République Algérienne Démocratique et Populaire

Ministère de L'Enseignement Supérieur et De la Recherche Scientifique
Université Mohamed El Bachir El Ibrahimi de Bordj-Bou-Arreridj
Faculté des mathématiques et d'informatique

MEMOIRE

Présenté en vue de l'obtention du diplôme de

Master en informatique

Spécialité : TECHNOLOGIE D'INFORMATION ET DE COMMUNICATION

THEME

L'analyse d'activité utilisateur dans les réseaux sociaux basé sur le Biclustering Par les règles d'association: Étude de cas réseau Gowalla

Présenté par :

DAOUD Assia

NACEF Lynda

Soutenu publiquement le : Juillet 2021
Devant le jury composé de :

Président: Mme : CHELLAKH	**MCB à L'U. El Bachir El Ibrahimi-BBA**
Examinatrice : Mme : BENABID.S	**MCB à L'U. El Bachir El Ibrahimi-BBA**
Encadreur : Mr: ATTIA Abdelouahab	**MCB à L'U. El Bachir El Ibrahimi-BBA**
Co-Encadreur : Mr : BELAIDI Yehia	**MCB à L'U. El Bachir El Ibrahimi-BBA**

Promotion : 2020/2021

Introduction Générale

A nos jours la communication ne peut plus être unidirectionnelle, dans un monde bouleversé par les NTIC (nouvelle technologie de l'information et de la communication), les réseaux sociaux prennent aujourd'hui une place de plus en plus primordiale dans la vie des gens, les réseaux sociaux ont pu remplacer tous les autres moyens de communication traditionnels, du fait qu'ils ont à faire à des clients et consommateurs qui créent des contenus, partagent des informations et interagissent entre eux.

La science des réseaux (networks science) tient ses origines en sociologie (SNA ou Social Networks Analysis) avec des travaux datant du début du 20e siècle mais a pris un essor nouveau ces quinze dernières années et impacte la plupart des disciplines.

Les techniques de regroupement des ensembles de données sont utilisées dans un certain nombre d'applications. En particulier, le biclustering est très important dans le domaine des données d'informatique. Les algorithmes de biclustering ont également des applications importantes dans classification des échantillons. De nombreuses méthodes de biclustering, et les algorithmes de clustering en général, utilisent des modèles ou stratégies heuristiques pour identifier le « meilleur » regroupement d'éléments selon certains métrique, définition de cluster, et donc aboutir à des clusters sous-optimaux.

Les méthodes du data mining aussi sont des outils nouvelles dans le domaine de l'analyse des réseaux sociaux, « la fouille des réseaux sociaux». Plusieurs travaux ont été utilisé la fouille de données comme une technique d'analyse dans les réseaux sociaux. L'application du data mining dans les réseaux sociaux permet d'analyser les comportements des utilisateurs. Dans ce travail, nous présentons une étude sur un réseau social avec des milliers d'utilisateurs. Ce réseau contient notamment des données qui circulent entre les utilisateurs à travers le site web. Notre objectif est de faire une étude sur les différentes recherches effectuées dans ce site afin d'extraire les activités d'utilisateurs dans ce réseau social.

Introduction

Par définition, un réseau social est un groupe d'individus ou d'organisations qui connectent des internautes entre eux par la communication. Cela leur permet de partager des opinions, des idées et même du contenu. Les fonctions de tous les types de réseaux sont similaires: après inscription, vous pouvez effectuer une recherche nominative. Lorsque vous trouvez un contact, cliquez simplement sur un bouton pour vous connecter avec cette personne. Internet a introduit les relations interpersonnelles dans le domaine virtuel, changeant complètement le concept des relations interpersonnelles. Les réseaux sociaux les plus connus aujourd'hui sont Facebook, Twitter, Instagram, etc. Ces services en ligne continuent d'attirer les internautes. Le premier réseau social s'appelait Classmates.com et a été créé par Randy Conrads en 1995 dans le but de renouer avec d'anciens camarades de classe. Néanmoins, Classmates.com ne fournit pas toute la disponibilité du réseau actuel

Sixdegrees.com sera le premier site web qui regroupera toutes les fonctions de base d'un réseau social en 1997. Pour des raisons économiques, le site a dû être fermé. Le réseau social est en constante évolution : -Facebook : L'un des plus célèbres créé par Mark Zuckerberg en 2004 compte à lui seul 900 millions de membres. Chaque internaute peut se créer un profil personnel avec un réseau d'amis. Le réseau peut également partager du texte, des photos, des liens vers d'autres sites et des vidéos. Il permet la communication avec plusieurs utilisateurs. Il est également utilisé par les entreprises et les artistes pour leur publicité.

1- Les réseaux sociaux

Dans le Domaine de la technologie, un réseau social consiste en un service qui permet de réunir différentes personnes pour communiquer sur des sujets précis ou non. Dans une certaine mesure, les réseaux sociaux sont issus de forums, de groupes de discussion et de forums de discussion introduits aux premiers jours d'Internet. Depuis le début des années 2000, l'existence de réseaux sociaux (appelés aussi réseaux communautaires) est devenue de plus en plus importante et tend à croître de façon exponentielle en fonction de diverses caractéristiques les premiers grands réseaux sociaux (MySpace et Facebook) se sont positionnés comme des généralistes. Le service peut partager le contenu de son choix avec ses contacts, quel que soit le sujet.

1.1. À quoi sert un réseau social ?

Tout dépend du but, les médias sociaux comme Facebook avec près de 40 millions d'utilisateurs en France fin 2019 sont nés de la volonté de créer une passerelle virtuelle entre l'utilisateur et ses amis. En d'autres termes, le succès croissant de l'entreprise de Mark Zuckerberg a incité l'entreprise à utiliser ce média social comme levier pour ses activités de marketing en ligne. Leurs objectifs : augmenter leur visibilité sur Internet, augmenter la fréquentation de leur site internet, cibler leurs clients potentiels pour les convertir en clients, fidéliser leurs clients existants et communiquer avec eux (newsletters, lancement de ventes à durée limitée) Pour un usage strictement professionnel, nous recherchons plutôt LinkedIn, qui permet de publier son CV et de consulter les offres d'emploi. Instagram repose d'avantage sur la vision.

Les photos et vidéos sont privilégiées, et la présence de la création artistique est forte (les artisans utilisent pour vendre leurs produits, et les peintres et photographes utilisent les tendances pour mettre en valeur leurs talents). Les célébrités communiquent avec leurs followers sur Twitter principalement. YouTube est la plateforme de référence pour le visionnage et téléchargement de vidéos.

Les autres médias sociaux les plus utilisés en France incluent Snapchat (vidéos et photos qui s'effacent automatiquement au bout de quelques secondes) et Copains d'avant (pour retrouver ses anciens camarades)

1.2. Comment fonctionne le réseau social ?

Les réseaux sociaux se nourrissent des émotions des gens, de leurs besoins de communication et d'interaction. Les fonctionnalités existantes, comme les fameux likes Facebook ou les hashtags Twitter, sont conçues pour faire réagir les utilisateurs et les inciter à partager des informations et des opinions.

Le besoin de réalisation de soi et les avantages émotionnels sont en effet réels et sont réalisés par des moyens virtuels. Le réseau de chaque utilisateur est généralement comparé à une étoile. Chaque branche correspond à une caractéristique qui la définit: contacts, centres d'intérêts, groupes, données personnelles... Une fois votre réseau tissé, la

plateforme utilisera vos liens forts (nos contacts proches) pour créer des liens faibles (vous Amis amis). Par conséquent, la communauté grandit et le réseau social s'épanouit.

1.3. Les différents types des réseaux sociaux :

Il existe plusieurs types des réseaux sociaux on site quelques exemple :

1.3.1. Les réseaux sociaux généralistes pour discuter :

- ***Facebook :*** chaque internaute peut créer un profil limité au réseau d'amis

(Personnes proches ou inconnues) Il a accepté. Il permet le partage : statut, photos, liens et vidéo. Il est également utilisé par les entreprises et les artistes pour leur promotion, grâce à une page de fans que tout le monde peut visiter. Le leader mondial.

- ***Twitter :*** outil Weibo qui permet d'envoyer des messages appelés « tweets » à suivez les internautes de chaque compte. Ce sont des « followers » ou abonnés.
- ***MySpace:*** Espace réseau personnalisé. La possibilité de fournir des informations personnelles et faire un blog. Parce qu'il existe de nombreux groupes de musique, ce réseau est particulièrement connu occupant cet espace, sa popularité a décliné ces dernières années.
- ***Snapchat* :** Il s'agit d'une plate-forme sociale de de partage photos et de vidéos disponible sur smartphone et qui permet de discuter avec des amis en utilisant des images. La particularité de ce réseau social est que chaque image ou vidéo envoyée ne peut être visible que durant une période de temps par son destinataire. Il vous permet d'explorer les nouvelles et mêmede consulter des histoires en direct qui se passent dans le monde.

1.3.2. Les réseaux Avec contenu:

Le concept de réseaux sociaux définit une communauté d'utilisateurs réunis Fonctions d'intérêts communs (loisirs, passions, musique, voyages, vie Professionnel...) via leur site Internet enregistré. Où ils créent un profil (Les utilisateurs décrivent leurs pages personnelles) et se connecter avec les autres Les utilisateurs inscrits sur un même site échangent entre eux des messages publics ou privés, Liens hypertexte, vidéos, photos ou utiliser des applications de collaboration

Figure (I-1) : Réseau d'intérêts communs

1.3.3. Réseau social professionnel d'entreprise:

Est un réseau social dédié aux professionnels, axé sur la promotion et la communication professionnelle de ses membres est différente des réseaux sociaux publics ordinaires aimez Facebook. Les plus connus sont Viadeo, Linkedin et XING.

- **LinkedIn :** C'est un réseau professionnel qui vous permet de publier et de partager votre CV.
- **Viadeo :** Il permet d'établir des liens professionnels et de vous faire connaître en le publiant reprendre. Il offre également des opportunités d'emploi.
- **Ziki :** Son objectif est d'aider les entreprises à trouver les meilleurs prestataires la réalisation d'un projet.
- **InterFrench :** Un réseau français mondial pour les projets à l'étranger.
- **Piwie :** Le premier chat d'entreprise.

1.3.4. Les réseaux de services

- **Ma-résidence** : échanger des adresses, des services et un endroit pour parler de votre relation voisins copains d'avant et Trombi: Permet de retrouver d'anciens camarades de classe.
- **RéseauxLycée et Etnoka**: Un réseau d'étudiants du secondaire et du collégial qui peut être discuté, Organisez des rassemblements et partagez des leçons.
- **BeGlob** : Dédié aux amoureux du voyage. Vous permet de communiquer des compétences, des suggestions, vivre

1.3.5. Un réseau avec une dynamique importante :

Un réseau dynamique est un réseau dont la topologie peut changer au cours de l'exécution.

Programme parallèle ou entre l'exécution de deux programmes ; c'est la topologie

La logique du réseau, c'est-à-dire son comportement. Les réseaux dynamiques sont généralement utilisés pour les multiprocesseurs à mémoire partagée. Dans ce cas, la connexion réseau, processeurs dans un référentiel central partagé : Il s'agit d'un réseau interconnecté.

Les instructions machine (Charger, Stocker, etc.) qui accèdent à la mémoire transitent par le réseau : On dit que le couplage est très fort. Par conséquent, le réseau doit être très efficace pour ne pas ralentissez trop les instructions machines les plus sensibles.

1.3.6. Les réseaux géo localisés :

Sont des sites qui basent sur la géo localisation. Foursquare et Gowalla : possibilité d'ajouter des amis lorsque l'on se rend quelque part avec possibilité de signaler sa présence ...

1.4. La présentation et la visualisation :

En raison de la diffusion et de la démocratisation de la technologie Internet, il est devenu de plus en plus Collecte facile de grandes quantités de données et extraction de grands réseaux sociaux listez-les. Par exemple, via des sites Web tels que Facebook ou Friendster, Internet

Les réseaux sociaux disponibles sont très vastes et étendus (millions d'utilisateurs), riches (ycompris Les multiples relations ou informations des participants au réseau), qui se développent et s'accroissent avec le temps.

Depuis 1930 la représentation visuelle des réseaux sociaux a émergé. Prenons un exemple : Jacob Moreno a été l'un des pionniers de l'utilisation de la représentation de type nœud de lien pour transmettre son travail. Dans cette représentation, un nœud représente un participant et un lien représente la relation du réseau. Bien que la visualisation soit

souvent utilisée dans cette situation, c'est-à-dire pour communiquer des résultats ou expliquer des théories, Ils peuvent également être utilisés pour analyser des données.

Statisticien Tukey, introduction l'analyse exploratoire des données montre que la représentation graphique peut être un puissant outil d'analyse. Son intérêt réside dans l'observation de représentations multiples de données brutes, ce qui permet de poser des questions ou d'établir des hypothèses qu'on ne peut imaginer a priori.

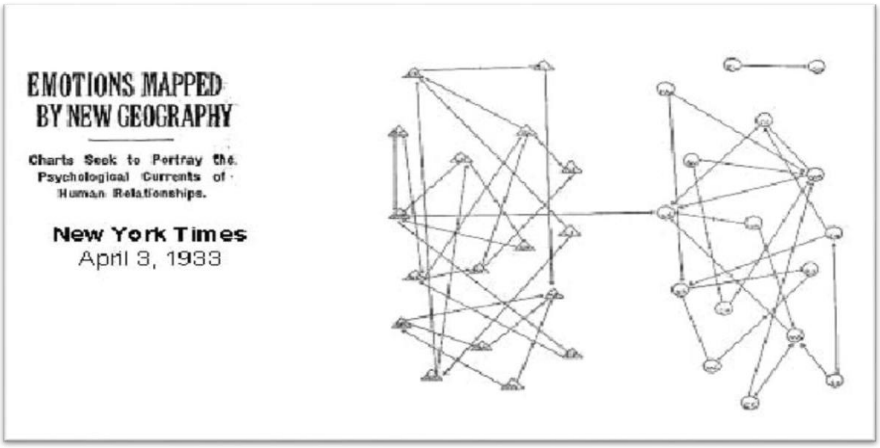


Figure (I-2) : Réseau d'amitiés entre garçons (triangles) et filles (cercle) créé par J.Moreno.

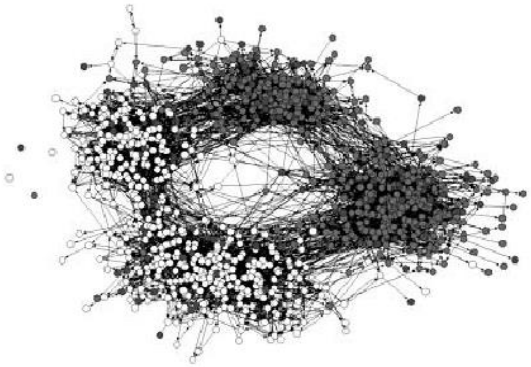


Figure (I-3) : Réseau d'amitié entre lycée J.moody

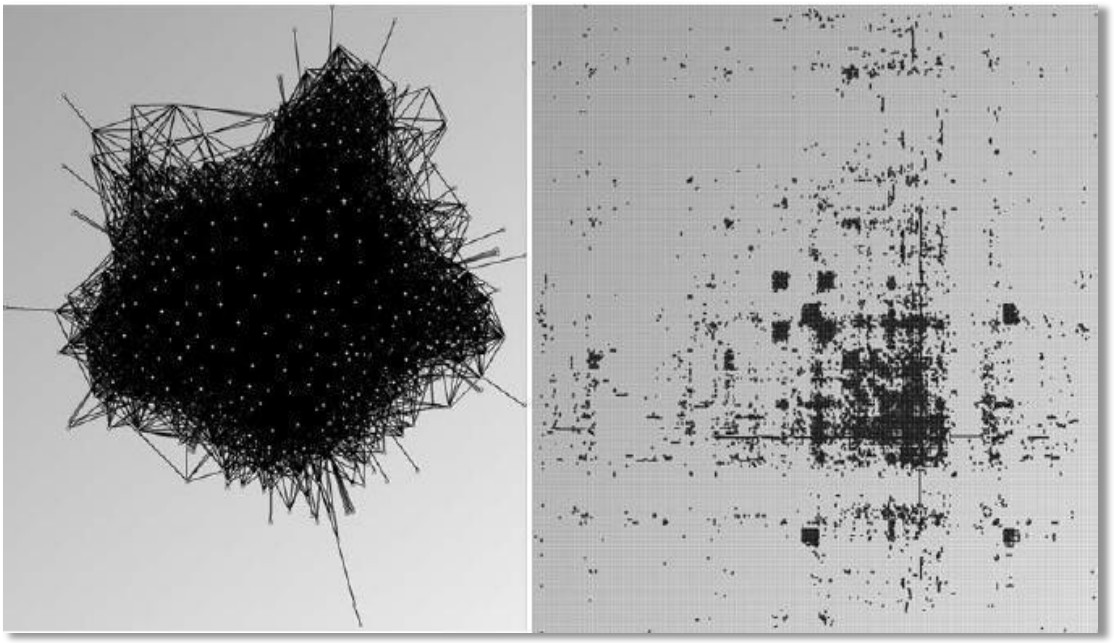


Figure (I-4) : Réseau d'échange de courriers électroniques entre plus de 450 Chercheurs, pendant un an. Noeud-lien (a), matrice (b)

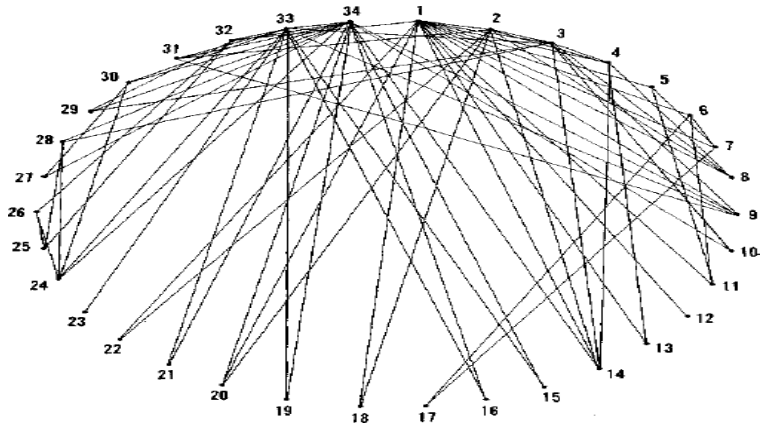


Figure (I-5) : Représentation graphique des relations sociales entre les 34 membres du club de karaté

La représentation la plus courante des graphiques ou des réseaux est la représentation des liens de nœuds, de sorte que cette représentation présente des avantages familiers à la plupart des chercheurs.

Cependant, lorsque le réseau représenté est large (nombreux nœuds) ou dense (nombreux maillons). Avec de plus en plus de données à analyser, ces problèmes d'amplification

deviennent plus importants ; avec l'émergence de réseaux sociaux en ligne larges et denses, ils se sont paralysés.

