INTRODUCTION GENERALE

Depuis quelques années, les entrepôts de données ont pris une place importante dans les préoccupations des utilisateurs des bases de données. Le marché estime a connu une croissance énorme, et de nombreux projets ont été développés au sein des entreprises qui utilise les entrepôts de données.

Les entrepôts de données sont dédiés aux applications d'analyse et de prise de décision. Le processus d'analyse est réalisé à l'aide de requêtes complexes comportant de multiples jointures et des opérations d'agrégation sur des tables volumineuses. Les performances de ces requêtes dépendent directement de l'usage qui est fait de la mémoire secondaire. En e et, chaque entrée-sortie sur disque nécessitant jusqu'_a une dizaine de millisecondes, l'accès à la mémoire secondaire constitue de ce fait un véritable goulot d'étranglement. L'administrateur, dans le but de minimiser le cout d'exécution de ces requêtes, sélectionne un ensemble de vues matérialisées et un ensemble d'index.

Les utilisateurs des systèmes OLAP formulent des requêtes décisionnelles pour répondre à des besoins d'analyse spécifiques pour l'aide à la décision.

Les outils OLAP sont connus pour être intuitifs car leurs utilisateurs finaux ne sont pas forcément informaticiens. Cependant, la grande volumétrie des données et la complexité des requêtes d'analyse qui impliquent beaucoup d'agrégations rendent plus difficile la tâche d'analyse aux utilisateurs. Il est donc nécessaire d'offrir à ces derniers des solutions mieux adaptées à leur mode de raisonnement. Dans cet article nous proposons une architecture d'un processus de personnalisation de requêtes par des règles d'association.

Introduction

L'information représente un capital immatériel dont la bonne gestion est un facteur primordial pour la réussite de toute organisation. Les systèmes d'information ont pour objectif de supporter la réalisation des activités d'une organisation. Ils sont construits à partir des exigences des métiers et des processus définis par l'entreprise afin de stocker, traiter et communiquer les informations.

Les systèmes d'information des entreprises peuvent accumuler au fil du temps un volume important de données stockées sur plusieurs sites internes à l'entreprise ou provenant de son environnement externe (partenaires, Web, …). Le problème des entreprises est d'exploiter efficacement ces données afin de permettre aux décideurs d'optimiser leurs choix et de leur faciliter le pilotage à moyen terme via une meilleure anticipation. Ainsi, le besoin d'une exploitation efficace des données dans une perspective décisionnelle a donné lieu à l'élaboration de nouveaux systèmes, dits systèmes d'aide à la décision, facilitant le stockage et le traitement synthétique de grands volumes de données. [H. Jerbi, 2012]

1. LES SYSTEMES D'AIDE A LA DECISION

Un système d'information permet de faciliter la mise en œuvre des stratégies de l'entreprise. Le rôle d'un système d'aide à la décision est plutôt d'aider à déterminer les bonnes stratégies.

Plus précisément, un système d'aide à la décision permet de développer la capacité de réflexion et d'action de l'entreprise en aidant à l'apprentissage (analyse et suivi des activités précédentes) et au pilotage des plans d'actions (prévision et planification des activités futures).

Les systèmes d'aide à la décision sont employés dans tous les domaines où la prise de décision est nécessaire, à savoir, les domaines du commerce (marketing, ventes), de la logistique, de la santé (aide à la décision médicale), de la science (par exemple en bio-informatique), des télécommunications, des transports (trafic autoroutier), des banques, [F. Abdelhadi, 2014]

1.1. LE SYSTEME D'AIDE A LA DECISION DANS L'ORGANISATION

Toute organisation peut être décomposée en trois systèmes :

- Le système opérant qui correspond à l'activité de production de l'organisation en transformant les flux primaires pour répondre aux besoins des clients,
- Le système de décision correspondant à l'ensemble des traitements et du personnel dirigeant qui contrôle, régule, pilote et adapte l'organisation par leurs décisions,
- Le système d'information permettant de collecter, conserver, traiter et restituer les données produites dans l'organisation ; il joue le rôle d'interfaces entre les deux systèmes précédents. [F. Abdelhadi, 2014]

La Figure 1 présente les trois systèmes et leurs interactions

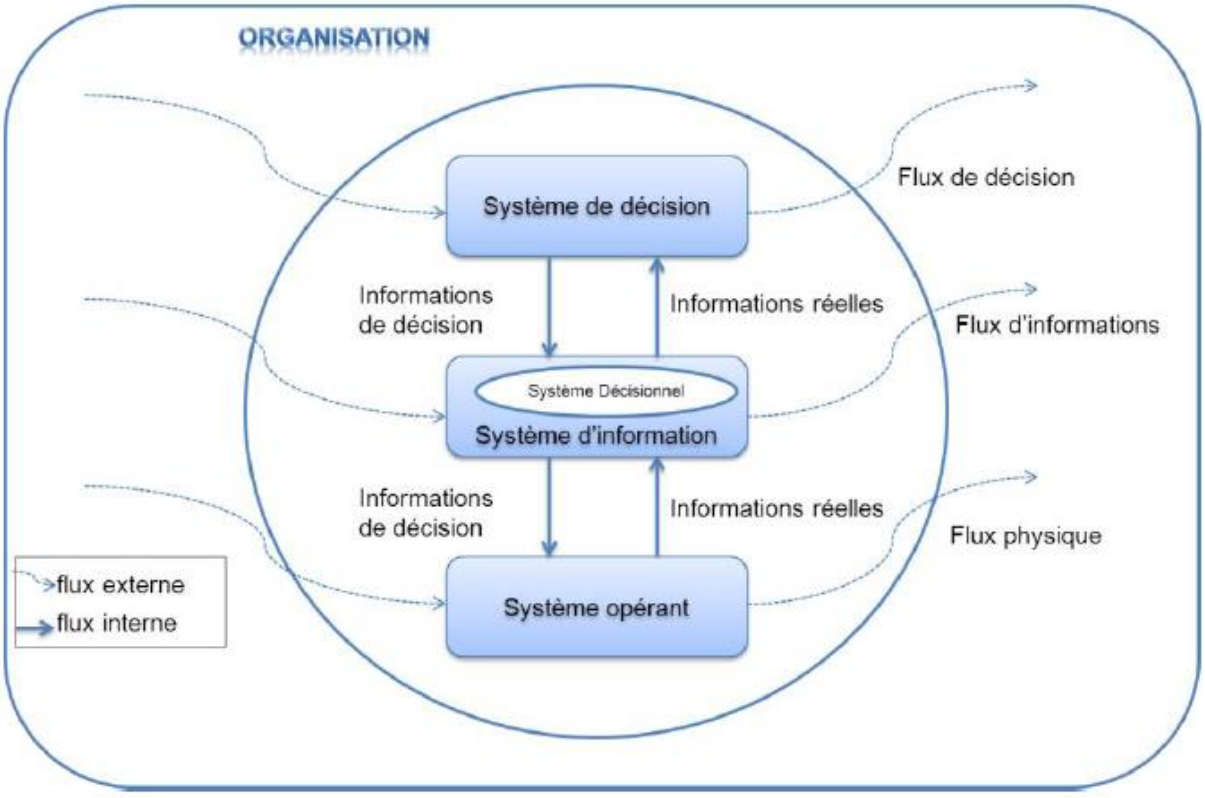

Figure 1. Positionnement d'un système décisionnel dans l'organisation[F. Abdelhadi, 2014]

Définition Un système d'aide à la décision est l'ensemble des outils matériels et logiciels qui permettent de collecter, de stocker et d'analyser des données issues du système d'information des entreprises dans le but de faciliter la prise de décision par les décideurs.

Les décideurs, utilisateurs de ces systèmes, sont des experts d'un métier chargé d'analyser les données décisionnelles pour le pilotage de l'organisation. Ils sont généralement non informaticiens. Dans la suite, nous les désignerons simplement par le terme usagers.

1.2 ARCHITECTURE DE SYSTEME D'AIDE A LA DECISION

Afin d'offrir une vision transversale de l'activité de l'entreprise, les systèmes d'aide à la décision collectent et stockent des données en provenance des bases de données des différents métiers de l'entreprise et de sources externes (sites web, emails,…). A notre sens, la conception d'un système d'aide à la décision doit être basée sur la séparation entre deux espaces de stockage : l'entrepôt qui regroupe toute l'information décisionnelle et les magasins qui contiennent une partie de cette information, dédiée à un thème, un métier, ou une analyse. [F. Abdelhadi, 2014]

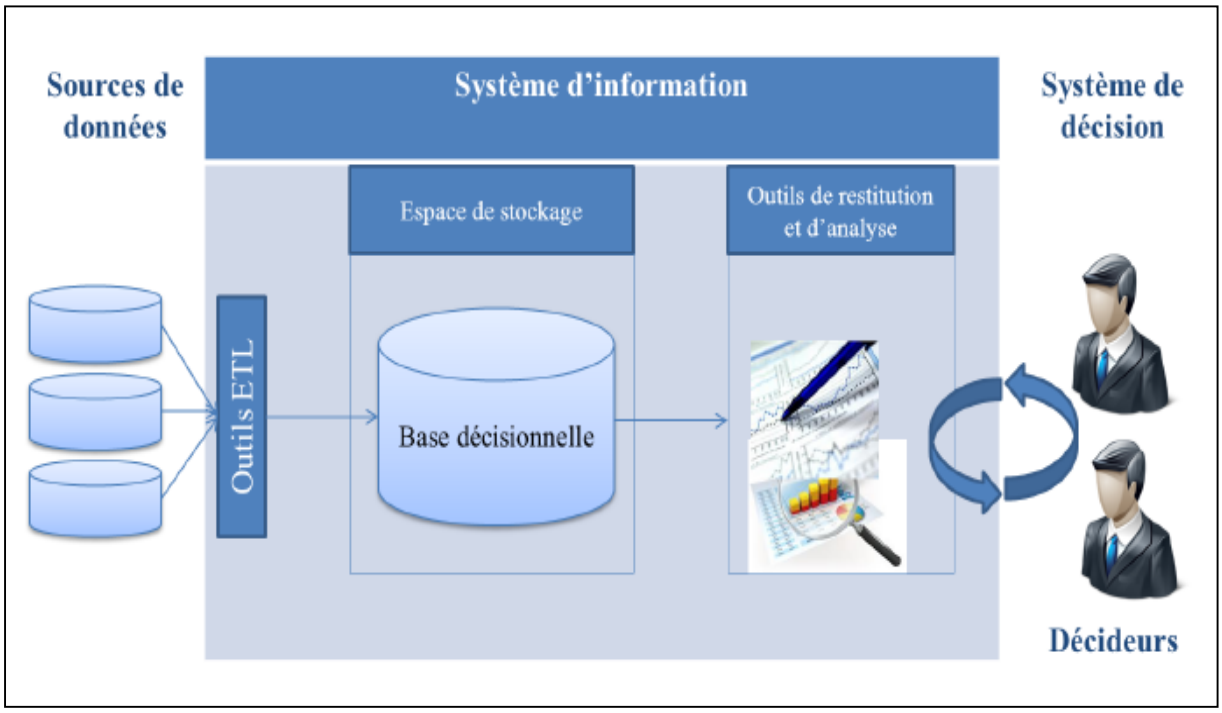

Figure 2. Le système d'aide à la décision[F. Abdelhadi, 2014]

1.3 STOCKAGE DE DONNEES

1.3.1 ENTREPOT

Bill Inmon définit l'entrepôt de données comme « une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision » (Inmon, 1996). D'après cette définition, les données sont :

- **Intégrées** : les données proviennent de plusieurs sources. Afin de les entreposer suivant une vision homogène, il faut les nettoyer, reformater et fusionner pour réduire leur hétérogénéité.
- **Orientées sujet** : les données sont regroupées et organisées en accord avec des thèmes ou des sujets d'analyse.
- **Non volatiles** : les données de l'entrepôt sont stables, c'est-à-dire, que les données déjà intégrées sont peu modifiées mais il est toujours possible d'ajouter des nouvelles données.
- **Historisées** : l'entrepôt garde une trace de l'historique des données

Définition. L'entrepôt de données est un espace de stockage centralisé qui permet de stocker et d'historiser des données hétérogènes qui sont pertinentes pour la prise de décision.

