

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTRE D'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE
« HOUARI BOUMEDIENNE »

FACULTE D'ELECTRONIQUE ET INFORMATIQUE



MEMOIRE

Présenté pour l'obtention du diplôme de MAGISTER

EN : INFORMATIQUE

Spécialité : Intelligence Artificielle et Base de Données Avancées

Par : DJIROUNE RAHMA

Sujet

UTILISATION DES RESEAUX BAYESIENS POUR LA RECHERCHE D'INFORMATION STRUCTUREE

Soutenu le 03/07/2005, devant le jury composé de :

Mr- A.AISSANI, Professeur, USTHB
Mme- Z.ALIMAZIGHI, Maître de conférence, USTHB
Mr- S.LARABI, Maître de conférence, USTHB
Mme- M.BOUKALA, Maître de conférence, USTHB
Mme- FZ. BESSAI, Chargée de recherche, CERIST

Président
Dteur de Thèse
Examineur
Examineur
Invité

Remerciements

Je tiens à remercier tout particulièrement mon directeur de thèse Madame Ali MAZIGHI Zaia. Je la remercie pour ses précieux conseils et pour son soutien pendant toutes ces années. Il a été un véritable plaisir de travailler avec elle. Sa bonne humeur et sa gentillesse ont également contribué au bon déroulement de mon travail.

Je suis très reconnaissant à Madame BESSAI F.Z Chargé de Recherche au CERIST qui m'a beaucoup enrichi par ses conseils, son expérience et sa disponibilité.

J'adresse mes plus sincères remerciements Mr AISSANI, qui me fait l'honneur de présider ce jury.

Je tiens aussi à exprimer ma reconnaissance envers Mr LARABI, Mme BOUKALA, d'accepter de prendre sur leur temps pour lire et juger mon mémoire.

Merci enfin à toutes les personnes qui m'ont encouragé, soutenu et aidé dans mon travail, tous mes amis ainsi que les membres de ma famille qui m'ont supporté et encouragé pour que ce mémoire puisse voir le jour.

Dédicace

Je dédie ce travail à :

La mémoire de mon père

Ma très chère Mère

Mon Mari Nouredine

Ma très chère fille Sarah

Mes frères et mes soeurs

Mes Beaux Parents

Mes Beaux frères et Mes Belles sœurs

Mes neveux et Mes Nièces

Mes Cousins et Mes Cousines

Mes Amis et Mes collègues

Table des matières

Introduction Générale

Partie A : Etat de l'art sur les Systèmes de Recherche d'Informations (SRI)

Introduction.....	1
Chapitre I – La recherche d'informations	2
I.1. Introduction.....	2
I.2. Naissance de la Recherche d'Information.....	2
I.3. Définition de la Recherche d'Information.....	3
Chapitre II – Les Systèmes de Recherche d'Informations (SRI)	4
II.1. Définition d'un SRI.....	4
II.2. Composants principaux d'un modèle de recherche d'information.....	5
II.2.1. Modèle de documents.....	6
II.2.2. Modèle de requêtes.....	6
II.2.3. Fonction de correspondance.....	6
II.2.4. Base de connaissance.....	6
II.3. L'indexation	7
II.3.1. Méthode manuelle	7
II.3.2. Méthodes Automatiques.....	7
II.4. L'interrogation.....	8
II.4.1. Les langages booléens.....	9
II.4.2. Les langages libres.....	9
II.4.3. Les langages graphiques.....	9
II.5.4. Le langage naturel.....	9
II.5. Les Modèles de Recherche d'Information.....	10
II.5.1. Les Modèles classique de la Recherche d'Information	10
II.5.1.1. Modèle booléen.....	10
II.5.1.2. Modèle vectoriel.....	11
II.5.1.3. Modèle probabiliste.....	13
II.5.2. Autres Modèles de recherche d'Information.....	13
II.5.2.1. Réseaux inférentiels Bayésiens.....	14
II.5.2.2. Réseaux sémantiques.....	15
II.5.2.3. Systèmes experts	16
II.5.2.4. L'approche connexionniste.....	17
II.6. Bilan sur l'approche classique des SRI	20
II.7. Méthodes d'évaluation d'un système de recherche d'information.....	20
Conclusion	22

Chapitre III – La Recherche d'Information Structurée (RIS)	23
III.1 Documents structurés	23
III.1.1 Définition.....	23
a- Contenu.....	23
b- Structure.....	23
c- Attribut externes.....	25
d- Sens de lecture.....	25
III.1.2 A quoi sert un document structuré ?.....	25
III.2. La recherche d'information dans les documents structurés.....	27
III.3. XML et la recherche d'information	27
III.3.1 Définition d'un corpus de document XML	28
III.3.2 La structure	28
III.3.3 Les relations structurelles	29
Conclusion	30

