UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES
ET INFORMATIQUE APPLIQUÉES

PAR
HASSANE HILALI

APPLICATION DE LA CLASSIFICATION TEXTUELLE POUR L'EXTRACTION DES RÈGLES D'ASSOCIATION MAXIMALES

AVRIL 2009

Université du Québec à Trois-Rivières

Service de la bibliothèque

Avertissement

L'auteur de ce mémoire ou de cette thèse a autorisé l'Université du Québec à Trois-Rivières à diffuser, à des fins non lucratives, une copie de son mémoire ou de sa thèse.

Cette diffusion n'entraîne pas une renonciation de la part de l'auteur à ses droits de propriété intellectuelle, incluant le droit d'auteur, sur ce mémoire ou cette thèse. Notamment, la reproduction ou la publication de la totalité ou d'une partie importante de ce mémoire ou de cette thèse requiert son autorisation.

Résumé

Ce mémoire traite de la problématique de la classification textuelle dans son application à l'extraction des règles d'association. En effet, avec l'avènement de l'informatique et l'augmentation de données stockées sur des supports électroniques, qui peuvent cacher des dépendances et des corrélations pertinentes, la conception et la mise en œuvre des outils d'analyse et de traitement automatique de corpus deviennent une nécessité absolue.

La première partie de ce mémoire est consacrée à la présentation et la définition des différentes notions qui interviennent dans le développement de notre projet. En effet, le premier chapitre donne une vue détaillée des principales méthodes de classification qui existent, ainsi que les avantages et les inconvénients de chacune d'entre elles. Dans le deuxième chapitre, nous exposons les différentes caractéristiques des règles d'association et les étapes de leur extraction. Nous traitons dans le troisième chapitre d'un cas particulier des règles d'association que l'on nomme les règles d'association maximales. Le quatrième et le cinquième chapitre sont consacrés à la présentation du système développé ainsi qu'à l'interprétation des différents résultats obtenus par l'analyse d'un ensemble de documents.

En général, la chaîne de traitement de notre système se déroule en deux phases. La première partie consiste à faire la classification d'un texte encodé en Unicode. Ensuite, nous utilisons les classes obtenues dans la première partie pour générer un type de règles d'association appelées les règles d'association maximales.

Remerciements

Ce mémoire est le résultat d'un long travail de trois années de recherche qui m'ont permis d'exploiter et de découvrir le monde de la lecture et de l'analyse des textes assistés par ordinateur.

La rédaction de ce travail de maîtrise n'aurait pas pu voir le jour sans l'aide et la collaboration d'un grand nombre de personnes, chacun ayant apporté une aide précieuse à sa finalisation.

Tout d'abord, je désire vivement remercier M. Ismaïl Biskri, mon directeur de recherche, pour son appui, sa supervision et sa disponibilité pendant les différentes étapes de ce travail. Merci Ismaïl de m'avoir accordé votre confiance et d'avoir pris le temps de répondre à toutes mes questions.

Un très grand merci à mes parents toujours présents, par leur soutien et leur encouragement, dans les moments difficiles de la réalisation de ce mémoire.

J'adresse également mes remerciements à mes meilleurs amis : Yousef Aichour, Zahra Lachgar et Ali Jouki, qui m'ont apporté leur soutien moral pendant ces années d'études.

Un énorme remerciement pour mes deux frères adorés, Khalid et Mohammed, qui ont été toujours disponibles pour moi.

Enfin, j'adresse mes sincères remerciements aux membres du jury. Je les remercie pour leur patience ainsi que pour leur lecture attentive de ce travail.

Table des matières

Liste des tableaux

Liste des figures

Chapitre 1 - Introduction

Avec l'avènement de l'informatique et l'augmentation du nombre de documents électroniques stockés sur des supports électroniques et sur le Web, l'intervention des outils d'analyse et de traitement automatique des textes est devenu plus que nécessaire, pour assister et aider le lecteur à explorer et à dégager des informations pertinentes, qui facilitent la compréhension rapide des corpus.

Nous présentons dans ce mémoire la classification mathématique des textes dans son application à l'analyse et l'extraction des règles d'association maximales.

En effet, plusieurs projets de recherche présentent de nombreuses méthodes issues seulement de la classification, mais nous pensons que l'utilisation de la classification jumelée avec les règles d'association maximales comportera plusieurs avantages par rapport à l'utilisation seule de la classification. Parmi ces avantages, on peut citer par exemple :

1. la détection des dépendances et des corrélations utiles entre les mots des différentes classes de notre corpus.
2. l'extraction des connaissances cachées, souvent très pertinentes, à partir d'un grand volume de données.
3. la combinaison des avantages des deux approches, la première, de la classification, où les résultats sont toujours variables selon les paramètres choisis au début, et la

deuxième, des règles d'association, où tous les algorithmes doivent découvrir, dans la plupart du temps, les mêmes règles d'association.

Notre objectif de recherche consiste à mettre en place une passerelle qui permet d'intégrer le processus d'extraction des règles d'association avec celui de la classification textuel et cela, dans l'objectif de profiter des avantages des deux approches.

Le présent mémoire est articulé autour de sept chapitres. Nous présenterons dans le deuxième chapitre la définition de la classification. Ensuite, nous exposerons quelques méthodes qui permettent de faire la catégorisation d'objets. On peut diviser ces techniques en deux types. D'abord, les méthodes de classification supervisées, qui consistent à classer les documents en se basant sur un ensemble d'exemples pré classés. Le deuxième type de méthodes, appelées non supervisées, permet de classer des documents sans l'utilisation des classes de références ou de base. Dans ce dernier type de classification, le nombre de classes n'est pas connu à l'avance; les classes sont créées au fur et à mesure avec l'avancement du processus classificatoire.

Dans le troisième chapitre, nous définirons les règles d'association et les différentes mesures utilisées pour les extraire, comme par exemple le support et la confiance. Nous montrerons par la suite les principales étapes d'extraction d'une règle d'association. Ce chapitre mettra aussi en évidence l'algorithme APRIORI, qui représente la base des algorithmes d'extraction des règles d'association. Il exposera, entre autres, les différentes étapes du déroulement de cet algorithme. Enfin nous conclurons ce chapitre par la présentation des avantages et des inconvénients liés à l'utilisation des règles d'association.

Le quatrième chapitre présentera un cas particulier des règles d'association, qu'on appelle les règles d'association maximales. Nous avons opté pour cette méthode, en

l'intégrant à notre projet, car elle offre plusieurs avantages par rapport aux autres méthodes. En effet, les règles d'association régulières permettent de trouver beaucoup d'associations intéressantes, mais elles ne sont pas capables d'extraire des associations moins intéressantes, cachées dans les données, et qui peuvent aider et faciliter la compréhension d'un corpus.

Dans le même chapitre, nous présenterons en détail les étapes d'extraction des règles d'association maximales ainsi que l'algorithme qui permet de les calculer.

Nous conclurons ce chapitre par la présentation d'un exemple concret qui permettra de faciliter la compréhension du processus d'extraction des règles d'association maximales.

Le cinquième chapitre sera consacré à la présentation de notre projet et de la méthodologie suivie pour intégrer les règles d'association maximales au processus classificatoire de la première étape de notre analyse.

La première partie de ce chapitre débutera par une courte introduction qui expliquera la méthodologie suivie pour combiner la classification textuelle (un processus dynamique), avec les règles d'association maximales (un processus statique). Le second volet de ce chapitre mettra en évidence l'architecture de notre système, ainsi que ces différentes composantes.

En effet, notre système se compose de deux couches. La première couche permet de faire la classification d'un document en utilisant le classificateur de neurone ART1. Ensuite, la deuxième couche utilise les résultats obtenus au niveau de la première étape pour calculer les règles d'association maximales.

Enfin, nous présenterons d'une manière détaillée la chaîne de traitement de notre système. Cette étape consistera, entre autres, à montrer le fonctionnement de notre plateforme par le biais d'un ensemble de prises d'écran, accompagnés d'explications pour chaque étape.

Le sixième chapitre sera consacré à la présentation des résultats statiques obtenus par le traitement d'un ensemble de documents. Nous démontrerons entre autres, comment notre système permet d'extraire d'importantes connaissances relatives aux documents traités.

Finalement, nous traiterons en conclusion de quelques réflexions qui pourront nous aider à améliorer la qualité des résultats obtenus par notre système.

Chapitre 2 - Classification

La classification automatique est le processus qui permet d'analyser et d'organiser un ensemble de données, selon leurs caractéristiques, dans des classes de similarité. Elle se base principalement sur des représentations classiques de données dont les limites de traitement sont connues et, qui dans la plupart du temps, demande un temps de calcul énorme. [3]

C'est au début des années soixante et avec l'avènement de l'informatique que les méthodes de classification ont connu de nouveaux développements méthodologiques qui ont permis l'apparition d'algorithmes d'analyse et de classification automatique de données. [3]

Les différents outils de classification développés se distinguent principalement par la méthode utilisée dans la classification, ainsi que par les différentes unités d'information (mot, n-gram, phrase, paragraphe, etc..) choisis pour l'extraction des données.

En 1798, l'Académie française a donné une définition du mot classification dans la cinquième édition de son dictionnaire comme la «distribution en classes et suivant un certain ordre». On retrouve de nombreuses autres définitions, souvent complémentaires; par exemple (Turenne, 2001) a ajouté la notion de hiérarchie de classes à travers certaines propriétés communes pour la définition de la classification. [4]

La quantité énorme des documents et l'explosion des ressources textuelles non structurées ont suscité dès les années 80 beaucoup d'intérêt pour les différentes techniques

de traitement automatique de documents. (Turenne, 2001) a montré le besoin urgent de ces techniques et les résultats pertinentes qu'on peut dégager de ces derniers, comme par exemple la classification de documents, la recherche d'informations pertinentes dans un texte, le résumé de documents, le filtrage des emails, etc. [4]

On peut représenter le problème de la classification sous un angle purement mathématique ((Jalam, 2003) et (Jaillet, 2005)). Le but étant d'identifier une fonction mathématique capable d'affecter les bonnes classes pour chacun des documents [4]. On peut représenter la fonction de classification Ω de cette façon [4] :

$$\Omega : A \times C \rightarrow (Vrai, Faux) \tag{2.1}$$

Où A est l'ensemble des documents et C est l'ensemble des classes. Cette fonction retourne vrai si le document A_i appartient à la classe C_k.

2.1 Les premières méthodes de classification

Les premières approches de classification avaient comme objectif d'utiliser les techniques qui existaient déjà en les adaptant aux besoins fonctionnels. [4]

[Anton 1988] a démontré après des observations que l'approche la plus naturelle était de décrire chaque document par un ensemble de mots-clés choisis par les experts du domaine. Mais cette méthode demeure très limitée et son utilisation a démontré certaines faiblesses. (Rocchio, 1971) et (Salton, 1975) ont développé des formules qui se basent principalement sur les occurrences des mots. Cette approche a été développée par la mise en œuvre de bases de règles (Serradura ,2002) et de systèmes experts ((Hardt, 1988) et (Vleduts, 1987)). Mais l'approche la plus connue est celle qui a été démontrée dans le

projet «CONSTRUE», réalisé par le «Carnegie Group», et dont la précision des résultats a été de 90 %. Le problème majeur rencontré dans l'utilisation de ces méthodes a été le temps de génération des catégories et leur faible capacité de construire de nouvelles classes. [4]

2.2 Les méthodes de classification supervisées

À cause de la grande quantité de documents échangés et stockés sur les supports électroniques, la classification automatique supervisée est devenue plus que nécessaire pour faciliter l'utilisation et l'analyse des données. À la différence de la classification non supervisée, où l'ordinateur doit trouver automatiquement les classes, la classification supervisée se base principalement sur le fait qu'il existe déjà une classification de documents, c'est-à-dire qu'on dispose d'un ensemble de données déjà classées qu'on appelle «ensemble d'apprentissage» et qu'on utilise comme base, pour classer le reste des données. On essaie dans ce type de classification de trouver le maximum d'informations à partir des ensembles d'apprentissage, pour permettre un meilleur groupement des données restant. [5,6]

Parmi les méthodes de classification supervisées les plus populaires, on peut citer par exemple [6]:

1. Les k plus proches voisins
2. les réseaux de neurones
3. les arbres de décision
4. les algorithmes génétiques
5. Naïve Bayes

6. Les machines à support de vecteurs.

2.2.1 *K plus proches voisins*

L'algorithme des k plus proches voisins sert dans plusieurs problèmes informatiques incluant la reconnaissance des formes, la recherche dans les données multimédia, la compression vectorielle, les statistiques informatiques et l'extraction des données.

Ce type de méthodes est très utilisé en algorithmique et plusieurs auteurs ont développé des algorithmes performants pour le résoudre. Il diffère essentiellement des autres méthodes par sa simplicité et par le fait qu'aucun modèle n'est introduit à partir des exemples pendant le processus de classification. [7]

Pour trouver la classe d'un nouveau cas, cet algorithme se base sur le principe suivant : il cherche les k plus proches voisins de ce nouveau cas, ensuite, il choisi parmi les candidats trouvés le résultat le plus proche et le plus fréquent. [6]

Cette méthode utilise principalement deux paramètres : [6] le nombre k et une fonction de similarité, qui permet de comparer les cas déjà classés avec le nouveau cas.

On peut représenter cette fonction de similarité par [6] :

$$d(y_i, y_j) = \sqrt{\sum_{k=1}^{n} (a_k(y_i) - a_k(y_j))^2} \qquad (2.1)$$

Cette formule représente la distance euclidienne entre deux documents y_i et y_j.

Avec

k : L'ensemble des attributs.

$a_k(y_i)$: Le poids du terme k dans le document y_i.

$a_k(y_j)$: Le poids du terme k dans le document y_j.

- **Exemple**

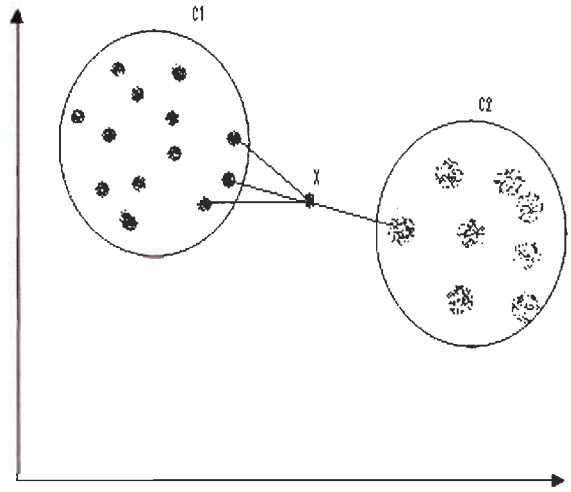

Figure 1 Distance euclidienne entre X et les deux classes c1 et c2

Dans l'exemple de la figure 1, on a deux classes de bases, c1 et c2. Le but est de classifier X dans l'une de ces deux classes.

Pour k=4 (c'est-à-dire on considère quatre plus proches voisins), on calcule la distance euclidienne entre l'élément X et les éléments de c1 et c2. Parmi les quatre plus proches voisins de X, nous avons trois qui appartiennent à c1 et un seul qui appartient à la classe c2, donc X est affecté à la classe majoritaire c1.

Il existe une autre fonction pour le calcul de la similarité, appelée cosinusoïdale, et qu'on définit par :

$$d(y_i, y_j) = \sum_{k=1}^{n} \frac{a_k(y_i) \times a_k(y_j)}{\sqrt{\sum_{k=1}^{n} a_k(y_i)^2 \times \sum_{k=1}^{n} a_k(y_j)^2}} \qquad (2.2)$$

Où

k : L'ensemble des attributs.

$a_k(y_i)$: Le poids du terme k dans le document y_i.

$a_k(y_j)$: Le poids du terme k dans le document y_j.

2.2.1.1 L'algorithme des k plus proches voisins

Soit : E un espace de dimension D

A un ensemble de N points dans cet espace

K un entier plus petit que N

On considère un point x de E qui n'appartient pas obligatoirement à A. La recherche des k plus proches voisins consiste à trouver quels sont les k points de A les plus proches de x. La définition complète de cet algorithme est [7]:

Pour i allant de 1 à k

Mettre le point D[i] dans proches_voisins

Fin Pour

Pour i allant de k+1 à N

Si la distance entre D[i] et x est inférieure à la distance d'un des points de proches_voisins à x alors :

1. On supprime du proches_voisins le point le plus éloigné de x
2. On met dans proches_voisins le point D[i]

Fin Si

Fin Pour

proches_voisins contient les k plus proches voisins de x

Figure 2.2 Algorithmes des k plus proches voisins [7]

2.2.1.2 Avantages de la méthode des k plus proches voisins

La méthode des k plus proches voisins présente plusieurs avantages parmi lesquels nous citons par exemple [56]:

1. La facilité de mise en œuvre de cet algorithme.
2. Son efficacité pour des classes réparties de manière irrégulière.
3. Son efficacité pour des données incomplètes.
4. La méthode des k plus proches voisins n'utilise pas de modèle pour classifier les documents.

2.2.1.3 Inconvénients de la méthode des k plus proches voisins

Le principal inconvénient de cette méthode est le temps d'exécution qu'elle met pour la classification d'un nouveau cas, car il faut calculer chaque fois la similarité entre les k exemples et le nouveau k, avant de décider quelle classe à choisir. [6]

Le deuxième inconvénient de cette méthode est la grande capacité de stockage qu'elle nécessite pour le traitement des corpus. [56]

En plus de ces deux inconvénients, l'algorithme des k plus proches voisins utilise de nombreuses données de références (les classes de bases) pour classifier les nouvelles entrées. [56]

2.2.2 *Les réseaux de neurones*

Depuis quelques années, les réseaux de neurones ont commencé à prendre de plus en plus une grande place dans divers domaines tels que: le traitement des signaux au niveau des télécommunications, la cryptographie, ainsi que le traitement des langues naturelles. Le principe de fonctionnement de ces réseaux est directement inspiré du fonctionnement de vrais neurones humains.

Le fait d'accepter que le cerveau humain fonctionne d'une façon totalement différente de celle d'un ordinateur a eu un impact très important sur le développement des réseaux de neurones. En effet, La quantité énorme des travaux effectués pour comprendre le fonctionnement du cerveau humain a mené la représentation de ce dernier par un ensemble de composantes appelées neurones, interconnectées les unes avec les autres. Le cerveau humain a la capacité d'organiser ces neurones, selon une organisation très complexe, non linéaire et extrêmement parallèle, afin d'accomplir des tâches très élaborées. [8]

Selon (Haykin ,1994) : «Un réseau de neurones est un processus distribué de manière massivement parallèle, qui a une propension naturelle à mémoriser des connaissances de façon expérimentale et de les rendre disponibles pour l'utilisation. Il ressemble au cerveau en deux points :

1. La connaissance est acquise au travers d'un processus d'apprentissage
2. Les poids des connexions entre les neurones sont utilisés pour mémoriser la connaissance». [8]

Cette définition est considérée comme une base pour l'élaboration des réseaux de neurones artificiels. [8]

De manière similaire à la nature, le fonctionnement d'un réseau de neurones est influencé par la connexion des éléments entre eux. On peut entraîner un réseau de neurones pour un rôle spécifique (traitement des signaux par exemple) en ajustant les valeurs des connections (ou poids) entre les neurones. [8]

Pour pouvoir utiliser les capacités d'un réseau de neurones dans la classification, il faut premièrement le construire. Ce processus se déroule en quatre étapes [9]:

1. Construire la structure du réseau.
2. Construire une base de données de vecteurs pour modéliser le domaine étudié, ce qui se fait en deux étapes : la première consiste en l'apprentissage du réseau et la deuxième aux différents tests de cet apprentissage.
3. La troisième étape consiste à paramétrer le réseau par apprentissage. Puisque les vecteurs de la base de données d'apprentissage sont présentés au réseau séquentiellement, un algorithme d'apprentissage interviendra pour ajuster les poids du réseau afin que les vecteurs soient correctement interprétés.
4. Reconnaissance : Cette phase consiste à utiliser une base de données de tests qui permettra de voir si les entrées de tests seront reconnues par le réseau construit.

Après l'exécution de plusieurs tests, si le réseau de neurones semble efficace dans le traitement des entrées, on peut l'utiliser pour de vraies applications.

Nous présenterons dans ce qui suit les propriétés et les algorithmes les plus connus des réseaux de neurones.

2.2.2.1 Un neurone formel

Le terme neurone formel désigne une fonction ou une méthode algébrique paramétré, à valeurs bornées, de variables réelles appelées des entrées. [10]

La valeur de cette fonction peut être calculée en deux étapes. La première étape consiste à calculer la combinaison linière des entrées par l'équation suivante [10] :

$$u = v_0 + \sum_{i=1}^{n} v_i y_i \tag{2.2}$$

Où :

v_i : Les poids synaptiques.

y_i : Les entrées du neurone.

v_0 : Appelé biais, qui est la pondération de l'entrée 0 fixée à 1.

u : Est le potentiel du neurone.

La deuxième étape permet de calculer la sortie du neurone par la fonction [10]:

$$y = f(u) = f(\sum_{i=0}^{n} v_i . y_i) \tag{2.3}$$

Où

u : Le potentiel du neurone.

v_i : Les poids synaptiques.

y_i : Les entrées du neurone.

On appel aussi cette fonction la fonction d'activation du neurone. Elle permet de définir l'état interne du neurone en fonction de son entrée totale. En effet, Chaque entrée y_i est multipliée par un poids synaptiques correspondant v_i.

Il existe plusieurs variantes de la fonction d'activation, parmi lesquelles on peut citer par exemple [10]:

1. La fonction sigmoïde (appelée aussi fonction logistique) :

$$f(u) = \frac{1}{(1+\exp(-u))} \tag{2.4}$$

La fonction logistique est une fonction bornée avec des valeurs réelles comprises entre 0 et 1. Elle possède deux propriétés importantes : elle n'est pas polynômiale et indéfiniment continûment dérivable, ce qui facilite le calcul de la dérivée et réduit le temps consacré à l'apprentissage du réseau de neurones.

2. La fonction identité :

$$f(u) = u \tag{2.5}$$

Dans les deux fonctions 2.5 et 2.6, u représente le potentiel du neurone.

La fonction identité est une fonction linéaire. Le calcul réalisé par le réseau de neurone en utilisant ce type de formules est similaire pour tous les neurones. Dans ce cas, le calcul par un seul neurone est suffisant pour donner des résultats équivalents.

2.2.2.2 Les réseaux de neurones non bouclés

On peut définir un réseau de neurones non bouclé comme une composition de fonctions construites par des neurones interconnectés les unes avec les autres. Il existe plusieurs combinaisons entre les neurones, mais l'architecture la plus connue est la perception multicouche dans laquelle les neurones sont organisés dans des couches. [10]

Dans ce modèle, chaque couche est construite de N_i neurones qui prennent leurs entrées sur les N_{i-1} neurones de la couche précédente. À chaque synapse est associé un poids synaptique, de sorte que les N_{i-1} neurones sont multipliés par ce poids, ensuite additionnés par les neurones de niveau i : cela est équivalent à multiplier le vecteur d'entrée par une matrice de transformation. [11]

Le processus qui permet de mettre les différentes couches les unes après les autres est équivalent à mettre en cascade plusieurs matrices de transformations. Ce processus pourrait être réduit à une seule matrice, produit des autres matrices, s'il n'y avait, à chaque couche, la fonction de sortie qui introduit une non linéarité. Donc, le choix d'une bonne fonction de sortie est très important, car un réseau de neurones dont les sorties seraient linéaires n'offrirait aucun intérêt. [11]

2.2.2.3 Apprentissage des réseaux de neurones

Le processus d'apprentissage des réseaux de neurones compte parmi les étapes les plus importantes pour construire un réseau efficace et fiable.

En général, le principe de l'apprentissage des réseaux de neurones se base principalement sur la logique suivante : pour une entrée particulière présentée au réseau, correspond une cible spécifique. Donc, on ajuste les connections (ou les poids) en faisant une comparaison entre la réponse du réseau et la cible, jusqu'à ce que la sortie corresponde au mieux à la cible. [12]

On peut définir aussi l'apprentissage par : «Le processus itératif et dynamique qui permet de changer les paramètres d'un réseau comme réponse à l'activation de l'environnement». [13]

Dans cette définition, on peut déduire qu'un réseau de neurones doit être activé par son environnement, qu'ils subissent des changements comme réponse pour l'activation, et que ceux-ci provoquent dans le futur une nouvelle réponse vis-à-vis de l'environnement. [13]

Dans le cas d'un exemple de régression, la tâche d'apprentissage consiste à approcher une fonction continue, tandis que dans le domaine de la classification supervisée, on doit trouver une surface de séparation. Il s'agit en général de minimiser une fonction de coût qu'on calcule à partir de la sortie du réseau de neurones et des exemples de la base d'apprentissage. [10]

L'apprentissage donne aux réseaux de neurones la capacité de réaliser des tâches très complexes dans différents domaines et types d'applications tels que : la classification et la reconnaissance de caractères. Dans la plupart de ces cas, les réseaux de neurones ont la

capacité d'apporter une solution simple pour des problèmes complexes qui ne peuvent pas être résolus rapidement par les systèmes actuels (insuffisance de la puissance de calcul ou le manque de connaissances). [14]

Le processus d'apprentissage se déroule en trois étapes [8] :

1. Un objet A est présenté à l'entrée du réseau de neurones, sans qu'on précise la classe à laquelle il appartient.
2. La recherche du neurone k'dont le vecteur de poids est le plus proche au vecteur d'entrée. On calcule k' par la fonction suivante:

$$k' : \min_{k} \frac{1}{2}(A - W_k)^T (A - W_k) \tag{2.6}$$

Où

k : Le nombre de neurones.

A : L'entrée présentée au réseau de neurone pour être classifiée.

W_k : Un vecteur de poids.

Cette fonction permet de comparer l'entrée A, présentée au réseau de neurone, à tous les vecteurs poids W_k. Le neurone gagnant k' est celui dont le vecteur de poids $W_{k'}$ est le plus proche de l'entrée A.

3. Cette troisième étape consiste à adapter le vecteur de poids de k' et aussi ceux de ces voisins topologiques pour qu'ils se rapprochent le plus possible du vecteur d'entrée :

$$W_k(\Gamma + 1) = W_k(\Gamma) + \mathrm{N}(A - W_k(\Gamma)) \quad \text{Si } k \in V(k') \tag{2.7}$$

Sinon

$$W_k(\Gamma+1) = W_k(\Gamma) \tag{2.8}$$

Où

A : L'entrée à classifier.

$\Gamma+1$: La fonction qui permet de faire la mise à jour des poids à l'instant $\Gamma+1$.

N : Représente le taux d'apprentissage. À chaque itération, cette fonction décroit au cours du temps, $0 \leq N \leq 1$.

2.2.2.4 Les fonctions d'activation

Les fonctions d'activation ou de seuillage servent essentiellement à introduire une non linéarité dans le processus de fonctionnement du neurone. [11]

Le fonctionnement des fonctions d'activation se déroule en trois phases [11]:

1. Si on est au-dessus du seuil, le neurone sera activé.
2. Si on est au-dessous du seuil, le neurone ne sera pas activé.
3. Si on est aux alentours du seuil, c'est une phase de transition.

Afin de bien comprendre le fonctionnement des réseaux de neurones, le reste de cette section sera consacré à la présentation d'un ensemble de méthodes et de fonctions de ce modèle.

2.2.2.5 Les réseaux de Perceptron

Développé par Rosenblatt en 1958, ce réseau est considéré parmi les premiers réseaux de neurones. Il a deux caractères, soit d'être monocouche et linéaire. Sa première couche d'entrée représente la rétine, tandis que les neurones de la couche suivante sont appelés les cellules d'association, enfin, la dernière couche représente les cellules de décision. [15]

Dans ce type de réseaux, seulement les poids de connexion entre la dernière couche et la couche d'association peuvent être modifiés, tandis que les sorites des neurones ne peuvent prendre que deux états : -1 et 1 ou 0 et 1. [15]

Ce réseau de neurones utilise comme méthode de modification les poids de la règle Widrow-Hoff. Cette méthode se base sur le principe suivant : le poids de la connexion entre un neurone dont la sortie est égale à la sortie désirée et le neurone d'association garde sa valeur et ne peut pas être modifié. Dans le cas contraire, le poids peut être modifié selon l'entrée. [15]

2.2.2.6 Les réseaux Hopfield

C'est un réseau qui est constitué d'un mélange de neurones de McCulloch et de Pitts, et qui prennent deux états : -1 et 1 ou 0 et 1. Il utilise la règle Hebb (Hebb, 1949) comme loi d'apprentissage. Cette loi se base sur le principe suivant : une synapse améliore son activité si et seulement si l'activité de ces deux neurones est corrélée. [15]

2.2.2.7 Les réseaux de kohonen

Dans ce modèle, on cherche toujours un type de neurones plus proche de la réalité. Il se base principalement sur l'observation biologique du fonctionnement des systèmes nerveux des mammifères. [15]

Ce réseau utilise une loi de Hebb (Hebb, 1949) modifié, pour l'apprentissage qui tient compte de l'oubli. Dans le cas où les neurones reliés ont une activité simultanée, la connexion entre ces derniers est renforcée, et elle est diminuée dans le cas contraire. [15]

Dans ce type de neurones, une loi d'interaction latérale est aussi modélisée. En effet, les neurones les plus proches interagissent positivement, tandis que les neurones un peu plus éloignés communiquent négativement. Enfin, les neurones éloignés ne communiquent pas. Dans ce dernier cas, et pour une entrée quelconque, une sortie particulière sera activée et pas les autres. [15]

2.2.2.8 Les réseaux du Perceptron multicouche

Ils représentent une amélioration du Perceptron contenant une ou plusieurs couches intermédiaires. Pour modifier leurs poids, ces perceptrons utilisent un algorithme de rétro-propagation du gradient, qui est une généralisation de la règle de Widrow-Hoff. L'objectif principal est de toujours minimiser l'erreur quadratique, ce qui est très simple avec l'utilisation d'une fonction dérivable, comme la sigmoïde par exemple. La modification des poids est propagée dans ce cas, de la couche de sortie jusqu'à la couche d'entrée. [15]

Les perceptrons multicouches agissent comme un séparateur non linéaire et peuvent être utilisés dans plusieurs domaines, comme : le traitement d'images, la classification, etc. [15]

2.2.2.9 Avantages des réseaux de neurones

Les réseaux de neurones sont considérés parmi les méthodes de classification les plus efficaces. En effet, ce type d'algorithmes représente plusieurs avantages parmi lesquels on peut citer par exemple :

1. La rapidité et l'efficacité de traitement des grands corpus. [53]
2. La possibilité de combiner ce type d'algorithmes avec d'autres méthodes de classification. [57]
3. un taux d'erreur très faible par rapport aux autres méthodes de classification. [57]
4. Les réseaux de neurones ne nécessitent pas l'utilisation de modèles mathématiques très complexes pour leur fonctionnement. En effet, grâce à leur capacité d'apprentissage, ils se basent principalement sur les modèles de données à traiter. [9]

2.2.2.10 Inconvénients des réseaux de neurones

Malgré les grands avantages que représentent les réseaux de neurones, ils ont aussi des inconvénients qu'on peut résumer dans les points suivants :

1. La lenteur d'apprentissage. [57]
2. Les résultats obtenus par la classification des réseaux de neurones ne sont pas interprétables. En effet, le réseau généré par ce type d'algorithmes est considéré comme une boîte noire, c'est-à-dire que l'utilisateur n'a aucune information explicite sur le fonctionnement interne. En cas d'erreurs, il est impossible de déterminer la cause cette erreur. [53]

3. La convergence des résultats des réseaux de neurones est incertaine. [53]

4. Les réseaux de neurones ne permettent pas l'intégration des connaissances a priori pour le traitement de nouvelles données. [53]

2.2.3 *Les arbres de décision*

Les arbres de décision sont considérés parmi les méthodes les plus populaires pour la classification textuelle. Parmi les algorithmes les plus connus, on peut citer ID3 (Quinlan, 1986) et C4.5 (Quinlan, 1993). [5]

Le fonctionnement des arbres de décision se base principalement sur des exemples. En effet, si on veut classer des documents dans des catégories, on doit construire un arbre de décision par catégorie. [5]

D'une manière générale, chaque nœud de l'arbre de décision exécute un test If / Then et les feuilles de l'arbre ont les valeurs de décision Oui ou Non. Les tests exécutés, observent les valeurs des attributs de chaque exemple. Pour un texte quelconque par exemple, l'attribut peut être un mot avec une valeur de 0 ou 1 selon que ce mot appartient à ce texte ou non. [5]

2.2.3.1 La construction des arbres de décision

Pour chaque échantillon considéré, il y a plusieurs arbres de décision qui peuvent le représenter. En général, l'arbre ayant la taille la plus petite possible est choisie parmi l'ensemble. En effet, plus que la taille de l'arbre est petite, plus qu'il contient les meilleurs attributs. L'expression «meilleurs attributs» désigne dans ce cas, les attributs qui divisent bien les données, c'est-à-dire les attributs significatifs et pertinents. [16]

Pour construire un arbre avec une taille plus petite, on doit mesurer les données D par une quantité quelconque, comme par exemple le gain d'information. Le faite de diviser D en $D[1],..., D[k]$, diminue cette quantité. [16]

Quand la quantité d'un arbre est égale ou se rapproche de 0, on arrête la procédure et les données sont bien divisées. [16]

Le procédé qui permet de construire un arbre de décision est le suivant : [16]

Méthode Construire-Arbre-Décision (D : données, N : nœud, M : méthode)

1. On applique M sur D afin de trouver le critère de division
2. En utilisant le critère de division, on divise le nœud N en sous nœuds
3. On met le nombre des sous nœuds de N dans la variable k

Si k>0 alors

On fabrique k sous nœuds $N_1,...N_k$ de N

On divise D en $D_1,...D_k$

Pour i=1 jusqu'à k faire

Construire-Arbre-Décision (D_i, N_i, M)

Fin pour

Fin Si

Fin de la procédure

Figure 2.3 Procédure de construction d'un arbre de décision

Il existe plusieurs algorithmes qui permettent de construire un arbre de décision parmi lesquels on peut citer par exemple:

1. ID3
2. C4.5

2.2.3.1.1 Algorithme ID3

Développé par Quinlan en 1986, ID3 est un algorithme de classification supervisé qui se base principalement sur des exemples déjà classés pour générer des arbres de décision. [5]

En se basant sur le théorème de Shanon (Shanon, 1948), ID3 utilise l'entropie pour mesurer le désordre des données. Shanon a utilisé ce théorème afin de calculer la taille minimale pour coder un message. [16]

Cet algorithme permet de construire récursivement un arbre de décision. Il calcule, parmi les attributs qui restent, celui qui va générer, le plus d'informations, qui permettront de classer les exemples d'un niveau quelconque de l'arbre de décision. [17]

Soit 'A' l'attribut utilisé pour diviser les données. On peut décrire l'algorithme ID3 par [16] :

Méthode ID3 (N : Nœuds, D : Données)
Si les éléments de D appartiennent à la même classe alors
Sortir
Fin Si
Pour chaque attribut A faire
Calculer *InformationObtenue*(D, A)
$A_{\max} \leftarrow \arg\max_A$ *InformationObtenue*(D, A)
On divise N en utilisant $A_{\max}$
K $\leftarrow$ nombre de sous nœuds de N
Si k>0 alors
Construire k sous nœuds $N_1,...,N_k$ de N
On divise D en $D_1,...,D_k$
Pour i=1 jusqu'à k
ID3 (N_i , D_i , M : méthode)
Fin Pour
Fin Si
Fin Pour
Fin de la procédure

Figure 2.4 Algorithme ID3

2.2.3.1.2 Algorithme C4.5

L'algorithme C4.5 est un algorithme de classification supervisé. Proposé par Quinlan en 1993, il se base principalement sur ID3 en lui apportant certaines améliorations. ID3 représente plusieurs inconvénients parmi lesquels on peut citer par exemple [16]:

1. La difficulté de son application pour des vrais projets. Il sert seulement à montrer comment construire un arbre de décision.
2. ID3 permet de traiter seulement les attributs continus et pas ceux qui sont discrets.