L'organisation des données au sein de l'entrepôt de données suit un modèle assurant la gestion efficace des données.

1.3.2 MAGASINS DE DONNEES

Définition. Un magasin de données constitue un extrait de l'entrepôt adapté à une classe de décideurs ou à un usage particulier et organisé suivant un modèle adapté aux traitements décisionnels

On parle généralement d'architecture à niveaux en raison des différents espaces de stockage considérés ; on distingue principalement l'architecture à 2 niveaux et l'architecture à 3 niveaux. Par exemple si l'on prend une architecture à 3 niveaux, on considère :

- Les sources de données, souvent hétérogènes et réparties, associées aux outils ETL permettent l'intégration et l'alimentation de l'entrepôt ;
- L'entrepôt contenant les données pour la prise de décision
- Le magasin extrait de l'entrepôt et dédié à une classe de décideurs ; il est organisé suivant un modèle multidimensionnel. [F. Abdelhadi, 2014]

Nous présentons une architecture à 3 niveaux dans la Figure 3

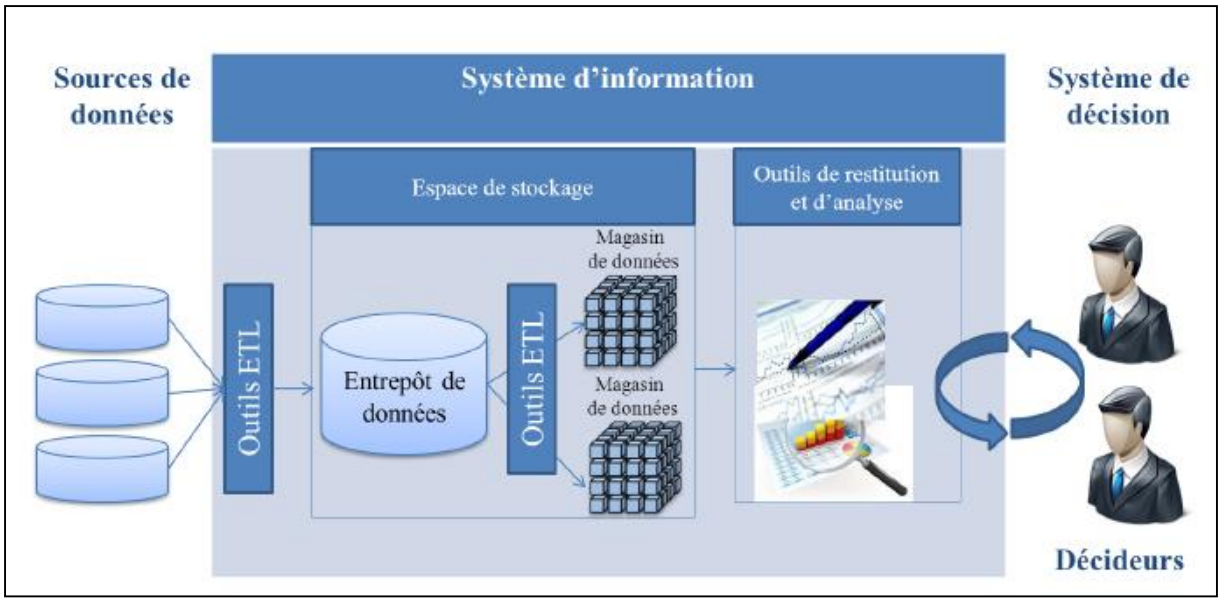

Figure 3. Architecture décisionnelle à 3 niveau[F. Abdelhadi, 2014]

Cette architecture sépare clairement les deux espaces de stockage ; "l'entrepôt" où les données sont représentées selon un modèle informatique et le « magasin » où les données sont décrites dans un modèle multidimensionnel.

1.4. LISTE DES PHASES D'UN PROJET DECISIONNEL

1.4.1 LA PHASE DE COLLECTE

La collecte s'effectue à partir de données appelées : **données sources**. Ces données peuvent se présenter sous différents formats. Il peut s'agir de fichiers "plats" (fichiers CSV avec séparateurs, fichiers XML, fichiers ASCII...) mais aussi de systèmes de bases de données (export de base MySQL, PostgreSQL, DB2, ORACLE...). Ces sources de données sont donc en général **hétérogènes** c'est pourquoi il va falloir passer par une phase dites d'**intégration** pour pouvoir les manipuler avant de les stocker dans notre système d'aide à la décision.

1.4.2 LA PHASE D'INTEGRATION

C'est à ce niveau qu'apparait la première couche logicielle de l'environnement décisionnel à savoir l'ETL. Cette couche offre des fonctions d'extraction de données issues de différents systèmes (internes ou externes), de transformation de ces données (homogénéisation, filtrage, calcul) et de leur chargement dans un *ODS* intermédiaire ou directement dans le DW (entrepôt de données). Elle garantit la délocalisation de la charge de calcul et une meilleure disponibilité des sources.

La deuxième couche logicielle est l'ODS qui fait office de structure intermédiaire destinée à stocker les données issues des systèmes de production opérationnelle. Ce sont en quelque sorte des zones de préparation avant l'intégration des données dans le DW : périodicité journalière, données qualifiées, premier niveau de filtrage et d'agrégat. En général, il existe deux types de schéma : un schéma "ODS brut" qui contient les tables qui reçoivent

les données brutes des différentes sources et un schéma "ODS final" qui contient des tables avec une structure (champs et contraintes associées) le plus proche possible du schéma du DW (même si les tables de celui-ci peuvent contenir plus de champs que les tables du DW) car ces données vont ensuite être figées dans l'entrepôt. L'ODS ne contient des données que sur une **faible période** et ces données vont être manipulées, transformées, traitées, modifiées plusieurs fois avant d'être copiées dans le DW. On peut se passer d'utilisation d'un ODS dans un seul cas : si les données du DW sont une simple copie (c'est-à-dire qu'il n'y a pas de traitements à faire et que les données extraites ne vont pas évoluer) des données de production (sources) ce qui n'est malheureusement pratiquement jamais le cas dans de grosses structures.

1.4.3 LA PHASE D'ORGANISATION

La troisième phase permet de **stocker** les données dans un entrepôt appelé : **Datawarehouse**. Cet entrepôt contient les données orientées métier, non volatiles (datées), historisées et documentées. Cette structure de données est volontairement généralement **dénormalisée** pour pouvoir optimiser les temps de réponses lorsque l'on fait des analyses de type OLAP qui se réfère à une base de données **multidimensionnelle** (aussi appelée cube ou hypercube). Elle est constituée de **dimensions** ou **axes d'analyse** (l'axe temporel ou géographie sont des exemples courant) et de **faits** ou **indicateurs** (tels que le chiffre d'affaires). Un élément important vient du fait que les données stockées dans le DW ne doivent plus changer une fois à l'intérieur. Ce sont des données consolidées et figées qui vont nous permettre de faire toute sorte d'analyses et statistiques.

Une fois ces données stockées dans le Datawarehouse, on va pouvoir créer des magasins de données appelés : **Datamarts**. Comme le Datawarehouse, c'est un entrepôt de données mais dédié à une fonction de l'entreprise pour des raisons d'accessibilité, de facilité d'utilisation ou de performance. Les données sont généralement équivalentes à celles présentes dans le DW principal mais elles sont représentées de façon adaptée aux besoins spécifiques de la fonction et/ou du domaine utilisateur (par exemple, on va créer un DM dédié pour le service Marketing ou Commercial). Le DM peut avoir une implémentation physique (cube) ou n'être qu'une vue logique ("multiprovider").

1.4.4 LA PHASE DE RESTITUTION

La dernière phase concerne la restitution des résultats, on distingue à ce niveau plusieurs types d'outils différents :

- Les outils de **reporting** et de **requêtes**
- Les outils d'**analyse**
- La phase de **Datamining**

Les outils de **reporting** et de **requêtes** permettent la mise à disposition de rapports périodiques, pré-formatés et paramétrables par les opérationnels. Ils offrent une couche d'abstraction orientée métier pour faciliter la création de rapports par les utilisateurs eux-mêmes en interrogeant le datawarehouse grâce à des analyses croisées. Ils permettent

également la production de tableaux de bord avec des indicateurs de haut niveau pour les managers, synthétisant différents critères de performance.

Les outils d'**analyse** OLAP permettent de traiter des données et de les afficher sous forme de cubes multidimensionnels et de naviguer dans les différentes dimensions. Cet agencement des données permet d'obtenir immédiatement plusieurs représentations d'un même résultat, en une seule requête sous une approche descendante des niveaux agrégés vers les niveaux détaillés (Drill-down, Roll-up).

Les outils de **Datamining** offrent une analyse plus poussée des données historisées permettant de découvrir des connaissances cachées dans les données comme la détection de corrélations et de tendances, l'établissement de typologies et de segmentations ou encore des prévisions. Le Datamining est basé sur des algorithmes statistiques et mathématiques, et sur des hypothèses métier.

