Les Mathématiques et l'information

YALAOUI Bilal
CEntre de Recherche sur l'Information Scientifique et Technique
yalaoui@mail.cerist.dz
AITHADDADENE Hocène
Université des Science et de Technologie Houari BoumédieneUSTHB
Aithaddadene@yahoo.fr

Introduction

Dès les premiers âges l'échanges de renseignements fut indispensable à la survie des petits groupes humains : l'usage des messagers et des espions fut, sans doute, une des formes premières de la civilisation et l'emploi de signaux codés visuels ou sonores ancêtres de nos télécommunications.

Donc à l'origine le besoin pressant de dire et d'avoir dit son opinion. La nécessité d'un terme général évoquant à la fois ce besoin et les moyens propres à le satisfaire est doc apparue.
Parallèlement, les progrès techniques ont fait de la publication des pensées et des connaissances une activité sociale organisée et éminente. C'est ainsi que le développement de la presse imprimée à partir du milieu du XIXe siècle, a fait d'elle le véhicule principal des informations et opinions.

Au XXe siècle, les découvertes scientifiques et l'évolution technologique donnèrent à l'image puis à la parole et à leurs combinaisons une amplification supérieure à celle données à l'écriture par la machine à imprimer. On parla de presse ou de journal parlé et de filmée. L'inadaptation du mot presse a conduit après 1945 à cherche d'autre termes ; dès 1950 le mot communication à été emprunté aux américains, en traduisant leurs formules de « mass communication » et « media of mass communication » par « communication de masse » et « moyens de communication de masse » ; depuis le mot « média » s'est lentement imposé pour désigner les moyens d'information.
Le terme information a connu une singulière fortune. Du langage courant, où il évoque à la fois l'acte de recueillir et celui de donner des renseignements, en passant par le langage judiciaire, où il désigne la procédure de recherche et de consultation d'une infraction. Il est le plus apparemment précis qui soit au langage scientifique puisque il a

servi à qualifier l'une des théories de cybernétique[1] : « traitement de l'information », puis a fourni le dérivé qui la désigne « l'informatique ».

Les divers emplois du terme « Information » ont entraînés quelques confusions et aussi quelques oublis de son sens original. Celui-ci exprime essentiellement l'idée de mise en forme – la mise en forme étant faite en vue d'une mise au courant (Fernand Terrou[2], 1962).

Son but est l'appréhension de sens ou d'être dans leur signification. Que ce soit pour le simple plaisir de connaître, d'être informé sur les événements politiques, sur les progrès de la science et de la technologie ou pour celui, moins simple, d'être au fait des derniers objets et résultats de la recherche, de suivre le front de la connaissance scientifique. Elle est transmissible sous différents types et suivant plusieurs critères à savoir :

- son contenu (économique, social …),
- son public (public, scientifique, pédagogique …),
- sa nature (sonore, textuel …),
- sa diffusion (interne, publiée …).

Le mot « inform. » signifie originellement « donner forme à » (Dahmane Madjid, 1990) []. L'information devra donc modifier la forme de la personne qui reçoit l'information. Du moins elle va bousculer les choses dans sa façon de voir les choses sur un sujet particulier. C'est le récepteur qui décidera si le message reçu constitue véritablement une information.

Alors que la communication est le processus intermédiaire qui permet l'échange d'information entre individus (Yves-François Le Coadic, 1994). Comme définit par Robert Escarpit[3] (1990) « La communication est un acte, un processus, une machinerie. L'information est un produit, une substance, une matière ».

[1] La cybernétique est une modélisation de l'échange, par l'étude de l'information et des principes d'interaction. Elle est issue en particulier du passage par la théorie entre l'étude du système nerveux et sa reproduction en intelligence artificielle. Le mot cybernétique formalisé en 1948 par Norbert Wiener est le résultat de tout un mouvement scientifique très largement interdisciplinaire et source d'une nouvelle école de pensée. Techniquement, c'est une méthode interdisciplinaire qui étudie l'évolution dynamique des systèmes

[2] *Fernand Terrou* (1951-1976), participait aux différents travaux dont la législation de la presse. Après avoir animé plusieurs groupes de recherche sur l'information dès 1947, il est nommé directeur de l'Institut Français de Presse, fondé en 1951 dans le cadre de l'Université de Paris. En 1965, il organise la préparation d'un doctorat de 3ème cycle consacré, pour la première fois en France, aux différents aspects de l'information. La création du D.E.A. en Sciences de l'information, en 1975, consacre l'effort entrepris par *Fernand Terrou*, près de 30 ans auparavant.

[3] *Robert Escarpit (*1918 - 2000), un universitaire, écrivain et journaliste français.

Notons aussi que l'information n'a aucune valeur tant qu'elle n'est pas utilisée. Cette valeur réside dans son utilité opérationnelle[4] en apportant la bonne information à la bonne personne au bon moment. Son usage est nombreux et est exprimée par les éléments suivants :

- Instrument de diffusion de connaissances,
- Stimule la réflexion et l'action, grâce à l'introduction des idées, de l'expérience et de la réalisation des autres personnes (avec toutes les interactions réciproques),
- Joue un rôle dans la détermination des rapports socioculturels entre les nations.

L'impact de l'usage de l'information se manifeste ainsi par la généralisation de recours de plus en plus à:

- L'utilisation des procédures informatisées,
- L'extension de réseaux informatisés donnant accès à un volume croissant d'informations diversifiées,
- L'usage de nouveaux supports et canaux de communication,
- Des possibilités d'interaction plus souples et plus étendues entre les utilisateurs et les systèmes.

En parallèle, le coût de traitement et de la diffusion des informations continue à décroître, permettant ainsi de faire face à l'augmentation exponentielle du volume d'informations et à la nécessité de banaliser l'accès à celle-ci.

Il faut dire à la fin que l'avènement de l'électronique, qui s'est traduit par le passage de supports matériels à des supports immatériels, puis de l'informatique et le développement de communication d'informations à distance n'a fait que renforcer la démultiplication, l'amplification et mémorisation de masses d'informations qui se poursuivent sans fin. A l'air des (nouvelles) technologies d'information et de la communication, nulle distance de fait plus obstacle à la vitesse, nulle frontière n'arrête plus l'information. Les ordinateurs travaillent en milliardième de seconde. Les satellites de télécommunications permettent d'atteindre en quelques secondes, par voie entièrement automatique, toutes les régions du monde.