3. Si on a des exemples avec attributs manquants, cela aura une influence sur les résultats obtenus. ID3 n'est pas capable de résoudre ce problème.
4. Le problème de sur satisfaction, c'est-à-dire, quand les données contiennent des erreurs au moment de la construction de l'arbre de décision. Dans ce cas, ID3 va satisfaire aussi les erreurs.

C4.5 apporte plusieurs améliorations qui permettront de résoudre les problèmes causés par ID3, comme par exemple : [18]:

1. Le traitement des attributs continus.
2. Le traitement des valeurs nulles pour un attribut.
3. C4.5 permet de grouper l'ensemble des valeurs discrètes nominales, pour un attribut quelconque, afin de supporter des essais plus complexes.

Pour utiliser C4.5, l'algorithme ID3 a subit certaines modifications, parmi lesquelles, on peut citer par exemple [19] :

1. La modification de la fonction qui permet de calculer le gain.
2. Après la construction de l'arbre de décision, l'algorithme C4.5 cherche à élaguer cet arbre avec une heuristique.

Soit P un ensemble d'apprentissage, l'algorithme C4.5 se déroule en deux phases [19]:

1. **La phase d'expansion :** Cette phase utilise la fonction entropie pour construire l'arbre de décision et cela, en divisant récursivement notre ensemble d'apprentissage P.

On peut diviser cette étape en trois sous phases :

I. Déterminer si un nœud est terminal ou non : On dit qu'un nœud b est terminal si $N(b) < n_0$ Où $i(b) < i_0$, et n_0, i_0 sont des paramètres qu'il faut fixer.

II. Association d'un test à un nœud : Dans cette étape, on calcule le gain par rapport à un test T et une position b.

Soit B_i la proportion d'éléments de l'ensemble des exemples associés à b qui vont sur le nœud en position b_i, et soit i la fonction entropie. On définit la fonction de gain par :

$$gain(b,T) = i(b) - \sum_{i=1}^{n} (B_i \times i(b_i)) \tag{2.9}$$

À une position b, on doit choisir le test qui va maximiser la fonction de gain.

III. Associé une classe à une feuille : Cette étape consiste à attribuer une classe majoritaire à une feuille.

2. **La phase d'élagage :** Dans cette phase, l'algorithme C4.5 se base sur l'ensemble d'apprentissage afin d'élaguer l'arbre obtenu à l'étape précédente. Dans ce cas, le critère d'élagage se base principalement sur une heuristique qui permet d'estimer l'erreur réelle sur un sous ensemble donné.

2.2.3.2 Avantages des arbres de décision

L'utilisation des arbres de décision représente plusieurs avantages parmi lesquels on peut citer par exemple :

1. Leur capacité à travailler sur des données symboliques. [58]
2. Leur grande capacité et efficacité à faire de la classification. [58]

3. Leur facilité d'apprentissage et d'utilisation. [59]

2.2.3.3 Inconvénients des arbres de décision

Malgré les avantages que représentent les arbres de décision, ils ont aussi des inconvénients qu'on peut résumer dans les points suivants :

1. Ce type d'algorithmes est très sensible aux points aberrants et au bruit. [58]
2. Leur sensibilité au changement des données. [59]
3. Une détection difficile des interactions entre les variables. [59]

2.2.4 *Les algorithmes génétiques*

Les algorithmes génétiques représentent un outil efficace et performant pour résoudre des problèmes d'optimisation et pour aider d'autres outils de prédiction à augmenter leurs performances. Les réseaux de neurones, par exemple, utilisent les algorithmes génétiques pour représenter tous les poids par un gène. Ainsi, après la construction de plusieurs gènes, l'algorithme permute les meilleurs de celle-ci pour obtenir un ensemble de poids optimal. [20]

Au début des années 1990, les algorithmes génétiques ont été utilisés pour résoudre plusieurs problèmes de classification.

Le fonctionnement de ces algorithmes se base principalement sur le comportement réel des espèces naturelles. En effet, la plupart de ces techniques essaient de développer un ensemble de solution (population) en leur appliquant certaines règles qui permettent de produire un comportement donné. Ce comportement est représenté dans la plupart des cas, sous forme d'une fonction appelée la fonction de fitness. [21]

2.2.4.1 Les concepts de base des algorithmes génétiques

Pour bien comprendre le fonctionnement des algorithmes génétiques on doit définir certains principes de base.

Dans la vraie vie, plusieurs opérations ont été mises en évidence pour permettre la reproduction des chromosomes. Ces techniques imitées par les algorithmes génétiques se basent principalement sur ces trois opérations [23]:

1. **Les sélections :** Cette étape consiste à choisir les individus qui permettent de donner les meilleurs résultats, ce qui est similaire à un processus naturel où les meilleures espèces arrivent à se produire, tandis que les autres meurent.
2. **Les croisements :** Cette étape se résume dans l'échange, de deux chromosomes, des parties de leurs chaînes pour produire de nouveaux chromosomes.

On peut distinguer deux types de croisements : simple et multiple.

I. Les croisements simples : Dans ce cas, le croisement et l'échange de l'ADN se fait en un seul point entre deux chromosomes. Ce type de croisement consiste à choisir dans un premier temps deux chromosomes avec une probabilité p, ensuite, les chaînes représentatives de ces deux chromosomes sont coupées, en une position aléatoire identique, pour donner naissance à deux segment queue et deux segments tête. La dernière phase consiste à permuter les deux segments queue pour obtenir deux enfants héritant des caractéristiques de leurs parents.

II. Les croisements multiples : À la différence du croisement multiple, ce type de croisement permet l'échange de l'ADN dans différents points entre deux

chromosomes. Dans ce cas, Plusieurs segments de gènes peuvent être interchangés entre les deux chromosomes.

La probabilité de croisement des algorithmes génétiques est toujours comprise entre 0 et 1.

3. **La mutation :** Cette phase permet la modification irréversible de l'information génétique dans la séquence d'un génome. Dans la plupart du temps, cette mutation est due à des erreurs de copie du matériel génétique au cours de la division cellulaire. Ces erreurs sont corrigées par des mécanismes complexes de réparation de génomes.

2.2.4.2 Le fonctionnement des algorithmes génétiques

La plupart des algorithmes génétiques se basent principalement sur la résolution de la fonction $f : Y \rightarrow \Re$, où Y est un espace de recherche et f la fonction de fitness. Cette fonction se comporte comme une boîte noire pour l'algorithme génétique. [24]

La première étape pour construire un algorithme génétique consiste à créer une population initiale. Ensuite, on doit coder chaque individu de la population par un chromosome (Holland, 1975). Dans ce cas, chaque chromosome code un point dans notre espace de recherche. Ce codage des chromosomes a une grande influence sur la qualité et la performance de l'algorithme génétique. Par exemple, l'algorithme génétique de Holland, se base sur la représentation de chaque chromosome sous forme de chaînes de bits contenant la description d'un point dans l'espace, pour faciliter les opérations de mutation et de croisements simples. [24]

Les algorithmes génétiques utilisent une fonction d'évaluation pour calculer le coût d'un point de l'espace de recherche. Le résultat de l'évaluation d'un individu ne se base pas sur celle d'un autre individu. En effet, la fonction d'évaluation permet d'accepter ou de refuser un individu, selon le coût obtenu par ce dernier.

Enfin, la mutation est utilisée pour éviter que la fonction de fitness de l'algorithme génétique converge vers des optima locaux. Pour cela, la mutation permet de générer des erreurs de recopie pour donner la vie à un nouvel individu. Si l'individu généré est faible, alors l'algorithme l'élimine sinon, il le garde. [24]

D'une manière générale, les algorithmes génétiques fonctionnent de la manière suivante : [24]

On doit choisir une population initiale p_0 de n individus.

1. À partir de la population initiale choisie, on sélectionne une population intermédiaire de n individus.
2. Les n individus se croisent deux à deux pour construire n nouveaux individus.
3. Les individus obtenus passent par un opérateur de mutation qui agit aléatoirement pour construire une nouvelle population.
4. On réitère le processus jusqu'à l'obtention d'une solution optimale.

Figure 2.5 Fonctionnement des algorithmes génétiques

2.2.4.3 Avantages des algorithmes génétiques

Les algorithmes génétiques représentent plusieurs avantages, parmi lesquels on peut citer :

1. Une visualisation graphique simple et claire des résultats. [20]
2. Leur capacité à faire plusieurs calculs en parallèle. [20]

3. L'optimisation des fonctions sans connaître le mécanisme de leur fonctionnement. [61]
4. La possibilité de combiner un algorithme génétique avec un autre algorithme pour résoudre un problème de classification complexe. [20]
5. La flexibilité de ces algorithmes, car ils permettent de s'adapter avec n'importe quel problème. [62]
6. À la différence des autres méthodes de classification, les algorithmes génétiques convergent vers le minimum global de la fonction objectif. [63]
7. Leur capacité à supporter de plus gros volumes de données. [20]

2.2.4.4 Inconvénients des algorithmes génétiques

Malgré les grands avantages que représentent les algorithmes génétiques, ils ont aussi des inconvénients, comme par exemple [60] :

1. Le temps énorme consacré au calcul et à la résolution des problèmes.
2. La difficulté de programmer ces algorithmes.
3. La difficulté de trouver une solution efficace ou exacte. Dans la plupart des cas, les algorithmes génétiques trouvent seulement des solutions rapprochées aux problèmes.

2.2.5 *L'algorithme de Naïve Bayes*

Les méthodes Naïve Bayes sont considérées parmi les modèles probabilistes les plus connus. Elles se basent principalement sur le théorème de Bayes (Bayes, 1963). [4]

Les algorithmes Naïves Bayes sont souvent utilisés dans la catégorisation et la classification de documents. Ils permettent d'estimer la probabilité de chaque classe parmi les exemples, étant donné un document, et affectent à ce dernier la classe la plus probable. On appel ce procédé «Prior probabilities». [5]

Si on considère par exemple la classe "UQTR" qu'on retrouve 3 fois dans 7 documents, alors le «Prior probabilities» est de 3/7. [5]

Pour classer un ensemble de documents, Naïve Bayes utilise comme entrée les mots qui se trouvent dans ces derniers, ensuite il calcule la fréquence de chaque mot dans les différents documents classés dans une classe donnée [5,55]. La procédure complète qui permet de calculer la fréquence d'un mot dans un ensemble de documents est la suivante [5,55]:

Soit D un ensemble de documents avec leurs valeurs prévues V, $P(w_k(v_j))$ la probabilité d'un mot w_k tiré au hasard à partir d'un document quelconque de la classe v_j et $P(v_j)$ la «Prior probabilities» de la classe v_j.

1. Vocabulaire ← La collection de tous les mots, ponctuations et d'autres caractères apparaissant dans l'ensemble des documents.
2. Calculer les termes $P(v_j)$ et $P(w_k \mid v_j)$.

Pour chaque valeur v_j de V faire

$docs_j \leftarrow$ le sous-ensemble de documents de D pour lesquels la valeur prévue est v_j.

$$P(v_j) = \frac{|doc_j|}{|D|}$$

$Text_j \leftarrow$ un document simple crée en concaténant tous les membres de doc_j.

$n \leftarrow$ le nombre total de positions des mots distincts dans $Text_j$.

Pour chaque mot w_k dans le vocabulaire

$n_k \rightarrow$ Le nombre de fois que le mot w_k apparaît dans $Text_j$

$$P(w_k \mid v_j) \leftarrow \frac{n_k + 1}{n + |vocabulaire|}$$

Classifiez_Texte_Naïve_Bayes(Doc)

Retourner la valeur prévue estimée du document Doc. a_i Dénote le mot trouvé à la ième position dans Doc.

Positions ← Toutes les positions des mots dans Doc qui contiennent des «token» appartenant au vocabulaire.

Retourner v_{NB} où

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i \in Positions} P(a_i \mid v_j)$$

Figure 2.6 Calcul de la fréquence d'un mot dans un ensemble de documents avec Naïve Bayes

2.2.5.1 Avantages des méthodes Naïve Bayes

Parmi les avantages des méthodes Naïve Bayes on peut citer par exemple [64] :

1. La facilité et la simplicité de leur implémentation.
2. Leur rapidité.
3. Les méthodes Naïve Bayes donnent de bons résultats.

2.2.5.2 Inconvénients des méthodes Naïve Bayes

À cause de l'hypothèse d'indépendance des mots dans ce modèle, on le qualifié souvent de naïf ou de simple. En général, ce type d'algorithmes permet de faire le même travail de classification que les autres algorithmes qui existent déjà, mais ces performances sont limitées quand il s'agit d'une grande quantité de lexiques à traiter. En effet, si le nombre de lexiques augmente, alors le nombre des dépendances entre l'ensemble des mots augmentent, et donc, la vérification de l'hypothèse de Naïve Bayes diminue. [5,55]

2.2.6 *Les machines à support de vecteurs (SVM)*

Les machines à support de vecteurs sont parmi les techniques les plus connues. Développé par Vapnik en 1995, elles sont considérées comme une alternative récente pour la classification. Elles se basent principalement sur l'utilisation des fonctions appelées kernel, qui facilitent la séparation des données. [25]

En général, les SVM peuvent être utilisées pour résoudre plusieurs problèmes réels, tels que, la classification textuelle et la régression. Pour cela, on doit construire une fonction f qui accepte un vecteur d'entrée x et qui retourne un vecteur de sortie y, avec :

$$y = f(x) \tag{2.10}$$

2.2.6.1 Le principe des SVM

Soit x et y deux classes d'exemples de données, le but des SVM est de trouver un classificateur capable de séparer les données et de maximiser la distance entre les deux classes. On appelle ce genre de classificateur : un classificateur linéaire hyperplan. [25]

Le schéma suivant montre un hyperplan qui sépare deux classes de points [26]:

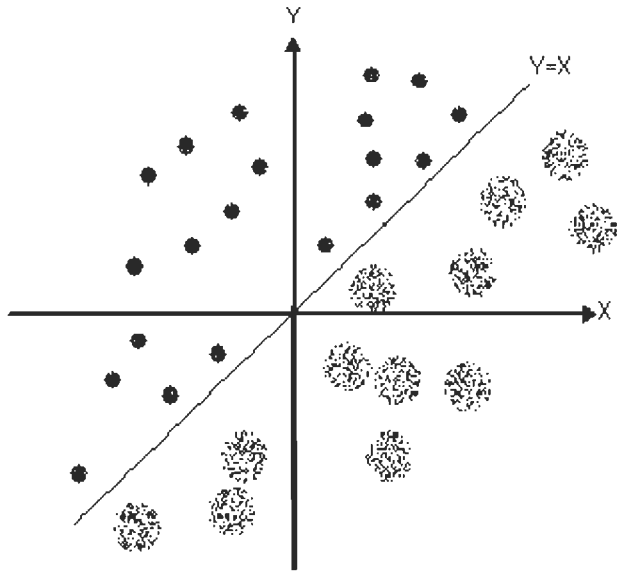

Figure 2.7 Hyperplan qui sépare deux classes de points

Malgré la grande quantité des hyperplans qu'on peut obtenir, les SVM choisissent seulement celui qui est optimal, c'est-à-dire celui qui passe au milieu des points des deux classes. Cela revient à chercher un hyperplan dont la distance (la marge entre l'hyperplan et les exemples d'apprentissage) est minimale par rapport aux données d'apprentissage. En effet, pour obtenir un hyperplan optimal, il faut maximiser la marge entre les données et l'hyperplan. [25]

Le schéma suivant montre un hyperplan (celui du milieu) optimal avec une marge maximale. [26]

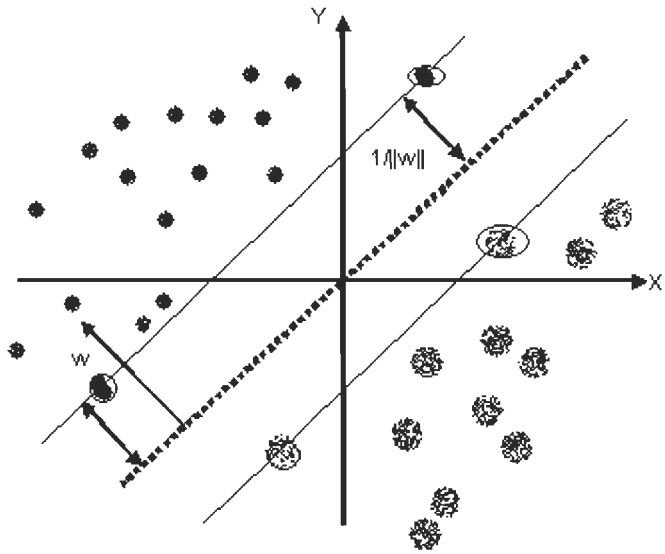

Figure 2.8 Hyperplan optimal avec une marge maximale

On peut distinguer deux types de SVM, les modèles linéairement séparables et celles non linéairement séparable. La première catégorie des SVM est très simple et permet de trouver rapidement le classificateur linéaire, mais pour des problèmes réels, il n'existe pas en général de séparatrice linéaire. [25]

Les deux schémas suivants montrent la différence entre les modèles linéairement séparables et celles non linéairement séparables [26] :

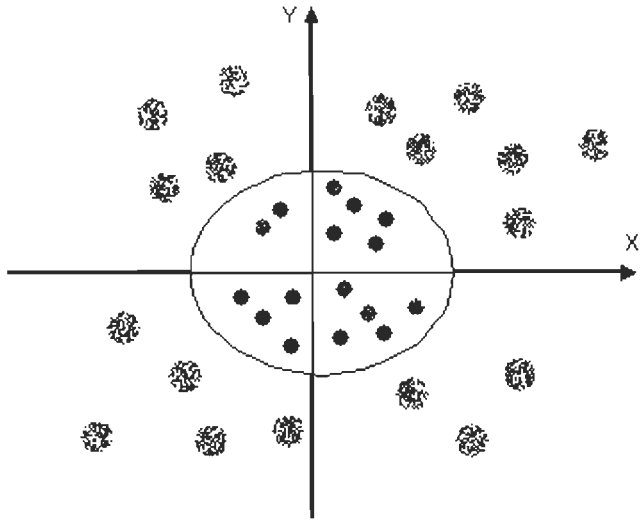

Figure 2.9 Problème de discrimination à deux classes avec une séparatrice non linéaire

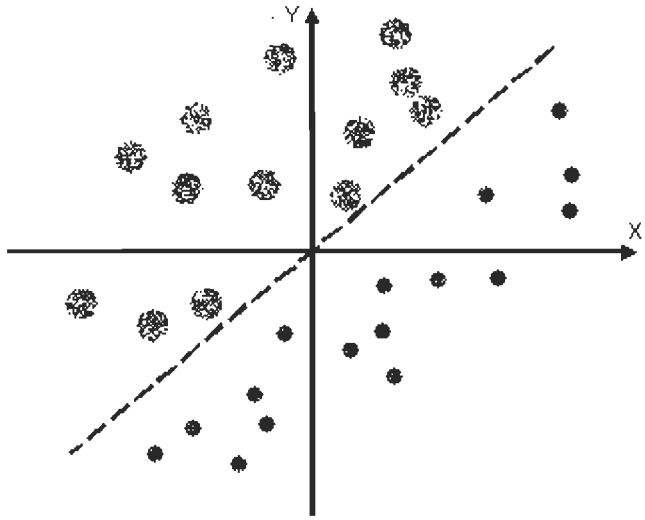

Figure 2.10 Problème de discrimination à deux classes avec une séparatrice linéaire

Pour résoudre le problème du non linéarité séparatrice, l'idée des SVM consiste à changer l'espace de données. Dans ce cas, la transformation non linéaire des données dans un autre espace permettra d'avoir une séparation linéaire de ces derniers. En effet, plus la dimension du nouvel espace est grande, plus la chance de trouver un hyperplan séparateur entre les données est grande aussi. [25]

Cette transformation de données d'un espace à un autre se fonde principalement sur l'utilisation de la fonction Kernel (noyau). Il existe plusieurs types de fonctions Kernel comme la fonction Gaussien, polynomiale et sigmoïde. Il revient donc à l'utilisateur des SVM de choisir celle qui est la plus convenable pour résoudre son problème. [25]

2.2.6.2 Avantages des SVM

Les SVM représentent plusieurs avantages parmi lesquels on peut citer par exemple [26]:

1. Leur capacité à manipuler de grandes quantités de données.
2. Le faible nombre d'hyper paramètres utilisés par ces méthodes.
3. Elles sont bien fondées théoriquement.
4. Les résultats pertinents qu'on obtient avec les SVM en pratique.

2.2.6.3 Inconvénients des SVM

Malgré leurs performances, les SVM représentent aussi des faiblesses, parmi lesquelles on peut citer :

1. Leur utilisation des fonctions mathématiques complexes pour la classification des corpus. [25]
2. Pour trouver les meilleurs paramètres, ce type d'algorithmes demande un temps énorme pendant les phases de test. [25]

2.3 Les méthodes de classification non supervisées

Les méthodes de classification non supervisées se basent principalement sur la séparation automatique des nuages de points dans un espace, sans le besoin de fournir des données d'apprentissage. Dans ce genre de méthodes, le nombre de classes est fixé au préalable par l'utilisateur. [27]

En général, le fonctionnement des algorithmes de classification non supervisée consiste à partitionner un ensemble d'objets en k sous-ensembles, où k représente le nombre de regroupements attendus par l'utilisateur. Il existe plusieurs stratégies qui permettent de trouver ces regroupements, comme les méthodes se basant sur les densités, sur le partitionnement, de même que des méthodes hiérarchiques et celles utilisant la quantification. [28]

Parmi les algorithmes de classifications non supervisées les plus connues, on peut citer par exemple [6]:

1. K-moyen
2. Single-pass
3. Suffix tree clustering
4. Hierarchical Agglomerative Clustering
5. Les cartes auto organisatrices de Kohonen
6. ART

2.3.1 *K-moyen*

Introduite par J. McQueen en 1971 et amélioré sous sa forme actuelle par E. Forgy, la méthode du k-moyen est considérée comme un outil de classification efficace qui permet de diviser un ensemble de données en k classes homogènes. En effet, cette méthode initialise k clusters avec k vecteurs qui servent comme centres de gravité pour le reste des vecteurs à classifier. Chaque vecteur est ajouté dans ce cas, au cluster dont le centre est le plus proche. Les k clusters sont produits de façon à minimiser la fonction objective suivante [31,32]:

$$E = \sum_{r=1}^{k} \sum_{X_i \in C_r} (X_i - g_r)^2 \qquad (2.11)$$

Où :

C_r : représente l'ensemble des classes.

X_i : Un point qui appartient à une classe C_r.

g_r : Le point moyen de la classe C_r.

Dans le domaine de la classification non supervisée, cet algorithme cherche à partitionner l'espace des données en classes isolées les unes des autres, et cela, en minimisant la variance entre ces derniers.

L'exécution de la méthode du K-moyen se déroule en trois étapes [30]:

1. Initialisation de tous les groupes.
2. Faire une première allocation des entités aux groupes de l'étape 1.

3. En cas de besoin, Réallocation des entités aux groupes pour minimiser un critère quelconque.

Il est important de combiner la méthode K-moyen avec un autre algorithme pour fournir une estimation des m groupes à obtenir. Ensuite, pour améliorer le regroupement, K-moyen utilise des conditions de transfert pour reclasser les entités. [30]

Parmi ces conditions, on peut citer [30] :

1. La vérification si un transfert est possible.
2. Arrêt de l'algorithme si aucun transfert a eu lieu ou si on a atteint le nombre maximum d'itérations.

On peut résumer le fonctionnement de l'algorithme K-moyen dans les étapes suivantes [32] :

1. On choisit k objets au hasard qu'on considère comme des centres pour les classes initiales.
2. On affecte chaque objet au centre le plus proche pour obtenir une partition de k classes.
3. On recalcule les centres de chaque classe.
4. La répétition des étapes 2 et 3 jusqu'à la stabilité des centres.

Figure 2.11 L'algorithme du K-moyen

La complexité de l'algorithme du K-moyen est de $O(lkn)$, où l est le nombre d'itérations, k le nombre des classes, et $k < n$. [32]

On retrouve différents types de K-moyen qui se distinguent, par la sélection initiale des k objets de bases, par la méthode de calcul des moyennes et des similarités entre les objets

de l'espace. Parmi ces méthodes on peut citer, par exemple, la méthode des centres mobiles et celle des nuées dynamiques. [32]

2.3.1.1 Avantages du K-moyen

La méthode du K-moyen représente plusieurs avantages, comme par exemple :

1. Sa complexité linière.
2. Sa facilité.
3. Sa convergence rapide.
4. Son adaptation à de larges bases de données.
5. L'ordre d'entrée des objets n'a aucune influence sur les résultats de cette méthode.

2.3.1.2 Inconvénients du K-moyen

Malgré les grands avantages de cette méthode, elle a aussi des inconvénients qu'on peut résumer dans les points suivants :

1. Le nombre d'objets k est fixé au début, ce qui influence les résultats.
2. Sa sensibilité aux éléments marginaux.
3. Sa mauvaise gestion pour les clusters mal isolés.

2.3.2 *Single-Pass*

Single-Pass est une méthode de classification qui traite séquentiellement les documents. Elle se base principalement sur le principe suivant [6]:

Elle commence par arranger un premier document dans un cluster. Ensuite elle regarde si ce document est similaire à un deuxième, c'est-à-dire si la similarité entre les deux documents est supérieure ou égale à un seuil fixé par l'utilisateur. Si oui, on ajoute ce document au cluster. On continue avec le même principe, tant que la similarité entre le nouveau document et le cluster satisfait le seuil défini par l'utilisateur.

Si nous ne dépassons pas le seuil, alors on aura un ensemble de documents non traités et un cluster. On réitère le processus en ajoutant au cluster un nouveau document non traité, ensuite nous calculons le seuil de similarité des documents qui restent.

L'algorithme de Single-Pass se résumera dans les étapes suivantes [6]:

Choisir un premier document

Faire

1. Calculer la distance qui sépare un document, non encore affecté, à chaque classe.
2. Si cette distance est inférieure au seuil défini par l'utilisateur, affecter le document à la classe et recalculer les représentants de cette dernière. Dans le cas contraire, créer une nouvelle classe utilisant ce document comme représentant.

Tant qu'il y a des documents à traiter

Figure 2.12 Algorithme Single-Pass

2.3.2.1 Avantages de Single-Pass

Parmi les avantages de la méthode Single-Pass, on cite [65]:

1. Sa simplicité d'utilisation.
2. Sa rapidité à classifier les documents.
3. Ce type d'algorithmes permet le traitement d'un grand nombre de données.

2.3.2.2 Inconvénients de Single-Pass

L'inconvénient majeur de cette méthode est que les résultats de la classification dépendent toujours de l'ordre de traitement séquentiel des documents. En effet, à cause de la quantité énorme des classes crées, les résultats obtenus, dans la plupart du temps, sont mauvais (Zamir & Etzioni, 1998). [6]

2.3.3 *Suffix Tree Clustering*

Suffix Tree Clustering (Zamir, 1997) est une méthode de classification linière. Pour construire les clusters, cet algorithme se base principalement sur l'identification des mots ou des expressions communes à plusieurs documents. [6, 34,35]

Contrairement aux autres méthodes ordinaires, la classification par arbre de suffixes ne cherche pas à classifier chaque document dans un groupe précis, mais le même document peut se retrouver dans plusieurs classes. [6]

On peut résumer le déroulement de la classification par arbre de suffixes dans les six étapes suivantes [6]:

1. La première étape consiste à nettoyer les documents à traiter.
2. Faire une lemmatisation rapide des documents.
3. Identification des phrases de chaque document.
4. En utilisant les index inversés, on associé chaque phrase avec l'ensemble des documents où elle appartient.

5. Pondération de chaque phrase; en effet, le résultat ou le score d'une phrase dépend essentiellement du nombre de mots qu'elle contient, ainsi que le nombre de documents dans lesquels elle apparaît.

6. Fusionner les clusters similaires. Le critère de ressemblance se base sur une fonction permettant de calculer la similarité entre deux clusters. En effet, cette fonction permet de compter le nombre de documents que deux clusters partagent.

