

Chapitre 1 : Introduction

— Master 2
Systèmes d'Information, Optimisation et Décision

Dr D. AKROUR

2024/2025

Présentation du cours

- **Crédit** : 4
- **Coefficient** : 2
- **Mode d'évaluation**: 50% examen + 50% TP
- **L'objectif du cours** est d'initier les étudiants aux concepts et questionnements propres aux données séquentielles, leur représentation et leur analyse:
 - Fouille des motifs séquentiels dans les bases de données transactionnelles
 - Fouille des séries temporelles
 - Fouille des motifs séquentiels pour les données biologiques

- **Référence**:
 - Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.
 - Aggarwal, C. (2015). Data Mining The Text Book.
 - Fournier-Viger, P., Lin, J. C. W., Kiran, R. U., Koh, Y. S., & Thomas, R. (2017). A survey of sequential pattern mining. Data Science and Pattern Recognition, 1(1), 54-77.
 - Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2009). Mining sequence data in R with the TraMineR package: A user's guide. Geneva: Department of Econometrics and Laboratory of Demography, University of Geneva.

Plan du cours

1. Introduction
2. Types de séquences
3. Domaines d'application
4. Visualisation de séquences
5. Représentation de séquences

Introduction

—

Introduction

- La **fouille de données** est la science qui permet **d'extraire** des **connaissances** à partir d'un grand nombre de données.
- Elle permet **d'explorer** des données complexes et brutes et ...
- **Découvrir** des **motifs fréquents** (non triviaux, implicites, précédemment inconnus et potentiellement utiles) afin de ...
- Prendre les **bonnes décisions**, la **planification** stratégique, **l'optimisation** et la **gestion** des ressources humaines et matérielles.

Introduction

- **2 modèles** de fouille de données:
 - **Prédicatif**: prédire une valeur ou une classe inconnue pour un élément en se basant sur des données existantes.
 - (Classification, Régression, analyse des séries temporelles, prévision)

Introduction

- **2 modèles** de fouille de données:
 - **Prédictif**: prédire une valeur ou une classe inconnue pour un élément en se basant sur des données existantes.
 - (Classification, Régression, analyse des séries temporelles, prévision)
 - **Descriptif**: décrire des motifs ou des relations cachés dans les données sans chercher à prédire une valeur ou un résultat spécifique.
 - (Clustering, règles d'association, découverte des séquences)

Introduction

- La **fouille de données séquentielles** est utilisée sur des base de données contenant des **séquences** (données **ordonnées**).
- Elle permet de **révéler** et d'**extraire** des motifs récurrent et des **tendances** cachées dans des ensemble de données ordonnées.
- **Ca sert à quoi ?**

Introduction

- La **fouille de données séquentielles** est utilisée sur des base de données contenant des **séquences** (données **ordonnées**).
- Elle permet de **révéler** et d'**extraire** des motifs récurrent et des **tendances** cachées dans des ensemble de données ordonnées.
- **Ca sert à quoi ?**
 - **Comprendre** les comportements des humains et les phénomènes complexes, mais aussi le déroulement de divers événements au fil du temps.

Introduction

- La **fouille de données séquentielles** est utilisée sur des base de données contenant des **séquences** (données **ordonnées**).
- Elle permet de **révéler** et d'**extraire** des motifs récurrent et des **tendances** cachées dans des ensemble de données ordonnées.
- **Ca sert à quoi ?**
 - **Comprendre** les comportements des humains et les phénomènes complexes, mais aussi le déroulement de divers événements au fil du temps.
 - **Prévoir** ce qui pourrait se passer dans les jours, les mois ou les années à venir et donc **planifier** à l'avance, **optimiser** les processus, **personnaliser** les services, ...

Introduction

- La **fouille de motifs séquentiels** est introduite par Agrawal et Srikant en 1996.
- La **question** posée par Agrawal et srikant (dans un contexte de marketing):

Introduction

- La **fouille de motifs séquentiels** est introduite par Agrawal et Srikant en 1996.
- La **question** posée par Agrawal et srikant (dans un contexte de marketing):

Comment extraire des sous-séquences fréquentes
à partir d'un ensemble de séquences ?

