Université du Québec à Trois-Rivières

MÉMOIRE PRÉSENTÉ COMME EXIGENCE PARTIELLE DE LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUE APPLIQUÉES.

PAR

WADII HAJJI

Les treillis de Galois et leurs applications dans la classification textuelle.

Avril 2003

Résumé

La classification des données textuelles pose des problèmes. Un nombre grandissant d'institutions accumulent très rapidement de très grandes quantités de documents qui ne sont souvent catégorisés que d'une façon très sommaire. Rapidement, les tâches de dépistage, d'exploration et de récupération de l'information présente dans ces textes, c'est-à-dire des connaissances, sont devenues extrêmement ardues, sinon impossibles. En raison de l'ampleur et de la dynamicité des corpus, cette extraction des connaissances devient de moins en moins possible dans des temps raisonnables ou faisables avec des ressources restreintes.

Nous proposons les Treillis de Galois comme solutions à ce problème. Cette approche permet d'appliquer les performances classificatoires à des corpus textuels et produire donc des regroupements susceptibles d'interprétations sémantiques. Le traitement par Treillis de Galois est intégré à des processus de pré-traitement du texte. Par ailleurs, la notion de n-grams, qui donne de bons résultats dans l'identification de la langue ou dans l'analyse de l'oral, est, par les recherches récentes, devenue un axe privilégié dans l'acquisition et l'extraction des connaissances dans les textes. La chaîne de traitement ainsi produite peut fournir à l'usager un outil précieux dans les tâches d'extraction des connaissances.

Mots-clés : Classification, les Treillis de Galois, analyse des textes, apprentissage non supervisé, réseaux de neurones ART.

Remerciements

Ce travail n'aurait pu voir le jour sans le soutien, les conseils et les encouragements de mon directeur de recherche, le professeur Ismaïl Biskri. Je te remercie Ismaïl, d'avoir accepté de diriger ce travail et de m'avoir orienté dans mes moments les plus critiques.

Je désire également remercier les membres du jury d'avoir accepté d'évaluer ce travail et d'y avoir consacré autant de temps et d'attention.

Finalement, je tiens à remercier tous mes collègues et amis, pour le support et la compréhension qu'ils ont si bien su me montrer.

Table des matières

Table des figures.

Liste des tableaux.

Introduction.

L'information de nos jours est devenue de plus en plus présente dans une multitude de domaines qui ne sont pas forcément scientifiques. L'essor du Web rend le phénomène encore plus perceptible.

Dans la situation économique actuelle, de nombreuses entreprises s'interrogent sur une meilleure utilisation de leur capital de connaissances. Elles disposent de dizaines d'années d'activités durant lesquelles se sont accumulées de nombreuses connaissances (expérience et savoir-faire), en général sous la forme de rapports, notes techniques... archivées et mal exploitées, d'où le besoin actuel de capitaliser les connaissances et l'expérience acquise, et un intérêt prononcé pour la **gestion des connaissances et du savoir-faire de l'entreprise**. D'un point de vue technique, on peut envisager de stocker ces diverses connaissances dans des **bases de données hypertextes et multimédia** (étant donnée la diversité des objets manipulés : textes, photos, dessins/schémas, plans, sons et parole, et parfois même séquences vidéo pour présenter un produit, un procédé ou l'activité d'un expert).

De ce fait l'échange d'information ne cesse de nécessiter des outils performants pour en faciliter l'accès. Ainsi l'indexation, le filtrage, la recherche d'information, l'extraction des connaissances etc. sont devenues autant de domaines que privilégie la recherche scientifique en traitement des langues naturelles. Tous ces travaux partagent une vision comme la préparation de l'information. En effet toutes les opérations précitées font appel à la classification textuelle. Nous pouvons définir la classification textuelle comme un traitement permettant de regrouper ce que nous appelons dans une première définition des documents similaires dans des classes. Cette similarité est perçue du fait de la régularité de cooccurrence d'unités d'information par exemple le mot dans les différents documents.

Parmi les algorithmes de classification que nous retrouvons dans la littérature :

- Les méthodes neuronales (Cartes auto organisantes de Kohonen, Réseau de hopfield, ART (Adaptive Resonance Theory)),

- Méthodes d'extraction de graphes (partition de graphe, motifs de graphe, Graphe connexe Arbres de décision, Treillis de Galois).
- Modèle de représentation des données (Modèle de vecteur d'objets, Modèle des N-grammes, Modèle de l'information syntaxique, Modèle de distance de similarité).
- Méthode des plus proches voisins (k-moyennes) (Méthodes avec simple passe (single-pass).
- Méthode de réallocation (reallocation), Méthode des nuées dynamiques et ces centres mobiles).
- Méthodes hiérarchiques descendantes, Méthode hiérarchiques ascendantes.

Le premier chapitre de ce mémoire est consacrée à une brève introduction à la classification par réseaux de neurones, nous décrivons dans quelques paragraphes l'idée principale qui a donné naissance aux réseaux de neurones et l'inspiration que les scientifiques ont utilisé pour créer le perceptron. À la fin du chapitre nous présentons le fonctionnement du réseau ART par un algorithme et des schémas explicatifs.

Dans le deuxième chapitre nous présentons de la littérature reliée à la théorie des treillis en général et celle des treillis de Galois en particulier. Nous donnons un fondement mathématique des treillis avec toutes les définitions et les démonstrations possibles, ensuite nous traitons le cas particulier des treillis appelés le treillis de galois.

Le troisième chapitre présente l'objectif principal de notre travail avec des détails qui s'imposent.

Dans le dernier chapitre on propose une interprétation des résultats de notre expérimentation, on discutera de la performance des treillis de Galois comparés au classifieur ART.

CHAPITRE 1

Réseaux de neurones

1.1 Avant-propos

L'approche algorithmique consiste à concevoir, tout le processus que la machine devra suivre dans le but de résoudre un problème. Lorsque celui-ci peut se traduire à quelques équations mathématiques, l'algorithmique remplit parfaitement son rôle. Cependant, dans des cas plus complexes, résoudre un problème de cette façon peut s'avérer coûteux en temps, et même impossible. De plus, ce type d'approche oblige le concepteur à prendre en considération tous les cas de figures envisageables, puisque la machine, étant résolument binaire et disciplinée, serait incapable de prendre seule des décisions face à une situation pour laquelle elle n'a pas été programmée.

C'est d'ailleurs l'envie de fournir à une machine les moyens nécessaires, afin de prendre des décisions, qui a fait voir le jour à une approche complètement différente, il s'agit de l'Intelligence Artificielle (IA) (ou *Artificial Intelligence* (AI) en anglais).

Définition 1.1 :

Intelligence artificielle : Discipline scientifique relative au traitement des connaissances et du raisonnement humain dans le but de les reproduire artificiellement et ainsi de permettre à un appareil d'exécuter des fonctions normalement associées à l'intelligence humaine : raisonnement, compréhension, adaptation, etc.

Cependant, face à des situations non prévue au moment de la conception, le fonctionnement reste limité et même trop éloignés des règles de base. L'Intelligence Artificielle a sans aucun doute une grande utilité pour les sciences dites exactes, telles que les mathématiques, l'électronique ou la physique, mais ses limitations se font nettement sentir dans le cas des sciences humaines, par opposition à exactes, comme la médecine ou la

psychologie, dont les principes de base sont avant tout empiriques et impossibles à traduire en expressions mathématiques.

1.2 L'humain, modèle de l'informatique?

Ces deux approches, algorithmique et base de connaissance, se révèlent malheureusement insuffisantes pour résoudre tous les problèmes existants. Il peut même s'avérer frustrant de voir une machine paradoxalement capable d'effectuer en quelques secondes un calcul qui prendrait toute une vie à la main et quasiment incapable d'accomplir des tâches tout à fait banales pour un humain, comme la compréhension du texte. Certains domaines d'application n'ont d'ailleurs jamais été pourvus de systèmes automatiques, faute d'approche logicielle capable de les appréhender avec la puissance actuelle des machines. Quelques essais algorithmiques ou à base de connaissances ont été effectués dans des domaines tels que la reconnaissance de formes, la compréhension du langage, ou la traduction automatique, mais sans jamais rencontrer le succès escompté.

Pour espérer un jour automatiser ce type de processus, il a fallu envisager un autre type d'approche. Les progrès de la médecine aidant, les informaticiens se sont penchés sur l'analyse du fonctionnement du cerveau humain pour tenter de reproduire certains de ses mécanismes sur une machine.

Le point de départ fut l'analyse de la capacité d'un humain à raisonner et à s'adapter à des situations nouvelles. L'étude des processus physiologiques a permis d'attribuer au réseau de neurones humains, connectés entre eux par des synapses, la capacité de réfléchir par le biais de la transmission de signaux actifs.

Une fois établie l'importance de la connexion des neurones, des réseaux neuraux artificiels, s'inspirant d'un modèle humain simplifié, ont vu le jour. Une science nouvelle est née de ces observations, le connexionnisme. Elle constitue d'ailleurs l'un des principaux centres d'intérêt de la recherche actuelle dans bon nombre de domaines, dont notamment l'informatique, la physique, la psychologie, la linguistique, la biologie ou la médecine. Sous le terme "connexionnisme" se cache un moyen particulier de traiter l'information, calqué sur les principes fondamentaux de fonctionnement du cerveau humain.

Le connexionnisme se structure autour d'un grand nombre de cellules primitives, ou unités, connectées entre elles par des liens, et fonctionnant en parallèle, sur les bases du modèle humain évoqué plus haut.

1.3 Les réseaux de neurones artificiels

Aujourd'hui de nombreux termes sont utilisés dans la littérature pour désigner le domaine des réseaux de neurones artificiels, comme connexionnisme ou neuromimétique. Pour notre part, il nous semble qu'il faut associer à chacun de ces noms une sémantique précise. Ainsi, les réseaux de neurones artificiels ne désignent que les modèles manipulés; ce n'est ni un domaine de recherche, ni une discipline scientifique. Connexionnisme et neuromimétique sont tous deux des domaines de recherche à part entière, qui manipulent chacun des modèles de réseaux de neurones artificiels, mais avec des objectifs différents.

L'objectif poursuivi par les ingénieurs et chercheurs connexionnistes est d'améliorer les capacités de l'informatique en utilisant des modèles aux composants fortement connectés. Pour leur part, les neuromiméticiens manipulent des modèles de réseaux de neurones artificiels dans l'unique but de vérifier leurs théories biologiques du fonctionnement du système nerveux central.

Définition 1.2 :

Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau.

1.4 Historique

Les réseaux de neurones artificiel sont "nés" il y a une cinquantaine d'années, des efforts combinés de scientifiques issus d'horizons divers et aux motivations variées. Leur histoire est jalonnée d'un certain nombre de publications clés, livres ou articles d'intérêt essentiellement historique, associés aux étapes décisives de leur développement.

1.4.1 Les débuts

Tout commence en 1943, lorsque deux biophysiciens de l'université de Chicago McCulloch et Pitts [McCulloch et Pitts 1943], s'inspirant des récentes découvertes en neurobiologie, conçoivent le premier modèle du neurone biologique, baptisé **neurone formel** ou automate à seuil. Un peu plus tard, un neurophysiologique renommé, Donald Hebb [Hebb 1949], propose en 1949 une formulation du mécanisme d'apprentissage, sous la forme d'une **règle de modification des connexions synaptiques** qui porte encore son nom. Finalement, c'est en 1958 que Rosenblatt, combinant les idées de ses prédécesseurs, conçoit le **Perceptron**, un réseau de neurones artificiels inspiré du système visuel, possédant une couche de neurones ``perceptive" et une couche de neurones ``décisionnelle". Ce réseau, qui parvient à apprendre à identifier des formes simples et à calculer certaines fonctions logiques, constitue le premier système artificiel exhibant une faculté jusque là réservé au vivant, la capacité d'apprendre par l'expérience; le premier réseau de neurones artificiel proprement dit.

Les travaux de Rosenblatt suscitent au début des années 60 un vif enthousiasme chez les scientifiques alors fortement impliqués dans la recherche sur l'intelligence artificielle. Cet enthousiasme se voit pourtant brusquement refroidi en 1969 lorsque deux scientifiques américains de renom, Minsky et Papert, publient un livre qui, au terme d'une analyse mathématique approfondie, met à jour les limites intrinsèques du Perceptron, en particulier son incapacité à résoudre les problèmes non linéairement séparables, tels que le célèbre problème du XOR (opération logique, ou exclusif). Ces conclusions plongent alors la recherche sur les réseaux de neurones artificiels dans une disgrâce qui ne prendra fin que 15 ans plus tard.

Ce qu'ont démontré Minsky et Papert c'est qu'un réseau de neurones de type perceptron, c'est-à-dire ne possédant qu'une couche de neurones (la couche des neurones d'entrée, "perceptifs") en plus de la couche de sortie, est incapable de résoudre tout une classe de problèmes simples (les problèmes non linéairement séparables). Certes l'utilisation de couches intermédiaires, "cachées", de neurones, permettrait de contourner cette limitation, à condition de disposer d'un mécanisme d'apprentissage approprié pour ces neurones additionnels. Mais c'est précisément ce mécanisme qui à l'époque fait cruellement défaut, ce

qui fait dire en substance aux deux savants américains, qu' «un réseau de type perceptron ne sera jamais capable de faire quoi que ce soit d'intéressant».

1.4.2 Le renouveau

Il faut attendre le début des années 80 pour voir un regain d'intérêt pour les réseaux de neurones artificiels. Celui-ci s'explique tout d'abord par les résultats des travaux de Hopfield [Hopfield 1982]qui démontre, en 1982, l'utilité des réseaux complètement connectés (les réseaux récurrents, avec ``feed-back", qui constituent la deuxième grande classe de réseaux avec les réseaux de type perceptron, aussi qualifiés de ``feed-forward") dans la compréhension et la modélisation des processus de la mémoire et rend manifeste la relation existant, sur le plan formel, entre ce type de réseaux et des systèmes physiques (tels que les verres de spin) pour lesquels la physique statistique fournit un cadre théorique parfaitement approprié. Parallèlement aux travaux de Hopfield, Werbos conçoit un mécanisme d'apprentissage pour les réseaux multicouches de type perceptron: c'est l'algorithme d'apprentissage par **"Back-propagation"** (rétropropagation de l'erreur) qui fournit un moyen simple d'entraîner les neurones des couches cachées. Cet algorithme sera réellement popularisé en 1986 par Rumelhart et al [Rumelhurt 1986] dans un article de Nature et un livre ("Parallel Distributed Processing") qui a longtemps constitué la « bible » des connexionnistes.

Cet algorithme, et ce livre, ont eu un impact considérable. Disposant d'un moyen simple d'entraîner les neurones cachés, les réseaux de type perceptron munis d'une ou plusieurs couches cachées, (appelés MLP pour Multi-layer Perceptron) qui, contrairement à leur célèbre ancêtre, ne souffrent d'aucune limitation théorique, ont pu être employés avec un succès grandissant pour résoudre toute une panoplie de problèmes complexes rencontrés dans de nombreux domaines à la fois scientifiques et techniques.

Depuis la fin des années 80, l'intérêt pour les réseaux de neurones artificiels ne s'est pas démenti, dans tous les milieux et sur tous les fronts. En pratique, les réseaux de neurones ont dors et déjà donné lieu à de très nombreuses applications.

1.5 Le modèle neurophysiologique

1.5.1 Le neurone

Les cellules nerveuses, appelées neurones, sont les éléments de base du système nerveux central. Celui-ci en posséderait environ 1000 milliards. Les neurones possèdent de nombreux points communs dans leur organisation générale et leur système biochimique avec les autres cellules. Ils présentent cependant des caractéristiques qui leur sont propres et se retrouvent au niveau des cinq fonctions spécialisées qu'ils assurent :

- Recevoir des signaux en provenance de neurones voisins.
- Intégrer ces signaux.
- Engendrer un influx nerveux (message nerveux).
- Le conduire
- Le transmettre à un autre neurone capable de le recevoir.

1.5.2 Structure des neurones.

Comme nous venons de dire le système nerveux compte plus de 1000 milliards de neurones interconnectes, avec 1000 à 10000 synapses par neurone. Bien que les neurones ne soient pas tous identiques, leur forme et certaines caractéristiques permettent de les répartir en quelques grandes classes. En effet, il est aussi important de savoir, que les neurones n'ont pas tous un comportement similaire en fonction de leur position dans le cerveau. Mais nous ne rentrerons pas plus dans les détails. Examinons un neurone.

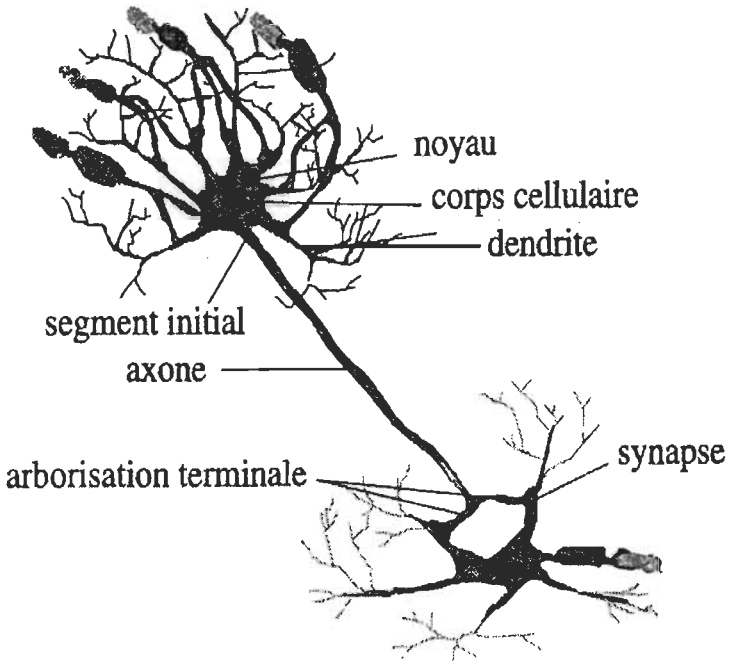

Figure 1.1 - Un neurone simple.

Nous pouvons le décomposer en trois régions principales:

Le corps cellulaire

Il contient le noyau du neurone et effectue les transformations biochimiques nécessaires à la synthèse des enzymes et des autres molécules qui assurent la vie du neurone. Sa forme est pyramidale ou sphérique dans la plupart des cas. Elle dépend souvent de sa position dans le cerveau. Ainsi les neurones du néo-cortex ont principalement une forme pyramidale. Ce corps cellulaire fait quelques microns de diamètre.

Les dendrites

Ce sont de fines extensions tubulaires qui se ramifient autour du neurone et forment une sorte de vaste arborescence. Les signaux envoyés au neurone sont captés par les dendrites. Leur taille est de quelques dizaines de microns de longueur.

L'axone

C'est le long de l'axone que les signaux partent du neurone. Contrairement aux dendrites qui se ramifient autour du neurone, l'axone est plus long et se ramifie à son extrémité ou il se connecte aux dendrites des autres neurones. Sa taille peut varier entre quelques millimètres à plusieurs mètres.

1.6 Fonctionnement des neurones

Chaque neurone est une cellule. Autour du noyau, on trouve le corps cellulaire (sommateur ou soma). Celui-ci se prolonge par un axone unique et comporte de nombreuses dendrites qui constituent son organe (d'entrée). Dans chaque neurone en repos, ils a une différence de potentiel entre l'intérieur du neurone et le monde extérieur, cette différence de potentiel est de -70 mV.

L'influx nerveux est assimilable à un signal électrique, se propageant dans les neurones de la manière suivante :

1. Les dendrites reçoivent l'influx nerveux d'autres neurones.
2. Le neurone évalue alors l'ensemble de la stimulation qu'il reçoit (c'est à dire sa dépolarisation par rapport à l'extérieur).
3. En fonction de cette stimulation (si la dépolarisation est suffisante > -50mV par exemple), le neurone transmet ou non un signal de type « tout ou rien » le long de son axone, selon une fréquence en fonction du niveau de dépolarisation. On dira alors que le neurone est ou non excité.
4. L'excitation du neurone est propagée le long de l'axone jusqu'aux autres neurones ou fibres musculaires qui y sont connectés via les synapses.

L'influx nerveux n'est pas atténué au cours de sa transmission et est invariant en forme et amplitude. Le seul paramètre variable est la fréquence de passage de l'influx

nerveux. Ce passage permet de modifier les deux premiers paramètres uniquement par effet de superposition ou addition au niveau des synapses. Ce train d'ondes dont l'amplitude électrique mesurable par électroencéphalogramme est de l'ordre de 50mV se déplace le long des fibres neuronales à une allure variant de 10 à 100 m/s.

La transmission du potentiel d'action au travers des synapses n'est pas similaire à une simple connexion électrique entre deux conducteurs. En effet, l'arrivée d'un potentiel d'action à l'extrémité d'un axone entraîne le passage en plus ou moins grande quantité de neurotransmetteurs chimiques dans l'intervalle synaptique.

1.6.1 Le cerveau

Le cerveau humain pèse environ 1,500Kg. Il est composé de trois couches successives :

- Le cerveau reptilien qui est la couche la plus ancienne.
- L'hippocampe (ou cerveau archaïque) est le siège de l'émotivité et de la sensibilité.
- Le cortex (couche la plus récente) est propre aux mammifères, c'est le siège du raisonnement et de la pensée conceptuelle.

Le cerveau peut aussi être décomposé en un certain nombre de régions, tant sur le plan morphologique que fonctionnel. Si la plupart des fonctions sont réparties de manière symétrique sur les deux hémisphères du cerveau, un certain nombre de fonctions (de haut niveau) ne sont en fait affectées qu'à un seul hémisphère. Il est aussi important de noter que plusieurs types de neurones peuvent exister au niveau du cortex (pyramidaux, stellaires, en panier, ...) et sont structurés à la fois horizontalement et verticalement.

La notion de Population de Neurones définie par Hebb a permis de décrire une sorte de **coopération entre neurones**. Chacun des neurones de cette population est nécessaire pour maintenir l'activation des autres. On retrouve l'opposé de cette notion pour décrire des neurones ayant un fonctionnement concurrentiel. Lorsque dans un ensemble de neurones concurrentiels seul un neurone reste actif, on dit qu'il y a eu un choix ou un phénomène de

type (winner takes all). Le système nerveux est aussi capable de s'adapter, c'est à dire de modifier les efficacités synaptiques. La règle de Hebb qui renforce les corrélations les plus importantes est l'une des premières à avoir été formalisée. Cette règle semble d'ailleurs être utilisée dans le cerveau.

Le cerveau n'est donc pas composé d'un seul type de neurones mais au contraire de plusieurs types de neurones avec des topologies et des ensembles de connexions variés en relation avec la fonction du neurone (ouïe, vue, mémoire, etc..).

1.6.2 Correspondance entre biologique et artificiel

La figure 1.2 (voir [Claude Touzet 1992]) montre la structure d'un neurone artificiel. Chaque neurone artificiel est un processeur élémentaire. Il reçoit un nombre variable d'entrées en provenance de neurones amont. A chacune de ces entrées est associé un poids représentatif de la force de la connexion. Chaque processeur élémentaire est doté d'une sortie unique, qui se ramifie ensuite pour alimenter un nombre variable de neurones aval. A chaque connexion est associé un poids.