1.5. RESTITUTION ET ANALYSE OLAP

Selon l'architecture à deux niveaux, un entrepôt de données est structuré suivant une modélisation multidimensionnelle. Ceci permet de représenter l'extension d'un entrepôt sous la forme de points dans un espace à plusieurs dimensions avec la métaphore du cube ou de l'hyper-cube de données. [F. Abdelhadi, 2014]. La Figure 4 présente un exemple de cube qui permet l'analyse des ventes de matériels Informatiques. L'analyse des montants de ventes s'effectue en fonction de trois dimensions : les magasins où ont été effectuées les ventes, les dates de ventes et les produits vendus. Chacune de ces dimensions est associée à des paramètres de granularité différente (pour la dimension Magasin : ville, pays et continent). Ces niveaux hiérarchiques permettent d'obtenir des visions plus ou moins synthétiques lors des analyses OLAP.[F. Abdelhadi, 2014]

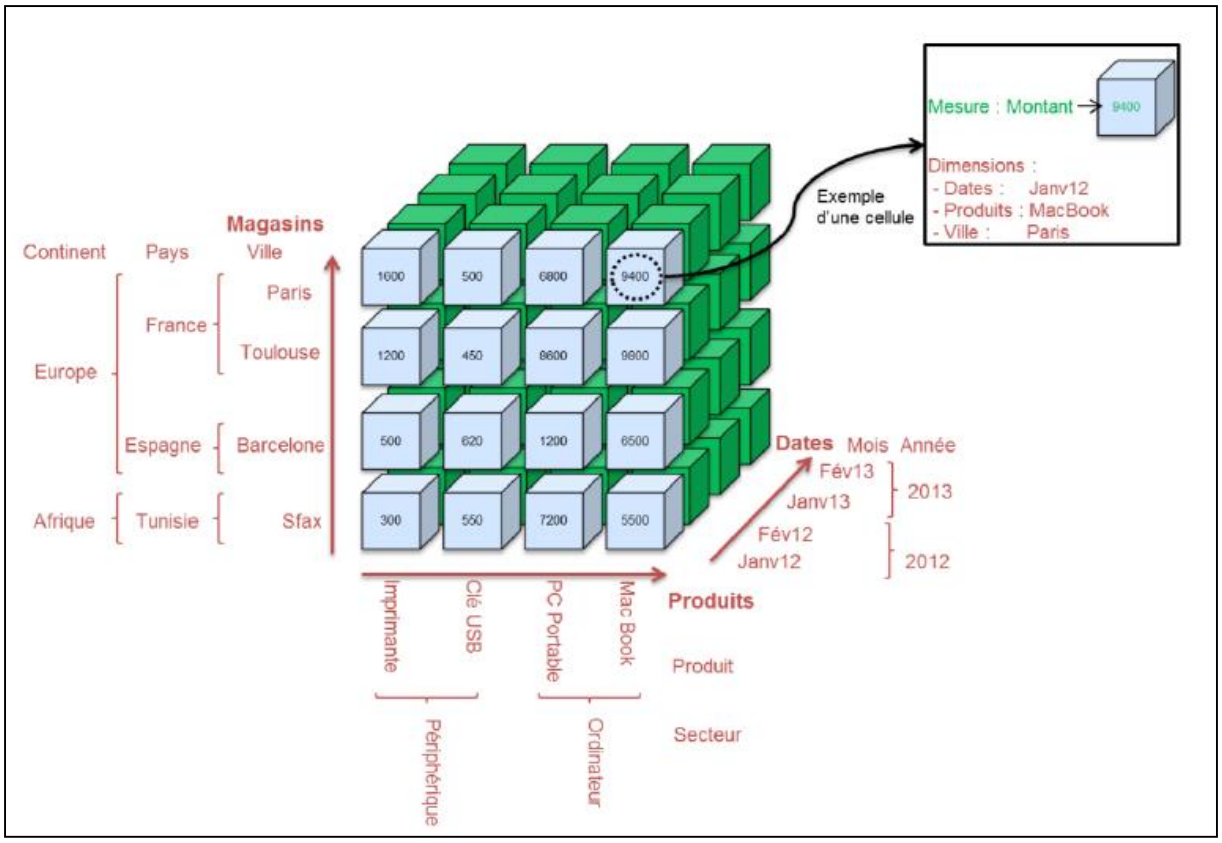

Figure 4. Exemple d'un cube représentant les ventes de matériels informatique [F. Abdelhadi, 2014]

La modélisation d'un entrepôt sous la forme d'un cube s'avère très limitée puisqu'elle se limite à trois dimensions (Torlone, 2003). Pour concevoir des schémas multidimensionnels plus élaborés, des structures plus avancées ont été définies ; elles permettent la modélisation de sujets d'analyse appelés faits, et d'axes d'analyse appelés dimensions (Kimball, 1996), (Abelló et al., 2001a) et (Abelló et al., 2001b). Les faits sont des regroupements d'indicateurs d'analyse appelés mesures. [F. Abdelhadi, 2014]

Les dimensions sont composées d'attributs, appelés paramètres, agencés de manière hiérarchique et qui modélisent les différents niveaux de détails des axes d'analyse. Un fait et ses dimensions associées composent un schéma en étoile (Kimball, 1996).

Les données des mesures sont appelées données factuelles car elles représentent un événement. Elles correspondent aux données des cellules du cube qui seront analysées en fonction des axes d'analyse. Le schéma multidimensionnel, associé à l'exemple de la Figure 4, est présenté en Figure 5.

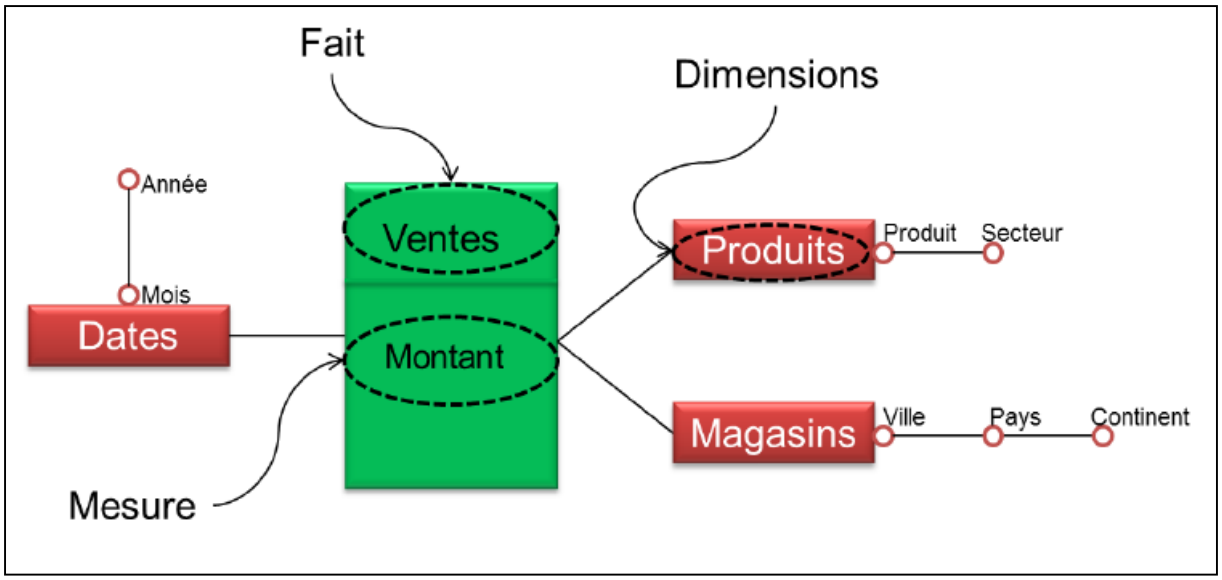

Figure 5. Exemple d'un schéma multidimensionnel [F. Abdelhadi, 2014]

Une analyse multidimensionnelle est une requête partant sur les données d'un entrepôt.

Généralement, le résultat d'une requête OLAP est représenté sous la forme d'une table à deux dimensions. La table multidimensionnelle de la Figure 6 représente le résultat d'une requête OLAP.

Dans cet exemple, la table contient les analyses des montants des ventes en fonction des pays auxquels appartiennent les magasins. La vente est restreinte aux ventes effectuées en janvier 2012 .[F. Abdelhadi, 2014]

Ventes SUM(Montant)			Magasins			
			Continent	Europe		Afrique
			Pays	France	Espagne	Tunisie
Produits	Secteur	Produit				
	Ordinateur	MacBook		19200	6500	5500
		PCPortable		15400	1200	7200
	Périphériques	CléUSB		950	620	550
		Imprimante		2800	500	300
Dates = Janv12						

Valeurs cumulées à partir des magasins situés à Paris et à Toulouse

Figure 6. Une table multidimensionnelle [F. Abdelhadi, 2014]

1.6. LES DEMARCHES D'ELABORATION DES ENTREPOTS

La construction d'un entrepôt est un processus qui comporte plusieurs étapes successives : la conception d'un schéma multidimensionnel, la création de l'entrepôt conforme à ce schéma et le chargement de l'entrepôt depuis les sources. La conception d'un schéma multidimensionnel peut être effectuée selon l'une des 3 démarches suivantes.

- La démarche ascendante utilise uniquement le schéma des sources pour générer des schémas multidimensionnels candidats sans prendre en compte, dans un premier temps, les besoins des décideurs. Ceux-ci choisissent ensuite le schéma le plus adapté à leurs besoins.
- La démarche descendante prend uniquement en compte les besoins des décideurs. Elle se base sur la spécification de ces besoins pour définir les sujets et les axes d'analyse. A l'issue du processus d'élaboration du schéma multidimensionnel, la correspondance entre le schéma résultat et la source de données est établie.
- La démarche mixte combine les deux démarches précédentes. En effet, cette démarche construit d'une part des schémas candidats à partir des sources de données (démarche ascendante) et d'autre part des schémas multidimensionnels à partir des besoins d'analyse (démarche descendante). L'informaticien doit confronter ces deux types de schémas pour obtenir un schéma multidimensionnel cohérent et répondant aux besoins des décideurs. [F. Abdelhadi, 2014]

Conclusion

Dans ce chapitre nous définis le système d'aide à la décision, son architecture dans l'organisme puis nous avons présenté les entrepôts de données et les magasins de données et leur démarche d'élaborations ainsi que la liste des phases du projet décisionnel dans le chapitre suivant nous voulons définir le DataMining et les règles d'association, ses mesures de qualité et ses applications.

Introduction

Le DataMining est une discipline qui se base essentiellement sur les pratiques et les expériences vécues par des entreprises qui ont décidé d'investir dans ce domaine. Le DataMining peut être vu comme la formulation, l'analyse et l'implémentation d'un processus de transformation de données en des connaissances. Les techniques de DataMining sont de plus en plus employées dans des domaines scientifiques et dans des domaines industriels. Parmi les méthodes utilisées par le DataMining sont les règles d'association, Nous présentons dans ce chapitre les règles d'association dyadique et triadique ainsi que leur mesure de qualité.

1.1. L'EXTRACTION DE CONNAISSANCES A PARTIR DE DONNEES

La donnée peut constituer une connaissance en tant que telle, mais peut également être traitée pour extraire de la connaissance. Jean-Paul Benzécri écrivait en 1977 que l'analyse des données avait pour objectif de dégager de la gangue des données le pur diamant de la véridique nature. Plus tard, Witten et al. ont défini la fouille de données comme l'extraction d'informations implicites, inconnues et utiles à partir des données. Cela a donné lieu à l'émergence d'une nouvelle discipline dans les années 80 appelée Knowledge Discovery in Databases (KDD) ou Extraction de Connaissances à partir des Données (ECD). Gregory Piatetsky-Shapiro 2 est un pionnier dans ce domaine. Fondateur des conférences KDD3, il a organisé le premier workshop *Knowledge*

Discovery in Databases en 1989. Depuis, nombre de travaux et de publications alimentent cette communauté et en enrichissent les techniques.