2.3.3.1 Avantages de la classification par arbre de suffixes

L'algorithme Suffix Tree Clustering représente plusieurs avantages, parmi lesquels on peut citer [6]:

1. L'ordre de présentation des documents n'a aucune influence sur le résultat attendu.
2. Le nombre de classes n'est pas défini au début.
3. La possibilité d'ajouter un nouveau document à ceux déjà traités.

2.3.3.2 Inconvénients de la classification par arbre de suffixes

Le seul problème avec la classification par arbre de suffixes est qu'il faut définir manuellement le coefficient de similarité pour fusionner deux classes. [6]

2.3.4 *Hierarchical Agglomerative Clustering(HAC)*

Cette méthode commence par assigner à chaque cluster un objet. Ensuite, pour construire un nouveau cluster, cet algorithme fusionne à plusieurs reprises les deux clusters les plus proches. Le processus est répété plusieurs fois, tant que les critères d'arrêts ne sont pas satisfaits (Manning et Schutze, 1999) et qu'il reste au moins deux clusters à traiter. [6]

Le fonctionnement de cet algorithme peut être résumé dans les étapes suivantes [6] :

1. Assigner chaque objet à un cluster.
2. Construire une matrice de distances entre les clusters.
3. Rechercher les deux clusters avec la distance minimale.
4. Supprimer ces deux clusters de la matrice des distances et les fusionner dans un nouveau cluster.
5. Évaluer toutes les distances de ce nouveau cluster par rapport aux autres clusters, et faire la mise à jour de la matrice des distances.
6. On réitère le processus jusqu'à l'obtention d'une matrice de distances avec un seul élément.

On peut distinguer plusieurs versions de la méthode HAC, comme par exemple : single linkage, group average linkage et complete linkage. [6]

2.3.4.1 Avantages des méthodes HAC

Le principal avantage des HAC réside dans leur capacité à générer des partitions emboîtées. En effet, ces méthodes proposent un ensemble de partitions (des solutions) représentées sous forme d'arbre, où l'utilisateur aura la possibilité de choisir la solution qui convient le mieux à ses besoins. [66]

2.3.4.2 Inconvénients des méthodes HAC

Les méthodes HAC représentent certaines faiblesses, comme par exemple [6]:

1. Le temps énorme consacré à la classification des documents.

2. L'utilisation de plusieurs types de Métriques, pour mesurer la distance entre deux clusters, peut générer des résultats différents.

2.3.5 *Les cartes auto organisatrices de Kohonen (SOM)*

Développée par T. Kohonen (Kohonen, 1984), la méthode SOM représente un cas particulier des réseaux de neurones. Elle constitue un outil efficace et performant qui permet de classifier des échantillons par rapport à leurs similarités. [36]

Cette méthode consiste à représenter dans un espace de 1, 2 ou 3 dimensions des prototypes, décrivant un petit nombre de points abstraits de l'espace d'observation. Chaque prototype est associé dans ce cas, à un sous-ensemble de données d'origine. [36]

Les SOM sont considérés comme des méthodes d'apprentissage compétitif, c'est-à-dire que le prototype gagnant est celui le plus proche de l'objet à traiter. À la différence de la méthode K-moyen par exemple, les cartes auto organisatrices permettent de préserver la topologie. En effet, les objets les plus semblables seront plus proches sur la carte, ce qui donne une meilleure visualisation des classes. [38]

Techniquement, la projection des données sur la carte de Kohonen se fait en fixant une topologie, c'est-à-dire, un nombre de cellules et leurs voisinages. Dans ce cas, l'algorithme de Kohonen associé à chaque case (i, j) dans la carte, un vecteur $C_{i,j}$ dans l'espace de données. [37]

On peut résumer le déroulement de l'algorithme de Kohonen dans les étapes suivantes [37,38]:

1. Initialisation de l'ensemble des prototypes (neurones) selon la topologie choisie.
2. La boucle de base.
   I. Lecture d'un vecteur d'entrée x.
   II. Repérage du neurone k le plus proche de x. Ceci se fait par le calcul de la distance euclidienne D_k entre la donnée et tous les vecteurs neurones w_k .Le neurone le plus proche de x est choisi dans ce cas comme gagnant.

   $k = \arg Min(D_k = \|x - w_k\|)$ (On choisi le k qui donne la plus petite distance)

   III. Tirer vers x les prototypes suivants : Le gagnant k et les autres prototypes du voisinage, centré sur k.
3. Le déplacement dans ce réseau de neurones se base sur deux paramètres :
   I. La fonction du voisinage : cette fonction utilise deux paramètres, la forme et le voisinage. Si la largeur est grande (un grand voisinage), alors il existe beaucoup de neurones qui se rapprochent au même temps du donnée actuelle, sinon, on doit adapter le voisinage, en commençant par un grand voisinage et en le réduisant progressivement par la suite.
   II. Le taux d'apprentissage : On utilise ce paramètre pour contrôler la vitesse d'apprentissage et le couple stabilité/plasticité, c'est-à-dire que, si le taux d'apprentissage est trop petit, on aura un apprentissage très lent. Dans le cas contraire, l'apprentissage sera très rapide et risque d'être instable et de ne jamais converger. Pour remédier à ce problème, on choisit au début, une très grande h, qui sera réduite par la suite. L'apprentissage va s'arrêter quand h prend la valeur 0.

Figure 2.13 L'algorithme des cartes auto organisatrices de Kohonen

2.3.5.1 Avantages des SOM

Les SOM représentent plusieurs avantages, parmi lesquels on citera [67, 68]:

1. La possibilité d'une visualisation graphique des résultats.
2. La représentation des données en plusieurs dimensions.
3. La facilité à comprendre ce modèle.

4. Leur capacité à classifier les corpus d'une façon non supervisée, et cela, en utilisant la fonction de densité de l'échantillon.

2.3.5.2 Inconvénients des SOM

Malgré les grands avantages que représentent les SOM, ils ont aussi des limites qu'on peut résumer dans les points suivants [67]:

1. Leur temps énorme de convergence.
2. La représentation des données dans ce type de méthodes n'est pas unique.
3. Il n'est pas sûr que ce type d'algorithmes converge dans un espace multidimensionnel.

2.3.6 *Réseaux de neurones ART, ART1et Fuzzy ART*

2.3.6.1 Présentation générale

Conçu par (Grossberg, 1987, Grossberg et Carpenter, 1987 et 1988), le réseau ART est un modèle de classification autonome, dynamique et incrémental à apprentissage par compétition. Il se base principalement sur l'auto organisation des données. En effet, ART est capable de générer de nouvelles classes d'une manière dynamique, sans recours a priori à des informations sur leur nombre. [1]

En général, dans un réseau à apprentissage par compétition, il n'est pas sûr que les catégories trouvées auparavant, restent stables pendant le processus de classification des étapes suivantes. Pour assurer cette stabilité, il faut que le coefficient d'apprentissage soit proche de zéro. Si ce coefficient s'approche de zéro, on rencontrera un autre problème qui est la perte de la plasticité. On peut définir ce problème par l'apprentissage d'un système à

un ensemble d'informations non pertinentes ou d'ignorer d'autres déjà apprises. En effet, le système reprend dès le début l'ensemble des calculs sans prendre en considération les résultats déjà trouvés. [1]

Pour résoudre ce dilemme de la stabilité-plasticité d'un système, le réseau ART est capable de normaliser les intrants en imposant le même seuil de transfert à tous les neurones. Par contre, ce type de réseaux ne permet pas la modification des vecteurs de poids que si, les inputs proposés sont plus proches des prototypes existants. Si l'input est loin des prototypes existants, alors une nouvelle classe est crée et l'entrée en question est utilisée comme une base pour la nouvelle classe crée. [1]

Pour mesurer la similarité entre le prototype de base et les inputs à classifier, l'algorithme ART utilise un paramètre de vigilance compris entre 0 et 1. Dans le domaine de la classification textuelle par exemple, si ce paramètre est grand, alors on a plus de classes contenants moins d'éléments, tandis que si la valeur de ce paramètre est petite, on obtiendra un nombre faible de classes avec un grand nombre de segments. [50]

2.3.6.2 ART1

ART1 représente l'une des dérivées de l'algorithme ART. Il appartient à la famille des algorithmes non supervisés. Ce modèle a la particularité de travailler sur des données binaires, ce qui lui rend plus efficace pour la classification des données textuelles. [50]

ART1 est un réseau de neurone à doubles couches. En effet, la première couche sert comme entrée et sortie pour les données, tandis que la deuxième, appelée couche cachée, est une couche d'activation compétitive avec des connexions à poids fixe entre les neurones. Dans cette architecture, chaque neurone de la couche d'entrée est relié à tous les

neurones de la couche cachée, tandis que les neurones de la couche cachée sont associés à tous les neurones de la couche de sortie. [2]

Dans le domaine de la classification textuelle par exemple, ART1 prend en entrée un ensemble de vecteurs (segments) qui caractérisent un corpus, ensuite, il applique une série d'opérations sur ces derniers (comme la comparaison des mots et les n-grams), pour aboutir à la fin de ce processus, à un ensemble de classes de vecteurs. [52]

2.3.6.3 Algorithme ART1

On peut résumer le déroulement de l'algorithme ART1 dans les étapes suivantes [52]:

1. Cette première étape permet d'initialiser le système ART1 en choisissant un paramètre de vigilance capable d'assurer la stabilité de la classification et qui permettra aussi de définir le nombre de clusters à construire. Ensuite, le système génère la liste des vecteurs prototypes qui seront utilisés comme une base pour la classification des autres entrées.
2. Cette étape consiste à introduire un nouvel input à classer.
3. On identifie le prototype le plus proche au nouvel input.
4. On calcule la distance entre la nouvelle entrée et le prototype.
5. Si le vecteur input est proche du prototype, alors on ajoute l'input à la classe de ce dernier, sinon, l'algorithme ajuste et mis à jour les prototypes en reprenant le calcul à partir de l'étape 3.
6. On répète le processus à partir de l'étape 2 tant qu'il y a de nouvelles entrées à traiter.

2.3.6.4 Les avantages et les limites de ART1

2.3.6.4.1 Les avantages

L'algorithme ART1 représente plusieurs avantages parmi lesquels, on peut citer par exemple [50,52]:

- Sa capacité à traiter les inputs d'une manière dynamique. En effet, ART1 est capable de construire les classes par étape et de s'adapter aux changements que connaît le corpus pendant le processus de classification.
- Le nombre de clusters n'est pas défini par l'utilisateur; il est construit dynamiquement pendant la phase de classification.
- ART1 contrôle la qualité des inputs pour aboutir à une meilleure classification des données.
- Il utilise un paramètre de vigilance capable de contrôler la stabilité de la classification.
- Cet algorithme permet de résoudre le problème de la plasticité. En effet, il est capable d'apprendre de nouvelles informations, sans ignorer les résultats déjà trouvés.

2.3.6.4.2 Les limites de ART1

Malgré les grands avantages que représente le réseau ART1, il a aussi des faiblesses, qu'on peut résumer dans les points suivants:

- Dans le domaine de la classification textuelle, ART1 représente l'inconvénient de ne pas pouvoir faire appartenir un même segment à plusieurs classes. [50]

- La valeur du paramètre de vigilance dépend essentiellement du choix de l'utilisateur. [50]
- L'influence de l'ordre dans lequel se fait l'apprentissage sur les classes produites par ART1. [53]
- Sa faiblesse à traiter des vecteurs intrants, constitués de valeurs pondérées : il n'accepte que des valeurs binaires. [53]

2.3.6.5 Fuzzy ART

Proposé par Carpenter et Grossberg, Fuzzy ART permet de catégoriser des entrées binaires et analogiques. Cet algorithme propose des calculs simplifiés pour former des classes sous forme d'hyper-boîte, contrairement à des classes circulaires comme on trouve dans la plupart des algorithmes de réseaux de neurones. Il se base principalement sur deux critères de distance, l'activation et le choix. Cet algorithme semble très efficace dans les traitements parallèles et offre de bons résultats de catégorisation avec une précision modérée sur les poids des neurones. [69]

Le Fuzzy ART est un réseau de neurone à doubles couches complètement inter-reliées. La première couche est utilisée pour la comparaison, tandis que la deuxième est une couche compétitive qui ressemble à celle de kohonen. Ces deux couches sont activées par une entrée E. Cet algorithme propose une catégorisation particulière avec des classes sous forme d'hyper-rectangle et un codage en complément des entrées. Dans ce type de réseaux, Chaque hyper-rectangle est représenté par un prototype. [69]

Fuzzy ART est contrôlé par trois paramètres dont deux peuvent être modifiés pendant le processus de classification [69]:

- Le paramètre de choix α , utilisé pendant la phase de compétition entre les neurones.
- Le paramètre ρ de vigilance qui définit la taille de l'hyper-rectangle.
- Le paramètre d'apprentissage β qui influence la vitesse d'apprentissage des neurones.

Chapitre 3 - Les règles d'association

3.1 Introduction

Introduites par Al. et Agwal en 1993, les règles d'association représentent un outil efficace et performant qui a fait ses débuts dans le domaine de l'analyse du panier de la ménagère. En effet, les règles d'association ont permis la découverte non supervisée des tendances implicatives entre les données dans des bases de données transactionnelles. Plus précisément, dans une base de données qui contient un ensemble d'articles. La relation $X \Rightarrow Y$ indique que les transactions qui contiennent les articles de l'ensemble X ont tendance à inclure les articles de l'ensemble Y. [39,40]

Il existe plusieurs algorithmes de recherche de règles d'association (comme l'algorithme APRIORI), qui permettent de trouver et d'extraire des informations intéressantes stockées dans de grandes bases de données. [39,40]

Les règles d'association sont utilisées dans plusieurs applications. Dans le domaine du commerce électronique par exemple, on utilise ce type de méthodes pour choisir les produits à mettre en promotion, pour l'organisation des catalogues des articles, ainsi que pour construire des catalogues personnalisés pour chaque client. Dans le domaine médical, les règles d'association sont utilisées pour déterminer le traitement efficace qui correspond à un ensemble de symptômes. [39,40]

L'extraction des règles d'association se base principalement sur deux mesures, le support et la confiance. En effet, la plupart des algorithmes utilisés dans ce domaine, parcourent les données pour trouver les éléments qui dépassent un support minimum défini par l'utilisateur, et extraire par la suite, les règles d'association dont la confiance dépasse une confiance minimum. [42]

3.2 Définitions

Cette section est consacrée à la définition de plusieurs termes utilisés dans la recherche et l'extraction des règles d'association [43, 39]:

1. **Transaction** : On considère un ensemble $E = \{E_1, ...E_n\}$ de n éléments (items) distincts. On appelle une transaction T le sous ensemble E' inclus dans E.

   Dans une base de données D, chaque transaction est identifiée par une clé unique.

2. **Une règle d'association**: Soit A et B deux sous-ensembles d'éléments qui appartiennent à un ensemble M. On appelle $A \Rightarrow B$ la règle d'association entre A et B. L'évaluation de cette relation se base principalement sur le calcul de deux paramètres qu'on appelle le support et la confiance.

3. **Itemset** : Un Itemset désigne un ensemble d'objets ou d'articles.

4. **K-Itemset** : Un K-Itemset est un Itemset de k Items.

5. **Support d'un Itemset :** Soit A un Itemset de n éléments. Dans une base de données transactionnelle D, le support de A est le nombre de transactions dans D incluant A divisé par le nombre total des transactions de D :

$$Support(A) = \frac{|\{t \in D / A \subseteq t\}|}{|D|} \tag{3.1}$$

6. **Support d'une règle d'association :** Dans une base de données D, le support d'une règle d'association $A \Rightarrow B$ est le nombre de transactions qui contiennent A et B divisé par le nombre total des transactions :

$$Support(A \Rightarrow B) = \frac{|\{t \in D/(A \cup B) \subseteq t\}|}{|D|} \tag{3.2}$$

7. **Confiance** : La confiance d'une règle d'association $A \Rightarrow B$ est le rapport entre le nombre de transactions de D contenant $A \cup B$, et le nombre de transactions de D contenant A .C'est-à-dire que :

$$Confiance(A \Rightarrow B) = \frac{|\{t \in D/(A \cup B) \subseteq t\}|}{|\{t \in D / A \subseteq t\}|} \tag{3.3}$$

Cette formule est équivalente à :

$$Confiance(A \Rightarrow B) = \frac{Support(A \cup B)}{Support(A)} \tag{3.4}$$

8. **Itemset Fréquent :** Un Itemset A est fréquent si et seulement si son support est supérieur à un support minimum défini par l'utilisateur.

3.3 Les étapes d'extraction des règles d'association

Le processus d'extraction des règles d'association se déroule en trois étapes [39] :

1. **Préparation des données** : Cette étape consiste à réduire la quantité des données en gardant seulement celles les plus pertinentes, et en transformant par la suite, ces derniers en un contexte d'extraction, c'est-à-dire une transformation en un triplet

constitué : d'un ensemble d'objets, d'un ensemble d'Itemsets ainsi qu'une relation binaire entre les deux. Cette transformation des données en données binaires permettra d'améliorer la qualité des règles d'association.

2. **Recherche des ItemSets fréquents** : Un Itemset fréquent est un ensemble d'éléments dont le support est supérieur ou égal à un certain support minimal spécifié par l'utilisateur. Cette étape est très coûteuse en temps d'exécution. Pour un ensemble de n items par exemple, le nombre d'Itemsets fréquents qui peut être générés est de 2^n .

3. **Production des règles d'association** : La génération des règles d'association consiste à déterminer les règles d'association dont le support et la confiance sont supérieurs ou égaux à un certain support et confiance minimaux définis par l'utilisateur.

3.4 L'algorithme Apriori

L'algorithme Apriori (Agrawal et Srikant, 1994) représente la base de tous les algorithmes de recherche des règles d'association. Il utilise une stratégie de recherche des Itemsets fréquents en commençant par les Itemsets les plus généraux vers les plus spécifiques. [43, 39,45]

Cet algorithme se base principalement sur les deux règles suivantes [43, 39, 45]:

Soit S un Itemset et $S' \subseteq S$ alors :

1. Si S est non fréquent, alors les Itemsets qu'on construit de S sont aussi non fréquents.

2. Si S est fréquent, alors S' est aussi fréquent.

Le déroulement de l'algorithme Apriori peut être décomposé en quatre étapes :

1. Recherche de tous les nouveaux candidats.
2. Pour chaque candidat trouvé, on calcule son support.
3. Évaluation du support calculé par l'algorithme par rapport au support minimum défini par l'utilisateur.
4. On supprime les candidats dont le support est inférieur au support minimum.

La description complète de l'algorithme Apriori se résume dans les étapes suivantes [45] :

Input : un support minimum et une base de données de transactions
Output : génération des itemsets fréquents

1. $M_1 = \Phi; i = 0$
2. $K_1 =$ tous les 1-itemsets dans la base de données
3. $F_1 =$ tous les itemsets fréquents de K_1

Tant que(M_{i+1} est non vide) Faire

1. $K_{i+1} =$ Candidate-gen(F_i)
2. $F_{i+1} =$ tous les itemsets fréquents de K_{i+1}
3. $i++$
4. Retourner l'union des M_i

Fin tant que

Figure 3.1 Algorithme Apriori

3.5 Avantages et inconvénients des règles d'association

3.5.1 *Avantages*

Les règles d'association représentent plusieurs avantages parmi lesquels on peut citer par exemple: [39, 45]

1. Leur application dans plusieurs domaines de la vie quotidienne, comme l'analyse du panier de la ménagère.
2. La découverte de connaissances utiles, cachées dans les grandes bases des données.
3. Leur simplicité, efficacité et facilité de compréhension.
4. Leur formalisme non supervisé et général.
5. Leurs résultats clairs et faciles à interpréter.

3.5.2 *Inconvénients*

Malgré les grands avantages que les règles d'association peuvent représenter, elles ont aussi des faiblesses qu'on peut résumer dans :

1. Le temps énorme consacré à la recherche des ItemSets fréquents. [29]
2. La grande quantité des règles d'association générées. [40]
3. La difficulté d'évaluer la qualité des règles d'associations par des indices statiques ou par l'expert du domaine. [44]
4. La production des règles triviales et inutiles qui n'apportent pas de nouvelles informations. [54]

Chapitre 4 - Les règles d'association maximales

4.1 Problématiques des règles d'association ordinaires

Les règles d'associations régulières fournissent des moyens de découvrir beaucoup d'associations intéressantes, mais elles ne permettent pas de trouver des associations moins fréquentes qui sont cachées à l'intérieur des corpus. [48]

En utilisant des algorithmes réguliers pour extraire des règles d'association, nous perdons certaines d'entre-elles. En effet, dans un ensemble de documents, des mots étroitement liés, apparaissent fréquemment ensembles. Le mot imprimante par exemple, peut apparaître souvent avec le mot papier. Dans ce cas, on aura des associations spécifiquement appropriées à un terme, mais pas à d'autres. Une association entre imprimante et encre, aurait une confiance basse, puisqu'il y a beaucoup de transactions où le mot imprimante est lié au mot papier (sans le mot encre). [48]

Pour montrer la faiblesse des règles d'associations régulières, on considère l'exemple suivant : [48]

Soit D une base de données des maladies et des symptômes d'un hôpital. Supposant par exemple qu'on a une maladie A qui apparaît dans 70 % des rapports et une autre maladie B qui apparaît seulement dans 15 % des rapports. Soit A la maladie avec les symptômes x, y, z et B la maladie avec les symptômes x ou y, mais pas tous les deux. Si nous cherchons des associations régulières, nous pouvons obtenir l'association entre A et (x, y, z), mais nous

manquerons l'association liant x seul et y seul avec B. La raison est qu'il y a beaucoup de cas reliant A à x, y et z, ce qui réduit la confiance de $x \Rightarrow B$ et $y \Rightarrow B$.

Pour remédier à ce problème, nous devons capturer la notion suivante : chaque fois que x apparaît seul alors B doit apparaître aussi, avec une haute confiance. Des règles d'association régulières ne sont pas capables de capturer de telles associations.

4.2 Présentation des règles d'association maximales

Les règles d'association maximales représentent un moyen efficace pour contourner le problème défini auparavant. Elles permettent d'extraire des relations moins intéressantes qui ne peuvent pas être capturées par des règles d'associations ordinaires. [48]

4.2.1 *Principes de bases des règles d'association maximales*

Les règles d'associations maximales représentent un outil efficace pour dégager des relations intéressantes dans un corpus; elles se basent principalement sur les notions suivantes : [48]

1. Pour une transaction t et un ensemble d'éléments X du même groupe, on dit que t supporte X si $X \subseteq$ t et on le dénote par $s_D(X)$, qui représente le nombre des transactions $t \in D$ qui supportent X.

Exemple 1 :

Considérons les 3 transactions suivantes :

C1 : A, 1, k

C2 : M, L, 2

C3 : A, 1, M

Et soit notre X= {A}

Alors les transactions qui supportent X sont C1 et C3, car ces deux ensembles contiennent l'élément X= {A}.

2. Le support de $X \Rightarrow Y$, où X et Y sont deux sous-ensembles d'éléments disjoints, est le support de $X \cup Y$ et sa confiance est :

$$c_D(X \Rightarrow Y) = \frac{s_D(X \cup Y)}{s_D(X)} \tag{4.1}$$

Considérons les transactions de l'exemple 1, et soit la règle d'association $A \Rightarrow 1$.

Le support de $A \Rightarrow 1$est le nombre de transactions qui contiennent A et 1. Dans notre cas, c'est C1 et C3 qui contiennent ces deux éléments, donc le support de cette règle d'association est 2.

3. Dans une règle d'association maximale $X \xrightarrow{\max} Y$, on s'intéresse à capturer la notion suivante : chaque fois X apparaît seul, Y apparaît aussi, mais avec une certaine confiance.

4. Élément apparaît seul : pour une transaction t, une catégorie g_i et un ensemble d'éléments $X \subseteq g_i$, nous disons que X est seul dans t, si et seulement si $t \cap g_i = X$, c'est-à-dire que X est le plus grand sous-ensemble de g_i qui appartient t. Dans ce cas, on dit que X est maximale dans t et que t M-Supporte X. On dénote le M-Support de X dans D par $s_D^{\max}(X)$, qui représente le nombre de transactions $t \in D$ qui M-Supporte X.

Exemple 2 :

Considérons les 3 transactions suivantes :

C1 : A, 1, K

C2 : M, L, 2

C3 : A, 1, 2

Et les deux catégories suivantes :

Lettres : {A, M, K, L}

Chiffres :{1, 2}

Considérons X= {A}

Dans la transaction C1 par exemple, X n'est pas seul dans C1 puisque :

(C1 $\cap$ lettre)= {A, K}.

Par contre, dans la transaction C3, X est seul puisque :

(C3 $\cap$ lettre)= {A}.

5. Une règle d'association maximale ou M-Association est la règle de la forme $X \xrightarrow{max} Y$, avec X et Y deux sous-ensembles qui appartiennent à deux catégories différentes g(X) et g(Y).

6. Le M-support : Le M-support de l'association maximale $X \xrightarrow{max} Y$ est : $s_D^{max}(X \xrightarrow{max} Y) = \{t : t \text{ M-supporte X et t supporte Y}\}$, où $s_D^{max}(X \xrightarrow{max} Y)$ est le nombre de transactions dans D qui M-supporte X et supporte Y dans le sens régulier.

Exemple 3 :

Considérons les 3 transactions suivantes :

C1 : A, 1, K

C2 : M, L, 2

C3 : A, 1, 2

Et les deux catégories suivantes :

Lettres : {A, M, K, L}

Chiffres :{1, 2}

Et soit la règle d'association maximale suivante :

$A \xrightarrow{max} 1$

Le M-Support de cette règle d'association est le nombre de transactions qui M-Supporte X et Supporte au même temps Y.

Dans notre exemple, seulement la transaction C3 qui M-Supporte A, puisque :

(C3 $\cap$ lettre)= {A}. (**'A'** est seul dans sa catégorie).

(C1 $\cap$ lettre)= {A, **K**}. (**'A'** n'est pas seul dans sa catégorie).

Tandis que dans la transaction C2 on ne trouve pas l'élément **'A'**.

Pour le même exemple on a : C1 et C3 **Supporte** Y= {1} puisque :

(C1 $\cap$ Chiffre)= {1}.

(C3 $\cap$ Chiffre)= {1, 2}.

Donc le nombre des transactions qui M-Supporte A $\xrightarrow{max}$ 1 est égale à 1, puisque seulement C3 qui vérifiée les deux conditions.

7. La M-confiance: soit D(X, g(Y)) le sous-ensemble de la base de donnée D constitué de toutes les transactions qui M-supporte X et qui contient au moins un élément de g(Y). le M-confiance de la règle $X \xrightarrow{max} Y$ est :

$$c_D^{max}(X \xrightarrow{max} Y) = \frac{S_D^{max}(X \xrightarrow{max} Y)}{|D(X, g(Y)|} \quad (4.2)$$

Exemple 4 :

Considérons les 3 transactions suivantes :

C1 : B, 2

C2 : M, L, 2

C3 : B, 1, 2

Et les deux catégories suivantes :

Lettres : {B, M, K, L}

Chiffres :{1, 2}

Considérons la règle d'association maximale suivante :

B $\xrightarrow{max}$ 1

De la même manière que l'exemple 3, le M-support S_D^{max} de la règle d'association B $\xrightarrow{max}$ 1 est 1(seulement la transaction C3 qui **M-Supporte** B et **Supporte** 1)

Pour calculer le M-Confiance, on doit trouver la valeur de $D(X, g(Y)$. Dans notre cas, nous gardons C1 et C3 puisque :

C1 : contient X= {B} seul et il contient aussi l'élément 2 qui appartient à la catégorie de Y= {1}.

C3 : contient X= {B} seul et il contient aussi les deux éléments 1 et 2 qui appartiennent à la catégorie de Y= {1}.

Donc $D(A, g(1)=2$

On peut conclure que la M-Confiance est égale dans ce cas à : $M - Confiance = \frac{1}{2} = 0.5 = 50\ \%.$

8. Le M-support minimum $\hat{s}$: Le M-support minimum $\hat{s}$ d'une règle d'association maximale est le support minimum défini par l'utilisateur que la règle d'association doit satisfaire pour être acceptée.

9. La M-confiance minimum $\hat{c}$: La M-confiance minimum $\hat{c}$ d'une règle d'association maximale, représente la confiance minimale définie par l'utilisateur que la règle doit satisfaire pour être acceptée.

10. Un ensemble X avec au moins le M-support minimum $\hat{s}$ est appelé ensemble M-Fréquent.

4.3 Exemple d'utilisation des règles d'association maximales

On considère cette base de données avec dix transactions : [48]

ID	Transaction
1	J,K,m,s,5
2	J,K,D,u,z,5,2,3
3	J,K,C,z,5
4	J,K,m,s,z,2,3,4
5	C,z,2,3
6	J,K,u,5,3
7	C,D,z,5,2
8	J,K,u,m,s,4
9	J,D,z,2,4
10	J,K,m,s,z,5

Tableau 4.1 Table des transactions

On regroupe les éléments en trois catégories :

Majuscules= {J, K, C, D}

Minuscules= {u, m, s, z}

Chiffres= {2, 3, 4, 5}

On choisi un M-support minimum $\hat{s}$ =3 et une M-confiance minimum $\hat{c}$ =75 %.

Donc on a pour :

1. $J, K \xrightarrow{max} m$

M-support=4 et M-confiance= $\frac{4}{5} \times 100 = 80\%$

2. $5 \xrightarrow{max} J, K$

M-support=3 et M-confiance= $\frac{3}{3} \times 100 = 100\%$

3. $z \xrightarrow{max} C$

M-support=3 et M-confiance= $\frac{3}{4} \times 100 = 75\%$

Si on utilise les règles d'association régulières, on obtiendra pour l'exemple précédent les résultats suivants :

1. $J, K \Rightarrow m$

Support=4 et confiance= $\frac{4}{6} \times 100 = 57\%$

2. $5 \Rightarrow J, K$

Support=5 et confiance= $\frac{5}{6} \times 100 = 83\%$

3. $z \Rightarrow C$

Support=3 et confiance= $\frac{3}{7} \times 100 = 42\%$

Dans ce cas, si on choisi une confiance minimale de référence de 60%, on ne gardera que le deuxième cas avec 83%, puisque :

Confiance = $83 > 60$

4.4 Comptage des règles d'association maximales

Le comptage des règles d'association maximales est plus facile que celui des règles d'associations régulières, car pour chaque catégorie, n'importe quelle transaction M-Supporte au maximum un seul ensemble d'éléments. En effet: [48]

1. Tous les ensembles M-fréquent peuvent être produits dans un petit nombre de passe sur la base de données.
2. Pour chaque catégorie, les ensembles M-fréquent divisent la base de données en sous bases de données, qui contiennent les transactions liées aux différents règles d'association.

4.4.1 *Algorithme*

Nous présentons dans ce qui suit le fonctionnement de l'algorithme des règles d'association maximales : [48]

Soit D une base de données, G le groupement des littéraux en des catégories, $\hat{s}$ le support minimum et $\hat{c}$ la confiance minimum.