[4] Dans le sens de l'utilisation objective et rationnelle de l'information

2- A l'origine : La théorie statistique de l'information de Claude Elwood Shannon ;

En 1948, C. Shannon et W. Wearver, et avant Hartley en 1928, vont avancer l'idée de l'entropie informationnelle. Les travaux de C. Shannon sont nés de l'étude au sein de la Compagnie Bell de problèmes particuliers aux télécommunications. Ils aboutissent à une théorie mathématique de communication dite théorie de l'information qui est essentiellement mécaniste (i.e. ne tenant pas compte de la signification des messages transmis). Le problème à résoudre était purement technique : quel codage optimal peut-on appliquer à des messages choisis dans un ensemble connu afin de les transmettre le plus fidèlement et le plus rapidement possible en présence de parasites ? Considérant que les messages sont transmis d'une source (émetteur) à un destinataire (récepteur) à travers une voie de communication, le message est codé afin de parcourir la voie puis décodé afin d'être restitué au destinataire. Shannon définit alors la quantité d'information contenue dans un message comme une fonction de la fréquence d'utilisation des différents symboles composant le message.

Cette théorie se révéla d'emblée très féconde dans son domaine d'origine. Mais très vite elle intéressa de nombreux chercheurs de toutes disciplines, comme en témoigne la tenue dès l'été 1950 à Londres du premier symposium international sur la « théorie de l'information » . Notons à la fin que la formule proposée est analogue à celle de l'entropie thermodynamique de Gibbs Boltzmann et qu'elle définit la quantité d'information sans définir la notion d'information.

3- Les mathématiques comme indicateurs métrique autour de l'information

Le suffixe « métrie » renvoie aussi bien à la mesure (l'évaluation d'une grandeur par comparaison à une unité) qu'à la métrique (la création d'une convention qui permette de définir une distance entre l'ensemble des éléments étudiés). Les travaux de recherche réalisés dans ce contexte sont fondés sur deux postulats :

- Un écrit scientifique est le produit objectif de l'activité d'une pensée. Dans un contexte scientifique, une publication est une représentation de l'activité de recherche de son auteur. Le plus grand effort de cet auteur est de persuader les autres scientifiques que ses découvertes, ses méthodes et techniques sont particulièrement pertinentes. Le mode de communication écrit fournira donc tous les éléments techniques, conceptuels, sociaux et économiques que l'auteur cherche à affirmer tout au long de son argumentation.

- L'activité de publication scientifique est une perpétuelle confrontation entre les propres réflexions de l'auteur et les connaissances qu'il a acquises par la lecture des travaux émanant d'autres auteurs. La publication devient par conséquent le fruit d'une communion de pensées individuelles et de pensées collectives. Ainsi, les chercheurs, pour consolider leur argumentation, font souvent référence à des travaux d'autres chercheurs qui font l'objet d'un certain consensus dans la communauté scientifique. Par conséquent, il existe une relation entre tous les travaux scientifiques publiés, que cette relation soit directe ou indirecte, reconnue ou dissimulée, consciente ou inconsciente, en accord ou en désaccord.

3.1- Bibliométrie, Scientométrie, Infométrie et Webométrie

Les définitions données à l'époque étaient très restrictives car elles ne prennent pas en compte l'étude de la circulation des documents mais seulement l'analyse quantitative des caractéristiques bibliographiques. C'est pour distinguer ces deux types d'application qu'un autre terme est apparu, celui de Scientométrie. Il est originaire d'un terme Russe signifiant l'application de méthodes quantitatives pour l'histoire des sciences (Dobrov & Korennoi,1969) . Brookes lors de la première conférence internationale de Bibliométrie, d'Infométrie et de Scientométrie en 1987, l'a clairement précisé :

« Alors que la Bibliométrie aurait pour objet d'étudier les livres ou les revues scientifiques et pour objectif de comprendre les activités de communication de l'information, la Scientométrie aurait pour objet l'étude des aspects quantitatifs de création, diffusion et utilisation de l'information scientifique et technique de et pour objectif la compréhension des mécanismes de la recherche comme activité sociale » .

Ainsi, la Bibliométrie regroupe l'ensemble des méthodes aidant à la gestion de bibliothèques et la Scientométrie recherche les lois qui régissent la science, d'où son appellation « Science de la science » par De Solla Price.

Un troisième terme l'Infométrie serait le terme générique embrassant les deux disciplines. Sa définition est plus large : c'est l'application des modèles et des méthodes mathématiques et statistiques de façon à dégager des lois relatives à l'information scientifique et technique (Rostaing Hervé, 1996).

Le terme Webométrie est une variante de l'Infométrie qui prend comme objet d'étude l'information sur le Web relative au support, au contenu et sa structure.

3.2- Un bref historique

La première étude Bibliométrique à été faite en 1917 par Cole et Eales , ils avaient réalisé une analyse statistique de la littérature de l'anatomie publiée entre 1850 et 1860 pour montrer les fluctuations d'intérêt pendant cette période.

En 1926, A. J. Lotka à constater qu'il existe une relation inverse entre le nombre de publications dans un domaine scientifique et le nombre de ses membres. Cette régularité est représentée par une fonction hyperbolique établie par Lotka, et a connue par la suite de nombreuses études (par exemple celle de R. Rousseau en 1992 et R. Wagner-Döbler & J. Berg en 1995). La loi de Lotka est une loi de production au sens économique du terme.

Une année après (1927), Gross P. L. K et Cross E. M comptabilisèrent les citations présentes dans les bibliographies d'articles de journaux en chimie, puis rangèrent les journaux dan l'ordre du nombre de citations reçues. Cette étude fut la première analyse des citations.

Ralph Hartley en 1928 propose une généralisation de la quantité d'information pour un ensemble ayant un nombre quelconque d'éléments.