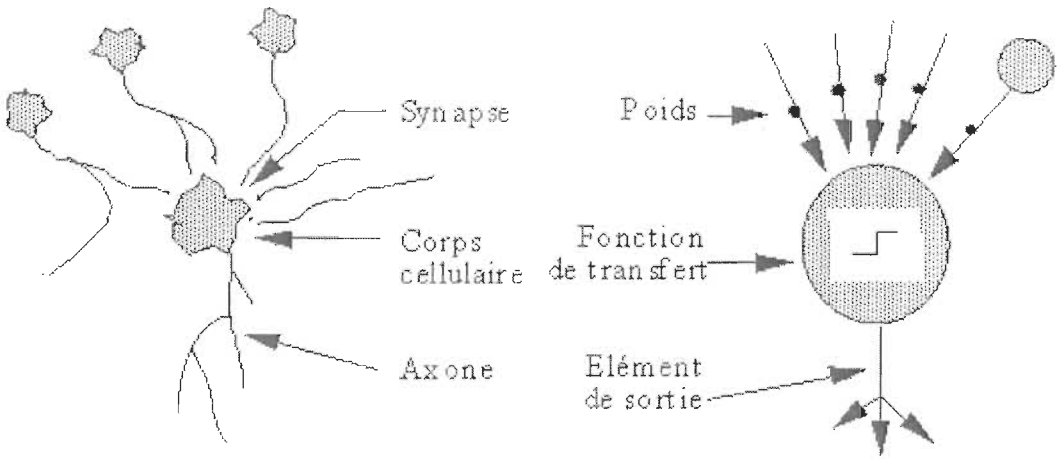

Figure 1.2 - Mise en correspondance neurone biologique / neurone artificiel.

1.7 perceptron ou neurone formel

La première modélisation d'un neurone ou perceptron date des années quarante. Elle a été présentée par Mac Culloch et Pitts.

Un neurone formel fait une somme pondérée des potentiels d'actions qui lui parviennent (chacun de ces potentiels est une valeur numérique qui représente l'état du neurone qui l'a émis), puis s'active suivant la sommation pondérée. Si cette somme dépasse un certain seuil, le neurone est activé et transmet une réponse (sous forme de potentiel d'action) dont la valeur et celle de son activation. Si le neurone n'est pas activé, il ne transmet rien.

Définition 1.3 :

Un perceptron à I entrées et à une seule sortie est défini par la donnée de n+1 constantes : les poids w_1, w_2,....., w_I et le seuil θ qui peuvent être, selon les variantes, des nombres réels ou des entiers.

Le perceptron calcule une sortie y en fonction de I d'entrée X_1,.....,X_I selon la formule :

$$y = \begin{cases} 1 & si \sum_i w_i I_i \geq \theta \\ 0 & \text{sinon} \end{cases}$$

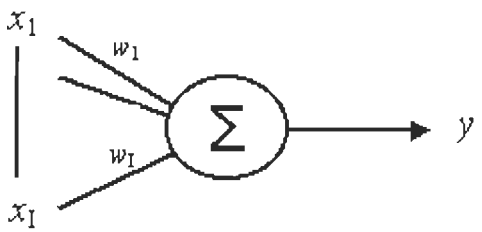

Figure 1.3 - Le perceptron

La quantité $\sum_i w_i I_i$ est le potentiel post-synaptique (ou l'entrée totale). La fonction d'activation utilisée pour calculer la sortie est appelée fonction de Heaviside:

$$f(x)=\begin{cases}1 & si\ x\geq 0\\ 0 & \text{sinon}\end{cases}$$

L'ensemble des variables d'entrée est parfois appelé rétine.

Exemple 1.1 :

Un perceptron qui calcule le « OU » logique. On remarque qu'il suffit de prendre $w_1=1$, $w_2=1$ et $\theta = 0$.

On voit que quelques uns des traits principaux des neurones réels ont été retenus dans la définition du perceptron : les entrées modélisent les dendrites, les impulsions en entrée sont pondérées par les coefficients synaptiques et l'impulsion émise, c'est-à-dire la sortie, obéit à un effet de seuil (pas d'impulsion si l'entrée totale est trop faible).

Un réseau de neurones en général peut donc être représenté par les poids synaptiques (w_i) des différents neurones. Ces poids varient au cours du temps, en fonction des entrées présentées (X_i). Le grand problème étant savoir comment modifier ces poids, pour que le perceptron nous donne la classification ou le résultat espérer.

1.8 Un réseau à architecture évolutive, ART

ART (Adaptive Resonance Theory) est un modèle de réseau de neurones à architecture évolutive développé en 1987 par Carpenter et Grossberg. Dans la plupart des réseaux de neurones, deux étapes sont considérées. La première est la phase d'apprentissage : les poids des connexions sont modifiés selon une règle d'apprentissage La deuxième est la phase d'exécution où les poids ne sont plus modifiés. Avec le réseau ART, ces deux étapes sont réalisées simultanément. Le réseau en phase de test, s'adapte à des entrées inconnues en construisant de nouvelles classes (ajout de neurones) tout en dégradant au minimum les informations déjà mémorisées. Il existe plusieurs versions de réseaux (ART1, ART2, ART3). Le réseau ART1 est un réseau à entrées binaires.

1.8.1 Structure

Le réseau ART1 est formé d'une couche d'entrée qui est aussi la couche de sortie et d'une couche cachée. Le terme de couche cachée est emprunté au réseau multicouche, il souligne le fait que cette couche n'est pas directement observable par l'utilisateur à la différence de l'entrée ou de la sortie. Il n'y a pas de connexion entre les neurones d'entrées. Par contre, la couche cachée est une couche d'activation compétitive, tous les neurones sont reliés les uns aux autres par des connexions inhibitrices de poids fixes. Chaque neurone de la couche d'entrée est relié à tous les neurones de la couche cachée et, réciproquement, chaque neurone de la couche cachée est relié à tous les neurones de la couche de sortie. A chaque connexion est associé un poids.

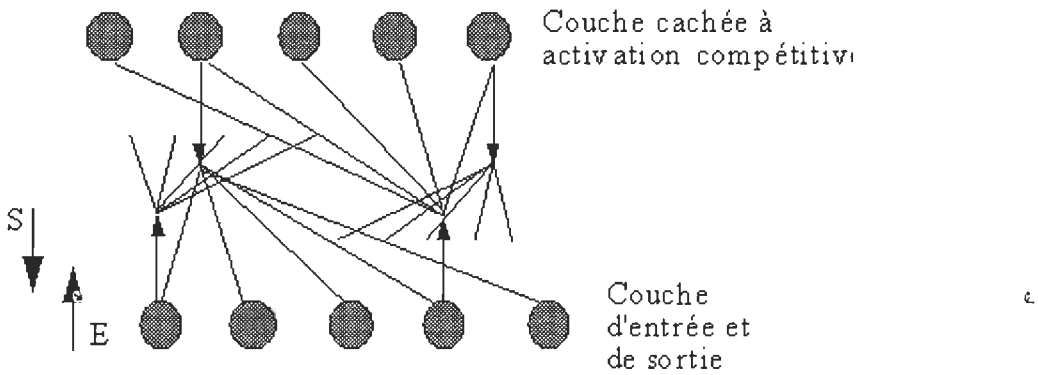

Figure 1.4 - Architecture du réseau ART 1.

La couche d'entrée est aussi celle de sortie. Tous les neurones de la couche d'entrée sont reliés à tous les neurones de la couche cachée et tous les neurones de la couche cachée à chacun de ceux de la couche de sortie. Il n'y a pas de relation entre les neurones d'entrée alors que la couche cachée est à activation compétitive.

1.8.2 Fonctionnement / Apprentissage.

La figure 1.4 (voir [Claude Touzet 1992]) montre un vecteur d'entrée E (1, 0, 1,1) soumis au réseau.

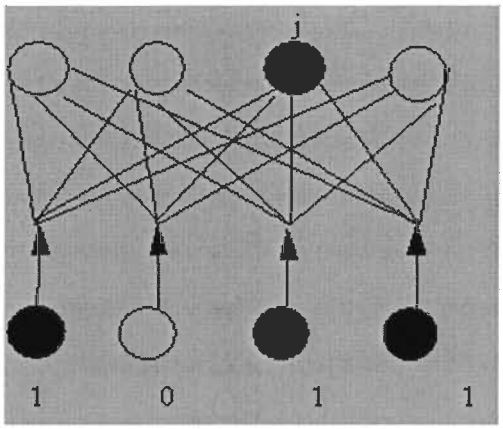

Figure 1.5 - Présentation du vecteur d'entrée E, un neurone gagnant j est sélectionné.

A cette entrée correspond, après compétition entre les neurones de la couche cachée, un unique neurone j gagnant. Ce gagnant est considéré par le réseau comme le plus représentatif du vecteur d'entrée E.

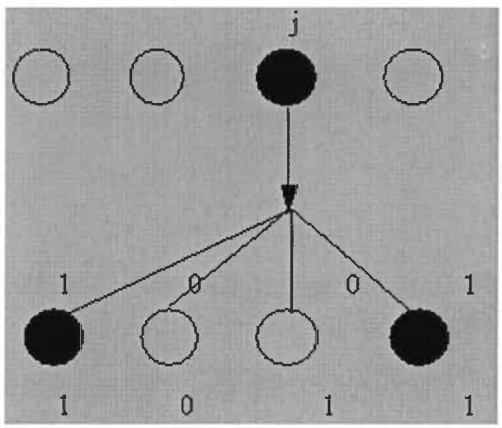

Figure 1.6 - Tentative d'unification entre S (retour du neurone j) et E.

Le neurone j génère en retour sur la couche de sortie un vecteur S (1, 0, 0,1) binaire (seuillage). S est ensuite comparé au vecteur d'entrée E. Si la différence est supérieure à un seuil fixé pour le réseau, le neurone gagnant j est inhibé.

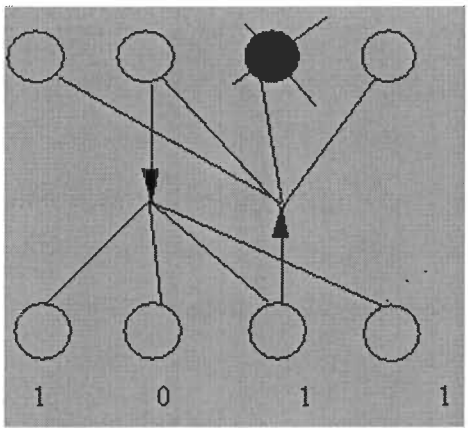

Figure 1.7 - Echec : suppression du neurone gagnant, présentation de E.

Le processus reprend avec les neurones de la couche cachée moins le neurone gagnant de l'étape précédente.

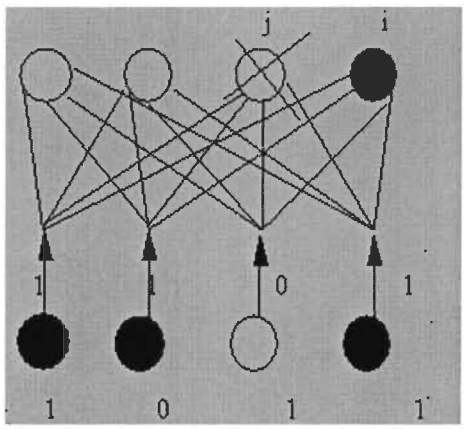

Figure 1.8 - Unification : le neurone i est un représentant de la classe du vecteur d'entrée E.

Le neurone i génère en retour sur la couche de sortie un vecteur S binaire (seuillage). S est ensuite comparé au vecteur d'entrée E. Si la différence est inférieure au seuil fixé pour le réseau, le neurone gagnant est considéré comme représentant de la classe du vecteur d'entrée. Dans ce cas, la modification des poids des connexions du neurone gagnant a pour effet de consolider ses liens d'activation avec l'entrée E; en fait l'adéquation entre ce vecteur d'entrée et cette classe est améliorée.

Si tous les neurones cachés sont passés en revue sans qu'aucun ne corresponde à E, un nouveau neurone caché est ajouté, qui est initialisé comme représentant de la classe du vecteur d'entrée E.

1.8.3 Algorithme

Ici, l'apprentissage consiste tout autant dans la détermination des poids que de la valeur du seuil d'unification ß.

1/ Initialisation des poids aléatoirement entre 0 et 1 et choix d'un seuil d'unification ß.

2/ Présentation d'un vecteur d'entrée El appartenant à la base d'apprentissage.

3/ Calcul du neurone gagnant sur la couche cachée Nj.

4/ Génération en retour d'un vecteur de sortie Sj issu de ce seul neurone Nj. Sj a été seuillé afin de le rendre binaire.

5/ Tentative d'unification entre Sj et El. Soit |S| est la norme de Sj égale au nombre de composantes à 1, par exemple | (1, 0, 1, 1)| = 3.

Si |Sj| / |El| = ß, l'unification est réalisée. Il faut modifier les poids : étape 7.

6/ Sinon |Sj| / |El| < ß, le neurone gagnant Nj est inhibé.

S'il y a encore des neurones non inhibés sur la couche cachée alors retour à l'étape 3.

Sinon un nouveau neurone caché est créé, initialisé comme représentant de la classe correspondant à la forme d'entrée El en utilisant la loi de modification des poids de l'étape 7.

7/ Modification des poids

Couche des poids montants :

h neurone de la couche d'entrée, j neurone gagnant de la couche cachée.

wjh = 1 / |Sj| si le neurone h est actif (valeur 1),

wjh = 0 sinon (valeur 0).

Couche des poids descendants:

j neurone gagnant de la couche cachée, k neurone de la couche de sortie.

wkj = 1 si le neurone k est actif,

wkj = 0 sinon.

Retour à l'étape 2.

8/ Quand le passage de tous les exemples de la base d'apprentissage n'occasionne plus aucun ajout de neurone, il faut mesurer les performances : contrôler le nombre et la qualité des

classes construites. Si le nombre est trop faible, retour à l'étape 1 avec une augmentation de la valeur de ß. Si ce nombre est trop élevé, retour à l'étape 1 en diminuant la valeur de ß.

La valeur du seuil contrôle le degré d'unification recherché entre les formes à classer et les prototypes des classes. Plus la valeur du seuil est grande, meilleure est l'adéquation recherchée. La valeur du seuil doit être choisie entre 0 et 1. Le neurone i est rattaché à une classe dont le prototype générique à priori ne correspond précisément à aucune des formes de la base d'apprentissage. L'unification est réalisée lorsque le nombre d'entrées à 1 est comparable avec le nombre de retours à 1 (coactivation statistique).

1.9 L'utilisation des réseaux de neurones dans le traitement textuelle.

[Carpenter 1991] et [Grossberg 1992] ont proposé une théorie pour modéliser l'apprentissage reposant sur l'auto organisation des connaissances en structures qui tend à résoudre le délicat dilemme stabilité-plasticité, la plasticité spécifiant la capacité du système à appréhender des informations nouvelles, et la stabilité, sa capacité à organiser les informations connues en structures stables. Cette théorie a donné naissance à plusieurs familles de modèles : ART1, ART2, *fuzzy* ART, ARTmap et *fuzzy* ARTmap. Ces modèles, comme les cartes auto-organisatrices de [Kohonen 1982], appartiennent aux réseaux de neurones à apprentissage non supervisé, dont les poids des interconnexions codent les prototypes des classes. Le modèle ART1 travaille avec des données binaires, ce qui le rend spécialement utile pour des tâches de classification textuelles (utilisé dans nos tests). Le nombre total de classes obtenu dépend du paramètre de vigilance p compris entre 0 et 1, fixé par l'opérateur. Plus p est proche de 1, plus les classes seront sélectives (comprendront moins d'éléments) et leur nombre important. Alors que pour des faibles valeurs de p le nombre de classes sera faible, chaque classe comportant un grand nombre d'éléments.

Chapitre 2.

Treillis de Galois.

2.1 Rappels mathématique

Une autre vision de la classification existe avec les treillis de Galois. Des définitions utiles sur le fondement mathématique des treillis de Galois sont nécessaires pour la bonne compréhension de ce chapitre.

Définition 2.1 d'une Relation Binaire.

Soit E un ensemble. ***Une relation binaire R sur E*** *est définie par la donnée d'une partie G de E×E.*

Si (x, y) ∈ G alors x R y (on lit : x est en relation avec y).

G est appelé le graphe de la relation R.

Propriétés 2.1 d'une relation binaire.

Soient E un ensemble et R une relation binaire sur E.

- R est dite réflexive si et seulement si $\forall x \in E, x R x$.
- R est dite symétrique si et seulement si $\forall x, y \in E, x R y \Rightarrow y R x$.
- R est dite anti-symétrique si et seulement si $\forall x, y \in E, x R y$ et $y R x \Rightarrow x = y$.
- R est dite transitive si et seulement si $\forall x, y, z \in E, x R y$ et $y R z \Rightarrow x R z$.

Définition 2.2 d'une Relation d'ordre.

Une relation binaire R entre élément d'un ensemble E est une relation d'ordre si et seulement si elle est réflexive, anti-symétrique et transitive.

Définition 2.3 d'un Ensemble ordonné.

Soit R une relation d'ordre, on dit que (E, R) (E est muni d'une relation d'ordre R) est un ensemble ordonné. On note souvent cet ordre par le symbole $\leq$.

Définition 2.4 Comparable.

Soit $(E, \leq)$ un ensemble ordonné, deux éléments x et y de E sont dits comparables si on a : $x \leq y$ ou $y \leq x$.

Définition 2.5 Ensemble Totalement Ordonné.

Soit $(E, \leq)$ un ensemble ordonné, on dit que $(E, \leq)$ est totalement ordonné si tous les éléments de $(E, \leq)$ sont comparable. On dit que $\leq$ est une relation d'ordre total.

Définition 2.6 Ensemble Partiellement Ordonné.

Soit $(E, \leq)$ un ensemble ordonné, on dit que $(E, \leq)$ est partiellement ordonné s'il existe au moins un couple d'éléments de E non comparable. On dit que $\leq$ est une relation d'ordre partiel.

Définition 2.7 Majorant, minorant d'une partie A d'un ensemble E ordonnée.

Soit $(E, \leq)$ un ensemble ordonné et A une partie de E. un élément de $a \in E$ est un majorant (resp. minorant) de A si et seulement si $\forall x \in A$ on a $x \leq a$ (resp. $a \leq x$).

Définition 2.8 Plus Grand, Plus Petit élément d'un ensemble ordonné.

Soit $(E, \leq)$ un ensemble ordonné S'il existe dans E un élément supérieur (resp. inférieur) á tous les éléments de E, on l'appelle le plus grand (resp. le plus petit) élément de E.

Propriété 2.2 :

Soit $(E, \leq)$ un ensemble ordonné, si le plus grand (resp. plus petit) élément existe, il est unique.

Démonstration :

Soit m le plus grand élément de E alors $\forall x \in E\ x \leq m$.

Soit m' un autre élément de E tel que : $\forall x \in E\ x \leq m'$ C'est-à-dire (càd) il existe un autre plus grand élément. On a, en particulier $m' \leq m$ et $m \leq m'$ donc $m = m'$. (D'après anti-symétrie de la relation d'ordre). De même pour le plus petit élément.

Définition 2.9 Borne Supérieur, Borne Inférieur d'une partie dans un ensemble ordonné.

Soit $(E, \leq)$ un ensemble ordonné et A une partie de E. Le plus petit majorant (resp. plus grand minorant) de A s'il existe est appelé borne supérieur (resp. borne inférieur) de A.

Remarque 2.1 :

La borne supérieure et la borne inférieure si elles existent, sont uniques et elles peuvent appartenir à E.

2.2 Diagramme de Hasse et Treillis.

Dans cette partie nous allons essayer de donnes une définition mathématique d'un treillis et ces propriétés.

2.2.1 Diagramme de Hasse.

Le diagramme de Hasse c'est une représentation graphe de tous les éléments d'un ensemble ordonnes, tous en respectant la définition mathématique si dessous.

Définition 2.10 Diagramme de Hasse.

Le diagramme d'un ensemble ordonné$(E, \leq)$ est un graphe dans le lequel il existe une arête entre deux élément $a, b \in E$ si et seulement si les conditions suivantes sont vérifier :

- *Ils sont distincts : $a \neq b$.*
- *Ils sont comparables : on peut supposer $a \leq b$.*
- *Il n'existe pas $x \in E / x \neq a$ et $x \neq b$, $a \leq x \leq b$.*

2.2.2 Treillis.

Définition 2.11 Treillis.

Soit $(E, \leq)$ un ensemble ordonner. On dira que cet ensemble ordonné est un treillis si $\forall x, y \in E$ le sous ensemble $\{x, y\}$ de E possède une borne inférieure et une borne supérieure.

Exemples 2.1 :

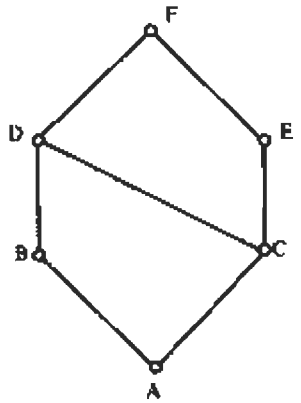

Figure 2.1 - Exemple d'un treillis.

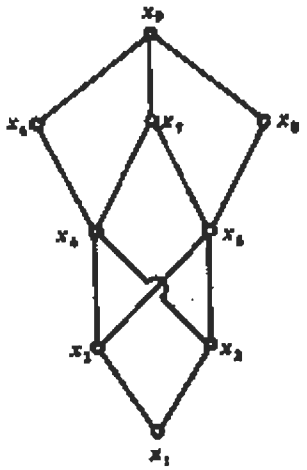

Figure 2.2 - Exemple d'un non treillis.

(Voir [A.kaufman et G.Boulaye 1978] Théorie Des Treillis en vue des applications)

Figure 2.1.

Cet exemple est un treillis car il vérifie les propriétés d'un treillis.

Figure 2.2.

Considérons la paire d'éléments $\{x_4, x_5\}$. Le sous-ensemble M de tous les minorants de $\{x_4, x_5\}$ est M= $\{x_1, x_2, x_3\}$ ce sous-ensemble M ne possède pas un plus grand élément. Donc $\{x_4, x_5\}$ n'a pas de borne inférieure et l'ensemble ordonné n'est pas un treillis.

Définition 2.12 Une Application.

Soient E et F deux ensembles. Une correspondance f de E dans F est dite application si $\forall x \in E$, *il existe un et un seul* $y \in F$ *tel que* $f(x)=y$.

Définition 2.13 Une Application Bijective.

Soit f une application d'un ensemble E dans un ensemble F. f bijective si $\forall y \in F$, *il existe un et un seul* $x \in E$ *tel que* $f(x)=y$.

Définition 2.14 Une Lois de Composition Interne.

Soit E un ensemble, une lois de composition interne notée (*) sur E est une application de $E \times E$ *dans E.*

$$C\grave{a}d \ \forall x, y \in E \ alors \ x * y \in E.$$

2.2.3 Treillis en tant que structure algébrique.

Utilisation de cette deuxième définition de treillis et le théorème d'isomorphisme, facilite quelques démonstrations des théorèmes que nous verrons par la suite.

Théorème 2.1 : Treillis Comme Structure Algébrique.

Soit E un ensemble muni de deux lois de composition interne * et *' et vérifiant les propriétés suivantes :

$\forall x, y, z \in E$:

A1. $x * y = y * x$.