Introduction

- La **fouille de motifs séquentiels** est introduite par Agrawal et Srikant en 1996.
- La **question** posée par Agrawal et srikant (dans un contexte de marketing):

Comment extraire des sous-séquences fréquentes
à partir d'un ensemble de séquences ?

- Par **exemple**:
 - trouver des séquences d'articles achetés par de nombreux clients au fil du temps.

Introduction

C'est quoi une séquence ?

Une séquence est une liste ordonnée d'événements.

Exemple : S est une séquence qui est composé de cinq événements.

$$S = \langle e1, e2, e3, e4, e5 \rangle$$

Introduction

C'est quoi une séquence ?

Une séquence est une liste ordonnée d'événements.

Exemple : E-Learning

- Un événement représente une ressource pédagogique (examens, exercices, cours, etc.) que les étudiants consultent.

$$\text{élève-143} = \langle \mathbf{R3}, \mathbf{R8}, \mathbf{R13}, \mathbf{R27}, \mathbf{R29} \rangle$$

- **élève-143** représente l'identifiant de l'élève et **Rn** représente l'identifiant d'une ressource pédagogique.

Introduction

C'est quoi une séquence ?

Une séquence est une liste ordonnée d'événements.

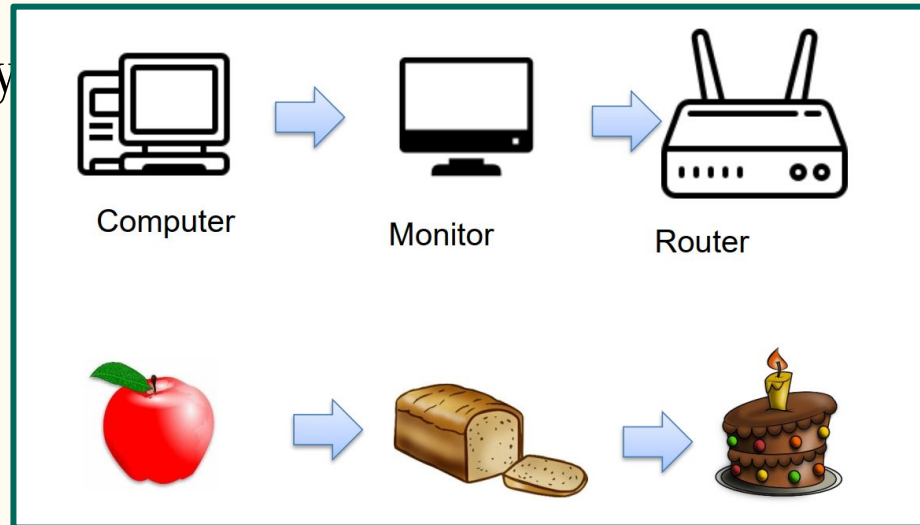
Exemple : Analyse de panier du marché

Introduction

C'est quoi une séquence ?

Une séquence est une liste ordonnée d'événements.

Exemple : Analyse



Introduction

C'est quoi une séquence ?

Une séquence est une liste ordonnée d'événements.

Exemple : Analyse de panier du marché

$$S = \langle \{A, B\}, C \rangle$$

- **S** représente une séquence de deux transactions effectuées par un client donné.
- Si **S** est une séquence fréquente cela veut dire que les clients achètent souvent les produit A et B, puis le produit C.
- L'entreprise peut proposer des promotions sur le produit B après l'achat des produits A et B.