Dans les cinq dernières années, le domaine de la visualisation d'information a connu plusieurs avancées et a permis de trouver des représentations alternatives (et complémentaires) aux représentations nœud-lien.

1.4.1. Représentation nœud-lien :

L'analyse des réseaux sociaux a débuté il y a plus de 70 ans, avec les travaux empiriques de Jacob Moreno**.** Il a été un des pionniers à utiliser la représentation nœud-lien pour communiquer sur ses travaux. Dans cette représentation, un nœud représente un acteur et un lien représente une relation du réseau. Il s'agit de la représentation la plus courante des réseaux. Wasserman et Faust. Présentent les diverses catégories de méthodes : analyses statistiques, structurelles et exploratoires. Freeman. Retrace l'historique des visualisations de réseaux sociaux et montre que les représentations visuelles peuvent être un outil efficace pour illustrer des concepts tels que les acteurs centraux ou les groupes sociaux.

Le très grand avantage des diagrammes nœud-lien est leur intuitivité : la grande majorité des lecteurs peut les comprendre. En revanche, qu'ils soient dessinés manuellement ou générés automatiquement, leur lisibilité dépend totalement du placement des nœuds dans le plan. Ce problème épineux a d'ailleurs donné naissance à un domaine de recherche à part entière nommé le dessin de graphes (graph drawing).

Exploration de réseau et passage à l'échelle :

Lorsque le réseau à représenter contient de nombreux nœuds et de nombreux liens entre ces nœuds, dans ce cas, il y aura un problème récurrent : le graphe de liens de nœuds est converti en un ensemble de lignes et de points difficiles voire impossibles à transformer en une représentation lisible, ni manuel ni automatique.

Il existe une technologie distincte à résoudre ce problème :

• **Réduisez la quantité d'informations représentées par le filtrage ou l'agrégation**

Il y a quelques Méthode d'échantillonnage réseau ou grappe de calcul (groupe Éléments similaires), qui peuvent ensuite être agrégés en un élément représentatif groupé. On parle ici de clustering, dans la méthode de data mining utilisée en créant des groupes d'éléments dans de grands réseaux, il permet d'analyser ces grands réseaux.

• **Fournir des représentations interactives pour explorer l'ensemble du réseau**

Plusieurs outils la visualisation ne représente qu'une partie du réseau et suggère la navigation

Explorez le reste du réseau de manière interactive.

Les stratégies varient : certaines sont concentrez-vous sur des nœuds spécifiques et permettez la navigation en suivant la connexion, d'autres reposent principalement sur le filtrage interactif ou permettent naviguez dans le réseau convergé.

• **Utilisez une autre notation pour le graphe de liens à 11 nœuds**

Cette stratégie Il s'agit d'utiliser différentes métaphores visuelles pour représenter le web.

L'idée est de trouver des représentations qui permettent "d'augmenter l'espace visuel",

Permet d'afficher plus d'informations de manière plus lisible.

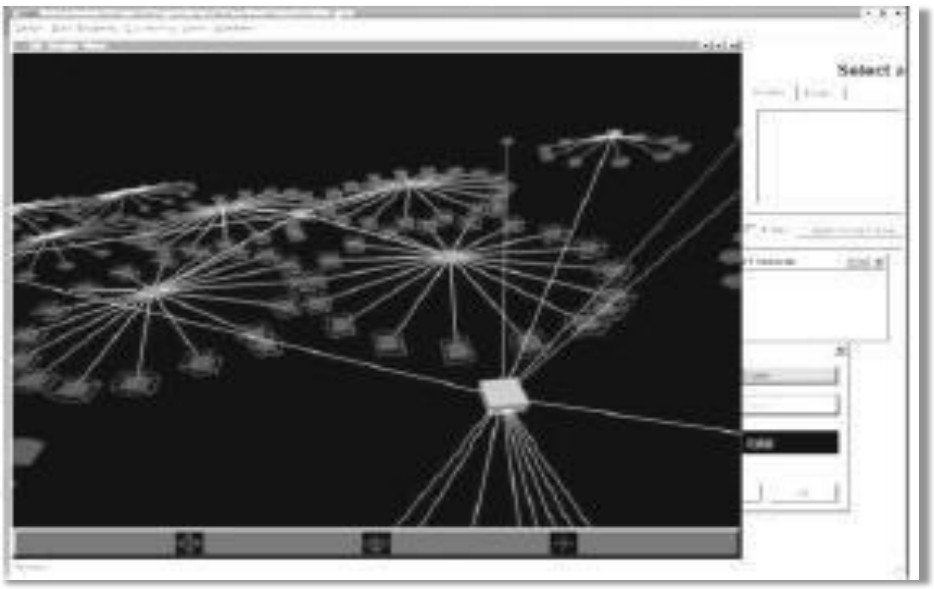


Figure (I-6) : Représentation d'un réseau en 3D

1.4.2. Représentation matricielle :

Les graphes ont deux représentations canoniques : les diagrammes nœud-lien et les matrices d'adjacence. Une matrice d'adjacence représente chaque sommet d'un réseau à la fois comme une ligne et comme une colonne. Si deux sommets sont connectés, la case correspondant à l'intersection de la ligne et de la colonne est marquée. Traditionnellement, on utilise une valeur numérique (0 marquant l'absence de connexion, 1 marquant la présence), soit par une marque graphique comme dans la figure 5. Le tableau 1 liste les principaux avantages et inconvénients des représentations matricielles par rapport aux diagrammes nœud-lien.

Avantages	Inconvénients
Absence de superpositions des nœuds (ce qui permet de pouvoir toujours lire les étiquettes portées par les nœuds)	Taille de l'espace visuel requis à un niveau de détail équivalent plus important que le diagramme nœud-lien
Absence de croisements des liens (ce qui permet de toujours identifier la source et destination des connexions)	Difficulté à effectuer des tâches de suivi de chemin (suivre les liens de Pierre à Paul passant par Marie)
Facilité avec laquelle il est possible d'identifier les absences de connexions	Manque de familiarité, les matrices paraissent moins intuitives que les diagrammes nœud-lien

Tableau (I-1) : Avantages et inconvénients de la représentation matricielle

Utiliser la représentation matricielle :

Deux facteurs principaux entrent en jeu lors qu'il s'agit d'utiliser les représentations matricielles pour explorer de grands réseaux :

2 être capable de réordonner leurs lignes et colonnes.

3 pouvoir naviguer dans des matrices de grande taille.

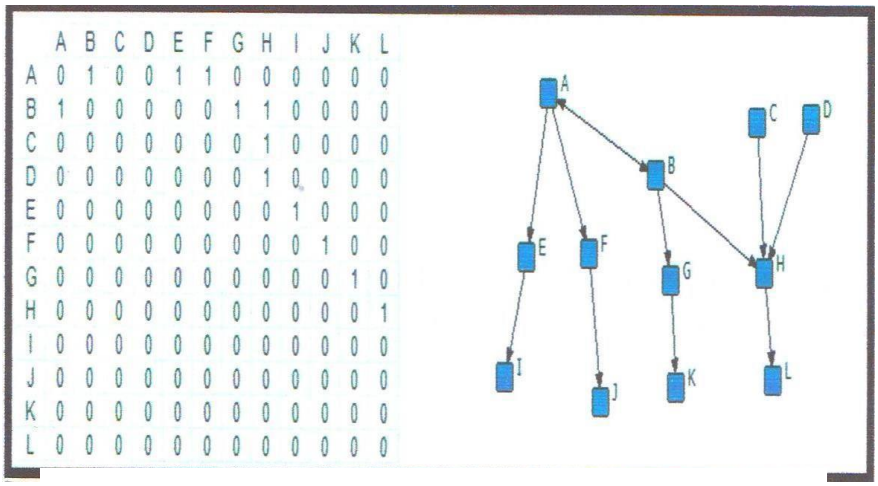
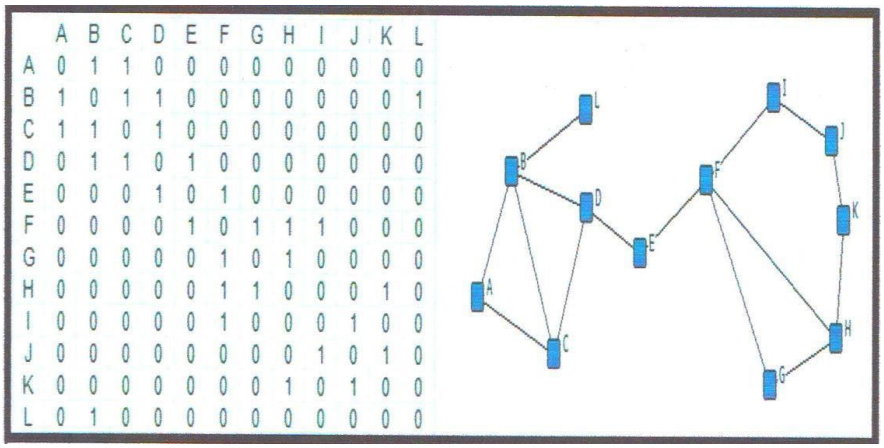


Figure (I-7): Matrice binaire avec une représentation par graphe orienté



	A	B	C	D	E	F	G	H	I	J	K	L
A	0	1	1	0	0	0	0	0	0	0	0	0
B	1	0	1	1	0	0	0	0	0	0	0	1
C	1	1	0	1	0	0	0	0	0	0	0	0
D	0	1	1	0	1	0	0	0	0	0	0	0
E	0	0	0	1	0	1	0	0	0	0	0	0
F	0	0	0	0	1	0	1	1	1	0	0	0
G	0	0	0	0	0	1	0	1	0	0	0	0
H	0	0	0	0	0	1	1	0	0	0	1	0
I	0	0	0	0	0	1	0	0	0	1	0	0
J	0	0	0	0	0	0	0	0	1	0	1	0
K	0	0	0	0	0	0	0	1	0	1	0	0
L	0	1	0	0	0	0	0	0	0	0	0	0

Figure (I-8) : Matrice binaire avec une représentation par un graphe non-orienté

2. Quelques logiciels de visualisation des réseaux sociaux

2.1. Gephi : est un logiciel gratuit d'analyse et de visualisation de réseau développé en Java est basée sur la plate-forme NetBeans.

Gephi est particulièrement utilisé pour les projets de recherche scientifique et de journalisme

Les données. Par exemple, il est utilisé pour visualiser la connectivité globale du contenu new-yorkais Times, Analyser les activités en réponse aux événements sur le réseau Twitter, et l'analyse des réseaux traditionnels.

Gephi a été sélectionné pour le Google Summer of Code pendant cinq années consécutives depuis 2009 et 2013.

En 2010, il a remporté le Duke's Choice Award d'Oracle dans la catégorie innovation.

Visualisation des données techniques. Le logiciel et ses algorithmes inspirés de Linkedln

Créez des InMaps et utilisez-les pour la visualisation du réseau de Truthy

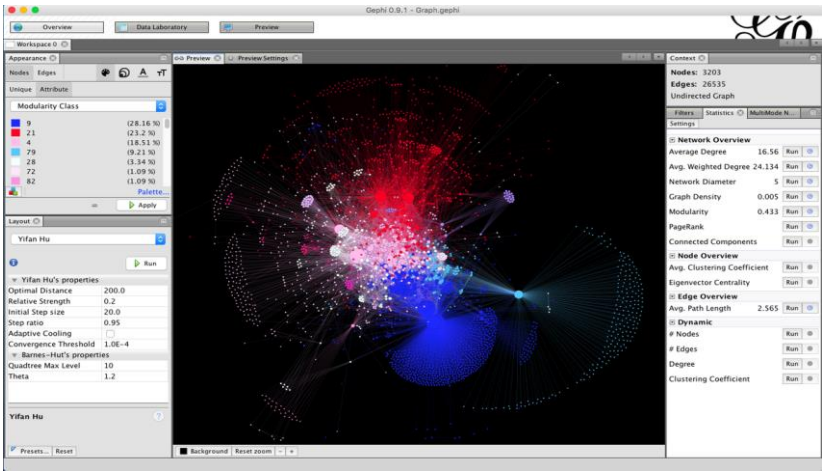


Figure (I-9) : Exemple visuel d'un réseau social sous Gephi

2.2. **Cytoscape** : Ce logiciel est bien évidemment adapté à tous types de réseaux (simple ou non, dichotomique ou monomode, valeur ou booléen), incluant plusieurs graphiques ou réseau à plusieurs niveaux. Depuis sa sortie en 2002, version 0.8, le logiciel a été développé en Amérique du Nord (États-Unis et Canada) mais a bénéficié d'un soutien plus large (Institut surtout Pasteur), qui est maintenant la version 2.8.3. Ils peuvent être développés par n'importe qui utiliser l'API ouverte Cytoscape basée sur la technologie JavaTM et encourager le développement de la communauté d'applications. La plupart des applications sont disponible gratuitement sur l'App Store Cytoscape.

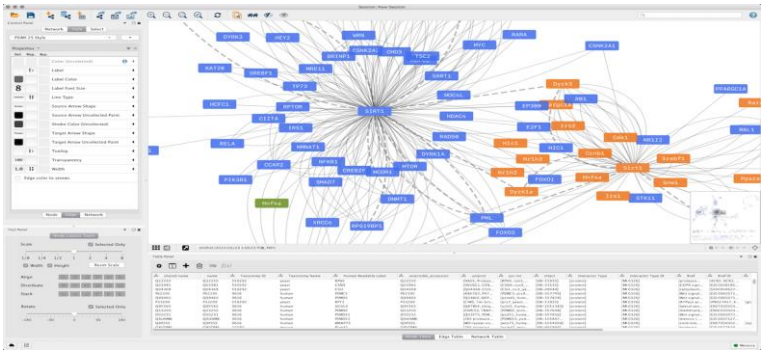


Figure (I-10) : Exemple visuel d'un réseau social sous Cytoscape

2.3. **Tulip** : Tulip v.2.0.4 est un logiciel qui a été écrit et est toujours mis à jour régulièrement

Mise à jour par l'Université de Bordeaux 1. Le logiciel dispose d'une licence GPL. Il permet d'éditer

Graphiques et effectuez quelques opérations dessus.

En utilisant ce logiciel, les utilisateurs peuvent créer des sommets et des arêtes orientées connectez-les. Le nombre de sommets est limité à 500 000. Concernant la version de la figure, nous vous pouvez utiliser les arêtes pour afficher ou non les arêtes, et vous pouvez même utiliser les flèches pour afficher leurs flèches.

Une fois que vous avez modifié les graphiques requis, vous pouvez tester différentes fonctionnalités :

• Si c'est simple

• Si c'est un arbre

• S'il est acyclique

• Si connecté

• S'il s'agit de 2 connexions

• S'il s'agit de 3 connexions

• S'il est plat

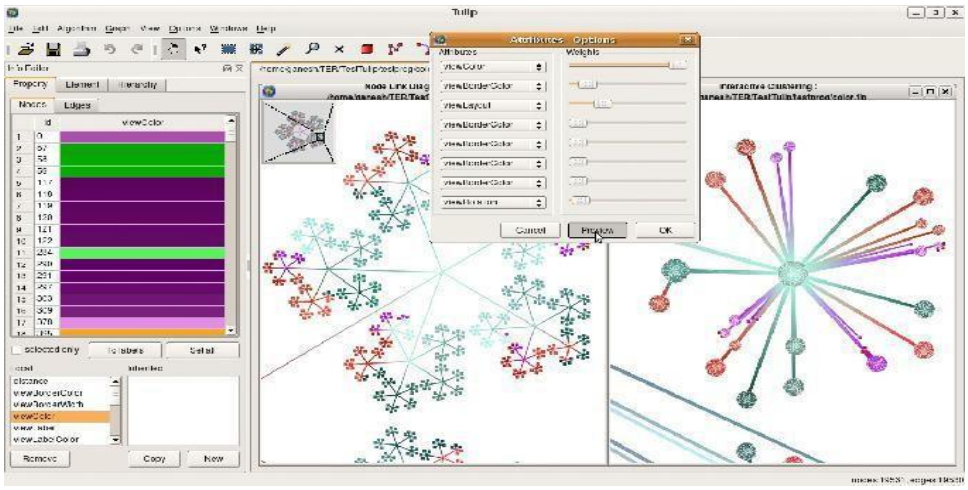


Figure (I-11) : Exemple de visualisation d'un graphe par le logiciel Tulip

Conclusion

Dans ce chapitre nous avons défini les concepts d'analyse des réseaux sociaux ; son développement historique. Par la suite nous avons également présenté une brève description sur la représentation et la visualisation des réseaux sociaux, ainsi que quelques exemples de logiciels de représentation et de visualisation des réseaux sociaux.

Cependant, les réseaux sociaux sont des axes de recherche très vaste et très large. Ils prennent l'attention de plusieurs chercheurs en raison de leur importance pour rapprocher les peuples. En plus ils sont joue un rôle très important dans la vie quotidienne de la vie moderne dans le monde. Ils aident aussi à la diffusion et la propagation de l'information entre les gens de façon plus rapide que les outils traditionnels.

Introduction

L'analyse des réseaux sociaux (ARS) est une méthode de visualisation et de modélisation des relations sociales sous forme de nœuds (individus, organisations) et de liens (relations entre ces nœuds). Il peut observer et calculer le degré, la force ou la densité des liens entre les participants du réseau. Ainsi, l'analyse des réseaux sociaux repose sur une approche structurée des relations entre les membres d'un environnement social organisé.il permet de décrire l'interdépendance entre acteurs, ayant ainsi une « représentation simplifiée d'un système social complexe ».

Selon les recherches de Divjak et Peharda l'ARS est définie comme la cartographie et la mesure des relations et des flux entre les personnes, les groupes, les organisations ou toute autre entité de traitement de l'information. Les nœuds du réseau sont des personnes ou des groupes, et les liens montrent la relation ou le flux entre les nœuds. Il permet une analyse visuelle et mathématique des relations interpersonnelles. L'analyse des réseaux sociaux peut être définie comme un moyen de collecter des données à partir des réseaux sociaux pour générer des informations stratégiques et des indicateurs de développement. **[13]**

1. Développement historique

L'ARS a une histoire de près de 70 ans. Elle peut être divisée en trois périodes principales : La base de la méthode ; l'élaboration de la méthode ; le développement actuel. Les fondations de ces différents bâtiments ont été construites entre les années 1940 et 1960. Dans les années 1960 et 1970, des études méthodologiques ont été développées pour assurer une mise en œuvre stricte. Des années 1980 à aujourd'hui, ils ont été révisés et améliorés, conduisant à la découverte de nouvelles approches. L'analyse du réseau repose sur deux cadres théoriques et méthodologiques. **[14]**

- **L'aspect théorique** repose sur un large éventail de concepts de structure sociale et un grand nombre d'études empiriques, qui montrent que le comportement des individus est lié à la structure dans laquelle ils se trouvent. La sociométrie a également contribué au développement de l'analyse des réseaux sociaux. L'apport empirique de la mesure sociale est en partie attribué aux travaux de Moreno, qui est considéré comme l'un des pionniers de l'analyse des réseaux et de la psychologie sociale. Enfin, l'analyse de réseau repose également sur l'apport des mathématiques

aux sciences sociales : dans le développement théorique de l'analyse de réseau, les chercheurs ont découvert l'utilisation de modèles mathématiques.