A2. $x *' y = y *' x$.

A3. $x *(x * z) = (x * y)*z$.

A4. $x *'(y *' z) = (x *' y)*'z$.

A5. $x * x = x$.

A6. $y * y = y$.

A7. $x * (x *' y) = x$.

A8. $x *'(x * y) = x$.

Alors E muni de la relation $R \subset E \times E$ définie par $\forall x, y \in E$ $x * y = x$ (ce qui par hypothèse veut dire x inférieur à y) est un treillis.

Remarque 2.2 :

On peut écrire la relation R par l'une ou l'autre des relations suivant.

$$\text{Càd } (x * y = x) \quad \Leftrightarrow \quad (x *' y = y).$$

Démontrons :

D'abord :

$(x * y = x) \Rightarrow (x *' y = y)$

Supposons que l'on ait : $x * y = x$ et montrons alors que $x *' y = y$.

Donc $x *' y = (x * y) *' y = y *' (x * y)$ d'après A2

$= y *'(y * x)$ d'après A1

$= y$ d'après A8

Supposons que l'on ait : $x *' y = y$ et montrons alors que $x * y = x$.

Donc $x * y = x * (x *' y) = x$ d'après A7.

Démonstration :

Soient $x, y \in R$ est définie tel que $x*y=x$ càd (x inférieur à y).

1) Premièrement montrons que R ainsi définie est une relation d'ordre.

Réflexive : Si on prend x=y on a : $x*x=x$ par définition.

Transitive : Montrons que si $x*y=x$ et $y*z=y$ alors $x*z=x$.

Donc $x*z=(x*y)*z=x*(y*z)$ d'après A3.

$=x*y=x$.

Anti-symétrie : Montrons que si $x*y=x$ et $y*x=y$ alors x=y.

D'après A1 $x*y=y*x$ alors x=y.

On déduit que R est une relation d'ordre.

2) Montrons que $\forall x, y \in E$ {x, y} admet une borne inférieur $x*y$.

Montrons que c'est un minorant de {x} et {y}.

Càd $(x*y)*x=x*(x*y)=(x*x)*y=x*y$.

Donc $x*y$ est un minorant de {x}.

De même pour le {y}.

Montrons maintenant que $x*y$ est la borne inférieur de {x, y}.

Soit m un autre minorant

Alors m*x=m et m*y=m.

Donc m(x*y)= (m*x)*y d'après A3.

=m*y=m.

Donc x*y le plus grand des minorants càd borne inférieur.

3) Montrons que $\forall$ x, y $\in$ E {x, y} admet une borne supérieur x*'y.

Montrons que c'est un majorant de {x} et {y}.

Càd x*(y*'x)=x. d'après A7.

Et y*(x*'y)=y. d'après A7.

Donc x*'y est un majorant de {x} et {y}.

Montrons maintenant que x*'y est la borne supérieur de {x,y}.

Soit M un autre majorant.

Alors x*'M=M et y*'M=M. d'après la remarque.

Donc (x*'y)*'M= x *' (y *' M) d'après A4.

= x *' M

= M.

Donc x *' y le plus petit des majorant càd borne supérieur.

L'ensemble (E, R) est un treillis.

Définition 2.15 : Isomorphisme.

Soient (E, $\leq$) et (F, $\leq$') deux ensembles ordonnés. On dit qu'ils sont isomorphes s'il existe une bijection f de E vers F telle que :

$\forall x, y \in E, x \leq y$ *si et seulement si* $f(x) \leq' f(y)$.

Théorème 2.2 :

Soient (E, $\leq$) et (F, $\leq$') deux ensemble ordonnés et f une application bijective entre E et F. alors (E, $\leq$) est un treillis $\Leftrightarrow$ (F, $\leq$') est un treillis.

Démonstration :

$\Rightarrow$ (De la même manière en montre l'autre sens).

Soit f E $\longrightarrow$ F une application bijective.

Supposons que $(E, \leq)$ est un treillis montrons que $(F, \leq')$ est un treillis.

Soient x, y$\in$ F montrons que l'ensemble {x, y} admet une borne supérieur dans F.

Soient $a = f^{-1}(x)$, $b=f^{-1}(y) \in E$ alors admet une borne supérieur c.

Donc $a = f^{-1}(x) \leq c$ et $b=f^{-1}(y) \leq c$ on applique l'isomorphisme de f alors $x \leq' f(c)$ et $y \leq' f(c)$.

Ce qui veut dire que f(c) est un majorant de x et y.

Maintenant montrons que f(c) est la borne supérieure.

Supposons que $\exists z \in F$ tel que $z \leq' f(c)$ tel que $x \leq' z$ et $y \leq' z$.

C'est-à-dire $f^{-1}(z) \leq c$, $a=f^{-1}(x) \leq f^{-1}(z)$ et $b=f^{-1}(y) \leq f^{-1}(z)$ ce qui est absurde car la borne supérieur pour a et b c'est c.

Ce que on vient de faire pour la borne supérieure on peut le faire pour la borne inférieur.

Donc l'ensemble $(F, \leq')$ est un treillis.

2.3 La correspondance de Galois.

La correspondance de Galois est la base de la construction des éléments de treillis de Galois. Ces éléments, nous allons les appeler par la suite les couples complets du treillis ou les concepts [wille 1982].

Définition 2.16 : L'ensemble des Partitions.

P(E) l'ensemble des partitions de E. càd P(E) c'est l'ensemble de tous les sous ensembles de E.

Définition 2.17 : Correspondance de Galois.

Soient E et F deux ensemble et R une relation binaire sur $E \times F$. soient f et g deux correspondance définies par :

1) $f : P(E) \longrightarrow P(F)$

$f(X)=\{y \in F / \forall x \in X, x R y\}$

$= \cap_{x \in X}\{y \in F / x R y\} = \cap_{x \in X} F_x$.

2) g : P(F) $\longrightarrow$ *P(E)*

$g(X)=\{x\in E / \forall x\in X, x R y\}$

$=\cap_{y\in Y}\{y\in E/ x R y\}=\cap_{y\in Y} E_y$.

Le couple (f, g) est appelé correspondance de Galois.

A noter que par définition, f (∅) = E et g (∅) =F.

Propriété 2.3 :

Les applications f et g de la correspondance de Galois sont monotones décroissantes pour l'ordre partiel (inclusion) défini dans P(E) et P (F).

càd $\forall$ X, X' éléments de P(E)et Y,Y' éléments de P(F), on a

(1) $X\subset X' \Rightarrow f(X') \subset f(X)$.

(2) $Y\subset Y' \Rightarrow g(Y') \subset g(Y)$.

Preuve :

(1) Supposons que $X \subset X'$ montrons que $f(X') \subset f(X)$

Soit $f(x) \in f(X')$ avec $x \in X'$ donc $f(x) \in \cap_{x\in X'}\{y\in F/ x R y\} \subset \cap_{x\in X}\{y\in F/ x R y\}$ car $X\subset X'$.

De même pour le numéro (2).

Notation : On notera les compositions de ces applications par : $h=g\circ f$ et $h'= f\circ g$.

Définition 2.18 : Fermeture.

Les applications h et h' sont les ***fermetures*** *de la relation binaire R sur E×F, et on dira qu'un sous-ensemble X de E (resp. Y de F) est* ***un fermé de la relation binaire R*** *si et seulement si h(X) = X (resp. h'(Y)=Y).*

Définition 2.19 : Un Couple Complet.

Soit (X, Y) ∈ E×F est couple complet si et seulement si h(X)=X ou h(Y)=Y.

Remarque 2.3 :

Le terme **complet** introduit par [Godin 1995] représente le fait que pour un couple (X, f(X)), s'il y a des éléments de E qui ne sont pas dans X mais qui sont reliés à tous les éléments de f(X), alors ils doivent être ajoutés à X pour que le couple soit complet.

À cause de la symétrie entre X et $Y\in$ F dans la définition, le même raisonnement s'applique à Y. Cette idée de complétude des couples est formalisée par la notion mathématique de fermeture dans les ensembles ordonnés.

Appellation : on appelle aussi le couple complet (X, f(X)) un concept [wille 1982].

Propriété 2.4 :

Les applications h : P(E) $\rightarrow$P(E) et h' : P (F) $\rightarrow$P (F) ont les propriété suivantes :

- Monotones croissante.

  $X\subset X' \Rightarrow f(X') \subset f(X) \Rightarrow g(f(X)) \subset g(f(X')) \Rightarrow h(X) \subset h(X')$.

  $Y\subset Y' \Rightarrow f(Y') \subset f(Y) \Rightarrow g(f(Y)) \subset g(f(Y')) \Rightarrow h(Y) \subset h(Y')$.

- Extensives.

  $X \subset h(X)$

  $Y \subset h'(Y)$

  Soit $x_0\in X$, montrons que $x_0\in h(X)$ càd $x_0\in g(f(X))$

  $f(X)= \cap_{x\in X} \{y\in F / x\ R\ y\}$ donc $\forall y\in f(X)$ et $\forall x\in X$ on a x R y

  en particulier $x_0 \in X$ càd que $x_0\in g(f(X))= \cap_{y\in f(X)}\{x\in E/ x\ R\ y\}$

  de même pour $Y \subset h'(Y)$.

- Idempotentes.

  $h\circ h=h$

  $h'\circ h'=h'$

  soit $X \in P(E)$ et montrons que $h\circ h(X) = h(X)$.

  on va utiliser une preuve géométrique :

soit la figure 1 et $X=\{x_3,x_4,x_5\}$ et la partie f(X) l'image de X par f, est l'intersection des segments x_1, x_2 et x_3. h(X) c'est la partie X plus x_2 et x_6 car x_2 et x_6 ont une relation avec f(X).

nous voyons d'après la figure 2.3 que l'image f (h(X))=f(X) et h (h(X))=h(X).

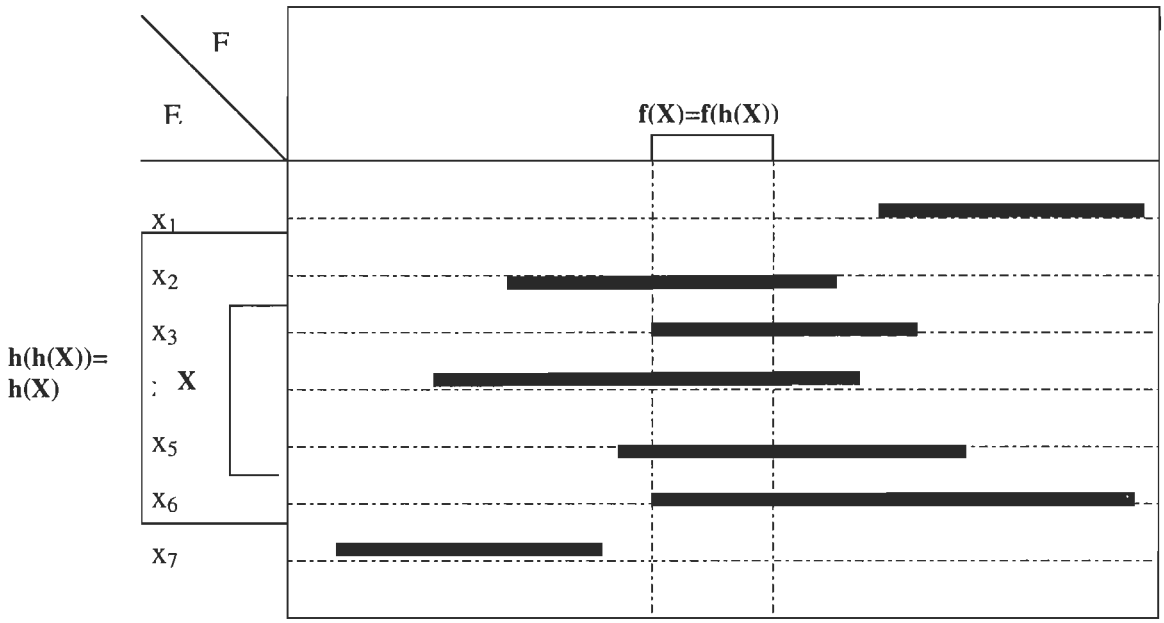

Figure 2.3 - Illustration de l'idempotence de l'application h.

2.4 Définition mathématique du treillis de Galois.

Notation : notons par C_E (resp. C_F) L'ensemble des fermés E (resp. de F)

Nous pouvons munir C_E (resp. C_F) de deux lois de composition * et *'. Définies comme suite :

$\forall$ A, B $\in$ C_E (resp. C_F).

- A * B= A $\cap$ B.
- A *'B=h (A $\cup$ B) (resp. h' (A $\cup$ B)).

Propriété 2.5 :

Les deux lois * et *' mentionnée si dessus, définissent deux lois de composition interne sur L'ensemble C_E (resp. C_F).

Démonstration :

Soient A et B $\in C_E$ alors h (A)=A et h (B)=B

Montrons que * et une lois interne pour C_E, ce qui reviens de montrer que h $(A \cap B)=A \cap B$.

- On a $A \cap B \subset A$ et $A \cap B \subset B$. donc d'après la Monotonie croissante de h
  h $(A \cap B) \subset h(A)$ et h $(A \cap B) \subset h(B)$ donc h $(A \cap B) \subset h(A) \cap h(B)= A \cap B$.

- On a $A \cap B \subset$ h $(A \cap B)$ d'après l'extensivité de h.

donc h $(A \cap B) =A \cap B$.

càd A * $B=A \cap B \in C_E$ donc * est une lois interne pour C_E.

Montrons que *' et une lois interne pour C_F.

ce qui reviens de montrer que $h \circ h (A \cup B)=h (A \cup B)$.

On a d'après Idempotence de h que $h \circ h (A \cup B)=h (A \cup B)$

càd $h(A \cup B) \in C_E$ donc *' est une lois interne pour C_E.

De même pour C_F.

Théorème 2.3 :

Soit l'ensemble C_E muni de deux lois de composition interne * et *'. Les deux lois vérifient les propriétés suivantes :

$\forall$ A, B et $C \in C_E$:

A1. A * B = B *A.

A2. A *' B = B *' A.

A3. A * (B * C) = (A * B) *C.

A4. A *' (B *' C) = (A *' B) *'C.

A5. A * A = A.

A6. B *' B = B.

A7. A * (A *' B) = A.

A8. $A *' (A * B) = A$.

Démonstration :

A1) $A * B = A \cap B = B \cap A = B * A$.

A2) $A *' B = h(A \cup B) = h(B \cup A) = B *' A$.

A3) évidant.

A4) $A *' (B *' C) = (A *' B) *' C$ càd $h(A \cup h(B \cup C)) = h(h(A \cup B) \cup C)$.

Démonstration par double inclusion.

Montrons que $h(A \cup h(B \cup C)) \subset h(h(A \cup B) \cup C)$

$A \subset A \cup B \subset h(A \cup B)$ d'après l'extensivités de h. Et $A \subset h(A \cup B) \cup C$.

D'après l'extensivités de h, on a donc $A \subset h(h(A \cup B) \cup C)$.

$B \subset A \cup B \subset h(A \cup B)$ d'après l'extensivités de h.

Donc $B \cup C \subset h(A \cup B) \cup C$.

D'après la monotonie de h, on a $h(B \cup C) \subset h(h(A \cup B) \cup C)$.

Donc $A \cup h(B \cup C) \subset h(h(A \cup B) \cup C)$ D'après la monotonie et l'idempotence de h.

$h(A \cup h(B \cup C)) \subset h(h(A \cup B) \cup C)$ ce qui fallait démontrer.

De la même manière on peut montrer que $h(h(A \cup B) \cup C) \subset h(A \cup h(B \cup C))$

A5) évidant.

A6) $B *' B = h(B \cup B) = h(B) = B$ car $B \in C_E$

A7) $A * (A *' B) = A \cap h(A \cup B) = A$ car $A \subset A \cup B \subset h(A \cup B)$.

A8) $A *' (A * B) = h(A \cup (A \cap B)) = h(A)$ car $A \cap B \subset A$.

Théorème 2.4:

C_E muni de la relation d'ordre définie par :

$\forall A, B \in C_E$ $A \cap B = A$ si et seulement si $A \subset B$ est un treillis.

Preuve :

D'après le théorème (2.1) il est clair que C_E est un treillis.

Remarque 2.4 : de même on peut démontrer que C_F est un treillis.

Notation :

On note $C = C_E \times C_F$ l'ensemble des couples complets qui sont définies comme suite : Soit $X \in C_E$ et $Y \in C_F$. les couples sont sous la forme (X, f(X)) et (Y, g(Y)).

Définition 2.20 :

En munissant l'ensemble C d'une relation d'ordre partiel $\leq$ tel que pour tous couple d'éléments $C_1=(X_1, Y_1)$ et $C_2(X_2, Y_2)$ de C on a :

$$C_1 \leq C_2 \Leftrightarrow X_2 \subset X_1 \text{ et } Y_1 \subset Y_2$$

Théorème 2.5 :

L'ensemble $(C = C_E \times C_F, \leq)$ et l'ensemble $(C_E, \subset)$ sont isomorphes bijective.

Preuve :

Soit $f\ C \rightarrow C_E$ définie comme suite $\forall\ C_1 = (X, Y) \in C$ on a $f(C_1)=X$.

L'application f est bijective par construction.

Montrons que $\forall\ C_1=(X_1, Y_1)$ et $C_2(X_2, Y_2)$ de C tel que $C_1 \leq C_2 \Leftrightarrow X_2 \subset X_1$

$\Rightarrow$

D'après la définition de la relation d'ordre définie sur C on a $\forall\ C_1=(X_1, Y_1)$ et $C_2(X_2, Y_2)$ de C tel que $C_1 \leq C_2$ on a : $X2 \subset X_1$.

$\Leftarrow$

Supposons que $X2 \subset X_1$ on sait que $f(X_1)=Y_1$ et $f(X_2)=Y_2$ donc $Y_2 \subset Y_1$ car f monotone décroissante. D'où $C_1=(X_1, Y_1) \leq C_2(X_2, Y_2)$.

Théorème 2.6 :

L'ensemble $(C = C_E \times C_F, \leq)$ est un treillis.

Preuve :

D'après le théorème (2.2) sur l'isomorphisme et le théorème (2.4) on en déduit que C est un treillis.

Appellation : On appel l'ensemble ($C = C_E \times C_F, \leq$) treillis de Galois de la relation binaire R définie entre E et F.

2.5 Treillis de Galois.

La notion de treillis de Galois d'une relation (ou treillis de concepts) est à la base d'une famille de méthodes de classification conceptuelle. Introduite par Barbut et Monjardet [Barbut 1970], cette approche a été popularisée par Wille qui a utilisé la notion de treillis de Galois comme base de l'analyse formelle de concepts [Wille 1982].

Wille propose de considérer chaque élément du treillis comme un concept formel et le graphe (diagramme de Hasse) comme une relation de généralisation/spécialisation entre les concepts. Wille a développé un nouveau domaine de recherche appelé : analyse formelle de concepts.

Le treillis est donc perçu comme une hiérarchie de concepts. Chaque concept est une paire composée d'une extension représentant un ensemble des domaines d'informations de l'application et d'une intention représentant les propriétés communes aux domaines d'informations (unités d'informations).

La notion de treillis a été l'objet d'études dans plusieurs domaines. C'est en théorie de graphes et des ensembles qu'on retrouve les premiers travaux formels sur les treillis.

Les treillis font le lien entre l'étude des relations d'ordre et l'étude de certaines structures algébriques. Ils fournissent les éléments permettant une conception algébrique des différents calculs logiques, classiques et non classiques. Ils interviennent en topologie, et géométrie et en théorie des anneaux.

Dans le cadre du dépistage d'information en sciences sociales (analyse de questionnaire), Barbut et Monjardet [Barbut 1970] introduisent le treillis d'une correspondance de Galois entre deux ensembles ou treillis de galois. Godin et al. [Godin 1995] organisent une base de domaines d'informations sous forme de treillis dans une application de recherche d'information à l'aide de mots-clés.

Exemple 2.2 :

Cette sous-section donne un exemple concret de la définition de base de treillis de Galois d'une relation binaire.

Soit deux ensembles finis, E et F, et une relation binaire, $R \subseteq E \times F$ (Tableau 2.1), entre ces deux ensembles, on peut représenter par un treillis de Galois les regroupements naturels des éléments de E et de F par rapport à la relation R (Figure 2.5). Chaque élément du treillis est un couple (appelé concept formel par Wille) noté, (X, X').

Quoique la définition soit totalement symétrique, dans les applications, on associe habituellement les éléments de E à des domaines d'informations[1] et ceux de F à des unités d'informations.

Définition 2.21 :

Diagramme de Hasse est une représentation graphique de l'ensemble de tous les couples complets, il forme une structure de base pour supporter une interface de navigation permettant à l'utilisateur de graduellement élargir ou spécialiser sa requête en terme des sous-ensembles de domaine d'information et de unités d'informations présents dans le treillis. (Voir figure 2.4).

R	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9
D_1	1	0	1	0	0	1	0	1	0
D_2	1	0	1	0	0	0	1	0	1
D_3	1	0	0	1	0	0	1	0	1
D_4	0	1	1	0	0	1	0	1	0
D_5	0	1	0	0	1	0	1	0	0

[1] Domaine d'information, dans le cas d'un texte, est une partie de ce texte qui peut être définie soit par une (ou plusieurs) phrase, paragraphe, page ou encore document.

Tableau 2.1 - Un exemple d'une Représentation matricielle de la relation R.

L'ensemble E c'est l'ensemble des domaines d'informations = $\{D_1, D_2, D_3, D_4, D_5\}$.
L'ensemble F c'est l'ensemble des unités d'informations = $\{U_1, U_2, U_3, U_4, U_5, U_6, U_7, U_8, U_9\}$.

La relation R qui est entre les domaines d'information et les unités d'information peut être représentée par le graphe biparti ci-dessous [Alaoui 1993].

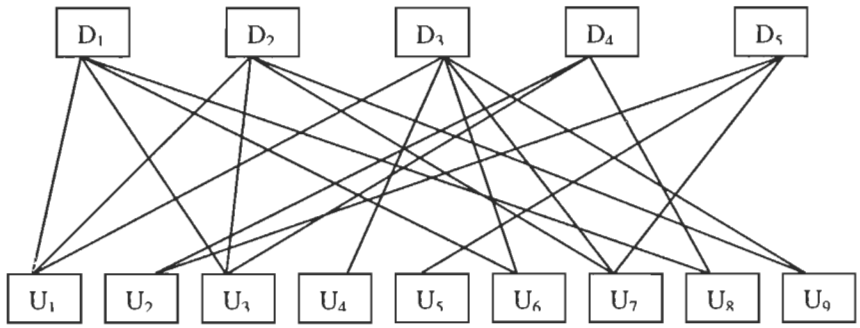

Figure 2.4 - Graphe biparti de la relation R.

Dans cet exemple f ($\{D_1, D_3\}$)= $\{U_1\}$ et g ($\{U_1\}$) = ($\{D_1, D_2, D_3\}$) donc $\{D_1, D_3\}$ n'est pas un fermé. Par contre h ($\{D_1, D_2\}$)= $\{D_1, D_2\}$ donc $\{D_1, D_2\}$ est un fermé et $\{D_1, D_2\} \times \{U_1, U_3\}$ est un couple complet (voir définition 2.19).