Introduction

CID	TID	Transactions
1	100	a,b,c,d
3	111	a,f,d,c
1	122	d,e,p
3	133	b,f,s,a
1	144	b,c,d,e
3	155	a,f,d,c
1	166	a,e,p

Base de données transactionnelles

La séquence de transactions
pour le client 1 est :

$S = \langle (a, b, c, d), (d, e, p), (b, c, d, e), (a, e, p) \rangle$

Introduction

Transformer une base de données transactionnelle en une base de données séquentielle

CID	TID	Transactions
1	100	a,b,c,d
3	111	a,f,d,c
1	122	d,e,p
3	133	b,f,s,a
1	144	b,c,d,e
3	155	a,f,d,c
1	166	a,e,p



CID	Sequences
1	(a,b,c,d)(d,e,p)(b,c,d,e)(a,e,p)
3	(a,f,d,c)(b,f,s,a)(a,f,d,c)

Introduction

- **Comment déterminer si une séquence est une séquence fréquente (significative) ou non?**
 - Calculer sa fréquence dans la base de données,
 - La comparer avec un **seuil de support minimum** (déterminé par l'utilisateur).

Introduction

- **Comment déterminer si une séquence est une séquence fréquente (significative) ou non?**
 - Calculer sa fréquence dans la base de données,
 - La comparer avec un **seuil de support minimum** (déterminé par l'utilisateur).
- **Méthodologie?**
 - Extraire toutes les séquences fréquentes dans un ensemble de données est une tâche très **complexe** en raison de **la taille de l'espace de recherche**.
 - Il faut utiliser les **techniques** de fouille de données séquentielles.

Types de séquences

—

Types de séquences (selon le type de données)

Données symboliques

- Une liste ordonnée de **valeurs nominales** (symboles)

Données de séries temporelles

- Une liste ordonnée de **valeurs numériques**

Types de séquences (selon le type de données)

Données symboliques

- Une liste ordonnée de **valeurs nominales** (symboles)
- **Représente** : séquence de mots, de caractères, d'articles achetés, de pages web visités, les données biologiques, ...

Données de séries temporelles

- Une liste ordonnée de **valeurs numériques**
- **Représente** : cours de bourse, température, consommation d'électricité, le trafic routier... (Évolution d'une quantité spécifique au cours du temps.)

Types de séquences (selon le type de données)

Données symboliques

- Une liste ordonnée de **valeurs nominales** (symboles)
- **Représente** : séquence de mots, de caractères, d'articles achetés, de pages web visités, les données biologiques, ...
- L'ordre peut ne pas être défini par la dimension du temps.

Données de séries temporelles

- Une liste ordonnée de **valeurs numériques**
- **Représente** : cours de bourse, température, consommation d'électricité, le trafic routier... (Évolution d'une quantité spécifique au cours du temps.)
- L'ordre est défini par la dimension du temps.

Types de séquences (selon le type de données)