- **L'aspect méthodologique** Il évoque l'utilisation de types de données quantitatives et qualitatives par les chercheurs, l'analyse et le traitement de ces données. En 1957, Elisabeth Bott publie ses recherches sur le système des relations familiales. Elle suppose que le degré de séparation des rôles entre mari et femme évolue dans le même sens que la densité des réseaux sociaux familiaux ; c'est-à-dire que dans les réseaux sociaux où les membres sont étroitement liés, la séparation évidente des tâches ménagères selon le sexe tend à être plus élevée. À ce jour, l'hypothèse de Bot est toujours valable et personne ne la réfute. De son côté, StanleyMilgram a mis en place en 1967 un dispositif d'investigation expérimentale, qui reste une référence pour la recherche du petit monde. Il a essayé de calculer le nombre moyen de liens qui sépare une personne de toute autre personne sur la planète. **[15]**

De nos jours, il existe de nombreux sujets de recherche en analyse de réseau, famille, relation de travail, amitié, etc. Cette méthode est actuellement également utilisée à des fins autres que la recherche scientifique, les consultants en relations professionnelles ou à des fins commerciales, telles que les projets FOAF (Friends of Friends). Les réseaux sociaux prennent de plus en plus d'importance dans notre vie quotidienne, et l'idée de les analyser pour en extraire des informations offre sans aucun doute des avantages importants dans de nombreux domaines. **[16]**

2. Définition de l’analyse des réseaux sociaux

De nos jours, il existe de nombreux sujets de recherche en analyse de réseau, famille, relation de travail, amitié, etc. Cette méthode est actuellement également utilisée à des fins autres que la recherche scientifique, les consultants en relations professionnelles ou à des fins commerciales, telles que les projets FOAF (Friends of Friends).

Les réseaux sociaux prennent de plus en plus d'importance dans notre vie quotidienne, et l'idée de les analyser pour en extraire des informations offre sans aucun doute des avantages importants dans de nombreux domaines. Cette simplification du dispositif d'interdépendance est volontaire, car l'analyse des réseaux sociaux se veut une « technique

d'exploration et de caractérisation ». Par ailleurs, l'analyse des réseaux sociaux s'intègre dans une réflexion plus large sur la sociologie. La compréhension de la structure des groupes sociaux repose sur l'étude des relations entre les membres de l'environnement social.

Cette analyse dite structurelle passe notamment par la description et l'analyse des différents schémas relationnels possibles : interdépendance des membres, réciprocité des relations, position centrale de certaines personnes, absence de relations qui créent des « trous » relationnels dans le réseau, fréquence Relation (forte relation et relation faible). L'avantage de l'analyse structurelle est qu'elle permet d'exprimer de manière simplifiée la complexité et la diversité des relations entre les participants.

La modélisation des systèmes interdépendants prend en compte l'imbrication progressive des acteurs dans la « forme » de la structure, qui évolue, se rétrécit ou s'agrandit selon les activités de ses membres. La plasticité du réseau est plus importante en raison de son absence de frontières claire.

3. **Utilisations de l'analyse des réseaux sociaux :**

La grande quantité d'informations stockées dans les réseaux sociaux peut être très utile Il peut être utilisé pour suivre la marque de son entreprise et fournir des informations sur: les opinions publiques sur les produits, la compréhension des conditions actuelles du marché, les entreprises compétitives et l'accès à de nouvelles stratégies marketing, grâce à ses données qui permettront à terme d'établir une comparaison des produits existants sur le marché, assurer une meilleure gestion de la gamme de produits, offrir un meilleur accompagnement aux clients, et être en mesure de suivre les influenceurs**. [17]**

Ces données peuvent également être utilisées lors d'élections pour solliciter l'opinion publique. Dans la suite de cette section, nous présentons les méthodes utilisées pour analyser les réseaux sociaux, puis comparons leurs méthodes, dans le but de révéler les inconvénients de chaque réseau social. Déterminez quelle personne est la mieux placée pour analyser les réseaux sociaux d'aujourd'hui

4. **Méthodes d'analyse des réseaux sociaux :**

Tous ces réseaux sociaux collectent beaucoup de données: amis, messages, Photos, fréquence d'utilisation, etc. Toutes ces communications et informations sont soigneusement

conçues écris le. Cela soulève la question de savoir comment utiliser beaucoup d'informations. Il a besoin de tout d'abord, modélisez le réseau sous forme mathématique. La structure de base est bien sûr Graphiques: l'analyse des graphiques générés peut extraire beaucoup d'informations et Prévoir partiellement le développement futur du réseau.

Mis à jour pour fournir Méthode d'analyse flexible, tenant compte de toutes les informations dans la structure Et le contenu. Ensuite, on peut distinguer deux grandes familles: traditionnelle et data mining **[18]**

4.1. Méthodes traditionnelles (classiques) :

Dans cette section, nous détaillons les mesures utilisées dans l'analyse traditionnelle Réseau social. Les mesures locales sont des indicateurs qui prêtent attention aux informations locales D'autre part, la mesure globale fournit des informations sur l'ensemble du processus Structure (réseau).

a. Les mesures locales :

La mesure s'effectue localement : en cas de conformité, veuillez-vous assurer que votre environnement linguistique est correct. Dans ce festival, nous 21 Actes Introduction Concepts La commune danoise : Centralisation et prestige Central **[19]**

- ✓ **Centralité** : informations de base importante et droite implicites des acteurs Les acteurs indépendants ont un large éventail de visages. Dans un réseau, on peut définir un participant Centralcom qui est impliqué dans plusieurs liens

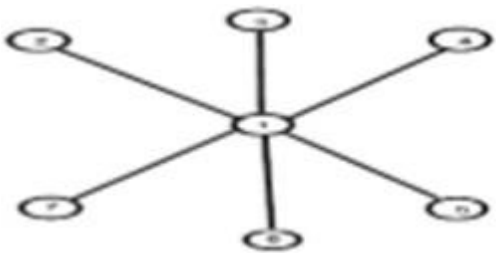


Figure (II-12) : réseau sociaux centre

Nous avons remarqué dans la figure que l'acteur 1 est l'acteur principal, car il / elle communiquait avec la plupart des autres acteurs et il est apparu sur les 6 chemins les plus courts reliant les 6 autres acteurs.

Il existe différents types de liens entre nœuds (participants), ce qui permet de générer plusieurs types de centralité, notamment:

- **Degré de Centralité**: Soit n le nombre total de nœuds (participants) dans le réseau (graphe). La centralité de l'acteur i notée CD (i) est le degré (le nombre d'arêtes) du nœud d'acteur noté d (i) normalisé au degré maximum n-1.

$$CD(i) = d(i) / (n-l) \quad (1)$$

- **Centralité de proximité:** Dans cette méthode, la centralité est définie par le concept de proximité ou de distance. Si je suis l'acteur principal, alors il peut facilement interagir avec d'autres acteurs. Par conséquent, la distance entre lui et les autres doit être courte. Nous utilisons donc la distance la plus courte pour calculer cette mesure.

- **Prestige:** Le concept de prestige est une autre façon de mesurer l'importance d'un nœud. Une sorte de Le nœud prestigieux est le nœud auquel de nombreux autres nœuds sont connectés- il en reçoit un grand nombre de liens entrants. On distingue les mesures suivantes:

  - **Degré de prestige**: dI (i) est le degré d'entrée du nœud i, et n est le nombre de nœuds Réseau, le degré de prestige est donné par la relation

$$PD(i) = dI(i) / (n-1)$$

  - **Prestige de tri**: les mesures proposées jusqu'à présent sont basées sur des liens Entrée et sortie d'un acteur donné. La mesure du prestige du classement prend en compte La réputation et l'importance des acteurs choisissent les acteurs i. Ce type d'algorithme est Utilisé pour trier les résultats de recherche, tels que Google (algorithme de page de classement)

b. Les mesures globales :

Afin de comprendre la structure globale du réseau, quelques mesures globales ont été formulées.

• Densité P du graphe G : Cette mesure peut exprimer le degré de connectivité dans le graphe G qui représente le réseau **[20].** C'est le nombre de liens dans le graphe G, normalisé par le nombre maximum de liens dans le graphe.

• La distance géodésique entre deux nœuds est le chemin le plus court entre les deux nœuds.

• La distance moyenne du graphe connecté est égale à la moyenne des distances géodésiques entre tous les participants.

• Le diamètre du graphe connecté est l'excentricité maximale qui peut exister entre deux de ses nœuds.

5. Fouille de données dans les réseaux sociaux :

Appliquée à la réalité sociale, la fouille de données s'avère être un outil très riche et puissant. La recherche sur les réseaux sociaux se concentre sur des méthodes graphiques détaillées qui permettent l'analyse de l'interaction entre les individus et le degré de connexion (l'arc de la somme). L'exploration de données a fourni de puissantes méthodes d'analyse de données pour les réseaux sociaux (graphiques d'interaction) pour découvrir leurs caractéristiques et leurs fonctions.

L'exploration de données répond à un ensemble de tâches dans le domaine scientifique en fournissant un ensemble de technologies qui répondent à ces tâches spécifiques. L'extraction des règles d'association est devenue l'une des tâches de base de la recherche. Les règles d'association sont la méthode la plus utilisée dans le domaine du marketing et de la distribution. Leur principe est de trouver le groupe d'articles le plus fréquent et de générer des règles d'association. Ces règles sont faciles à comprendre et comportent des probabilités, ce qui en fait un outil agréable qui peut être utilisé directement par les utilisateurs professionnels.

Le data mining sur les réseaux sociaux représente une véritable mine d'or d'informations sur des sujets marketing. Dans cette partie, nous étudierons comment les entreprises utilisent ces informations pour trouver de nouveaux clients ou cibler des entités influentes pour mettre en place des stratégies marketing dites « virale ». On distingue le terme « **DataMining** » et le terme « Extraction de conaissance », même si ces termes sont utilisés pour définir la découverte du savoir due à l'abus de langage : **KDD** (Knowledge Discovery in Databases). L'extraction de connaissances est effectuée en raison du processus spécial d'exploration de données, il semble donc qu'il s'agisse d'une méthode d'extraction. L'exploration de données peut être définie comme un processus d'exploration de données visant à découvrir des relations et des faits nouveaux et significatifs sur de grands ensembles de données.

6. Les tâches de la fouille de données :

Une multitude de problèmes d'ordre intellectuel, économique ou commercial peuvent être regroupés, dans leur formalisation, dans l'une des tâches suivantes :

1.1. La classification :

Est une méthode supervisée qui consiste à définir une fonction qui attribue une ou plusieurs classes à chaque donnée. Dans cette approche on suppose qu'un expert fournit auparavant les étiquettes pour chaque donnée, les étiquettes sont des classes d'appartenance. Selon [Govaert, 2003] : « (la classification supervisée (appelée aussi classement ou classification inductive) a pour objectif « d'apprendre » par l'exemple. Elle cherche à expliquer et à prédire l'appartenance de documents à des classes connues a priori. Ainsi c'est l'ensemble des techniques qui visent à deviner l'appartenance d'un individu à une classe en s'aidant uniquement des valeurs qu'il prend)» Dans le cas d'extraction de données à partir des réseaux sociaux, cette méthode nous permettra de regrouper les intérêts d'une personne par classe (personnage, produit, achat, consommation, activité...etc.). D'une autre façon nous pouvons dire que cette méthode descriptive permet de décrire de façon simple une réalité complexe en la résumant.

1.1.1. Le clustering

Le travail du clustering consiste à regrouper les données en classe ; nous obtenons par ce biais une forte similarité intra-classe et une faible similarité inter-classe. Il est possible d'utiliser un nombre important d'algorithmes classiques de détection de groupes basés sur les procédés suivants :

- Partitionnement de graphe (par coupures minimales) mais cela suppose de connaître à l'avance le nombre de partitions à réaliser.
- Clustering hiérarchique mais il s'agit d'une méthode en général coûteuse en O (n 2 log n) et qui n'est bien sûr pas adaptée aux structures n'ayant pas de hiérarchie à la base
- Clustering en partitions (k-means) mais comme pour les procédés de partitionnement de graphe, le nombre de partitions est à définir à l'avance ;
- clustering spectral utilisant les représentations matricielles des graphes ainsi que leurs propriétés (valeurs propres).

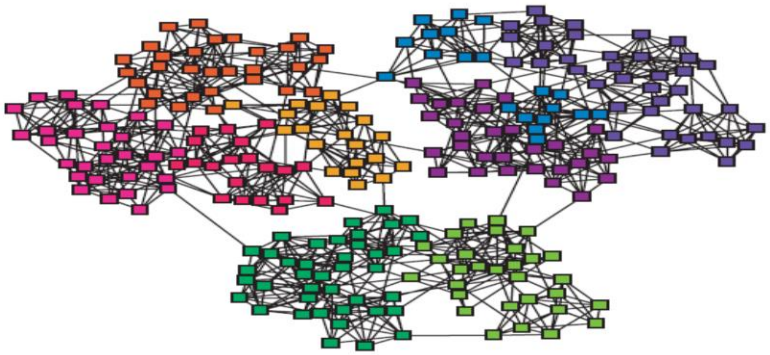


Figure (II-13) : Représentation des clusters dans réseau social.

1.1.2. Biclustering :

Le biclustering est une méthode d'apprentissage non supervisée pour détecter simultanément des groupes informatifs d'objets et d'attributs qui a été proposée par Hartigan (1972) avec le terme « clustering direct ». Le nom « biclustering » a été utilisé par Mirkin (1996). D'autres noms tels que co-clustering, block clustering, two-mode clustering, twoway clustering et simultané clustering sont également utilisés.

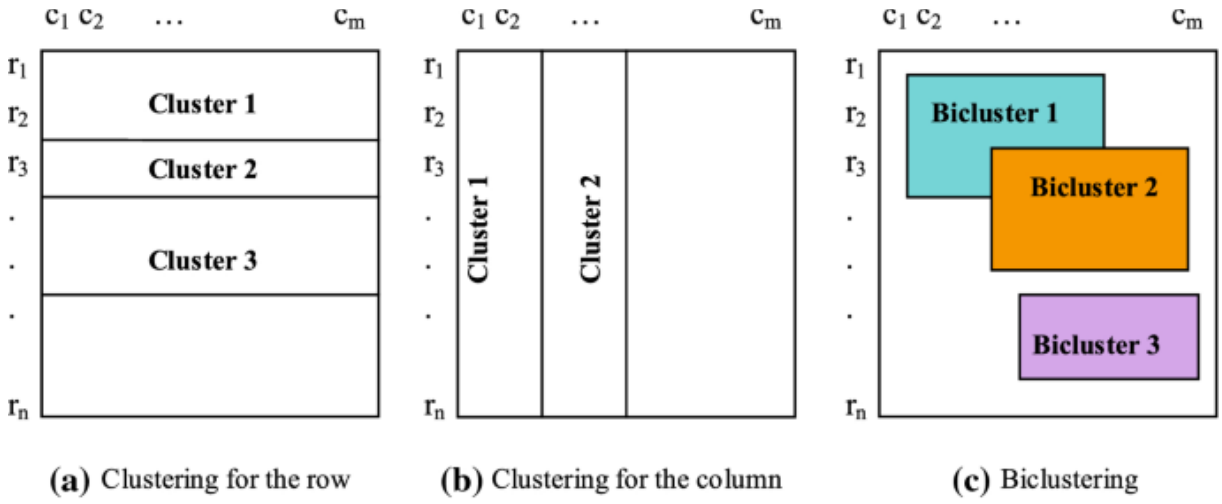


Figure (II-14) Différence entre la solution de clustering traditionnelle et solution de biclustering

Chaque méthode de biclustering suppose des structures et des types de données spécifiques. La figure (15) illustre certaines structures de bicluster définies par Madeira et Oliveira (2004) :

(a) bicluster unique.

(b) biclusters à rangées exclusives.

(c) biclusters non chevauchants avec structure arborescente.

(d) biclusters superposés placés arbitrairement.

Les biclusters peuvent être divisés en un type de valeur continue et un type de valeur catégorielle selon les types de données. Une valeur d'élément du *je* l'objet et le *j* L'attribut dans un bicluster de type valeur continue peut être modélisé comme un modèle additif

$$Zij = \mu + \alpha je + \beta j + \varepsilon je$$

Où μ désigne la moyenne globale en bicluster, αje et β j désignent les ajustements pour le *je* l'objet et le *je* attribut, respectivement, et εje représente l'erreur aléatoire.

Dans cet article, seuls les modèles additifs sont traités car un Le modèle multiplicatif peut être transformé en modèle additif par transformation logarithmique

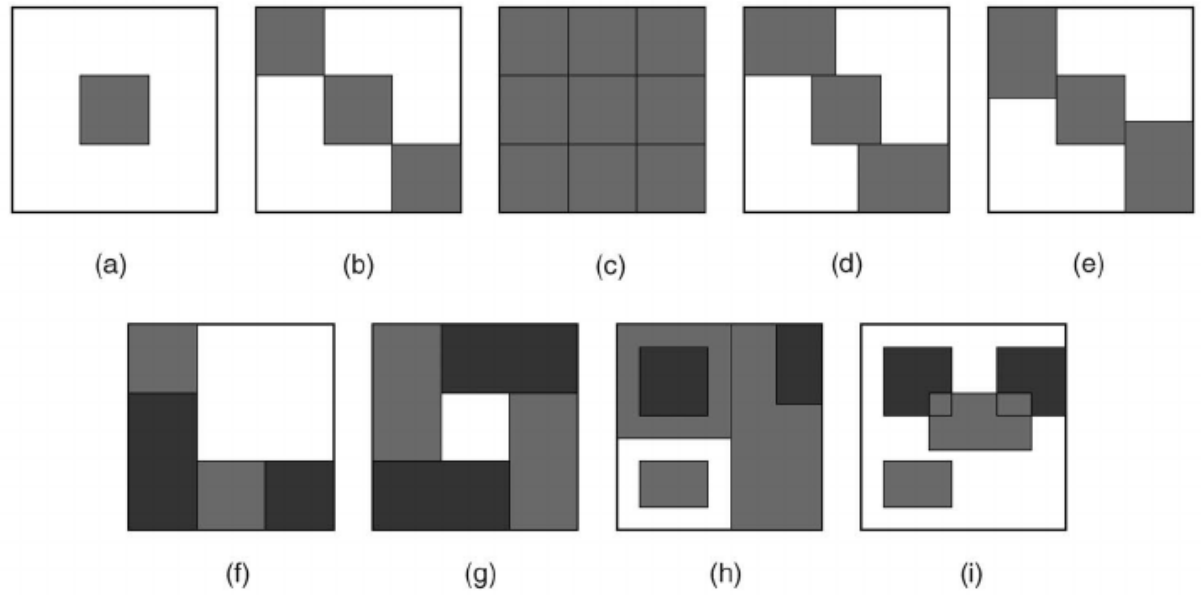


Figure (II-15) : Quelque structure de biclustering

1,0	1,0	1,0	1,0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0

(une)

1,0	1,0	1,0	1,0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0

(b)

1,0	2.0	3.0	4.0
1,0	2.0	3.0	4.0
1,0	2.0	3.0	4.0
1,0	2.0	3.0	4.0

(c)

1,0	2.0	3.0	4.0
2.0	3.0	4.0	5.0
3.0	4.0	5.0	6.0
4.0	5.0	6.0	7.0

(ré)

Figure (II-16) : Quatre types de bicluster

De plus, les biclusters basés sur l'additif modèle peut-être classé en quatre types selon les valeurs de αje et β j comme le montre la figure (16) :

(a) constante valeurs bicluster ($\alpha je = , 0, \beta j = 0$) ,

(b) bicluster à lignes constantes bicluster ($\alpha je \neq 0, \beta j = 0$)

(c) (ré) colonnes-constantes valeurs-cohérentes ($\alpha je = 0, \beta j \neq 0$) ($\alpha \neq 0, \beta \neq 0$) **. [21]**

Algorithmes :

Le but des algorithmes de classification double est de trouver, s'il existe, le plus grand « bi-cluster » contenu dans une matrice, en maximisant une fonction objectif. On peut prendre comme fonction, avec les notations adoptées ci-dessus :

$$f_1 = |I| + |J| \quad \text{ou} \quad f_2 = |I| * |J|$$

Function: biclust

The main function of the package is

Biclust (data, method=BCxxx(),number,...)