Il doit apparaître comme un nœud du diagramme de Hasse du Treillis de Galois de la relation R, dans ce cas le graphe (2.5) est appelé sous graphe complet maximal de $\{D_1, D_2\} \times \{U_1, U_3\}$.

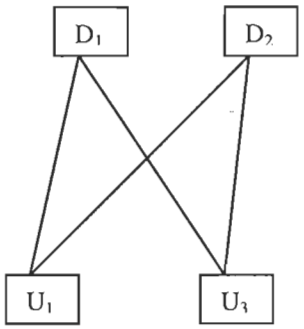

Figure 2.5 - Sous graphe complet maximal de $\{D_1, D_2\} \times \{U_1, U_3\}$.

Vu les petites cardinalités des ensembles E et F, on peut anticiper un peu et engendrer tous les éléments du diagramme de Hasse et ce, en utilisant, par exemple, l'algorithme de Ganter décrit dans la section suivante.

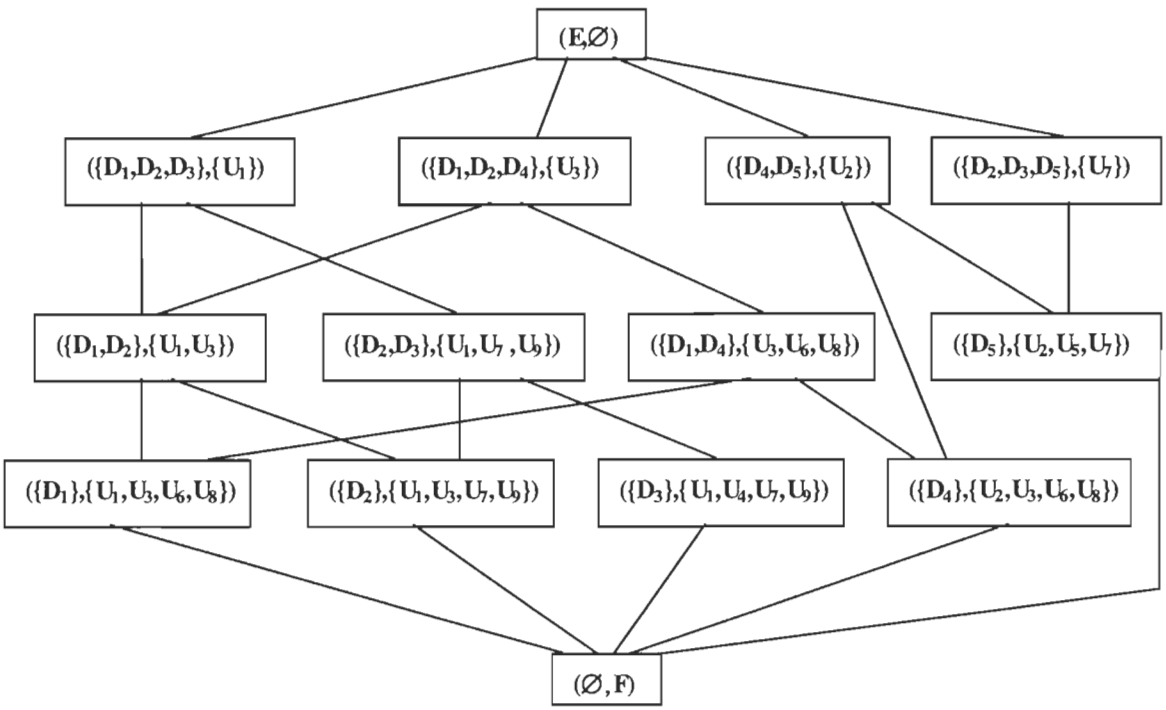

Figure 2.6 - Diagramme de Hasse de la correspondance R.

Nous avons tracé à la figure 2.4 le graphe biparti de la relation R et comme on l'a mentionné, les éléments du diagramme de Hasse sont exactement l'ensemble des sous graphes bipartis complets maximaux. Donc l'élément $\{D_1, D_4\} \times \{U_3, U_6, U_8\}$ correspond au sous graphe biparti complet maximal suivant :

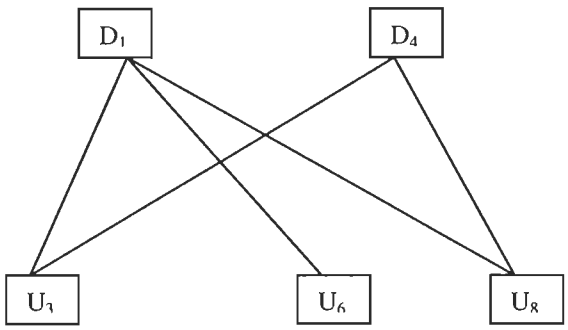

Figure 2.7 - Sous graphe complet maximal de $\{D_1, D_4\} \times \{U_3, U_6, U_8\}$.

Dans un diagramme de Hasse, tel que celui de la figure 2.6, les lignes représente la relation d'ordre exemple le couple $C_1 = \{D_1, D_2\} \times \{U_1, U_3\}$.et le couple $C_2 = \{D_1, D_2, D_3\} \times \{U_1\}$ donc d'après la définition 2.20 $\{D_1, D_2\} \subset \{D_1, D_2, D_3\}$ donc $C_2 \leq C_1$.

D'après la définition mathématique du diagramme de Hasse, nous pouvons mettre une arête entre deux couples complets, si le premier couple plus grand ou plus petit directement du deuxième couple.

2.6 Algorithmes de construction de treillis.

La construction du treillis de concepts implique deux tâches distinctes :

1. la découverte des concepts.
2. la construction de la relation entre ces concepts.

Certains algorithmes abordent ces deux problèmes ensemble, tandis que la majorité vont se spécialiser dans l'une des deux tâches. De plus, certains algorithmes vont effectuer le traitement en général, tandis que d'autres permettent une construction incrémentale [Godin 1991].

[Guénoche 1990] présente quatre algorithmes pour une étude comparative dont notamment celui de [Bordat 1986] qui est le seul utilisant une approche globale et simple pour la construction du diagramme de Hasse du treillis.

Ce tableau résume les algorithmes qui existent dans la littérature et il nous donne une idée sur chaque algorithme.

Algorithmes	Trouver les concepts	Trouver les liens entre les concepts	Traitement global	Traitement incrémental
Chein	Oui	Non	Oui	Non
Ganter	Oui	Non	Oui	Non
Nouris et Raynaud	Oui	Oui	Oui, pour les liens entre les concepts	Oui, pour les concepts
Bordat	Oui	Oui	Oui	Non
Godin	Oui	Oui	Non	Oui

Tableau 2.2 - quelques algorithmes de construction de treillis de concepts. (Les plus connus)

Les algorithmes de Bordat, Godin et de Nourine et Raynaud construisent à la fois les concepts et les liens entre les concepts tandis que ceux de Chein et Ganter sont spécialisés dans la découverte des concepts.

Dans notre mémoire on ne va pas détailler la complexité des algorithmes, notre but étant de construire le diagramme de Hasse. Nous focalisons notre attention sur l'algorithme de Ganter car étant le plus simple, il nous permet de trouver les concepts et les liens entre les concepts simultanément et d'une manière globale.

L'algorithme de ganter calcule les éléments du treillis C_F sans construire les arêtes. Pour obtenir le treillis $C=C_E \times C_F$, il faut calculer chaque fermé de C_F, son image par l'application g, afin d'avoir les couples complets de C. Cependant la méthode de ganter présente un très gros avantage : le faible encombrement mémoire lors de la construction des nœuds du treillis C_F, l'algorithme n'ayant besoin que du dernier fermé trouvé et de la relation binaire pour générer le prochain fermé. On va voir l'utilisation de cette méthode dans le chapitre trois.

2.6.1 Algorithme de Bordat.

Dans [Bordat 1986], Bordat décrit sa méthode de construction des treillis de Galois de façon assez complexe. Cet algorithme construit les concepts du treillis et les liens entre ceux-ci, en partant du concept supérieur. Il insère un à un les concepts du treillis et retrouve leurs concepts inférieurs.

La généralisation d'un concept direct qui aura été déjà inséré dans le treillis implique une recherche de ce concept et la construction d'une arête entre ce concept et le concept père, tandis que pour un concept qui n'a pas été généré préalablement, l'algorithme construira un sous treillis dont ce concept serait le concept supérieur.

[Nguifo 1999] par exemple développe une approche de construction de treillis de concepts approximatifs, basée sur l'algorithme de Bordat et la théorie des sous ensembles flous.

2.6.1.1 Principe de la méthode de Bordat :

Soit $(X, Y) \in C=C_E \times C_F$ on sais que $(C_F, \subset)$ est un ensemble partiellement ordonné par l'inclusion, $Y \in C_F$, cherchons l'ensemble Y' qui contient immédiatement l'ensemble Y.

si on connaît l'ensemble Y' on peut trouver le couple complet (g (Y'), Y') qui est inférieur immédiatement à (X, Y) d'après la définition (2.20).

Question : Comment on peut trouver Y' ?

Pour cela nous allons donner une première définition :

Définition 2.22 :

Soit $(X, Y) \in C=C_E \times C_F$.

Y'est une partie maximale de $g(Y) \times (F/Y) \Leftrightarrow \neg\exists Z \subset F/Y$ *tel que* $Y' \subset Z$ *et* $g(Y') \cap g(Y)=g(Z) \cap g(Y)$.

Théorème:

$\forall (X, Y) \in C=C_E \times C_F$. (X, Y) est le supérieur immédiat de $g(Y') \cap g(Y) \times (Y \cup Y')$ $\Leftrightarrow$ Y' est une partie maximale de $X \times F/Y$.

La démonstration de ce théorème se trouve en annexe dans [Bordat 1986].

2.6.1.2 Description de l'algorithme.

L'algorithme de Bordat consiste à construire le treillis en partant du couple (E,∅) (si il n'a pas un élément de F qui est en relation avec tous les éléments de E).cette élément va se placer au sommet du treillis. On appliquant l'algorithme de Bordat sur ce couple on trouve les couples qui viens immédiatement après, et on fait la même procédure pour chaque concept(X, Y). Ainsi on construit notre treillis de Galois.

Il a deux étapes dans l'algorithme de Bordat. Premier étape consiste à calculer un vecteur vectC, et la deuxième utilise le vectC pour extraire les couples complets qui vient après le concept du départ (X, Y) on appel ces couples la couverture du concept(X, Y).

Calcul du vecteur vectC d'un concept (X, Y) quelconque.

Soit vect$Y=y_1,\ldots,y_p$ tel que $(y_1,\ldots,y_p)$les éléments de F\Y indicés dans l'ordre croissant des colonnes du tableau $X \times F\backslash Y$, c'est-à-dire les cardinaux croissants des $g_x(y_i)=g(y_i) \cap X$. on construit le vecteur vectC de la manière suivante :

- c(i)=0 si $g_x(y_i)$ est strictement inclus dans $g_x(y_j)$ pour j>i .
- c(i)=j si $g_x(y_i)= g_x(y_j)$ avec j>i.
- c(i)=-1 sinon.

Calcul de la couverture de X × Y:

Cette étape consiste à balayer le vecteur vectC de X×Y, de gauche à droite. À chaque composante du vecteur vectC qui est égale à -1 va correspondre un couple complet de la couverture de X×Y. La deuxième partie d'un couple complet de la couverture de X×Y correspondant à la $j^{ième}$ composante de C va être composée des éléments de Y, y_j et de tous les y_i de vectY tel que $g_x(y_i) = g_x(y_j)$.

Exemple de calcul de la couverture d'un couple complet:

Soient E = {D_1, D_2, D_3, D_4, D_5} et F = {U_1, U_2, U_3, U_4, U_5, U_6, U_7, U_8, U_9}. Deux ensembles en relation par R et X×Y = ({D_1, D_2, D_3}, {U_1}) un couple complet du treillis de Galois de la relation R décrite dans le tableau 2.1. Cette relation est décrite partiellement par R' dans le tableau suivant sur X ×F\Y.

R	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9
D_1	0	1	0	0	1	0	1	0
D_2	0	1	0	0	0	1	0	1
D_3	0	0	1	0	0	1	0	1

Tableau 2.3 - Tableau de la relation partielle R' de R sur X ×F\Y

Alors pour le couple complet ({D_1, D_2, D_3}×{U_1}) on a

vectY=($\mathbf{U}_2$,$\mathbf{U}_5$,$\mathbf{U}_4$,$\mathbf{U}_6$,$\mathbf{U}_8$,$\mathbf{U}_3$, $\mathbf{U}_7$, $\mathbf{U}_9$)

vectC = (0, 0, 0, 0, 0, -1, 8,-1)

et la couverture de ({D_1, D_2, D_3}×{T_1}) sera formée de deux couples complets suivants:

- de vectC(6) on obtient Y'= $\{\mathbf{U}_1,\mathbf{U}_3\}$ d'où le couple complet $(\{D_1,D_2\}\times\{\mathbf{U}_1,\mathbf{U}_3\})$.
- de vectC(8) on obtient Y' = $\{\mathbf{U}_1,\mathbf{U}_7,\mathbf{U}_9\}$ d'où le couple complet $(\{D_2,D_3\},\{\mathbf{U}_1,\mathbf{U}_7,\mathbf{U}_9\})$.

Les couples complets $(\{D_1,D_2\}\times\{\mathbf{U}_1,\mathbf{U}_3\})$ et $(\{D_2,D_3\}\times\{\mathbf{U}_1,\mathbf{U}_7,\mathbf{U}_9\})$ de la couverture de $(\{D_1,D_2,D_3\}\times\{\mathbf{U}_1\})$ seront respectivement générés, une deuxième fois par algorithme, lors du calcul de la couverture des couples complets $(\{D_1,D_2,D_4\}\times\{\mathbf{U}_3\})$ et $(\{D_2,D_3,D_5\}\times\{\mathbf{U}_7\})$. Cette deuxième génération permettra juste de dresser les arêtes entre les générateurs et les couples complets de leur couverture, dans le diagramme de Hasse.

2.6.1.3 Algorithme de Bordat.

{Input: la relation binaire)}
{Output: le treillis de Galois)

```
1   L := E x ∅  {Ensemble des noeuds du treillis}
2   A :=∅       {Ensemble des arêtes du treillis}
3   Pour Tout X ×Y∈ L Faire
        vectY :=(y1,...,yp)
        {les éléments F\Y indicés dans l'ordre croissant des gx(yi) }
5       Pour Tout i < j en variant i et j de 1 à p Faire
6           Si gx(yi)⊂ gx(yj) Alors vectC(i) := 0
7           Sinon
8               Si gx(Yi) = gx(Yj) Alors c(i) := j
9               Sinon c(i) :=-1
10              FinSi
11          FinSi
12      FinPourTout
13      Pour i =1 à p Faire
14          vectY' :=∅
```

15 j := i
16 **Tant Que vectC**(j) > 0 **Faire**
17 **vectY'** := **vectY'** ∪ {y_j}
18 k := j; j := **vectC**(j); **vectC**(k) := 0
19 **FinTantQue**
20 **Si vectC** (j)=-1 **Alors**
21 **vectY'** := **vectY'** ∪ {y_j}
{Faire le lien entre le générateur et l'élément généré}
22 A:= A ∪ {(X × Y, g(**vectY'**) ∩ g(**vectY**) × (**vectY** ∪ **vectY'**)) }
23 **Si** g(**vectY'**) ∩ g(**vectY**) × (**vectY** ∪ **vectY'**) ∉ L **Alors**
24 L:= L ∪ {g(**vectY'**) ∩ g(**vectY**) × (**vectY** ∪ **vectY'**)
25 **FinSi**
26 **vectC(j) .= 0**
27 **FinSi**
28 **FinPour**
29 **FinPourTout**

Trace de algorithme de Bordat :

La trace sera faite sur la correspondance de l'exemple 2.1. Les couples complets seront numérotés dans l'ordre de leur génération par l'algorithme. À chaque élément X×Y nous allons faire correspondre son sous tableau X×F\Y et les parties maximales Y' ∪ Y qu'il génère. Les éléments ainsi générés seront traités ultérieurement dans la trace. Les arêtes du graphe de Hasse seront construites progressivement entre les générateurs et les éléments qu'ils génèrent.

1 : E×∅

R	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9
D_1	1	0	1	0	0	1	0	1	0
D_2	1	0	1	0	0	0	1	0	1
D_3	1	0	0	1	0	0	1	0	1
D_4	0	1	1	0	0	1	0	1	0
D_5	0	1	0	0	1	0	1	0	0

2: $\{U_1\}$ 3: $\{U_2\}$ 4: $\{U_3\}$ 5: $\{U_7\}$

2: $\{D_1,D_2, D_3\}\times\{U_1\}$

R	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9
D_1	0	1	0	0	1	0	1	0
D_2	0	1	0	0	0	1	0	1
D_3	0	0	1	0	0	1	0	1

6: $\{U_1,U_3\}$ 7 : $\{U_1, U_7\ U_9\}$

3: $\{D_4, D_5\}\times\{U_2\}$

R	U_1	U_3	U_4	U_5	U_6	U_7	U_8	U_9
D_4	0	1	0	0	1	0	1	0
D_5	0	0	0	1	0	1	0	0

8: $\{U_2, U_3, U_6, U_8\}$ 9: $\{U_2, U_5, U_7\}$

4: $\{D_1, D_2, D_4\}\times\{U_3\}$

R	U_1	U_2	U_4	U_5	U_6	U_7	U_8	U_9
D_1	1	0	0	0	1	0	1	0
D_2	1	0	0	0	0	1	0	1
D_4	0	1	0	0	1	0	1	0

6: $\{U_1, U_3\}$ 10: $\{U_3, U_6, U_8\}$

5: $\{D_2, D_3, D_5\} \times \{U_7\}$

R	U_1	U_2	U_3	U_4	U_5	U_6	U_8	U_9
D_2	1	0	1	0	0	0	0	1
D_3	1	0	0	1	0	0	0	1
D_5	0	1	0	0	1	0	0	0

7 : $\{U_1, U_7\ U_9\}$ 9: $\{U_2, U_5, U_7\}$

6:$\{D_1,D_2\} \times \{U_1,U_3\}$

R	U_2	U_4	U_5	U_6	U_7	U_8	U_9
D_1	0	0	0	1	0	1	0
D_2	0	0	0	0	1	0	1

11: $\{U_1, U_3, U_6, U_8\}$ 12: $\{U_1, U_3, U_7, U_9\}$

7: $\{D_2, D_3\} \times \{U_1,U_7,U_9\}$

R	U_2	U_3	U_4	U_5	U_6	U_8
D_2	0	1	0	0	0	0
D_3	0	0	1	0	0	0

12: $\{U_1, U_3, U_7, U_9\}$ 13: $\{U_1, U_4, U_7, U_9\}$

10: $\{D_1, D_4\}\times\{U_3,U_6,U_8\}$

R	U_1	U_2	U_4	U_5	U_7	U_9
D_1	1	0	0	0	0	0
D_4	0	1	0	0	0	0

11: $\{U_1, U_3, U_6, U_8\}$ 8: $\{U_2, U_3, U_6, U_8\}$

8:$\{D_4\}\times\{U_2,U_3,U_6,U_8\}$,9:$\{D_5\}\times\{U_2,U_5,U_7\}$,1:$\{D_1\}\times\{U_1,U_3,U_6,U_8\}$,
2:$\{D_2\}\times\{U_1,U_3,U_7,U_9\}$ et 13:$\{D_3\}\times\{U_1,U_4,U_7,U_9\}$, Génèrent tous le concept:
14: $\varnothing\times F$.

La méthode de Bordat a été utilisée dans plusieurs systèmes qui font appel à la structure des treillis de Galois pour représenter les données et les connaissances d'un contexte. La popularité de cette méthode vient du fait que c'est une des rares méthodes complètes de construction de hiérarchies de concepts.

2.6.2 Algorithme de Ganter.

Définition 2.23 : Le vecteur caractéristique.

Soit $A= \{U_i\}_{i\in Q}$ *avec Q un sous ensemble de {1,..., m}, m = cardinal de E et les* $U_i \in F$, *le vecteur caractéristique de A est le vecteur* $(a_1,..., a_m)$ *avec* a_j *qui prend la valeur 1 à la position k si* U_k *est présent dans A et 0 sinon.*

Définition 2.24 de la relation d'ordre sur les vecteurs caractéristique :

Soit $<_i$ *la relation d'ordre partiel strict sur l'ensemble des vecteurs caractéristiques, définie comme suit :*

Soient A et B deux sous ensembles de P (F) et leurs vecteurs caractéristiques VA $(a_1,\ldots, a_m)$ *et VB* $(b_1,\ldots, b_m)$.

$A <_i B \Leftrightarrow$ *les vecteurs caractéristiques de A et B coïncident pour tous les éléments de rang strictement inférieur à i et* $a_i < b_i$.

Définition 2.25 de l'ordre lexicographique :

Soient A et B deux sous-ensembles de P (F), $A < B$ *pour l'ordre lexicographique si et seulement si* $\exists i,\ 1 \leq i \leq m$ *et* $A <_i B$.

Soit A un sous ensemble de P (F) alors on pose :

$A_{i=} (a_1, a_2,\ldots, ai, 0,\ldots,0)$, $A_i' = (a_1, a_2,\ldots, a_{i-1}, 1, 0,\ldots,0)$, *et* $A \oplus i = (f \circ g)(A_i')$.

Proposition[2] :

Soit A un élément de C_F. Le plus petit fermé supérieur à A pour l'ordre lexicographique est $A \oplus i$ ou i est le plus petit indice tel que $A <_i A \oplus i$

Introduction :

L'algorithme de Ganter est fondé sur cette proposition, pour la construction des fermés d'une relation binaire, qui utilise le calcul des vecteurs caractéristiques et l'ordre lexicographique de ces vecteurs. En partant d'un fermé A, représenté par son vecteur caractéristique $(a_1, a_2, \ldots, a_m)$, l'algorithme de Ganter trouve le plus petit fermé se trouvant juste après A pour l'ordre lexicographique sur les vecteurs caractéristiques.

```
1   A:=(0,..., 0, 0)
2   i:= m
3   Tant Que A < (1,..., 1, ,1) Faire
        {Calcul du prochain vecteur qui est potentiellement fermé}
4       Tant Que a_i =1 Faire
5           i : = i - 1
```

[2] Pou la preuve voir [Ganter 1984]

6 **FinTant Que**

7 $a_i := 1$

8 **Pour** j := i + 1 à m **Faire**

9 **a_j :=0**

10 **fin pour**

{Évaluation de A' par l'application composée h' = f ∘ g de la correspondance}

11 A' := h'(A)

{Vérification de la fermeture de A' et préparation de l'indice i pour le calcul du prochain fermé éventuel} .

12 j:= 1

13 **Tant Que** j < i et aj ≥aj' **Faire**

14 j:=j+1

15 **FinTant Que**

16 **Si** aj < aj' **Alors**

17 i:= i - 1

18 **Sinon**

{Le vecteur A est un fermé de la correspondance).

19 i := m

20 A := A'

21 **FinSi**

22 **FinTant Que**

Trace de l'algorithme de Ganter :

Ci-dessous une simulation de l'algorithme de Ganter pour la génération des fermés appliquée sur l'exemple du tableau2.1.