La définition de l'extraction de connaissances à partir des données a été donnée en 1996 par Fayyad et al. : « KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. » L'ECD est un domaine distinct de l'apprentissage automatique (*Machine Learning*), car ce dernier met l'accent sur l'induction de modèles pour la prédiction, par exemple pour reproduire une tâche humaine ou pour adapter un comportement en fonction de résultats de calculs. L'extraction de données combine des techniques issues de disciplines variées, comme les bases de données, l'intelligence artificielle, les statistiques, les découvertes scientifiques et la visualisation. Elle est donc à l'intersection de ces disciplines et va puiser dans chacune d'elles. Une de ses caractéristiques est la taille des bases de données qu'elle traite, qui atteindra le yottaoctet avec la mise en service du Utah Data Center de la N.S.A. [G. Bothorel, 2014]

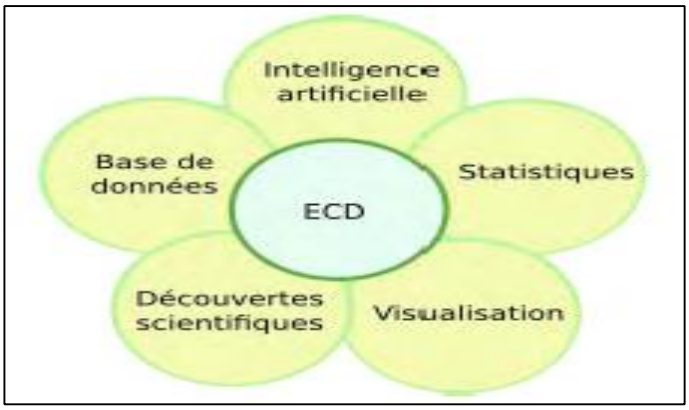

Figure 7 – L'extraction de connaissances à partir de données est à l'intersection de plusieurs disciplines.[G. Bothorel, 2014]

L'extraction de connaissances à partir de données est un processus interactif et itératif, contenant plusieurs niveaux de décisions de la part de l'utilisateur . Il est constitué des neuf étapes suivantes (Figure 8) :

1. Développer une compréhension du domaine d'application et identifier le but du processus ECD du point de vue de l'utilisateur ;
2. Sélectionner un jeu de données sur lequel l'extraction va être réalisée ;
3. Nettoyer et prétraiter les données. Cela concerne par exemple l'extraction du bruit et la mise en œuvre de stratégies dans le cas de données manquantes ;
4. Réduire et projeter les données, en identifiant des caractéristiques pour les représenter en fonction du but de la tâche. Le nombre de variables peut ainsi être fortement diminué ;
5. Faire correspondre les buts de l'extraction avec une méthode de fouille de données particulière, comme la classification, le regroupement (clustering) ;
6. Analyser de manière exploratoire et choisir l'algorithme de fouille de données et de la méthode de sélection ;
7. Réaliser la **fouille de données**. Il s'agit de rechercher des motifs intéressants ;
8. **Interpréter les motifs trouvés**. Cette étape comprend également la **visualisation des motifs** ;
9. Valoriser la connaissance acquise, en l'utilisant directement, ou en l'intégrant dans un autre système pour un futur processus. Ce processus est itératif, de manière globale ou entre différentes étapes, comme indiqué sur la figure 5 La notion de fouille de données ou Data Mining varie selon la littérature. La définition de ce concept peut aller de l'extraction de motifs jusqu'au processus global d'Extraction de Connaissances à partir des Données. [G. Bothorel, 2014]

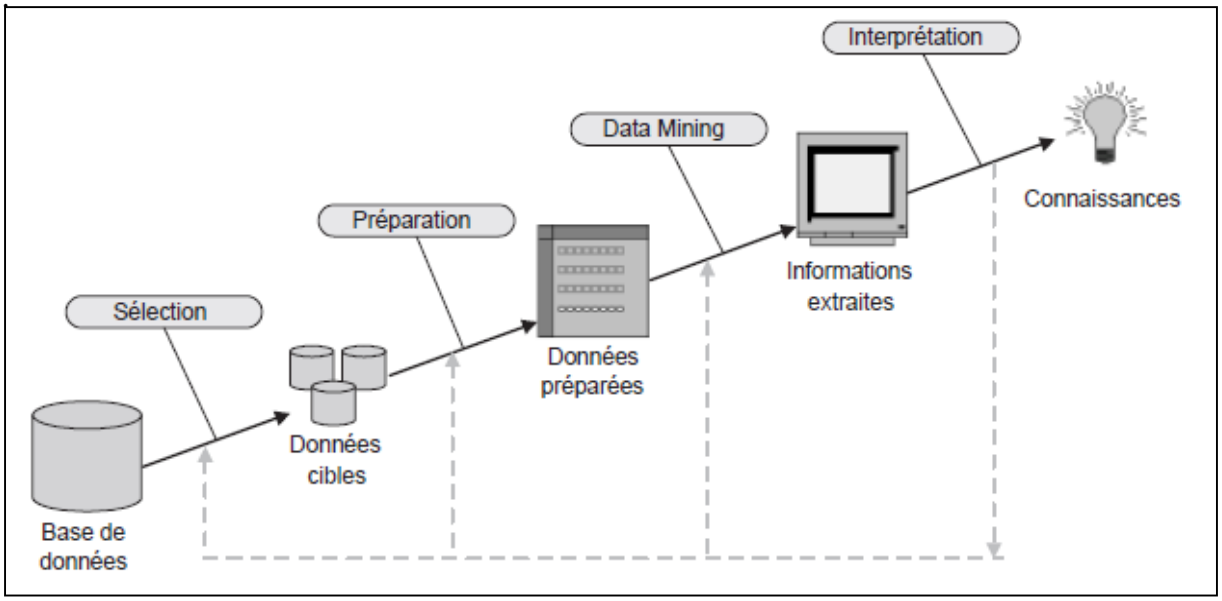

Figure 8 – Les neuf étapes de l'Extraction de Connaissances à partir de Données [N. Pasquier, 2010]

1.2. DATA MINING

C'est un outil d'exploration des données décisionnelles

Définition : Le *Data Mining* est en fait un terme générique englobant toute une famille d'outils facilitant l'exploration et l'analyse des données contenues au sein d'une base décisionnelle de type Data Warehouse ou DataMart. Les techniques mises en action lors de l'utilisation de cet instrument d'analyse et de prospection sont particulièrement efficaces pour extraire des informations significatives depuis de grandes quantités de données.

À quoi ça sert ?

Principe : En peu de mots, l'outil de prospection Data Mining est à même de trouver des structures originales et des corrélations informelles entre les données.
Il permet de mieux comprendre les liens entre des phénomènes en apparence distincts et d'anticiper des tendances encore peu discernables.

Comment on l'utilise ?

A contrario des méthodes classiques d'analyse statistique, Cet instrument d'analyse est particulièrement adapté au traitement de grands volumes de données. Avec l'augmentation de la capacité de stockage des supports informatiques, un maximum de renseignements seront captés, ordonnés et rangés au sein du Data Warehouse. Comportement des acheteurs, caractéristiques des produits, historisation de la production,

désormais plus rien n'échappe à la collecte. Avec le Data Mining, ces "*téra-nesque*" bases de données sont exploitables.

Les techniques mises en œuvre

Différentes techniques sont proposées. Elles sont à choisir en fonction de la nature des données et du type d'étude que l'on souhaite entreprendre

- Les méthodes utilisant les techniques de classification et de segmentation
- Les méthodes utilisant des principes d'arbre de décision assez proches des techniques de classification
- Les méthodes fondées sur des principes et des règles d'associations ou d'analogies
- Les méthodes exploitant les capacités d'apprentissage des réseaux de neurones
- Et pour les études d'évolution de populations, les algorithmes génétiques
- *Algorithmes Naïve Bayes, séries chronologiques, régression linéaire...[6]*

1.3. Processus Data Mining

Le principe : une démarche (simplifiée et didactique) en 5 temps majeurs.

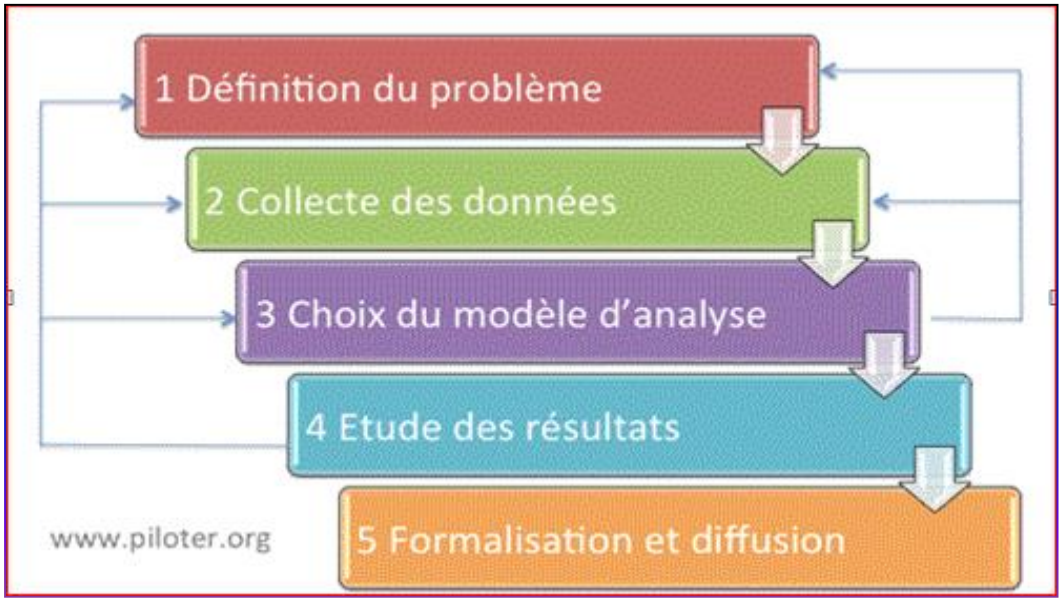

Figure 9 – Processus Data mining [6]

2.3.1 DEFINITION DU PROBLEME

Quel est le but de l'analyse, que recherche-t-on ? Quels sont les objectifs ? Comment traduire le problème en une question pouvant servir de sujet d'enquête pour cet outil d'analyse

bien spécifique ? A ce sujet, se souvenir que l'on travaille à partir des données existantes, la question doit être ciblée selon les données disponibles.

2.3.2 COLLECTE DES DONNEES

Une phase absolument essentielle. On n'analyse que des données utilisables, c'est à dire "propres" et consolidées. On n'hésitera pas à extraire de l'analyse les données de qualité douteuse. Bien souvent, les données méritent d'être retravaillées. S'assurer au final que la quantité de données soit suffisante pour éviter de fausser les résultats. Cette phase de collecte nécessite le plus grand soin.

2.3.3 CONSTRUIRE LE MODELE D'ANALYSE

Ne pas hésiter à valider vos choix d'analyse sur plusieurs jeux d'essais en variant les échantillons. Une première évaluation peut nous conduire à reprendre les points 1 ou 2. [6]

2.3.4 ETUDE DES RESULTATS

Il est temps d'exploiter les résultats. Pour affiner l'analyse on n'hésitera pas à reprendre le point 1, 2 ou 3 si les résultats s'avéraient insatisfaisants.

2.3.5 FORMALISATION ET DIFFUSION

Les résultats sont formalisés pour être diffusé. Ils ne seront utiles qu'une fois devenus une connaissance partagée. C'est bien là l'aboutissement de la démarche. C'est aussi là que réside la difficulté d'interprétation et de généralisation...[6]

2.4. REGLE D'ASSOCIATION

2.4.1 ITEMSET

- **Principe et définition**

  Les données, appelées également *transactions*, sont composées d'attributs. Leur ensemble constitue la base de données.

  Nous appelons $\mathcal{A} = \{A_1, A_2, ..., A_m\}$ l'ensemble des ***m*** attributs possibles d'une base. Ils peuvent être regroupés de multiples manières pour former des *itemsets* définis de la manière suivante :[G. Bothorel, 2014]

Définition 1 *(itemset) Un itemset* $\mathcal{I}$ *est un sous-ensemble d'attributs :*

$$\mathcal{I} = \{I_1, I_2, ..., I_n\}, \mathcal{I} \subset \mathcal{A}$$

Il est également noté *k*-itemset pour préciser son cardinal *k*, qui est appelé son *ordre*. Ainsi, un 1-itemset, c'est-à-dire un itemset d'ordre 1, correspond à un seul attribut.