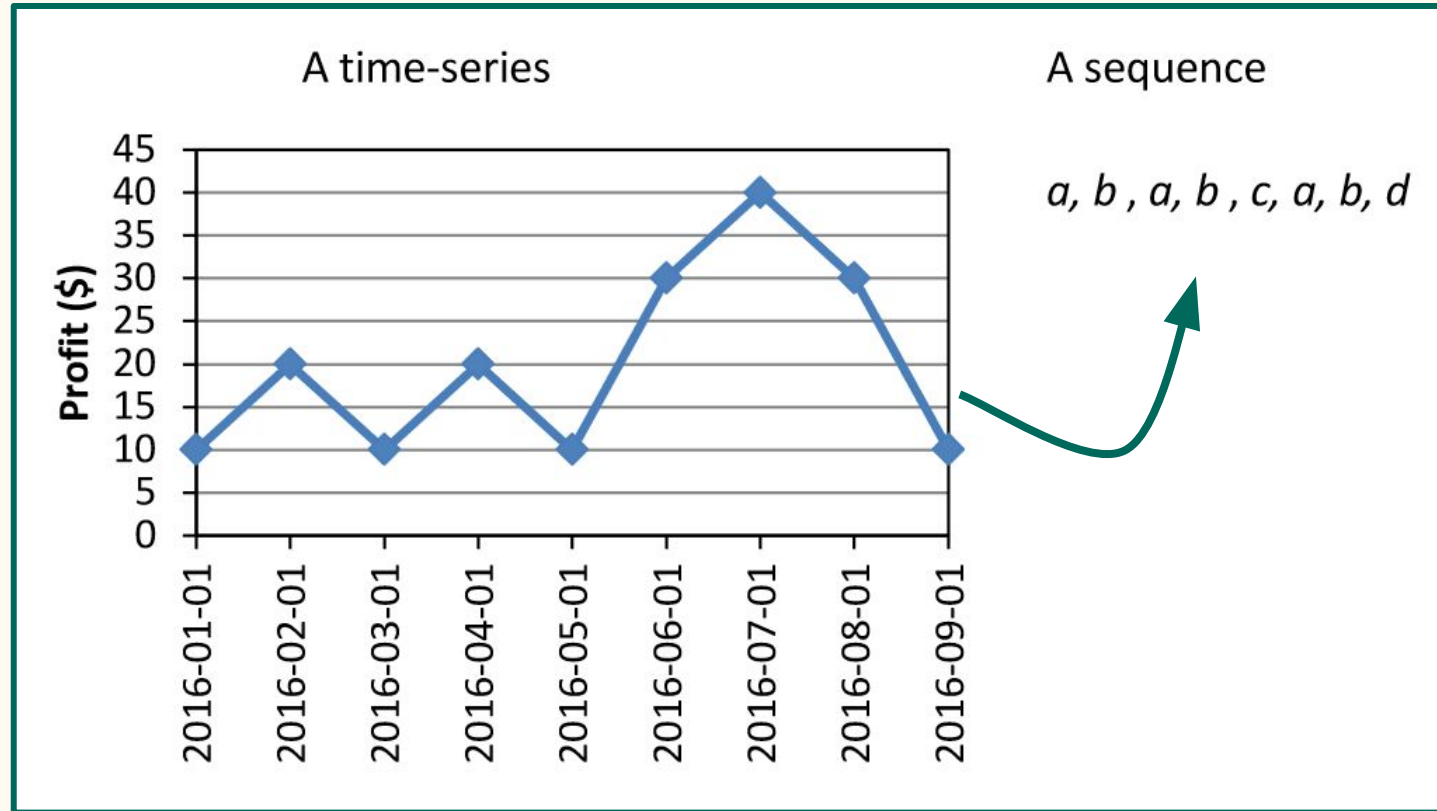
Données symboliques

- Une liste ordonnée de **valeurs nominales** (symboles)
- **Représente** : séquence de mots, de caractères, d'articles achetés, de pages web visités, les données biologiques, ...
- L'ordre peut ne pas être défini par la dimension du temps.
- **Application**: fouille de motifs séquentielle

Données de séries temporelles

- Une liste ordonnée de **valeurs numériques**
- **Représente** : cours de bourse, température, consommation d'électricité, le trafic routier... (Évolution d'une quantité spécifique au cours du temps.)
- L'ordre est défini par la dimension du temps.
- **Application**: prévision

Types de séquences (selon le type de données)



Types de séquences (selon les caractéristiques des événements qu'elles décrivent)

- **Les séquences de séries temporelles :**
 - Longues séquences de données numériques, enregistrées à des intervalles de temps égaux.

Types de séquences (selon les caractéristiques des événements qu'elles décrivent)

- **Les séquences de séries temporelles :**
 - Longues séquences de données numériques, enregistrées à des intervalles de temps égaux.
- **Les séquences de données symboliques :**
 - Longues séquences d'événements qui ne sont généralement pas observées à des intervalles de temps égaux. Pour de telles séquences, les intervalles entre les événements enregistrés importent peu.

Types de séquences (selon les caractéristiques des événements qu'elles décrivent)

- **Les séquences de séries temporelles :**
 - Longues séquences de données numériques, enregistrées à des intervalles de temps égaux.
- **Les séquences de données symboliques :**
 - Longues séquences d'événements qui ne sont généralement pas observées à des intervalles de temps égaux. Pour de telles séquences, les intervalles entre les événements enregistrés importent peu.
- **Les séquences biologiques**
 - Des séquences d'ADN et de protéines. Ces séquences sont généralement très longues et portent un sens sémantique important, complexe mais caché. Dans ce cas, les intervalles sont généralement importants.