With:

- ✓ data: The preprocessed data matrix
- ✓ method: The algorithm used (E. g. BCCC() for CC)
- ✓ number: The maximum number of bicluster to search for
- ✓ ... : Additional parameters of the algorithms

Returns an object of class Biclust for uniform treatment.

De nombreux algorithmes ont été développés notamment par la bio-informatique, dont :

- Bimax(Barkow et al. (2006)): Groups with ones in binary matrix
- CC (Cheng and Church (2000)): Constant values
- Plaid (Turner et al. (2005)): Constant values over rows or columns
- Spectral (Kluger et al. (2003)): Coherent values over rows and columns
- Xmotifs (Murali and Kasif (2003)): Coherent correlation over rows and columns

BicatYeast

Matrice de données de puces à ADN pour 80 expériences avec l'organisme Saccharomyces Cerevisiae extraites de l'ensemble de données d'exemple BicAT

```
> data(BicatYeast)
>x <-discretize(BicatYeast)
> Xmotif <-biclust(x, method=BCXmotifs(), number=50, alpha=0.05, + nd=20, ns=2)
> Xmotif
An object of class Biclust
Call:
        biclust(x = x, method = BCXmotifs(), number = 50, alpha = 0.05)
 Number of Clusters found: 15
 First Cluster size:
                        Number of Rows: 175
                        Number of Columns: 6
```

BCXmotifs :

Les biclusters aux évolutions cohérentes sont représentés par l'algorithme Xmotifs de Murali et Kasif (2003). Cet algorithme recherche les lignes avec une constante valeurs sur un ensemble de colonnes. **[22]**

Usage

##S4 method for signature 'matrix, BCXmotifs' biclust(x, method=BCXmotifs(), ns=10, nd=10, sd=5, alpha=0.05, number=100)

Arguments

X Data Matrix.

Method Here BCXmotifs, to perform Xmotifs algorithm

Ns Number of columns choosen.

Nd Number of repetitions.

Sd Sample size in repetitions.

Alpha Scaling factor for column result.

Number Number of bicluster to be found.

Value

Returns an object of class Biclust.

Extends

Class "BiclustMethod", directly.

Fonction de BCXmotifs :

Input : Expression Matrix EM; Thersholds ɣ

Output : List of Biclusters *L*

Preprocess the missing values of *EM*

List *L* = Ø

Bicluster *B*

Repeat n times

B = EM

B_γ = multiple node deletion phase (B, ɣ)

B'_γ = simple node deletion phase (B_γ, ɣ)

B''_γ = addition phase (B'_γ)

$L = L \oplus B''_\gamma$

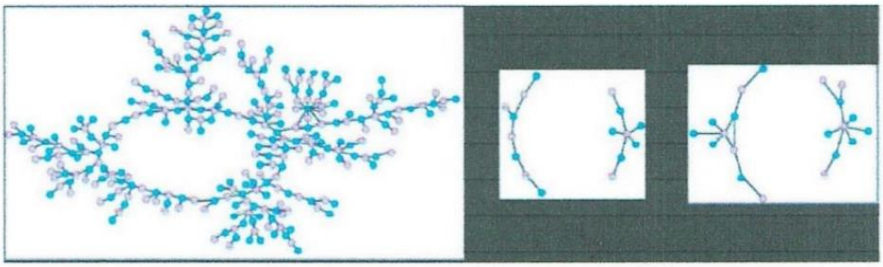
Substitution phase (B''_γ, *EM)*

End_repeat

Return L

1.2. La recherche des motifs fréquents :

Comme troisième type d'application des techniques de la fouille de données dans les réseaux sociaux, baptisé la recherche des motifs fréquents pour la recherche des sous-graphes fréquents dans les réseaux.



Figure(II.17) : Représentation des motifs fréquents dans les réseaux

2. Quelques travaux utilisant les techniques Data Mining dans l'analyse des réseaux sociaux :

Nom de l'auteur	Titre de papier	Les techniques	Année
Gauri Joshi1, Mr. SamadhanSonawane	Filtrage et Classification des utilisateurs sur la base de données de médias sociaux à l'aide de méthodes Bayetic et Memetic	Algorithme de classificateur multi-étiquettes Naive Bayes et algorithme Memetic	**2015**
A Sharma, MK Sharma, RK Dwived	Analyse documentaire et défis des techniques d'exploration de données pour l'analyse de réseaux sociaux	Data Mining Techniques	**2017**
	EXPLOITATION DES DONNÉES	K-Means Clustering	

PoojaSikka	DES RÉSEAUX SOCIAUX À L'AIDE DECLUSTERING BASED-SVM	Based SVM (KMCB-SVM)	**2015**
Thai Le, Phillip Pardo and William Claster	Application du réseau de neurones artificiels à l'analyse de données sur les médias Sociaux	Réseau de neurones artificiels (ANN)	**2016**
R.Adaikkalam and Dr. A. Shaik Abdul Khadir	Une enquête sur les techniques d'exploration dedonnées pour l'analyse de réseau social	Data Mining Technique	**2016**
Kanika Mathur	Réseau social en ligne	Naive Bayes Text Classifier	**2016**

Tableau (II.2) : présente les différentes publications analyse des réseaux sociaux

3. Les logiciels d'analyse des réseaux sociaux :

De nombreux logiciels qui existent dans le domaine d'analyse des réseaux sociaux. Nous nous présentons dans cette partie quelques logiciels qui ont été développés, et qui ont les plus utilisés par les chercheurs.

1.1. Pajek :

Un logiciel d'analyse et de visualisation de réseaux pour Windows. Développé par deux chercheurs slovènes, Vladimir Batagelj et Andrej Mrvar, il a le grand mérite d'être gratuit et puissant, ce qui lui a permis de s'imposer comme un des outils les plus utilisés en analyse des réseaux.

L'interface de Pajek:

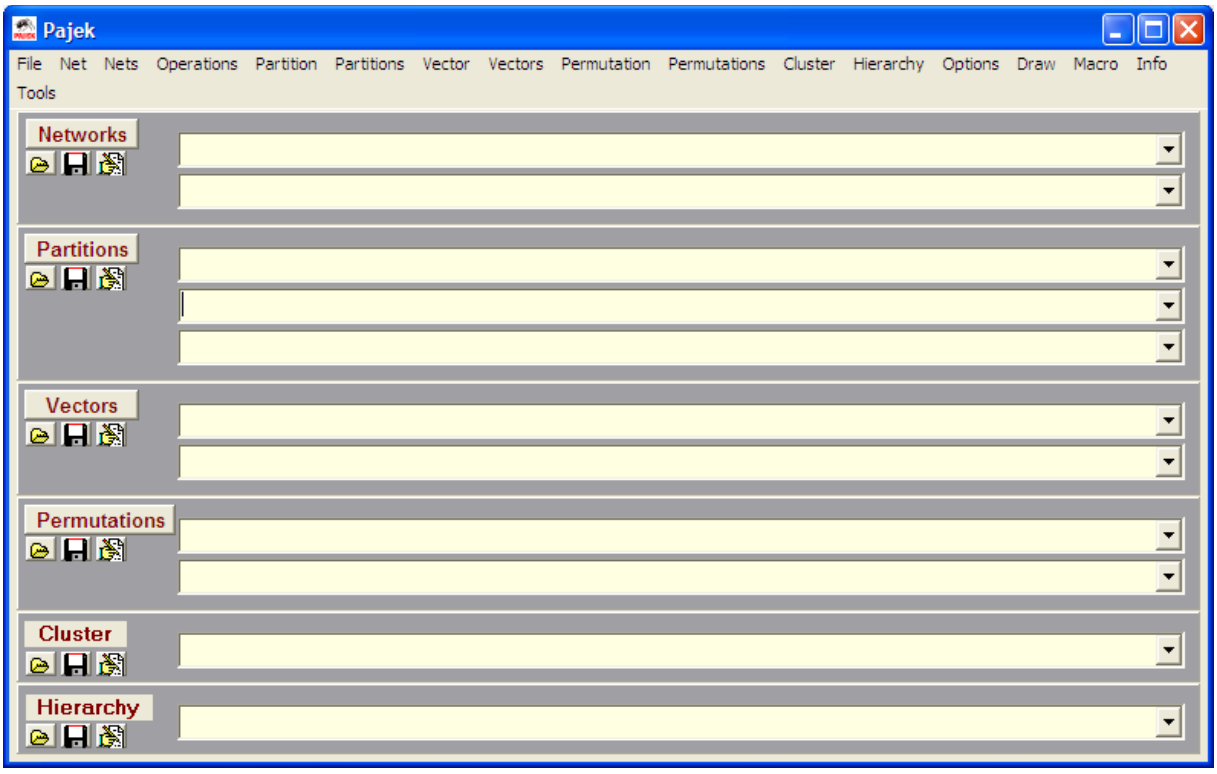


Figure (II.18) : L'interface de logiciel Pajek.

La Visualisation des réseaux sous Pajek: La visualisation des graphes, des partitions et/ou des vecteurs se représente dans la figure suivante : fait en sélectionnant « Draw » dans le men u « Draw » dans la fenêtre principale de Pajek

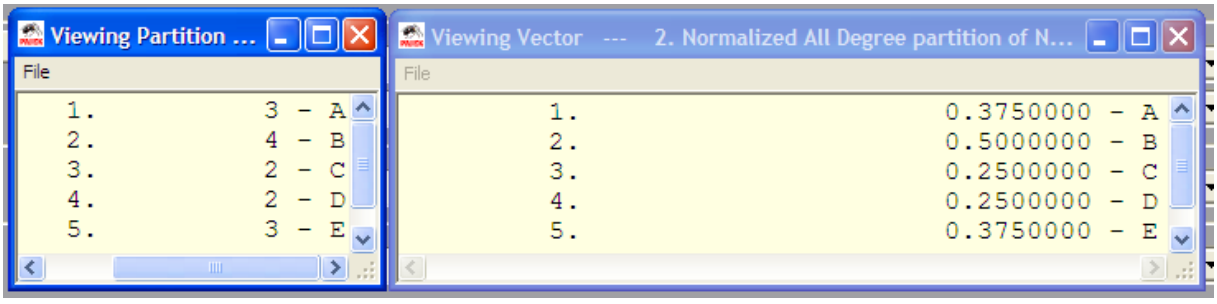


Figure (II.19) La visualisation d'un graphe sous Pajek.

1.2. Le Langage R :

Le logiciel R est un logiciel de statistique créé par Ross Ihaka & Robert Gentleman. Il est à la fois un langage informatique et un environnement de travail : les commandes sont exécutées grâce à des instructions codées dans un langage relativement simple, les résultats sont affichés sous forme de texte et les graphiques sont visualisés directement dans une fenêtre qui leur est propre. C'est un clone du logiciel S-plus qui est fond´e sur le langage de programmation orient´e objet S, développé par AT&T Bell Laboratoires en 1988. Ce logiciel sert à manipuler des données, à tracer des graphiques et à faire des analyses statistiques sur ces données.

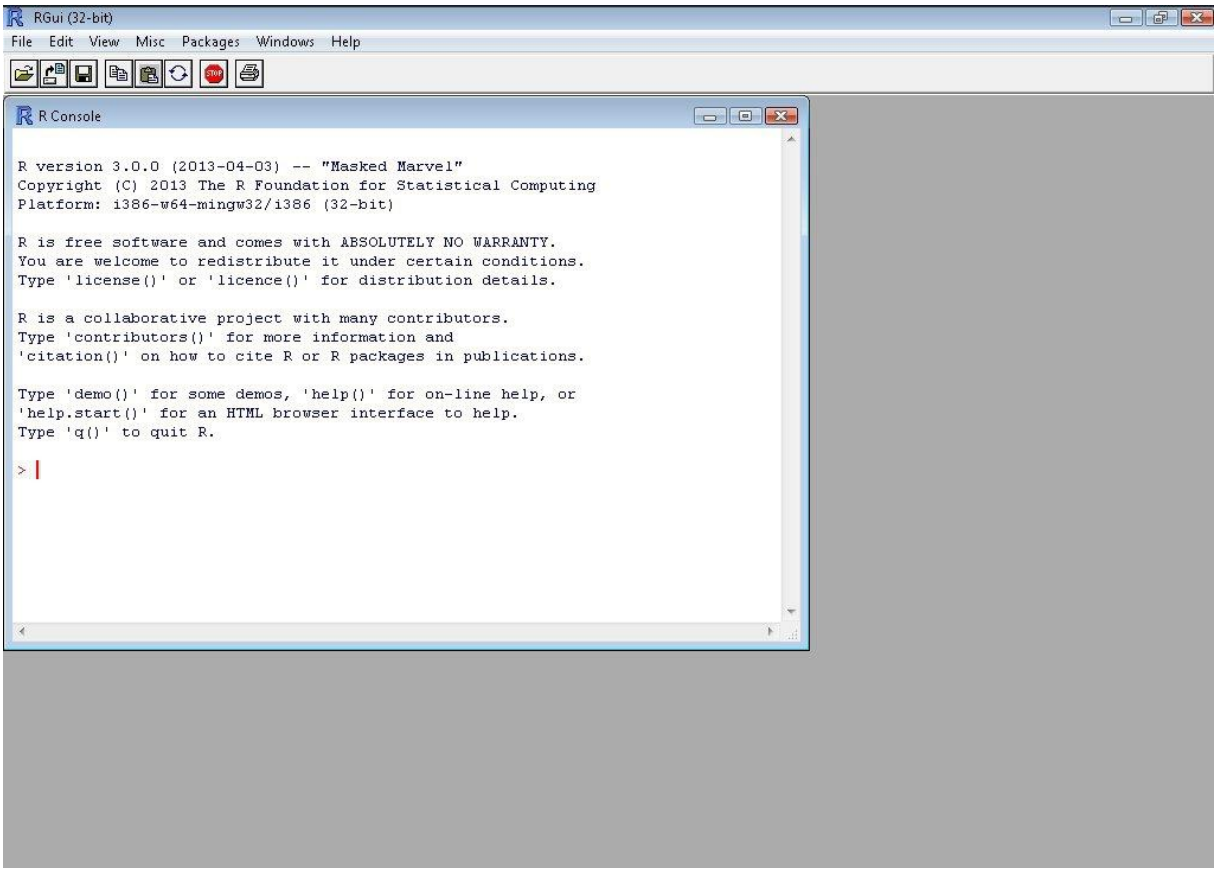


Figure (II.20) Interface graphique R

Conclusion :

Dans ce chapitre, nous avons abordé l'analyse des réseaux sociaux, ses origines et sa définition, après on a présenté les deux grands familles de l'analyse des réseaux et la méthode biclustering. Aussi de quelques logiciels d'analyse les plus populaires.

L'analyse des réseaux sociaux devenus aujourd'hui une nouvelle technique pour la socialisation de l'homme. Est donc tous des outils permettant de visualiser et modéliser les relations sociales. Alors, la force des méthodes d'analyse des réseaux réside dans sa capacité à représenter de façon simplifiée la complexité et la diversité des relations entre acteurs.

Introduction

L'extraction des règles d'association est devenue l'une des tâches les plus importantes aujourd'hui. L'exploration de données populaire en raison des travaux d'Agrawal et al. Éthique analytique le panier des femmes au foyer est l'une des applications de l'extraction des règles d'association. Allié L'objectif est d'identifier les relations compréhensibles entre les attributs de la base de données. Dans L'extraction des règles d'association de notre étude de cas permet d'analyser le comportement Un utilisateur d'un réseau social appelé **gowalla** pour comprendre ses habitudes de déplacement, Leur loisir et leur consommation, c'est pourquoi dans ce chapitre nous présentons ces processus en détail Extraire les règles d'association du contexte d'exploration de données.

1. Les regels d'associations :

Les règles d'association sont des règles extraites de la base de données des transactions (ensemble d'éléments), qui décrivent l'association entre certains éléments. Cette technologie permet de mettre en évidence le lien entre les produits de base (produits de base, produits que les clients voyagent) et les produits complémentaires. Il est possible de formuler une stratégie commerciale visant le développement et une stratégie commerciale visant à augmenter et à augmenter. Ces algorithmes peuvent résoudre le problème dit de comptage d'ensembles fréquents (FSC). Les règles d'association sont une forme de mappage ➔ Y ou X et Y sont des ensembles d'éléments disjoints. Les règles d'association ne transforment que la cooccurrence plutôt que la causalité. La force des règles de l'association se mesure à son soutien et à sa confiance

Support, s (X7 Y) = (8(XUY)/N).

Confiance, c (f-7Y) = (8(XUY)/8(X)).