Partie B : Etude des Réseaux Bayésiens

Introduction	31
Chapitre I – Fondations	32
I.1 Une représentation graphique de la causalité.....	32
I.1.1 Introduction	32
I.1.2 Circulation de l'information dans un graphe causal.....	33
I.1.3 D-séparation (blocage)	35
I.2 Une représentation probabiliste associée.	37
I.2.1 Définition des variables.....	37
I.2.2 Définition des probabilité	37
I.3 Réseaux bayésien	38
I.3.1 Histoire des réseaux bayésiens.....	38
I.3.2 Origine	38
I.3.3 Définition	39
I.3.4 Fonctionnement des réseaux bayésiens.....	40
I.3.5 Propriétés des réseaux bayésiens.....	41
I.3.6 Utilisation et difficultés réseaux bayésiens.....	41
Chapitre II – Techniques & Algorithmes	42
II.1 Inférence	42
II.1.1 Approche intuitive.....	42
II.1.2 Cas général	44
II.1.3 Les principaux algorithmes.....	46

II.2 Apprentissage.....	47
II.2.1 Les principaux algorithmes.....	47
Chapitre III – <i>Mise en œuvre des réseaux bayésiens</i>	49
III.1 Pourquoi utiliser des réseaux bayésiens ?.....	49
III.1.1 Acquisition des connaissances.....	49
III.1.2 Représentation des connaissances	50
III.1.3 Utilisation des connaissances	51
III.1.4 Limites des réseaux bayésiens.....	53
III.1.5 Comparaison avec d'autres techniques	54
III.2 Où utiliser des Réseaux bayésiens ?	55
III.3 Comment utiliser des Réseaux bayésiens ?	55
III.3.1 Identification des variables et de leur espaces d'états.....	57
III.3.2 Définition de la structure du réseaux bayésiens.....	57
III.3.3 Loi de probabilité conjointe des variables	58
Conclusion	59

Partie C : Modélisation de notre prototype de RIS utilisant les RB

1. Introduction.....	60
2. Modélisation du prototype de RIS utilisant les RB.....	61
2.1 Construction du RB de notre prototype de RIS.....	61
2.1.1 Identification des variables et de leurs espaces d'états	61
2.1.2 Définition de la structure du RB.....	61
2.1.3 Loi de probabilité conjointe des variables	63
2.2. Utilisation du réseau bayésien	63
2.2.1 Indexation	63
2.2.2 La recherche d'information (interrogation).....	65
2.2.2.1. L'inférence	65
2.2.2.2. L'apprentissage.....	67
Recherche par contenu.....	67
Recherche par structure.....	67
3. Exemple	68
Conclusion	80
Application	81

Conclusion Générale

ANNEXE

Annexe A	88
1. Les descripteurs	88
2. Le thésaurus	89
3. Les différentes approches pour la Recherche d'Informations	90
Annexe B	100
1 Espace probabilité.....	100
2 Variable aléatoire.....	102
3 Théorie des graphes.....	103
Annexe C	104
Probabilité conditionnelle.....	104

Références bibliographiques

Abstract

The recent Web expansion has given chance of access to a great quantity of information. An effective management of information needs new standards document representation in contact with the Web. The structured representations are particularly developed with the standards proposal coming from XML language.

Emergence of structured documents needs model of information retrieval and browsing able to merge information coming from different media.

The aim is to allow a user to find in a structured document base, documents or parts of documents which meet a specific need.

The representation of a structured document allows the researching process to send back some components such as: a section or a chapter instead of providing the whole document.

We set out to study new methods adapt for information retrieval on corpus of structured documents to make the most of content and the logical structure of these documents.

The model we set out in this article makes for treating the problem of document representation and structured corpus in a unified way, and meets the information requirements: simple and/or structured using bayesian networks.

Keywords: Structured document, structured information retrieval, bayesian networks, XML document.

Résumé

La récente explosion du Web donne accès à une quantité d'information énorme. Une gestion efficace de l'information nécessite de nouveaux standards de représentation des documents en liaison avec le Web. En particulier, les représentations structurées se développent avec la proposition de standards issus du langage XML.

L'émergence de documents structurés, nécessite la proposition des modèles de recherche d'information et de navigation capables de fusionner des informations provenant de différents médias.

L'objectif est de permettre à un utilisateur de retrouver dans une base de documents structurés les documents ou les parties de documents qui répondent à son besoin d'information. Cette fonctionnalité devient nécessaire dans la mesure où il est courant de trouver des environnements de production de documents qui s'articulent autour de leur structure logique.

La représentation d'un document structuré permet au processus de recherche de renvoyer certains composants, par exemple : une section, un chapitre au lieu de fournir le document en entier.

Nous nous sommes attachés à étudier de nouvelles méthodes adaptées à la recherche d'information sur des corpus de documents structurés de façon à exploiter au mieux le contenu et la structure logique de ces documents.