En 1930, S.C. Bradford, un bibliothécaire travaillant dans un centre de documentation, s'est intéressés à la répartition des articles scientifiques pour un domaine précis dans les périodiques. En 1934 il a déterminé à l'aide d'un modèle mathématique simple une méthode qui va quantifier le désordre dans la documentation . Il a montré que les articles scientifiques sont distribués avec une régularité remarquable dans les revues. Cette théorie qui aura été à l'origine d'un grand nombre d'articles en bibliométrie n'a pas fait beaucoup fait d'émules à cette époque. De nombreuses formulations mathématiques vont essayer par la suite de modéliser ces travaux de Bradford notamment ceux de Burrell en 1988 .

En 1948, C. Shannon et W. Wearver, vont avancer l'idée de l'entropie informationnelle. Leurs travaux s'inscrivent dans le cadre de la théorie des probabilités qui repose sur la théorie des ensembles et de la mesure en mathématiques.

En 1949, G.K. Zipft constate en étudiant des corpus de données textuelles des régularités sur la fréquence d'apparition des mots. Très grossièrement, nous pouvons dire que si nous rangeons les termes suivant leurs fréquences décroissantes, nous nous apercevrons qu'il existe une relation entre le rang et la fréquence : le produit rang fréquence est à peu près constant. Cette loi est nommée aussi sous le terme de loi du moindre effort.

Au début des années 60, E. Garfield créa à Philadelphie (Etats Unis) l'ISI (Istitute for Scientific Information) qui permet aux travaux de Bibliométrie de prendre un nouvel essor. Une nouvelle école de pensées fondée sur l'étude des citations, se greffe autour de cet institut.
PH Morse utilise en 1968 les méthodes de prévision issues de la recherche opérationnelle pour gérer le flux dans les bibliothèques.

Au début des années 70, toujours au tour de l'ISI, la bibliométrie étend son domaine d'application aux technologies grâce à la fondation CHI (Computer Horizons Inc).

De Solla Price en 1976, va construire un modèle probabiliste, dit lois des avantages cumulés , qui va expliquer différents phénomènes observés dans des circonstances de production bibliométriques différentes. Ce principe peut s'énoncer ainsi : plus une source (journal, chercheur, texte …etc.) produit des éléments plus grande est sa chance d'en produire. De Solla Price montre que les lois empiriques précédentes ne sont que des limites de son modèle.

Les propriétés mathématiques de ces distributions statistiques vont être étudiées par S. D. Haitun en 1982 et vont recevoir le nom de « Zipfiennes » qu'on opposera au « Gaussiennes ».

Lors de la première conférence internationale de Bibliométrie, d'Infométrie et de Scientométrie en 1987, l'organisateur L. Egghe énumère les champs d'intérêts suivant : les statistiques, la recherche opérationnelle, les lois bibliométriques, l'analyse des citations, la théorie de la circulation, la théorie de l'information, et en fin les aspects théoriques en recherche d'information documentaire. Tous ces thèmes ont comme point commun d'être abordés de manière quantitative en utilisant l'outil mathématique et statistique.

De nombreux travaux théoriques et applicatifs ont montré par la suite l'intérêt de ces outils et techniques pour la veille, l'évaluation de la recherche, la recherche et la structuration de l'information …etc. plus généralement l'analyse et l'exploration métriques de l'information.

4- Les mathématiques comme langage pour l'information

De structures hiérarchiques au réseau sémantiques, en passant par les formalismes logiques, les mathématiques ont été très utilisées comme langage pour communiquer de l'information. Dans ce contexte nous parlons de la représentation des connaissances à

l'aide des formalismes mathématiques, qui sont considérés comme modèles pour la représentation des connaissances et d'informations.

La conception d'un système d'information à base de connaissance conduit nécessairement au développement d'un modèle de représentation de ces connaissances pour qu'elles aient la capacité d'être facilement interprétées par la machine ainsi que les algorithmes employant ce modèle. Il est à savoir que, l'approche employée en intelligence artificielle est de formuler le modèle conceptuel en termes de connaissance, pour décrire (par des déclarations) les objets du domaine et leurs relations. La manière dont le modèle conceptuel est conçu est connue sous le nom de « représentation de la connaissance » et c'est la base sur laquelle des applications intelligentes se développent. Cette discipline est un domaine de recherche de l'Intelligence Artificielle, qui est elle-même un domaine de l'informatique dont le but est de faire accomplir par l'ordinateur des tâches effectuées par l'être humain et qui demandent de l'intelligence.

Pratiquement, une représentation de connaissances est un système définissant une série de symboles et une série d'opérations sur ces symboles. Pour décrire formellement quelque chose, on a besoin d'un médiateur : un langage.

Un langage est défini par deux aspects fondamentaux :
(1) Une syntaxe ; Règles selon les quelles les éléments d'un langage sont assemblés.
(2) Une sémantique ; Règles selon les quelles les expressions syntaxique se voient assignées un sens.

4.1- Connaissances intentionnelles et extensionnelles

Luger et Subblefield affirment que la distinction entre connaissances intentionnelles et extensionnelles est la clé pour la représentation des connaissances. Le terme extension est utilisé pour désigner l'ensemble de tous les choses/objets exprimés par le concept (exemple « une boule» à son extension vers l'ensemble de toutes les formes de boule), alors que l'intention d'un terme détermine l'abstraction du sens désigner (pour « une boule» c'est sa forme ronde et son habilité à rouler) bien que ces deux notions caractérisent le même concept, leur rôle est différent. Décrire les connaissances en extension est relativement une tâche facile, mais décrire les connaissances intentionnelles est une tâche difficile.

A un autre niveau, la capacité d'avoir des connaissances sur les choses qu'ont connaît, souvent référée comme « méta-connaissance », est également souhaitable pour une représentation de connaissances. La méta-connaissance est importante si nous souhaitons raisonner au sujet de la connaissance elle-même.