Remarque2.5 :

Si on veut trouver les documents il faut travailler avec le vecteur A (0, 0, 0, 0, 0) et dans l'algorithme on va utiliser h au lieux de h'.

On commence notre algorithme par le vecteur nul.

1. A (0, 0, 0, 0, 0, 0, 0, 0, 0)

2. i= 9

7. A= (0, 0, 0, 0, 0, 0, 0, 0, 1)

11. A'= (1, 0, 0, 0, 0, 0, 1, 0, 1)

17. i= 8

4. a_8 =1

9. A = (0, 0, 0, 0, 0, 0, 0, 1, 0)

11. A'= (0, 0, 1, 0, 0, 1, 0, 1,0)

17. i= 7

4. a_7= 1

9. A= (0, 0, 0, 0, 0, 0, 1, 0, 0)

11. A' = (0, 0, 0, 0, 0, 0, 1, 0, 0)

19 I=9

20 A=A' (le premier vecteur qui appartient au treillis)

Voici tous les vecteurs caractéristiques qui génèrent l'algorithme de notre treillis.

h'(0, 0, 0, 0, 0, 0, 1, 0, 0) = (0, 0, 0, 0, 0, 0, 1, 0, 0) d'où $\{U_7\}$

h'(0, 0, 1, 0, 0, 0, 0, 0, 0) = (0, 0, 1, 0, 0, 0, 0, 0, 0) d'où $\{U_3\}$

h'(0, 0, 1, 0, 0, 1, 0, 0, 0) = (0, 0, 1, 0, 0, 1, 0, 1, 0) d'où $\{U_3, U_6, U_8\}$

h'(0, 1, 0, 0, 0, 0, 0, 0, 0) = (0, 1, 0, 0, 0, 0, 0, 0, 0) d'où $\{U_2\}$

h'(0, 1, 0, 0, 1, 0, 0, 0, 0) = (0, 1, 0, 0, 1, 0, 1, 0, 0) d'où $\{U_2, U_5, U_7\}$

h'(0, 1, 1, 0, 0, 0, 0, 0, 0) = (0, 1, 1, 0, 0, 1, 0, 1, 0) d'où $\{U_2, U_3, U_6, U_8\}$

h'(1, 0, 0, 0, 0, 0, 1, 0, 0) = (1, 0, 0, 0, 0, 0, 0, 0, 0) d'où $\{U_1\}$

h'(1, 0, 0, 0, 0, 0, 1, 0, 0) = (1, 0, 0, 0, 0, 0, 1, 0, 1) d'où $\{U_1, U_7, U_9\}$

h'(1, 0, 0, 1, 0, 0, 0, 0, 0) = (1, 0, 0, 1, 0, 0, 1, 0, 1) d'où $\{U_1, U_4, U_7, U_9\}$

h'(1, 0, 1, 0, 0, 0, 0, 0, 0) = (1, 0, 1, 0, 0, 0, 0, 0, 0) d'où $\{U_1, U_3\}$

h'(1, 0, 1, 0, 0, 0, 1, 0, 0) = (1, 0, 1, 0, 0, 0, 1, 0, 1) d'où $\{U_1, U_3, U_7, U_9\}$

h'(1, 0, 1, 0, 0, 1, 0, 0, 0) = (1, 0, 1, 0, 0, 1, 0, 1, 0) d'où $\{U_1, U_3, U_6, U_8\}$

h'(1, 1, 0, 0, 0, 0, 0, 0, 0) = (1, 1, 1, 1, 1, 1, 1, 1, 1) d'où F

Donc, l'algorithme de Ganter fournit à la sortie une liste de vecteurs caractéristiques représentant les fermés du treillis de Galois de la relation binaire entrée en input. Cette liste

sera topologiquement ordonnée selon l'ordre lexicographique induit par l'inclusion des fermés. (Voir définition 2.25)

2.7 Domaine d'application : Recherche documentaire.

L'utilisation du treillis de Galois généré à partir d'une relation d'indexage est une alternative attrayante parce que deux modes d'interaction peuvent être combinés dans un système intégré et cohérent à partir du même espace de recherche. Chaque classe du treillis correspond à un ensemble de domaines d'informations décrits par les unités d'informations communes. Dans la perspective de la recherche, chaque classe peut être vue comme une requête formée de la conjonction des unités d'informations.

Le graphe représente une relation de généralisation/spécialisation entre les requêtes. La recherche est effectuée par une combinaison libre de

(1) la spécification directe de termes d'index, résultant en un saut dans la classe la plus générale incorporant les termes spécifiés et les termes de la classe de départ.

(2) la navigation libre en suivant les arcs du graphe du treillis. Le parcours d'un arc correspond à un élargissement (généralisation) ou un raffinement (spécialisation) minimal par rapport à la requête correspondant à la classe courante.

Le second mode d'interaction offre une solution au problème de raffinement de requête des systèmes booléens.

Une question importante pour l'applicabilité de cette approche est la complexité de la structure et des algorithmes de construction du treillis. Moyennant une hypothèse raisonnable dans le contexte de la recherche documentaire, le treillis croît linéairement par rapport au nombre de documents. D'autre part, plusieurs algorithmes ont été conçus pour générer le treillis de Galois [Bordat 1986, Carpineto 1993, Chein 1969, Fay 1975, Ganter 1984, Godin 1995, Norris 1978]. Parmi ceux-ci, des algorithmes incrémentaux [Carpineto 1993, Godin 1995] permettent de mettre à jour le treillis et le graphe.

Cette caractéristique est importante en recherche documentaire où de nouveaux documents sont fréquemment ajoutés à la collection. Il y a plusieurs implémentations d'algorithmes incrémentaux et des données empiriques provenant de plusieurs applications ont démontré que l'adjonction d'un nouveau document est faite en temps linéaire par rapport au nombre de documents [Godin 1991], [Godin 1993]. Avec l'hypothèse d'une borne supérieure sur le nombre de termes d'index par document, l'analyse dans le pire des cas montre aussi une complexité linéaire.

Chapitre 3

Classification textuelle au moyen des treillis de galois.

Dans ce chapitre, nous proposons une méthode variante originale des modèles de classification connexionniste pour résoudre le problème. Cette approche permet d'appliquer les performances classificatoires des réseaux de neurones et treillis de Galois à des *corpus* textuels et de produire donc des regroupements susceptibles d'interprétations sémantiques, et ce en utilisant le concept des n-grams de caractères.

Normalement la première étape dans un processus de traitement d'un gros corpus au moyen d'un outil statistique est de subdiviser le texte à traiter en plusieurs unités d'information appelées *tokens* qui sont, traditionnellement, des mots simples. Ce processus de *tokenisation* pose une question primordiale : sur le plan informatique, comment repérer un mot ? En d'autres termes, quels sont les indicateurs formels de surface, non ambigus, qui peuvent délimiter un mot ? Si pour le français ou l'anglais littéraire, ou des langues apparentées, la réponse est presque triviale — à savoir que toute chaîne de caractères précédée et suivie d'un espace est considérée comme un mot simple — il en va tout autrement pour d'autres langues. Dans le cas de termes composés en langue allemande comme, par exemple, *lebensversicherungsgesellschaftsangestellter* ("employé d'une compagnie d'assurance vie"), ou pour la langue arabe dans laquelle les pronoms sujets et compléments sont dans certains cas attachés aux verbes et une seule chaîne de caractères représente ainsi une phrase comme, par exemple, kathabthouhou ("je l'ai écrit"), cette notion de *tokens* devient carrément inadéquate [Manning & Schütze, 1999].

Si le mot simple ne convient pas à toutes les langues, quelle est donc l'unité d'information atomique la plus adéquate pour segmenter un texte ? [Balpe *et al.* 1996] soulignent que dépendant de l'objectif de lecture et de compréhension que nous nous donnons, la définition de l'unité d'information dépend de l'usage qui en est attendu. Dans

une perspective de classification numérique à des fins d'extraction de connaissances, la définition d'une unité d'information est tributaire des contraintes suivantes :

- L'unité d'information doit être une portion du texte soumis à l'analyse numérique.
- Il doit être facile sur le plan informatique de repérer les unités d'information.
- La définition d'une unité d'information doit être indépendante de la langue dans laquelle le texte est écrit. Une telle définition permet à l'analyse numérique, moyennant des modifications minimes, de couvrir un large éventail de langues.
- Les unités d'information doivent être statistiquement comparables. Il doit être aisé d'en calculer les fréquences d'apparition dans les différentes parties du texte et par conséquent d'estimer leur distribution et la régularité à laquelle plusieurs unités cooccurrent dans les mêmes parties du texte.

Que l'unité linguistique dans une analyse de classification numérique soit linguistiquement comprise dans une certaine mesure hors de son contexte n'est pas en soit une contrainte lors de la phase de classification. Toutefois à l'affichage des résultats, ce facteur devient important dès lors que l'interprétation faite par l'utilisateur en dépend.

La plupart des analyseurs statistiques fondés sur le calcul de la fréquence des cooccurrences utilisent le mot comme unité d'information, même si celui-ci ne répond pas à toutes les contraintes énumérées ici. Cependant, l'importance de l'ergonomie interprétative des mots a prévalu sur tout autre aspect, particulièrement ceux liées aux aspects multilingues de l'analyse. Ce dernier facteur devient aujourd'hui incontournable : l'essor du Web confirme ce besoin de multilinguisme. Il semble donc impératif que les modèles pour l'analyse de corpus, qu'ils soient numériques ou linguistiques, tiennent compte du caractère multilingue des textes à analyser.

3.1 Les n-grams de caractères

Bien qu'ayant été proposée depuis longtemps et utilisée principalement en reconnaissance de la parole, la notion de **n-grams de caractères** prit davantage d'importance avec les travaux de [Greffenstette 1995] sur l'identification de la langue, de [Damashek 1995] sur le traitement de l'écrit. Autre autres, ils prouvèrent que ce découpage, bien que différent d'un découpage en mots, ne faisait pas perdre d'information. Parmi des applications plus récentes des n-grams on retrouve des travaux sur : l'indexation [Mayfield & McNamee, 1998] ; l'hypertextualisation automatique multilingue avec les travaux de [Halleb et Lelu 1998] qui, à travers une méthode de classification thématique de grandes collections de textes, indépendante du langage, construisent des interfaces de navigation hypertextuelle ; ou encore l'analyse exploratoire multidimensionnelle en vue d'une recherche d'information dans des corpus textuels [Lelu *et al.*, 1998].

On définira un n-gram de caractères par une suite de n caractères : bi-grams pour n=2, tri-grams pour n=3, quadri-grams pour n=4, etc. Il n'est plus question de chercher un délimiteur comme c'était le cas pour le mot.

Un découpage en n-grams de caractères, quelque soit n, reste valable pour toutes les langues utilisant un alphabet et la concaténation comme opérateur de construction de texte.

Le choix des n-grams apporte un autre avantage très important : il permet de contrôler la taille du lexique et de la maintenir à un seuil raisonnable. La taille du lexique était jusqu'à présent l'aspect le plus controversé et considéré comme une limite des techniques fondées sur la comparaison des chaînes de caractères. En effet, un découpage en mots fait que la taille du lexique est d'autant plus grande que le corpus est grand. Cette limite subsiste malgré certains aménagements tels le "nettoyage" des mots fonctionnels, la lemmatisation et la suppression des hapax.

Un lexique obtenu suite à un découpage en n-grams de caractères ne peut dépasser la taille de l'alphabet à la puissance n. Le choix d'un découpage en quadri-grams pour une langue de 26 caractères donnerait une taille maximale de 26^4 entrées, soit un lexique de 456 976 quadri-grams possibles. Si on élimine les combinaisons qu'il est impossible de rencontrer (p.ex. AAAA, ABBB, BBBA, etc.), ce nombre diminue de façon considérable.

D'ailleurs ce nombre est estimé par [Lelu *et al.* 1998] à quelques 13 087 quadri-grams pour un texte de 173 000 caractères.

Dans une approche avec découpage en n-grams de caractères, contrairement aux approches avec découpage en mots, il n'est pas question d'utiliser la lemmatisation pour réduire le lexique. La lemmatisation (qui consiste à remplacer une forme fléchie par son lemme) est, d'une part, relativement lourde à mettre en œuvre sur le plan informatique mais en plus, impose un traitement spécifique à chaque langue. Qui plus est, plusieurs lemmatiseurs ne semblent pas être en mesure de ramener des termes comme informatisation, informatique, et informatiser à un même concept qu'est l'informatique. Or souvent dans les corpus, on utilise des expressions ayant quasiment le même contenu informationnel comme, par exemple, dans les segments suivants : "l'informatisation de l'école", "informatiser l'école" et "introduire l'informatique à l'école". Le découpage des trois segments en n-grams est suffisant pour classer les trois segments dans la même classe car, outre le mot école qui est redondant dans les trois expressions, les tri-grams inf, nfo, for, orm, rma, mat et ati, permettent par un calcul de similarité d'affirmer que c'est d'informatique dont il est question. Par ailleurs, les tri-grams susmentionnés apparaissent aussi dans le découpage des mots information, informationnel, etc., ce qui peut être considéré à juste titre comme du bruit, à moins bien sûr que l'on évoque une interprétation sémantique particulière de l'informatique comme étant une science de l'information.

L'extraction de connaissances d'un texte est aussi un processus itératif comprenant plusieurs étapes avec plusieurs décisions (Figure 3.1) pouvant être prises par l'utilisateur-analyste, avec retour aux étapes précédentes en cas de non satisfaction. Ces étapes sont :

1. Prétraitement.
2. Traitement.
3. Interprétation.

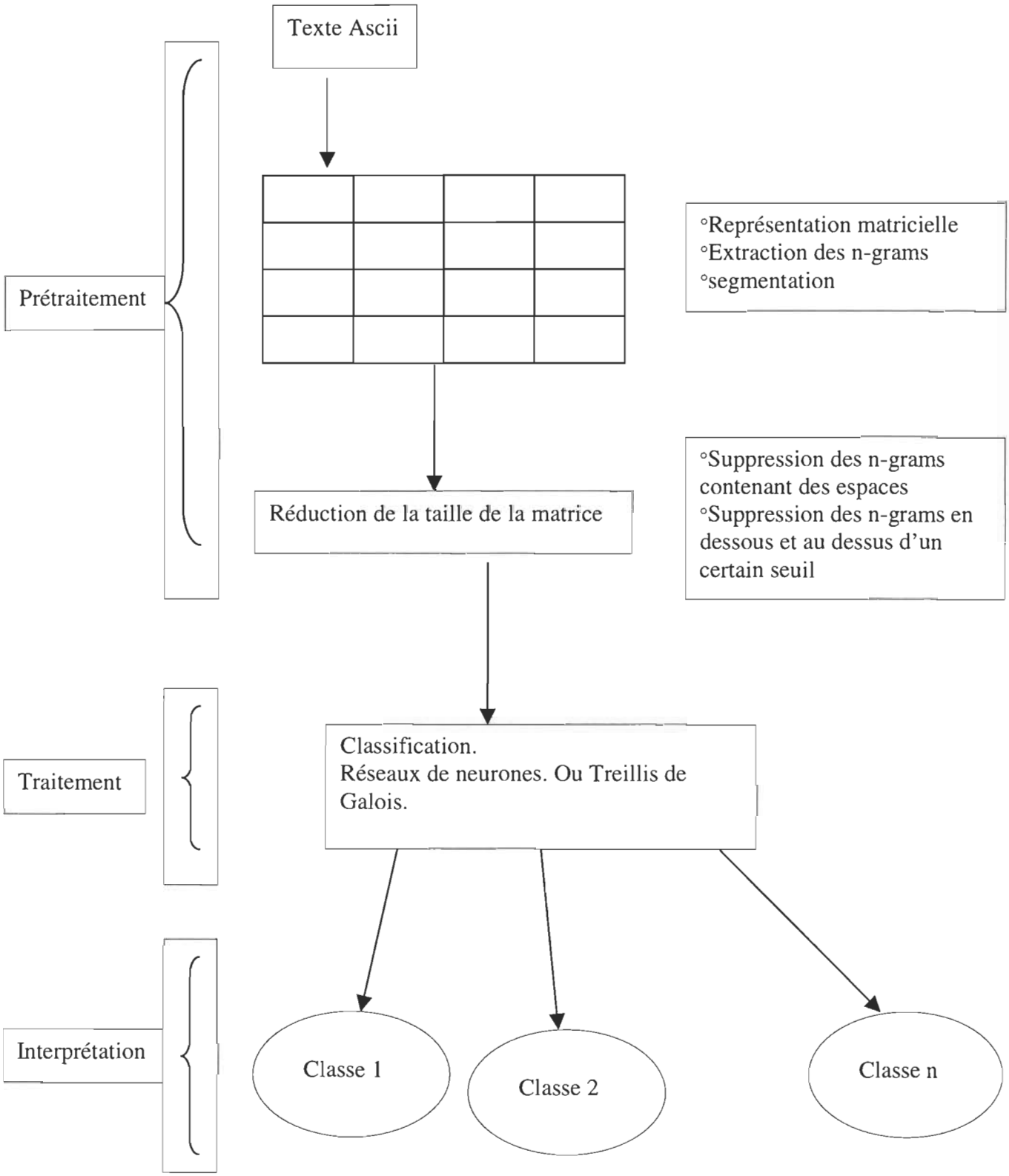

Figure 3.1 - La représentation des trois étapes de notre travaille.

3.2 Prétraitement

GRAMEXCO est un outil logiciel que nous avons développé pour la classification numérique des gros corpus et l'extraction de connaissances sur le contenu des textes. La classification numérique s'effectue au moyen d'un réseau de neurones ART et Treillis de Galois. L'unité d'information considérée est le n-gram de caractères, la valeur de n étant paramétrable. L'objectif visé est de fournir la même chaîne de traitement, peu importe la langue du corpus, avec toutefois des aménagements dans la présentation des résultats pour en permettre une relative facilité de lecture comme nous le verrons plus loin. Le fonctionnement de GRAMEXCO n'est pas totalement automatique. Le choix de certains paramètres est fait par l'utilisateur en fonction de ses propres objectifs. Du choix de ces paramètres dépend l'interprétation des résultats qui se fait par l'utilisateur en fonction de sa subjectivité. GRAMEXCO prend en entrée un texte brut (non indexé) sous format ASCII. Il s'en suit trois grandes étapes où l'utilisateur peut paramétrer certains traitements.

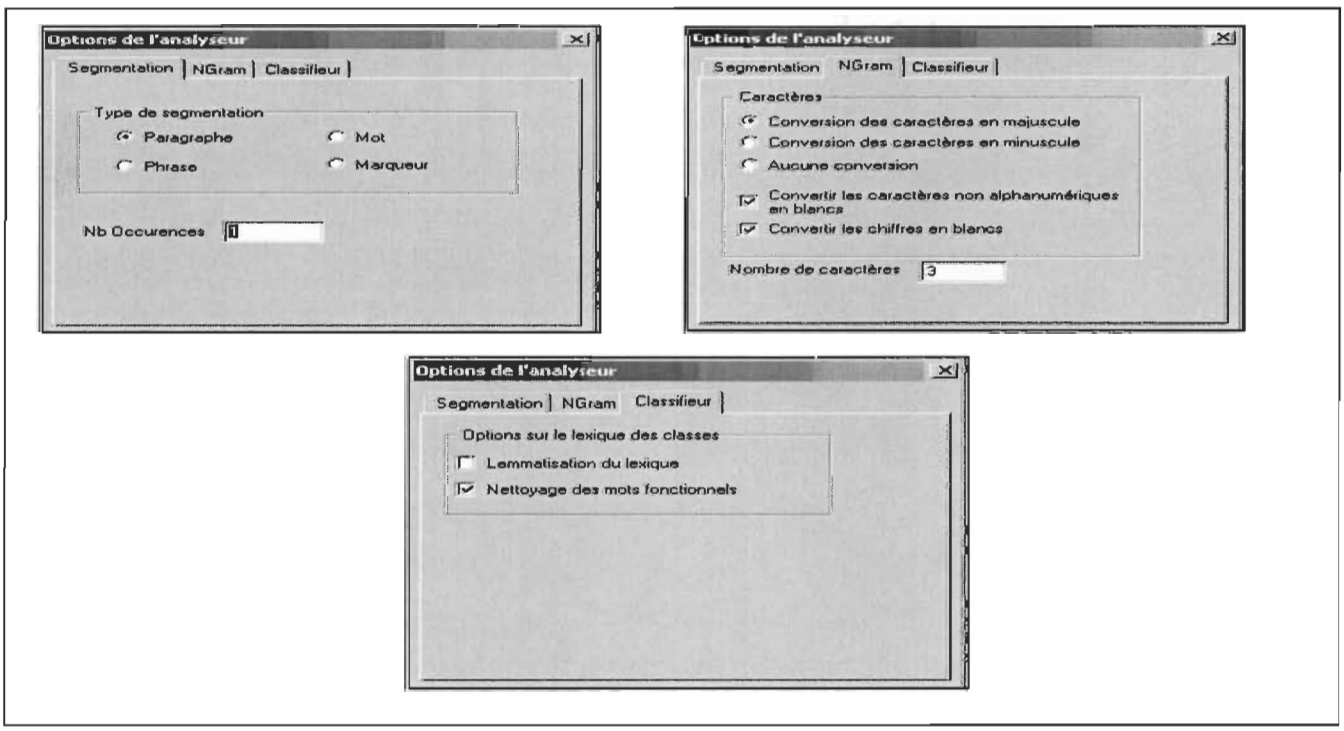

Figure 3.2 - Paramétrage de l'outil GRAMEXCO

La **première étape** consiste à construire la liste des n-grams de caractères contenus dans le texte ainsi qu'à partitionner le corpus en plusieurs segments. Les deux opérations se

faisant simultanément, nous récupérons en sortie une matrice où seront répertoriés les fréquences d'apparition de chaque n-gram dans les différents segments.

Le choix de la valeur du n (bi-gram, tri-gram, quadri-gram, etc.) dépend de l'utilisateur et de l'expertise qu'il veut mener. Outre la valeur du n, d'autres paramètres (voir Figure 3.2) ont la possibilité d'effectuer la conversion des caractères non alphanumériques en caractère espace, ou encore la conversion des chiffres en caractère espace.

Ces deux paramètres répondent aux besoins d'une analyse pour laquelle les chiffres, la ponctuation ou encore d'autres caractères spécifiques seraient importants pour la qualité des résultats. Dans un texte technique par exemple, il serait peut être intéressant de savoir si version1 est différente de version2 et, par conséquent, les chiffres pourraient avoir autant d'impact informatif que les caractères alphabétiques.

Le dernier paramètre pour les n-grams est en rapport avec la conversion des caractères majuscules en minuscules, ou vice versa. Si aucune de ces conversions n'est choisie, alors GRAMEXCO distinguera les lettres minuscules des majuscules.

L'autre aspect important de cette première étape est le paramétrage de la segmentation. Ainsi, nous pouvons partager le texte soit en des sections formées d'un nombre déterminé de phrases, de paragraphes ou de mots, ou tout simplement des sections séparées par un caractère spécial. Ce paramètre est toujours choisi par l'utilisateur.

Le pseudo-lexique formé de n-grams subit au cours de cette première étape un nettoyage (voir Figure 3.3) soit, l'élimination des "n-grams hapax" dont la fréquence est inférieure à un certain seuil ou supérieure à un autre seuil, l'élimination de n-grams spécifiques sélectionnés dans la liste (par exemple des n-grams contenant des espaces) ou encore, si on veut pousser les choses plus loin, l'élimination de certains n-grams considérés comme fonctionnels, particulièrement les suffixes.