1. $M \leftarrow Ensembles - M - Fréquents(\hat{s})$
2. Pour chaque $X \in M$ faire
3. Pour chaque $g \in G$ faire
4. $D' \leftarrow D(X, g)$
5. $\bar{s} \rightarrow \max(\hat{s}, \hat{c}.|D'|)$
6. $F \leftarrow Ensembles - Frequents(D', \delta)$
7. Pour chaque $Y \in F$ faire
8. Sortie $X \overset{\max}{\Rightarrow} Y$

9. Sortie M-support$=s_{D'}(Y)$ M-Confiance$=\frac{s_{D'}(Y)}{|D'|}$

10. Fin pour

11. Fin pour

12. Fin pour

Procédure des ensembles M-fréquents ($\hat{s}$)

1. $l\arg e \leftarrow \Phi$

2. Pour chaque $t \in D$ Faire

3. Pour chaque $g \in G$ Faire

4. $X \leftarrow t \cap g$

5. if $X \neq \Phi$ alors

6. $Hash(X)++$

7. Fin Pour

8. Fin Pour

9. Pour chaque $t \in D$ Faire

10. Pour chaque $g \in G$ Faire

11. $X \leftarrow t \cap g$

12. if $X \neq \Phi$ et $Hash(X) \geq \hat{s}$ alors

13. $S^{\max}(X)++$

14. if $S^{\max}(X) \geq \hat{s}$ alors

15. $l\arg e \leftarrow l\arg e \cup \{X\}$

16. Pour chaque $g' \in G$ Si $g' \cap t \neq \Phi$ Alors

17. $D(X,g') \leftarrow D(X,g') \cup \{t \cap g'\}$

18. Fin pour

19. Fin Pour

20. Fin Pour

21. Retourner Large

$Ensembles - Frequents(D', \delta)$

Rechercher et calculer le support de tous les ensembles d'éléments, avec $Support \geq \bar{s}$.
Il existe plusieurs algorithmes pour trouver les ensembles fréquents.

Figure 4.1 Algorithme des règles d'association maximales

D'une manière plus simple, on peut résumer le fonctionnement de cet algorithme dans les étapes suivantes : [48]

1. L'algorithme commence par générer tous les ensembles M-fréquents en utilisant la procédure **Ensembles-M-fréquents**.

2. Pour chaque ensemble M-Fréquent X, et chaque catégorie g, la méthode **Ensembles-M-fréquents** sélectionne tous les sous bases de données D(X, g) qui contiennent entre autres, la portion de t incluse dans g (t représente les transactions qui M-supporte X).

3. L'algorithme calcule dans cette étape la partie droite de la règle d'association $X \xrightarrow{\max} Y$. Soit X un ensemble M-Fréquent et $X \xrightarrow{\max} Y$ une M-association, avec $Y \subseteq g$. Et Soit D' une sous base de données de D(X, g) et $s_{D'}(Y)$ le support de Y dans D'. Le M-support de $X \xrightarrow{\max} Y$ est $s_{D'}(Y)$ et sa M-confiance est $\frac{s_{D'}(Y)}{|D'|}$. Ainsi, pour trouver les M-associations avec un M-support minimum $\hat{s}$ et une M-confiance minimum $\hat{c}$, nous devons rechercher dans D' tous les ensembles Y avec un support $\bar{s} = \max(\hat{s}, \hat{c}.|D'|)$ (Ligne 5 et 6).

4. La procédure **M-fréquent-sets()** utilise deux passes sur la base de données. Le premier passe sert à réduire le nombre des candidats de l'ensemble M-fréquent (en utilisant une table de hachage sur X), tandis que le second passe sert à générer les ensembles M-fréquents et les sous bases de données adjacentes D(X, g).

5. Dans le premier passe (ligne 2 à 8), la base de données est scannée en séquence pour chercher les ensembles **supportés** par t. Pour chaque X trouvée, on incrémente la valeur de Hash(X) par 1. Ensuite, dans le deuxième passe (ligne 9 à 19), la base de donnees et scannée une deuxième fois pour conserver seulement les ensembles d'éléments X qui **M-supporte** t et dont la valeur du paramètre Hash(X) est supérieure ou égale à $\hat{S}$.

Chapitre 5 - Système développé

5.1 Introduction

Ce mémoire traite de la problématique de la classification textuelle dans son application à l'extraction des règles d'association maximales. En effet, ce projet a pour objectif de classifier un texte encodé en Unicode et d'utiliser par la suite, l'algorithme des règles d'association maximales, pour calculer le M-support et la M-confiance d'un ensemble de mots choisi par l'utilisateur.

À la différence des autres travaux qui utilisent des transactions statiques pour calculer les règles d'association, nous pensons que l'utilisation des classes, comme des transactions, s'avère très efficace dans la recherche des relations intelligibles entre les attributs d'un corpus. En effet, nous intégrons les règles d'association maximale dans le système de classification Gramexco (Biskri et Delisle, 2002) pour profiter des avantages des deux approches. L'utilisation de cette technique permet, entre autres, de ne pas dépendre des connaissances préétablies, qui ne sont pas toujours disponibles, mais plutôt de s'appuyer sur des connaissances dynamiques, non connues à priori. Dans notre approche, l'utilisation des techniques de classification combinées aux règles d'associations donnera, chaque fois, des classes et des règles d'association variables, selon les paramètres choisis par l'utilisateur pendant la phase de classification, comme par exemple: le type de découpage (paragraphe, mot, phrase, etc.), le nombre des N-Gram et la valeur du paramètre de vigilance.

Nous pensons que la combinaison de la classification avec les règles d'association maximales représente beaucoup d'avantages. Elle permettra, entre autres, La découverte des connaissances cachées, souvent utiles, à partir de gros volumes de données, et cela, en calculant le M-Support et la M-Confiance pour mesurer les dépendances entre les mots.

Un des points forts de notre travail est son utilisation de l'encodage UTF (Unicode transformation format) qui vise à représenter les caractères de n'importe quelle langue, par un identifiant et un nom. En effet, Cette utilisation de l'Unicode facilitera le traitement de plusieurs langues, comme l'arabe par exemple. [30]

Nous présenterons, dans un premier temps, la méthode utilisée pour l'analyse et la classification des corpus. Pour cela, nous avons opté pour Gramexco (Biskri et Delisle, 2002), en lui apportant quelques modifications. Cette transformation consiste à utiliser l'encodage Unicode pour représenter les caractères spécifiques à chaque langue, ainsi que l'ajout de certaines fonctionnalités qui n'existaient pas auparavant.

Dans la deuxième partie, nous montrerons l'application des règles d'association maximales sur les classes générées par Gramexco. En effet, cette étape consistera à utiliser les classes trouvées dans la première étape, pour calculer le M-Support et la M-Confiance entre deux ensembles de mots.

5.2 Architecture du système développé

L'idée principale de notre modèle est celle d'un système d'interaction à doubles couches. Le premier niveau permet de faire la classification d'un document, tandis que le

deuxième niveau utilise les classes générées par la première couche, en tant que transactions, pour calculer le M-Support et la M-Confiance entre deux ensembles de mots.

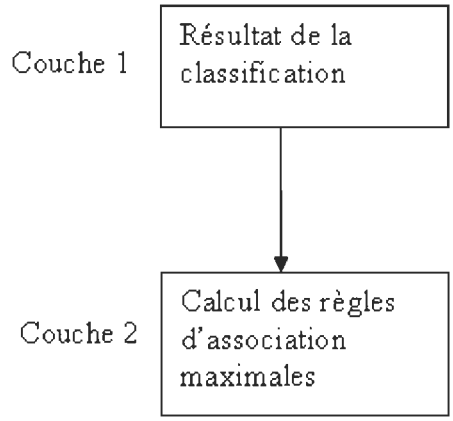

Figure 5.1 Architecture du système développé

5.3 Fonctionnement du système développé

Le processus d'extraction des règles d'association maximales à partir des classes générées par Gramexco se déroule en sept étapes :

1. Choix du type de découpage

Cette étape consiste à choisir un type de segmentation pour le texte sélectionné par l'utilisateur. À la différence de la version précédente de Gramexco, où les documents doivent être en ASCII, la nouvelle version n'accepte que des textes encodés en Unicode, et cela, pour permettre le traitement de plusieurs langues, tel que l'arabe par exemple.

La segmentation d'un texte peut prendre plusieurs formes. Ainsi, on peut diviser un document en mots, paragraphes, phrases ou selon un marqueur unitaire (point, virgule, etc.).

Dans la nouvelle version de Gramexco, nous avons ajouté un nouveau type de segmentation qui consiste à choisir un mot comme séparateur.

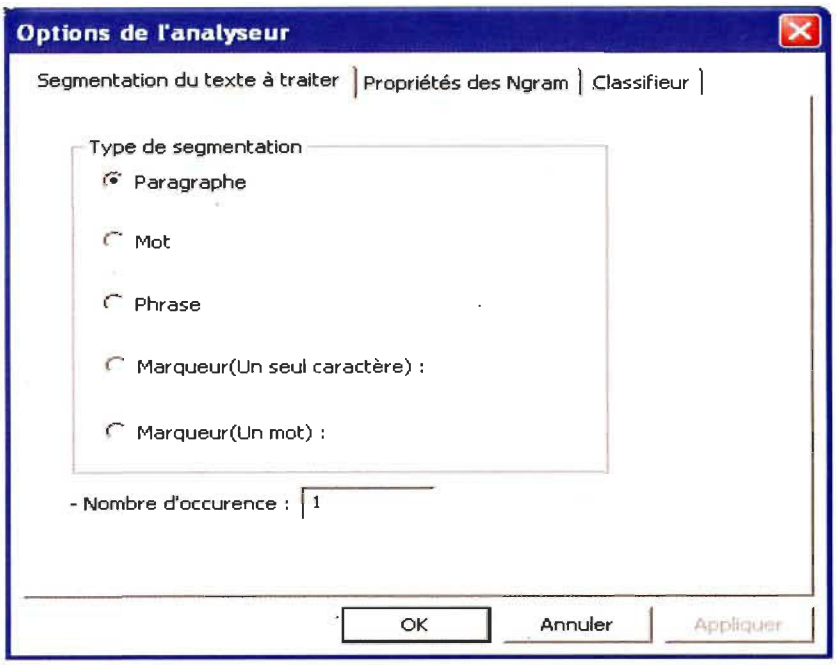

Figure 5.2 Choix du type de segmentation d'un texte

En parallèle avec la segmentation, le texte est divisé en N-Grams pour permettre la comparaison des différents segments de notre corpus. Selon (McNamee et Mayfield, 1998), il s'agit de découper un mot ou un ensemble de mots en séquence de n lettres en gardant les espaces entre ces derniers. En respectant ce principe, et pour n=3, le mot 'Segmentation' par exemple, sera représenté par : Seg, egm, gme, men, ent, nta, tat, ati, tio, ion. Ce type de découpage représente plusieurs avantages, surtout pour certaines langues comme l'arabe ou l'allemand (Biskri et Delisle, 2001). Pour l'allemand par exemple, il est difficile d'identifier un mot simple dans une expression ou une phrase, car la notion d'espace utilisée dans les autres langues, pour repérer les mots simples, est absente dans certaines expressions de la langue allemande, comme par exemple l'expression «Nachbargemeinden » qu'on peut traduire par «Des communes voisines». (Biskri et Delisle, 2001)

De plus, cette première étape nous donne la possibilité de choisir d'autres paramètres tels que:

1. La conversion des N-grams en majuscules ou en minuscules.
2. La conversion des chiffres en blancs pour des corpus qui ne nécessitent pas des analyses sur les chiffres.
3. La conversion des caractères non alphanumériques en blancs.
4. La possibilité de supprimer les mots fonctionnels.
5. La lemmatisation d'un mot, qui consiste à réduire un mot à sa forme racine ou canonique. Dans ce cas, on transforme les verbes en infinitif et les autres mots au masculin singulier.

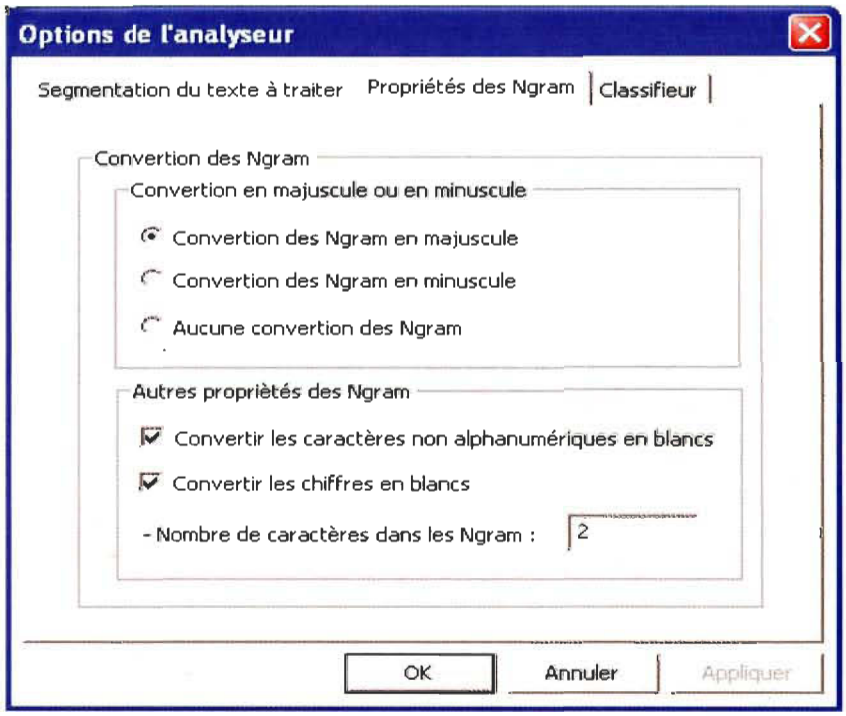

Figure 5.3 Propriétés des N-grams

Figure 5.4 Option du lexique des classes

2. Analyse du texte choisi par l'utilisateur

Cette étape consiste à diviser notre texte en plusieurs segments. Le type de segmentation est choisi par l'utilisateur pendant la première phase de paramétrage de notre système.

En parallèle avec la segmentation, le texte est divisé en N-grams pour permettre la comparaison des différents segments de notre corpus.

Le résultat de la division de notre corpus en segments et en N-grams est enregistré par la suite dans une matrice, qui contiendra, les numéros de segments, les numéros des N-grams, ainsi que la fréquence de ces derniers dans les différents segments. Cette matrice sera utilisée dans l'étape suivante par ART pour classifier les différents segments de notre corpus.

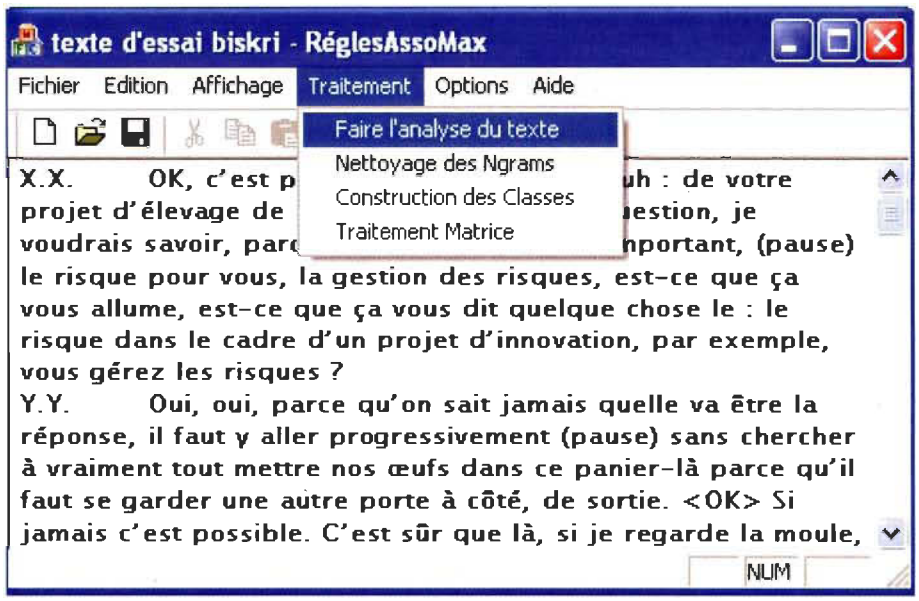

Figure 5.5 Analyse d'un texte choisi par l'utilisateur

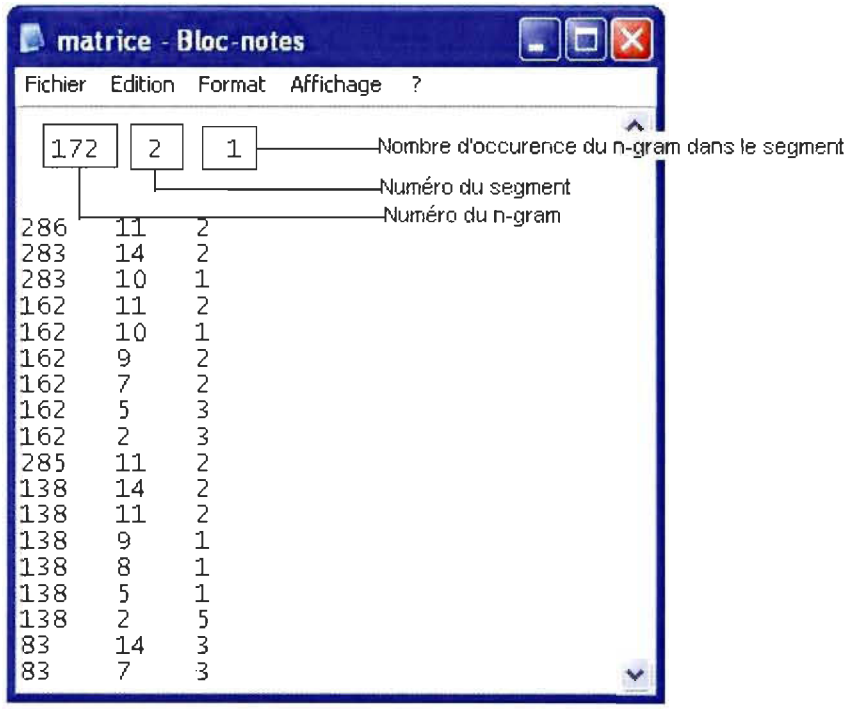

Figure 5.6 La matrice qui contient les résultats de la segmentation

3. Nettoyage des N-grams

On élimine dans cette étape tous les N-grams fonctionnels, ceux avec espace, ainsi que les N-grams dont la fréquence est inférieure ou égale à un certain intervalle choisi par l'utilisateur.

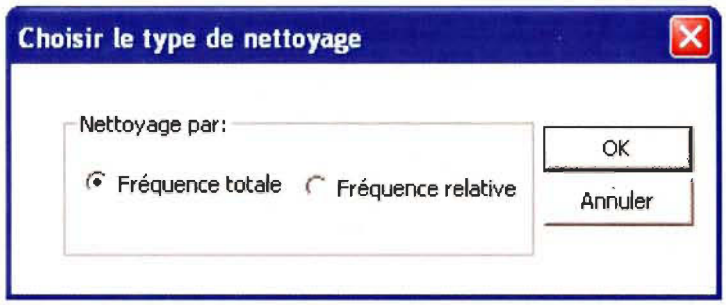

Figure 5.7 Choix d'un type de nettoyage

Notre système donne le choix entre 2 types de nettoyages :

1. Nettoyage par fréquence totale : Ce type de nettoyage consiste à éliminer les N-grams dont la fréquence totale dans notre corpus se situe dans un intervalle choisi par l'utilisateur.

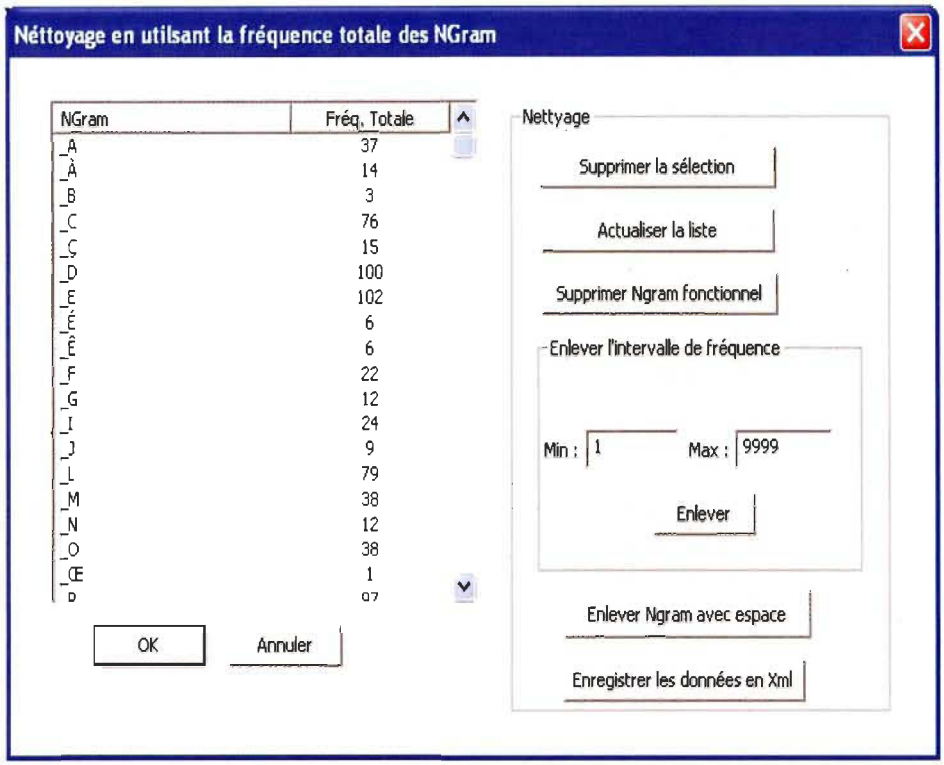

Figure 5.8 Nettoyage par fréquence totale

À la différence de Gramexco qui permet seulement de faire un nettoyage par fréquence totale, notre système permet de faire un autre type de nettoyage qu'on appelle le :

2. Nettoyage par fréquence relative : Il s'agit de supprimer les N-grams dont la fréquence, dans chaque segment, est inférieure ou égale à un intervalle choisi par l'utilisateur.

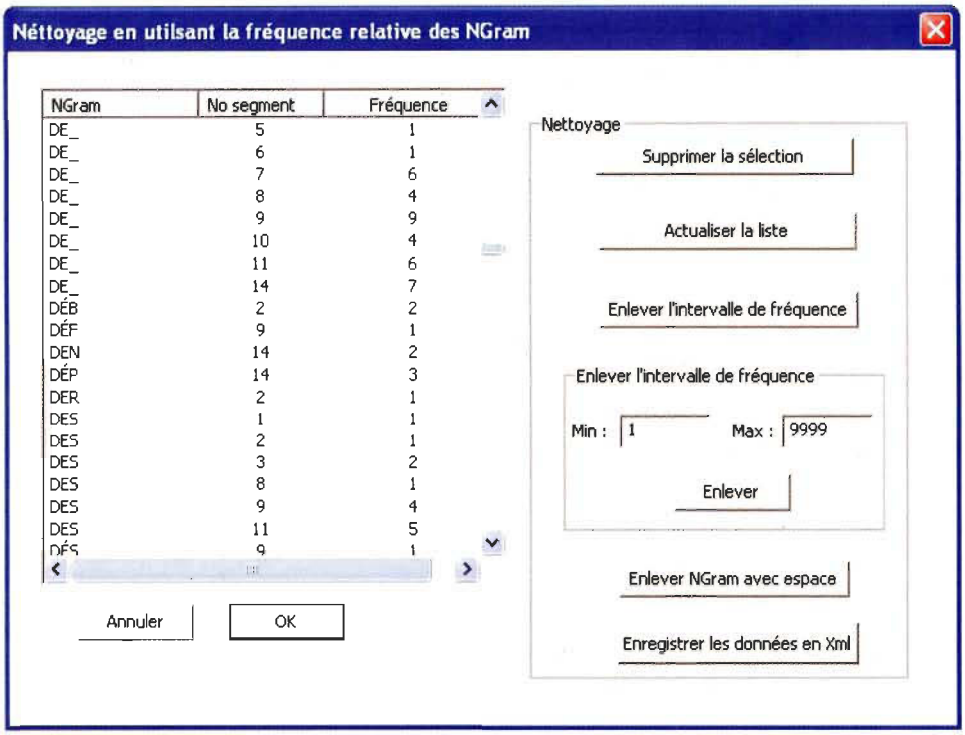

Figure 5.9 Nettoyage par fréquence relative

L'autre nouveauté de notre système est sa capacité d'enregistrer les informations nettoyées en format XML pour faciliter l'échange de ces données via le Web. Puisque les applications qui existent sont nombreuses et illimitées, en utilisant XML on est certains de pouvoir communiquer nos informations auprès de systèmes hétérogènes.

```
<?xml version="1.0" encoding="utf-16" standalone="no" ?
>
<!-- Liste des Ngram et leur Fréquence  -->
- <NGram>
- <InformationSurLesNgram>
    <valeur>ABL</valeur>
    <Fréquence>2</Fréquence>
  </InformationSurLesNgram>
- <InformationSurLesNgram>
    <valeur>ABR</valeur>
    <Fréquence>1</Fréquence>
  </InformationSurLesNgram>
- <InformationSurLesNgram>
    <valeur>AÇO</valeur>
```

Figure 5.10 Enregistrement des N-grams en format XML

4. Construction des classes

Le classificateur de neurones ART1 intervient à cette étape pour comparer et classifier les différents segments obtenus dans les étapes précédentes.

Cette classification se base principalement sur la comparaison des N-grams qui appartient aux différents segments. On dit que deux segments appartiennent à la même classe si et seulement si ils contiennent les mêmes N-grams avec les mêmes fréquences (Biskri et Delisle, 2001).

La qualité de la classification dépend essentiellement d'un paramètre qu'on appelle le paramètre de vigilance v, où $0 \leq v \leq 1$. Si ce paramètre est grand, alors on a plus de classes qui contiennent moins d'éléments, tandis que si cette valeur est petite, on obtient un nombre faible de classes avec un grand nombre de segments. [50]

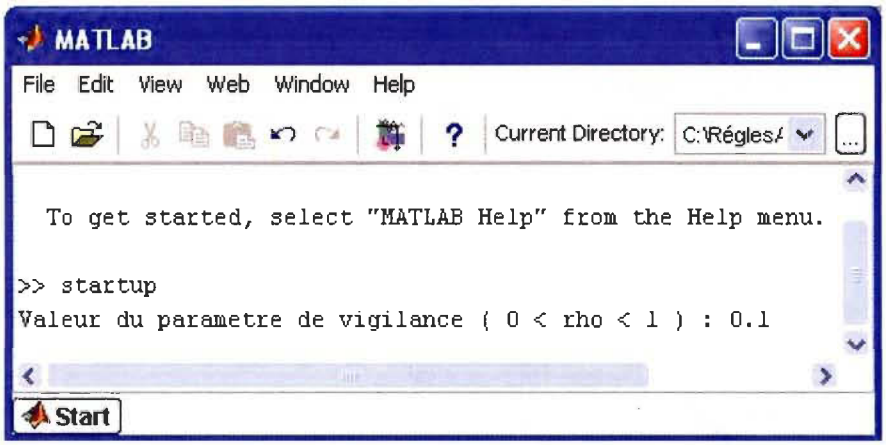

Figure 5.11 Construction des classes par MATLAB

5. Affichage des classes

Cette étape consiste à générer les résultats de notre analyse. Dans un premier temps, nous affichons les classes trouvées par ART1 ainsi que les différents segments qui leur appartiennent. Ensuite, pour chaque segment trouvé, nous affichons l'ensemble de ses mots. À ce niveau, nous pouvons aussi appliquer certaines opérations sur les segments. Parmi ces opérations nous citons (Biskri et Delisle, 2001):

1. **L'union** : Nous utilisons cette opération pour afficher l'union des mots de chaque classe. L'objectif principal de cette tâche est de déterminer le sujet partagé par l'ensemble des segments.
2. **La différence** : Cette opération consiste à produire et afficher les mots non partagés par les différents segments.
3. **L'intersection** : On utilise l'opération d'intersection pour connaître le thème partagé par les différents segments.
4. **L'occurrence** : L'opération d'occurrence permet d'afficher les mots des segments dont la fréquence est supérieure ou égale à un seuil choisi par l'utilisateur.

6. Affichage des règles d'association maximales

Pour bien comprendre le fonctionnement de notre système, l'exemple suivant montrera en détail le processus d'extraction des règles d'association maximales à partir des classes générées par Gramexco :

Soit E={x, a, b, c, d, e, f} l'ensemble du lexique de notre corpus, et considérons les trois classes suivantes :

Classe 1 : {**x**, a, b, c}

Classe 2 : {a, c, d}

Classe 3 : {**x**, e, f, d}

Le processus d'extraction et de calcul des règles d'associations maximales à partir de ces classes se déroule en trois étapes :

I. Choix de X :

Nous choisissons parmi la liste des éléments de E le lexique qui va représenter notre X. Considérons par exemple X=x.

II. Choix de Y :

Cette étape consiste à calculer le M-Support et la M-Confiance de X par rapport aux éléments qui appartient à la même classe que X. Elle donne aussi à l'utilisateur, la possibilité d'éliminer ou de grader les Y qui ne sont pas des noms.

Dans l'exemple précédent, nous avons deux classes qui contenaient X, soit la classe 1 et la classe 3. Selon la combinaison C choisi par l'utilisateur, la construction de l'ensemble Y à partir de ces deux classes se fait de la manière suivante:

- **Pour la Classe 1 : {x, a, b, c} nous aurons les possibilités suivantes:**

- Si C=1 alors Y prend les valeurs : {a}, {b}, {c}.
- Si C=2 alors Y prend les valeurs : {a}, {b}, {c}, {a, b}, {a, c}, {b, c}.
- Si C=3 alors Y prend les valeurs suivantes : {a}, {b}, {c}, {a, b}, {a, c}, {b, c}, {a, b, c}.

- **Pour la Classe 3 : {x, e, f, d} nous aurons les possibilités suivantes:**

- Si C=1 alors Y prend les valeurs : {e}, {f}, {d}.

- Si C=2 alors Y prend les valeurs : {e}, {f}, {d}, {e, f}, {e, d}. {f, d}.
- Si C=3 alors Y prend les valeurs : {e}, {f}, {d}, {e, f}, {e, d}. {f, d}, {e, f, d}.

III. Calcul du M-Support et de la M-confiance :

Nous calculons dans cette étape le M-Support et la M-Confiance de X par rapport aux différents valeurs de Y trouvées dans l'étape précédente.

Le calcul de ces deux paramètres se déroule en quatre étapes :

1. Nous choisissons dans cette première étape une combinaison C parmi la liste des combinaisons de Y déjà trouvées.
2. On suppose que X est seul dans sa catégorie, on doit construire la catégorie g_i de chaque combinaison Y. En effet, la catégorie de Y représente l'intersection des catégories de chaque élément de Y.

Pour Y= {a, c} par exemple, on construit sa catégorie de la manière suivante:

- **Construction de la catégorie de 'a' et de 'c'** :

On considère les trois classes de notre exemple :

Classe 1 : {**x**, a, b, c}

Classe 2 : {a, c, d}

Classe 3 : {**x**, e, f, d}

La catégorie de **'a'** est l'union des ensembles incluant **'a'**, c'est-à-dire l'union de la **classe 1** et la **classe 2** sans l'élément **'x'** :

$$catégorie(a) = classe1 \cup classe2 = \{a,b,c\} \cup \{a,c,d\} = \{a,b,c,d\}$$

De la même façon, la catégorie de **'c'** est l'union des classes incluant **'c'**, c'est-à-dire l'union de la classe 1 et la classe 2 sans l'élément **'x'** :

$$catégorie(c) = classe1 \cup classe2 = \{a, b, c\} \cup \{a, c, d\} = \{a, b.c, d\}$$

Donc La catégorie de Y= {a, c} est l'intersection de la catégorie de **'a'** avec celle de **'c'**, tel que:

$$catégorie(Y) = Catégorie(a, c) = Catégorie(a) \cap Catégorie(c) = \{a, b, c, d\}$$

3. Calculer le M-support et la M-Confiance de chaque combinaison de Y. Le calcul de ces deux paramètres se décompose en deux étapes :

   I. Parcourir l'ensemble des classes que Gramexco a généré. Dans notre exemple, On parcourt les trois classes : Classe1, Classe2 et Classe3.

   II. Pour chaque Classe, vérifier s'elle contient X (par hypothèse, X est unique dans sa catégorie) et s'elle contient aussi Y avec $Y \subseteq catégorie(Y)$; si oui, incrémenter le M-Support P de 1. De la même façon, on vérifie si cette classe contient X (X est seul de sa catégorie) et s'elle contient au moins un élément de la catégorie de Y; si oui, incrémenter le paramètre F, de la M-Confiance, de 1.