Durant les trente dernières années une variété de formalismes de représentation de connaissances a été développée. Elles peuvent être classées (selon l'approche utilisée) en trois principales approches :

- Approche logique ; dont le principe est d'utiliser la logique mathématique comme outil.
- Approche sémantique ; utilisée à l'origine par les linguistes pour représenter la sémantique des phrases.
- Approche Hybride ; entre la sémantique et la logique avec par exemple la notion de Schéma (Frame) mise en évidence par les chercheurs en psychologie.

4.2- L'approche logique

Le formalisme logique a été l'un des premiers formalismes proposés pour représenter de la connaissance, et constitue toujours la base de nombreuses recherches en Intelligence Artificielle. L'approche logique est fondée sur La logique symbolique traditionnelle qui elle même se subdivise en deux branches : Calcul de propositions et Calculs des prédicats.

Le calcul des propositions traite les rapports (appelés les propositions) comme « Lila est la mère de Adel », composées de symboles atomiques simples (typiquement P, Q, …etc.). Il n'y a aucune manière d'accéder aux différents composants de l'affirmation (i.e. de la proposition) ni d'exprimer les relations entre eux.

Cependant, le calcul des prédicats fournit des capacités d'analyse des rapports dans des composants plus fins en les représentants au moyen d'attributs. Un prédicat est un symbole indiquant la relation spécifique entre les entités. Pour continuer notre exemple, la proposition pour décrire le rapport mère – enfant entre Lila et Adel serait exprimée par : Mère(Lila,Adel)

Le langage de la logique des prédicats du premier ordre est plus riche que celui des propositions ; il inclut des symboles de fonctions. Son principe consiste à utiliser les symboles de prédicats et les symboles de fonctions de parité quelconque pour représenter la connaissance.

Dans la logique des prédicats, on interprète les formules syntaxiquement correctes du langage sur l'ensemble des valeurs de vérité : « vrai » et « faux », et on exploite des règles d'inférence valides, c'est-à-dire inférant des formules qui prennent la valeur vraie dès lors que les formules parentes prennent aussi la valeur vraie.

Les techniques les plus répandues concernent la logique des prédicats, dite du premier ordre, sont fondées sur l'emploi d'une seule règle d'inférence : principe de résolution de Robinson (1965), notamment utilisé dans PROLOG . Le vocable premier ordre signifie que les prédicats contiennent des variables, paramètres quantifiés universellement, qui peuvent être remplacés par n'importe quelle expression bien formée du langage. L'emploi de logiques d'ordre supérieur permet par exemple de faire porter les quantificateurs sur des prédicats ou des fonctions.

Exemples :

Les livres sont des objets :

$(\forall x)$ (livre(x) $\supset$ objet(x))

Les livres sont écrits par des auteurs :

$(\forall x)$ (livre(x) $\supset$ (($\exists y$) (auteur(y) $\wedge$ a_écrit(y,x)))

Celui qui écrit un livre est un auteur :

$(\forall x)$ $(\forall y)$ (a_écrit(y,x) $\wedge$ livre(x) $\supset$ auteur(y))

Un livre est édité par une entreprise d'édition :

$(\forall x)$ (livre(x) $\supset$ (($\exists y$) (éditeur(y) $\wedge$ a_édité(x,y)))

$(\forall x)$ (éditeur(x) $\supset$ entreprise(x))

$(\forall x)$ (livre(x) $\supset$ ((entreprise(y) $\wedge$ (($\exists z$) (a_pour_activité(y,z) $\wedge$ édition(z))))

La logique dite classique ne reconnaît comme modalité que le vraie et le faux, d'autres théories de logique ont été proposées telles que :

- La logique floue (Lotfi.A Zadeh) dans laquelle la véracité d'une proposition est un nombre réel dans l'intervalle [0,1].
- Les logiques modales (Lewis 1918) introduisent des modalités telles que la possibilité ou la nécessité, mais aussi des modalités temporelles telles que le passé ou le futur.
- Certaines formules peuvent être considérées comme prouvables dans un ensemble d'axiomes A1, mais non prouvables dans un ensemble A2 contenant A1, c'est la non monotonie. Les logiques non monotones sont nées à partir des années 1970.
- Par ailleurs, des travaux (Girard 1987) ont permit de définir une logique linéaire dans laquelle les propositions jouent le rôle de ressources consommables.

Outre que théories logiques nous trouvons aussi les systèmes à base de règles de production. Où les règles de production sont des connaissances déclaratives de la forme : « SI Condition ALORS Conclusion ». Condition et Conclusion sont des expressions logiques telles que Condition doit être vérifiée pour que la règle s'applique.

Exemple :

SI (le moteur cale) et (l'allumage correct) et (réservoir d'essence non vide)
ALORS (vérifier carburation)

Le noyau d'un système à base de connaissances utilisant des règles de production comprend trois parties principales : 1-Une base de règles (constituant la connaissance permanente), 2-Un moteur d'inférences (contient le mécanisme de raisonnement exploitant les règles), 3-Une base de faits (une mémoire de travail du système qui contient les données initiales décrivant le problème, les hypothèses émises et les faits apparus avant de parvenir à la résolution du problème).

Le moteur d'inférence fonctionne selon un cycle, il itère sur ce cycle jusqu'à l'obtention d'une (ou de toutes les) solution(s) ou éventuellement un échec (i.e. pas de réponse). La résolution se fait par exploration d'un graphe (dit espace de recherche) ou les sommets sont les états du problème et les arcs les règles applicables. Le principe consiste à choisir parmi les stratégies d'exploration de graphes celle qui aboutira à la résolution (i.e. une réponse).

4.3- L'approche sémantique

L'utilisation des graphes en représentation des connaissances pour l'Intelligence Artificiel vient de l'idée de représenter graphiquement des concepts et leurs liens. Les réseaux sémantiques ont pour origine des expériences de psychologie qui ont montré que l'homme semble mémoriser les informations selon un principe d'économie avec un modèle de mémoire associative représentant des relations entre concepts.

Un réseau sémantique est un graphe dont les nœuds sont des concepts ou des objets modélisés, et les arcs sont des relations sémantiques (dites aussi liens) Les concepts sont alors définis par leur position dans le réseau d'où l'appellation, i.e. nous savons ce qu'est un concept en le positionnons par rapport aux autres.