Le modèle que nous présentons dans ce travail permet de traiter de façon unifiée le problème de la représentation des documents et des corpus structurés, ainsi que la réponse à des demandes d'information simples et/ou structurées en utilisant les réseaux bayésiens.

Introduction Générale

La Recherche Documentaire est un domaine d'importance croissante depuis déjà un certain nombre d'années. Avec l'augmentation constante du volume d'information accessible électroniquement, la notion de document électronique a considérablement évolué. Nous sommes passés, en peu de temps d'un, monde où le concept dominant était celui du document plat constitué de suite de mots ou de phrases à un monde où le document est devenu un objet plus complexe, comportant différentes sources d'information (multimédia). Cette évolution a été accélérée par le développement du web qui propose lui même ces formats structurés pour organiser et décrire ses propres ressources.

Avec cette évolution de la nature des sources d'information, sont apparus de nouveaux besoins qui visent à exploiter la richesse présente dans ces documents. Le format d'un document est aujourd'hui défini par une structure logique décrite par des primitives du langage XML. Des formats particuliers sont proposés pour standardiser certains types de documents et en faciliter ainsi leur exploitation. Le traitement de ces nouvelles sources d'information révolutionne les domaines traditionnels de l'accès à l'information. Il faut élaborer de nouveaux concepts pour la représentation, le stockage, le traitement de cette information, pour l'interaction avec l'utilisateur, pour la manipulation de très grosses masses de données a fort contenu sémantique. Les outils classiques pour le traitement des informations textuelles sont mal adaptés à l'exploitation de cette information bien plus riche et doivent être adaptés ou plus probablement redéfinis. C'est dans ce cadre que se situe le travail présenté dans ce mémoire.

Nous nous sommes attachés à étudier de nouvelles méthodes adaptées à la recherche d'information sur des corpus de documents structurés de façon à exploiter au mieux le contenu et la structure logique de ces documents. Nous nous sommes limités à des corpus textuels, mais les méthodes proposées sont assez génériques et peuvent être facilement adaptées à d'autres types de contenus.

Le modèle que nous présentons dans ce mémoire permet de traiter de façon unifiée le problème de la représentation des documents et des corpus structurés, ainsi que la réponse à des demandes d'information simples et/ou structurées en utilisant les réseaux bayésiens.

Nous présentons ce document en trois parties :

La première partie présente les bases des systèmes de recherche d'informations. Dans cette partie nous rappelons les concepts de base de la recherche d'informations. En particulier, nous présentons les mécanismes d'indexation utilisés pour obtenir une représentation des données textuelles. Nous présentons, également, les modèles de recherche, ainsi que les modèles de représentation des données sous forme de réseaux (réseaux inférentiels bayésiens, réseaux sémantiques).

Nous concluons cette première partie par un bilan qui montre les apports et les limites des approches utilisées en recherche d'informations.

La deuxième partie présente une étude sur les réseaux bayésiens. Dans cette partie, nous présentons de façon intuitive les fondements des réseaux bayésiens. Ceci nous permet de les situer par rapport aux autres techniques de modélisation, de décrire leurs applications potentielles, et les principaux algorithmes de l'inférence et d'apprentissage.

La troisième partie présente la modélisation de notre prototype de recherche d'information structurée utilisant, les réseaux bayésiens. Cette partie décrit la représentation des informations sous forme de réseau bayésien. Nous expliquons également le fonctionnement de ce réseau pour répondre à des demandes d'informations simples et/ou structurées.

En conclusion, nous dressons un bilan de nos travaux et nous introduisons les perspectives d'évolution de ces travaux.

Introduction

L'objectif des systèmes de recherche d'informations est de proposer des outils et des mécanismes assurant la gestion automatique des informations documentaires.

Un Système de Recherche d'Information (SRI) efficace doit permettre à un utilisateur de retrouver seulement les informations qui l'intéressent parmi un ensemble d'informations. Un tel système doit donc posséder des fonctions de mémorisation et d'organisation des informations ainsi que des mécanismes de recherche et de consultation de ces informations.

Les techniques assurant ces fonctionnalités sont très diversifiées et dépendent souvent du type d'informations gérées et du type d'utilisateur.

Lorsqu'il s'agit d'unités d'information textuelle, l'approche largement utilisée consiste à construire une représentation de chacune des unités sous forme d'une liste de mots clés représentatifs de leur contenu. La recherche des informations susceptibles de satisfaire la demande d'un utilisateur est réalisée par un calcul de ressemblance entre la requête de l'utilisateur et les représentations des documents.

Des mécanismes permettent d'améliorer ce processus de recherche en prenant en compte des relations existantes entre les unités d'information ou entre les termes représentatifs de ces unités.