1.1. Quelque définition :

Item, item set, item fréquent set : Dans la base de données, un item est un couple (attribut, valeur), son attribut a un nom, le nom doit porter une valeur, et cette dernière doit appartenir à un domaine clairement défini. Un groupe d'éléments constitue un groupe d'éléments ; si et seulement si la prise en charge est supérieure ou égale au seuil minimum fixé par l'utilisateur, l'ensemble d'éléments est dit fréquent. Les ensembles d'éléments sans support suffisant sont considérés comme non pertinents, Les symboles suivants peuvent être utilisés :

- **D**: Ensemble des transactions, une transaction notée Ti (Ti est l'Ième transaction)
- **Lk**: Ensemble des itemsets fréquents de taille k.
- **Ck**: Ensemble des itemset candidats de taille k.
- **k-itemsets**: les items set détaillé k

SUPPORT : indicateur de la « fiabilité » d'une règle ; le support est important car les règles faibles ne peuvent être observées que par hasard. Le support est généralement utilisé pour éliminer les règles inintéressantes. Mesuré par le pourcentage de transactions avec A et B

$\boldsymbol{Supp} = (\boldsymbol{A} \rightarrow \boldsymbol{B}) = \boldsymbol{m}\ |\boldsymbol{T}|$

$\boldsymbol{|T|}$: est le nombre total de transaction de la base de données

M : le nombre de transaction où (A et B) apparaissent en même temps dans la même transaction du point de vue probabiliste, chaque sous-ensemble d'items se voit associé l'événement selon lequel la transaction contient les items de ce sous- ensemble. Le support s'exprime donc par la Probabilité de réaliser simultanément les événements A et B : $Supp = (A \rightarrow B) = (A \cap B) = P(B/A) \times P(A)$ ou (A) événement «la transaction contient tous les items de l'ensemble A» et B est l'événement «la transaction contient tous les items de l'ensemble B ».

CONFIANCE : indicateur de la « précision » d'une règle ; la confiance mesure la pertinence des inférences faites par une règle. Couplé à la confiance élevée de XY, la probabilité d'observer Y avec X est élevée. Estimation de probabilité de confiance donnée condition Y sait autrement : transaction de confiance $condition$ & $ésulta$) le rapport entre le nombre de $freq$ ($condition$) =

mc (**4**) pointe vers la confiance égale à la probabilité que l'événement B se produise DE savoir A Réalisé : $Conf$= (A->B) = P (B /A) =($\boldsymbol{A} \cap \boldsymbol{B}$) $\boldsymbol{P}(\boldsymbol{A})$ = $Supp(A \rightarrow B)$ $Supp(A \rightarrow B)$

LIFT : Lift représente le rapport de l'indice d'origine P(AB) de la règle à l'indice d'indépendance entre les parties constituant la règle d'association, qui est égal au produit des deux PA) x P (B). La règle (A-> B)), l'ascenseur peut simplement comprendre sa "distance" par rapport à l'indépendance. Sa formule de calcul est (ascenseur (A-> B) = P (AB) / P (A) x P (B)). Si la promotion d'une telle règle est inférieure ou égale à 1, alors la règle n'a pas de sens. Par exemple, la règle d'un lift égale à 2 (A-> B) indique qu'un individu avec l'attribut A est deux fois plus susceptible d'avoir l'attribut B qu'un individu moyen. Cette mesure est symétrique, donc les règles (A-> B) et (B -> A) ne peuvent pas être distinguées.

1.2. Le principe des règles d'association :

L'association consiste à savoir quelles valeurs des variables sont assemblées. Par exemple, une certaine valeur d'une variable coïncide avec la valeur d'une autre variable. La forme de la règle d'association est : si d'abord, alors le résultat. Aucune variable cible n'est définie pour l'association. Toutes les variables peuvent être des variables prédictives et des variables cibles. Ce type d'analyse est également appelé « analyse d'affinité ». Intérêt : meilleure compréhension des comportements.

1.3. Problèmes de règles d'association :

L'un des plus grands défis de l'exploration de bases de données est la vitesse et l'efficacité du développement d'algorithmes capables de gérer de grandes quantités de données. Comme la plupart des algorithmes d'exploration l'estiment, le nombre total de bases de données et généralement de très grandes bases de données.

• Si la base de données est volumineuse, l'extraction de schéma (plus ou moins) peut être numériquement coûteuse.

• Certaines associations peuvent être fausses ou intéressantes, elles arrivent juste par accident.

1.4. Domaine d'application

Les règles d'association s'appliquent à de nombreux domaines, tels que les sociétés commerciales (Sociétés commerciales), car les commerçants sont intéressés par l'analyse des données pour mieux connaitre le comportement d'achat de leurs clients, et s'appliquent également à d'autres domaines tels que la bio-informatique (génétique), le diagnostic médical et web mining .

2. Processus d'extraction des règles d'association :

Le processus d'extraction des règles d'association s'effectue généralement en quatre étapes ; la figure illustre l'enchaînement de ces étapes.

2.1. Sélection et préparation des données : les données utilisées par l'algorithme d'extraction des règles d'association doivent d'abord passer par deux étapes : la première étape peut sélectionner des données dans la base de données, et dans certains cas, considérer le choix de l'utilisateur comme réduire le nombre de données traitement La taille est utilisée pour assurer l'efficacité de l'algorithme, et la seconde est utilisée pour la conversion de ces données.

2.2. **Trouver des itemsets fréquents** : Cette étape est la plus coûteuse en termes de temps d'exécution, car le nombre d'items fréquents est lié de manière exponentielle au nombre d'éléments traités. Pour N éléments, nous avons 2N ensembles d'éléments fréquents potentiels.

2.3. **Génération de règles d'association** : Afin de générer des règles d'association, nous sélectionnons le seuil de support minimum (Minsup) et le seuil de confiance minimum (Minconf). Acceptez uniquement les règles dont les niveaux de prise en charge et de confiance dépassent les seuils spécifiés. Ce type d'opération utilisant ces métriques est un problème exponentiel avec la taille de l'ensemble d'éléments fréquents.

3. Induction des règles associatives :

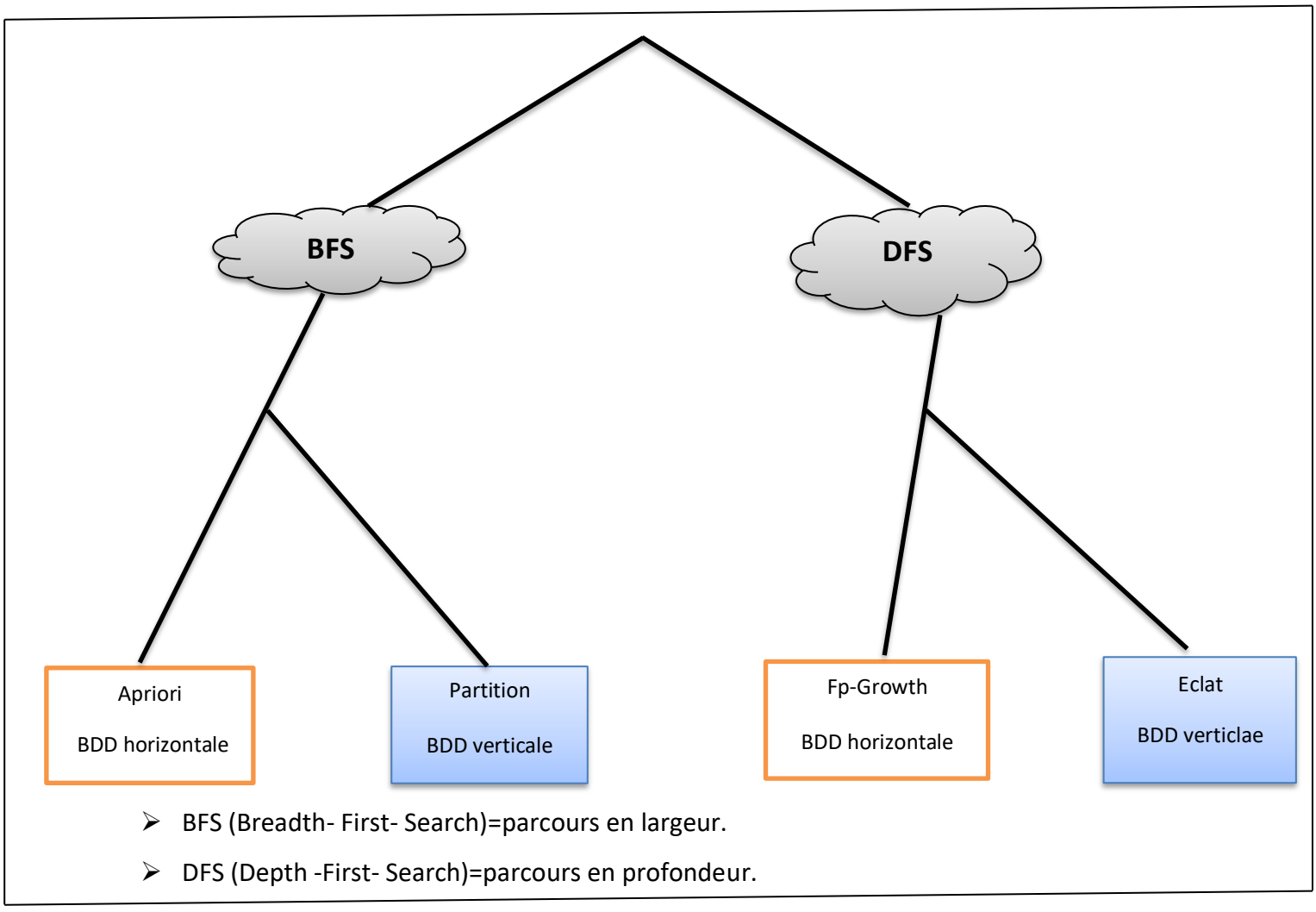


Figure (III-21) : Les algorithmes de frequant Set Counting (FSC)

L'algorithme suivant utilise les concepts de support et de confiance pour déterminer la pertinence des associations.

Si la prise en charge de la règle d'association générée est supérieure à minSupp et que la confiance est supérieure à minConf, alors la règle d'association générée est conservée. Ces deux constantes, dépendantes de la base de données, sont définies empiriquement par les utilisateurs du système.

3.1. Apriori :

Apriori est un algorithme de recherche de règles d'association classique. Comme tous les algorithmes de découverte d'association, il est applicable aux bases de données de transactions (enregistrements de transaction). Afin de révéler la pertinence des règles, nous utilisons deux concepts, à savoir le soutien et la confiance. Pour être retenue, chaque règle doit avoir un support supérieur à minSupp et une confiance supérieure à minConf. Ces deux valeurs sont définies empiriquement par l'utilisateur du système.

L'algorithme commence par une liste des produits les plus fréquents dans la base de données qui satisfont au seuil de prise en charge. Générez un ensemble de règles (règles candidats) à partir de cette liste. Les candidats sont testés sur la base de données (on cherche des exemples de règles générées et leur apparence), et les candidats qui ne respectent pas minSupp et minConf sont retirés. L'algorithme répète ce processus et augmente la taille du candidat d'une unité à chaque fois qu'il trouve une règle pertinente. Enfin, fusionnez les ensembles de règles découverts. La sélection des candidats se fait en deux étapes :

1) La jointure.
2) L'élagage.

La jointure consiste en un ensemble de règles d'éléments k-1 sur elle-même, ce qui se traduit par un ensemble de candidats k-éléments. Enfin, l'élagage supprime les candidats qui n'ont pas au moins l'une des sous-chaînes d'éléments k-1 dans l'ensemble de règles d'éléments k-1. L'algorithme est très puissant, mais si l'ensemble d'éléments fréquents est trop grand, il sera affecté. De plus, le modèle d'analyse rapide et répétée d'une base de données donnée devient un obstacle aux performances des grandes bases de données.

3.2. Le principe Apriori :

Si un groupe d'éléments est fréquent, alors tous ses sous-ensembles sont également fréquents. Inversement, si l'ensemble {a, b} n'est pas commun, alors tous les éléments contenants {a, b} sont définis. Donc tant que l'on sait que {a, b} est rare, on peut éliminer tous les ensembles le contenant a priori.

Cette stratégie est appelée élagage basé sur les médias. Un attribut clé des métriques de support rend cette stratégie possible : le support d'un ensemble de projets ne sera jamais plus

grand que le support de sa collection. Définition (monotonie) Soit I un ensemble d'items, J = 21 puissance de l'ensemble I une mesure f est une monotone si vx, Y E J: X ç Y f(X) < f(Y)

Une mesure f est anti monotone si vx, Y E J : X ç Y f(X) > f(Y)

Toute mesure admettant une propriété d'anti-monotonie peut être intégrée dans un algorithme de recherche d'ensemble d'items fréquents.

3.3. Algorithme Apriori :

L'algorithme Apriori est un algorithme d'exploration de données conçu en 1994, par Rakesh Agrawal et Ramakrishnan Sikrant et le premier algorithme de recherche de règle d'association incluant des étapes d'élagage pour tenir compte de la croissance exponentielle du nombre de l'ensemble d'items candidats

Input : D : l'ensemble des transactions

Minsup : seuil minimum de support

Output :

L : les Itemsets fréquents.

Algorithme

1 L1 = {1-itemsets fréquents)

2 k=2 ;

3 **Tant que** Lk-1 non vide **faire**

4 Ck=Apriori-Gen (Lk) ;

5 **Pour** chaquet de B **faire**

6 Ct = Subset (Ck, t) ; {les candidats contenus dans Ck}

7 **Pour** chaque c de Ct **faire**

8 c.count++ ;

9 **Fin pour**

10 **Fin pour**

11 Lk = {c de Ct / c.count>= minsup};

12 k++ ;

13 **Fin du tant que**

14 Return ULk;

Pour que l'algorithme trouve les itemsets fréquents, il analyse la base de données plusieurs fois. Dès le premier passage, l'algorithme génère des itemsets fréquents représentés par L1. Afin de générer des itemsets fréquents à partir de k égal à deux, le processus est réalisé en trois étapes :

- Étape 1 : L'ensemble d'items fréquents (k-1) trouvé lors du passage (k-1) est utilisé pour générer un ensemble d'items candidats de taille k (algorithme APRIORI-GEN).
- Étape 2 : utilisez la fonction Sous-ensemble qui fournit tous les ensembles d'éléments Ck inclus dans la transaction pour rechercher dans la base de données afin de calculer la fréquence des candidats à chaque fois.
- Étape 3 : Enfin, nous choisissons la fréquence supérieure ou égale au seuil minsup.

3.4. Exemple d'utilisation de l'algorithme Apriori :

1temsets fréquent par Apriori

Afin de valider l'algorithme, nous allons l'appliquer sur une table de transactions. La table Représente un ensemble de quatre transactions contenant cinq Items respectivement (A, B, C, D, E).

Ti	**Items**				
100	**A**	**B**	**C**	**D**	**-**
200	**-**	**B**	**C**	**-**	**E**
300	**A**	**B**	**C**	**-**	**E**
400	**-**	**B**	**-**	**-**	**E**

Tableau (III-3) : Ensemble de transaction

Nous fixons le seuil minimum de support à 2. (**Minsup** =2) c'est-à-dire 50%. Le déroulement des étapes de l'algorithme est schématisé comme ci-après :

La première étape L1 L généré les itemset fréquent on calcule les supports des items pour K =1

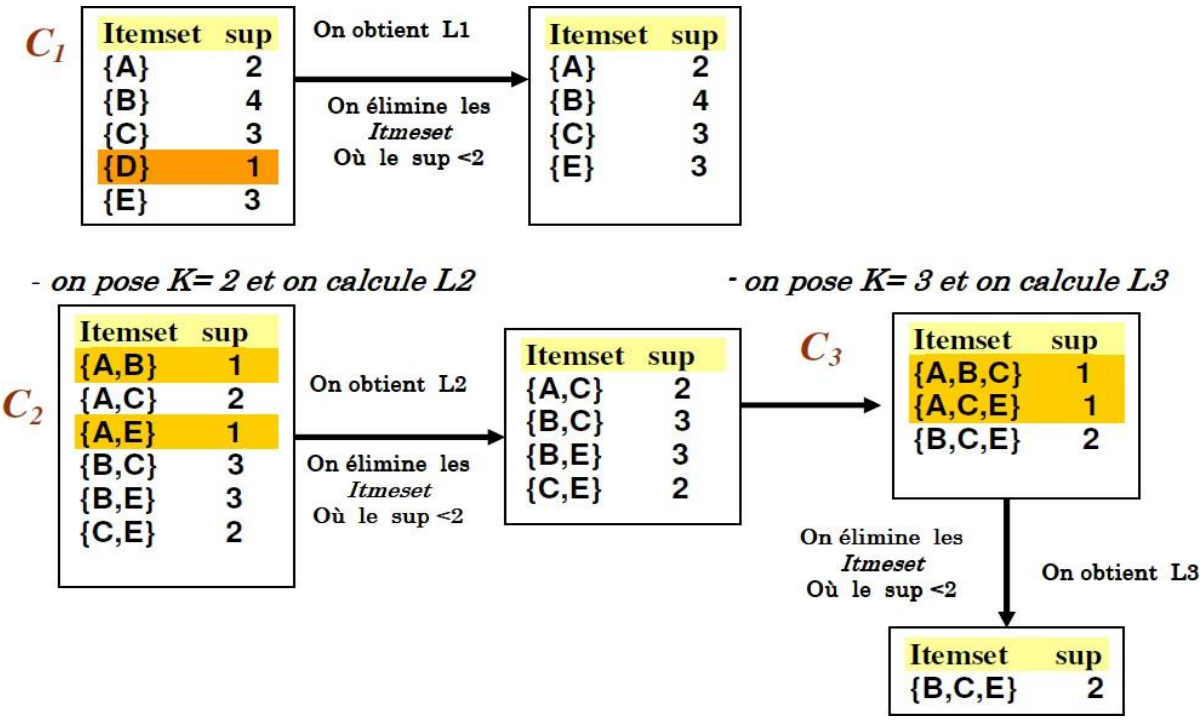


Figure (III-22) : Les défirent itemsets fréquents génère par APRIORI

Règles associative générés :

Pour la génération des règles d'association, nous fixions le seuil minimum de la confiance par exemple à 50%, les règles extraites par Apriori sont décrites par la table

règles	Support	Confiance	règles	Support	Confiance	règles	Support	Confiance
A→C	50%	100%	E→C	50%	66.67%	C,E→B	50%	100%
C→A	50%	66.67%	C→E	50%	66.67%	E→C,B	50%	66.67%
B→C	75%	75%	B→C,E	50%	50%			
C→B	75%	100%	B,C→E	50%	100%			
B→E	75%	75%	B,E→C	50%	66.67%			
E→B	75%	100%	C→B,E	50%	66.67%			

Figure (III-23) : Règles genre par APRIORI

Par exemple en calculant le lift des règles générées décrites en tables, afin de choisir des règles intéressantes. Dans cet exemple toutes les règles d'association ayant un lift ≤ 1 signifient que la règle n'est pas intéressante.

règles	Confiance	Lift	règles	Confiance	Lift	règles	Confiance	Lift
A→C	**100%**	**2**	E→C	66.67%	0,88	C,E→B	100%	1
C→A	**66.67%**	**2**	C→E	66.67%	0,88	E→C,B	66.67%	0,88
B→C	75%	1	B→C,E	50%	1			
C→B	75%	1	B,C→E	100%	0,88			
B→E	75%	1	B,E→C	66.67%	0,88			
E→B	100%	1	C→B,E	66.67%	0,88			

Figure (III-24) : Les règles générées par REGLES GENEREES PAR APRIORI

4. Problématique de l'algorithme Apriori :

Le but de la recherche de la relation entre les attributs (règles d'association) est de générer toutes les règles d'association d'intérêt, c'est-à-dire des règles avec un support et une confiance supérieurs ou égaux aux seuils minimaux **(Minsupp, Minconf)** fixés par l'utilisateur Affichage, impression, etc.) pour permettre à l'utilisateur de spécifier les règles appropriées.