Si nous considérons par exemple la relation $x \xrightarrow{\max} a, c$.

En utilisant les classes suivantes comme de transactions :

Classe 1 : {x, a, b, c}; Classe 2 : {a, c, d}; Classe 3 : {x, e, f, d}

Et avec $Catégorie(a, c) = \{a, b, c, d\}$

Le M-support P de cette règle d'association est 1, puisque seulement la classe 1 contient X= {x} et Y= {a, c}.

Le paramètre F de la M-Confiance est égale dans cette situation à 2, puisque la classe1 et la classe3 contiennent toutes les deux X={x} et aussi contiennent des éléments qui appartiennent à la catégorie de Y ($Catégorie(a,c) = \{a,b,c,d\}$).

Donc la valeur de la M-Confiance est : $M - Confiance = \frac{P}{F} \times 100 = \frac{1}{2} \times 100 = 50\%$

4. Affichage des résultats, avec le M - Support = P et la M - Confiance $= \frac{P}{F} \times 100$.

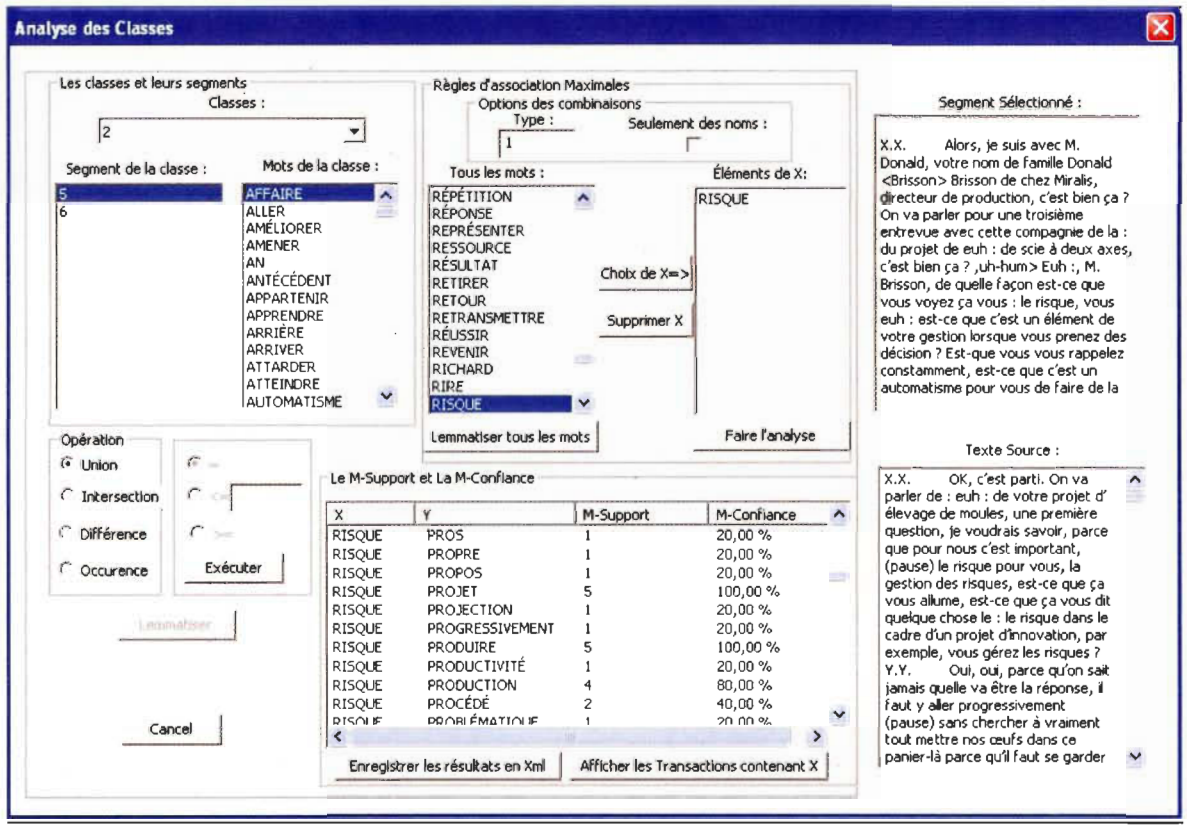

Figure 5.12 Analyse et affichage des résultats

7. Enregistrement des résultats en XML

Cette dernière étape consiste à enregistrer les résultats trouvés, c'est-à-dire X, Y, le M-support et la M-confiance en format XML.

```
- <InformationSurLeMSupportEtLaMConfiance>
    <ElementX>RISQUE</ElementX>
    <ElementY>PROJET</ElementY>
    <M-Support>2</M-Support>
    <M-Confiance>100,00 %</M-Confiance>
  </InformationSurLeMSupportEtLaMConfiance>
- <InformationSurLeMSupportEtLaMConfiance>
    <ElementX>RISQUE</ElementX>
    <ElementY>PRODUCTION</ElementY>
    <M-Support>1</M-Support>
    <M-Confiance>50,00 %</M-Confiance>
  </InformationSurLeMSupportEtLaMConfiance>
- <InformationSurLeMSupportEtLaMConfiance>
    <ElementX>RISQUE</ElementX>
    <ElementY>PROBLÈMES</ElementY>
    <M-Support>2</M-Support>
    <M-Confiance>100,00 %</M-Confiance>
```

Figure 5.13 Enregistrement du M-Support et de la M-Confiance en XML

Chapitre 6 - Expérimentations et résultats

Dans ce chapitre, nous présentons une description détaillée des résultats obtenus lors des phases d'expérimention. Nous avons effectué des tests sur un jeu de données qui appartiennent à trois domaines différents; Celui de la gestion, l'histoire et les sciences.

- **Exemple 1 : Gestion du Risque dans les entreprises**

Dans cet exemple, nous proposons d'illustrer notre approche sur un jeu de données portant sur la gestion du risque au sein des entreprises québécoises. Ce texte se base, entre autres, sur une série d'entrevues avec des Dirigeants d'entreprises pour connaître leurs points de vue sur le risque (voir Annexe A). Lors de cette première expérimentation, nous avons mené une série de tests permettant de calculer le M-support et la M-confiance du Mot 'Risque' par rapport au reste du lexique. Pour cela, nous avons opté pour une segmentation par 'Mot' pour diviser notre texte. Le mot choisit est 'StopFin' qui séparent les différents entrevues. Pour le reste des paramètres de la classification, nous avons considéré :

- 2 caractères pour la taille des N-gram.
- 0.1 pour le paramètre de vigilance au niveau de MatLab.
- Les lettres minuscules sont identiques aux lettres majuscules.
- La suppression des N-gram avec espace.

En ce qui concerne la génération des règles d'associations maximales nous avons choisi les paramètres suivants :

- Le mot '**Risque**' pour représenter le 'X' de la règle d'association $X \xrightarrow{\max} Y$.
- ignorer les termes qui ne sont pas des noms.
- Une combinaison C=2 pour générer les différents combinaisons **un à un** et **un** à **deux** pour l'élément Y de l'association $X \xrightarrow{\max} Y$.

Les résultats obtenus après l'analyse de ce texte sont :

X	Y	M-Support	M-Confiance
Risque	Client	1	10 %
	Actionnaires Coût	1	10 %
	Client Projet	1	10 %
	Décision Produit	2	20 %
	An	2	20%
	Marchés Prix	2	20 %
	Scie	3	30 %
	Entrevue Études	3	30 %
	Fonction	4	40 %

	Façon Niveau	5	50%
	Produit	5	50%
	Question	6	60 %
	Entrevue Risque	6	60 %
	Niveau X	7	70%
	Gestion	7	70%
	Gestion Projet	7	70 %
	Pause Risques	8	80%
	Projet Risques	8	80%
	X	10	100 %
	Pause	10	100 %
	Projet X	10	100 %
	Pause X	10	100 %
	Projet	10	100 %

Tableau 6.1 Calcul du M-Support et de la M-Confiance pour X = Risque

Malgré la présence de données bruit, comme par exemple : 'Pause X', 'Pause', 'X', On voit bien que les résultats obtenus sont fort intéressants, surtout entre le mot 'Risque' et :

- **Projet :** 100 %
- **Gestion Projet :** 75 %
- **Gestion :** 70 %
- **Produit :** 50 %

Ces résultats sont étroitement cohérents avec le texte qu'on a utilisé pour notre expérimentation. En effet, selon les réponses des Dirigeants, le risque est toujours lié à la façon de gérer un projet, ainsi qu'au produit de l'entreprise.

➢ **Exemple 2 : Hassan II, le Maroc et l'histoire**

Nous avons choisi pour ce deuxième exemple un texte publié dans la revue « L'histoire en juin 2004 ». Ce texte traite de l'histoire d'Hassan II, qui a régné sur le Maroc de 1961 À 1999 (voir Annexe A).

Lors de cette deuxième expérimentation, nous avons appliqué des opérations de prétraitement différentes de celles qui furent appliquées dans notre premier exemple. En effet, notre texte a été segmenté en paragraphes. Pour le reste des paramètres de classification, nous avons opté pour :

- 2 caractères pour la taille des N-gram.
- 0.6 pour le paramètre de vigilance au niveau de MatLab.
- Les lettres minuscules sont identiques aux lettres majuscules.
- La suppression des N-gram avec espace.

Pour la génération des règles d'associations maximales, nous avons choisi :

- Le Mot **'Hassan'** (nom du roi du Maroc) pour représenter le 'X' de la règle d'association $X \xrightarrow{\max} Y$.
- une combinaison C=1 qui permet de construire les combinaisons un à un pour l'élément Y de l'association $X \xrightarrow{\max} Y$.

Nous présentons dans le tableau suivant un échantillon des résultats obtenus après l'analyse de notre système :

X	Y	M-Support	M-Confiance
Hassan	Docteur	1	7.69 %
	Professeur	1	7.69 %
	Espagne	1	7.69 %
	Tunisie	1	7.69 %
	Espagnol	2	15.38 %
	Journaliste	3	23.08 %
	Histoire	3	23.08 %
	PRÉPARER	3	23,08 %
	TITRE	4	30,77 %
	FRANCE	5	38,46 %
	France	5	38.46 %
	Politique	6	46.15 %
	ANNÉE	7	53,85 %
	Roi	8	61.54 %
	Maroc	8	61.54 %

	II	13	100 %

Tableau 6.2 Calcul du M-Support et de la M-Confiance pour X = Hassan

Les résultats trouvés précédemment sont cohérents avec la description du Mot 'Hassan'. En effet, Le roi Hassan II a régné sur le Maroc de 1961 à 1999, ce qui justifié la forte relation entre ce mot et les termes :

- **II :** 100 %
- **Roi :** 61.54 %
- **Maroc :** 61.54 %

➢ **Exemple 3 : Définition de l'informatique**

Nous proposons d'illustrer dans cet exemple le fonctionnement de notre système sur un jeu de données traitant de l'histoire de l'informatique (voir Annexe A). En effet, ce texte donne une définition de l'informatique, son évolution, ainsi que sa relation avec le domaine de la science.

Nous avons choisi pour cet exemple une segmentation par paragraphe.

Pour le reste des paramètres de classification, nous avons choisi :

- 0.1 pour le paramètre de vigilance au niveau de MatLab.
- 2 caractères pour la taille des N-gram.
- Les lettres minuscules sont identiques aux lettres majuscules.
- La suppression des N-gram avec espace.

En ce qui concerne la génération des règles d'association maximales, nous avons choisi :

- une combinaison C = 1 pour construire nos combinaisons.
- Le Mot **'Informatique'** pour représenter le 'X' de la règle d'association $X \xrightarrow{max} Y$.

Les résultats obtenus après l'analyse de ce texte se résume dans le tableau suivant :

X	Y	M-Support	M-Confiance
Informatique	Administration	1	2.33 %
	Architecture	2	4.65 %
	Américain	3	6.98 %
	Trouver	4	9,30 %
	consister	5	11,63 %
	Nécessaire	5	11,63 %
	Voir	6	13,95
	Computer	6	13.95 %
	Pouvoir	7	16,28 %
	Premier	9	20,93

	Faire	10	23,26 %
	Machine	10	23.26 %
	Ordinateur	10	23.26 %
	Science	15	34.88 %

Tableau 6.3 Calcul du M-Support et de la M-Confiance pour X = Informatique

Les résultats obtenus dans le tableau précédent montrent très bien l'étroite relation du mot **'Informatique'** et les trois mots :

- **Science** : 34.88 %
- **Ordinateur** : 23.26 %
- **Machine** : 23.26 %

En effet, en lisant le texte, on voit très bien que la définition du mot informatique est fortement liée à ces trois termes, comme par exemple dans les phrases suivantes :

- l'informatique désigne l'ensemble des **sciences** et techniques en rapport avec le traitement de l'information.
- La discipline scientifique désignée par le terme informatique fait partie des **sciences** formelles.
- L'informatique n'est pas plus la **science** des ordinateurs.

- Le second acte de naissance de l'informatique est bien entendu la réalisation concrète des premiers **ordinateurs** dans les années 1940.
- L'informatique (information automatique) désigne l'automatisation du traitement de l'information par un système, concret (**machine**).

➢ **Exemple 4 : OPEC**

Ce quatrième exemple traite le cas d'un texte en arabe sur L'organisation des pays exportateurs de pétrole OPEC (voir Annexe A). Il permet de montrer la relation entre le mot «أوبك» (OPEC) et les autres mots de ce texte. Nous avons utilisé pour notre analyse les paramètres suivants :

- Un séparateur de type point pour la segmentation.
- 0.1 pour le paramètre de vigilance au niveau de MatLab.
- 2 caractères pour la taille des N-gram.
- Les lettres minuscules sont identiques aux lettres majuscules.
- La suppression des N-gram avec espace.

Pour la génération des règles d'association maximales, nous avons choisi :

- une combinaison C = 2 pour le type de combinaison.
- Le Mot 'أوبك' (OPEC) pour représenter le 'X' de la règle d'association $X \xrightarrow{\max} Y$.

Nous résumons dans le tableau suivant les résultats obtenus par notre analyse :

X	Y	M-Support	M-Confiance
(OPEC) أوبك	(mécanismes) آليات	1	9,09 %
	(Paris Pays) الدول باريس	1	9,09 %
	(Création tarifs) أسعار إنشاء	2	18,18 %
	(création tarifs) أسعار إنشاء	2	18,18 %
	(pétrole) البترول	3	27,27 %
	(pays membres) الأعضاء الدول	3	27,27 %
	(tarifs) أسعار	3	27,27 %
	(organisation tarifs) أسعار منظم	3	27,27 %
	(création) إنشاء	3	27,27 %

	(membres) اعضاء	4	36,36 %
	Sommet قمة	4	36,36 %
	(monde) العالم	4	36,36 %
	(organisation pays) منظمة الدول	4	36,36 %
	organisation منظمة	6	54,55 %
	(pays) الدول	7	63,64 %
	(dans) في	9	81,82 %

Tableau 6.4 Calcul du M-Support et de la M-Confiance pour X = **أوبك**

Le tableau précédent montre bien l'étroite relation entre le mot أوبك (OPEC) et les deux mots منظمة (Organisation) et الدول (Pays). En effet, en lisant le texte, on voit très bien que le mot أوبك (OPEC) apparaît dans plusieurs endroits avec ces deux termes, comme par exemple dans les phrases suivantes :

- Organisation des pays exportateurs du pétrole OPEC منظمة الدول المصدرة للبترول «أوبك».

- Dirigé par l'Organisation des pays de «l'OPEC»«أوبك» منظمة دول رأسها وعلى .
- L'organisation de l'OPEC a décidée «أوبك» منظمة قررت وقد.
- L'organisation de l'OPEC fait beaucoup d'efforts جهودا «أوبك» منظمة وتبذل.
- La contribution des pays de l'OPEC «أوبك» دول إسهام لكن.

➢ **Exemple 5 : Barack Obama**

Nous proposons dans ce cinquième exemple un texte en arabe résumant la bibliographie du nouveau président des états unis, Barack Obama. Nous avons choisi de calculer pour cet exemple le M-Support et la M-Confinace du nom «اوباما » (Obama) par rapport aux autres mots. Ainsi, les paramètres choisis pour l'analyse de ce texte sont les suivants :

- Une segmentation par paragraphe.
- 0.2 pour le paramètre de vigilance au niveau de MatLab.
- 2 caractères pour la taille des N-gram.
- Les lettres minuscules sont identiques aux lettres majuscules.
- La suppression des N-gram avec espace.

X	Y	M-Support	M-Confiance
(Obama) اوباما	(candidat dernier) حشدم رخأ	1	33,33 %
	(armes) ةحلسأ	1	33,33 %

	(Vie président) الرئيس حياة	1	33,33 %
	(washington américain) أمريكي واشنطن	1	33,33 %
	(comme) مثل	2	66,67 %
	(de) من	2	66,67 %
	(états unis) المتحدة الولايات	2	66,67 %
	(origines africaines) أصول أفريقية	2	66,67 %
	(barack) باراك	3	100,00 %

Tableau 6.5 Calcul du M-Support et de la M-Confiance pour X = اباراو

Les résultats obtenus dans le tableau précédent montre très bien la forte relation entre le nom et le prénom du nouveau président des états unis (M-Confiance =100%) Barack Obama.Il y a aussi une étroite relation entre le nom Obama et les combinaisons suivantes:

- المتحدة الولايات (États Unis) : 66,67 %
- أصول أفريقية (Origines africaines) : 66,67 %

En lisant la bibliographie de Obama on voit très bien qu'il est d'origine africain et qu'il est le président actuel des états unis, ce qui justifié les résultats obtenus précédemment.

Chapitre 7 - Conclusion

Le projet réalisé dans le cadre de ce mémoire se base principalement sur l'utilisation combinée de deux méthodes de fouille de données, la classification textuelle et les règles d'association. Dans un premier temps, et précisément dans le deuxième chapitre, nous nous sommes intéressé à dresser un état de l'art des techniques de classification qui existent. De manière générale, cette étape est divisée en deux parties, une consacrée à la présentation des méthodes de classification supervisées et l'autre dédiée aux techniques de classification non supervisées.

Cet état de l'art sur les techniques de classification nous a mené à prendre connaissance de leur complexité. En effet, avec le nombre incroyable des techniques de clustering qui existent, le choix de l'une d'entre elles est devenu très difficile. De plus, la plupart de ces techniques sont bien plus liées au jeu de données qu'à l'exactitude théorique de la technique elle même [51]. Le choix d'une méthode de clustering dans ce cas dépend essentiellement des résultats obtenus.

Le troisième chapitre a été consacré à la présentation des règles d'association. Nous avons tout d'abord présenté les différents paramètres utilisés pour mesurer la qualité d'une règle d'association. Nous avons procédé par la suite à l'exposition des différentes étapes d'extraction de ces règles à partir d'une base de données transactionnelle. Ensuite nous

avons présenté l'algorithme APRIORI, qui représente la base de tous les algorithmes de recherche des règles d'association. Enfin, nous avons cité dans ce chapitre quelques avantages et inconvénients liés à l'utilisation des règles d'association. Les inconvénients majeurs de ces méthodes résident essentiellement dans : le temps énorme consacré au traitement d'un grand volume de données, la grande quantité des règles d'association générées, la difficulté de leur évaluation et enfin, la génération des règles d'association inutiles qui n'apportent pas de nouvelles informations.

Dans le quatrième chapitre, nous avons mis le focus sur un cas particulier des règles d'association qu'on appelle les règles d'association maximales. Cette méthode a été intégrée dans notre projet pour permettre l'extraction des relations pertinentes entre les mots d'une même classe ou ceux de classes différentes.

L'avantage d'utiliser des règles d'association maximales est leur capacité à extraire des relations pertinentes que les règles d'association ordinaires ignorent.

Dans le cinquième chapitre, nous avons présenté le système développé. À cet effet, nous avons montré le procédé détaillé qu'il faut suivre pour trouver l'ensemble des classes d'un corpus, ainsi que l'utilisation de ces classes comme des transactions pour générer les règles d'association maximales. Notre explication de chaque étape a été accompagnée de captures d'écran qui facilitent la compréhension de notre projet.

Enfin, le sixième chapitre a été consacré à la présentation des résultats obtenus en analysant un ensemble de documents.

En général, notre travail semble très concluant quant à l'importance de l'utilisation des règles d'association maximales avec la classification textuelle. Par contre, ces résultats sont

accompagnés d'un coût énorme en termes de temps de calcul du M-Support et de la M-Confiance. En effet, plus le nombre de combinaisons augmente plus le temps de traitement augmente également.

En général, ce travail de recherche peut être prolongé dans un certain nombre de directions, tel que:

- L'amélioration du temps de calcul du M-Support et de la M-Confiance.
- Calcul du M-Support et de la M-Confiance pour les mots les plus significatifs de notre corpus. Pour cela, on calcul le nombre d'occurrence de chaque terme dans notre texte et on garde à la fin ceux avec un taux d'occurrence élevé.
- l'application de ce modèle pour la recherche documentaire sur le Web. En effet, pour rechercher un ensemble de documents relié à un document **D**, l'utilisateur peut calculer le M-Support et la M-Confiance pour l'ensemble des mots de **D** et d'utiliser par la suite, ceux avec une forte corrélation, pour la recherche des documents similaire à **D**.

Bibliographie

[1] Cantin, A.-A. (1998). Implantations du réseau de neurones Fuzzy Art Montréal, Université du Québec à Montréal.

[2] Touzet, C. (1992). Les réseaux de neurones artificiels, introduction au connexionnisme.

[3] Mathian, H. and L. Sanders (2006). Les méthodes de classification de données spatiales.

[4] Nakache, D. (2007). Extraction automatique des diagnostics à partir des comptes rendus médicaux textuels. Laboratoire CEDRIC – équipe ISID. Paris, Conservatoire National des Arts et Métiers: 219.

[5] Denoue, L. (2003). Classification supervisée de documents.

[6] Quang, C. T. (2005). Classification automatique des textes vietnamiens Hanoi, Institut de la Francophonie pour l'informatique.

[7] Lechevalier, Y. (1980). Méthodes de discrimination non paramétrique asymptotiquement efficaces au sens de bayes, IRINA.

[8] GOSSELIN, B. (1996). Application de réseaux de neurones artificiels à la reconnaissance automatique de caractères manuscrits, Faculté Polytechnique de Mons: 231.

[9] Balmisse, G. (2002). Les réseaux de neurones, ENSTA, Centre de Documentation Multimédia

[10] Stricker, M. (2000). Apprentissage des réseaux de neurones et régularisation, Laboratoire d'Électronique de l'ESPCI.

[11] Chettaoui, A. (2008). Le développement d'un réseau de neurones pour l'analyse en forme de signal dans l'expérience DVCS Clermont Ferrand.

[12] Laboratoire de météorologie physique, L. (2006). "Introduction aux réseaux de neurones."

[13] Département de génie électrique et de génie informatique. (2007).Processus d'apprentissage. Québec, Université Laval.

[14] Dreyfus, G. (2004). Spécial réseaux de neurones, Laboratoire d'électronique de PC.

[15] Labo algo. (2004). "Réseaux de neurones artificiels.".

[16] Hai Anh, H. (2004). Usage des arbres de décision, Institut de la francophonie pour l'informatique.

[17] J. Ross Quinlan, Machine Learning, 1986, « Induction of decision trees », p. 81-106

[18] Hanoune, M. and F. Benabbou (2005). Modélisation Informatique de Clients Douteux, En utilisant les Techniques de DATAMINING. Paris.

[19] NRSA, Méthodologie projet décisionnel. "Les arbres de décision.".

[20] Tufféry, S. (2005). Data mining et statistique décisionnelle: l'intelligence dans les bases de données.

[21] Nasri, M. and M. El Hitmy. Algorithme génétique et critère de la trace pour l'optimisation du vecteur attribut : application à la classification supervisée des images de textures. Oujda, École Supérieure de Technologie.

[22] M. Mitchell(1996). An introduction to genetic algorithm, MIT press.

[23] Lafleur, A. (2008). Les mécanismes de l'évolution. Paris Université Pierre et Marie Curie.

[24] Sabrina, T. (2003). Introduction aux algorithmes génétiques Paris, Université Pierre et Marie Curie-Paris6.

[25] Mohamadally, H. and B. Fomani (2006). SVM : Machines à vecteurs de support ou séparateurs à vastes marges. Versailles St Quentin.

[26] John, Shawe-Taylor (2000), Nello Cristianini, Support Vector Machines and other kernel-based learning methods, Cambridge University Press.

[27] Bolon, P., I. Pétillot, et al. TP de Traitement d'Images, Données multi composantes, Classification. Savoie, Université Chambery Annecy Savoie.

[28] Barbu, E., P. Héroux, et al. (2005). Classification non supervisée hiérarchique incrémentale basée sur le calcul de dissimilarités. 12e rencontre de la société francophone de classification. Mont-Saint-Aignan, Université de Rouen.

[29] Ben Yahia, S. and E. Mephu Nguifo (2004). Approches d'extraction de règles d'association basées sur la correspondance de Galois. Lens, Centre de Recherche en Informatique de Lens.

[30] Gafner, T. (1991). Analyse critique des méthodes classiques et nouvelle approche par la programmation mathématique en classification automatique. Faculté de Droit et des Sciences économiques. Neuchâtel, Université de Neuchâtel: 119.

[31] Ganaoui, O. E. and M. Perrot (2004). Segmentation par régions : une méthode qui utilise la classification par nuées dynamiques et le principe d'hystérésis. Paris.

[32] Nakache, J.-P. and J. Confais (2005). Approche pragmatique de la classification.

[33] Bellot, P. (2003). Méthodes de classification et de catégorisation de textes. Avignon, Université d'Avignon et des Pays de Vaucluse.

[34] Chevalier, M., C. Julien, et al. (2001). Un système de gestion des besoins web au sein d'un groupe d'utilisateurs. XIXe Congres de l'informatique des Organisations et Systèmes d'Information et de Décision. Martigny-Suisse.

[35] Alusse, A., J.-C. Lamirel, et al. (2006). Un outil d'aide à la découverte du contenu des documents et à la création de dossiers. Semaine du document numérique. Université de Fribourg.

[36] El golli, A., R. Fabrice, et al. (2006). "Une adaptation des cartes auto-organisatrices pour des données décrites par un tableau de dissimilarités." Revue de statistique appliquée ISSN 0035-175X CODEN RVSTA7.

[37] Aaron, C. Un algorithme de normalisation des données à l'aide de graphes pour le traitement non-linéaire des données : Application à l'optimisation des cartes de Kohonen. Paris, Université Paris I.

[38] Jouini, W. (2002). Les méthodes et techniques d'Extraction de Connaissances de Base s de données Paris, École centrale: 38

[39] Abdelali, M. and O. Hicham (2003). Création de règles d'association. Caen, Ensicaen.

[40] Marinica, C., F. Guillet, et al. (2008). Vers la fouille de règles d'association guidée par des ontologies et des schémas de règles. QDC 2008. École polytechnique de l'université de Nantes.

[41] Vaillant, B., P. Meyer, et al. (2006). "Mesurer l'intérêt des règles d'association" Revue des Nouvelles Technologies de l'Information (Extraction et gestion des connaissances: État et perspectives).

[42] Lallich, S. and O. Teytaud (2003). "Évaluation et validation de l'intérêt des règles d'association." Revue des nouvelles Technologies de l'information.

[43] Gay, J.-S. (2004). Application des arbres de radicaux à des algorithmes parallèles du datamining. Amiens, Université de Picardie Jules Verne.

[44] Cherfi, H. and Y. Toussaint (2002). Adéquation d'indices statistiques à l'interprétation de règles d'association. Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles. Saint-Malo.1.

[45] Diop, C. T., M. Lo, et al. (2007). Intégration de règles d'association pour améliorer la recherche d'informations XML. Quatrième conférence francophone en Recherche d'Information et Applications. École Nationale Supérieure des Mines de Saint-Étienne.

[46] NRSA, Méthodologie Projet Décisionnel. "Analyse Du Panier De La Ménagère."

[47] Cherfi, H., A. Napoli, et al. (2005). "Deux méthodologies de classification de règles d'association pour la fouille de textes." Revue des nouvelles technologies de l'information

[48] Amir, A., Y. Aumann, et al. (2005). "Maximal association rules: a tool for mining associations in text." Kluwer Academic Publishers Hingham, MA, USA **25**(3).

[49] ANDRIES, P. (2008). Unicode 5.0 en pratique, Codage des caractères et internationalisation des logiciels et des documents.

[50] Torres-Moreno, J.-M., P. Veláquez-Morales, et al. (2000). Classphères : un réseau incrémental pour l'apprentissage non supervisé appliqué à la classification de textes. 5es Journées Internationales d'Analyse Statistique des Données Textuelles. Lausanne, Suisse.

[51] Beck, N. (2006). Application de méthodes clustering traditionnelles et extension au cadre multicritère. Faculté des Sciences appliquées. Bruxelles, Université libre de Bruxelles.

[52] Forest, D. (2002). Lecture et analyse de textes philosophiques assistées par ordinateur : application d'une approche classificatoire mathématique à l'analyse thématique du Discours de la méthode et des Méditations métaphysiques de Descartes. Montréal, Université du Québec à Montréal: 144.

[53] Forest, D. (2006). Application de techniques de forage de textes de nature prédictive et exploratoire à des fins de gestion et d'analyse thématique de documents textuels non structurés. Montréal, Université du Québec à Montréal: 302.

[54] Aby, K. and A. EL Kourri (2003). Post traitement de règles d'association. Caen, ISMRA ENSI Caen.

[55] Zeitouni, K. (2006). Analyse et extraction de connaissances des bases de données spatiotemporelles. Versailles. Université de Versailles Saint-Quentin-en-Yvelines.

[56] Fertil, B. (2006). Reconnaissance des formes: Classement d'ensembles d'objets.

[57] Talbi, E.-G. Fouille de données (Data Mining), Un tour d'horizon. Lille, Laboratoire d'informatique fondamentale de Lille.

[58] Moutarde, F. (2008). Brève introduction aux arbres de décision. Paris, Centre de Robotique (CAOR), École des Mines de Paris.

[59] Geneva, lab. (2007). Les arbres de décision.

[60] Ren (2001). Les algorithmes génétiques. Paris, École centrale paris.

[61] Baptiste, J. (2005). Concepts fondamentaux des algorithmes évolutionnistes.

[62] Trémeaux, J.-M. (2005). Algorithmes génétiques pour l'identification structurelle des réseaux bayésiens. Lyon, Université Lumière - Lyon II.

[63] Drant, Y., M. Lagacherie, et al. (2006). Présentation d'algorithmes de Datamining, Epita.

[64] Graham-Cumming, J. (2006). "Interview de John Graham-Cumming, l'auteur du logiciel antispam PopFile ".

[65] Bellot, P. (2003). Méthodes de classification et de catégorisation de textes. Avignon, Université d'Avignon et des Pays de Vaucluse.

[66] Tanagra (2008). "Traitement de gros volumes – CAH Mixte ".

[67] Leray, P. (2006). Quelques Types de Réseaux de Neurones- La rétropropagation. Rouen, INSA Rouen -Département ASI - Laboratoire PSI.

[68] Barthelemy, S. and J.-B. Filippi (2001). Typologie d'Entreprises et Self-Organizing Maps. Corte, Université de Corse.

[69] Cantin, M.-A. (1998). Implantations du réseau de neurones, Université du Québec à Montréal.

Annexe A – Textes utilisés pour l'expérimentation

- **Exemple 1 : Gestion du Risque dans les entreprises**

X.X. OK, c'est parti. On va parler de : euh : de votre projet d'élevage de moules, une première question, je voudrais savoir, parce que pour nous c'est important, (pause) le risque pour vous, la gestion des risques, est-ce que ça vous allume, est-ce que ça vous dit quelque chose le : le risque dans le cadre d'un projet d'innovation, par exemple, vous gérez les risques ?