Dans cette première partie, nous présentons les concepts de base de la recherche d'informations. En particulier, nous décrivons comment les informations textuelles sont représentées, les différents modèles pour la recherche d'informations et les outils de consultation des collections d'informations, ainsi que les différentes approches d'un système de recherche d'information.

I. La Recherche d'information

Les connaissances n'ont cessé d'augmenter, et les volumes de documents s'accroissent très rapidement. Nous arrivons ainsi à une situation parfaitement contradictoire : Jamais il n'y a eu autant d'informations disponibles, mais trouver dans cette accumulation précisément ce que l'on recherche devient de plus en plus ardu. La Recherche d'Informations (RI) devient donc de plus en plus cruciale [RADH03].

I.1. Naissance de la Recherche d'Information (RI) [NIE04]

La RI n'est pas un domaine récent. Il date des années 1940, dès la naissance des ordinateurs. Au début, la RI se concentrait sur les applications dans des bibliothèques, d'où aussi le nom "automatisation de bibliothèques". Depuis le début de ces études, la notion de pertinence a toujours été un but. Dans les années 1950, de petites expérimentations utilisant des petites collections de documents (références bibliographiques) ont été développées. Le modèle utilisé est le modèle booléen qui utilise un calcul d'appariement exact entre la représentation de la requête et celle des documents. Dans les années 1960 et 1970, des expérimentations plus larges ont été menées, elles ont abouti à une méthodologie d'évaluation du système qui est aussi utilisée maintenant dans d'autres domaines. Des corpus de test ont été conçus pour évaluer des systèmes différents. Ces corpus de test ont beaucoup contribué à l'avancement de la RI, car on pouvait les utiliser pour comparer différentes techniques et mesurer leurs impacts en pratique. Le système qui a le plus influencé le domaine est sans aucun doute SMART, développé à la fin des années 1960 et au début des années 1970. Les travaux sur ce système ont été dirigés par G. Salton, professeur à Cornell. Certaines nouvelles techniques ont été implantées et expérimentées pour la première fois dans ce système (par exemple, le modèle vectoriel qui évalue une ressemblance entre la représentation vectorielle de la requête et celle des documents).

Dans les années 1980, les techniques de l'Intelligence Artificielle (IA) ont été intégrées dans le processus de recherche d'information comme par exemple le système expert, etc.

Dans les années 1990 (surtout à partir de 1995) sont les années de l'Internet. Cela a pour effet de propulser la RI en avant scène de beaucoup d'applications. La venue de l'Internet a aussi modifié la RI. La problématique est élargie. Cependant, les techniques de base utilisées dans les moteurs de recherche sur le web restent identiques.

I.2. Définition de la Recherche d'Information [RADH03]

La Recherche d'Information (RI) concerne les mécanismes qui facilitent l'accès à une base d'informations. C'est une démarche faite par un utilisateur pour obtenir, à l'aide du Système de Recherche d'Informations (SRI), les informations (ou les références vers les informations) qui peuvent répondre à son besoin.

La recherche documentaire consiste à retrouver les documents enregistrés dans la base et qui sont pertinents par rapport à une requête donnée.

Son automatisaion a été entreprise pour différentes raisons parmi lesquelles nous citons :

- Impossibilité de canaliser le flux d'information à partir d'un certain seuil.
- Impossibilité de maîtriser les volumes de données des fonds documentaires, s'accroissant linéairement avec le temps.
- Impossibilité d'assurer une certaine précision informative à la recherche.
- Impossibilité d'accéder rapidement à l'information dès lors que le volume couvrant l'information recherchée devient important et que la question devient complexe.

La figure 1, montre les étapes d'une opération de recherche.

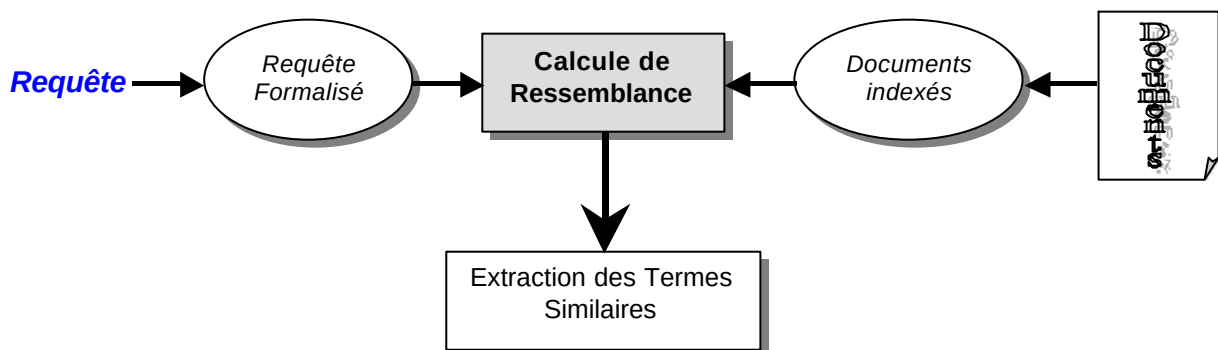


Figure 1: Les étapes d'une recherche

II. Les Systèmes de Recherche d'Informations (SRI)

Devant la masse croissante d'information et de données accessibles, de nombreuses recherches ont été mise en place au niveau de la recherche automatique d'information. Le but de ces recherches étant de mettre à la disposition des demandeurs d'information, des outils capables de trouver à partir d'un lot important de documents présents dans une ou plusieurs bases de données, quelques documents pertinents traitant d'un sujet donné. Cela abouti à la création de logiciels et d'outils de recherche documentaire.