Les relations les plus courants entre les nœuds sont les suivantes : « est-un », « partie-de », « type-de » (sous classe, généralisation/spécialisation), « instance-de » (appartenance à une classe), « est-une-partie-de », « propriété », « conséquence ».

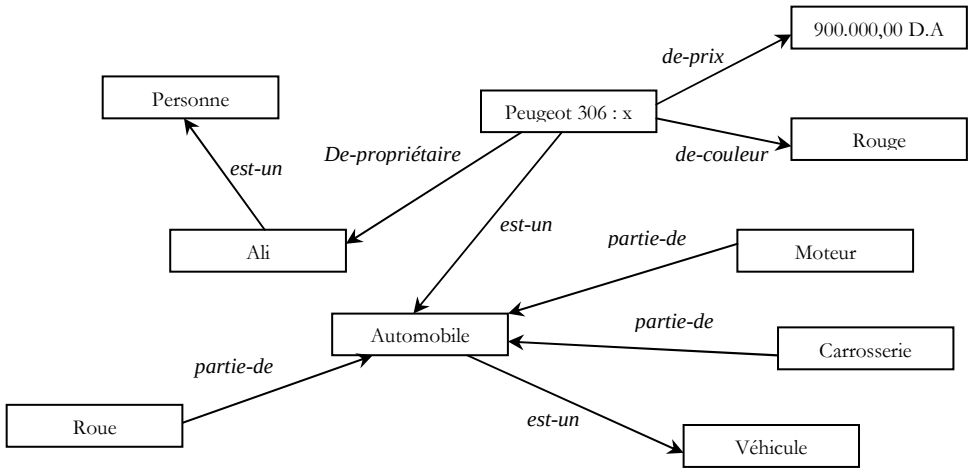

Figure 1 : Exemple de réseau sémantique

Les réseaux sémantiques ont été employés pour la représentation de la connaissance au commencement de l'intelligence artificielle. En effet, les premiers travaux dans ce domaine ont été faits par Charles Sanders Peirce (1839-1914), un logicien, mathématicien, et le premier psychologue expérimental moderne en Amériques . Il a développé un système graphique de la logique appelé graphe existentiel, il a utilisé ces graphes pour enregistrer systématiquement des observations du monde autour de lui. Les réseaux sémantiques contemporains ont une grande ressemblance avec ces graphes qui ont été l'inspiration pour beaucoup de chercheurs dans le domaine de l'intelligence artificielle et philosophie.

Par la suite, ils furent développés par Quillian en 1968 dans l'analyse du langage naturel pour mettre en évidence les relations existantes entre la signification de termes connectés entre eux. Les réseaux sémantiques furent redéfinis méthodologiquement par Woods (1975) pour éliminer les ambiguïtés liées à la nécessité de mieux définir les liens et les propriétés qu'on leur assigne. Grâce aux travaux de Woods, il fut possible de les rendre équivalents aux propriétés de la logique, en héritant toutefois les désavantages caractéristiques de ce domaine (en particulier leur maniement limité)

Ces liens sont graphiquement suggérés différemment pour que leur nature soit explicite et donc ils se comportent de façons différentes. Cette spécificité de comportement doit être formalisée chaque fois qu'on introduit un nouveau lien. Par exemple, le lien est-un hérite au nœud d'origine toutes les caractéristiques du nœud de destination comme subordonné hiérarchique. Au contraire, le lien partie-de n'a pas cette caractéristique d'héritage.

Par la suite, les réseaux sémantiques ont été l'objet d'un grand intérêt motivé par la recherche des méthodes d'organisation et de consultation de bases de connaissances volumineuses et complexes. Les recherches en programmation orientée objet et en bases de données orientées objet a également orienté l'étude sur les aspects centrés objet des réseaux sémantiques, particulièrement sur la hiérarchie et l'héritage.

Notons à la fin que le terme « réseau sémantique » entoure une famille entière des représentations visuelles à base de graphe. Toutes ces représentations partagent l'idée fondamentale de représenter la connaissance de domaine sous forme de graphe, mais il y a quelques différences au sujet de notation, d'appellation des règles et les inférences applicables. Le terme « réseau sémantique » est également souvent employé comme synonyme pour les graphes conceptuels qui vont être présentés ci-dessous.

4.4- Les Graphes Conceptuels (GC)

Le modèle des Graphes Conceptuels est un formalisme de représentation des connaissances fondé sur la définition de concepts et de relations entre concepts, il a été introduit par John F. Sowa en 1984. Un graphe conceptuel est un graphe biparti avec deux types de nœuds : concepts et relations. Les nœuds sont liés par des arcs orientés, et un arc relie toujours un nœud concept et un nœud relation, un nœud peut être isolé (tout seul).

Un GC représente une formule logique les noms et les arguments sont représentés par des nœuds. Les arcs du graphe relient les noms des prédicats à leurs arguments. On utilise des rectangles pour les concepts (arguments) et des cercles pour les relations (prédicats), la représentation peut avoir une signification littérale d'une phrase par exemple : « Ali va à Alger avec une voiture » peut être représenté sous la forme graphique ci-dessous :

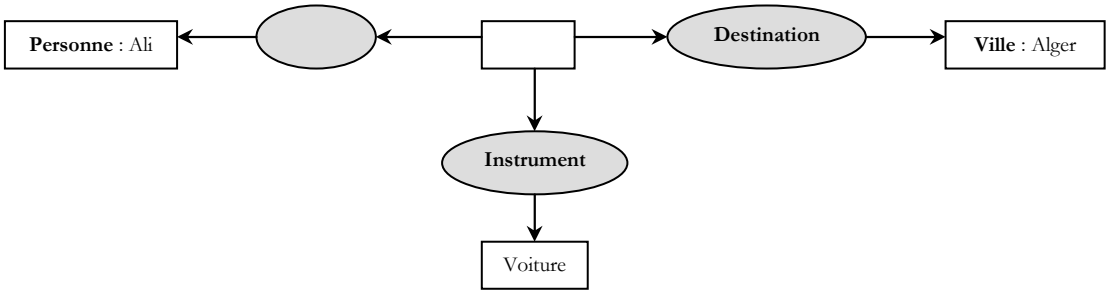

Figure 2 : Exemple de représentation graphique

Cette représentation est dite représentation graphique d'un GC (Display Form), d'autres représentations sont possibles :