Définition 2 *(support) Le support* $s(\mathcal{I})$ *d'un itemset* $\mathcal{I}$ *est le nombre d'occurrences de celui-ci dans la base de données.*

Constituant la mesure principale caractérisant un itemset, il est exprimé, soit par le nombre d'occurrences, soit, ce qui est le plus fréquent, en indiquant sa fréquence d'apparition. En d'autres termes, il donne une mesure de la généralisation de l'itemset dans la base. Par exemple, un itemset ayant une fréquence égale à 0,20 apparaît dans 20% des enregistrements de la base.[G. Bothorel, 2014]

2.4.2 REGLE D'ASSOCIATION

Une règle d'association est définie, à partir d'un itemset $\mathcal{I}$, par la relation $X \Rightarrow Y$ où $X \cup Y = \mathcal{I}$ et $X \cap Y = \emptyset$.

Cela peut se traduire par : « Si ***X*** est présent dans la transaction, alors ***Y*** l'est également ». Notons que ***X*** et ***Y*** peuvent être composés de plusieurs attributs, mais un attribut ne peut pas figurer simultanément dans les deux parties de la règle. La partie gauche de la règle s'appelle la *prémisse* ou l'*antécédent* ou le *corps*. La partie droite s'appelle la *conclusion* ou le *conséquent* ou la *tête*.

Définition 4 *(confiance) La confiance* ***c*** *de la règle d'association* $\boldsymbol{X \Rightarrow Y}$ *est*

définie par :

$$c(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$$

S'agissant de la mesure de base caractérisant une règle d'association, elle peut être assimilée à la probabilité conditionnelle *P(Y |X)*, c'est-à-dire la probabilité d'avoir *Y* sachant ***X***. La confiance peut donc être écrite de la manière suivante :

$$c(X \Rightarrow Y) = \frac{P(XY)}{P(X)}$$

Elle indique la validité de la règle. Pasquier et al. [Pasquier 99a] définissent la notion de *règle valide* si sa confiance est supérieure à un seuil. Elle est *approximative* si elle est inférieure à 1, et *exacte* si elle est égale à 1. Dans ce dernier cas, $P(XY)$ est égal à $P(X)$, ce qui signifie que si X est présent dans les transactions, alors Y l'est également. En d'autres termes, l'ensemble des transactions contenant X est inclus dans l'ensemble des transactions contenant Y.

Pour illustrer le support et la confiance, nous considérons la règle (*pain,saucisse*) ⇒ *hotdog*, pour laquelle ils sont égaux respectivement à 0,6 et 0,8. Les trois attributs *pain*, *saucisse* et *hot dog* sont ainsi présents simultanément dans 60% de la base de données, et 80% des transactions, contenant les attributs *pain* et *saucisse*, contiennent également l'attribut *hotdog [G. Bothorel, 2014]*

2.5 MESURE DE QUALITE DES REGLES D'ASSOCIATION

Afin d'extraire les règles d'association d'une base de données, il est nécessaire de fixer le support seuil, pour déterminer les itemsets fréquents, ainsi que la confiance seuil, pour trouver les règles valides. En fonction du besoin de l'utilisateur, la taille maximale de la règle peut être fixée, ainsi que la taille de la prémisse ou de la conclusion. Cependant, le support et la confiance ne sont pas toujours suffisants pour trouver des règles pertinentes.

En effet, en fonction des valeurs seuil, les algorithmes peuvent générer un nombre de règles très important, qui peut, dans certaines situations de seuil trop bas, dépasser le nombre de transactions initiales. De même, si le minimum est trop élevé, alors des règles intéressantes à faibles supports peuvent ne pas être détectées. De plus, ces deux mesures sont souvent insuffisantes pour prouver l'intérêt d'une règle, parce qu'elles ne prennent pas en compte $P(Y)$ ni les contre-exemples $P(X\bar{Y})$. Par exemple, si $c(X \Rightarrow Y) = P(Y)$, cela signifie que X et Y sont indépendants, parce que $P(X)P(Y) = P(XY)$.

Cette règle n'est donc d'aucun intérêt, même si le support et la confiance sont élevés.

Il est donc nécessaire de caractériser les règles d'association par des mesures supplémentaires dites de qualités ou d'intérêt. Celles-ci sont nombreuses et ont fait l'objet de multiples publications.

Piatetsky-Shapiro [Piatetsky-Shapiro 91a] a défini la notion de bonne mesure, en fonction de sa valeur par rapport à 0 [Lallich 04]. Ainsi, une bonne mesure est :

- Nulle dans le cas de l'indépendance ;
- Positive en cas d'attraction entre X et Y, c'est-à-dire dépendance positive :

- $P(XY) > P(X)P(Y)$;
- Négative en cas de répulsion entre X et Y , c'est-à-dire dépendance négative : $P(XY) < P(X)P(Y)$. **[G. Bothorel, 2014]**

2.5.1 LIFT

La mesure la plus connue est le ***lift***, défini par :

$$l(X \Rightarrow Y) = \frac{P(XY)}{P(X)P(Y)}$$

Il indique la dépendance entre la prémisse et la conclusion. S'il est inférieur à 1, alors la règle est considérée sans intérêt. S'il est égal à 1, alors, comme $P(XY) = P(X)P(Y)$, X et Y sont indépendants, c'est-à-dire que la présence de l'un n'apporte rien à la présence de l'autre. Puis, plus il est élevé, plus un lien entre X et Y est probable. Ainsi, si le lift est égal à **3**, cela signifie que $P(Y/X) = 3P(Y)$ et que $P(X/Y) = 3P(X)$, c'est-à-dire que si nous avons X, la probabilité d'avoir Y est trois fois plus grande que la probabilité d'avoir Y en général. Il en est de même en inversant X et Y . Il s'agit donc d'un indicateur de pertinence de la règle.

2.5.2 CORRELATION LINEAIRE DE PEARSON

Elle est définie par :

$$r(X,Y) = \frac{P(XY) - P(X)P(Y)}{\sqrt{P(X)P(\overline{X})P(Y)P(\overline{Y})}}$$

Elle permet de mesurer la force de la liaison entre X et Y. Si elle est nulle, alors cela signifie que X et Y sont indépendants.

Une valeur positive forte indique que X et Y sont corrélés.

Une valeur négative forte indique que X et Y sont corrélés négativement, c'est-à-dire que X et Y sont corrélés.

2.5.3 LOEVINGER

Elle est définie par :

$$LO(X \Rightarrow Y) = 1 - \frac{P(X\overline{Y})}{P(X)P(\overline{Y})} = \frac{P(Y/X) - P(Y)}{P(\overline{Y})}$$

Cette mesure est considérée comme un indice d'écart à l'indépendance et prend la valeur nulle en cas d'indépendance. Elle augmente au fur et à mesure que le nombre de contre exemples diminue, c'est-à-dire quand $P(X\overline{Y})$ décroît, pour atteindre la valeur 1 quand il n'y

en a plus. Elle décroît avec le support, et permet de rejeter des règles peu intéressantes, malgré une confiance élevée.[G. Bothorel, 2014]

2.5.4 CONFIANCE CENTREE

Elle est définie par :

$$CC(X \Rightarrow Y) = c(X \Rightarrow Y) - P(Y) = P(Y/X) - P(Y)$$

Dans le cas de l'indépendance, la confiance est égale à ***P(Y)***. En la recentrant par rapport à ***P(Y)***, la confiance centrée devient alors nulle à l'indépendance, ce qui est vrai quelle que soit la probabilité de ***Y*** .

2.5.5 CONVICTION

Elle est définie par :

$$CO(X \Rightarrow Y) = \frac{P(X)P(\overline{Y})}{P(X\overline{Y})} = \frac{1 - P(Y)}{1 - c(X \Rightarrow Y)}$$

Brin et al. [Brin 97b] ont créé cette mesure, car le lift ne mesure qu'une cooccurrence de ***X*** et ***Y*** et pas une implication. En effet, ***l(X ⇒Y) ⇒l(Y ⇒X).*** La conviction est une mesure d'écart à l'indépendance où elle est égale à 1. Dans le cas où ***X ⇒Y*** est toujours vérifié, alors ***P(X$\overline{Y}$)*** est nulle et la conviction est infinie. Elle mesure donc bien l'implication. De plus, elle est un indicateur du nombre de contre-exemples d'une règle, car, s'il augmente, alors la conviction diminue.

2.5.6 SEBAG ET SCHOENAUER

Elle est définie par :

$$SS(X \Rightarrow Y) = \frac{P(XY)}{P(X\overline{Y})} = \frac{c(X \Rightarrow Y)}{1 - c(X \Rightarrow Y)}$$

Comme dans le cas de la conviction, quand le nombre de contre-exemples augmente, sa valeur diminue. Si ***SS*** est égale à 3, alors ***P(XY) = 3 P(X$\overline{Y}$),*** ce qui signifie qu'en cas de présence de ***X***, le nombre de chances d'avoir ***Y*** est 3 fois plus élevé que le nombre de chance de ne pas l'avoir. Autrement dit, si ***X*** est présent, alors la quantité de chances d'avoir ***Y*** est de 75%.

A l'indépendance, la mesure est égale à $\frac{P(Y)}{P(\overline{Y})}$ [G. Bothorel, 2014]

2.5.7 PIATETSKY-SHAPIRO

Cette mesure, appelée Rule Interest par Piatetsky-Shapiro, est définie par :

$$PS(X \Rightarrow Y) = n(P(XY) - P(X)P(Y)) = nP(X)(c(X \Rightarrow Y) - P(Y))$$

A l'indépendance, sa valeur est nulle. Comme le lift, il s'agit d'une mesure symétrique qui est donc la même que pour la règle $\boldsymbol{X \Rightarrow Y}$

2.5.8 MULTIPLICATEUR DE COTE

Il est défini par :

$$MC(X \Rightarrow Y) = \frac{P(XY)P(\overline{Y})}{P(X\overline{Y})P(Y)}$$

Cette mesure d'écart à l'indépendance est une variante de celle de **Sebag & Schoenauer :**

$$MC(X \Rightarrow Y) = \frac{P(\overline{Y})}{P(Y)} SS(X \Rightarrow Y)$$

Cela lui permet d'être égale à 1 à l'indépendance.

Elle peut également s'exprimer en fonction du **lift** et de la **conviction** :

$$MC(X \Rightarrow Y) = l(X \Rightarrow Y)CO(X \Rightarrow Y)$$

2.5.9 ZHANG

Elle est définie par :

$$ZH(X \Rightarrow Y) = \frac{P(XY) - P(X)P(Y)}{Max\{P(XY)P(\overline{Y}); P(Y)P(X\overline{Y})\}}$$

Elle est nulle à l'indépendance.

2.5.10 SURPRISE

Elle est définie par :

$$SU(X \Rightarrow Y) = \frac{P(XY) - P(X\overline{Y})}{P(Y)}$$

A l'indépendance, elle est égale à

$$-2P(X)-\frac{P(X)}{P(Y)}.$$

2.5.11 PEARL

Le mesure est définie par :

$$PE(X \Rightarrow Y) = P(X)|P(Y/X) - P(Y)| = P(XY) \pm P(X)P(Y)$$

A l'indépendance, elle est nulle. Elle rappelle la mesure de **Piatetsky-Shapiro**, mais ne prend pas en compte l'effectif de la base et ne différencie pas l'attraction et la répulsion.