L'objectif des logiciels documentaires, regroupés sous la dénomination plus générique de **Système de Recherche d'Information (SRI)**, est de proposer des outils et des mécanismes assurant la recherche automatique des informations documentaires. Un tel système doit donc posséder des fonctions de mémorisation et d'organisation des informations ainsi que des mécanismes de recherche et de consultation de ces informations. Les SRI assurent donc le stockage et permettent la recherche et la consultation d'informations bibliographiques ou textuelles [PILL00].

II.1. Définition d'un SRI [PILL00]

Le système de recherche d'information (SRI) efficace est un système qui permet de retrouver les **documents pertinents** à une **requête** d'utilisateur, à partir d'une base de documents volumineux dans les meilleurs délais.

Dans cette définition, il y a trois notions clés: documents, requête, et pertinence.

Document: Un document peut être un texte, un morceau de texte, une page Web, une image, une bande vidéo, etc. On appelle document toute unité qui peut constituer une réponse à une requête d'utilisateur. Dans notre étude, nous nous sommes limités aux documents textuels.

Requête: Une requête exprime le besoin d'information d'un utilisateur. Elle est, en général, de la forme suivante: "Trouvez les documents qui ...".

Pertinence: Le but de la RI est de trouver seulement les documents pertinents. La notion de pertinence est très complexe. De façon générale, dans un document pertinent, l'utilisateur doit pouvoir trouver les informations dont il a besoin. C'est cette notion de pertinence qui permet à un système de juger si un document doit être donné à l'utilisateur comme réponse.

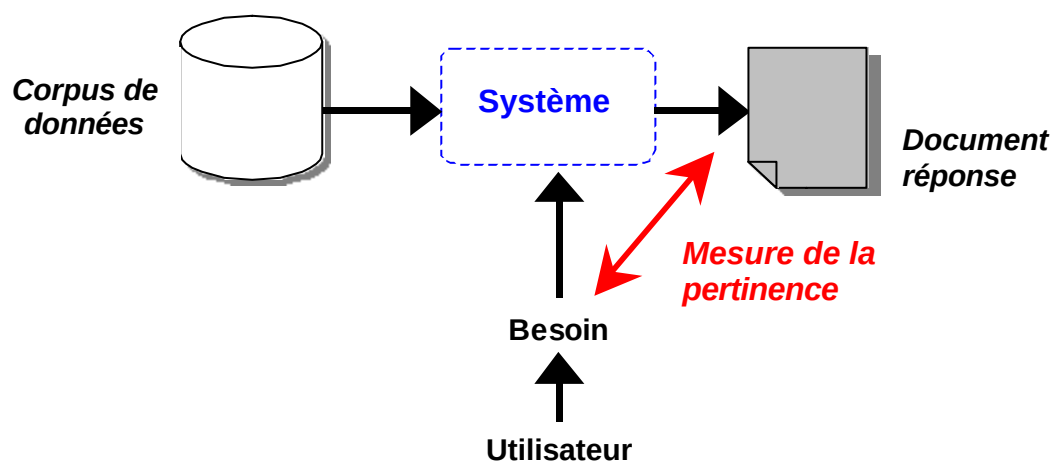


Figure 2: Fonctionnement général d'un Système de Recherche d'Information

Un SRI est un système qui traite un ensemble de documents sous forme électronique dans le but de permettre aux utilisateurs de retrouver ceux dont le contenu correspond le mieux à leur besoin d'information.

Nous distinguons donc les deux tâches principales d'un SRI :

L'indexation, c'est-à-dire l'extraction et le stockage du contenu sémantique des documents du corpus. Cette phase nécessite un modèle de représentation de ce contenu sémantique, appelé *modèle de documents*.

L'interrogation, c'est-à-dire l'expression du besoin d'information de l'utilisateur sous la forme d'une requête, la recherche dans le corpus, et la présentation des résultats. Cette phase nécessite un modèle de représentation du besoin de l'utilisateur, appelé *modèle de requête*, ainsi qu'une *fonction de correspondance* qui doit évaluer la pertinence des documents par rapport à la requête.

Avant d'expliquer, plus en détail, ces deux tâches nous présentons d'abord les composants principaux d'un modèle de recherche d'information.