-Représentation sous forme linéaire (Linear Form) ; pour le même exemple ont obtient :

[Aller]-
(Agent) ->[Personne : Ali]
(Destination) -> [Ville : Alger]
(Instrument) ->[Bus]

-Il existe d'autres possibilités avec les formats d'échange de graphes conceptuels entre machines tel que CGIF (Conceptual Graph Interchange Format) :

*CGIF : [Aller *x](Agent ?x [Personne : Ali]) (Destination ?x [Ville : Alger])*
(Instrument ?x [Voiture])

Un GC peut être traduit par une fonction notée φ (Sowa, 1984) en une formule bien formée de la logique des prédicats du 1er ordre. Par exemple pour celui ci-dessus:

$$(\exists x)\,(\exists y)\,(\,Personne(Ali) \wedge Aller(x) \wedge Ville(Alger) \wedge Voiture(y) \wedge Agent(x,Alger) \wedge Destination(x,Nice) \wedge Instrument(x,y)\,)$$

La hiérarchie des types peut se traduire avec des implications logiques : pour tous les couples (t, t') de types de concepts, si $t < t'$, alors $\forall x\ t'(x) \Rightarrow t(x)$. De même, si G' est une spécialisation de G, on a : $\varphi(G') \Rightarrow \varphi(G)$.

Deux niveaux de représentation des connaissances sont distingués dans le modèle :
1-Le niveau terminologique ; qui comprend la définition du vocabulaire conceptuel du domaine visé (le support) et des connaissances générales sur ce domaine en terme de définitions, contraintes ou règles.

2-Le niveau assertionnel ; qui comprend des graphes étiquetés dédiés à la représentation de faits. Ces graphes sont construits à partir du vocabulaire conceptuel du niveau terminologique. Le vocabulaire conceptuel du modèle des GC est composé de :

- Un ensemble ordonné de types de concept (un treillis : Tc)
- Un ensemble ordonné de types de relation (une hiérarchie/arborescence : TR)
- Un ensemble de marqueurs (ou référents : M)

Les deux hiérarchies de types sont structurées par des relations Sorte-de qui permettent d'exprimer qu'un concept (ou qu'une relation) est plus spécifique qu'un (qu'une) autre. Les faits sont représentés par des graphes conceptuels simples, c'est-à-dire des graphes finis (non nécessairement connexes). Les sommets concepts représentent des instances de concepts et les sommets relations indiquent la nature des liens qui lient ces instances entre elles.

Chaque sommet concept est étiqueté par un type de concept et un référent qui peut être un marqueur individuel (dénotant alors une instance précise du type spécifié) ou le marqueur générique, noté *, qui dénote une instance quelconque du type spécifié.

Chaque sommet relation est étiqueté par un type de relation. Les arêtes sont étiquetées par des entiers qui indiquent l'ordre des arguments des sommets relations (pour les relations binaires, un arc entrant vers un sommet relation indique le concept premier argument et un arc sortant le deuxième). Les étiquettes des sommets concepts voisins d'un sommet relation doivent respecter la signature de l'étiquette du sommet relation.

Remarque : Les prédicats binaires jouent un rôle particulier en représentation de connaissance, un prédicat m-aires peut être facilement transformé en une série de prédicats binaires. Par exemple :

« envoie(Mourad,Lila,Livre) » peut être remplacée par :

« expéditeur(envoi,Mourad) ∧ destinataire(envoi,Lila) ∧ objet(envoi,livre) »

John Sowa (1984), distingue Entre les graphes conceptuels et des réseaux sémantiques ; pour lui chaque graphe conceptuel affirme une seule proposition simple, alors que les réseaux sémantiques sont beaucoup plus grands. John Sowa suggère que les réseaux sémantiques soient des entités qui incluent des graphes conceptuels. Il précise alors que dans un réseau sémantique les concepts et les relations d'un graphe conceptuel sont liés à un contexte, une langue, une émotion et à la perception. Des concepts peuvent être associés aux percepts pour expérimenter le monde et les mécanismes moteurs pour agir sur elle; ils peuvent être associés aux mots et aux règles grammaticales d'une langue.