2.5.12 IMPLICATION

Elle est définie par :

$$IM(X \Rightarrow Y) = \sqrt{n}\frac{P(X\overline{Y}) - P(X)P(\overline{Y})}{\sqrt{P(X)P(\overline{Y})}}$$

A l'indépendance, elle est nulle. Cette mesure est utile pour étudier les contre-exemples, car elle augmente au fur et à mesure que leur nombre augmente.[G. Bothorel, 2014]

2.5.13 J-MESURE

Elle est définie par :

$$JM(X \Rightarrow Y) = P(XY)log(\frac{P(XY)}{P(X)P(Y)}) + P(X\overline{Y})log(\frac{P(X\overline{Y})}{P(X)P(\overline{Y})})$$

A l'indépendance, elle est nulle. Elle prend en compte la généralité de la règle et sa capacité de prédiction [Lallich 04]. Son évolution en fonction des contre-exemples ne correspond pas à une fonction monotone. Donc il n'est pas aisé de déterminer si la règle est de meilleure qualité que son contre-exemple [Gras 10]. Une autre particularité de cette mesure est qu'elle a la même valeur pour la règle et pour son contre-exemple.

2.4.14 INTENSITE D'IMPLICATION

Elle est définie par :

$$II(X \Rightarrow Y) = P[Poisson(nP(X)P(\overline{Y}) \geq nP(X\overline{Y})]$$

L'implication d'intensité mesure la surprise statistique de trouver la règle.[G. Bothorel, 2014]

2.6 APPLICATION DES REGLES D'ASSOCIATION

L'extraction de règles d'association a pour but d'identifier les relations significatives entre les données des base de donnés. Les relations ainsi identifiés peuvent être utiles pour de nombreux organismes commerciaux, scientifiques, industriels et gestion de l'information, afin d'améliorer leurs résultats dans leur activités. Plusieurs systèmes de KDD utilisant l'extraction des règles d'association ont été utilisés pour des applications réelles dans diverses domaines. Nous présentons dans la suite une liste non exhaustive des applications dont les résultats ont put être améliorées par l'analyse des règles d'association extraites.

2.6.2 PLANIFICATION COMMERCIALE

L'identification des articles achetés fréquemment ensemble apporte une aide importante dans le placement des articles. Un problème proche de celui-ci est la définition des catalogues.

Les règles d'association permettent aux sociétés de vente par correspondance de déterminer quels articles il est préférable de placer sur la même page d'un catalogue.

Ces informations sont aussi utilisées afin de déterminer quels articles en promotion pourront inciter les clients à effectuer d'autres achats Dans le cas de transactions de ventes dans lesquelles le client est identifié, les règles d'association permettent de définir des catalogues personnalisés en se basant sur les achats précédents du client.

Elles permettent également de réduire les couts des mailings en identifiant les clients les plus susceptibles de répondre à chaque mailing selon leurs achats précédents.

2.6.3 RESEAUX DE TELECOMMUNICATIONS

Les bases de données d'alarmes détectées dans les réseaux de télécommunications sont constituées de rapports de situations anormales dans les composants des réseaux.

Classiquement, ce sont plusieurs milliers d'alarmes qui sont détectées chaque jour, plusieurs milliers de types d'alarmes pouvant être distingué. Dans ce cadre, les règles d'association ont été utilisées avec succès dans le système TASA pour le filtrage des alarmes non informatives, l'identification des causes d'anomalie, et la détection et la prédiction d'incidents dans les processus de télé-maintenance, afin de limiter les couts des interventions manuelles et d'améliorer la qualité du service.

2.6.4 RECHERCHE MEDICALE

La plupart des organismes médicaux (hôpitaux, laboratoire d'analyse, cabinets médicaux, ect) stockent systématiquement les informations relatives à leurs patients dans des bases de données. Ces informations sont les résultats de consultations auprès des médecins,

les résultats de mesures indiquant la condition du patient et de données sur l'évolution de la condition du patient pendant le traitement.

L'extraction de règles d'association dans ces bases de données permet d'apporter une aide au diagnostic en identifiant les symptômes ou maladies précurseurs d'une maladie.

Une aide dans la définition de traitements en déterminant les symptômes ultérieurs ou les effets secondaires possibles, l'identification de population à risque vis-à-vis de certaines maladies, etc.

Les règles d'association ont également été utilisées dans le cadre de la prédiction de résultats d'analyses médicales. Les règles d'association extraites ont permis d'identifier les analyses fréquemment pratiqués sur les mêmes patients, et de prédire les résultats de certaines analyses par combinaison de caractéristiques des patients et de résultats d'autres analyses.[N. Pasquier, 2010]

2.6.5 ANALYSE DE DONNEES SPATIALES

Les bases de données spatiales sont largement utilisées dans les systèmes d'information géographiques, en cartographie, en astronomie et pour les études de l'environnement. Elles stockent des informations spatiales et non-spatiales relatives aux objets (forme, dimensions, positions, couleurs, température, etc.) qui occupe un espace. Du fait du développement des outils automatiques d'acquisition et des outils d'acquisition à distance leur nombre et leur volume ne cesse de croitre. Afin de découvrir des informations utiles enfouie dans ces bases de données, des techniques d'analyse de grands volumes de données, telle que l'extraction de règles d'association, sont nécessaires. Les règles d'association extraites depuis ces données définissent des relations entre des caractéristiques spatiales ou non-spatiales des objets. Elles ont été utilisées, notamment dans le système GeoMiner, pour l'aide à la prédiction d'événements naturels (éruptions, tremblements de terre, ouragans, etc.) et l'aménagement du territoire, pour prévision météorologique et les études biologiques, démographique et géographiques. [N. Pasquier, 2010]

2.6.6 MULTI –MEDIA ET INTERNET

Des quantités croissantes de données de diverses types (images, audio, vidéo, etc.), appelées données multi-médias, sont stockées dans des bases de données dont le nombre ne cessent d'augmenter. L'extraction de règles d'association à partir de données multi-médias a donné lieu de nombreuses études, principalement dans le cadre de l'analyse d'images. Las applications concernant la reconnaissance militaire, le filtrage des données parasites, la prévision météorologique, l'imagerie médicale, l'aide dans les enquêtes criminelles, etc. de même, un grand nombre de ressources sont accessibles par le réseau internet et un nombre important d'accès à ces données sont réalisés chaque jour par des millions d'utilisateurs. La taille et le nombre croissants des sites internet entrainent d'importants besoins d'outils pour la réorganisation de ces sites en fonction des cheminements des usagers, l'aide à la navigation

dans les systèmes de gestion d'informations, la recherche et la sélection des sites (moteurs de recherche), etc. l'extraction des règle d'association à partir des historiques des accès par les usagers aux ressources des sites Internet ont été utilisées dans ce cadre pour l'aide à la conception et l'organisation des sites. [N. Pasquier, 2010]

2.6.7 Analyse de données statistiques

L'analyse de données statistiques constitue un défi important pour le KDD de par sa difficulté et l'intérêt des informations qui peuvent être extraites de ces données. La difficulté provient e la nature des données statistique, qui sont fortement corrélées et denses, ce qui pose d'important problèmes d'efficacité. L'intérêt tient au nombre d'applications pouvant bénéficier de l'analyse de données statistiques qu'elles utilisent. Les organismes financiers, de recherche et les administrations stockent de nombreuses données de ce type (résultats de recensements, de sondages et d'étude par exemple). L'analyse de ces données constitue une part importante de l'activité de ces organismes, et les règles d'association peuvent constituer des indicateurs utiles dans ce cadre.

L'extraction de règles d'association a été utilisée pour de nombreuses autres applications, car elle constitue un module important des systèmes de KDD commercialisés, parmi lesquels on Fraser (Canada).[N. Pasquier, 2010]

2.7 REGLES D'ASSOCIATION TRIADIQUE

Définition

Une implication triadique a la forme suivante : $(X \Rightarrow Y)$ c. Cette implication est vraie si

"Chaque fois que **X** est vrai sous toutes les conditions dans **C**, alors **Y** est aussi vrai sous toutes ces conditions"

2.7.1 CONTEXTE DYADIQUE ET CONTEXTE TRIADIQUE

Contexte triadique : En analyse formelle de concepts, un contexte triadique est un quadruplet **K := (R,U, A, Y) où R, U, A et Y** .

- **R, U,A** définissent respectivement des Requêtes, des Utilisateurs et des Attributs (descripteurs et mesures)

- **Y = R×U ×A** représente une relation ternaire, où chaque **y ⊆ Y** représente un triplet : **y = {(r, u, a)|r ⊆ R, u ⊆ U, a ⊆ A}.** Autrement dit, une requête r est lancée par un utilisateur **u** et qui implique l'attribut a.[6]

Contexte dyadique : Un contexte formel dyadique est un triplet $\mathbf{K^{(1)}}$ **:=(G,M, I)** où **G** est un ensemble d'objets, **M** un ensemble d'attributs et **I** une relation binaire entre **G** et **M** . Le

contexte dyadique obtenu après aplatissement du contexte triadique que nous avons défini est formé de **G = R** et **M = U × A.**

Le tableau 1 représente le contexte dyadique $\mathbf{K^{(1)}}$ obtenu à partir du contexte triadique

K ainsi : $\mathbf{K^{(1)}}$) := (R,U × A, $Y^{(1)}$) avec ((a_i, (a_j , a_k)) ∈ $Y^{(1)}$ ⇔ (a_i, a_j , a_k)∈ Y).

valeur 1 pour la première ligne et la première colonne 1 signifie que l'utilisateur U_1 lance la requête R_1 qui implique l'attribut a_1 dans la table suivante **(a)** le contexte dyadique et **(b)** le contexte triadique [6]

𝕂	U_1	U_2	U_3	U_4
R_1	$a_1a_2a_4$	$a_1a_2a_4a_5$	a_1a_3	a_1a_5
R_2	$a_1a_4a_5$	$a_2a_3a_4$	$a_1a_2a_4a_5$	a_4a_5
R_3	$a_1a_2a_4$	a_4a_5	a_1a_2	a_1a_5
R_4	$a_1a_2a_4a_5$	a_2a_4	a_1a_2	a_4a_5
R_5	$a_1a_4a_5$	$a_1a_4a_5$	$a_1a_2a_4a_5$	a_1a_5

(a)

$\mathbb{K}^{(1)}$	U_1					U_2					U_3					U_4				
	a_1	a_2	a_3	a_4	a_5	a_1	a_2	a_3	a_4	a_5	a_1	a_2	a_3	a_4	a_5	a_1	a_2	a_3	a_4	a_5
R_1	1	1		1		1	1		1	1	1		1			1				1
R_2	1			1	1		1	1	1		1	1		1	1				1	1
R_3	1	1		1					1	1	1	1				1				1
R_4	1	1		1	1		1		1		1	1							1	1
R_5	1			1	1	1			1	1	1	1		1	1	1				1

(b)

Table 2 contexte dyadique et contexte triadique

Conclusion

Dans ce chapitre, nous avons présenté les règles d'associations, leurs mesures de qualité et les règles d'association triadique. Dans le chapitre suivant nous essayons de présenter une vue générale de processus de personnalisation des requêtes décisionnelle par les règles d'association

INTRODUCTION

La personnalisation de l'information constitue un enjeu majeur pour l'industrie informatique, son but est de faciliter l'expression du besoin de l'utilisateur et de lui permettre d'obtenir des informations pertinentes lors de ses interactions avec un système d'information.