L'auteur a proposé plusieurs algorithmes, mais nous nous concentrons principalement sur l'algorithme Apriori car il est à la base d'autres algorithmes.

Apriori reçoit en entrée une table d'éléments contenant un grand nombre d'enregistrements, et son objectif est de générer tous les itemsets fréquents qui dépassent le seuil (Minsup). A partir de ces ensembles ; l'algorithme génère des règles d'association avec une confiance supérieure ou égale au seuil (Minconf)

5. Evaluation d'une règle d'association

Lors de l'extraction de connaissances à partir de données, il est évident qu'à des fins décisionnelles ou organisationnelles, des règles d'association doivent être utilisées à la fin du processus d'extraction de règles, ce qui nécessite d'évaluer la qualité de la connaissance. Pour cela, nous avons besoin de suffisamment d'exemples pour vérifier cette règle, ainsi que

d'un grand nombre de contre-exemples, qui n'affectent pas le sens de cette règle dans son contexte d'extraction.

Le processus d'extraction des règles d'association peut générer des règles d'association triviales, évidentes et connues, et n'apportant pas d'informations supplémentaires

(**Exemple :** SI achat d'une café ALORS achat de sucre, etc.). Ainsi, on peut générer des règles inutiles qui sont difficiles à interpréter.

Nous avons trouvé plusieurs critères d'évaluation des règles d'association dans la littérature, qui se divisent en deux catégories. La première dite subjective (les experts en la matière savent quels attributs il souhaite que les règles d'association aient), qui sont rarement utilisées car elle viole l'objectif d'EDC. Le second s'appelle l'objectif et comprend le nombre d'exemples et de contre-exemples de recherche. Les règles qui en résultent ne doivent pas être trop générales ou trop évidentes, dans ce cas, il n'y a rien de nouveau ! Il ne doit pas non plus être trop spécifique, car s'il ne provient que de valeurs aberrantes, il n'a aucune valeur.

Dans ce travail, nous nous limitons aux standards des algorithmes de type Apriori, qui sont associés à chaque règle d'association de la forme suivante basée sur le support et la confiance : La métrique A->B permet de comprendre sa qualité Le support est une sorte En tant que mesure d'utilité, la confiance est une mesure appelée exactitude.

A la fin on obtient les meilleures règles qui ont le support et la confiance supérieure ou égale à des seuils minimaux définis par un expert en fonction de ses objectifs et du type de données traitées.

Conclusion

Ce chapitre nous permet d'explorer le concept de règles d'association. Nous avons déjà vu que le premier algorithme (Apriori) pour traiter la découverte de règles d'association a été principalement développé à des fins de marketing. À l'heure actuelle, les règles de

l'association ont retenu l'attention dans tout processus décisionnel, ce qui a conduit à certaines contraintes imposées au niveau des normes d'évaluation.

Les règles d'association sont généralement des outils efficaces pour identifier les relations entre les attributs dans la base de données. De même, ils peuvent permettre aux analystes de découvrir des connexions inattendues.

Introduction

Nous nous intéressons dans ce chapitre à deux parties principales, la première partie concerne le jeu de données que nous avons utilisé, leur source et le modèle d'analyse proposé. Dans La deuxième partie on présente l'implémentation de notre modèle, en discutant les résultats obtenus.

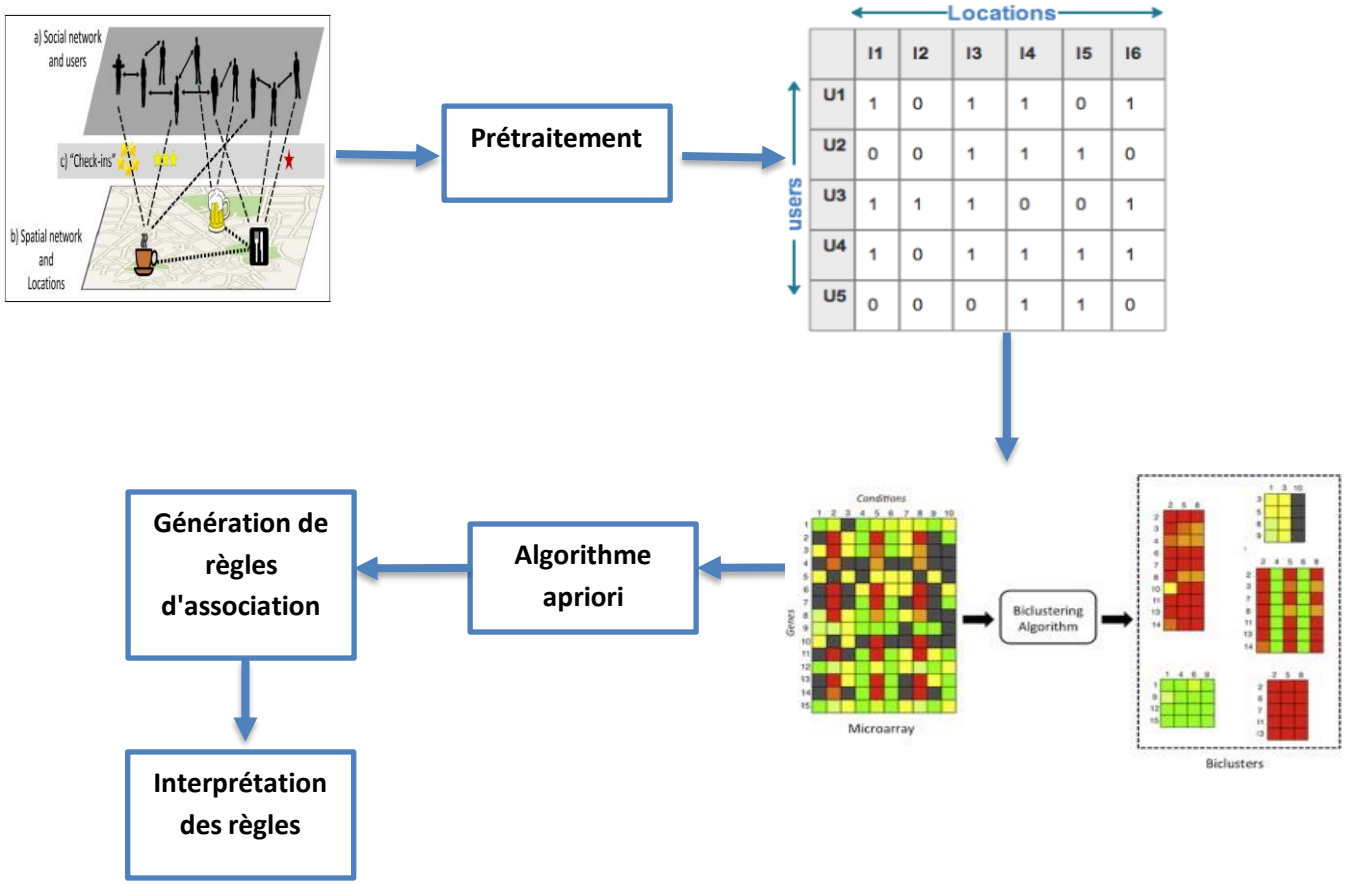
1. Architecture Globale :

La structure globale décrit les différents éléments de notre travail, leurs interrelations et interactions de manière symbolique et schématique. Ce modèle décrit comment notre méthode est conçue pour répondre aux spécifications. Cette architecture vise à prédire l'activité des utilisateurs dans les réseaux sociaux et représente nos principales réalisations. D'une manière générale, nous avons :

- En entrée, on localise la source des données (site gowalla).
- Ensuite, il y a les étapes de prétraitement et les étapes d'exploration de données.
- Dans la sortie, nous montrons une explication des règles obtenues.

L'exploration de données (règles d'association) utilise une énorme base de données pour extraire des connaissances pouvant être utilisées dans différents domaines.

Cependant, si vous choisissez un petit indice de support et de confiance, le nombre de règles générées en appliquant l'algorithme Agrawal peut être énorme. Nous pouvons réduire le nombre de règles en choisissant un grand indice de support et de confiance, mais ici nous pouvons perdre certaines règles importantes qui n'apparaissent pas dans le résultat final. Après le succès de l'algorithme Agrawal, de nombreuses améliorations ont été apportées à l'algorithme, parmi lesquelles des améliorations nous aident à filtrer et à donner de la pertinence aux règles, même celles dont les indices de confiance ne sont pas satisfaisants. Dans notre travail, nous nous concentrerons sur un exemple de ces règles. La figure suivante décrit le modèle de notre ensemble de données de recherche et la façon dont nous traitons les données.



Figure(VI.25) : Schéma fonctionnel des règles d'association basées sur l'activité de l'utilisateur

La source des données (Gowalla)

Gowalla est un service de réseaux sociaux basé sur la localisation, où les utilisateurs partagent Leurs emplacements par le check-in. Créé au début de 2009, permet aux utilisateurs d'ajouter des amis, de partager leur emplacement et d'afficher des notes ou des photos .Des utilisateurs check-in à lieux à travers une application mobile dédiée, disponible pour plates formes mobiles comme Androïde de Google, l'i Phone d'Apple et BlackBerry, ainsi que via un navigateur Web mobile. Ces dispositifs utilisent le GPS et d'autres technologies de détection pour automatiquement détecter leur emplacement Les développeurs de Gowalla fournissent une API publique pour permettre à d'autres applications d'intégrer à leur service : en particulier, ils fournissent des informations sur les profils des utilisateurs, les listes d'amis, le check-in de l'utilisateur et son lieu. Les données de Gowalla ont été collectées en 2010 et 2011. Pour chaque utilisateur, il ya l'identificateur, la liste d'amis et la liste de tous les endroits où il s'est connecté.

Nous avons téléchargé ces données à partir du site principal de Gowalla La table suivante résume l'ensemble des propriétés des données de Gowalla.

Le total des utilisateurs	**19182**
Le nombre des utilisateurs active	**8334**
Le total de spot id connecté	**30366**
Le nombre de hometown connecté	**11129**
Le total de la fréquence spot id	**357753**
Le total des types de spot	**43**

Tableau (IV.4) : l'ensemble des propriétés des données de Gowalla

Processus de prétraitement et nettoyage

La phase de prétraitement des données est souvent la plus laborieuse, elle demande le plus de temps, ceci est dû en particulier à l'absence de structuration et à la grande quantité de bruits existants dans les données brutes d'usage.

Les données que nous souhaitons analyser par techniques de data mining sont incomplètes (absence de valeurs d'attributs ou certaines caractéristiques d'intérêt), les erreurs bruyantes (contenant des valeurs aberrantes qui s'écartent de l'attendu), et incohérentes. En outre, les données Manquantes, en particulier pour les lignes avec des valeurs manquantes pour certains attributs, pourraient avoir besoin d'être déduites.

Les étapes de prétraitement sont :

- **Elimination du bruit et des erreurs** : Nous éliminons les données bruitées (c'est-à-d ayant des valeurs d'attributs incorrects). Aussi les erreurs qui se produisent à l'entrée de données. Les données qui se produisent des erreurs dans la transmission des données sont éliminées aussi.
- **Nettoyer et structurer** : Ce processus consiste à nettoyer et structurer les données existantes afin de les préparer pour une future analyse. Les données utilisées (**Gowalla**) étant souvent du texte, c'est pourquoi l'un des objectifs de cette étape est de convertir ces données et les stocker dans les champs respectifs de la base de données.

La génération de nouveaux attributs

Après l'élimination des bruits, la suppression des erreurs, et l'élimination des données n'ayant pas de sens, l'étape suivante est de procéder la transformation de ces données par le bais de requêtes de calcul, dans l'objectif de créer de nouveaux attributs (le nombre de connexion de chaque utilisateur,

la fréquence de chaque **hometown** connecté au site…) afin de mieux analyser ce genre de donnée.

Finalisation du traitement

Après le processus de La génération de nouveaux attributs, on obtient une seule table pour l'analyse qui se compose de trois colonnes (**userid, spotname, is-topspot**). Ces trois colonnes sont copiées vers l'Excel (**2019**) afin de construire une table croisée dynamique qui nous permettra la visualisation des transactions dans les lignes (Les lignes représentent l'ensemble des utilisateurs actifs), et les colonnes désignent les différents types de places **spotname**, Les valeurs de cette matrice sont les nombres de visites des places par utilisateur. **Voir Figure 25**

A	B	C	D	E	F	G	H	I	J	K	L
August	aged	inoteca, li	1 Dag Ham	1 Penn	1 Republi	1 Times S	1 World T	10 Avenue	101 Pub	103 St-Co	10th and \
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

Figure (VI.26) : La représentation de la matrice

Important :

Microsoft Access, ne permet de copier ou de transférer d'une table croisée dynamique que presque 65000 enregistrements. C'est cette contrainte qui nous a obligés de transférer les données vers Microsoft Excel. Les données résultantes de cette opération sont copiées dans un fichier en format **CSV**, qui est importé par le **logiciel R**

Le Logiciel R :

Le logiciel R est un logiciel statistique créé par Ross Ihaka et Robert Gentleman. Il est Langage informatique et environnement de travail : l'exécution de la commande est attribuée à Les instructions sont codées dans un langage relativement simple, les résultats sont affichés sous forme de texte et les graphiques sont

visualisés directement dans leur propre fenêtre. C'est un clone du logiciel S-plus basé sur un langage de programmation orienté objet Développé en 1988 par AT&T Bell Laboratories. Le logiciel est utilisé pour traiter les données, dessiner des graphiques et effectuer des analyses statistiques sur ces données.

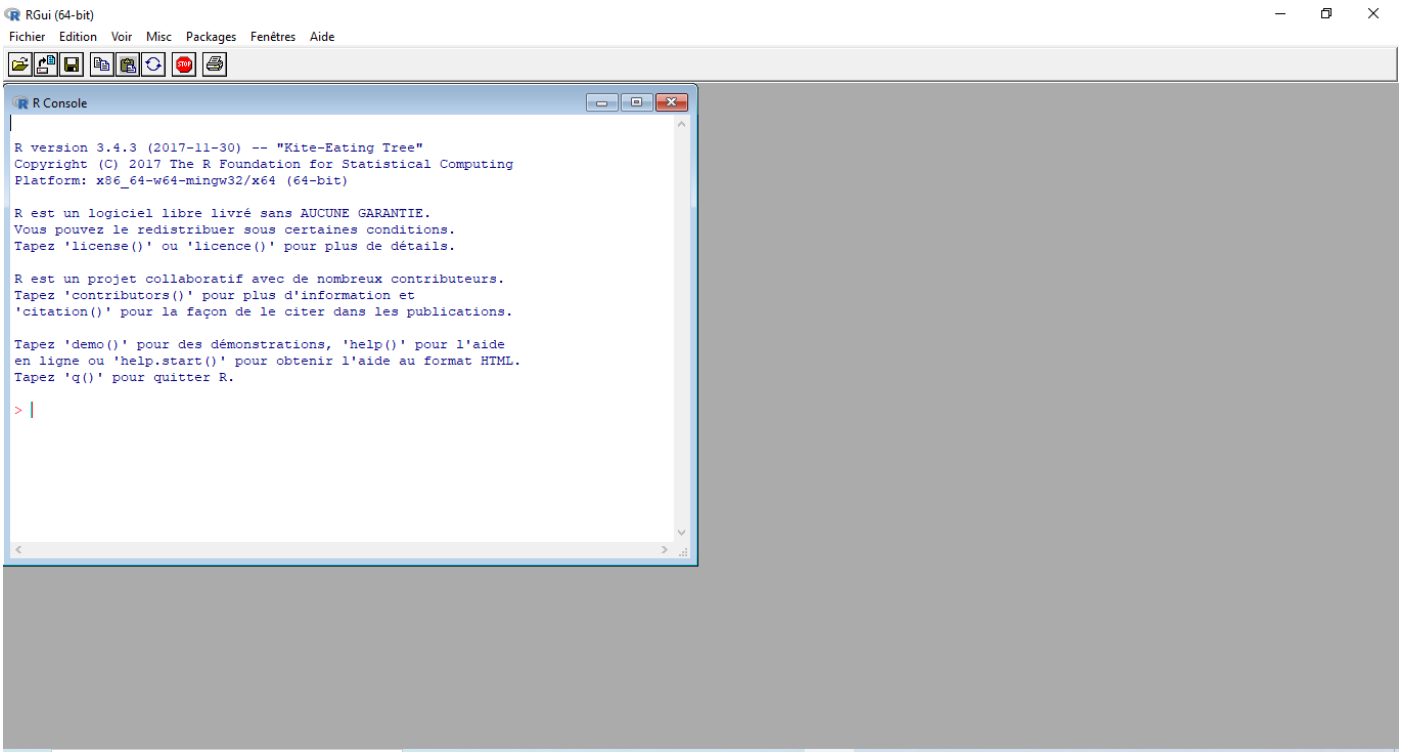


Figure (VI-27) : Interface graphique R

R et statistiques :

R est un logiciel dans lequel de nombreuses techniques statistiques modernes et classiques ont été implémentées.

Les méthodes les plus courantes pour effectuer des analyses statistiques, telles que :

- Statistiques descriptives;
- test hypothétique ;
- analyse de variance;
- Méthodes de régression linéaire (simple et multiple);
- et beaucoup plus.

- **R et les règles d'associations :**

  La recherche des **Itemsets** fréquents et des règles d'association est une approche populaire et bien documenté pour découvrir des relations intéressantes entre les variables

dans les grandes bases de données, dans R deux paquets sont mise en œuvre pour cette raison, **<arules> et <arulesViz>**.

- **Application de l'algorithme A-priori dans R :**

L'exécution de l'algorithme Apriori sur la base des données se précède toujours par la définition des paramètres suivants :

- **maxlen** : longueur de la règle (c.-à-d. le nombre d'items)
- **supp** : support de la règle.
- **Conf**=confiance de la règle

1- **Chargement du package « arules ».** La première étape consiste à installer le package. Puis, nous le chargeons en faisant appel à la commande **library(.)**

Library ("arules")

2- **Importation et transformation des données** Nous chargeons le fichier au format CSV avec **read.csv2, summary(.)** donne un aperçu rapide des caractéristiques des variables donne un aperçu rapide des caractéristiques des variables.