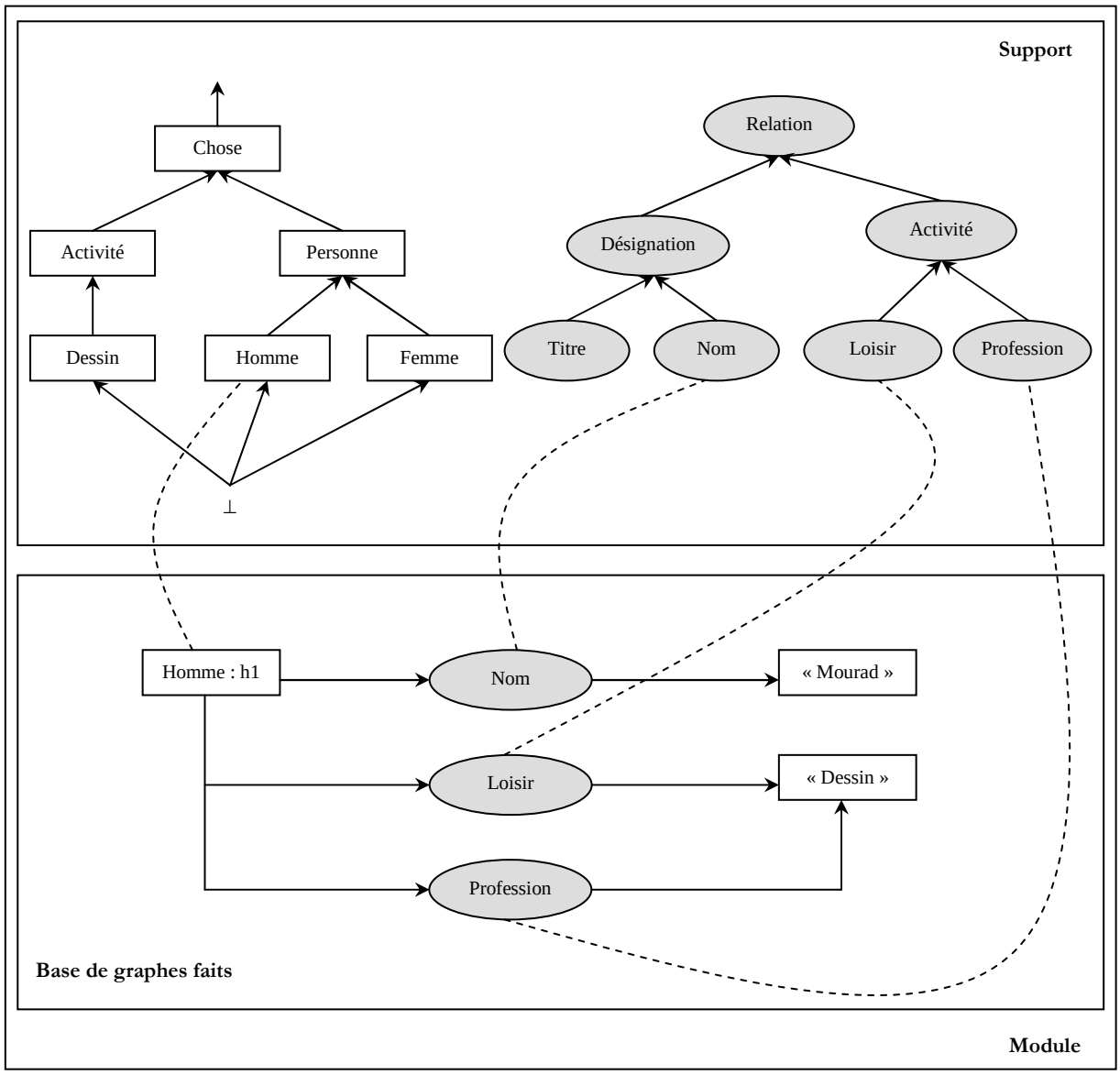

Figure 4 : Exemple de Module de GC

5- Les mathématiques comme outil pour l'exploration et la découverte du contenu de l'information

Face au rythme de croissance effréné du volume des informations disponibles et accessibles. Il est devenu impossible de lire et d'assimiler efficacement (en termes de temps et coût) le contenu par des méthodes classiques de lecture. Il est ainsi important d'assister l'utilisateur, par des outils, afin qu'il puisse passer moins de temps à chercher l'information et d'avantage à en exploiter le contenu essentiel.

C’est dans ce sens que les techniques et méthodes mathématiques ont été utilisées dans des systèmes informatiques. Nous trouvons l’illustration dans le domaine de data mining et les domaines dérivés tels que le text-mining (qui sera présenté dans la section suivante) ou le web mining.

Le data mining est un sujet brûlant. Il dépasse aujourd’hui le cercle restreint de la communauté scientifique pour susciter un vif intérêt dans le monde des affaires. Il peut être défini comme étant l’exploration et l’analyse, par des moyens automatiques ou semi-automatiques, d’un large volume de données afin de découvrir des tendances ou des règles (Michael J. A. Berry).

Le schéma ci-dessous nous donne un résumé sur les principaux techniques utilisées, une présentation détaillée peut être consultéE dans les ouvrages de : R. Lefébure & G. Venturi , M. Berry & G. Linoff ou J. Han & M. Kamber .

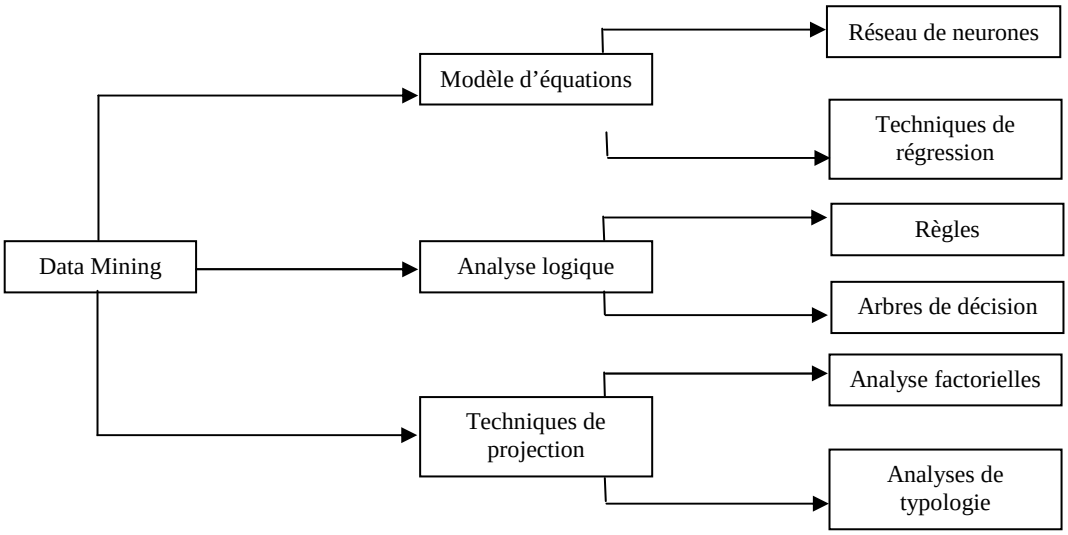

Figure 5 : Technique d’analyse de données en Data Mining

La découverte de connaissances à partir d’importantes masses de données réparties et hétérogène débouche le plus souvent sur l’analyse relationnelle. La recherche d’informations stratégiques s’appuie en effet sur les liens fonctionnels et sémantiques entre documents, acteurs, terminologie et concepts d’un domaine sans oublier le paramètre temps.
C’est dans ce contexte qu’un ensemble de visualisations interactives de graphes sont proposées, et dont la manipulation permet une découverte de connaissances intuitive et basé sur un langage graphique naturelle.

Le concept graphe est généralement utilisé comme modèle de représentation dès que les données sont intrinsèquement liées. Ce type de données peut être vu comme un graphe dont les arêtes représentent les relations entre ces données.