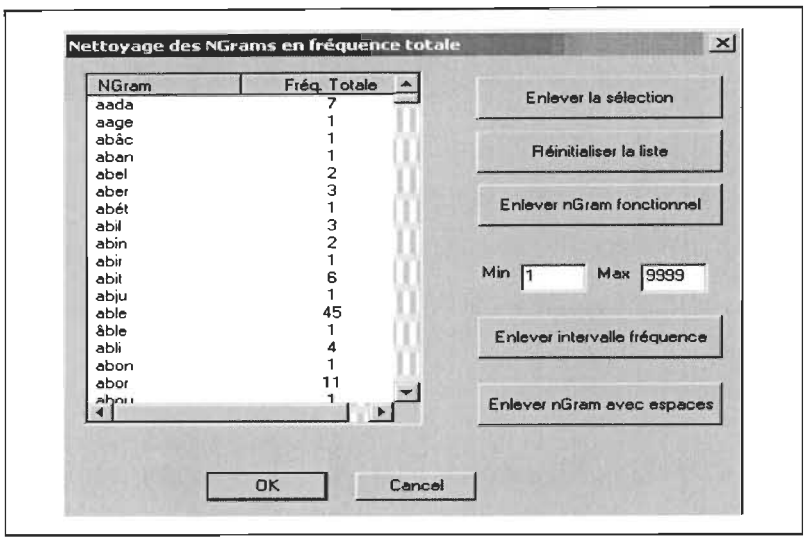

Figure 3.3 - Nettoyage de la liste des n-grams produits par GRAMEXCO

3.3 Traitement.

Nous faisons le traitement avec un réseau de neurones ou un treillis de Galois. Toutefois nous nous intéresserons ici au traitement avec les treillis de Galois, plusieurs publications ayant donnée le détail des traitements avec le réseau de neurones ART [Meunier & al. 1997]. Nous donnerons par contre au chapitre suivant une comparaison qualitative et quantitative des résultats obtenues avec les deux classifieurs.

3.3.1 La notion de la norme et seuil.

Ce que nous appelons traitement est l'opération qui consiste à regrouper l'ensemble des parties d'un corpus ayant des similarités lexicales. Aussi l'input de cette classification est une matrice à deux entrées comme celle qui suit :

	U_1	U_2	.	.	U_j	.	U_{n-1}	U_n
D_1	$X_{1,1}$	$X_{1,2}$			$X_{1,j}$		$X_{1,n-1}$	$X_{1,n}$
D_2	$X_{2,1}$	$X_{2,2}$			$X_{2,j}$		$X_{2,n-1}$	$X_{2,n}$
.								
.								
D_i	$X_{i,1}$	$X_{i,2}$			$X_{i,j}$		$X_{i,n-1}$	$X_{i,n}$
.								
D_{m-1}	$X_{m-1,1}$	$X_{m-1,2}$			$X_{m-1,j}$		$X_{m-1,n-1}$	$X_{m-1,n}$
D_m	$X_{m,1}$	$X_{m,2}$			$X_{m,j}$		$X_{m,n-1}$	$X_{m,n}$

Tableau 3.2 - La Représentation matricielle d'un texte.

D_i : domaine d'information.

U_j : unité d'information.

$X_{i,j}$: fréquence de U_j dans D_i.

n : nombre d'unité d'information dans le texte.

m : nombre de domaine d'information dans le texte.

Définition 3.1 d'une matrice binaire :

Soit la matrice M définie comme suit :

	U_1	U_2	.	.	U_j	.	U_{n-1}	U_n
D_1	$X_{1,1}$	$X_{1,2}$			$X_{1,j}$		$X_{1,n-1}$	$X_{1,n}$
D_2	$X_{2,1}$	$X_{2,2}$			$X_{2,j}$		$X_{2,n-1}$	$X_{2,n}$
.								
.								
D_i	$X_{i,1}$	$X_{i,2}$			$X_{i,j}$		$X_{i,n-1}$	$X_{i,n}$
.								
D_{m-1}	$X_{m-1,1}$	$X_{m-1,2}$			$X_{m-1,j}$		$X_{m-1,n-1}$	$X_{m-1,n}$
D_m	$X_{m,1}$	$X_{m,2}$			$X_{m,j}$		$X_{m,n-1}$	$X_{m,n}$

La matrice M est dite binaire si $X_{i,j}$=0 ou 1.

Il est possible de passer d'une matrice de fréquence à une matrice binaire. Ceci est important en raison de la contrainte que posent les treillis de Galois en l'occurrence leur choix d'une Matrice binaire. Le passage d'une matrice de fréquence à une matrice binaire sera appelé **normalisation**.

Pour cela nous avons besoin de quelques définitions mathématiques.

Définition 3.2 de la présence :

$P(x_{i,j})$ c'est la présence de l'unité d'information $x_{i,j}$ dans le segment D_i.

$$P(x_{i,j}) = x_{i,j} / \sum_{j \leq n} x_{i,j} .$$

La présence mesure l'importance de chaque unité d'information par rapport aux autres unités.

Remarque 3.1 :

Nous pouvons donner une autre formule de la présence pour mesurer l'importance d'une unité d'information cette mesure est la fonction de la présence d'une unité d'information par rapport a l'ensemble d'un texte et non uniquement par rapport à un segment.

$$P(x_{i,j}) = x_{i,j} / \sum_{i \leq m} x_{i,j} .$$

Définition 3.3 seuil :

Le seuil « S » : *C'est une valeur réelle, il va nous permettre de déterminer si à une unité d'information, selon sa présence, on lui attribue la valeur 0 ou 1. Le seuil est choisi par l'usager.*

Définition 3.4 de la fonction d'attribution :

La fonction A (attribution) *est définie de* $N \rightarrow \{0,1\}$.

$$A(x_{i,j}) = \begin{cases} 1 & \text{si } P(xi, j) \succ S \\ 0 & \text{si } P(xi, j) \leq S \end{cases}.$$

3.3.2 Exemple d'application.

Exemple 3.2:

R	U_1	U_2	U_3	U_4	U_5	U_6
D_1	10	6	3	0	0	2
D2	0	1	6	5	0	0
D3	0	0	0	2	1	4

Tableau 3.3 - Une représentation matricielle.

1 cas :

On suppose que le seuil : S = 0. Donc la fonction attribution =1 pour toutes les unités d'information même si elles n'apparaissent qu'une fois dans les segments. Sinon la fonction d'attribution = 0.

Donc le tableau 3.3 après normalisation laisse place au tableau3.4 ci-dessous.

R	U_1	U_2	U_3	U_4	U_5	U_6
D_1	1	1	1	0	0	1
D2	0	1	1	1	0	0
D3	0	0	0	1	1	1

Tableau 3.4 - Une Représentation matricielle après la normalisation avec S=0.

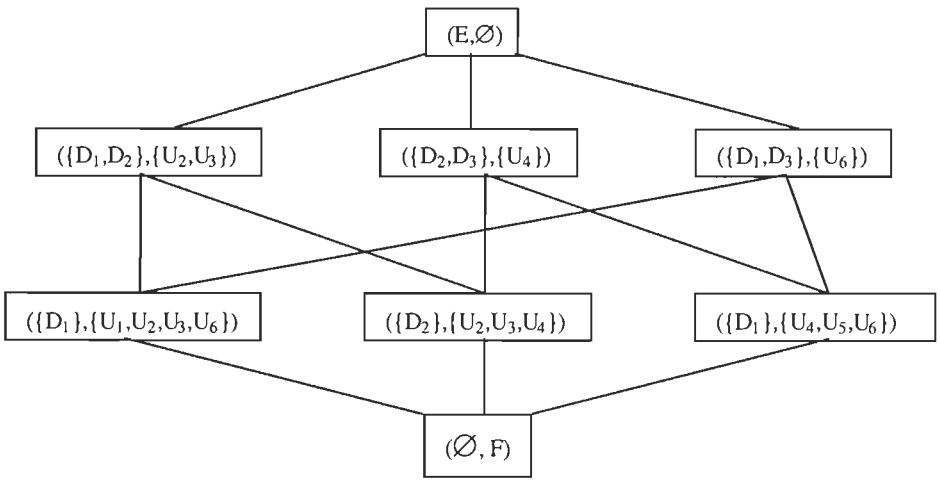

Figure 3.4 - Diagramme de Hasse de la correspondance R du tableau 3.4.

2éme cas

On suppose que le seuil: $S = 0.1$, Pour calculer la présence P $(x_{1,1})$ de l'unité d'information U_1 dans le document D_1.

$$P(x_{1,1}) = 10/(10+6+3+2) = 0.47 \succ S .$$

Pour calculer la présence P $(x_{2,2})$ de l'unité d'information U_2 dans le document D_2.

$$P(x_{2,2}) = 1/(1+6+5) = 0.08 \leq S$$

Donc le tableau 3.3 après normalisation donne le tableau 3.5 ci-dessous.

R	U_1	U_2	U_3	U_4	U_5	U_6
D_1	1	1	1	0	0	0
D2	0	0	1	1	0	0
D3	0	0	0	1	0	1

Tableau 3.5 - Une Représentation matricielle après la normalisation avec S=0.1.

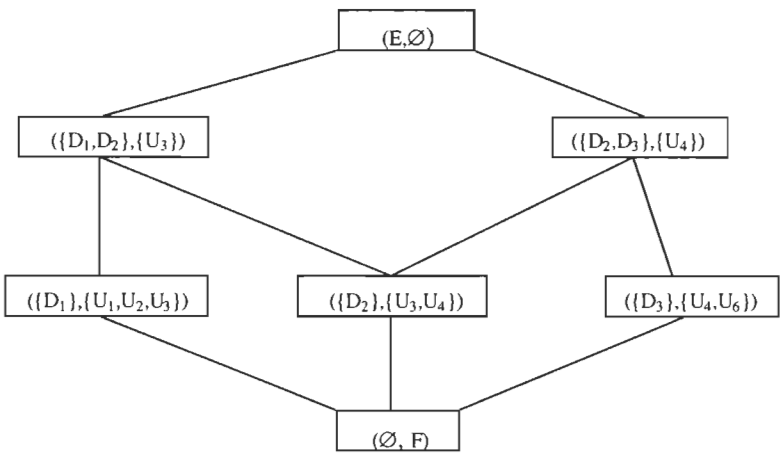

Figure 3.5 - Diagramme de Hasse de la correspondance R du tableau 3.5.

3éme cas

On suppose que le seuil: **S = 0.4**. Après avoir calculé les fonctions d'attribution on trouve :

R	U_1	U_2	U_3	U_4	U_5	U_6
D_1	1	0	0	0	0	0
D2	0	0	1	1	0	0
D3	0	0	0	0	0	1

Tableau 3.6 - Une Représentation matricielle après la normalisation avec S=0.4.

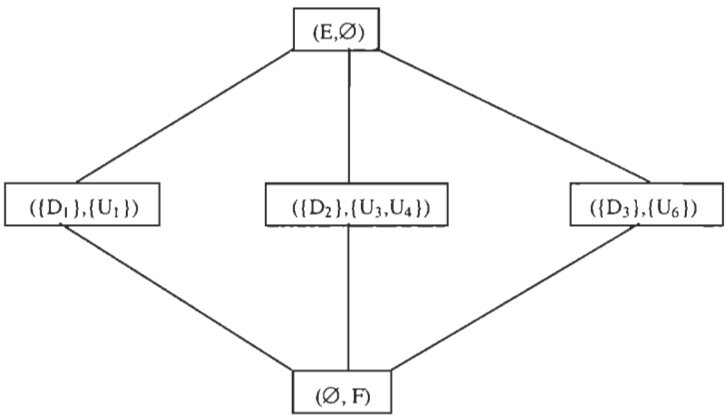

Figure 3.5 - Diagramme de Hasse de la correspondance R du tableau 3.5.

Remarque 3.2 :

Le choix du seuil peut changer la construction du treillis et le nombre de concepts dans un treillis. Plus le seuil S est grand plus le nombre de concepts diminue étant donné qu'il y aura moins de relations entre les segments.

3.3.3 Contraintes et solutions.

Nous avons besoin du numéro de la classe et les domaines d'informations associés à cette classe donc nous allons utiliser l'algorithme de Ganter pour trouver juste les domaines d'informations qui appartiennent à chaque concept du treillis. Car l'algorithme de ganter calcule les éléments du treillis C_E sans construire les arêtes.

Définition 3.5 Taille du domaine d'information :

Taille du domaine d'information c'est le nombre de n-grams ou de mots existant dans chaque domaine.

Si la taille du domaine d'information est très grande, c'est à dire contient beaucoup plus de n-grams que l' autres domaines, celui ci sera en relation avec beaucoup plus de domaines. Il sera considéré comme un bruit. Soit il sera dans la derniere classe qui contiendra tous les domaines de notre texte, soit ce domaine va être seul dans une classe.

Solution :

(1) augmenter la taille des n-grams.

Augmenter la taille des n-grams dans la phase de prétraitement du texte avec comme conséquence la diminution des n-grams.

(2) définir pour chaque domaine d'information un seuil.

Exemple 3.3 :

Soit $D_i = (x_{i,j})$ un segment, on calcule $(P(x_{i,1}),P(x_{i,2}),\ldots,P(x_{i,n}))$ avec $P(x_{i,j})$ le vecteur de présence pour chaque éléments de D_i.

$$S_i = \frac{\sum_{0\le j\le n} P(x_{i,j})}{\sum_{0\le j\le n} \delta(x_{i,j})} \text{ avec } \delta(x_{i,j}) = \begin{cases} 1 & \text{si } x_{i,j} \neq 0 \\ 0 & \text{sinon} \end{cases}.$$

Cette deuxième solution est à l'étude, pour notre travail on a considéré la premiere solution.

3.4 Interprétation.

La configuration du résultat de la classification numérique se présente par l'affichage des classes de segments et, pour chaque classe, l'affichage des segments qui la constituent d'une part, et du lexique qui la forme d'autre part (voir Figure 3). À cette **troisième étape** la notion de n-gram n'est plus de mise.

Il serait en effet impossible à un utilisateur d'interpréter des résultats et de donner des thèmes aux différentes classes à partir d'une seule liste de n-grams. Comme le souligne [Turenne 2000], l'interprétation de telles classes est déjà un exercice non trivial en lui-même, dépendant *des* points de vue de l'utilisateur : il ne serait donc pas utile de lui rendre cette phase moins intuitive en utilisant une liste de n-grams.

Le lexique de chaque classe est formé par les mots que contiennent les différents segments de cette classe. L'utilisateur pourra considérer le lexique comme l'union des mots des segments pour déterminer le thème global des classes, leur intersection pour déterminer le thème commun partagé par les segments, leur différence pour identifier des gains informationnels, ou encore tous ceux dont la fréquence est au dessus d'un certain seuil, etc.

L'utilisateur peut également lemmatiser le lexique des classes comme il peut en retirer les mots fonctionnels. L'utilisateur peut appliquer l'opération de lemmatisation à l'ensemble des lexiques de toutes les classes où seulement au lexique d'une seule classe, ceci en fonction de contraintes de temps. Il est à retenir que la lemmatisation et la suppression des mots fonctionnels n'interviennent que pour améliorer l'aspect des résultats et n'interviennent nullement avant la classification à proprement parler.

Toutes ces configurations du lexique sont à même d'aider l'utilisateur à proposer son interprétation des résultats. En effet, il demeure que GRAMEXCO, comme nous l'avons souligné plus haut, ne propose pas d'interprétation automatique. Il ne fait que faire ressortir les similarités et les régularités découvertes dans le corpus.

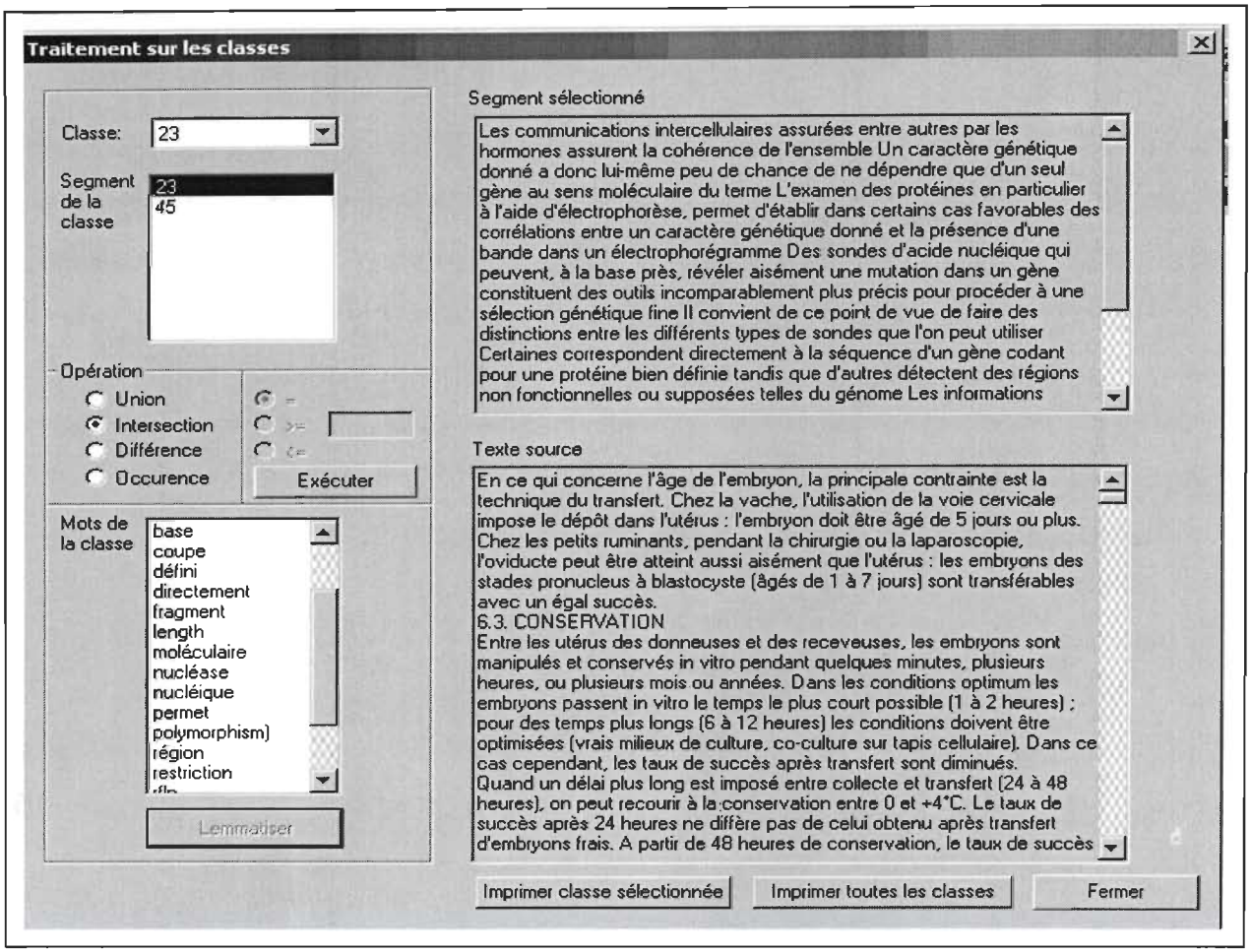

Figure 3.7 - Configuration des résultats

Dépendant des paramètres choisis, les résultats de GRAMEXCO peuvent servir à plus d'une finalité. Comme nous le verrons à l'aide des exemples de la prochaine section, nous pouvons :

- déterminer le contenu lexical des segments similaires, et ainsi connaître le thème principal de ces segments.
- déterminer l'acception et la signification d'un mot de par les mots qui lui sont associés dans une classe donnée.

Chapitre 4

Évaluations et Commentaires.

Nous avons mené deux évaluations principales. La première évaluation avec le classifieur Treillis de Galois et la deuxième par réseau de neurones ART. Les évaluations voulaient aboutir à une comparaison entre les deux classifieurs, et montrer le comportement d'une classification numérique fondée sur les n-grams de caractères. Elles ont été réalisées sur un corpus formé de 60 pages (format ASCII) construit à partir d'extraits de documents trouvés sur le web. Ces documents couvrent divers domaines. Ils permettent une hétérogénéité du contenu du corpus et, par conséquent, une meilleure compréhension des résultats de la classification.

Pour les opérations préliminaires des évaluations, soit la segmentation et l'extraction des n-grams, nous avons opté pour les paramètres suivants :

- 3 paragraphes pour déterminer la taille d'un segment
- 4 (caractères) pour déterminer la taille des n-grams[3].
- De plus, à l'aide des paramètres de GRAMEXCO, nous avons considéré les lettres majuscules identiques aux lettres minuscules, et nous avons remplacé les caractères non alphanumériques par des espaces.

Nous avons ainsi récupéré 103 segments et 13 349 quadri-grams, après un "ménage" de la liste des n-grams qui a consisté à supprimer les n-grams contenant un ou plusieurs espaces et les n-grams ayant une fréquence égale à 1.

[3] Selon Damashek (1995), les quadri-grams donneraient les meilleurs résultats pour l'anglais. Lelu *et al.* (1998) semblent confirmer cela pour le français.

Pour une meilleure comparaison on va utiliser le même texte en input, que ce soit pour une classification avec ART ou avec treillis de Galois. Le premier classifieur dans notre évaluation est un treillis de Galois, le deuxième est le réseau de neurones ART.

4.1 Classification au moyen de treillis de Galois.

La première classification, au moyen du treillis de Galois. Apres avoir mené plusieurs évaluations, nous avons choisi le seuil 0.01 car il donne de meilleurs résultats, le seuil qui est inférieur à 0.01 donnes plus de classes. Le seuil supérieur à 0.01 donnes moins de classes mais en perd de la précision dans la classification. notre choix du seuil (qui est égale à 0.01) donne lieu à la production de 284 classes examinons maintenant quelques résultats :

****

Classe 281

****

Segment 1

Jeudi soir, lors du Conseil communal, le législatif a accepté que certains bâtiments communaux soient chauffés à distance, et au bois. Ainsi, nous ne dépendrons plus des pays producteurs de pétrole, relevait un conseiller.

Le développement des installations se fera par étapes Dans un premier temps, les deux bâtiments scolaires et la grande salle sont seuls concernés. Le coût de cette installation s'élève à 180 000 francs. A terme, d'autres bâtiments pourront se raccorder au système existant: l'administration communale, la cure, la Bourse des Pauvres, entre autres, mais aussi nombre de particuliers.

Les bâtiments susceptibles d'être raccordés feront l'objet d'études et de demandes de crédits ultérieures. Les permis de construire pour les nouveaux bâtiments seront délivrés avec une demande expresse qu'ils soient raccordés au réseau s'ils sont construits dans le périmètre d'exploitation. Ce projet présent l'avantage de sauvegarder l'environnement et d'utiliser des énergies renouvelables, soulignait le syndic Jurg Hofer.

Segment 2

Le législatif a accepté un second préavis, concernant la création d'une société anonyme pour le chauffage à distance, Genolier CAD SA. La commune disposera de la majorité des voix à l'assemblée générale.