Regle<-read.csv ("Gowalla.csv")

3- **Application de la méthode Biclust** La première étape consiste à installer le package, **install.packages("biclust") .** Puis, nous le chargeons : **library("biclust")**

4- **Chargement de Data** Nous chargerons le data **data(BicatYeast)**

lynass<- discretize(BicatYeast)

On applique la fonction XMotifs : **res<- biclust(lynass, method=BCXmotifs(), alpha=0.05, number=9)**

```
An object of class Biclust
call:
        biclust(x = lynass, method = BCXmotifs(), alpha = 0.05, number = 9)
Number of Clusters found:  9
First  5  Cluster sizes:
                  BC 1 BC 2 BC 3 BC 4 BC 5
Number of Rows:     168   71   31   41   29
Number of Columns:    6    6   10    8    7
```

- On affecter le résultat dans un bibliothèque **: bib<-biclusternumber(res)**
  Et pour afficher les cluster : **biclusternumber(res)**

Exemple de résultat :

```
$Bicluster1
$Bicluster1$Rows
  [1]   2   9  14  18  19  24  25  26  27  29  33  34  37  48  55  58  60  61
 [19]  62  64  67  69  70  71  73  78  80  84  88  90  92  93  99 100 102 109
 [37] 111 112 114 115 116 120 121 122 128 129 130 132 133 134 135 136 138 141
 [55] 142 157 159 160 165 166 167 168 169 170 172 173 174 175 176 182 183 184
 [73] 185 193 194 198 204 207 208 209 212 214 215 216 218 219 221 222 223 224
 [91] 225 226 227 232 233 234 235 236 243 244 246 247 253 254 261 274 276 277
[109] 278 284 286 287 288 289 290 291 294 295 296 297 298 299 300 332 333 337
[127] 338 340 342 344 345 346 347 348 349 352 354 356 360 364 366 368 369 370
[145] 374 377 383 384 385 386 387 388 392 396 397 398 399 401 402 403 404 406
[163] 410 412 413 414 416 418

$Bicluster1$Cols
[1] 15 16 23 28 53 62
```

5- **Affichage des clusters :**

M=bib$B1cluster1 // donner un nom a le cluster

X= bib$Bicluster1$Rows // donner un nom a les lignes

Y= bib$Bicluster1$Cols // donner un nom a les colonnes

cluster1= lynass[X,Y]

lynass[aa,bb] //afficher le cluster

Résultat :

	Apotheke	Apple.Store..SoHo	Babe.Ruth.Plaza	Bank.of.America.Tower	Café.Café	Dutch.Kills	Empire.State.Building	Federal.Hall.National.Memorial
312	0	0	0	0	0	0	0	
319	0	0	0	0	0	0	0	
321	0	0	0	0	0	0	0	
323	0	0	0	0	0	0	0	

	Flushing.Meadows.Corona.Park	Four.Seasons.Hotel.New.York	Frying.Pan	George.Washington.Bridge	Grand.Central.Terminal	Grand.Hyatt.New.York	Hotel.Pennsylvania
312	0	0	0	1	0	0	
319	0	0	0	1	0	0	
321	0	0	0	0	0	0	
323	0	0	0	0	0	0	

	Houndstooth.Pub	Hunan.Delight	JFK.John.F..Kennedy.International	JFK.Terminal.5	Junior.s	Katz.s.Delicatessen	L.I.C. LGA.LaGuardia.Airport	Liberty.Island.Ferry
312	0	0	0	0	0	0	0	0
319	0	0	0	0	0	0	0	0
321	0	0	0	0	0	0	1	0
323	0	0	0	0	0	0	1	0

	Lil..Frankie.s.Pizza
312	0
319	0
321	0
323	0

6- Extraction des règles. L'étape suivante est l'extraction des règles.

```
M=as.matrix(lynass)
M= as (M,"transactions")
Rules = apriori (M,parameter=list(supp=0.00002,conf=0.75,minlen=2))
```

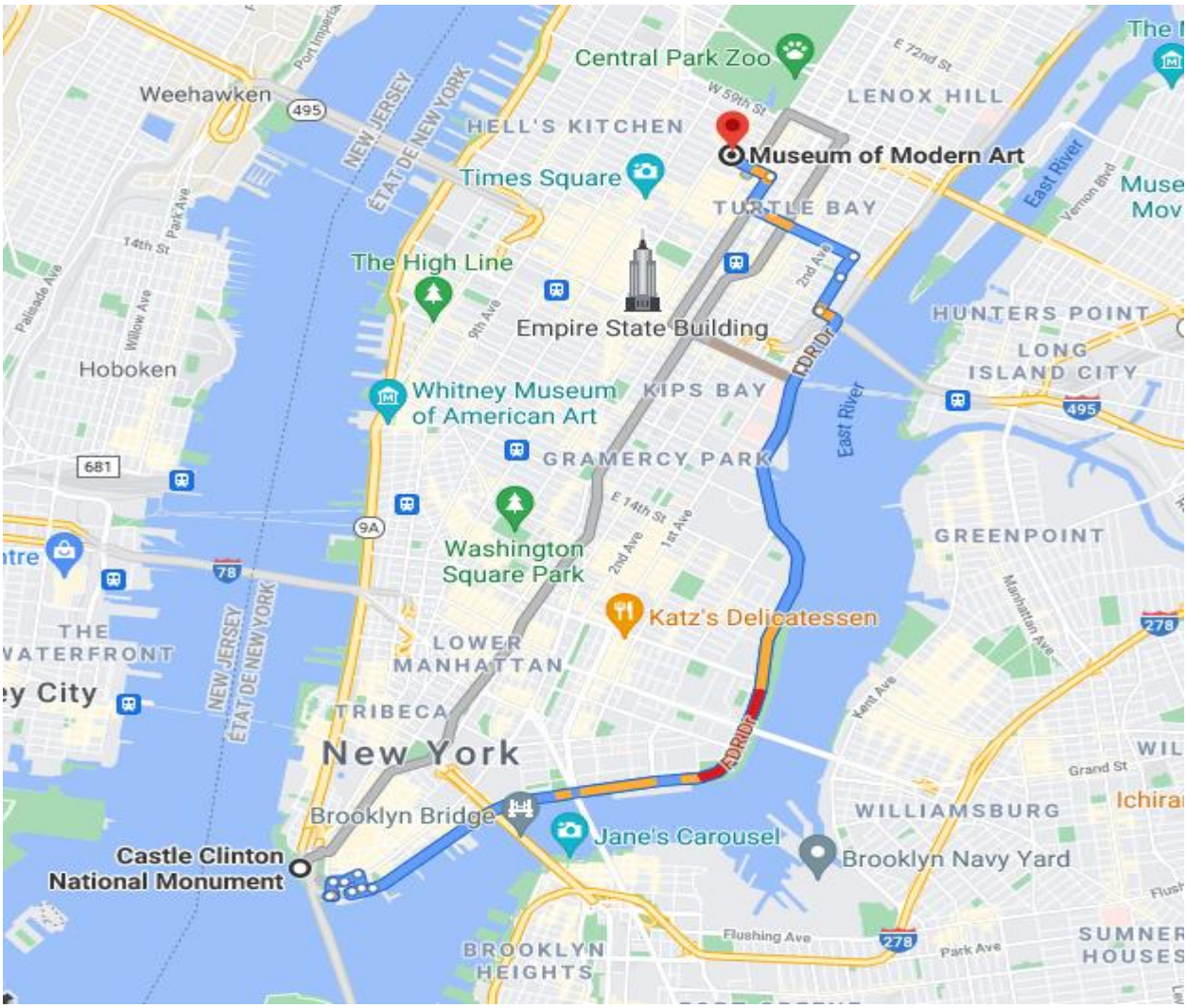
- On applique l'Apriori sur chaque cluster on a affiché

Numéro	Antecedent	Consequent	Lift	Support (%)	Confiance(%)
1	"Castle.Clinton.National.Monument =true"	"The.Museum.of.Modern.Art..MoMA=true "	24,3	0.91	75
2	"Ellis.Island =true"	"Statue.of.Liberty =true"	40.82667	0.261	80
3	"New.Rochelle.Station=true"	"Grand.Central.Terminal=true"	11.7	0.91	60
4	"Radio.City.Music.Hall, The.New.York.Public.Library=true"	"Times.Square =true"	45.3	0.131	66,66
5	"Grand.Central.Terminal=”true” ’ The.New.York.Public.Library=true"	"Rockefeller.Center=true"	23	0.131	66,66
6	"Grand.Central.Terminal=”true” ’ Rockefeller.Center=true"	"The.New.York.Public.Library=true"	40.8	0.131	66,66
7	"JFK.John.F..Kennedy.International =true" – "Madison.Square.Garden=true"	"Grand.Central.Terminal=true"	19	0.131	100
8	"Grand.Central.Terminal=true", "Rockefeller.Center =true"	"TimesSquare=true"	3.6	0.131	66,66
9	"The.Stonewall.Inn=true"	"Washington.Square.Park=true"	2.01	0.131	50
10	"Ed Sullivan Theatre =true"	"X1.World.Trade.Center=true"	13.5	0.131	100

Tableau (VI-4): Les règles d'associations (10 règles de différents cluster)

2. Description des règles :

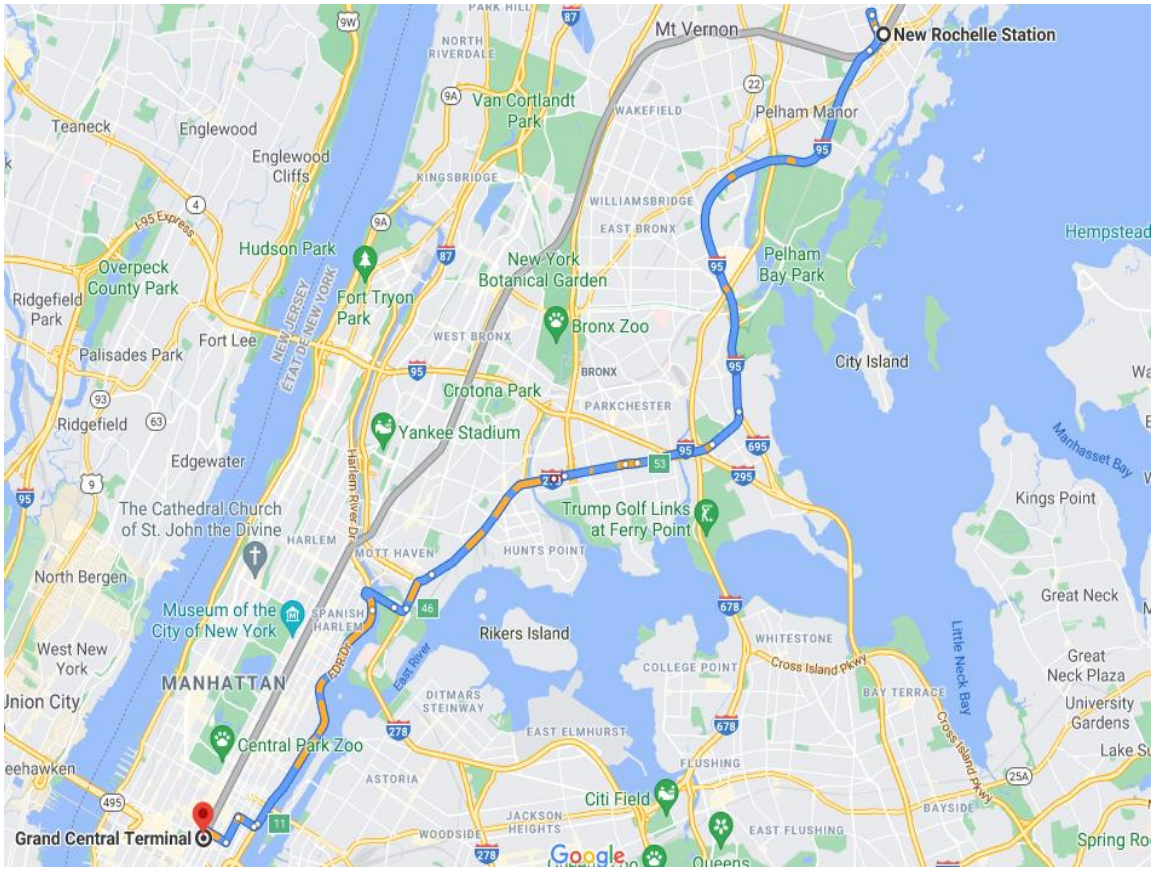
Castle.Clinton.National.Monument => The.Museum.of.Modern.Art.MoMA



Cette règle avec un support de 1.90 % et une confiance de 75% parmi plus de 817 nous indique que tous ceux qui visite le théâtre **Castle.Clinton.National.Monument** vont visiter le musée d'art moderne,

Cette règle même avec une confiance que de 42,85% inférieures à plusieurs autres règles mais son lift est de 668,85 ce qui lui donne une pertinence élevée et plus de chances d'être appliquée.

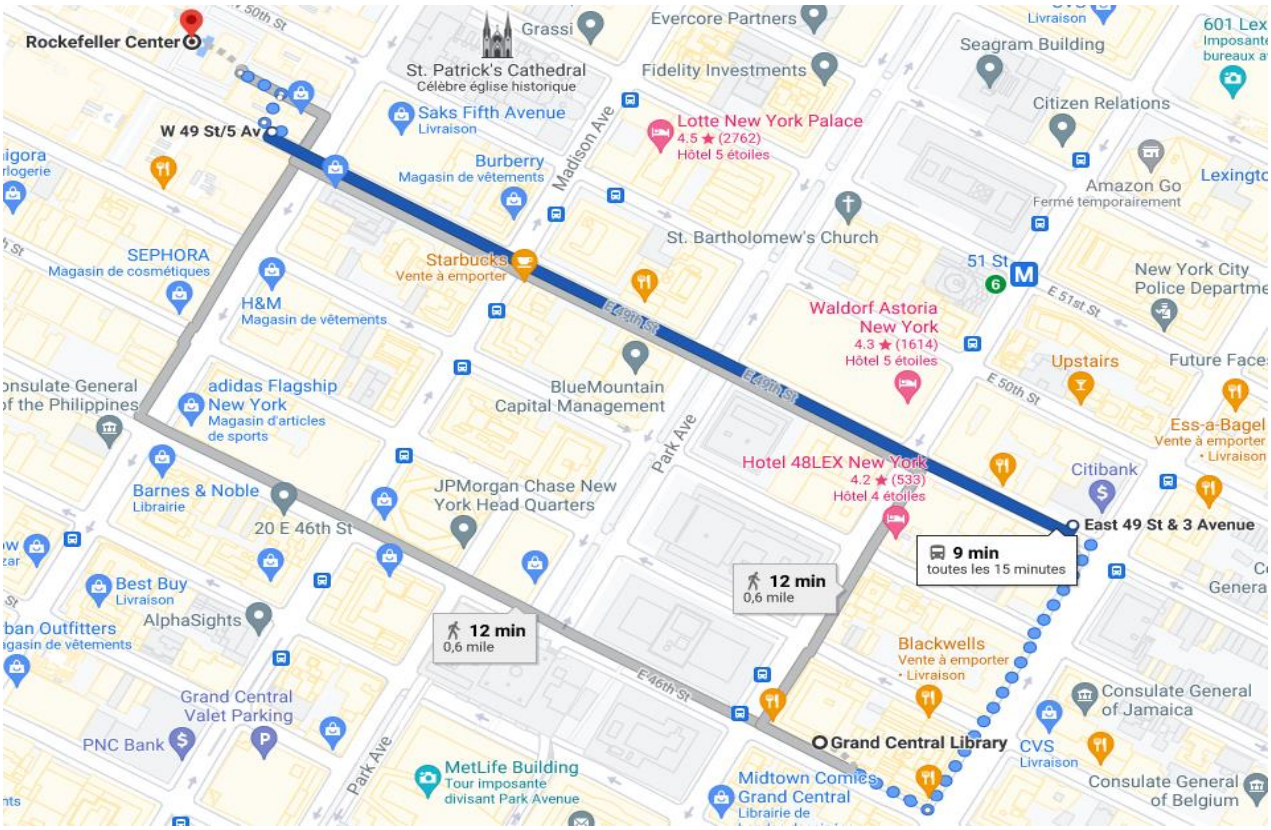
New.Rochelle.Station} => {Grand.Central.Terminal



Cette règle présente que 1.9 % des utilisateurs Gowalla préfèrent visiter New.Rochelle.Station et Grand.Central.Terminal avec 60 % precision

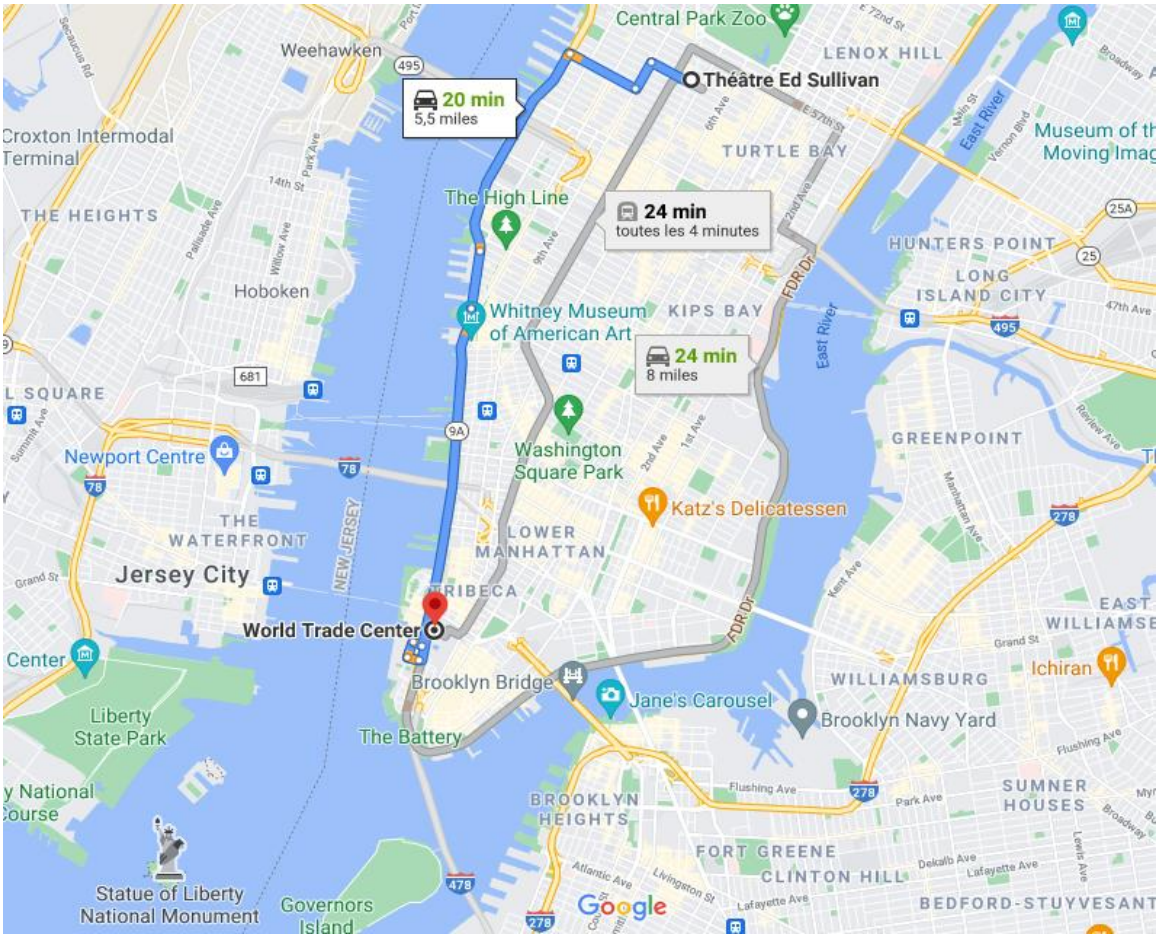
Grand.Central.Terminal ,The.New.York.Public.Library => Rockefeller.Center