Par ailleurs, la représentation des données par des graphes est largement utilisée dans différents domaines qui traitent de l'information textuelle. Par exemple en Sociologie pour l'analyse des réseaux sociaux, en text-mining pour l'analyse des mots-associés, en Scientométrie et Bibliométrie pour l'analyse des réseaux de co-citations ou en Web mining pour l'analyse de la distribution d'hyperliens. L'analyse de ces réseaux a pour objectif de concevoir des représentations synthétiques qui puissent exprimer l'interaction entre les différentes entités représenté.

6- Conclusion

Le terme « Information » est à considérer dans un double sens ; il exprime les deux éléments constitutifs du phénomène « la mise en forme » et « la divulgation » (Fernand Terrou, 1962). L'information désigne ainsi toute publication sous une forme appropriée d'éléments de connaissance ou de jugements à l'aide de mots, de sons ou d'images ou de tous signes accessibles au public visé.

Nous pouvons donc décrire une information comme une association significative et subjective d'un ensemble ou d'une collection de données organisées, représentée par des signes et symboles qui sont des éléments du langage (signe alphabétique, mot, signe de ponctuation), inscris sur un support et visant à transmettre un message d'un émetteur à un récepteur. Elle est supposée changer la façon dont le récepteur perçoit quelque chose, avoir un impact sur son jugement et son comportement. Le but est de « l'informer », il s'agit de données qui font une différence.

L'objet de la science de l'information (i.e. l'information) est une matière qui envahit l'espace professionnel. C'est une ressource vitale. Son contenu, marqué par le sceau de l'interdisciplinarité, est un savant dosage de sciences mathématiques et physiques et de sciences sociales et humaines.

L'objet de cet article été de montré d'une part, l'usage et la contribution des sciences mathématiques en science de l'information…..

Références :

1 A. J. LOTKA, « The frequency distribution of scientific productivity », Journal of the Washington Academy of Sciences, 16, 1926, p. 317-323.

2 Brooks B C, « Biblio-, sciento-, info-métrics ??? What are we talking about? », Informetrics 87/88, Proceedings of the diepenbeek conference, Amsterdam: Elsevier, 1988.

3 C. SHANNON, W. WEAVER, « Théorie mathématique de la communication », Bibliothèque du CEPL, 1975, 188 pages.

4 Cole F J., Eales N B, « The history of comparative anatomy. Part I: a statistical analysis of the literature », Science progress, London, April 1917, Vol 11, p.578-596

5 D.S. PRICE, « A general theory of bibliometric and other cumulative advantage processes », Journal of the American Society for Information Science, Vol.27, N°5, 1976, p. 292-306.

6 Dahmane Madjid. Contribution à l'étude des systèmes d'information scientifique et techniques. Thèse de doctorat université de Bordeaux II, Institut des sciences de l'information et de la communication. 1990.

7 Dobrov G M & Korennoi A A, « The information basis of scientometrics », A I Michailov et al. (eds), On theoretical problems of informatics, Moscow VINITI for FID, 1969, p 165-191

8 G. Sabah, "L'intelligence Artificielle et le langage", Vol. 2, représentation des connaissances, hermès, 1990

9 G.K ZIPF, « Human Behavior and the Principle of least Effort: An Introduction to Human Ecology Reading », Mass: Addison-Wesley, 1949.

10 Gérard Sabah, « L'IA et le langage – Représentation des connaissances, Editions Hermès, 1998
Tomi KANKAANPÄÄ, « Design and implementation of conceptual network and ontology editor », HELSINKI University of technology, ivl\STER'S Thesis, Department of Computer Science 2.6.1999

11 Gross P L K, Gross E M, College libraries and chemical education Science", October 1927, V 66, p 1229-1234

12 Hervé Rostaing, « La bibliométrie et ses techniques », Sciences de la société Collection Outils et méthodes, Co-édition Science de la Société Toulouse & Centre de Recherche Rétrospective de Marseille, 1996.

13 Jaiwei Han & Micheline Kamber (2006), Data Mining : Concepts and Techniques, Ed Morgan Kaufmann Publishers

14 Jea-Bernard Marino, « utilisation de la théorie mathématique de la communication en science de l'information », thèse docteur en 3e cycle en sciences de l'information, Ecole des hautes études en science sociales, France, 12 janvier 1984.

15 Luger. G , Studbblefield.W 1998, Artificial Intelligence, Structures and strategies for complex Problem Solving, Third Edition, Addison-Wesley Longman Inc, 824p

16 M. Berry & G. Linoff (2004), Data Mining Techniques, Ed. Willey

17 M.R.Quillian, « Semantic memory », in « Semantic information processing », Minsky editor, MITPress, 1968

18 PH. MORSE, ibrary effectivness, The M.I.T. Press, Cambridge, 1968.

19 Q.L. BURRELL, « Predictive aspects of some bibliometric process », Informetrics 87/88: Select proceedings of the first international conference on bibliometrics and theorical aspects of information retrieval, Elsevier, Amsterdam 1988.

20 R. V. L. Hartley. « Transmission of Information ». Bell System Technical Journal, July 1928

21 René Lefébure & Gilles Venturi (2001), Ed. Eyrolles

22 S. C. BRADFORD, « Sources of information on specific subject », Engineering, p. 85-86, 26 janvier 1934.

23 S. D. HAITUN, « Stationary Scientometric Distributions», Scientometrics n°4, 1982, Part I p.5-25, Part II p.89-104, Part III p.181-194.

24 Sowa J. (1984) “Conceptual Structures – Information”, Processing in Mind and Machine, Addison-Wesley, Reading Mass.

25 Sowa J. (1984) “Conceptual Structures – Information”, Processing in Mind and Machine, Addison-Wesley, Reading Mass

26 Yalaoui B. ; Dziri Ghouas A. - Sur la connaissance et l'ingénierie des connaissances .- Alger : CERIST, 2002. 20p.

27 Yalaoui Bilal, « Sur les modèles de représentation des connaissances », Rapport de recherche CERIST, Décembre 2002