Y.Y. Oui, oui, parce qu'on sait jamais quelle va être la réponse, il faut y aller progressivement (pause) sans chercher à vraiment tout mettre nos œufs dans ce panier-là parce qu'il faut se garder une autre porte à côté, de sortie. <OK> Si jamais c'est possible. C'est sûr que là, si je regarde la moule, la première expérience au début, bon (pause) le Québec il n'en voulait pas. (3sec) Un, on a commencé à (XXX), pis on est allé avec des petits volumes, pis on a même coupé les prix pour apprendre (pause) fait que là, on sait au moins, l'alternative qu'on a, on peut vendre la moule en coquille. <OK> Mais, par contre, (pause) le but est pas de vendre la moule en coquille, le but est de faire de la pasteurisation. Avec un produit fini, cuisson et pasteurisation pour débarquer et aller chercher une plus value à notre produit.

X.X. Par rapport à ce projet-là, si on parle donc, vous avez eu des : des problèmes techniques, c'est-à-dire au niveau de la production, mettre en production la moule, parce que vous faites l'élevage, c'est-à-dire la culture en mer, vous prévoyez la mettre, de la faire en deuxième transformation, la pasteurisation entre autres, c'est quoi pour vous les risques associés à ça ?

stopfin

X.X. Alors, je suis avec M. P de la compagnie M. Votre titre, M. L, c'est ?

Ph. L. Directeur de l'informatique.

X.X. Directeur de l'informatique. On va parler d'un projet euh : gros projet de chez M, c'est-à-dire la scie à deux axes qui est un projet qui représente une complexité technique, technologique importante, qui a été fabriqué à l'interne, c'est important de le dire. Euh : M. L, dans vos mots euh : dans vos propres mots, euh : le risque (pause) qu'est-ce que le risque pour vous euh : lié à un projet comme celui dont on va parler. De quelle façon vous percevez ça, est-ce que c'est important de le gérer ou euh : de quelle façon vous voyez les risques ?

Ph. L. Les risques euh : je les vois en termes de: euh : d'évaluer euh : la méthode qu'il faut prendre pour euh : solutionner les problèmes de : euh : les problèmes techniques qu'on

est capables d'envisager, (pause) d'être capables d'envisager la liste des problèmes qu'on : (pause) que ça peut euh : ça peut engendrer pis d'avoir plus de solutions.

X.X. Est-ce que c'est (pause) la méthode (pause), on va dire que c'est de la gestion des risques. (XXX) la méthode va être définie que c'est de la gestion des risques, parce qu'on en fait tous de la gestion de risques, je veux dire, même dans notre vie courante <On en fait tous>, on en fait tous, c'est euh :, chauffer un automobile c'est de la gestion des risques constante. Est-ce que vous sentez que dans le cadre de chez M c'est quelque chose qui est intégré dans les procédés de gestion, la gestion des risques est (XXX) systématiquement, on : on se…

Ph. L. C'est que ça se fait de façon non euh :, ça se faisait de façon non systématique (pause)et plutôt euh : chacun euh : chacun de son côté euh : (pause) avec ses réflexions, amenait ses réflexions pour que : euh : par la suite euh : on essaie de converger ça mais c'était un euh : un effort qui souvent euh : qui était euh : individuel.

X.X. OK. Au niveau de < Mais la euh : euh :, comment le dire...> Mais il en projection je crois là ? <Oui> Vous êtes en train d'installer des : des euh : des façons de faire ou des : de la planification, meilleure planification, pis euh : <Oui> Au niveau, on va tomber maintenant dans les problèmes de catégories de risque, est-ce qu'on sorte les risques techniques et au projet de la scie à deux axes au niveau technologique ou technique, les incertitudes liés à ce projet-là…

stopfin

X.X. Alors une entrevue euh : pour euh : qui concerne le développement d'un nouveau produit qui s'appelle un : thermos chauffant. On va commencer l'entrevue (pause) je voudrais connaître votre perception du risque. C'est quoi pour vous prendre un risque en affaires, ou : euh : comment vous percevez le risque.

Y.Y. Un risque euh :, ma perception de : <Dans un :> c'est simple là, ça dépend euh : ça dépend de ce que tu : veux (XXX) où tu veux en sortir, mais le risque c'est ce : euh : il y a tout le temps d'incidence mon affaire, en tout cas, ça peut-être de très grande importance, ça peut-être de moindre importance, pis : ça peut faire un succès ou un « flop » euh : monumental, mais regarde euh : ça dépend de ce que tu : veux mettre en marché ou ce que tu veux : (pause) ce dans quoi tu veux investir, c'est sûr, il y a une incidence monétaire toujours.

X.X. Est-ce que la première partie de votre définition ça vient de dire : plus c'est nouveau, plus il y a de risque ? Sur le marché en tout cas (pause) C'est radical, on dit, plus c'est nouveau, plus l'innovation est importante …

Y.Y. Euh : ça peut être nouveau pis :, pis t'as pas de risque (pause) t'as faite une : euh : une étude appropriée, pis même des fois, si t'as fait une étude appropriée, même si c'est nouveau (pause) (XXX) tu te dis : « Bon, ça va finir un succès monstre, ça fait un flop pareil ».Euh :, t'sais, je veux dire euh : c'est très difficile à évaluer <Pour euh :> et (XXX) pas de boule de cristal-là…

X.X. Oui, effectivement c'est dur à évaluer (pause) est-ce que euh : (pause) pour revenir au thermos chauffant, au niveau de euh : la faisabilité technique…

stopfin

X.X. Une entrevue avec euh : directeur général chez A euh : un projet Innovation Marketing (pause) qui concerne euh : la refonte des emballages dont il va nous parler. Vas-y N euh : ce que tu me disais là ?

N. X. C'est-à-dire que le projet était hautement stratégique parce que (3 sec) Ben : euh : c'est autrement un contexte (pause) nous autres à l'époque euh : où on prend la décision euh : (pause) on devait trouver ; une solution à une problématique qui était liée nous autres à l'augmentation de nos coûts (pause) euh : Nous autres, on travaille avec une matière première qui : qui : dont les prix varient à la semaine, donc c'est pas congelé, donc pis euh : les coûts de ce qu'on offrait ont augmenté de euh : quelque chose comme 60% en un an et puis euh : notre mix de produits était lourdement axé sur les produits S puis, étant donné qu'on n'avait pas de reconnaissance de : de euh : de la marque A et Frères dans le marché, pis on n'avait pas d'emballage attrayant, le monde veut pas nous différencier de nos compétiteurs dont on ne pouvait retransmettre pour les considérations du marché, les hausses des coûts qu'on subissait en amont de la production.<OK> Donc le choix nous autres qui a été fait, qui était un choix très, très laborieux mais stratégique, a été de dire : ben, nous allons travailler sur des aspects commerciaux, en occurrence l'aspect markéting, pour bien implanter la marque A et Frères pour que la marque soit reconnue comme étant une marque euh : une entreprise qui fait des produits de qualité, une entreprise qui est considérée comme leader de produit (pause) et que : les emballages traduisent la qualité des produits.

X.X. Donc, le projet d'innovation marketing, d'emballage a une saveur hautement commerciale, c'est-à-dire pour changer votre euh : (pause) vision de la mise en marché et votre façon de mettre en marché votre produit ? <Complètement.> C'est complètement ça ?

N.X. C'est complètement ça. Nous euh :, en fait, ce qu'on veut là-dedans c'est que :,(pause) que on :, la seule chose dont on est sûr c'est que la qualité la : la euh : la qualité des produits est (pause) on est convaincus que les produits sont d'une excellente qualité pis que tout le monde (XXX) les aime. Si on part de là, pis on veut se faire reconnaître nous dans nos marchés euh : dans nos marchés donc il faut non seulement qu'il y ait le bout de l'incidence de la marque, donc le relevé qui soit fait, il faut qui ait du packaging qui soit fait, il faut qu'il ait du merchandising qui soit arrêté <OK> donc on a investi dans ces aspects-là.

X.X. Donc, vous euh : le euh :, la façon la stratégie que vous avez adopté, c'est une stratégie de, qu'est-ce tu m'as expliqué tout à l'heure ?

N.X. De soph. <Soph> Donc, c'est un positionnement stratégique de soph qui fait en sorte que nous euh : on assume et on assume très bien le fait que : on a un produit qui est façonné artisanalement qui coûte plus cher à faire, mais qui est de meilleure qualité tant au niveau des qualités intrinsèques des produits qu'au niveau organoleptique, au niveau des goûts (pause) alors euh : nous ça nous (XXX) pas, pas en toute de travailler avec un positionnement de prix qui soit un peu plus élevé, mais encore faut-il que dans la tête du consommateur, que leur aperçu de la marque soit (pause) soit plus élevée, qu'il soit capable de mettre un :

X.X. Alors, nous, on met le doigt sur un facteur de risque qui est important, c'est-à-dire qu'il faut changer la perception du consommateur par rapport au produit <Tout à fait> Ça

c'est (pause) est-ce que tu considères que c'est un risque majeur par rapport à : (pause) un risque majeur, un risque <Oui> (XXX) du consommateur ?

N.X. Il y a de l'évangélisation à faire, donc euh :, c'est pour ça que là on est au stade du déploiement de la stratégie et puis on : on est à même de constater l'ampleur du travail de terrain qu'il y a à faire. (pause) Euh :, l'image fait pas tout, faut qu'il y ait du travail du terrain (XXX) comment faire des techniciens, des poissonniers, du détaillant direct pour qu'on puisse euh : pour qu'on puisse vraiment (XXX) nos efforts puis que : mettre le produit en bouche du monde, puis après ça, aussitôt, si on met le produit est en bouche, pis ils associent à un beau packaging, pis un : euh : un beau euh : <Une image de marque forte> Une image de marque forte, on : euh : on solidifie.

X.X. Une petite question pour nous mettre en contexte, pour toi le risque euh : qu'est-ce que ça représente ? Est-ce que c'est un langage qui est commun (pause) est-ce que c'est euh : gestion des risques ?

N.X. (XXX) un langage qui est commun, parce que dans le fond nous autres on est financés en fonction des risques, fait que : (rire) <OK> tout projet nous autres que : (pause) si on n'a pas de ressources même à l'interne de : de, des actionnaires même (pause) il faut aller les chercher (pause), il faut aller chercher d'aide financière.

X.X. Vous êtes financés en fonction de : euh : que vous démontrez que vous gérez vos risques ?

N.X. Non, en fonction du risque évalué du projet. <OK> Dans le fond, nous, on a un projet X donc euh : comme euh : on fait (pause) on va faire un exercice de vraiment pour planifier les coûts à moyen (pause) moyen et long terme de l'entreprise on parle de mois de septembre (pause) Ce plan d'affaires-là va comporter certains projets dans différents champs de : de de l'entreprise, dans les différentes fonctions, chacun des projets euh : (pause) est associé à un certain risque et donc euh : il y a un mix de financement qui va se faire en fonction de : des, des <OK> Et les risques sont (pause) sont évalués en collégialité non seulement à l'interne, mais avec des acteurs externes soit dans (pause) dans le financement par la dette (XXX) où c'est à peu près sans risque le schéma par prise de garantie ou le capital de risque et conventionnel.

X.X. Vous utilisez les deux ?

N.X. Les deux. <OK>

X.X. Pour en revenir à notre projet donc ce qu'on dit, c'est qu'on a une stratégie de euh : tu parles de soph qui implique une modification au niveau du consommateur ce qui implique évidement le projet d'innovation qui a été d'innover au niveau de votre packaging, de vos emballages. Au niveau (pause) allons-y au niveau technique.

stopfin

X.X. Alors, je suis avec M. D, votre nom de famille D.B de chez M, directeur de production, c'est bien ça ? On va parler pour une troisième entrevue avec cette compagnie de la : du projet de euh : de scie à deux axes, c'est bien ça ? ,uh-hum> Euh :, M. B, de quelle façon est-ce que vous voyez ça vous : le risque, vous euh : est-ce que c'est un élément de votre gestion lorsque vous prenez des décision ? Est-que vous vous rappelez constamment, est-ce que c'est un automatisme pour vous de faire de la gestion de risque ?

D.B. Calculer les risques ? Pas nécessairement. (Pause) Pas, euh : mais pas : (pause) on fait pas une grosse étude du risque, souvent euh : quand même souvent les projets qu'on a fait on a quand même une certaine expérience en arrière, pis euh : on n'a pas de : c'est pas un nouveau produit qui n'a pas de : (pause) pas d'antécédents, c'est pas un nouvel équipement qu'on n'a JAMAIS euh : qu'on n'a JAMAIS vu ou : t'sais comme : la scie, c'est quand même répétition d'une autre mais on a amélioré beaucoup <OK> donc (pause) le risque est quand même MOINS élevé parce qu'on a quand même une expérience.

X.X. L'expertise pour la scie. <Oui, c'est ça.> Pis donc pour euh : en me :,en gestion des risques, c'est pas systématique d'en faire vous-autres, parce que vous :..

D.B. Non, on n'a pas un système de gestion de risques ou euh :

X.X. C'est-tu parce que euh : il y a un manque (pause) on peut dire que : (pause) c'est pas qu'il y a un manque de ressources, mais parce que , souvent les PME ce qu'on voit, c'est qu'il y a un manque oui de ressources humaines, à l'interne, <Oui> ça c'est définitif <Oui> Est-ce qu'on peut dire que c'est une des causes <Oui:> c'est ce qu'on est pas capables de faire pis :< Oui, c'est sûr qu'on est > vous êtes …

D.B. Oui euh : (pause) la : eu : <XXX> c'est sûr qu'on saute pas mal à conclusion euh : (pause) qu'elles ont (pause) on fait quand même une bonne étude pour savoir ça vaux-tu la peine les objectifs qu'on veut atteindre dans le(XXX) (pause) dans le cas présent-là <OK, OK> Euh :, les principes qu'on utilise, c'est-tu des bonnes principes, la mécanique on a-tu amélioré <OK> t'sais, pour pas arriver à faire une grosse machine pis que : dans un an elle ne soit plus (pause) elle ne soit plus en fonction-là, être hors (pause). Alors ça, on a évalué, si on peut considérer ça comme des risques <OK> je pense que c'est quand même des risques tout ce qui est au niveau mécanique-là, c'est quand même (pause) pas trop euh : c'est toujours réparable-là, c'est moins pire que de créer un produit pis de euh : (XXX)

X.X. On va tomber euh : parce ce que ce que vous me parlez là, ce sont des risques techniques et technologiques..

stopfin

X.X. Alors (pause) c'est avec le PDG de chez euh : , en fait c'est PDG <VP exécutif> VP exécutif < En fait DG> chez Pre, on va discuter du même projet, c'est-à-dire le thermos chauffant. Monsieur le Directeur, je veux pas laisser le nom, vous comprenez <Oui> pour vous, le risque, euh : comment vous entrevoyez ça, le risque dans le cadre de votre euh : de vos (pause) de votre gestion, les affaires courantes, de quelle façon est-ce que vous : (pause) vous confrontez le risque ou : (pause) êtes-vous en mesure de : (pause) est-ce que vous vous êtes déjà attardé au risque dans ça, à penser qu'un risque en affaires peut : euh : ?

Y.Y. Ben nécessairement, toute entreprise qui est innovante doit faire face à l'élément de risque, évidemment ce risque-là faut qu'il soit évalué en fonction des données qu'il y a (pause) et toutes les entreprises ne sont pas nécessairement dotées de meilleurs outils et (pause) comme je te disais tantôt, souvent c'est instinctif, intuitif (pause) euh : par le bagage :, ou par l'éducation (XXX) et tout ça, et : euh : le risque est inhérent à toute entreprise et euh : comme je disais à toutes les entreprises ici à Rivière-du-Loup, c'est pas : (pause), en fait, c'est pas naturel , qu'une entreprise à Rivière-du-Loup, euh : (pause) que ses marchés soient à Montréal et Toronto, ce qui fait que : à Rivière-du-Loup (pause) Pr a

réussi à croître à Rivière-du-Loup, c'est parce qu'elle a innové, pis à ce moment-là (XXX) perdu les marchés. Fait que le fait (XXX) qu'on soit en région, a forcé l'entreprise à innover pis a euh : créer des nouveaux produits (pause) et euh : aujourd'hui Pr de Rivière-du-Loup est pas mal l'entreprise la plus (XXX) dans le verre au Canada, mais ça a coûté (pause) il y a des sommes investies là-dedans là euh :

X.X. Est-ce qu'on peut dire que si aujourd'hui vous auriez à relancer (pause) à mettre des investissements majeurs comme bâtisse pour le nouveau Pr aujourd'hui vous ne seriez pas à Rivière-du-Loup. <Il ne serait pas à Rivière-du-Loup, il ne serait pas à Rivière-du-Loup> Parce qu'il y a un risque important…

Y.Y. Parce qu'il a des coûts (pause) il y a des coûts intérieurs de ça quand ton marché est à Toronto, à Montréal pis t'es Gaspésien, il y a un coût important ça, écoute le bout Québec, Rivière-du-Loup-Montréal là, la compétition l'a pas là <OK> il y a un coût (pause) c'est ce que je disais aux) employés, il faut ; il faut maintenir notre position de leader pis de (pause) de trouver toujours de produits et il faut se démarquer et : c'est : là je (XXX) mais j'ai vite compris ça, écoute :, parce que : lorsque j'ai été recruté, je veux dire, que c'est une entreprise que : le siège social qui est à Rivière-du-Loup, que c'est qu'ils font là ? Remarque que, quand j'ai pris le poste j'ai compris tout ça, vraiment il y a eu une croissance interne qui s'est faite (pause) pis cette croissance-là s'est faite par le biais vraiment d'innovation et de produits et ou il y avait des risques effectivement (pause) le verre courbé (pause), il y avait des risques, ils ont pété de verre ici, ils l'ont pété en sacrement, jusqu'à un point que (pause) le président à l'époque disait : « Aye, hostie, on avait tout ça-là!» <uh-hum> et c'est (pause) ce risque-là qui a amené la : l'obtention du contrat de Pr Cadre qui est le plus important contrat ici <uh-hum. Contrat qui était fait auparavant en Europe-là, en (XXX) et en Finlande (pause) oui, des risques de prix, mais les gens y ont cru (pause) pis il avait un client aussi qui nous mettait de la pression pis qui nous garantissait évidemment euh : du volume (pause) alors le risque appartient (XXX) à la vie de Pr, et euh :, en fait, c'est pas ma décision, c'est inhérent, c'est :, c'est euh :, c'est dans la nature de l'entreprise euh :

X.X. Donc, l'entreprise est en mesure de faire face à une somme de risques importante (pause) donc, euh : est (XXX) pour faire face à : euh

Y.Y. Définitivement, elle est (XXX) pis elle est expérimentée aujourd'hui.

X.X. Parce que, comme vous dites, l'entreprise existe, parce qu'elle a innové constamment, donc, innover, c'est prendre un risque en soi.

Y.Y. Exactement, et (XXX) dans le marché (XXX). Pr est reconnue comme L'ENTREPRISE la plus innovante au niveau du verre.

X.X. Cette culture de risque-là, lorsque vous êtes arrivé en fonction, est-ce que vous euh : elle vous a été transférée ou vous étiez vous-même quelqu'un qui était comme : euh : avec une bonne capacité de prendre des risques (pause) ou vous avez appris la culture Pr de (XXX) et d'innovation.

Y.Y. En fait, à moins que je sois très naïf, si tu veux occuper un poste de direction générale ou bon (pause) de haute direction, il faut que tu sois capable de gérer les risques, sinon, écoute, tu n'es pas utile à l'entreprise. Une entreprise a une direction pour la faire avancer (pause) pis tu avances pas si tu prends pas de risques, alors euh :(pause) pour moi

c'est inhérent à la direction de l'entreprise. Il y a des gens qui en prennent MOINS y en a qui en prennent PLUS, mais euh : c'est toujours prendre des risques et c'est euh : c'est inhérent à : toute entreprise si tu veux (pause) tu veux y embarquer.

X.X. Pour revenir à notre projet de thermos chauffant <uh-hum> euh : selon vos : connaissances (pause) votre connaissance des dossiers et vos perceptions, au niveau technologique…

stopfin

X.X. On va avoir une entrevue avec euh : excuse, c'est quoi ton nom de famille déjà ?

Y.Y. R

X.X. R de l'entreprise Pa ??? On va parler d'un projet majeur d'innovation, innovation de procédé qui consiste en une refonte complète de la : ligne de production (pause) qui a augmenté euh : beaucoup leur capacité de production dont il va nous parler. Euh : ce qui (pause), une première question, je voudrais savoir pour toi (pause) pour toi c'est quoi le risque ? T'sais, dans tes mots à toi, qu'est-ce que c'est un risque ou prendre un risque ? Ça te dit quoi ça ?

Y.Y. Un risque (pause) ben euh : , dans (pause) dans toute euh : toute situation où tu dois prendre décision euh : le risque est là, que ce soit dans n'importe quel domaine que ce soit en affaires ou : (pause) ou : personnel, pis euh : le risque euh : à : ce moment-là je pense que : ce qui est bon dans la prise de décision, c'est euh : toujours de l'évaluer. <OK> Essayer de connaître tous les euh : les variables pis essayer de euh : de faire une évaluation pour euh : pouvoir prendre la euh : la meilleure décision possible.

X.X. Pis, dans ton euh : dans le contexte de : euh : de, toi, est-ce que (pause) nous, on fait une étude sur la gestion des risques <uh-hum> on veut connaître un peu les perceptions des entrepreneurs, donc, ce qui est par rapport au risque parce qu'on s'aperçoit que euh : c'est pas tous les entrepreneurs qui font systématiquement la gestion des risques, par exemple. <uh-hum> La question du risque, est-ce que c'est quelque chose que tu : (2 sec) systématiquement sera évalué les risques, on a, à la base, constamment ?

Y.Y. Constamment. Pourquoi euh : dans une négociation euh : avec les employés, avec un employé ou même avec les syndicats, ou : une décision d'affaires : un achat d'équipement, (pause) de : des décisions au niveau euh : des clients, des nouveaux produits euh :, d'ajustement des prix <OK> Le risque est toujours là, pis euh : j'essaye de toujours en prendre en : en considération.

X.X. OK. Maintenant on va tomber spécifiquement sur notre projet. Le projet de euh : d'innovation de procédé, c'est-à-dire de euh : de refonte de la ligne de production <Pas de problème>

stopfin

X.X. OK, entrevue avec euh : M. S D chez Ver un projet de euh : d'extrusion en stainless steel pour des verres de bateaux, c'est un tout nouveau produit. Euh :, on va parler des risques(pause) risques liés à ce projet-là. Euh : je voudrais juste avoir votre perception euh : à propos d'une définition du risque, pour vous, qu'est-ce que c'est un risque en affaires, ou le risque en général, de quelle façon vous voyez ça ?

S.D. B :, en fait, le risque on peut le (pause) moi, de ma façon de le voir, on peut le partager en plusieurs catégories, on a le risque financier, qui correspond à : on investit de l'argent et est-ce qu'on en va retirer un bénéfice, on a aussi le risque technique (pause) on investit de l'argent pour euh : développer une nouvelle technique, un nouveau produit (pause) est-ce qu'on va RÉUSSIR à le faire le nouveau produit (pause) donc, techniquement, est-ce qu'on est capable de le faire pis monétairement, est-ce qu'on peut en retirer quelque chose ? Moi, je vois le risque avec ces deux facettes-là.

X.X. OK, on va arrêter dans les catégories de : euh : de risque un peu plus spécifiques relatifs au projet d'extrusion en stainless steel. …

stopfin

X.X. Alors j'ai une entrevue avec un dirigeant de chez Pr qui concerne le projet de euh : thermos chauffant. (pause) Euh :, première question que j'aimerais savoir c'est quoi votre perception du risque euh : (pause) comment vous entrevoyez ça le risque : dans le cadre de euh : vos fonctions, est-ce que c'est quelque chose qui est systématique, que vous évaluez le risque dans un projet ? C'est quoi le risque pour vous ?

Y.Y. B, ça dépend de quel risque et la : (pause) au niveau de (XXX) de risque, c'est euh : il y a deux types de risque : il va y avoir un risque qui peut être euh : pour la sécurité des gens et l'autre risque est la durabilité à long terme. (pause) Comme, par exemple, à l'heure actuelle on travaille sur un projet important à l'aéroport de Montréal. Euh : (pause) il faut « design »-er des : des, des fenêtres très spéciales pour euh : la façade du : du building. (3sec) La préoccupation (pause) au niveau de ce risque-là c'est est-ce que ça va durer 5 ans, 6 ans, 10 ans de garantie, parce qu'il y a des coûts monétaires. Alors, ça c'est que : il faut que je gère le risque en fonction de la garantie, tandis que : on fait euh :, pour le même projet, on fait un plancher de verre. Alors là, il y a des gens qui vont se plier là-dessus (pause) et là, c'est la sécurité des gens. Alors, à ce niveau-là, on est encore beaucoup plus (pause) exigeants dans nos façons de faire. Ça veut dire qu'on a des étapes et une première à euh : comme euh : lorsqu'il y a une demande de faisabilité, ça va à R et D et euh : la R et D en fonction des informations qu'ils ont (pause) donnent leur aval euh : oui, c'est faisable ou non, c'est pas faisable. (pause) Si c'est faisable, alors on passe à l'étape supérieure et la : euh : finalement moi, je passe là-dessus ou Stéphane Fournier, un ingénieur qualifié, lui il va passer dessus. <OK> Pourquoi ? (pause) Parce qu'il y a des risques euh : pour la vie des gens, alors ça-là :, c'est euh : (3 sec) c'est euh :, c'est crucial.

X.X. OK. Donc, vous euh : (XXX) vous pratiquez une : gestion de risque euh :…

Y.Y. Oui, on n'a pas le choix, de toute façon, on a une : (5sec)

X.X. On va aller maintenant au projet de :, spécifiquement au projet du :, du verre chauffant, du thermos chauffant. <Oui> Ce projet-là, est-ce qu'il présentait des risques techniques, technologiques particuliers…

stopfin

X.X. Oui, une entrevue avec M. J P , PDG chez Mir. On va parler de l'innovation de procédé qui est une scie à deux axes. C'est un projet majeur pour l'entreprise. Donc, on est prêt à débuter. La première question, M. L, euh : comment vous percevez le risque, qu'est-ce qui est le risque pour vous dans le cadre de vos affaires ?

J-P. L. Le risque (toux) pour moi c'est (3 sec) c'est-à-dire que c'est : le : le risque c'est le montant euh : d'investissement , l'énergie, le temps, (pause) que : qu'on va consacrer à un projet innovateur (pause) TOUJOURS en relation avec les : les résultats que le projet va donner à euh : est-ce que : il va y avoir un retour sur l'investissement, est-ce qu'on : va euh : augmenter la productivité, est-ce qu'on : va augmenter la qualité euh : etc., etc. Donc euh : c'est le ratio ou le rapport un sur l'autre (pause) pour que : le projet soit euh : soit définitivement rentable, pour que le projet donne une euh : une augmentation de la capacité, une augmentation de la qualité, mais (pause) TOUJOURS dans le but d'augmenter la productivité. (pause) <OK> OK, c'est ça le euh : pour moi le risque.

X.X. OK. Donc, c'est très en relation avec l'atteinte de résultats, le risque de ne pas atteindre nos résultats, si on veut.

J-P. L. On peut réussir à : euh :t'sais on peut investir 1 000 $ pour euh : (pause) pour en bénéficier de 500 $, mais on peut investir 1 million si ça nous rapporte euh : 2 millions sur euh : 2 ans, (XXX) <OK> C'est toujours le rapport entre l'un et l'autre.

X.X. OK, alors, plus spécifiquement on va parler maintenant des euh : du projet de la scie à deux axes.

stopfin

X.X. Alors euh :, pour débuter la première question, euh : j'aimerais savoir pour vous le risque (pause) qu'est que (pause) en affaires ou euh : le risque général, c'est quoi pour vous le risque ?

Y.Y. (5sec) C'est presque de l'inconscience (pause) Je te dirai que : quand qu'on voit un projet intéressant, je pense qu'il faut euh :, on : (pause) on pèse le pour et le contre (pause) RAPIDEMENT, mais euh : il y a quelque chose en dedans de nous autres (pause) je crois une question d'instinct qu'on dit qu'on y va : (pause) on sait qu'il y a un risque, mais euh : (pause) on y va pareil, on se garrache, c'est intéressant, ça a l'air bon, je vois pas pourquoi qu'on ne marchera pas, fait que : c'est (pause) je te dirai que c'est presque de l'inconscience de toute façon, mais il faut peser les pros et les cons, on dirait qu'on finit TOUJOURS : par trouver des raisons pour pas le faire (rire) <OK> Fait que : non (pause) Pis euh : (pause), moi, je suis un individu qui marche beaucoup (pause), je suis pas fait cartésien (pause), je marche beaucoup par instinct et dans le temps j'ai appris à faire confiance à mon instinct, fait que : (pause) pour moi, c'est ça le risque.

X.X. OK. On va débuter avec le projet de euh :

- **Exemple 2 : Hassan II, le Maroc et l'histoire**

Blois, jeudi 18 décembre 2003

HASSAN II, Un Sultan au XXe siècle,

(Titre d'un article publié dans la revue L'Histoire, juin 2004)

Le roi Hassan II a régné sur le Maroc de 1961 à 1999. Ce règne controversé s'inscrit dans une tradition historique très ancienne, alors que la monarchie marocaine était confrontée à la nécessité d'une modernisation sociale et économique.

Cette confrontation entre tradition politique et modernité ne pouvait être que douloureuse.

Par PIERRE VERMEREN,

Historien, auteur du Maroc en transition, (2002, Poche, La Découverte).

Compte-rendu par ARMAND LASRY, professeur d'histoire géographie

Anne GUÉNÉGUÈS, coordinatrice des cafés historiques, présente le dernier café historique de l'année 2003. Elle inaugure une formule inédite avec une conférence débats Hassan II et le Maroc, présentée par Pierre VERMEREN, suivi par la projection du film Mille mois de Faouzi BENSAÏDI, primé au festival de Cannes, et programmé par l'association Cinéfil au cinéma les Lobis. Le titre du film, qui se rapporte par ailleurs au contexte sombre des émeutes de Casablanca en 1981, évoque une sourate du Coran en rapport avec la nuit sacrée du Ramadan, Leilat el Qadr (la Nuit du destin). Une famille se réfugie dans le Haut Atlas, milieu villageois soumis à l'arbitraire des autorités, suite à l'arrestation du père à Casablanca, faisant croire à son petit garçon qu'il est parti travailler en France.

Pierre VERMEREN a publié quatre ouvrages sur le Maroc, la Tunisie et le Maghreb, dont Maghreb, La démocratie impossible ? Fayard, avril 2004. P. Vermeren a vécu 7 ans au Maroc, notamment en 1987-1988, puis de 1996 à 2002. Il a donc vécu dans le Maroc de Hassan II (mort le 23 juillet 1999), en particulier le Maroc des années 80, celui des années de plomb, marquées par une vive tension politique, économique et sociale. Puis il a assisté à la progressive libéralisation des années 90, à l'alternance gouvernementale dès 1997, et à la succession dynastique en 1999.

Hassan II, le Maroc et l'histoire…

Malgré son long règne de 38 ans (1961-1999) et des liens très forts avec la France, Hassan II reste un personnage peu connu, mystérieux, qui suscite débats contradictoires et réactions passionnées.

Il n'existe pas encore en France de biographie de Hassan II, hormis le pamphlet de Gilles Perrault Notre ami le roi (1990), et les mémoires du Roi, dictées par Hassan II lui-même au journaliste E. Laurent. La personnalité de ce roi reste énigmatique, en France comme au Maroc, où pendant tout son règne, il était impossible de commenter les options et les décisions du roi.

La répression politique a été forte jusqu'au début des années 1990, en particulier dans les années 1975-1990 (dites « années de plomb »). On estime de 50 à 60 000 le nombre de victimes durant son règne (disparus, prisonniers politiques, tués…). Cela constitue un puissant traumatisme dans la société marocaine. Le règne a notamment été marqué par de grandes émeutes urbaines à Casablanca (1965, 1981), Oujda (1984) et Fès (1990), mais il n'y a pas eu de grande révolte populaire nationale.