Domaines d'applications

—

Domaines d'application

Séquences symboliques

- Analyse du panier du marché
 - Comprendre le comportement et les habitudes d'achat des clients.
 - Identifier les séquences fréquentes d'articles achetés.
 - Lorsque vous achetez un produit “x” aujourd’hui, quels articles êtes-vous le plus susceptible d'acheter dans les prochains jours ?
 - **40%** des clients achètent en **premier** lieux un ordinateur, **ensuite** une caméra et enfin un téléphone portable dans les **prochains six mois**
 - **Utilité commerciale** ? diffuser des annonces ciblées pour des clients similaires, planifier et s'approvisionner à l'avance, ranger et emballer les articles ensemble.

Domaines d'application

Séquences symboliques

- Text Mining
 - Comprendre la structure du texte et donc classifier automatiquement des documents ou obtenir un aperçu du contenu d'un document sans le lire.
 - Comprendre le contexte dans lequel une personne parle en fonction de la séquence de mots prononcé auparavant (NLP).
- Conséquences de catastrophes naturelles
 - Identifier l'acheminement d'événements suite à un événement majeur (inondation, ...) ainsi réduire les dégâts et préparer un plan de secours

Domaines d'application

Séquences symboliques

- Utilité scientifique dans le secteur médical
 - En identifiant un motif séquentiel fréquent et prédominant
 - Séquence de traitements effectués pour guérir une maladie particulière (traitement 1 -> Traitement 2 -> Traitement 3 -> Traitement 4)
 - On peut comprendre le progrès de santé des patients depuis un traitement à un autre
 - Comprendre quel devrait être le prochain traitement
 - Séquence comportant l'acheminement de symptômes précédant une maladie (symptôme 1 -> symptôme 2 -> symptôme 3 -> symptôme 4)
 - Prévenir la maladie

Domaines d'application

Séquences symboliques

- **Web Mining:** séquences de pages Web visitées par les utilisateurs "clickstream"
 - **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 : concepts 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.

Domaines d'application

Séquences symboliques

- Bioinformatique
 - Détecter des motifs dans l'ADN qui sont associés à des maladies génétiques ou à des caractéristiques spécifiques. Cela aide dans le développement de traitements ou la compréhension des processus biologiques.
- Etude de transition professionnel
 - Séquences d'emplois ou de statuts professionnels qu'une personne occupe au cours de sa carrière (chômage, emploi, formation, ...).
 - L'analyse de séquences aide à comprendre les chemins typiques d'évolution de carrière, comme les transitions fréquentes entre secteurs d'activité, et à concevoir des politiques d'accompagnement professionnel.

Domaines d'application

Séquences de série temporelles

- Dans le domaine économique
 - **Le taux de chômage:** détecter des périodes de forte augmentation ou de diminution, et de corréler ces données à des événements économiques ou politiques et aussi prévenir et lutter contre le chômage.
 - **Les cours des actions (bourse):** décider du moment optimal pour acheter ou vendre des actions ou bien identifier les risques de chute du marché et prendre des mesures pour limiter les pertes.

Domaines d'application

Séquences de série temporelles

- Dans le domaine démographique :
 - **Les flux migratoires** : prévoir les besoins en termes de logement, d'emploi ou de soins de santé dans les zones à forte immigration, aussi prédire les destinations les plus probables des migrants suite à des guerres ou des catastrophes naturelles et planifier les ressources nécessaires.

Domaines d'application

Séquences de série temporelles

- Dans le domaine énergétique
 - **Température:** détection d'anomalies ou bien prévision et anticipation des événements extrêmes comme les canicules ou les vagues de froid.
 - **Consommation d'électricité:** identifier les périodes de forte demande et les schémas d'utilisation qui peuvent varier selon l'heure de la journée, le jour de la semaine ou la saison. Ainsi on peut mieux gérer leurs ressources et éviter les coupures de courant et réduire les coûts pendant les périodes de pic.