II.2. Composants principaux d'un modèle de recherche d'information

À l'origine de l'utilisation d'un *système de recherche d'information* (SRI), il y a un *besoin* d'information. Pour cela, la plupart des modèles demandent à l'utilisateur d'exprimer son besoin en utilisant le langage de *requête* du modèle. À partir de la requête, la *fonction de correspondance* du modèle extrait de la base les informations qui sont susceptibles de répondre au besoin. Habituellement, cette fonction n'utilise pas les informations de la base, mais les *indexations* qui sont des représentations d'informations, dont le but est d'améliorer les performances de la fonction de correspondance (temps et qualité des résultats). Nous distinguons donc quatre composants principaux (Figure3) [GERY02].

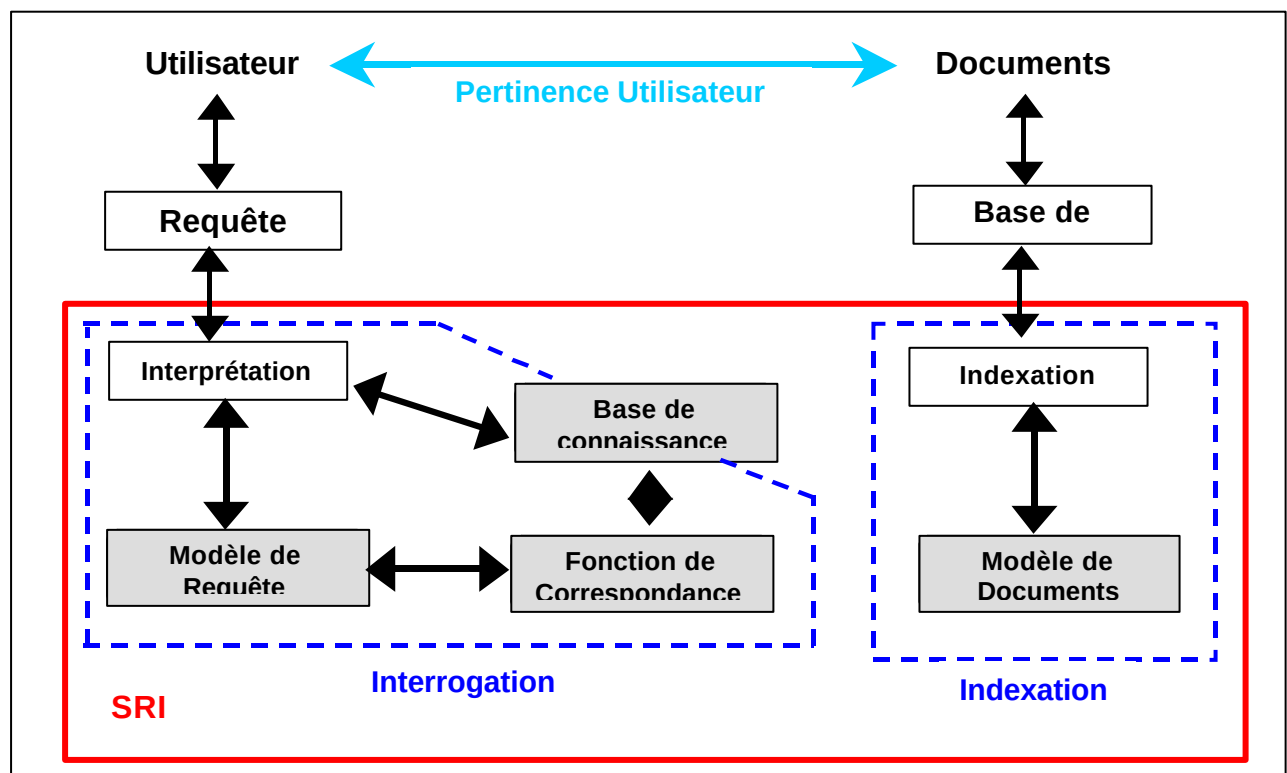


Figure 3: Schéma général d'un modèle de recherche d'information

II.2.1 Modèle de documents : correspond à la modélisation du contenu sémantique des documents, dans le formalisme de représentation de connaissances. Le choix du formalisme utilisé est crucial, mais il est toujours difficile, sinon impossible, d'obtenir une modélisation exprimant parfaitement l'idée initiale de l'auteur [GERY02].

II.2.2 Modèle de requête : correspond à la modélisation du besoin d'information de l'utilisateur, dans le formalisme de représentation de connaissances. Ce formalisme limite souvent la précision de définition du besoin. De plus, la "qualité" de la requête exprimée par l'utilisateur varie considérablement avec sa connaissance du domaine et avec son aptitude à définir son besoin. Il y a donc souvent une importante perte d'information entre le besoin et son expression [GERY02].