Elle sera dirigée par un conseil d'administration de 5 à 7 membres, dont un membre devra être issu du Conseil communal. La commune, l'Etat de Vaud et l'Association forestière vaudoise et du Bas-Valais en seront membres.

Les forêts ne sont pas exploitées correctement

Segment 15

Alimentation: le bon choix de matières grasses dans votre assiette

La chasse au cholestérol n'est plus d'actualité pour lutter contre les maladies cardio-vasculaires: l'accent est mis sur un choix judicieux des matières grasses. Explications et conseils.

Dans les pays industrialisés, les maladies cardio-vasculaires constituent une des premières causes de mortalité précoce. Des relations entre l'alimentation et ces pathologies ont été clairement établies: une consommation élevée de matières grasses accroît le risque de maladie, alors que la consommation de fruits, légumes et poisson le réduit.

Segment 16

Par maladies cardio-vasculaires, on entend, entre autres, l'infarctus, l'athérosclérose (modification de la paroi des artères) et ses complications, thrombose et embolie. Les troubles des lipides (ou graisses) sanguins qui correspondent au terme barbare de «dyslipidémies» et dont l'excès de cholestérol est un exemple constituent un facteur de risque des maladies cardio-vasculaires.

LIEN DE CAUSALITÉ

Quel est le lien entre les graisses alimentaires et les maladies cardio-vasculaires? Une consommation de graisses excessive de même que la qualité des graisses consommées (origine animale ou végétale) influencent les transformations biochimiques qui ont lieu dans l'organisme. Des déséquilibres entre les graisses sanguines peuvent alors survenir: trop de «mauvais» cholestérol, pas assez de «bon», par exemple.

Segment 93

`Tel était le thème de l'émission Controverse de ce dimanche sur RTL-TVi. La réponse ne peut être que négative! Quelle raison objective aurions-nous d'attaquer l'Irak? C'est un régime dictatorial? Absolument! Combien d'autres dictatures existent actuellement dans le monde? Ne nous leurrons pas. Les inspecteurs en

désarmement trouveront ce que Bush voudra qu'ils trouvent et le Conseil de sécurité de l'Onu votera ce que Washington lui conseillera de voter.´

SOCIÉTÉ -

Pourquoi s'arrêteraient-ils?

Lexique de la classe selon l'occurence 3 des mots dans les segments.

Bâtiment cardio communal graisse gras maladie matière membre raccorder vasculaire

Commentaire :

Cette classe est mixte. Elle contient deux thèmes : l'alimentation et les maladies cardio vasculaire. C'est normal car elle est parmi les dernières classes de la classification.

****

Classe 43

****

Segment 76

La devise européenne s'est appréciée de plus de 21% par rapport au dollar depuis son introduction fiduciaire le 1er janvier 2002.

En 2002, sa progression, qui a été de 15,5%, s'est exacerbée avec les tension en Irak, alors que l'euro s'est apprécié de 5,03% sur la première partie de l'année contre 10,46 % sur les six derniers mois de 2002.

`Le fait que la monnaie unique se soit fortement appréciée par rapport à la livre sterling conforte ce fait, alors que la Grande-Bretagne est le plus proche allié des Etats-Unis´, ajoute Audrey Childe Freeman.

Segment 77

L'euro évolue actuellement à son plus haut niveau depuis juin 1999 face à la livre sterling, au dessus du seuil de 0,6640. Mais `les indicateurs économiques fondamentaux de la zone euro ne justifient aucunement une telle appréciation´, souligne cette économiste.

Le chef économiste de la Banque centrale européenne, Otmar Issing, s'est lui-même inquiété récemment de la faiblesse de l'économie allemande.

`Bien sûr, le fait que le plus grand pays de la zone euro soit à la traîne en termes de croissance et de création d'emplois est pour nous une source de préoccupation´, avait-t-il déclaré dans une interview à l'AFP.

Lexique de la classe selon l'occurence des mots dans les segments.

Apprécier euro.

Commentaire :

Apprécier ici est compris dans son sens valoriser et non dans son sens affectionner du fait qu'il est accompagné du mot euro.

****

Classe 70

****

Segment 65

EIJSDEN Avant de rallier le Standard, le nom d'Ole-Martin Aarst fut cité, à plusieurs reprises, en Angleterre. `J'admire le football des Iles, avoue-t-il. Je suis un grand supporter de Tottenham. La plupart de mes amis encouragent Liverpool ou Manchester United. Depuis ma plus tendre enfance, mon coeur bat au rythme des Spurs. L'origine de cet engouement remonte à mon enfance. Lorsque j'étais âgé de huit ans, mon père me ramena d'un de ses périples dans la capitale anglaise une vareuse de Tottenham. J'étais très fier de ce cadeau et j'ai vibré aux exploits de Glenn Hoddle. Le médian britannique a toujours été un modèle pour moi. A tel point même que je souhaitais porter le même numéro que lui dans mon équipe. Cette année, Tottenham ne réalise pas une grande saison mais cela ne m'empêche pas de suivre ses résultats à la loupe. Lorsque j'aurais le temps et l'opportunité, je me rendrai à White Hart Lane pour assister à une rencontre de mon équipe favorite. C'est une certitude d'autant que mon beau-frère est, lui aussi, un supporter acharné des Spurs. Pour l'instant, nous devons nous contenter de matchs télévisés...´

Ole-Martin Aarst aurait aimé se mesurer à Tottenham mais il n'en eut pas l'occasion. `Lorsque j'ai décidé de quitter le Sporting d'Anderlecht, j'ai demandé à mon manager de me trouver un club en Angleterre´, avoue le Norvégien du Standard. `Aucune piste concrète ou alléchante ne put être exploitée. Et puis, ma vie a changé quand Elin, ma compagne, est venue me rejoindre en Belgique il y a trois ans. Les virées nocturnes entre amis sont, dès lors, devenues de plus en plus rares (rires). Mais cela ne m'a jamais posé le moindre problème car je n'étais quand même pas un coutumier du fait. Son arrivée m'a permis de me stabiliser. J'apprécie, désormais, les soirées entre nous à la maison au chaud et au calme. Et puis, nous avons eu Mina. Notre fille est âgée d'un an et demi mais elle est très évoluée. Elle sait se faire comprendre et répète tous les mots que nous prononçons. Nous devons désormais faire très attention à tout ce que nous disons à la maison. Sa présence a bouleversé notre vie. Quand je rentre de l'entraînement ou d'un match, j'oublie tous mes soucis quand elle m'offre son sourire. J'aurais pu tenter une autre expérience à l'étranger mais je dois aussi penser à elle et à son éducation. Je sais que la plupart des footballeurs n'hésitent pas à placer leurs enfants dans des écoles privées. Ils m'ont dit que les enfants

pouvaient s'adapter. Moi, j'ai envie de voir grandir ma fille. C'est une des raisons pour laquelle je souhaite revenir en Norvège à l'issue de mon contrat avec le Standard.´

ANGOULÊME Il n'est pas encore midi et une foule compacte se presse déjà devant le buste d'Hergé, au coeur d'une rue piétonne de la cité angoumoisine. Le prince Philippe et la princesse Mathilde sont attendus de pied ferme par des centaines, des milliers d'habitants de cette petite ville de Charentes qui, l'espace de quatre, jours s'approprie le titre de capitale internationale de la bande dessinée.

Segment 69

Tout y est. Les classiques: biographie, actualités (on y apprend que Bjork offrira, non pas un mais deux concerts à Paris Bercy, les 16 et 17/6), galeries photos, liens, livre d'or,... Mais aussi et surtout les plus: la galerie photos comporte plus de 500 images toutes superbes et notamment des dizaines de pictures pour chaque clip mais il y a aussi une galerie d'art. Celle-ci rassemble les oeuvres de fan.

Toutes font apparaître la chanteuse dans des montages visuels magnifiques que vous pourrez mettre en fond d'écran.

A noter aussi, les paroles de toutes ses chansons, y compris lorsqu'elle chantait au sein du groupe The Sugarcubes, et les photos de ses collaborations avec les plus grands stylistes de la planète.

Segment 72

Ce n'est pas son genre, il n'a aucune imagination. Il n'est pas historien, il sait des choses, il connaît bien ses archives et la généalogie locale. Mais il n'est pas romancier. Il est précis et technique. Et ça, j'aime bien. Quand il me parle de la calèche et qu'il me la montre en photo, elle est exactement comme dans son souvenir.´

Vous parlez de divertissement en évoquant ce livre. Cela a donc été plus amusant à écrire?

`Oui. Mais cela ne veut pas dire pour la cause que c'est parce que je me sentais plus proche d'elle que de mes autres victimes , qui sont à cheval sur le 18e et le 19 e siècle. Aglaé, elle ne fait rien, il ne se passe rien. Elle est fille, femme, mère d'officier. Elle tient bien son rôle, elle comprend, elle est une femme comme il y en avait des millions en France. Elle fait partie de la petite gentilhommerie terrienne pas riche mais solide, bien ancrée dans ses valeurs, son patrimoine culturel. Aglaé fait partie de ces femmes qui ont été des âmes et des piliers de la société. Ce ne sont pas des petites bonnes femmes qui divorcent tous les trois mois.

Elles ont une constance, une stabilité, un entendement qui explique pourquoi la société, jusqu'à la guerre de 14, n'a pas été un tissu social décomposé. Je ne suis pas féministe, mais des femmes comme elles étaient formidables.´

Segment 73

Vous évoquez l'existence de photos d'Aglaé, mais vous choisissez de mettre une reproduction de peinture sur la couverture, pourquoi?

`Ce n'était pas possible. C'était des photos de Nadar, en sépia. Aglaé était déjà assez forte. Mais ça aurait donné quelque chose de triste. J'ai donc choisi le portrait de la duchesse de Montebello dans le grand tableau des Dames d'honneur de l'impératrice Eugénie parce qu'elle est plausible. D'abord, elle est complètement d'époque et elle lui ressemble un petit peu.´

Françoise Wagener, Aglaé, Ed. Flammarion.

Segment 87

Au moment où son album solo Read My Lips en est à son troisième single (Get over you), nous n'avons pas manqué de lui faire remarquer que Groovejet était peut-être un petit peu plus tendance et moins popu que ce qu'elle livre désormais. Sa réponse révèle tout de suite un caractère bien trempé.

`Désolée mais il ne me semble pas que Groovejet soit moins commercial que Murder on the dancefloor. Il l'est même plus! Fais-moi confiance, je sais qu'un succès grand public n'est pas un gage de qualité. Je n'ai jamais agi dans cette optique et la première personne que je dois satisfaire quand je fais de la musique, c'est moi. Ce qui m'intéresse, c'est de créer des tempos variant d'un titre à l'autre, un pour le vendredi soir, un pour le dimanche après-midi ou un autre pour le mercredi matin. Et, au niveau des textes, faire preuve de maturité. Comme Bowie ou Debbie Harry.´

A bientôt 24 ans (le 10 avril), Sophie n'est donc pas qu'une belle plante lookée 50ies (`j'adore cette époque, Rita Hayworth, Bette Davis, Vivien Leigh, Marilyn´) mais se passionne et se déchaîne pour un tas de choses. A commencer par la vie! Dans son discours, elle encourage les femmes à saisir les opportunités qu'elles méritent. Dans ses chansons, et notamment sur Move the mountain, elle soulève que le sexe fort gère moins facilement que le sexe faible les difficultés du quotidien. Et puis, porte-parole de luxe, elle est actuellement en campagne contre la fourrure naturelle. Répondant à l'appel de l'association PETA.

Segment 88

`Il n'y a rien de pire qu'un artiste qui se regarde le nombril et ne communique que sur son ego. Moi, je m'enflamme, je m'indigne, mon cerveau est sans cesse en ébullition, je suis impatiente et du genre à prendre les devants et il est donc normal que tout cela se traduise en musique. Même si je ne me prends jamais au sérieux. Murder on the dancefloor est un bon exemple. Cette chanson dénonce le côté compétitif de notre société mais c'est la légèreté et l'ironie qui l'emportent.´

Si l'on ajoute qu'elle est très fan de Nabokov (`j'aime son esprit, très noir et très drôle´), on comprend mieux pourquoi Sophie s'énerve quand on la bassine avec son soi-disant passé de mannequin. `J'ai fait ça pendant six semaines, il y a trois ans, en me disant que c'était plus fun que de bosser dans une boulangerie. Mais, je le répète, rien ne me fait autant vibrer que la musique.´

Elle parle français!

Lexique de la classe selon l'occurence des mots dans les segments.

Aglaé avoir devoir faire femme galerie photo pouvoir standard tottenham

Commentaire :

Galerie signifie une galerie de photo et non un balcon avec des escaliers.

****

Classe 157

****

Segment 27

Pour nous, la deuxième solution serait beaucoup plus intéressante. Techniquement, il s'agit de devenir un opérateur de téléphonie mobile sur réseau virtuel. Ce qui signifie que sans avoir notre propre réseau nous pourrions utiliser l'un de ceux déjà existants et proposer nos propres tarifs. Ce système existe dans d'autre pays, comme au Danemark, en Hollande ou en Autriche. Mais en Suisse ce n'est pour l'instant pas possible, car aucun des trois opérateurs ne veut nous héberger. Et je le répète : trois opérateur sur un marché, ce n'est pas suffisant pour créer une situation de concurrence. Résultats : les tarifs de téléphonie fixe à mobile ou de mobile à mobile sont les plus élevés d'Europe.

Selon vous la libéralisation n'a donc pas vraiment fonctionné en Suisse ?

Non, car des erreurs ont été commises lors de l'ouverture du marché et de la distribution des licences pour la téléphonie mobile. Il a visiblement été oublié de préciser que les exploitants de réseaux devaient héberger d'autres opérateurs. Je sais que c'était une volonté initiale de l'Office fédéral de la communication (OFCom) de rendre cet hébergement possible. Mais quelque chose n'a visiblement pas fonctionné. J'ai obtenu une licence GSM, mais elle ne me sert à rien. Ce qui est possible dans d'autre pays est visiblement très difficile en Suisse.

Segment 53

Chaque religion a apporté sa réponse. Pour les judéo-chrétiens, Dieu, infiniment bon, ne peut être à l'origine du mal. D'où le vieux mythe biblique du péché originel commis par Adam et Ève en mangeant le fruit défendu. Saint Paul et surtout saint Augustin appuient leur enseignement sur ce récit fondateur.

Au XVIe siècle, le concile de Trente en fait un dogme qui devient le fondement du christianisme : le Christ serait venu sur terre pour sauver les hommes des conséquences du péché d'Adam, et le baptême serait pour chacun l'instrument de ce rachat. Toutefois les hommes resteraient foncièrement mauvais et ne pourraient échapper à l'enfer qu'en luttant contre le mal, avec la grâce de Dieu le Père.

Georges Minois conte avec érudition et clarté la longue histoire du péché originel : ses interprétations diverses, au sein de l'Église et hors d'elle, de saint Augustin à Teilhard de Chardin, de Paul Ricœur à Jean Paul II, en passant par Luther, Pascal, Voltaire.

Segment 89

Pour couronner le tout, il arrive à Sophie de s'exprimer dans un français plus que potable. Ce qui tombe bien puisqu'elle sera sur la scène de l'Ancienne Belgique le 17 février. Nous lui avons demandé d'énumérer dans notre langue ses coups de coeur belges. `J'aime beaucoup La Madone de Bruges Michel-Ange of course et les moules marinières. Avec des frites, de la mayonnaise et une bière blonde.´ Chapeau! Sûr qu'à la sortie de son prochain album (`on n'en est qu'aux démos´), aucun admirateur belge n'aura oublié cet exploit. Merci Sophie.

HOLLYWOOD Steven Spielberg possède une passion pour l'histoire, celle de l'humanité ou des Etats-Unis. `Transmettre la mémoire constitue une de mes plus grandes missions´, répète-t-il souvent. Jurassic Park, La couleur pourpre, L'empire du soleil, Amistad, La liste de Schindler et Il faut sauver le soldat Ryan sont là pour le prouver. A cette liste impressionnante, il va falloir prochainement ajouter un nouveau long métrage, consacré à la vie d'Abraham Lincoln. L'élection à la présidence des Etats-Unis, en 1861, de cet anti-esclavagiste convaincu fut le détonateur de la guerre de Sécession. Une étape capitale de l'histoire américaine, trop peu connue des jeunes Américains. Une lacune que le wonder boy d'Hollywood compte combler cinématographiquement. Le scénariste John Logan travaille déjà sur le projet d'arrache-pied. Pour le rôle principal, Spielberg n'envisagerait qu'un seul acteur: Tom Hanks. Et pas seulement en raison de leur amitié et de leur passion commune pour les grands faits historiques. D'après de très sérieuses recherches, Tom Hanks serait relié généalogiquement à Nancy Hanks, la maman du seizième président des Etats-Unis, à savoir Abraham Lincoln.

Ce projet a été confirmé par Steven Spielberg cette semaine: `Je pense que l'histoire est capitale. Je veux faire un film sur Abraham Lincoln; je travaille d'ailleurs au développement d'un script à ce sujet avec Frederick Douglass. Je le réaliserai probablement dans les deux prochaines années.´ Voire, d'après son entourage, avant Indiana Jones 4, dont la sortie est prévue pour 2004. Et Tintin? Il passera sans doute après le président des USA

Lexique de la classe selon l'occurence des mots dans les segments.

Abraham avoir etats hanks histoire lincoln mobile paul péché possible réseau saint spielberg suisse téléphonie uni visiblement.

Commentaire :

Il s'agit d'Abraham Lincoln et non d'une voiture de marque lincoln.

****

Classe 206

****

Segment 15

Alimentation: le bon choix de matières grasses dans votre assiette

La chasse au cholestérol n'est plus d'actualité pour lutter contre les maladies cardio-vasculaires: l'accent est mis sur un choix judicieux des matières grasses. Explications et conseils.

Dans les pays industrialisés, les maladies cardio-vasculaires constituent une des premières causes de mortalité précoce. Des relations entre l'alimentation et ces pathologies ont été clairement établies: une consommation élevée de matières grasses accroît le risque de maladie, alors que la consommation de fruits, légumes et poisson le réduit.

Segment 16

Par maladies cardio-vasculaires, on entend, entre autres, l'infarctus, l'athérosclérose (modification de la paroi des artères) et ses complications, thrombose et embolie. Les troubles des lipides (ou graisses) sanguins qui correspondent au terme barbare de «dyslipidémies» et dont l'excès de cholestérol est un exemple constituent un facteur de risque des maladies cardio-vasculaires.

LIEN DE CAUSALITÉ

Quel est le lien entre les graisses alimentaires et les maladies cardio-vasculaires? Une consommation de graisses excessive de même que la qualité des graisses consommées (origine animale ou végétale) influencent les transformations biochimiques qui ont lieu dans l'organisme. Des déséquilibres entre les graisses sanguines peuvent alors survenir: trop de «mauvais» cholestérol, pas assez de «bon», par exemple.

Segment 18

Les graisses alimentaires peuvent être visibles (graisses d'adjonction) comme les huiles, le beurre, les margarines et minarines, la crème, la mayonnaise, les noix, noisettes, olives, etc. Ou alors elles peuvent être cachées dans les aliments, par exemple dans les pâtisseries et douceurs (chocolat, glaces, gâteaux, etc.), dans les laitages (lait, yogourt, fromages, etc.), dans les viandes, le poisson, les oeufs. Les aliments du groupe des féculents (pain, pâtes, pommes de terre, etc.) ne contiennent pas de matières grasses, sauf si on en rajoute (pommes frites, pâte à pain avec des corps gras, croissants, etc.). Les légumes et fruits ne contiennent pas de graisses.

Pour respecter les recommandations nutritionnelles, il ne faudrait pas dépasser 30 à 35 g de graisses visibles par jour, le reste étant caché dans les aliments (viande, laitages, etc.). Cela représente environ 2 cuillères à soupe d'huile pour la ration de notre exemple à 2000 kcal. Afin que les différents acides gras soient représentés dans

de bonnes proportions, les graisses visibles de source végétale devraient être préférées (huiles ou graines comme les noix, amandes, etc.).

RECOMMANDATIONS NUTRITIONNELLES

Segment 20

Farineux: préférez les produits non gras (pain au lieu de croissants); gardez les spécialités (tresse, cuchaule, pâte à tarte) pour les occasions (le dimanche par exemple).

Fruits et légumes: ils ne contiennent pas de graisses, profitez-en!

Non seulement un excès de graisses peut s'avérer défavorable à la santé, mais la qualité des graisses est également déterminante. C'est pourquoi il convient de choisir avec soin les graisses d'assaisonnement (voir les recommandations), tout en faisant la part belle aux fruits, légumes et féculents non gras dans l'alimentation. Tous les aliments ont un rôle à jouer dans la prévention des maladies cardio-vasculaires!

Lexique de la classe selon l'occurence des mots dans les segments.

Aliment cardio graisse gras maladie matière vasculaire visible

Commentaire :

Thème : L'alimentation et les maladies cardio vasculaire.

****

Classe 229

****

Segment 10

La hausse serait de 14% pour les exploitations spécialisées dans la viande et de 4% en élevage laitier en raison de la baisse de 2,5% du prix du lait sur l'année 2002.

La hausse des prix des bovins «finis» en 2002 serait de 4%, mais celle des bovins maigres serait de 25% après le marasme de l'année 2001. Les cours de bovins restent inférieurs d'environ 10% à ce qu'ils étaient avant la crise. Les éleveurs compensent cette situation en déstockant une part substantielle de leur cheptel. De plus, le revenu profite de l'augmentation sensible des aides européennes (+17%) ainsi que de quelques reliquats d'aides françaises versées après la crise de la vache folle.

Le résultat moyen diminuerait de 9% en viticulture d'appellation et de 8% dans les autres exploitations viticoles avec un résultat plus favorable dans la région de Cognac. Le niveau des prix en viticulture d'appellation pour la campagne 2002-2003 serait stable en moyenne annuelle avec variations selon les régions. La récolte serait en

baisse de 5% par rapport à celle de 2001. Pour les vins non AOC, la récolte serait aussi fortement en baisse, notamment dans les régions touchées par les inondations. Les prix devraient se redresser de près de 10%.

Segment 77

L'euro évolue actuellement à son plus haut niveau depuis juin 1999 face à la livre sterling, au dessus du seuil de 0,6640. Mais `les indicateurs économiques fondamentaux de la zone euro ne justifient aucunement une telle appréciation´, souligne cette économiste.

Le chef économiste de la Banque centrale européenne, Otmar Issing, s'est lui-même inquiété récemment de la faiblesse de l'économie allemande.

`Bien sûr, le fait que le plus grand pays de la zone euro soit à la traîne en termes de croissance et de création d'emplois est pour nous une source de préoccupation´, avait-t-il déclaré dans une interview à l'AFP.

Segment 90

ZAVENTEM Après l'UBCNA, l'association Bruxelles Air Libre a dit lundi tout le mal qu'elle pense de l'accord intervenu au sein du Comité Etat Régions au sujet de la gestion des nuisances sonores liées aux vols de et vers l'aéroport de Bruxelles-National. Selon Bruxelles Air Libre, il fallait en effet choisir entre les vols de nuit ou un accroissement de 20% du trafic de jour au-dessus de Bruxelles.

Cette décision d'organiser des décollages nocturnes traversant de part en part la Région la plus densément peuplée du pays est `tout simplement écoeurante´, a insisté l'association qui voit en celle-ci l'anéantissement de 10 ans de lutte des associations et de quatre ans de travail du cabinet Durant.