La pertinence de l'information délivrée et son adaptation aux préférences des utilisateurs sont des facteurs clés du succès ou du rejet d'un tel système. Pour rendre la personnalisation effective, on a besoin de collecter et sauvegarder l'ensemble des critères et des préférences personnalisables spécifiques à chaque utilisateur.

Ces données sont souvent regroupées sous forme de profil. Le contenu du profil d'un utilisateur peut varier selon les approches et les applications. Le profil de l'utilisateur peut être construit explicitement ou implicitement. Dans l'approche explicite, l'utilisateur fourni lui-même les informations le concernant (données personnelles, préférences, etc.). Dans l'approche implicite, les informations du profil sont acquises par des techniques d'apprentissage ou de datamining exploitant les historiques des actions et des choix passés des utilisateurs.

Lors de notre recherche bibliographique, nous avons remarqué et constaté beaucoup de travaux de recherche qui travaillent sur l'optimisation des requêtes c.à.d. chercher à réduire le nombre de réponses d'une requête posée par l'utilisateur en introduisant les préférences des utilisateurs par des recherches avancées comme pour Google ; Facebook ; Yahoo ; …etc. cependant nous n'avons pas constaté des travaux qui répond à ce souci par des architectures ou des approches génériques permettant la prise en charge de profil de l'utilisateur.

Ainsi ce travail propose une solution générique comportant une architecture d'intégration des préférences des utilisateurs dans la formalisation des requêtes décisionnelles en se basant sur les règles d'associations.

1.1 DEFINITION :

Le terme **personnalisation** est utilisé pour expliquer comment recevoir à partir d'une grande quantité d'informations seulement la partie qui intéresse un individu ou un groupe d'individus. Pour connaître ce qui intéresse l'utilisateur, il faut connaître son profil (ses intérêts, ses préférences ou même ses contraintes, ses comportements, etc.).

Le terme personnalisation est alors utilisé pour la prise en compte du profil utilisateur.

1.2 LES DOMAINES DE LA PERSONNALISATION

Parmi les domaines qui font appel à la personnalisation :

- **Commerce électronique**

Ce domaine d'application recouvre la vente de produits de toute nature y compris les contenus multimédias « à la demande » pour lesquels la démarche commerciale tire un grand profit des techniques de personnalisation. Les apports de la personnalisation s'échelonnent de la mémorisation des données personnelles pour éviter leur re-saisie systématique jusqu'à la recommandation de produit basée sur les achats précédents et/ou le comportement de l'utilisateur sur le site web de vente.

- **La dissémination sélective d'information**

Concerne la diffusion d'information culturelle et d'actualité, la personnalisation permet le filtrage des informations en tenant compte d'un profil traduisant le centre d'intérêt, la langue, la religion et la position géographique de l'utilisateur.

- **Apprentissage assisté par ordinateur**

Ce domaine d'application concerne tous les environnements informatiques pour l'apprentissage humain ou de veille technologique. La personnalisation permet de définir des objectifs et des formations sur mesure (selon les connaissances, le style d'apprentissage préféré, etc.) et de suivre l'apprenant aux cours de sa formation afin d'adapter la réaction du système d'apprentissage à son état d'avancement et à son comportement.

- **Accès aux bibliothèques électroniques**

La personnalisation s'applique à différents niveaux:

- Limiter l'accès aux seuls documents auxquels l'abonné a souscrit.
- Guider la navigation de l'utilisateur au sein de ces documents selon sa requête du moment et recommander les nouveautés en fonction du profil.

- **Configuration des logiciels (réseau, composantes)**

La technologie informatique elle-même fait appel de plus en plus aux techniques de la personnalisation telle que la configuration de systèmes d'exploitation, de protocole réseaux ou de services web en fonction des besoins des utilisateurs.

- **Systèmes d'informations mobiles**

Les services accessibles via les téléphones mobiles et les systèmes embarqués requièrent d'un coté une personnalisation leur permettant de limiter l'effort et le temps passé à la recherche de l'information pertinente, et une adaptabilité aux contraintes physiques ou techniques [S.A. Selmane, F. Bentayeb, O. Boussid, 2014]

(écran de petite taille, clavier absent ou réduit, bande passante limitée, etc.). Les vendeurs de service présentent souvent ces techniques comme des moyens d'aide à la décision pour l'utilisateur.

3.3 PROCESSUS DE PRISE EN COMPTE DE L'UTILISATEUR DANS LES SYSTEMES D'INFORMATION

La prise en compte de l'utilisateur dans un système d'information doit suivre un

processus bien défini qui peut différer d'un domaine à l'autre mais qui se présente

généralement comme illustré dans la Figure 10. En effet, la prise en compte de l'utilisateur dans un système d'information comporte deux grandes phases à savoir

(1) la définition du profil utilisateur et (2) l'exploitation de ce dernier. La première phase se réalise par la collecte d'information concernant l'utilisateur. Cette collecte peut être implicite ou explicite. La deuxième phase à savoir l'exploitation du profil utilisateur déjà défini peut se dérouler par deux manières différentes : la personnalisation ou la recommandation. [8]

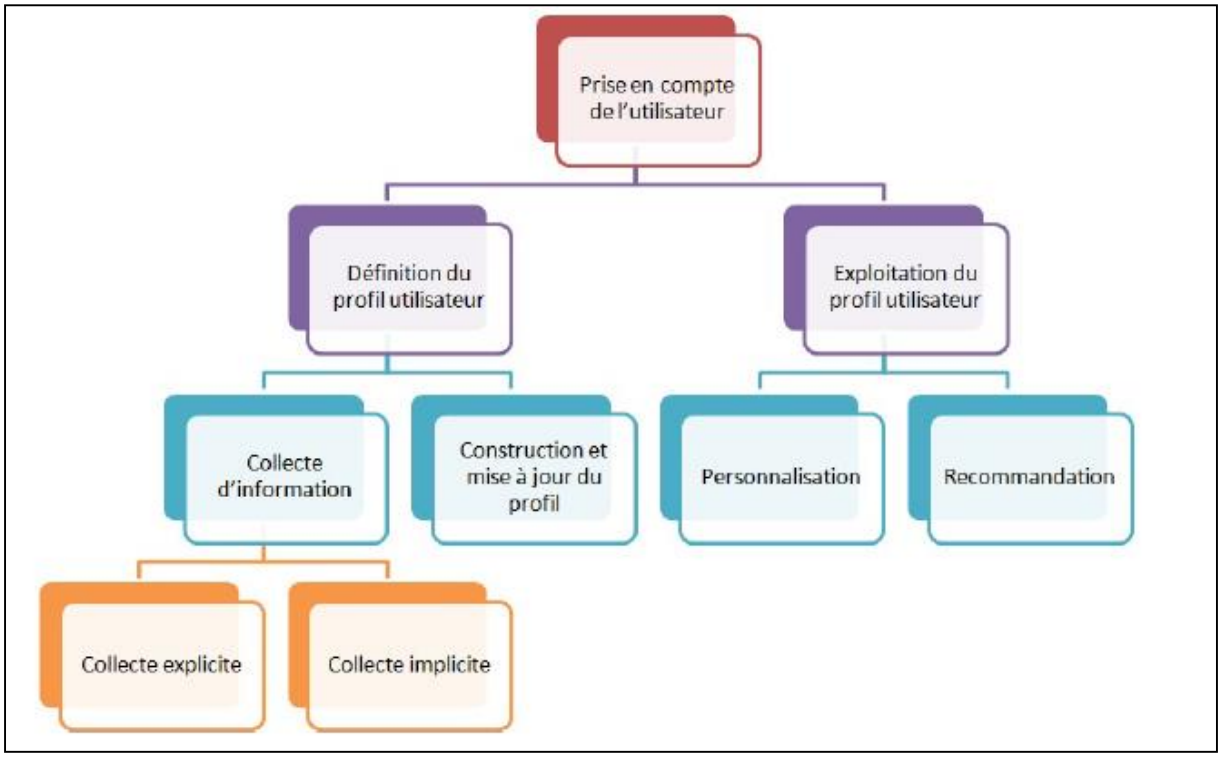

Figure. 10 Processus de personnalisation [S.A. Selmane, F. Bentayeb, O. Boussid, 2014]

3.4 VERS L'ADAPTATION DU PROCESSUS DE PERSONNALISATION PAR LES REGLES D'ASSOCIATION

Notre travail que nous proposons se base de la personnalisation des règles plus compact appelées règles d'association triadique, ces règles ont un sémantique plus riche que les règles

d'association classiques car elles sont formées en plus de la prémisse et de la conclusion, d'une condition qui s'ajoute à la règle.

Pour la prise en charge les préférences de l'utilisateur dans les requêtes décisionnelles Nous présentons dans La figure 11 une vue générale sur le processus de personnalisation par les règles d'association :

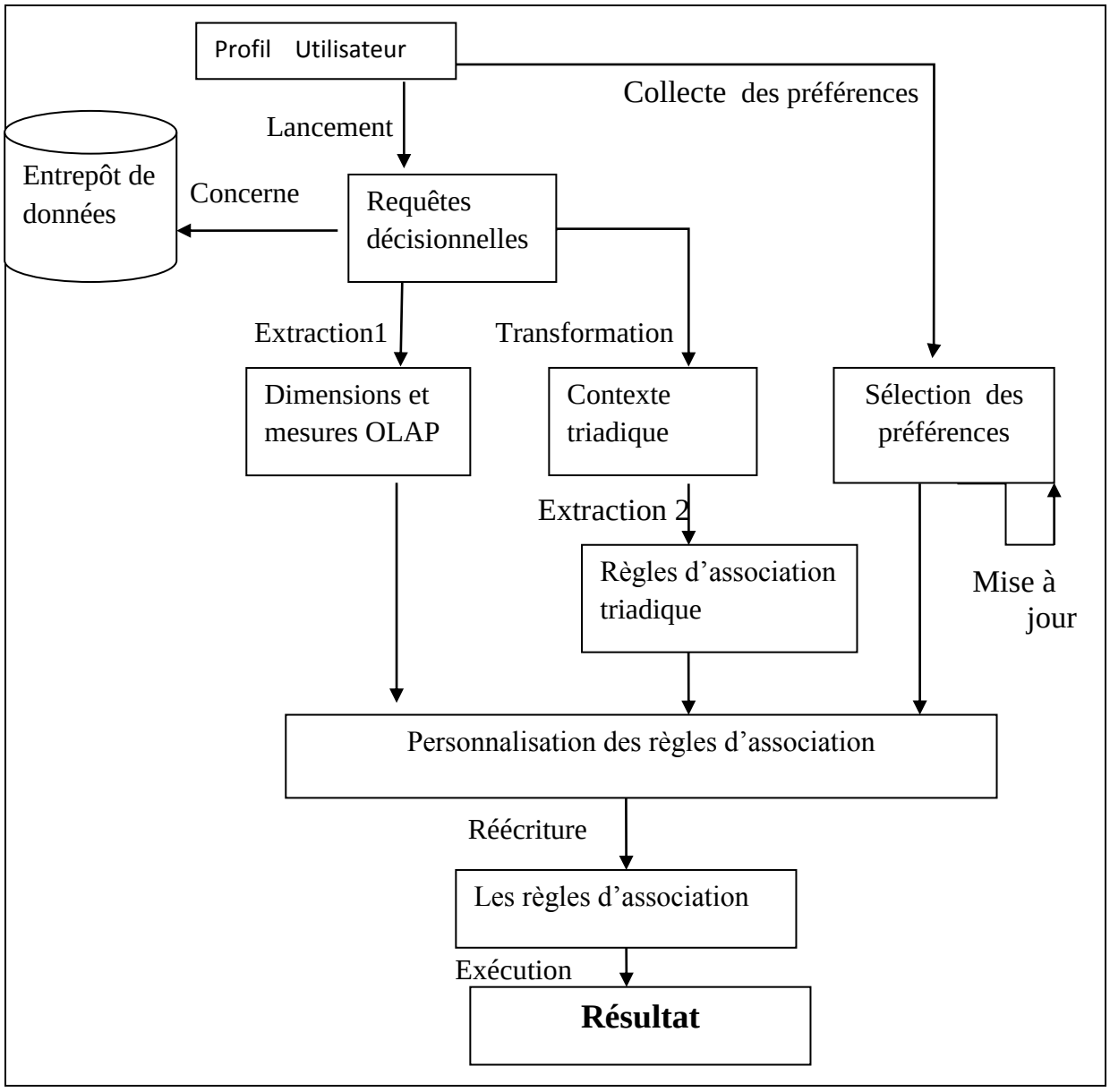

Figure 11 Vue générale sur le Processus de la personnalisation par les règle d'association