II.2.3 Fonction de correspondance : le système évalue la pertinence (la valeur de correspondance) des documents par rapport à la requête. La fonction de correspondance est un élément clé d'un SRI, car la qualité des résultats dépend de l'aptitude du système à calculer une pertinence des documents la plus proche possible du jugement de l'utilisateur [GERY02].

II.2.4 Base de connaissances : un thésaurus, composé de concepts apparaissant dans le corpus, reliés entre eux par diverses relations (*Spécificité/Généricité*, *Synonymie*, etc.). En considérant par exemple les relations de synonymie entre les concepts, il est ainsi possible de retrouver, pour une requête composée du terme "voiture", des documents traitant de voiture ou d'automobile.

Ces quatre éléments permettent de modéliser un processus de recherche d'information : ils forment ce qu'on appelle un **modèle de RI**, tel que représenté dans la *Figure 3* [GERY02].

II.3. L'indexation

L'indexation est le processus destiné à représenter par les éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou d'une question. On désigne également ainsi le résultat de cette opération.

L'indexation consiste donc, d'une part, à identifier les informations caractéristiques du contenu d'un ou plusieurs textes. D'autre part, il s'agit de représenter ces informations sous une forme compacte, homogène et manipulable, c'est-à-dire utilisable par un système de recherche d'information.

On distingue deux types d'indexation : [POUL02]

- **L'indexation manuelle** : une personne a préalablement désigné les termes d'indexation : les descripteurs (voir annexe A) associés à chaque texte
- **L'indexation automatique** : (par opposition appelée parfois "non-supervisée"), le programme fonctionne sans intervention humaine.

Et deux types de langages d'indexation :

- **Langage contrôlé** : Utilise un lexique de descripteurs figé, l'indexation est le plus souvent manuelle, (un professionnel choisit un ou plusieurs descripteurs pour représenter le document), parfois semi-automatique.
- **Langage libre** : Les descripteurs sont extraits automatiquement des documents ou de la requête de l'utilisateur (le plus souvent un document est indexé par la liste des mots qui le composent).

II.3.1. Indexation Manuelle

L'indexation manuelle est ajustée par des corrections humaines et est principalement adoptée par les moteurs de recherche de type thématique. Il s'agit d'un contrôle manuel des informations récoltées soit par un robot de recherche soit par soumission des utilisateurs.

Les informations sont ensuite classées et cataloguées par grandes catégories. Les catalogues thématiques, ainsi créés, sont construits à la main par des personnes filtrant les sites en fonction de leur qualité, pertinence et fiabilité. Cette méthode manuelle apporte une valeur ajoutée certaine mais la mise à jour est moins rapide et la couverture beaucoup moins large. Par ailleurs, l'intervention humaine peut biaiser la couverture géographique, linguistique ou thématique de la base [ISNET01].

II.3.2. Indexation Automatique

L'indexation automatique est un processus informatisé dont le but est de déterminer les mots clés représentatifs des concepts contenus dans un texte. Les méthodes d'indexation peuvent être classées en : [ISNET01]

II.3.2.1. Indexation en texte intégral

Elle utilise, pour l'indexation, tous les mots lexicaux (noms, verbes, adjectifs) présents dans le document. Un anti-dictionnaire élimine les mots ayant peu de valeur documentaire (mots vides) [ISNET01].

II.3.2.2. Indexation partielle du document

Elle consiste à n'utiliser que quelques termes jugés significatifs. On s'intéresse à quelques renseignements sur le document : son titre et/ou les premières lignes du texte par exemple.

La qualité de la recherche est évidemment directement proportionnelle à la qualité des renseignements partiels [ISNET01].

II.3.2.3. Indexation par méthodes statistiques [ISNET01]

Les méthodes statistiques partent de deux principes :

- Il existe une relation entre la fréquence d'un terme à l'intérieur d'un document et son importance pour la représentation du document.
- Pour l'indexation du document, il existe une relation inversement proportionnelle entre l'importance d'un terme et le nombre total de documents de la base de données contenant ce terme. Les termes rares seront de ce fait privilégiés.

Le document est indexé par ses n descripteurs, suivi de leur pondération qui est obtenue par produit du facteur d'importance interne au document et du facteur d'importance externe :

- Le facteur d'importance interne c'est la fréquence du descripteur dans le document.
- Le facteur d'importance externe c'est l'inverse du nombre de documents de la base qui comportent ce descripteur.

On définit une qualité de discrimination d'un terme dans la base en distinguant :

- ✓ Les termes très fréquents, peu spécifiques qui "attirent" les documents autour d'eux.
- ✓ Les termes très rares, utiles pour la précision de la recherche mais pas pour la recherche elle-même, vu qu'ils "rejettent" les documents.
- ✓ Les termes de fréquence moyenne qui ont tendance à regrouper des documents relativement semblables.