De son côté, l'Union belge contre les nuisances des avions, association composée de riverains, de bourgmestres et mandataires des principales communes de la Région de Bruxelles-Capitale et de la Région flamande avait quant à elle contesté le contenu de l'accord, quelques heures après sa conclusion.

Lexique de la classe selon l'occurence des mots dans les segments.

Association baisse bovin bruxelles euro prix région

Commentaire :

Le thème économique et en particulier le cours des bovins.

4.2 Classification au moyen de réseau de neurones ART.

La classification, au moyen du réseau de neurones ART avec un paramètre de vigilance de 0.05, donne lieu à la production de 53 classes de segments présentants des similarités. Examinons maintenant quelques résultats :

****

Classe 3

****

Segment 8

Filiale du groupe Ruag, Mecanex réalise la moitié de ses activités dans le domaine spatial, l'autre moitié dans le secteur industriel et militaire. Connue pour la réalisation de ses collecteurs de courant pour les panneaux solaires de satellites, la firme nyonnaise participe au projet de sonde interplanétaire Rosetta. Ce satellite, dont le lancement avec Ariane 5 prévu mi-janvier a été reporté, ira se poser sur la tête d'une comète. Mecanex a mis au point notamment le générateur d'ions de l'appareil Rosina. Celui-ci est installé sur le satellite et sera utilisé par l'Université de Berne pour mesurer la densité de la tête de la comète. la promet. Le revenu 2002 est en diminution de 1%; forte baisse dans l'élevage hors sol.

Segment 9

En dépit de bonnes récoltes et d'une sortie progressive de la crise de la vache folle, le revenu agricole 2002 est en diminution de 1%. Ainsi en a conclu la Commission des comptes de l'agriculture réunie le mercredi 18 décembre. Cause principale de cette diminution globale: la très forte contre-performance des productions hors sol (porcs et volailles) dont le résultat par actif, déduction faite des effets de l'inflation, est en diminution de 37%. Les experts notent également une régression sensible du résultat agricole pour la viticulture, qu'il s'agisse des vins AOC (-9%) ou des vins courants (-8%). Le revenu de l'élevage hors sol, de nature très fluctuant, diminuerait de 37%. Les prix du porc pour l'année 2002 sont en retrait de 21% par rapport à 2001. La production avicole devrait baisser de 2% en volume et de 3% en prix. Les deux secteurs paient le contrecoup d'une sortie de la crise de la vache folle: la demande a chuté après deux années de hausse où elle avait profité du désintérêt pour la viande de boeuf. En dépit des deux années précédentes de hausses, le résultat moyen des élevages hors sol revient sensiblement à son niveau du début des années nonante. Il reste néanmoins au-dessus de la moyenne des exploitations professionnelles. Après la crise bovine en 2000 et début 2001, le revenu serait en progression de 7% en élevage bovin (lait et viande).

Segment 10

La hausse serait de 14% pour les exploitations spécialisées dans la viande et de 4% en élevage laitier en raison de la baisse de 2,5% du prix du lait sur l'année 2002. La hausse des prix des bovins «finis» en 2002 serait de 4%, mais celle des bovins maigres serait de 25% après le marasme de l'année 2001. Les cours de bovins restent inférieurs d'environ 10% à ce qu'ils étaient avant la crise. Les éleveurs compensent cette situation en déstockant une part substantielle de leur cheptel. De plus, le revenu profite de l'augmentation sensible des aides européennes (+17%) ainsi que de quelques reliquats d'aides françaises versées après la crise de la vache folle. Le résultat moyen diminuerait de 9% en viticulture d'appellation et de 8% dans les autres exploitations viticoles avec un résultat plus favorable dans la région de Cognac. Le niveau des prix en viticulture d'appellation pour la campagne 2002-2003 serait stable en moyenne annuelle avec variations selon les régions. La récolte serait en baisse de 5% par rapport à celle de 2001. Pour les vins non AOC, la récolte serait aussi fortement en baisse, notamment dans les régions touchées par les inondations. Les prix devraient se redresser de près de 10%.

Lexique de la classe selon l'occurence des mots dans les segments.

Année baisse bovin crise diminution élevage prix région résultat revenir satellite sol

Commentaire :

Le thème commun entre ces segments est l'élevage des bovins.

****

Classe 33

****

Segment 58

irakien". Ils étaient des milliers à le clamer lors du meeting de solidarité organisé dimanche par le Comité national marocain de soutien à l'Irak. Un meeting auquel ont participé les partis politiques, les organisations de droits de l'Homme et les associations professionnelles et éducatives. Des personnalités publiques, des ministres, des leaders politiques comme M.Abderrahman Youssoufi, premier secrétaire de l'USFP, M.Ismaïl Alaoui, premier secrétaire du PPS, ou encore M.M'Hamed Boucetta, membre du Conseil de présidence du Parti de l'Istiqlal, pour ne citer que ceux-là, ont tenu à être présents lors de cette manifestation de

solidarité en vue d'exprimer leur soutien inconditionnel au peuple irakien. Touria Jabrane, artiste connue pour son engagement en faveur des peuples irakien et palestinien a ouvert le meeting par une prestation qui illustre les souffrances de ces peuples meurtris par l'embargo et par la guerre. C'est un rassemblement où l'art dans toutes ses formes, musique, chants, poésie, a remplacé les discours habituels prononcés dans ce genre d'événements. Les artistes, chacun à sa manière, ont exprimé leur soutien au peuple irakien et leur indignation face à la montée du spectre de la guerre contre l'Irak. Le doyen de la chanson marocaine Abdelwahab Doukali a ouvert le bal par sa célèbre chanson "Souk Al Bacharia".

Segment 59

La poésie a pris le relais et laissé les mots exprimer toutes les souffrances et tous les affres en courus de l'Homme. D'abord à travers un poème de Driss El Meliani puis un autre en langue amazighe d'Ibrahim Lachguer. La voix de l'enfance a été également présente et a tenu, elle aussi, à exprimer son soutien aux enfants irakiens. Les Rbatis, qui ont répondu spontanément et massivement à l'appel des organisateurs, ont pu écouter la voix d'une jeune artiste qui a entonné une chanson libanaise qui en dit long sur ces causes arabes ou encore l'enfant Said, accompagné de l'artiste Benbrahim, qui a enchanté le public par sa voix angélique. Un autre artiste, "Mouhsine", certes méconnu du public, a réussi à l'assistance par sa chanson "Baghdad", qui nous rappelle, dans sa composition, la célèbre chanson sur "Baghdad" de Kadem Saher. Place après au luth et à l'excellent Haj Younes qui a clos ce programme musical signant ainsi l'engagement des artistes marocains en faveur de la cause irakienne. Bien entendu, ce programme animé par Touria Jabrane, était à chaque fois ponctué de slogans clamés par le public. Encore un crime crapuleux à Casablanca. Samedi matin dans la décharge de Médiouna, on découvre le cadavre découpé d'un homme. La gendarmerie Royale trouvera au total, durant les fouilles, 8 sachets en plastique contenant des parties du corps mutilé et décapité. L'enquête préliminaire a permis de déterminer l'identité de la victime : A B., d'une quarantaine d'années, célibataire et ayant vécu avec une amie pendant sept ans. Le lieu où a eu le crime est, selon les enquêteurs, le domicile même de la victime qui se trouve à Hay Al Ahd Al Jadid sur la route de Médiouna.

Segment 60

A présent, les enquêteurs sont à la recherche de celui ou de ceux qui auraient commis le crime. Deux hommes et deux femmes, ayant eu un lien avec la victime, sont, actuellement, interrogés par la gendarmerie. L'enquête s'annonce longue et pourrait, vraisemblablement, buter sur le trou noir pour compléter cette série de meurtres aussi mystérieuse qu'ignoble qu'à connue la ville casablancaise, depuis septembre 2002. Les quartiers de la Gironde et Hay Hassani ont été le théâtre de découvertes macabres dont les protagonistes étaient de jeunes femmes sauvagement assassinées, découpées et réparties dans des caisses en plastique. Des meurtres qui transformaient l'identification de la victime, une étape décisive dans l'orientation de l'enquête, en une tâche

éprouvante et difficile. En rappel aux faits, le 04 décembre, le cadavre découpé d'une femme a été découvert dans une décharge à Hay Hassani. Son état en décomposition avancé a posé de gros problèmes à l'autopsie qui n'est parvenue qu'à avancer l'estimation de son âge en le situant à la quarantaine. Quant aux causes de la mort, elles ne sont pas déterminées à 100%, car les examens pour ce cas-là s'orientaient au départ vers la strangulation avant d'émettre des doutes.

Lexique de la classe selon l'occurence des mots dans les segments.

Artiste avoir chanson femme irakien meeting peuple public soutien voix.

Commentaire :

Thème : Artiste pour soutenir le peuple irakien.

****

Classe 5

****

Segment 15

Alimentation: le bon choix de matières grasses dans votre assiette La chasse au cholestérol n'est plus d'actualité pour lutter contre les maladies cardio-vasculaires: l'accent est mis sur un choix judicieux des matières grasses. Explications et conseils. Dans les pays industrialisés, les maladies cardio-vasculaires constituent une des premières causes de mortalité précoce. Des relations entre l'alimentation et ces pathologies ont été clairement établies: une consommation élevée de matières grasses accroît le risque de maladie, alors que la consommation de fruits, légumes et poisson le réduit.

Segment 16

Par maladies cardio-vasculaires, on entend, entre autres, l'infarctus, l'athérosclérose (modification de la paroi des artères) et ses complications, thrombose et embolie. Les troubles des lipides (ou graisses) sanguins qui correspondent au terme barbare de «dyslipidémies» et dont l'excès de cholestérol est un exemple constituent un facteur de risque des maladies cardio-vasculaires. LIEN DE CAUSALIT? Quel est le lien entre les graisses alimentaires et les maladies cardio-vasculaires? Une consommation de graisses excessive de même que la qualité des graisses consommées (origine animale ou végétale) influencent les transformations biochimiques qui ont lieu dans l'organisme. Des déséquilibres entre les graisses sanguines peuvent alors survenir: trop de

«mauvais» cholestérol, pas assez de «bon», par exemple.

Lexique de la classe selon l'occurence des mots dans les segments.

Cardio graisse gras maladie matière vasculaire.

Commentaire :

Les maladies cardio-vasculaires et la graisse.

****

Classe 6

****

Segment 17

Les éléments de base des graisses sont les acides gras. Comme un collier peut être composé de perles de différentes couleurs, une huile ou du beurre ne contiennent pas forcément les mêmes perles (ou acides gras). Et deux huiles peuvent contenir des perles de couleurs identiques, mais qui ne seront pas arrangées de la même façon! Ces différences de composition vont conférer des propriétés particulières aux aliments. Par exemple, à température ambiante une huile est fluide, le beurre est solide, etc. (voir le tableau 1). VISIBLES OU CACHÉES La quantité totale de graisses fournie par l'alimentation devrait avoisiner 30% des apports énergétiques. Une quantité plus élevée favorise l'apparition de l'obésité, des troubles des lipides sanguins (élévation du taux de cholestérol par exemple), etc. Concrètement, cela représente environ 60 à 70 g de graisses pour une ration à 2000 kcal. De plus, étant donné leurs effets sur les lipides sanguins, les différents acides gras doivent être représentés dans certaines proportions.

Segment 18

Les graisses alimentaires peuvent être visibles (graisses d'adjonction) comme les huiles, le beurre, les margarines et minarines, la crème, la mayonnaise, les noix, noisettes, olives, etc. Ou alors elles peuvent être cachées dans les aliments, par exemple dans les pâtisseries et douceurs (chocolat, glaces, gâteaux, etc.), dans les laitages (lait, yogourt, fromages, etc.), dans les viandes, le poisson, les oeufs. Les aliments du groupe des féculents (pain, pâtes, pommes de terre, etc.) ne contiennent pas de matières grasses, sauf si on en rajoute (pommes frites, pâte à pain avec des corps gras, croissants, etc.). Les légumes et fruits ne contiennent pas de graisses. Pour respecter les recommandations nutritionnelles, il ne faudrait pas dépasser 30 à 35 g de graisses

visibles par jour, le reste étant caché dans les aliments (viande, laitages, etc.). Cela représente environ 2 cuillères à soupe d'huile pour la ration de notre exemple à 2000 kcal. Afin que les différents acides gras soient représentés dans de bonnes proportions, les graisses visibles de source végétale devraient être préférées (huiles ou graines comme les noix, amandes, etc.). RECOMMANDATIONS NUTRITIONNELLES

Segment 19

Matières grasses: préférez les huiles de colza, noix, olive, soja ou associez deux huiles pour équilibrer les acides gras (colza + noix; colza + olive; colza + soja ou noix + olive); le beurre a sa place en petites quantités, sur les tartines ou pour affiner un mets. Evitez les margarines (elles contiennent souvent des acides gras «trans»). Gardez la crème pour des plats de fête occasionnels. Méfiez-vous des pâtes à tartiner (Nutella, Parfait) et de la mayonnaise, elles sont grasses! Viandes, poissons, oeufs: préférez les viandes maigres afin de limiter les graisses animales; cuisinez du poisson deux fois par semaine, il est riche en acides gras protecteurs; vous pouvez manger deux oeufs par semaine en cas d'hypercholestérolémie. Laitages: mangez le fromage à la place de la viande (repas sans viande par exemple).

Segment 20

Farineux: préférez les produits non gras (pain au lieu de croissants); gardez les spécialités (tresse, cuchaule, pâte à tarte) pour les occasions (le dimanche par exemple). Fruits et légumes: ils ne contiennent pas de graisses, profitez-en! Non seulement un excès de graisses peut s'avérer défavorable à la santé, mais la qualité des graisses est également déterminante. C'est pourquoi il convient de choisir avec soin les graisses d'assaisonnement (voir les recommandations), tout en faisant la part belle aux fruits, légumes et féculents non gras dans l'alimentation. Tous les aliments ont un rôle à jouer dans la prévention des maladies cardio-vasculaires!

Lexique de la classe selon l'occurence des mots dans les segments.

Acide aliment colza graisse gras noix olive perle viande visible.

Commentaire :

Produit alimentaire.

4.3 Comparaison.

4.3.1 Indices quantitatifs :

Le texte qu'on a choisi est un mélange de sujets différents. On a fait une segmentation par 3 paragraphes. Nous avons comparé les résultats des classifieurs ART et treillis de Galois, en constatant d'abord qu'ART trouve moins de classes avec 53 classes pour un paramètre de vigilance égal à 0,05 contre 284 classes pour les treillis de Galois avec un seuil égal à 0,01, voici quelques résultats statistiques.

	Treillis de Galois	ART
1 segment	78 classes	26 classes
2 segments	105 classes	11 classes
3 segments	43 classes	10 classes

Tableau 4.1 - Nombre de classes trouvées par Treillis de Galois et ART en fonction du nombre de segments dans la classe.

Ce tableau nous montre une comparaison statistique du nombre des classes trouvées par les deux classifieurs, en fonction du nombre de segments dans la classe.

Bien que puissant pour des applications textuelles, ART présente l'handicap de ne pas pouvoir faire appartenir un même segment à plusieurs classes.

Le tableau ci-dessous illustre quelques segments et leur fréquence d'apparition dans les différentes classes.

Numéro de segments	Fréquence d'apparition des segments
Segment N 2	16 classes
Segment N 5	10 classes

Segment N 10	13 classes
Segment N 15	17 classes
Segment N 20	13 classes

Tableau 4.2 - Illustration du nombre d'apparition des segments dans les classes avec le Treillis de Galois.

Parmi les avantages de la classification par le Treillis de Galois, le fait qu'un segment puisse appartenir à plusieurs classes, la question qui se pose est : est-ce-que cet avantage apportera un plus pour la classification?

La réponse à cette question doit faire l'objet de recherches futures.

4.3.2 Resultats qualitatifs.

Le texte que nous avons choisi est un mélange de plusieurs textes appartenant à des auteurs différents. Examinent d'une manieres qualitative des classes.

Treillis de Galois.

Classe 281 regroupe les segments 1, 2, 15, 16 et 93. Le lexique de cette classe formé de l'intersection des lexiques des segments de la classe est l'ensemble vide. Le segments 1 et 2 traite le problème du chauffage des bâtiment a distance et au bois. Les segments 15 et 16 parle des maladies cardio-vasculaires, tandis que le segment 93 parle de la guerre en Irak.

Classe 43 regroupe les segments 76 et 77. L'intersection des lexiques des deux segments est forme par : euro européenne faire grand livre sterling. Les deux segments parlent effectivement de la devise européenne euro.

Classe 206 regroupe les segments 15, 16, 18, 20. Le lexique issu de l'intersection des lexiques de segments de la classe est l'ensemble vide. Par contre les quatre segments parlent de l'alimentation.

Classe 229 regroupe les segments 10, 77 et 90. Le lexique de l'intersection de cette classe est l'ensemble vide. Le segment 10 parle de la hausse des prix des boivin. Le segment 77 dit que l'euro évolue actuellement à son plus haut niveau depuis juin 1999 face à la livre sterling, pendant que le segment 90 parle des nuisances sonores liées aux vols aériens vers l'aéroport de bruxelles. Quand nous avons vu dans la premier interprétation des résultats le lexique de cette classe selon une occurrence de mots inférieur à 3 nous avant trouvé : Association baisse bovin Bruxelles euro prix région et nous avons dit que le thème économique et en particulier le cours des bovins pour cette classe.

ART

Classe 3 regroupe les segments 8, 9 et 10. Le lexique de cette classe formé de l'intersection des lexiques est constitué par : élevage revenir effectivement les trois classes parlent de l'élevage des boivin.

Classe 5 rassemble les segments 15 et 16. Le lexique de l'intersection de ces segments donne : avoir cardio cholestérol consommation constituer maladie risque vasculaire. Les deux classes parlent de l'alimentation et les maladies cardio-vasculaires.

Classe 33 regroupe les segments 58,59 et 60. Le lexique de l'intersection de ces segments est : avoir. Les segments 58 et 59 exprime le soutien des Artistes, des parties politiques...etc. Au peuple irakien. Le segment 60 parle d'une enquête sur un crime a casablanca.

Conclusion.

Ce mémoire traite du problème de la classification textuelle en général et de la place des Treillis de Galois en particulier. L'extension des algorithmes de construction (une normalisation des vecteurs à classifier) des treillis de Galois et la comparaison avec un réseau de neurones ART ont été les objectifs principaux de notre travail.

- Premier résultat pour les deux classifieurs, le nombre de classes augmente si la taille des segmentations diminue.
- Nombre de classes :

  Le Treillis de Galois a plus de classes que le réseau ART. Bien qu'avec le Treillis de Galois, en diminuant le paramètre S on obtienne moins de classes, on perd par contre de l'information pertinente sur le texte.
- La Stabilité du treillis :

  Même si on change l'ordre des segments, le treillis de Galois est indépendant de l'ordre des segments à classer. Par contre ART change sa classification.
- La Méthodologie :

  La classification par treillis de Galois ne commence pas par des poids aléatoires comme le réseau ART. La classification est algébrique (logique).
- ART a l'interdiction d'avoir des segments appartenant à des différentes classes.

Ceci dit, il serait hasardeux d'arriver à une conclusion dans laquelle on affirmerait qu'un classifieur est meilleur qu'un autre. Ce qui est sur c'est que chaque classifieur a des propriétés particulières qui vont se refléter sur la qualité du résultat. C'est ce que d'ailleurs nous avons tenté de montrer.

Bibliographies.

[Barbut & Monjardet 1970], *Ordre et classification.* Algèbre et Combinatoire, **tome II**, Hachette.

[Balpe & al 1996], *Techniques avancées pour l'hypertexte.* Paris, Hermes.

[Bordat 1986], *Calcul Pratique du Treillis de Galois d'une Correspondance*, Mathématiques et Sciences Humaines, 96, 1986, p. 31-47.

[Carpineto 1993], *GALOIS: An Order-Theoretic Approach to Conceptual Clustering, Proceedings* of *thé Machine Learning Conférence* (1993), p. 33-40.

[Chein 1969], *Algorithme de Recherche des Sous-Matrices Premières d'une Matrice,* Bull. Math. Soc. Sci. Math. R.S. Roumanie, 13, 1969, p. 21-25.

[Claude 1992], *les reseaux de neuronnes Artificiels. Introduction Au connexionnisme.*

[Damashek 1995], *Similarity with n-Grams : Language-Independent Categorization Of Text*, Science, 267, 843-848.

[Fay 1975], *An Algorithm for Finite Galois Connexions*, Journal of' Computational Linguistic and Languages, 10, 1975, p. 99-123.

[Ganter 1986], *Conceptual Measurement and Many-Valued Contexts*, in Classification as a Tool of Research, W. Gaul, M. Schader (Eds.), NorthHolland, p. 169-176, 86.

[Godin 1991], *Learning Algorithms Using a Galois Lattice Structure,* Proceedings of the Third International Conference on Tools for Artificial Intelligence (1991), p. 22-29.

[Godin 1995], *Méthodes de classification conceptuelle basées sur les treillis de Galois et applications,* Revue d'Intelligence Artificielle, 9(2):105-137, 1995.

[Greffenstette 1995], *Comparing Two Language Identification Schemes*, Actes de JADT-95, 85-96

[Guénoche 1990], *Construction de treillis de Galois d'une relation binaire*, Mathématiques et sciences Humaines, 1990,109 :41-53.

[Halleb & Lelu 1998], *Hypertextualisation Automatique Multilingue à Partir des Fréquences de n-Grammes*, Actes de JADT-98, Nice, France.

[Hebb 1949], *The Organisation of Behavior*, New York: Wiley, 1949.

[Hopfield 1982] *Neural networks and physical systèms with emergent collective abilities*, Proceedings of the National Academy of Sciences, USA 81, pp.3088-3092, 1982.

[kaufman & Boulaye 1978], *Théorie Des Treillis en vue des applications*. Paris, Masson, 1978.

[Lelu, Halleb & Delprat 1998], *Recherche d'information et Cartographie dans des Corpus Textuels à Partir des Fréquences de n-Grammes*, Actes de JADT-98, Nice, France.

[Manning & Schütze 1999], *Foundations of Statistical Natural Language Processing*, MIT Press.

[McCulloch 1943], *A logical calculus of the ideas immanent in nervous activity*, Bull Math. Biophysics, 5;113-115, 1943.

[Norris 1978], *An Algorithm for Computing thé Maximal Rectangles in a Binary Relation*, Revue Roumaine de Mathématiques Pures et Appliquées, 23, n° 2, 1978, p. 243-250.

[Rumelhurt 1986], *Learning internal representations by error propagation.* Parallel Distributed Processing Explorations in the Microstructure of Cognition. Eds. D. E.

[Rumelhart & McClelland 1986], Cambridge, MA, MIT Press, *Bradford Books*, vol. 1, pp. 318-362, 1986.

[Turenne 2000], *Apprentissage statistique pour l'extraction de concepts à partir de textes (Application au filtrage d'informations textuelles),* thèse de doctorat en informatique, Université Louis-Pasteur, Strasbourg, France.

[Will 1982], *Restructuring the lattice theory : An approach based on hierarchies of concepts.* In I.Rival,editor,Ordored Set, Dordrecht-Boston, Reidel,1982,pages 445-470.