Notre processus de personnalisation est composé de cinq étapes :

1. Lancement :

L'utilisateur obligé de poser plusieurs requêtes afin d'obtenir un résultat le plus proche possible de son besoin , les requêtes peuvent être extraite de deux manières : la première est automatique ou implicite parce que l'utilisateur n'a pas besoins de préciser les informations sur son profile, ces information peuvent être extraite a partir des actions faites par l'utilisateur exemple les fichiers log, la deuxième est manuelle ou explicite ou l'utilisateur doit saisir manuellement les informations concernant son profil.

Il faut noter que la notion de profil est différente de celle du log. Le log est composé de requête ou de sessions de requêtes, alors qu'un profil est composé d'éléments de requêtes auxquels sont rajoutées éventuellement des caractéristique de l'usager telles que ses données démographiques, dans notre approche nous proposons d'écrire ou sauvegarder les requêtes dans un fichier à part séparée du fichier log du système.

La modélisation du fichier sauvegardé du serveur d'analyse OLAP avec un contexte triadique : il sera formé en trois ensembles : l'ensemble des utilisateurs, l'ensemble des requêtes, l'ensemble des attributs issus de la clause SELECT (descripteurs et mesures) des requêtes et d'une relation ternaire qui lie ces trois ensembles.

2. Concerne :

Il est parfois difficile aux utilisateur de traduire leurs besoins d'analyse par des requêtes textuelles structurées ou graphiques Ceci nécessite une maitrise d'un langage d'interrogation et une compréhension approfondie du schéma multidimensionnel, alors que l'utilisateur qui est généralement non informaticien, ne peut pas maitriser parfaitement le schéma multidimensionnel des entrepôt de données, Ainsi les systèmes OLAP doivent faciliter la tache de l'utilisateur en l'assistant durant le processus d'analyse décisionnelle.

3. Sélection des préférences :

Une étape de sélection des préférences est nécessaire pour déterminer celles qui seront utilisées dans le processus de personnalisation.

Une première méthode est entrée sur l'applicabilité de la préférence. Une préférence P est applicable à une requête Q si l'exécution de Q combinée conjonctivement avec P ne renvoie pas un résultat vide.

Une deuxième méthode à été proposé en bases de données ou les préférences sont sélectionnées si leurs contextes apparient avec le contexte de la requête.

- Si une préférence P est rattachée a un contexte interne C, la sélection de P dépend d'une confrontation entre le contexte C et le tuple de la base de données ou les attributs de la requête.
- Dans le cas de contexte externe, le contexte courant de l'usager CC(U) est d'abord détecté. Une préférence est sélectionnée si son contexte apparie avec CC(U). Par

exemple, si l'usager est localisé à toulouse au moment de la requête, les préférences qui sont associés à l localisation France sont sélectionnes.

4. Extraction 1 :

L'extraction de dimension et mesure OLAP, les dimensions contiennent chacune une liste d'attributs, où chacun d'eux est mappé vers une propriété de classe. Les dimensions permettent de filtrer, regrouper et d'étiqueter les données par exemple, nous pouvons filtrer les ordinateur par système d'exploitation, et les groupes de personnes des catégories par âge ou par sexe, les données peuvent etre classées par hiérarchies naturelles et catégories afin de permettre une analyse plus approfondie, les dimensions peuvent également avoir des hiérarchies naturelles permettant aux utilisateurs d'obtenir un plus grand niveau de détail, par exemple, la dimension date possède une hiérarchie pouvant etre détaillée par année, trimestre, semaine et jour. Les mesures sont des valeurs numériques que les utilisateurs peuvent découper, agréger et analyser. Elles constituent l'une des principales raisons pour lesquelles les cubes OLAP doivent etre créés à l'aide de l'infrastructure d'entreposage des données, les mesures sont des valeurs qui sont généralement mappées vers les colonnes numériques d'une table de faits de l'entrepôt de données, mais peuvent également être créées dans les attributs des dimensions.

Lorsque l'utilisateur extrait des données, il a accès à toutes les transactions ayant contribué à l'agrégation des données de cube OLAP. Il peut récupérer des données d'avantage résumées pour une valeur de mesure donnée. Par exemple, lorsque il accède aux chiffres de vente relatifs à un mois et à une catégorie de produit en particulier, il peut extraire ces données pour afficher la liste de chacune des ligne de table contenues dans cette cellule de données.

L'extraction permet d'accéder directement à niveau de détail mois élevé, et récupérer un ensemble de ligne provenant d'une source de données et ayant été agrégées au sien d'une même cellule

5. Transformation :

L'utilisateur U lance une requêtes R qui contient plusieurs attribut a, cette relation ternaire est transformée e contexte triadique :

Le quadruplet K = (R, U, A, Y) où R définit l'ensemble des requêtes, U définit l'ensemble des utilisateurs, A définit l'ensemble des attributs et Y définit la relation ternaire.

6. Collection des préférences

Pour tous les systèmes de personnalisation développés jusqu'à nos jours, la collecte de données relatives aux utilisateurs représente une phase clé dans le processus de personnalisation. Il s'agit de déterminer comment sont collectées les informations liées aux utilisateurs et à leurs besoins. Cette collecte peut se faire de manière explicite ou implicite.

- **Collecte manuelle :**

La collecte manuelle d'information concorde à toute donnée qui a été saisie ou fournie directement par l'utilisateur. On l'appelle aussi déclaration puisque c'est à l'utilisateur de déclarer ses informations. C'est la méthode la plus directe pour acquérir des informations sur l'utilisateur. Généralement, le mode d'acquisition manuel des données est le plus facile à mettre en œuvre. En effet, au niveau le plus basique, l'utilisateur est invité à saisir manuellement des informations utilisées dans la construction de son profil. Vraisemblablement, la demande explicite peut être le meilleur moyen pour recueillir des informations de bonne qualité, qui peuvent refléter les besoins subtils et les préférences des utilisateurs. Cependant, la collecte manuelle est une tâche fastidieuse pour l'utilisateur qui a tendance à être prudent quant à l'intervention directe pour exprimer ses intérêts. L'effort supplémentaire imposé à spécifier explicitement ses besoins est indésirable pour les utilisateurs. Ils peuvent de plus avoir l'impression que le processus prend trop de temps pour une amélioration peu perceptible du service offert par le système. Certains utilisateurs développent une crainte d'être surveillés et évitent les systèmes les incitant à fournir des données, ou laissant entendre que leur comportement est poursuivi pour personnaliser le contenu. La crainte d'une violation de leur vie privée les évite de fournir des informations. D'autre part, si l'utilisateur fournit ses informations, la valeur de ce type d'information n'est pas toujours fiable car rien ne garantit la véracité de ses informations.

- **Collecte automatique**

Par opposition à la collecte explicite, la collecte de l'information peut se faire implicitement sans la sollicitation directe de l'utilisateur. Ainsi, les informations sont récoltées sans que l'utilisateur n'ait besoin de préciser ces informations (comme par exemple les traces d'usage dans les fichiers logs).

Dans ce type de collecte d'information, l'observation des interactions de l'utilisateur Avec le système est indirecte, elle repose sur des données ou des appréciations '' automatique '' déduites à partir des actions réalisées par cet utilisateur. Ces actions sont souvent appelées ''les traces d'usage''. Ces traces représentent une suite d'actions effectuées par un utilisateur, elles sont souvent déduites à partir des fichiers logs.

Ainsi, il est possible d'enrichir les informations déclaratives des utilisateurs grâce à l'enregistrement de données par des moyens techniques. Le procédé de collecte d'informations le plus important concerne en effet les fichiers log qui constituent une source importante des données de l'utilisateur. En effet, dans le contexte des bases de données, l'exploitation des logs de requêtes par exemple peut aider à personnaliser les requêtes futures des utilisateurs ou à proposer des recommandations.

Le mode d'acquisition automatique repose sur des techniques d'extraction des informations basées sur des mesures de pertinence implicite (fréquence, temps d'exploration, etc) appliquées sur l'historique d'interactions de l'utilisateur.

7. Extraction 2 :

L'extraction des règles d'association triadique est définit par Le passage de contexte triadique vers un contexte dyadique et enfin la production de règle d'association dyadique conventionnelle de type prémisse → conclusion. Ensuite, puis la génération de règle d'association triadique de type (prémisse → conclusion) (condition) qui peut etre de deux types (BCAAR et BACAR) comme nous avons précédemment défini dans le deuxième chapitre.

8. Mise à jour des préférences :

Chaque utilisateur peut changer les informations de ses préférences que ce soit les informations collectées manuellement (exemple un utilisateur change la langue de recherche de la lange française vers la lange araba veut dire que le profile de l'utilisateur a été changé.

9. Personnalisation des règles d'association

Après avoir déterminé les préférences ; le système introduit ces préférences dans la requête ceci doit être faite automatiquement par le système. L'objectif attendu est de réduire au maximum le nombre de réponses et l'évitement des réponses inutiles et qui ne conviennent pas avec le profil des utilisateurs.

Conclusion

Dans ce chapitre nous avons essayé de définir la personnalisation avec une vue générale du processus de personnalisation par des règles d'association qui est enfin, la précision de profile des utilisateurs ou leur préférence afin de les introduire dans les requêtes décisionnelles pour avoir des réponses plus précises

Conclusion Générale

Dans le premier chapitre nous avons présente des définitions sur les Système d'aide à la décision, et leur position dans l'organisme puis les phases d'un projet décisionnel puis les entrepôts de données et les magasins de données l'analyse et la restitution OLAP

.Dans le deuxième chapitre nous avons définir le Datamining, le processus datamining

Et les règles d'association dyadique et triadique et leur application et mesures de qualité

Dans le troisième chapitre nous avons définis la personnalisation et nous avons essayé de présenter une architecture de personnalisation des requetés décisionnelle