Suite à la chute du « Mur de Berlin » et à la fin de la Guerre Froide, le roi Hassan II réoriente son pays dans une optique de libéralisation, mettant fin en quelques années à la répression politique, et préparant l'alternance gouvernementale de 1997-98. Le pays s'ouvre progressivement (via le tourisme, les médias par satellites…), dans un contexte international bouleversé, annonçant la transition des années 1997-2000.

Hassan II a eu une très forte personnalité, et il a su faire compter son pays sur la scène internationale (notamment par sa position de médiateur dans le conflit israélo-arabe). À l'intérieur, il a dû s'adapter et composer avec les forces en présence, son armée, ses élites et la classe politique (qui lui a longtemps contesté son monopole politique). La religion est un des moyens qu'il utilise pour gouverner, ayant à la fois une solide formation coranique, et étant par sa fonction le Commandeur (émir) des croyants (et descendant du Prophète ou chérif). Les oulémas (docteurs de la foi), traditionnels conseilleurs et électeurs du Sultan (devenu Roi à l'indépendance), sont mis de côté, le Roi s'accaparant l'essentiel du pouvoir religieux. L'institution sultanienne est alors devenue une monarchie héréditaire.

L'histoire du Maroc, telle qu'enseignée dans les lycées et à l'Université, s'arrêtait officieusement en 1912 (date du protectorat de la France sur le Maroc), afin d'occulter ce que l'histoire officielle ne voulait pas retenir. Or le Maroc a connu une histoire difficile et troublée jusqu'aux années 1960-61, ses élites politiques et économiques ayant beaucoup profité du Protectorat, et plus encore de l'après indépendance et du maintien de liens très forts avec l'ancienne Métropole, ce sur quoi les acteurs refusent le plus souvent de réfléchir.

Après la mort du roi en juillet 1999, cette histoire occultée a commencé à apparaître dans le débat public, dans le contexte nouveau d'une effervescence de publications (journaux, mémoires, témoignages...). En France aussi, des ouvrages importants sont parus, écrits notamment par des journalistes du Monde, Oufkir de S. Smith et Le dernier roi de J.-P. Tuquoi (qui, malgré son titre, évoque pour l'essentiel la vie privée de Hassan II).

Plusieurs témoignages dramatiques reviennent sur la violence et l'arbitraire des « années Oufkir » (quand il dirigeait l'Intérieur jusqu'à sa mort, après l'échec du coup d'État de 1972), et de la répression qui a frappé les opposants, notamment des témoignages écrits par les enfants du Général Oufkir. Des proches du Roi, des prisonniers politiques ont aussi publié des témoignages. Mais le livre le plus réussi revient à Ahmed Merzouki, officier berbère qui a raconté le coup d'Etat de Skirat de 1971, qui l'a conduit pour près de 20 ans dans le bagne mouroir de Tazmamart. Le livre Tazmamart, cellule 10, fut un succès de librairie sans précédant au Maroc, avec un tirage record de 35 000 exemplaires en 2000-2001 !

Le journaliste français de l'AFP Ignace Dale prépare une biographie d'Hassan II à partir de nombreux témoignages de ses proches qui doit paraître prochainement chez Fayard.

Hassan II, un personnage de roman...

Le Prince Moulay Hassan est né en 1927. Son père Mohammed V est autoritaire et fait élever durement ses enfants. Mohammed V s'impose peu à peu comme le Père de l'indépendance marocaine, réussissant à apparaître comme le Sultan de l'Istiqlâl (l'indépendance), le grand parti nationaliste marocain créé en janvier 1943. Mohammed V a été élevé comme un sultan médiéval, au Palais de Fès. En revanche, son fils, selon la volonté de son père, reçoit une double formation, à la fois traditionnelle, dont les bases sont l'Arabe et l'Islam, et française. Son père veut faire de lui une personnalité moderne. C'est pourquoi Hassan II a fait ses études au collège impérial de Rabat créé pour lui, puis il fait son droit à Bordeaux (1948-1951). Il a de grands professeurs de droit comme Maurice Duverger.

Il mène alors une vie d'étudiant aisé, voyageant beaucoup (il était déjà souvent venu avec son père en France), fréquente les grands couturiers parisiens, se baigne à Royan, La Baule. C'est un séducteur qui aime les belles femmes, les voyages et la fête.

Après avoir subi l'exil avec son père en 1953-55, il rentre au Maroc et devient le créateur et dirigeant des FAR (Forces armes royales), embryon de l'armée marocaine, très marqué par l'armée coloniale française dont elle est d'abord issue. C'est alors que le régime s'affirme et parvient à écarter tous les compétiteurs de la monarchie du pouvoir. le Prince héritier se montre très actif.

Puis Moulay Hassan devient Hassan II en 1961. Les années 60 apparaissent comme de « belles années » pour le régime qui s'est consolidé. La croissance économique du pays est honnête, mais les atteintes aux droits de l'Homme apparaissent. Ben Barka est assassiné en 1965, et l'opposition se heurte à un mur.

En 1971 et 1972, deux coups d'État militaires mettent le régime en péril. La situation devient dangereuse pour la pérennité de la Monarchie. Hassan II réagit alors très vigoureusement, et prépare un tournant entre 1973-1975. Après l'exécution d'Oufkir, les biens coloniaux sont nationalisés et redistribués à la bourgeoisie et aux officiers fidèles. C'est alors qu'Hassan II conçoit puis organise la « Marche Verte » (1975), à l'issue de laquelle le Maroc s'empare du Sahara espagnol, 200 000 km2 de terres très convoités. Le Maroc vit alors au rythme de l'Union sacrée, et il est strictement interdit de s'opposer au consensus saharien, sous peine des pires maux. En échange, le Roi s'engage à préparer une alternance politique... mais il faudra attendre plus de vingt ans.

Hassan II devient le « réunificateur » comme son père a été le « libérateur ». Mais l'Algérie refuse ce fait accompli et déclenche une guerre qui remet en cause l'unité du Maghreb et aggrave les tensions entre les deux pays. Il se sépare ensuite de son ministre de l'Intérieur Dlimi (exécuté en 1985 ?). Driss Basri, Ministre de l'Intérieur depuis 1979, s'impose, et exerce un ascendant croissant jusqu'à la fin du règne. Il est le numéro deux du régime, l'homme à tout faire qui veille sur la sécurité du Roi et la stabilité du pays, quel qu'en soit le prix. Le pays est gouverné d'une main de fer, en dépit de l'amorce de libéralisation dès 1990-91.

Après l'exposé du Professeur P. Vermeren, le public de la Brasserie de la Halle pose les questions suivantes :

Les deux coups d'État ont-il été le point de bascule du régime ?

Les coups d'État de 1971 et 1972 sont des points d'aboutissement plus que de bascule.

Un acte fondateur : Moulay Hassan, chef d'État major des FAR, réprime la révolte du Rif en 58-59, sous Mohammed V. Oufkir s'est alors affirmé, les armes à la main, comme l'homme de confiance du régime. La période du Rif a créé chez Hassan II la conscience d'un risque qu'il fallait mater pour la survie du régime. Dans les années soixante, le régime s'appuie paradoxalement sur les officiers berbères, issus de l'armée française, alors qu'il vient d'écraser la révolte berbère du Rif, pour neutraliser la bourgeoisie urbaine politisée. Tous le contre-pouvoirs sont neutralisés. Les coups d'état marquent la fin de cette alliance entre le Trône et les officiers berbères coloniaux. Le Roi doit trouver d'autres soutiens.

Comment expliquer la violence politique dans un pays où le pouvoir et l'État apparaissent comme historiquement légitimes ?

Il y a les émeutes de Casablanca en 1965, 1981... Mais une politique de force s'est imposée dès l'indépendance dans l'intérêt de la survie du régime et de l'unité du pays. Car jamais le pays, malgré l'ancienneté du Sultanat, n'a été unifié aussi certainement qu'à l'époque du Protectorat (il a fallu 22 ans de guerre pour soumettre toutes les tribus). À l'indépendance, nombre de tribus, partis et régions pensent recouvrir leur autonomie..., comme avant. La violence a donc précédé les émeutes urbaines (dues pour l'essentiel à la misère de l'exode rural) et les coups d'État. Au point que pour certains observateurs, les années de plomb sont étendues à tout le règne.

Quelle est la situation actuelle Pour le « Sahara occidental » ?

La position du Maroc sur cette question est marquée par une très grande fermeté. Il estime que le Sahara occidental fait partie intégrante de son territoire. L'Algérie n'accepte pas ce qu'elle considère comme des prétentions du Maroc, et elle soutient à ce titre le Front Polisario. Aujourd'hui, on est dans une situation de ni guerre ni paix, au grand dam de l'ONU. Cette question est un obstacle à la création d'un Maghreb uni. La réconciliation ne s'est pas faite encore.

Quelles sont les oppositions rencontrées par le pouvoir sur cette question du Sahara ?

Très peu. D'une part, on ne peut pas dire que l'intelligentsia berbère se serait opposée à Hassan II sur ce terrain (alors que les Sahraouis marocains sont arabes, contrairement aux touaregs berbères). Les officiers et l'armée (largement berbères) y ont au contraire trouvé un terrain d'expansion... et des primes. L'opposition à la politique du Sahara marocain a donc été très politisée à l'extrême gauche, et à ce titre minoritaire dans le pays. Et malgré le contexte de guerre dans les médias depuis 25 ans, la population se sait très proche du peuple algérien (en dépit de vieilles animosités). Mais ce sont les pouvoirs qui mènent la politique et les relations entre les deux pays.

Pour le peuple marocain, le problème majeur consiste à remplir le panier de la ménagère... Or avec la crise économique, la misère, le chômage et la pénurie de logements, tout cela est très difficile. L'opposition prend un appui religieux plus que politique. L'islamisme, c'est la rencontre entre un désarroi culturel et moral, et une pauvreté économique sans perspectives d'ascension. Mais parce que la guerre a coûté cher au Maroc (il a fallu entretenir une armée de 200 000 hommes à 1 500 km de Rabat), la population reste persuadée du bon droit marocain.

Quel a été le rôle du colonisateur vis-à-vis du Sahara ?

L'Espagne et la France ont joué un grand rôle. Le partage inégal du Sahara entre le Maroc et l'Algérie a été fait par la France quand elle pensait garder l'Algérie. La France a aussi créé la Mauritanie, peut-être à cause de la bauxite et du fer. C'est un pays qui n'a pas d'histoire précoloniale. Le partage du Sahara a été fait de façon arbitraire. Autrefois, il n'y avait pas de frontière du Sahara d'où la complexité de la question. Aussi, la réalité d'un Sahara sahraoui, espagnol marocain, algérien ou touareg est très relative et contingente. C'est une histoire post-coloniale. Ce quine la rend que plus complexe...

- **Exemple 3 : Définition d'informatique**

Aller à : Navigation, Rechercher L'informatique (information automatique) désigne l'automatisation du traitement de l'information par un système, concret (machine) ou

abstrait (on peut parler d'automate). Dans son acception courante, l'informatique désigne l'ensemble des sciences et techniques en rapport avec le traitement de l'information. Dans le parler populaire, l'informatique peut aussi désigner à tort ce qui se rapporte au matériel informatique (l'électronique), et la bureautique.

À ce sujet on attribue une phrase à Edsger Dijkstra qui résume assez bien cela :

« L'informatique n'est pas plus la science des ordinateurs que l'astronomie n'est celle des télescopes.

(En anglais : Computer science is no more about computers than astronomy is about telescopes.) »

La discipline scientifique désignée par le terme informatique fait partie des sciences formelles comme les mathématiques ou la logique. Aujourd'hui, la distinction entre ces trois disciplines est floue, mais l'on peut identifier l'informatique à travers les principales questions abordées :

qu'est-ce que le calcul ?

que peut-on calculer ?

comment calculer efficacement ?

comment décrire un algorithme de calcul ?

comment représenter un certain objet pour pouvoir le traiter?

On peut trouver des racines à la science informatique dans de nombreux domaines anciens des mathématiques (systèmes de numération, division euclidienne, construction à la règle et au compas, etc.). Cependant, la discipline n'a émergé qu'à partir des années 1930 à travers une série de travaux fondateurs[1][2][3] (Church, Gödel, Herbrand, Kleene,Turing) qui ont abouti à la première formalisation générale de ce qu'est le calcul. La force de cette formalisation est de faire converger plusieurs points de vue vers une même et unique notion :

Un point de vue mécanique avec les machines de Turing universelles qui constituent un véritable modèle d'ordinateur (améliorable en termes d'efficacité, mais toujours valable aujourd'hui en termes de capacités),

Un point de vue fonctionnel avec le lambda calcul qui a donné naissance à de nombreux langages de programmation,

Un point de vue arithmétique avec les fonctions récursives qui a notamment permis de relier la notion de calcul au raisonnement mathématique (voir à ce sujet le théorème d'incomplétude de Gödel).

Ces trois points de vue donnent un sens très général à la notion de calcul et la thèse aujourd'hui largement acceptée est qu'elle capture tout traitement réalisable mécaniquement.

Le second acte de naissance de l'informatique est bien entendu la réalisation concrète des premiers ordinateurs dans les années 1940, puis le développement de leur fabrication avec l'avènement de l'électronique numérique pour en faire aujourd'hui un domaine

technologique à part entière. Il s'agit bien d'un second acte de naissance car, si les capacités intrinsèques des ordinateurs actuels sont les mêmes que celle du modèle théorique des années 30, leur rapidité de traitement, leur coût (financier, mais aussi en termes de place, de ressources nécessaires, etc.), leur longévité et leur fiabilité ont été considérablement améliorés, ouvrant un vaste champ de possibilités auparavant impensables.

Des professions aussi diverses que concepteur, analyste, développeur, responsable d'exploitation, ingénieur système, technicien de maintenance matérielle ou logicielle, chercheur en informatique ou directeur d'un centre de calcul, relèvent du domaine de l'informatique. Néanmoins, le terme informaticien désigne le plus souvent ceux qui conçoivent, déploient et mettent en œuvre des solutions.

Sommaire [masquer]

Terminologie [modifier]

Origine [modifier]

Voir « informatique » sur le Wiktionnaire.

Le terme allemand Informatik est créé en 1957 par Karl Steinbuch qui a publié un essai intitulé Informatik: Automatische Informationsverarbeitung (Informatique : traitement automatique de l'information)[4].

Le terme informatique est utilisé pour la première fois en France en mars 1962 par Philippe Dreyfus, ancien directeur du Centre National de Calcul Électronique de Bull dans les années 1950, qui, en 1962, a utilisé pour la première fois ce terme dans la désignation de son entreprise « Société d'Informatique Appliquée » (SIA). à partir des mots « information » et « automatique »[5].

En France, l'usage officiel du mot a été consacré par Charles de Gaulle qui, en Conseil des ministres, a tranché entre « informatique » et « ordinatique », et le mot fut choisi par l'Académie française en 1967 pour désigner cette nouvelle discipline. En juillet 1968, le ministre fédéral de la Recherche scientifique d'Allemagne, Gerhard Stoltenberg, prononça le mot Informatik lors d'un discours officiel au sujet de la nécessité d'enseigner cette nouvelle discipline dans les universités de son pays, et c'est ce mot qui servit aussitôt à nommer certains cours dans les universités allemandes [réf. nécessaire]. Le mot informatica fit alors son apparition en Italie et en Espagne, de même qu'informatics au Royaume-Uni.

Pendant le même mois de mars 1962 Walter F. Bauer inaugura la société américaine Informatics Inc. qui, elle, déposa son nom et poursuivit toutes les universités qui utilisèrent ce nom pour décrire la nouvelle discipline, les forçant à se rabattre sur computer science, bien que les diplômés qu'elles formaient fussent pour la plupart des praticiens de l'informatique plutôt que des scientifiques au sens propre. L'Association for Computing Machinery, la plus grande association d'informaticiens au monde, approcha même Informatics Inc. afin de pouvoir utiliser le mot informatics pour remplacer l'expression computer machinery, mais l'entreprise déclina l'offre. La société Informatics Inc. cessa ses activités en 1985, achetée par Sterling Software [réf. nécessaire].

Évolution récente

L'évolution récente tend à employer plutôt l'expression STIC en français, pour sciences et technologies de l'information et de la communication.

Le mot communication tend à donner une importance excessive aux échanges et aux accès, par rapport aux contenus des bases de données de connaissances, dans une optique de gestion de connaissances (knowledge management).

C'est la raison pour laquelle certains experts, comme Bernard Besson, préfèrent remplacer TIC par l'expression TICC, pour technologies de l'information, de la communication et de la connaissance.

Terminologie anglo-saxonne

La traduction anglaise de informatique est computer science, littéralement « science du calculateur ». En français, l'expression science du calcul (computing science) fait plutôt penser à informatique scientifique.

En anglais les termes distincts suivants sont utilisés :

Informatics (science de l'information) : Ce qui ressort de l'étude des systèmes, biologiques ou artificiels, qui enregistrent, traitent et communiquent l'information. Ceci comprend l'étude des systèmes neuraux, aussi bien que les systèmes informatiques.

Computer science (l'Informatique théorique) : Ce qui ressort de l'épistémologie procédurale, soit notamment de l'étude des algorithmes, et donc indirectement des logiciels et des ordinateurs.

Computer engineering (Génie informatique) : ce qui ressort de la fabrication et de l'utilisation du matériel informatique.

Software engineering (Génie logiciel) : Ce qui ressort de la modélisation et du développement des logiciels ; ceci comprend deux aspects : les données et les traitements ; les deux aspects sont liés dans la mise en pratique des traitements de données (Data Processing). En France, en pratique, l'expression ingénierie informatique correspond plutôt à software engineering, soit l'ingénierie logiciel.

Information technology engineering (Génie des technologies de l'information) : ce qui ressort de l'intégration des techniques et des technologies relatif à l'information et reliées à l'informatique ainsi qu'à l'internet (par exemple : le e-business)

Information Technology (Technologies de l'information) : Représente l'évolution des techniques et des technologies reliées à l'informatique.

Il existe plusieurs termes anglais pour désigner le concept d'« informatique ». Certains comme automatic data processing ou electronic data processing et leur abréviation reflètent une vision plus ancienne et ne sont plus guère utilisés. Même data processing est parfois considéré par certains informaticiens professionnels comme propre à la langue des administrateurs et des non informaticiens (dans le jargon du métier, costards ou, en anglais, suits). Quant à informatics, il est davantage employé en Europe, selon certaines sources.[6]

On trouve d'autres variantes peu attestées; c'est le cas de computing science, electronical data processing, ordinatique, technologie des ordinateurs ou science de l'informatique.

Il faut dire que les concepts et la terminologie ont suivi l'évolution de la réalité. Ainsi, les ordinateurs, qui effectuaient autrefois des opérations relativement simples de calcul sur des données, traitent de façon de plus en plus complexe, aujourd'hui, de l'information autrement plus significative (connaissances et savoir-faire). De la désignation informatique, on est passé peu à peu à celle de technologies de l'information. On voit poindre, dans certains milieux, des appellations comme technologies ou nouvelles technologies de l'information et de la communication qui cèderont peut-être leur place à une autre dénomination qui reflètera le traitement des connaissances, des savoir-faire et même de « l'intelligence ». Progressivement le terme informatique glisse vers un sens plus restreint relié aux aspects techniques.

Article détaillé : Histoire de l'informatique.

Les origines

Depuis des millénaires, l'Homme a créé et utilisé des outils l'aidant à calculer (abaque, boulier, etc.). Parmi les algorithmes les plus anciens, on compte des tables datant de l'époque d'Hammurabi (env. -1750). Les premières machines mécaniques apparaissent entre le XVIIe et le XVIIIe siècle. La première machine à calculer mécanique réalisant les quatre opérations aurait été celle de Wilhelm Schickard au XVIe siècle, mise au point notamment pour aider Kepler à établir les tables rudolphines d'astronomie.

En 1642, Blaise Pascal réalisa également une machine à calculer mécanique qui fut pour sa part commercialisée et dont neuf exemplaires existent dans des musées comme celui des Arts et métiers et dans des collections privées (IBM).

La découverte tardive de la machine d'Anticythère montre que les Grecs de l'Antiquité eux-mêmes avaient commencé à réaliser des mécanismes de calcul en dépit de leur réputation de mépris général pour la technique (démentie d'ailleurs par les travaux d'Archimède).

Cependant, il faudra attendre la définition du concept de programmation (illustrée en premier par Joseph Marie Jacquard avec ses métiers à tisser à cartes perforées, suivi de Boole et Ada Lovelace pour ce qui est d'une théorie de la programmation des opérations mathématiques) pour disposer d'une base permettant d'enchaîner des opérations élémentaires de manière automatique.

La mécanographie

Une autre phase importante fut celle de la mécanographie, avec l'apparition des machines électromécaniques alimentées par cartes perforées de l'Allemand Hollerith, à la fin du XIXe siècle. Elles furent utilisées à grande échelle pour la première fois par les Américains lors du recensement de 1890 aux États-Unis, suite à l'afflux des immigrants dans ce pays lors de la seconde moitié du XIXe siècle. Les Allemands étaient probablement bien équipés en machines mécanographiques avant la Seconde Guerre mondiale. Ces équipements, installés par ateliers composés de trieuses, interclasseuses, perforatrices, tabulatrices et calculatrices connectées à des perforateurs de cartes ont dû leur apporter une certaine supériorité pour la construction des armements. Toutefois, ceci n'a pas été examiné en profondeur par les historiens. Leur moindre mérite n'est pas la réussite du programme. On ne pouvait pas encore parler d'informatique, car les traitements étaient exécutés à partir de techniques électromécaniques et basés sur l'usage de lampes radio ; anodes, cathodes, triodes etc. La chaleur dégagée par ces lampes rendait ces ensembles peu fiables.

Science des nombres et Système de numération [modifier]

Examen de quelques systèmes numériques de jadis à nos jours.

Du système binaire inventé par G. Boole, utilisé par les ordinateurs au dénombrement restreint du 1-2-3 pour dire beaucoup à partir de 4 en usage chez les primitifs, les hommes ont utilisé des systèmes numériques qui indiquent le degré de culture ou de vigilance des peuples auprès desquels ils étaient en application.

Les Sumériens utilisaient le système sexagésimal encore utilisé de nos jours pour mesurer l'heure qui compte 60 minutes divisées en 60 secondes. Il fut repris par les Grecs pour leurs calculs astronomiques, dans le calcul des angles et du temps.

Les Romains se servaient de leurs 10 doigts et pratiquaient le système décimal pour constituer des centuries de légionnaires. Ils marquaient les milliers par un cercle barré verticalement. Déformé ce signe a donné le « M » pour désigner 1 000 et la moitié de ce symbole pour le « D » pour désigner 500. Le système décimal en application chez les Romains sans la connaissance des chiffres arabes, ne facilitait pas la tâche arithmétique des intendants chargés de faire les comptes.

Les Celtes pour leur part allaient jusqu'à utiliser en plus les dix doigts de pieds, ce qui élargissait leur système numérique à 20. Les derniers Celtes sur le continent, de nos jours, utilisent encore ce système pour apprécier toutes les valeurs quantitatives de la vie courante.

Lors d'un safari poils, plumes et aux phacochères au Sénégal, on peut découvrir le système quincal utilisé par les Sérères. Pourquoi émanant des Sérères ? Parce que les Sérères incarnent la tribu de chasseurs au Sénégal. La population sénégalaise comprend quatre ethnies principales. En plus des Sérères, on y rencontre des Wolofs l'ethnie dominante, leur langue étant reconnue langue nationale et le français fait office de langue officielle. Les Toucouleurs et les Peuls forment les deux autres ethnies. Aux Sérères incombaient traditionnellement le rôle de pourvoyeurs de gibier à l'égard des autres ethnies, elles de culture agricole et pastorale. Par hordes, les Sérères effectuaient des déplacements en chassant de village en village armé de courts bâtons, d'arcs et de lances. Le gibier abattu ; pintades, perdreaux et pigeons, l'étaient par ce bâton qu'ils lançaient en virtuose dès que le gibier traqué prenait son envol dans la savane. Phacochères et autres bêtes à sabots étaient chassés à l'arc, à la lance et achevés à la sagaie ou à l'épieu par les porteurs et les traqueurs. Le système numérique originel des Sérères est quincal. Il est aisé de deviner le pourquoi de cette pratique. Dans l'exécution de leur activité, l'une des mains seulement était disponible pour des occupations annexes. L'autre restait à serrer toujours une arme, le bâton, la sagaie, la lance... pour compter. Den - niet - nient - njar - gurun ; a cinq se fait le report : gurun-den etc. Le produit de la chasse était troqué contre des produits agricoles, textiles, en somme contre des produits de première nécessité pour ces chasseurs ambulants qui cultivaient le souci de se faire respecter. Les bijoux et la céramique occupaient une place de choix dans le troc, ces dames occupent encore et toujours lors de rencontres de tout genre le premier rang pour ne pas figurer parmi les laissés-pour-compte.

Les Juifs pratiquaient le système numéral le plus élaboré. En exploitant les 12 phalanges des huit doigts décomptés à l'aide des pouces. Ils ont institué la grosse, une unité de mesure encore en vigueur pour certaines marchandises. On y arrive en se servant par exemple du pouce de la main gauche pour décompter les phalanges. Arrivé à la dernière phalange du petit doigt, 4 × 3, permettent d'enregistrer la première douzaine à l'aide du pouce de la main droite qui pointe sur la première phalange de l'index de la main droite. Ce système permet ainsi de compter jusqu'à 144.

Des tablettes d'argile servaient de document aux scribes, assis sur les quais de débarquements pour les prises en charge des arrivages ou dans les entrepôts lors de la sortie ou de l'entrée des marchandises. Le Romain marquait d'un trait à 10, le Celte à 20 et le juif à 144 arrivé à chaque fin de son système numéral pour enregistrer jusqu'à la dernière pièce les mouvements à enregistrer.

George Boole, mathématicien anglais (1815-1864), fut l'inventeur du système binaire. Sans son système, il n'y aurait pas d'ordinateurs transistorisés qui fonctionnent grâce à des 0 et des 1, qui permettent d'aller en calculs à l'infini.

L'informatique moderne [modifier]

L'ère des ordinateurs modernes commença avec les développements de l'électronique pendant la Seconde Guerre mondiale, ouvrant la porte à la réalisation concrète de machines opérationnelles. Au même moment, le mathématicien Alan Turing théorise le premier ce qu'est un ordinateur, avec son concept de machine universelle de Turing.

L'informatique est donc un domaine fraîchement développé, même s'il trouve ses origines dans l'antiquité (avec la cryptographie) ou dans la machine à calculer de Blaise Pascal, au XVIIe siècle. Ce n'est qu'à la fin de la Seconde Guerre mondiale qu'elle a été reconnue comme une discipline à part entière et a développé des méthodes, puis une méthodologie qui lui étaient propres.

Son image a été pendant quelque temps surfaite : parce que les premiers à programmer des ordinateurs avaient été des ingénieurs rompus à la technique des équations différentielles (les premiers ordinateurs, scientifiques, étaient beaucoup utilisés à cette fin), des programmeurs sans formation particulière, parfois d'ailleurs issus de la mécanographie, cherchaient volontiers à bénéficier eux aussi de ce label de rocket scientist afin de justifier des salaires rendus confortables par :

le prix élevé des ordinateurs de l'époque (se chiffrant en ce qui serait des dizaines de millions d'euros aujourd'hui compte tenu de l'inflation, il reléguait au second plan les considérations de parcimonie sur les salaires) ;

l'aspect présenté comme peu accessible de leur discipline et un mythe de difficulté mathématique entretenu autour. En fait, les premiers ordinateurs ne se programmaient pas de façon très différente de celle des calculatrices programmables utilisées aujourd'hui dans les lycées et collèges, et maîtrisées par des élèves de quatorze ans mais le domaine était nouveau et l'algorithmique nécessite un certain degré de concentration associé, peut-être à tort, à la réflexion pure.

L'émergence d'un aspect réellement scientifique dans la programmation elle-même (et non dans les seules applications scientifiques que l'on programme) ne se manifeste qu'avec la série The Art of Computer Programming de Donald Knuth, professeur à l'Université de Stanford, à la fin des années 1960, travail monumental encore inachevé en 2004. Les travaux d'Edsger Dijkstra, Niklaus Wirth et Christopher Strachey procèdent d'une approche également très systématique et elle aussi quantifiée.

On demandait à Donald Knuth dans les années 1980 s'il valait mieux selon lui rattacher l'informatique (computer science) au génie électrique — ce qui est souvent le cas dans les universités américaines — ou à un département de mathématiques. Il répondit : « Je la classerais volontiers entre la plomberie et le dépannage automobile » pour souligner le côté encore artisanal de cette jeune science.

Toutefois, la forte scientificité des trois premiers volumes de son encyclopédie suggère qu'il s'agit là plutôt d'une boutade de sa part. Au demeurant, la maîtrise de langages comme Haskell, Ocaml ou même APL demande un niveau d'abstraction tout de même plus proche de celui des mathématiques que des deux disciplines citées.

La miniaturisation des composants et la réduction des coûts de production, associées à un besoin de plus en plus pressant de traitement des informations de toutes sortes (scientifiques, financières, commerciales, etc.) a entraîné une diffusion de l'informatique dans toutes les couches de l'économie comme de la vie de tous les jours.

En France, l'informatique a commencé à vraiment se développer seulement dans les années 1960, avec le Plan Calcul. Depuis lors, les gouvernements successifs ont mené des politiques diverses en faveur de la Recherche scientifique, l'Enseignement, la tutelle des Télécommunications, la nationalisation d'entreprises clés.

La science informatique

Article détaillé : Informatique théorique.

Technologies de l'information et de la communication

Domaines d'application de l'informatique

Le traitement de l'information s'appliquant à tous les domaines d'activité, on pourra les trouver associés au mot informatique. Ainsi on pourra parler d'informatique médicale quand ces outils sont utilisés par exemple dans l'aide au diagnostic, et ce champ d'activité se rapportera plutôt à l'informatique scientifique décrit ci-dessous; ou bien on parlera d'informatique bancaire; il s'agira alors soit des systèmes d'information bancaire qui relèvent plutôt de l'informatique de gestion, de la conception et de l'implantation de produits financiers qui relève plutôt de l'informatique scientifique et des mathématiques, ou encore de l'automatisation des salles de marché qui en partie relève de l'informatique temps réel. On peut schématiquement distinguer les grands différents types suivants :

L'informatique de gestion : elle consiste à piloter les processus de gestion et de management dans les entreprises, dans tous les domaines d'activité : payes (employés, ouvriers, cadres) et gestion des ressources humaines, administration des ventes, des achats (déclaration de TVA) gestion de la relation client, gestion de la production et des approvisionnements, tenue de stocks ; des entrepôts de produits usinés, des en cours de fabrication, de l'inventaire permanent et des inventaires de fin d'exercice, carnet de commandes, marketing, finances... Ce domaine est de loin celui qui représente la plus forte activité, ce qui n'a pas toujours été perçu en France.

Jusqu'en 1965, la mécanographie, et par la suite la simple mécanisation de la mécanographie connue sous le vocable « informatique fiabilisée par la transistorisation », savait faire tous ce qui est énuméré ci-dessus, sauf de la comptabilité générale avec suivi des créances innové par Le lettrage conversationnel 1962 en Grandes entreprises.

Gilbert Bitsch, chef de projets à la SACM de Mulhouse, réalisa le premier positionnement de compte sur une tabulatrice IBM 421, réalisation qui ouvrait la comptabilité à l'informatique. Cette révolution en gestion mit fin à l'ère des ateliers de machines comptable en grandes entreprises.

L'informatique scientifique, qui consiste à aider les ingénieurs de conception dans les domaines de l'ingénierie industrielle à concevoir et dimensionner des équipements à l'aide de programmes de calcul : réacteurs nucléaires, avions, automobiles (langages souvent employés : historiquement le Fortran, de plus en plus concurrencé par C et C++). L'informatique scientifique est surtout utilisée dans les bureaux d'étude et les entreprises

d'ingénierie industrielle car elle permet de simuler des scénarios de façon rapide et fiable. La Scuderia Ferrari s'est équipée en 2006 avec un des plus puissants calculateurs du monde afin de permettre les essais numériques de sa formule 1 et accélérer la mise au point de ses prototypes.