Visualisation de séquences

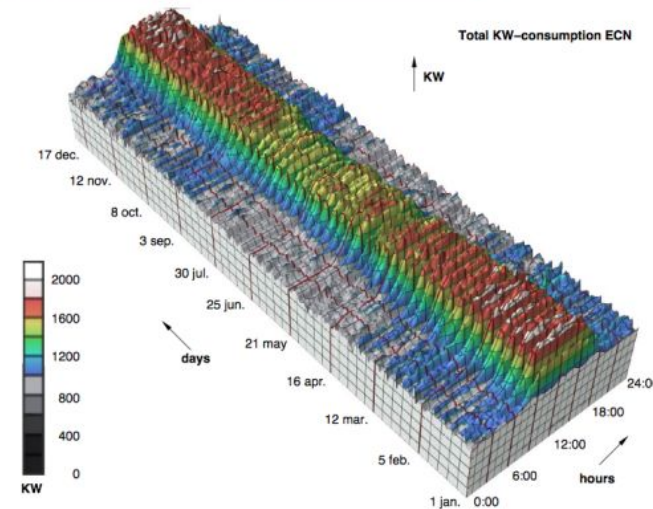
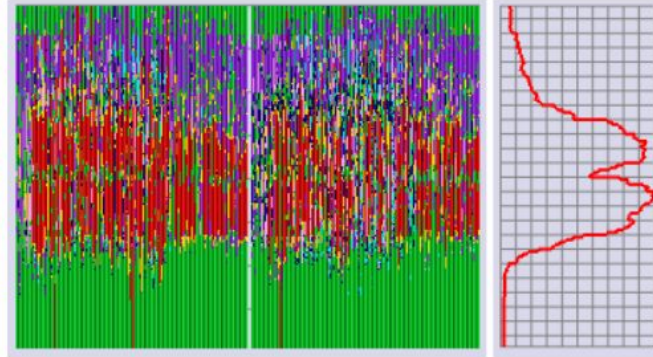
—

Visualisation de séquences

- La visualisation des données est un moyen efficace pour identifier les motifs cachés, les observations inhabituelles, le changement dans le temps, les relations entre les points de données, etc.
- Différentes visualisations aident à analyser les données et génèrent des informations utiles.
- Selon le type de données
 - données **numériques** continues (histogramme, boxplots, ...)
 - données **catégorielles** (diagrammes en barres, ...)
 - données **séquentielles** Représenter chaque item par une couleur différente et afficher les données à une grande résolution permet d'avoir un aperçu général des données

Visualisation de

- La visualisation de motifs cachés, les tendances dans le temps, les relations
 - Différentes visualisations des informations
 - Selon le type de données
 - données numériques
 - données catégorielles
 - données séquentielles
- d'avoir un aperçu



face pour identifier les
changement dans le
etc.

données et génèrent

boxplots, ...)

...)

m par une couleur
résolution permet

Représentation de séquences

—

Représentation des données

- Pour une analyse efficace, les bases transactionnelles sont souvent transformées et représentées sous une forme plus aisée pour l'analyse,
 - **Base de séquence**

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 }



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

Représentation des données

- Pour une analyse, les données doivent être transformées et représentées
 - **Base de données**

TID	Date
1	12/09/2011
2	17/09/2011
3	19/09/2011
4	28/09/2011
5	03/10/2011
6	06/10/2011
7	08/10/2011

- **Séquences simples**
- **Séquences avec des timestamps:** chaque événement est associé à une date ou une heure
- **Séquences à événements multiples (séquences multidimensionnelles) :** chaque événement peut contenir plusieurs dimensions ou attributs

d }

Représentation des données

- Pour une analyse efficace, les bases transactionnelles sont souvent transformées et représentées sous une forme plus aisée pour l'analyse,
 - **Bitmap**

Les bitmaps compressent les informations en les exprimant sous forme de bits (1 et 0), ce qui permet une utilisation optimisée de la mémoire et rend les opérations logiques rapides.

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

Questions?