,



Ces règles montrent que 1,30 % des utilisateurs situés à Grand.Central.Terminal et The.New.York.Public.Library à Rockefeller.Center avec une précision de 66.66%

Ed Sullivan Theatre => X1.World.Trade.Center



Cette règle montre que les 1,31% des utilisateurs de Gowalla qui sont situés dans le spot "Ed Sullivan Theatre" ont été localisés dans le World Trade Center avec une précision de 100%

3. La visualisation des règles d'association :

Parmi les façons de la visualisation nous citons le graphe de nuage de point :

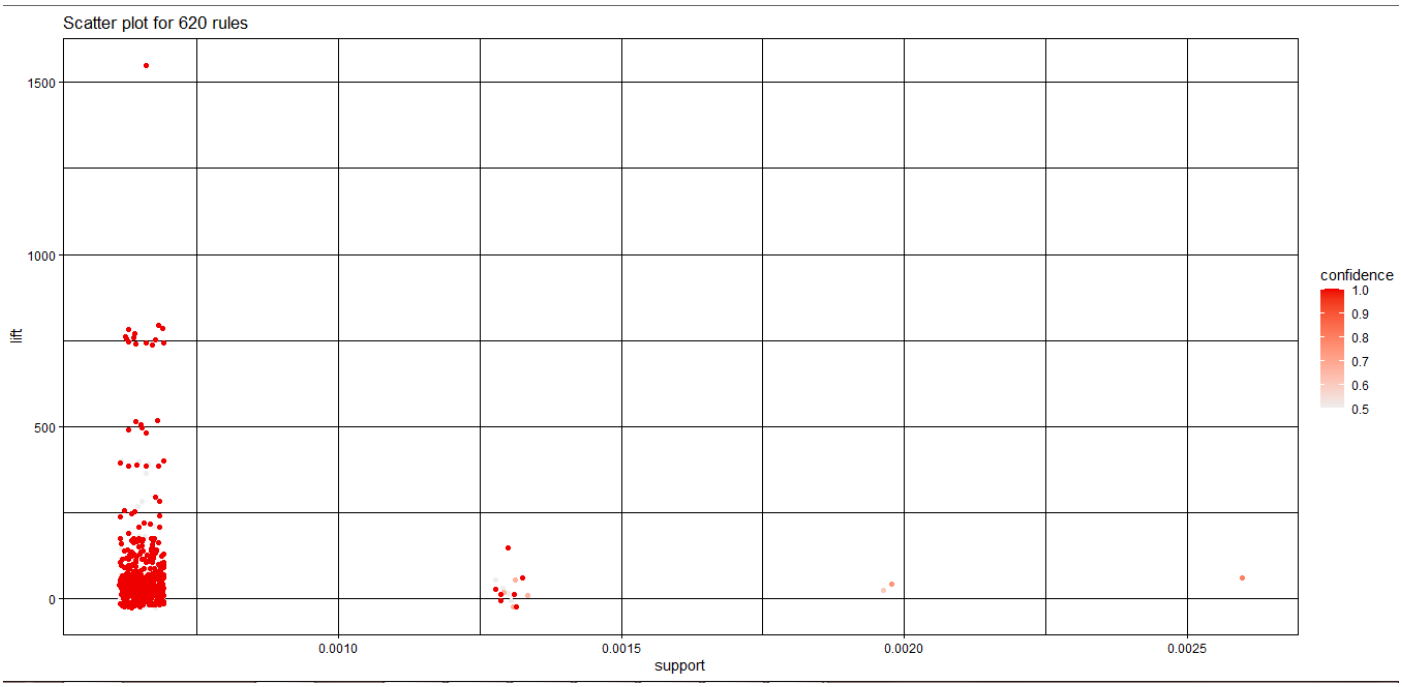


Figure (VI.28) : Le nombre des règles par rapport la variation support et confiance

Cette figure représente l'ensemble des règles trouvé par Apriori dans un Cluster. Il est au nombre 620 règles.

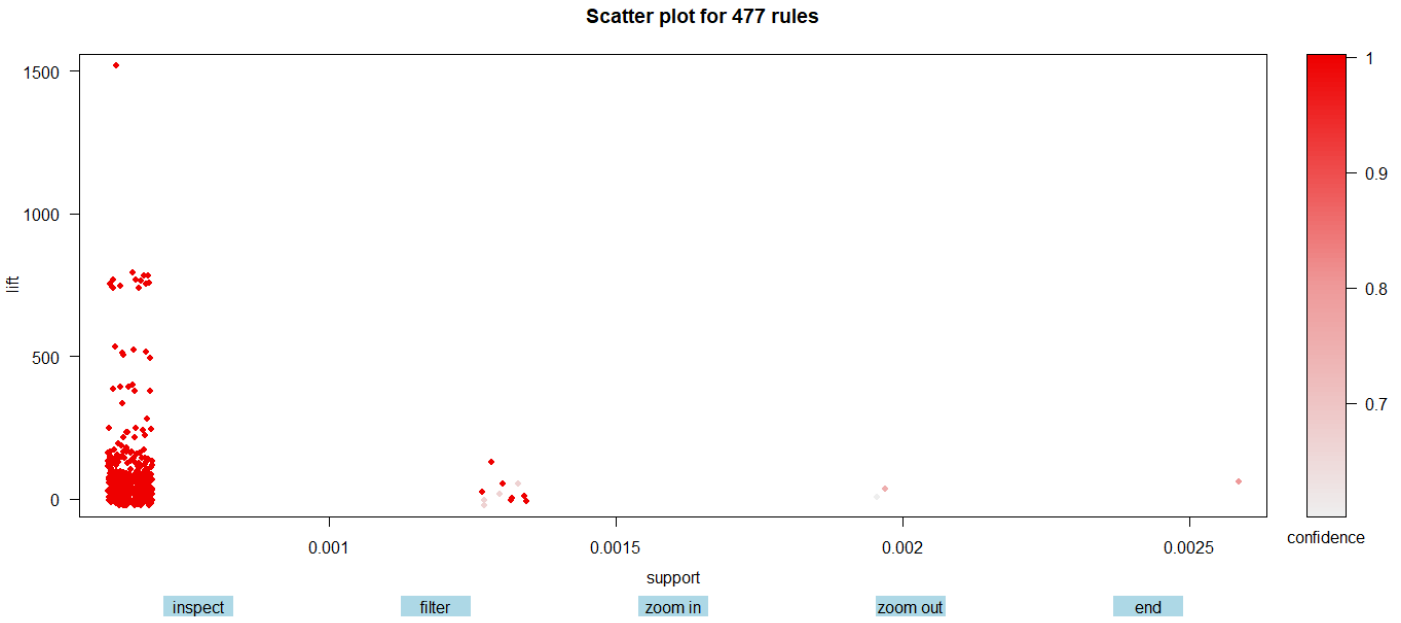


Figure (VI.29) : Scatter Plot de 7eme cluster pour 477 règles d'association

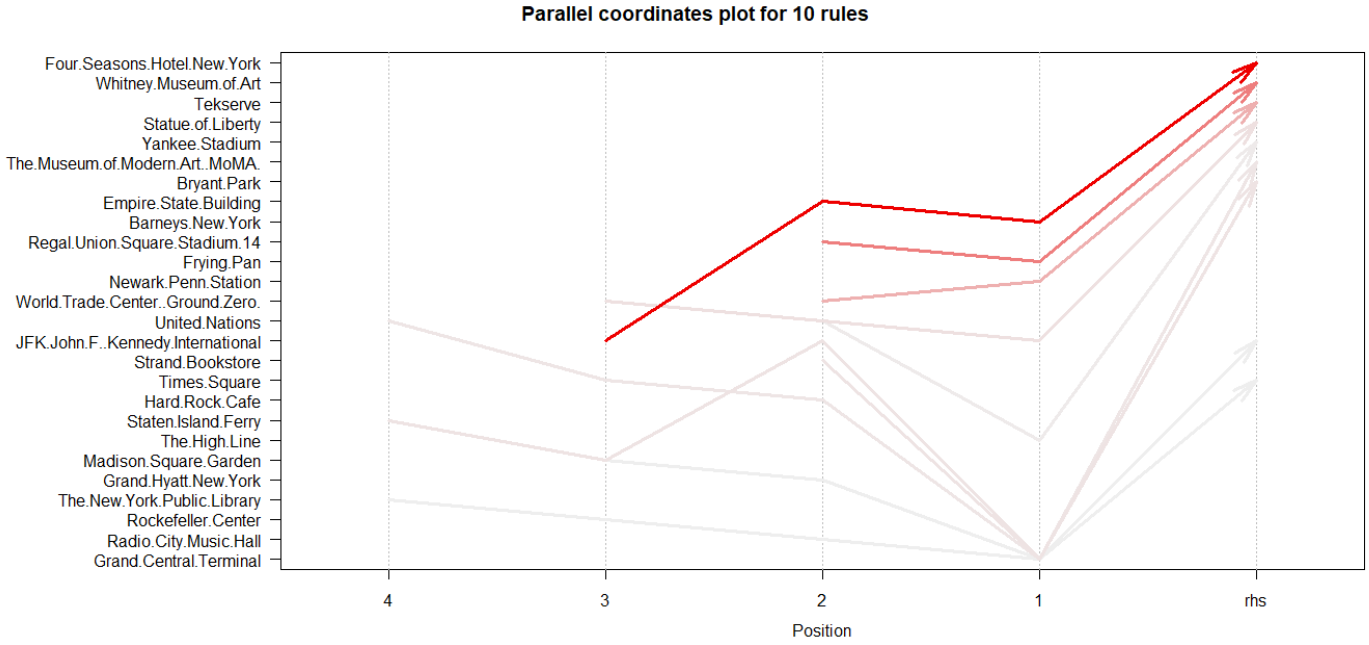


Figure (VI.30) : graphe parallèle coordinate pour 10 règles

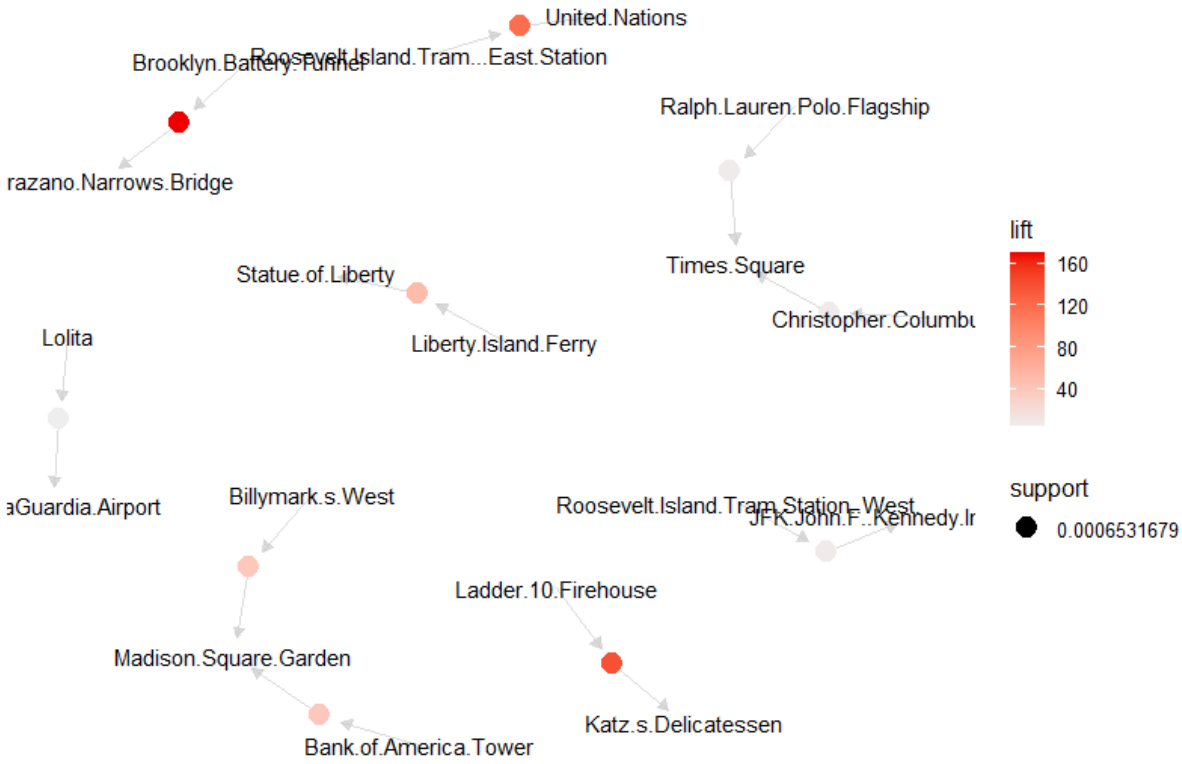
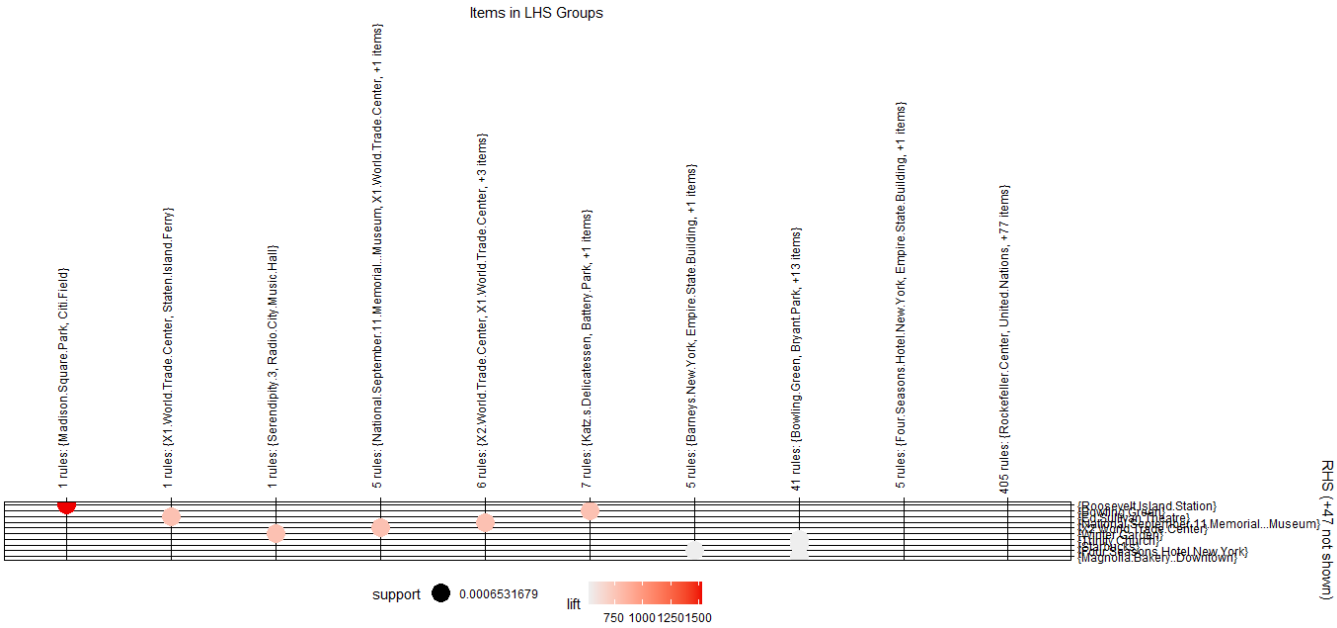


Figure (VI-31) : graphe two-key Plot



Figure(VI.32) : graphe des items pour 10 règles (Cluster7)

Conclusion

Le choix de notre échantillon de règles est guidé par le lift élevé et une confiance inférieure ou égale à 75% ce choix est motivé par le fait, qu'on considère que les règles avec une confiance élevée est un choix évident ou automatique.

On a voulu mettre la lumière sur ces règles qui font ressortir des liens qui ne sont pas évidents à voir par la simple application de l'algorithme Agrawal et all. Ces règles choisis nous disent que même quelquefois que la distance qui sépare ces sites géographiques est importante mais une dépendance forte les lie entre elles, ou dans d'autres règles des liens étranges ex (école de la mode et de l'art, le magasin des jeux pour enfants et le magasin de tissu).

On remarque d'après ces règles une hétérogénéité des utilisateurs et des lieux visités, mais en peut ressortir des liens forts qu'on peut utiliser pour influencer certaines décisions, choix ou tout simplement booster certain commerce, ou bien même mieux gérer les transports en communs et bien organiser et maitriser les événements culturels

Conclusion Générale

Conclusion

Dans ce mémoire nous nous sommes intéressés à l'étude sociologique des comportements des utilisateurs participants au site Gowalla, qui s'inscrivent volontairement au site est participent aux différentes activités de partages et de contributions au même temps ils nous indiquent leurs positionnements géographiques (spots) a partir de ces spots et utilisateurs Gowalla a pu collecter des informations sur les différents utilisateurs et les endroits visiter.

Notre étude est focalisée sur cette base de données Gowalla collecter entre 2007 et 2012 dans la ville de New York sur un échantillon plus de 600000 utilisateurs. En application du Biclustering sur les la base de données pour obtenu des résultats précis et correcte, et après on applique le data mining est précisément les règles d'associations avec l'algorithme Agrawal et all et les améliorations apporter à cet algorithme le Lift, pour pouvoir prédire sur le comportement futur des utilisateurs de ce site, et bien sûr la généralisation de ces connaissances sur l'ensemble des habitants et visiteurs la ville de New York.

Nous avons commencé par les différentes définitions des réseaux sociaux et l'analyse des réseaux sociaux, puis une étude sur le biclustering et ses fonctions sur le langage R. Puis sur le datamining les règles d'associations, l'algorithme Agrawal et all et quelques améliorations apporter à cet algorithme.

Dans la partie réalisation nous avons basé sur trois parties. La partie qui concerne le jeu de données nous avons décrit égalent dans cette partie l'architecture globale, le clustering des données c'est la deuxième partie. Nous avons utilisé le logiciel R pour appliquer le Biclustering et extraire les règles d'association de chaque cluster.

Alors que la troisième partie présente l'interprétation de notre modèle et la discussion de résultats obtenus. D'après les règles d'association qui ressortent de notre application de l'algorithme Agrawal par R, en remarque que la ville de New York est une ville très animé par des différents types : cultuelles, sportifs et économiques.