L'informatique temps réel : elle consiste à définir les logiciels de pilotage de systèmes en prise directe avec le monde physique : historiquement d'abord dans l'aéronautique, le spatial, l'armement, le nucléaire, mais maintenant universellement répandu avec la miniaturisation des circuits : automobile, machine à laver, etc.

L'ingénierie des connaissances (en anglais knowledge management) : il s'agit d'une forme d'ingénierie informatique qui consiste à gérer les processus d'innovation, dans tous les domaines, selon des modèles assez différents de ceux jusqu'alors employés en informatique de gestion. Cette forme d'ingénierie permettra peut-être de mieux mettre en cohérence les trois domaines gestion, temps réel, et scientifique dans l'organisation des entreprises. Elle s'intéresse plus au contenu et à la qualité des bases de données et de connaissances qu'à l'automatisation des traitements. Elle se développe déjà beaucoup aux États-Unis, mais ceci n'est pas encore tout à fait perçu en France.

Il faut enfin citer les applications du renseignement (intelligence en anglais) économique et stratégique, qui font appel aux technologies de l'information, notamment dans l'analyse du contexte, pour la recherche d'informations (moteurs de recherche). D'autre part, dans une optique de développement durable, il est nécessaire de structurer les relations avec les parties prenantes, ce qui fait appel à d'autres techniques telles que les protocoles d'échange et les moteurs de règles.

Approche fonctionnelle

Comme énoncé ci-dessus, l'informatique est le traitement automatisé de données par un appareil électronique : l'ordinateur ; les germanophones parlent de elektronische Datenverarbeitung / EDV (« traitement électronique de données »), les anglophones d'information technology / IT (« technologies de l'information »), c'est-à-dire : données ou informations : in fine, l'ordinateur manipule des suites ou des paquets de 0 et de 1 qui servent à représenter toutes sortes d'informations : des... nombres bien évidemment, dans le cas de calculs scientifiques (flottants) ou comptables (décimal, ou binaire entier)... ; un texte, des lettres (caractères), que l'on peut mettre en forme avec un traitement de texte, imprimer, envoyer par courrier électronique... ; du dessin vectoriel (CAO, logiciels d'illustration, et de typographie) ; des images statiques (photographies) ou animées (vidéo), des hologrammes ; des sons, enregistrés (technique du direct to disk) ou bien fabriqués par l'ordinateur (synthétiseur), que ce soient des bruitages, de la musique (cf. musique et informatique) ou de la parole ; la conversion de ces informations en suite de nombres pose le problème du format des données, du codage et des formats normalisés (par exemple, représentations des nombres entiers ou à virgule flottante, encodage des textes en ASCII, Unicode, format TeX ou RTF et polices PostScript ou TrueType pour les textes, formats bitmap, TIFF, JPEG, PNG, etc. pour les images fixes, formats QuickTime, MPEG pour les vidéos, interface MIDI pour la musique...).

automatisé : l'utilisateur n'intervient pas, ou peu, dans le traitement des données ; le traitement est défini dans un programme qui se déroule tout seul, l'utilisateur se contente de

fournir des paramètres de traitement ; le programme automatique se déroule selon un algorithme, l'établissement de ce programme est le domaine de la programmation.

Les moyens et techniques d'archivage varient en fonction de la durée de conservation souhaitée et des quantités de données en jeu : mémoires électroniques, bandes magnétiques, disques magnétiques ou optiques ;

Les moyens de restitution dépendent de la nature des données : écrans ou imprimantes pour le texte et les images, haut-parleurs ou instruments MIDI pour les sons...

Approche organisationnelle

L'informatique pour l'organisation est un élément d'un système de traitement d'information (les entrées peuvent être des formulaires papier par exemple) et d'automatisation. Depuis Henry Ford, l'automatisation des tâches ayant été identifiée comme un avantage concurrentiel, la question est : que peut-on automatiser ?

Autant il est relativement facile d'automatiser des tâches manuelles, autant il est difficile d'automatiser le travail intellectuel et parfois créatif. L'approche de l'informatique dans une organisation commence donc par l'élucidation des processus, c'est-à-dire la modélisation du métier. Après validation, la MOA (Maîtrise d'Ouvrage) fournit les spécifications fonctionnelles de (l'ouvrage) qui vont servir de référence dans la conception pour la MOE (Maîtrise d'œuvre).

Cette conception sera alors effectuée dans le respect d'un Cycle de développement qui définit les rôles et responsabilités de chaque acteur. Ainsi, les échanges entre MOA et MOE ne se résument pas à la maîtrise des chantiers (tenue des délais et des coûts, et validation des livrables), la MOA et la MOE sont garantes (éventuellement responsables sur un plan juridique) de la cohérence des systèmes d'information, et de l'adéquation des solutions informatiques avec les problèmes utilisateurs finals initialement constatés.

Matériel [modifier]

Article détaillé : Matériel informatique.

On utilise également le terme anglais hardware (littéralement « quincaillerie ») pour désigner le matériel informatique. Il s'agit de tous les composants que l'on peut trouver dans :

1. Les ordinateurs et leurs périphériques : un ordinateur est un ensemble de circuits électroniques permettant de manipuler des données sous forme binaire, représentées par des variations de signal électrique. Il existe différents types d'ordinateurs :

Un IBM PC 5150 datant de 1981, Système d'exploitation IBM-DOS 2.0Les micro-ordinateurs.

De bureau ou portables. Ils sont composés d'une unité centrale : un boîtier contenant la carte mère, l'alimentation, des unités de stockage. On y ajoute une console : un écran et un clavier. Divers périphériques peuvent leur être ajoutés, une souris, une imprimante, un scanner, etc.

Des micro-ordinateurs particulièrement puissants et chers, utilisés uniquement pour des besoins professionnels pointus (conception assistée par ordinateur). Ce terme était particulièrement en vogue dans les années 1980-1990. Depuis les années 2000, il n'est

guère possible de concevoir une station de travail plus puissante qu'un micro-ordinateur haut de gamme ;

Les mainframes.

Une armoire abrite l'unité centrale et l'alimentation, une ou plusieurs autres les périphériques de stockage (disque dur, sauvegarde) tandis que les moyens de communication et réseau (routeur, hubs, modem) sont dans la même pièce, mais dans des racks séparés. Une console d'administration (écran, clavier, imprimante) est généralement située dans ce même local ;

Les Serveurs.

Ce sont des ordinateurs qui proposent souvent à des entreprises un endroit de stockage universel pour les utilisateurs connectés aux serveurs. Les serveurs peuvent effectuer des tâches telles que : servir de Pare-Feu, héberger un serveur web (page internet partagée sur le World Wide Web) ou tout simplement pour partager un nombre important d'imprimantes et de périphériques. Les prix des Serveurs sont élevés car le Serveur a été conçu pour rester allumé en permanence, alors le matériel est durable et performant. ;

Les PDA (Personal Digital Assistant, encore appelés organiseurs).

Ce sont des ordinateurs de poche proposant des fonctionnalités liées à l'organisation personnelle (agenda, calendrier, carnet d'adresse, etc.). Ils peuvent être reliés à Internet par différents moyens (réseau Wi-Fi, Bluetooth, etc.). ;

Les Centre multimédia (Media Center).

Ce sont des micro-ordinateurs intégrant tous les périphériques nécessaires pour capter, enregistrer, la télévision, lire des films sur support numérique, écouter de la musique, en association avec un écran de télévision, avec généralement une manette à distance. Ce genre de PC est un divertissement familial accessible, bien que leurs prix ont eu tendance à être hauts ces derniers temps, ce type d'ordinateur devient de plus en plus accessible à tous.

Dans le domaine de l'informatique embarquée : téléphone, électroménager, automobile, armements militaires, etc. Les cartes à puces, ou l'informatique industrielle.

Logiciel

Le logiciel désigne la partie à première vue immatérielle de l'informatique, l'organisation et le traitement de l'information : les programmes. On s'est en effet vite rendu compte que des machines techniquement très avancées pour leur époque, comme la Bull Gamma 60, restaient invendables tant qu'on n'avait pas de programmes à livrer pour les rendre immédiatement opérationnelles. IBM lança entre 1968 et 1973 une sorte d'ancêtre du logiciel libre avec son ordinateur 1130, politique qui assura à celui-ci par effet boule de neige un succès immédiat et planétaire, mais les conclusions d'un procès antitrust lui interdirent de distribuer bénévolement du logiciel.

Le monde des mainframes classe les logiciels en catégories suivantes: systèmes d'exploitation, comme MVS ou GCOS bases de données, comme DB2, Ingres ou Oracle ; programmes de communication, comme NCP ou RSCS ; moniteurs de télétraitement ; systèmes transactionnels, comme CICS ou open UTM ; systèmes de temps partagé, utilisés pour le calcul ou le développement ; compilateurs traduisant les langages en instructions

machine et appels système ; tout le reste entrait en une catégorie nommée Logiciels applicatifs. Plus simplement on distingue généralement trois types de logiciels (par ordre de proximité du matériel) :

On classe aussi les logiciels en libre et propriétaire, bien que les deux soient parfois panachés à des degrés divers. Certains ont une fonction bureautique ou multimédia comme par exemple les jeux vidéo. Certains logiciels ont acquis des noms connus de tous.

Le noyau du système d'exploitation crée le lien entre le matériel et le logiciel. Un logiciel, quand il est fourni sous sa forme binaire, serait utilisable uniquement avec un système d'exploitation donné (car il en utilise les services), et ne fonctionnerait que sur un matériel spécifique (car il en utilise le code d'instructions). Une conception plus récente, depuis le milieu de années 1980, consiste à distribuer les logiciels tous binaires confondus, et à les munir d'un système de licences par jetons ou tokens permettant l'usage de N copies simultanées du logiciel sur le réseau, tous matériels confondus. Cette approche est majoritaire dans le monde UNIX.

À l'initiative de Richard Stallman et du GNU, à partir de 1985, une mouvance de programmeurs refuse cette logique propriétaire et ceux-ci se muent en concepteurs inventifs pour se lancer dans le développement d'outils et de bibliothèques système libres et compatibles avec le système UNIX. C'est pourtant le projet indépendant Linux, initié par Linus Torvalds, basé sur les travaux et les outils du GNU, qui aboutira dans la création d'un système d'exploitation complet et libre appelé GNU/Linux.

Une bonne partie des logiciels actuels fonctionnent dans un environnement graphique pour interagir avec l'utilisateur. La diversité des systèmes informatiques a fait apparaître une technique visant à combiner le meilleur de chacun de ces univers : l'émulateur. Il s'agit d'un logiciel permettant de simuler le comportement d'un autre système dans celui que l'on utilise,

Le terme anglais est software, à l'origine un jeu de mot entre hardware (« quincaillerie », pour désigner le matériel) et l'opposition soft/hard (mou/dur), opposition entre le matériel (le dur) et l'immatériel (le mou). Les traductions françaises matériel et logiciel rendent parfaitement cette opposition et cette complémentarité.

Le logiciel réalise normalement une fonction attendue de ses utilisateurs. Néanmoins, des effets secondaires (parfois nommés par contresens de traduction effets de bord) existent. Parfois même, certains logiciels sont destinés à nuire, comme les virus informatiques, nommés en anglais, par analogie avec software : malware (qu'on pourrait traduire par le néologisme nuisiciel, ou logiciel malveillant).

La création des logiciels [modifier]

Un projet informatique s'inscrit dans un cycle de développement qui définit les grandes étapes de la réalisation (planification), de la manière dont on passe d'une étape à l'autre (modèle incrémental, en V, en spirale, méthode up, extreme programming, etc.). Pour les petits projets (ou les petites équipes de développement), cette réflexion est souvent négligée (on se répartit les modules et chacun développe dans son coin). Ceci est une cause fréquente d'erreurs (bogues) et de non-conformité (le produit final n'est pas conforme aux attentes de l'utilisateur). Mais même les énormes projets, avec beaucoup de moyens, sont victimes de cette négligence ; ainsi, l'échec du premier vol d'Ariane 5 fut dû à un problème

de logiciel, etc. Un projet peut alors intégrer une approche de la qualité et de la sûreté de fonctionnement des systèmes informatiques afin de contrôler autant que possible le produit final.

Un projet comprend les étapes suivantes (selon le modèle incrémental) :l'établissement d'un cahier des charges qui définit les spécifications auxquelles devra répondre le logiciel ; la définition de l'environnement d'exécution (architecture informatique) : type(s) d'ordinateur sur lequel le logiciel doit fonctionner (station de calcul, ordinateur de bureau, ordinateur portable, assistant personnel, téléphone portable, guichet automatique de banque, ordinateur embarqué dans un véhicule ; type et version du(des) système(s) d'exploitation sous-jacent ; périphériques nécessaires à l'enregistrement des données et à la restitution des résultats (capacité de stockage, mémoire vive, possibilités graphiques...) ; nature des connexions réseau entre les composants (niveau de confidentialité et de fiabilité, performances, protocoles de communication...) ; la conception de l'application et de ses constituants, et notamment de l'interactivité entre les modules développés : structure des données partagées, traitement des erreurs générées par un autre module... : c'est le domaine du génie logiciel ; la mise en place d'une stratégie de développement : répartition des tâches entre les développeurs ou les équipes de développement, qui vont assurer le codage et les tests ;

le plan de test du logiciel, pour s'assurer qu'il remplit bien la mission pour laquelle il a été écrit, dans toutes les conditions d'utilisation qu'il pourra normalement rencontrer, mais aussi dans des cas limites.

Après chacune de ces phases, on peut avoir une étape de recette, où le client va valider les choix et les propositions du maître d'œuvre.

La phase de programmation consiste à décrire le comportement du logiciel à l'aide d'un langage de programmation. Un compilateur sert alors à transformer ce code écrit dans un langage informatique compréhensible par un humain en un code compréhensible par la machine, le résultat est un exécutable. On peut également, pour certains langages de programmation, utiliser un interpréteur qui exécute un code au fur et à mesure de sa lecture, sans nécessairement créer d'exécutable. Enfin, un intermédiaire consiste à compiler le code écrit vers du bytecode. Il s'agit également d'un format binaire, compréhensible seulement par une machine, mais il est destiné à être exécuté sur une machine virtuelle, un programme qui émule les principales composantes d'une machine réelle. Le principal avantage par rapport au code machine est une portabilité théoriquement accrue (il « suffit » d'implanter la machine virtuelle pour une architecture donnée pour que tous les programmes en bytecode puissent y être exécutés), portabilité qui a fait, après sa lenteur, la réputation de Java. Il convient de noter que ces trois modes d'exécution ne sont nullement incompatibles. Par exemple, OCaml dispose à la fois d'un interpréteur, d'un compilateur vers du bytecode, et d'un compilateur vers du code natif pour une grande variété de processeurs. Une fois écrit (et compilé si nécessaire), le code devient un logiciel.

Pour des projets de grande amplitude, nécessitant la collaboration de beaucoup de programmeurs, voire de plusieurs équipes, on a souvent recours à une méthodologie commune (par exemple MERISE) pour la conception et à un atelier de génie logiciel (AGL) pour la réalisation.

Au cours de la programmation et avant la livraison du produit final, le programme est testé afin de vérifier qu'il fonctionne bien (y compris dans des cas d'utilisation en mode dégradé) et qu'il est conforme aux attentes de l'utilisateur final. Les tests intermédiaires permettent de s'assurer que chaque module de code réalise correctement une fonction : ce sont les tests unitaires. Les tests finals qui vérifient le bon enchaînement des modules et des traitements sont des tests d'intégration.

Pour certaines applications demandant un haut niveau de sûreté de fonctionnement, les tests sont précédés d'une étape de vérification, où des logiciels spécialisés effectuent (généralement sur le code source, mais parfois aussi sur le code compilé) un certain nombre d'analyses pour vérifier partiellement le bon fonctionnement du programme. Il n'est toutefois pas possible (et des théorèmes mathématiques montrent pourquoi), de garantir la parfaite correction de tout logiciel par ce moyen et la phase de test reste donc nécessaire. Elle se complète aussi, lorsqu'il s'agit d'une évolution d'une application existante, de nombreux tests automatisés de non régression. Les tests non plus ne pouvant pas garantir totalement l'absence d'erreurs, il est bon de les compléter par des phases de vérification par relecture : des techniques existent pour essayer de rendre cette vérification exhaustive.

- **Exemple 4 : OPEC**

منظمة أوبك .. تاريخ حافل بالأزمات

تعد منظمة الدول المصدرة للبترول «أوبك» من أشهر المنظمات الدولية في العالم، فهي تزود العالم بما نسبته 40٪ من احتياجاته. ومن ثم فالكل يعلم دورها وأهميتها ويترقب اجتماعاتها ويدرس بعناية قراراتها. ولذلك ليس غريباً أن تحظى قمم أوبك ببالغ الاهتمام الاعلامي على صعيد العالم كافة. وفي هذا السياق، حظيت قمة «أوبك» الثالثة التي عقدت بالرياض يومي 17 و18 نوفمبر 2007م الماضي باهتمام بالغ، وسط مطالب وضغوط لزيادة إنتاج المنظمة ، بهدف تهدئة الأسعار في أسواق البترول العالمية، والتي تشهد ارتفاعاً غير مسبوق إذ قاربت عتبة مائة دولار أمريكي للبرميل، مما ألقى بظلاله على اقتصادات الدول المستهلكة وزاد من معدلات التضخم والغلاء في مختلف أنحاء العالم.

وفيما تتوجه الدول الصناعية الكبرى باللوم وتحميل المسؤولية وراء ذلك الارتفاع، للدول المنتجة، وعلى رأسها دول منظمة «أوبك»، ترى الأخيرة أن ارتفاع أسعار البترول لا يعود مطلقاً إلى نقص في الإمدادات، وإنما لأسباب أخرى؛ كالمضاربات، والأوضاع السياسية، ونقص طاقات التكرير، ودور الشركات العالمية الكبرى، وضرائب الكربون التي تفرضها الدول الغربية، فضلاً عن آليات السوق التي أضرت بأسعار البترول وأوصلتها إلى مستوى أقل من عشرة دولارات أمريكية للبرميل عام 1998م.

بيد أن قمة الرياض لم تتخذ قراراً بشأن الأسعار والإنتاج، وإنما ركزت بحثها على ثلاث قضايا رئيسية هي؛ توفير إمدادات الطاقة، تدعيم الاقتصاد العالمي، وحماية البيئة. وهي بذلك أعطت رسالة قوية تؤكد التزام المنظمة بتوفير إمدادات كافية من البترول، والمساهمة في تحقيق رخاء اقتصادي عالمي.

لكن ما هي حدود قدرات «أوبك» على الوفاء برسالتها؟ وهل يتسق ذلك الالتزام مع تاريخها منذ بدايته وإلى اليوم؟ بالطبع لا يمكن الاجابة على هذين السؤالين دون استعراض تاريخ أوبك، تلك المنظمة التي تشغل العالم ليل نهار.

إنشاء أوبك

لقد تأسست منظمة «أوبك» خلال مؤتمر عقد في العاصمة العراقية بغداد خلال الفترة من 10 إلى 14 سبتمبر 1960م، بمشاركة ممثلين عن أهم الدول المصدرة للبترول في حينه، وهي المملكة العربية السعودية وإيران والعراق والكويت وفنزويلا. ثم زاد عدد الأعضاء فيما بعد إلى 13 عضواً، حيث انضمت كل من؛ قطر (1961م)، إندونيسيا (1962م)، الجماهيرية العربية الليبية الشعبية الاشتراكية العظمى (1962م)، الإمارات العربية المتحدة (1967م)، الجزائر (1969م)، نيجيريا (1971م)، الجابون (1975-1994م)، أنجولا (2007م)، والاكوادور التي انسحبت منها عام 1992م احتجاجاً على نظام الحصص في المنظمة قبل أن تستعيد عضويتها خلال قمة «أوبك» الثالثة التي عقدت مؤخراً بالعاصمة السعودية الرياض.

ويتمثل الهدف من إنشاء منظمة «أوبك» في تنسيق السياسات البترولية بين الدول الأعضاء وتوحيدها من أجل تأمين أسعار عادلة ومستقرة للبترول للدول المنتجة، وتأمين إمدادات فاعلة واقتصادية ومنتظمة للدول المستهلكة، وكذلك تأمين عوائد مجزية للقطاعات المستثمرة في هذه الصناعة.

وكانت مدينة جنيف السويسرية مقراً لمنظمة «أوبك» خلال السنوات الخمس الأولى من عمرها، ثم انتقلت إلى فيينا عاصمة النمسا عام 1965م. وتتكون «أوبك» حالياً من ثلاثة أجهزة هي؛ المجلس الوزاري (يمثل السلطة العليا)، ومجلس المحافظين المكون من مندوبي الدول الأعضاء، وأخيراً الأمانة العامة وتضم الأمين العام ورؤساء الإدارات وبقية الموظفين.

وقد قررت منظمة «أوبك» خلال مؤتمرها الوزاري السابع في جاكرتا بإندونيسيا عام 1964م، إنشاء اللجنة الاقتصادية كجهاز لمساعدة المنظمة على مراقبة وتحقيق الاستقرار في أسعار البترول العالمية. كما أنشأت خلال اجتماع عقد في باريس لوزراء مالية الدول الأعضاء عام 1976م، «صندوق أوبك للتنمية» بهدف «تقديم العون المالي للدول النامية ومساعدتها بشروط ميسرة». ومنذ ذلك التاريخ قدم الصندوق مساعدات مالية لـ119 دولة، وهو يمثل اليوم كياناً قائماً بذاته، وقد حصل العام الماضي على مقعد مراقب في الأمم المتحدة.

ويعتبر اجتماع قمة الرياض الأخير اجتماعاً نادراً في تاريخ منظمة «أوبك»، فهو الثالث منذ تأسيس المنظمة قبل 47 عاماً، وقد سبقته قمة الجزائر التي أصدرت عام 1975م «إعلان المبادئ الأول»، وعقد اجتماع القمة الثاني في العاصمة الفنزويلية كاراكاس عام 2000م ليتمخض عن «إعلان المبادئ الثاني». أما «إعلان المبادئ الثالث» والذي ختمت به قمة الرياض أعمالها؛ فقد أعاد التأكيد على إعلاني الجزائر وكاراكاس، وعلى «الاستمرار في توفير البترول بشكل كاف وموثوق ومناسب للأسواق العالمية»، و«العمل مع جميع الأطراف من أجل أسواق عالمية متوازنة للطاقة»، و«التأكيد على أهمية السلام العالمي في تقوية الاستثمار في مجال الطاقة»، و«تأكيد العلاقة المتبادلة بين الأمن العالمي لتوفير البترول وأمن الطلب»، و«تقوية وتوسيع الحوار بين المنتجين والمستهلكين للطاقة»، و«استمرار التنسيق مع الدول الأخرى المصدرة للبترول».

وتبذل منظمة «أوبك» جهوداً كبيرة للحفاظ على استقرار أسواق البترول العالمية، وذلك من خلال نظام الحصص الملزم، حيث يبلغ إجمالي ما تنتجه دولها حوالي 30 مليون برميل يومياً، تتصدرها حصة المملكة العربية السعودية بإنتاج وصل في شهر نوفمبر 2007م الماضي إلى 9 ملايين برميل في اليوم.

لكن إسهام دول «أوبك» مجتمعة لا يتجاوز 40% من إجمالي الإنتاج البترولي العالمي، حيث يستطيع منتجون كبار

مثل روسيا، التأثير من جانبهم في اتجاهات السوق البترولية العالمية. بيد أن أهم عامل مؤثر في السوق هو حجم الطلب المتزايد على البترول من جانب الدول ذات الاقتصادات الناهضة، لاسيما الصين والهند والمكسيك. لكن تظل الولايات المتحدة الأمريكية أكبر مستورد للبترول، إذ تستهلك حوالي 12 مليون برميل يومياً، منها نسبة 45% قادمة من دول «أوبك».

ولهذا السبب حاول مجلس الشيوخ الأمريكي في شهر يونية 2007م الماضي تمرير خطة تمكن الإدارة الأمريكية من القيام بإجراءات قانونية ضد «أوبك»، ومقاضاة دولها الأعضاء أمام محاكم أمريكية، بتهمة التلاعب بالأسعار وإرباك السوق. وهي ثالث محاولة من نوعها للكونجرس الأمريكي منذ عام 2005م، لكن البيت الأبيض كان يسقطها في كل مرة.

أوبك ... تاريخ من الأزمات

لقد مرت منظمة «أوبك» منذ إنشائها بمحطات مهمة وخطيرة. ففي الستينيات الميلادية من القرن العشرين الماضي، مثلت هذه الحقبة الفترة التي تم فيها تأسيس المنظمة، في مسعى من الدول الأعضاء إلى تجسيد حقوقها الشرعية في سوق البترول الدولية التي تهيمن عليها شركات البترول المتعددة الجنسيات. وكانت المنظمة تمارس أنشطتها بصفة عامة في ذلك الوقت في الظل، حيث قامت في تلك الفترة بتحديد أهدافها، وتأسيس الأمانة العامة، التي انتقلت من جنيف إلى فيينا في عام 1965م، كما عملت على تبني عدد من القرارات، وإجراء المفاوضات مع الشركات.

وفي فترة السبعينيات الميلادية بزغ نجم منظمة أوبك، حيث تمكنت الدول الأعضاء في المنظمة من التحكم في قطاعات الصناعات البترولية المحلية في بلدانها، وأصبحت لها الكلمة العليا في ما يتعلق بمسألة تحديد أسعار البترول الخام في أسواق البترول الدولية. وقد شهدت هذه الفترة أزمتين اثنتين في ما يتعلق بمسألة أسعار البترول، الأولى كانت بسبب حظر البترول الذي فرضته الدول العربية المنتجة للبترول في عام 1973م، والثانية بسبب اندلاع الثورة الإيرانية في عام 1979م، والتي غذتها في الوقت ذاته التقلبات الكبيرة في سوق البترول، وقد شهدت أسعار البترول خلال هاتين الأزمتين ارتفاعا حاداً.

وعقدت القمة الأولى لرؤساء ورؤساء حكومات الدول الأعضاء في المنظمة في الجزائر في شهر مارس 1975م.

وكانت ذروة الأحداث التي مرت بها المنظمة في فترة الثمانينيات الميلادية من القرن الماضي، حيث وصلت أسعار البترول في بداية هذا العقد إلى ذروتها، وذلك قبل أن تبدأ في الانحدار والتدهور إلى مستويات متدنية جداً، الأمر الذي أدى إلى انهيار شامل في الأسعار في عام 1986م، وهي الفترة التي شهدت الأزمة الثالثة في ما يتعلق بأسعار البترول، ولكن الأسعار عادت لترتفع في الأعوام الأخيرة من هذا العقد، ولكن من دون الوصول إلى المستويات العليا التي وصلت إليها في بدايات هذه الحقبة. كما ارتفع في هذه الفترة مستوى الوعي بضرورة القيام بعمل مشترك من قبل الدول المنتجة للبترول إذا ما أرادت تحقيق استقرار في سوق البترول الدولية بأسعار معقولة في المستقبل. كما شهدت هذه الفترة بداية بروز المسائل البيئية المتعلقة بصناعة البترول على الساحة الدولية.

وفي بداية فترة التسعينيات الميلادية من القرن العشرين الميلادي، تم تفادي حدوث أزمة رابعة في أسعار البترول، وذلك بسبب غزو العراق للكويت عام 1990م، حيث تمكنت الدول الأعضاء في منظمة أوبك من الوصول بالأسعار إلى مستويات معقولة في الأسواق البترولية العالمية، التي شهدت ارتفاعاً حاداً مفاجئاً بسبب انتشار الذعر والهلع في تلك

الأسواق، بسبب تلك الأحداث، وذلك من خلال قيام دول المنظمة بزيادة الإنتاج من البترول في تلك الفترة.
وبعد ذلك بقيت الأسعار مستقرة نسبياً حتى عام 1998م، عندما حصل انهيار آخر في أسعار البترول بسبب الانهيار الاقتصادي الذي شهدته دول جنوب شرق آسيا، ولكن رد الفعل الذي قامت به الدول الأعضاء في منظمة أوبك، إلى جانب عدد من الدول المنتجة الكبرى من خارج المنظمة، من خلال القيام بعمل جماعي مشترك لمجابهة هذه الأزمة، أدى في نهاية الأمر إلى العودة بالأسعار إلى مستويات مرضية.

مؤتمرات قادة «أوبك» السابقة

تبنت المنظمة في شهر يونية 1968م «إعلان المبادئ الخاص بالسياسة البترولية في الدول الأعضاء». ودعا الإعلان الدول الأعضاء إلى اتخاذ الخطوات التالية، كلما كان ذلك ممكناً:

* التنقيب المباشر عن الموارد الهيدروكربونية وتطويرها.
* المشاركة في ملكية اتفاقيات الامتياز القائمة.
* التخلي التدريجي والمتسارع عن المساحات الموجودة في المناطق المتعاقد عليها حالياً.
* وضع قوانين حماية وصيانة من أجل تقيد شركات البترول العاملة في دول المنظمة بها.
* تحديد أسعار معلنة أو فرض ضرائب على الأسعار من قبل حكومات الدول الأعضاء في المنظمة، من أجل منع تدهور العلاقة بين أسعار البترول وأسعار البضائع المصنعة التي تتم المتاجرة فيها دولياً.

القمة الأولى: عقدت القمة الأولى لقادة دول «أوبك» في الجزائر في شهر مارس 1975م وتم خلالها تبني «إعلان مبادئ Declaration Solemn» أكد على ضرورة اتباع إرشادات جديدة في ما يتعلق بالسياسات البترولية، وذلك في ضوء تغير نمط العلاقات بين الدول المنتجة والدول المستهلكة.

وقد أشارت تلك الإرشادات إلى أنه يتعين على منظمة أوبك، من خلال التشاور والتنسيق مع بقية دول العالم، أن تسعى إلى تأسيس نظام اقتصادي عالمي جديد، مبني على أسس العدالة والتفاهم المشترك والاهتمام برفاهية جميع الشعوب وازدهارها.

القمة الثانية: عقدت القمة الثانية في عام 2000م في العاصمة الفنزويلية كاراكاس. وقد أكدت الدول الأعضاء في المنظمة في «إعلان المبادئ الثاني» التزامها بالمبادئ الاسترشادية للمنظمة، وذلك من أجل تحقيق نظام واستقرار دائمين في أسواق البترول العالمية، بأسعار معقولة وعوائد مجزية للمستثمرين.

وقد تطرق رؤساء ورؤساء حكومات الدول الأعضاء في المنظمة إلى تعزيز دور البترول في مجال الطلب العالمي على الطاقة في المستقبل، وأكدوا الرابطة القوية التي تربط بين أمن الواردات وأمن الطلب على البترول وشفافيته، كما أكدوا الحاجة إلى تحسين سبل الحوار والتعاون بين جميع الأطراف العاملة في هذا القطاع.

كما أعادوا النقاش في مسألة الخدمة التي يقدمها البترول للعالم بصفة عامة، والحاجة إلى ربط واردات الطاقة بالتنمية الاقتصادية والتوافق والانسجام في المسائل البيئية، وذلك من أجل المساعدة في تقليل المعاناة وحالات الفقر التي تواجهها الدول النامية، وكذلك من أجل تحفيز اقتصاداتها على النمو والازدهار.

محطات رئيسية في تاريخ «أوبك»

* **1960م تأسست أوبك بعضوية كل من المملكة العربية السعودية والكويت والعراق وإيران وفنزويلا.**
* **1965م انتقل مقر منظمة أوبك من سويسرا إلى فيينا عاصمة النمسا.**
* **1973م حظر بترولي عربي رفع أسعار البترول وأدى لأزمة في الاقتصاد العالمي.**
* **1979م أزمة في أسعار البترول بسبب اندلاع الثورة الإيرانية.**
* **1998م سعر برميل البترول يهوي إلى عشرة دولارات أمريكية.**
* **2000م منظمة أوبك تخفض الإنتاج لتعزيز الأسعار.**
* **2001م منظمة أوبك تضغط على الدول المصدرة للبترول غير الأعضاء فيها لكي تقلل إنتاجها.**