Ainsi, en ajoutant à la liste des termes pondérés, le regroupement des termes rares et des descripteurs composés constitués des termes peu discriminants, on améliore la couverture de recherche. Ceci se fait au détriment des temps de calculs qui sont à répéter à chaque indexation d'un nouveau document.

II.3.2.4. Indexation par les citations [ISNET01]

Dans cette méthode, on considère qu'un document répond à une question de deux manières :

Par un apport personnel de la part de l'auteur.

Par un ensemble d'autres documents cités en relation avec la question.

De même, on indexe le document de la façon suivante :

- Un ensemble de termes d'indexation.
- Un ensemble de documents indexant.

On peut calculer la similitude de deux documents à partir des termes d'indexation et des relations de co-citation qui induisent des agrégations de documents. On distingue trois types de co-citations:

- Un document A cite un document B.
- Les documents A et B sont cités simultanément par un document C.
- Deux documents A et B citent le même document C.

II.3.2.5. Indexation par des méthodes sémantiques [ISNET01]

L'indexation est composée d'une liste de descripteurs adaptés et une représentation du document dans un langage de description tenant compte des relations sémantiques entre les termes.

Pour qualifier l'environnement sémantique d'un terme, on utilise un thesaurus qui comporte les relations de type hiérarchique, d'équivalence ou d'association.

Les réseaux sémantiques sont en fait, une extension de la notion de thesaurus. Dans ces réseaux, les termes d'indexation sont des nœuds d'un graphe et les arcs sont porteurs d'une relation sémantique.

L'indexation est donc composée de la liste des descripteurs qui comportent leurs propriétés :

- "est un"
- "sorte de"
- "partie de"
- "synonyme de"

Les descripteurs héritent des propriétés de la classe à laquelle ils appartiennent, en fonction de la relation sémantique.

II.4. L'interrogation [KAHN00]

Nous introduirons deux dimensions de l'interrogation :

- la possibilité d'énoncer des requêtes à partir de la structure des documents. Il s'agit ici d'exprimer des requêtes telles que : quels sont les documents qui contiennent l'élément x? Combien de documents contiennent l'élément x ? et des combinaisons complexes de ces requêtes ;
- la possibilité d'énoncer des requêtes à partir des mots clés des documents.

Idéalement, ces deux dimensions sont combinées. Nous allons donc considérer les différents langages d'interrogation au regard de ces deux dimensions.

Nous distinguons quatre types de langage d'interrogation : les langages booléens, les langages graphiques, les langages libres et le langage naturel.

II.4.1. Les langages booléens [KAHN00]

Le langage booléen nécessite l'utilisation de différents opérateurs (le OU éventuellement inclusif ou exclusif, le ET, les opérateurs d'inclusion, les troncatures droite ou gauche, les délimiteurs d'éléments, etc.), pour formuler une requête. Ceci, tant pour les interrogations relatives à la structure que pour les combinaisons de mots à associer aux mots clés. Ce type de langage nécessite de la part du lecteur l'apprentissage du langage et sa syntaxe. La formulation des requêtes peut être simplifiée par l'ajout d'une couche graphique au langage booléen.

Par exemple, pour formuler la requête : « les décrets qui depuis 1984 légifèrent les changements de résidence » le lecteur devra taper [type: "Décret"] AND [date: ≥ 31/12/1984] AND [intitule: "changement de résidence] OR [corps de texte: "changement de résidence" OR (mutation AND remboursement.

II.4.2. Les langages libres [KAHN00]

Les langages libres correspondent à tout mode d'interrogation où le lecteur saisit sa requête sans contraintes autre que celles du langage courant. Les techniques d'appariement, employées pour associer la requête aux documents, sont variées (langage booléen, analyse statistique ou linguistique). Cependant, ces techniques n'intègrent pas de niveau sémantique et pragmatique.

Concernant l'accès aux éléments des documents, ces langages utilisent généralement une interface qui fige l'accès à certains d'entre eux.

Ces langages libres peuvent être combinés à un langage, booléen ou graphique, pour associer la formulation des requêtes en langage libre à la structure du document.

Par exemple, le lecteur formulera la requête : « les décrets qui depuis 1984 légifèrent les changements de résidence », et le système cherchera les documents qui contiennent tout ou partie des mots clés utilisés dans la requête. L'association des mots peut être de type booléen, ou émaner d'une analyse syntaxique, etc.

II.4.3. Les langages graphiques [KAHN00]

Les langages graphiques tendent à aider le lecteur à visualiser les éléments de la base d'information et leurs relations. Certains systèmes permettent de réaliser des zooms et ainsi de visualiser la structure du document à différents niveaux de granularité. L'interrogation revient de l'un ou de plusieurs des éléments (au moyen de langage booléen ou de langage libre). La requête "graphique" ainsi constituée est ensuite interprétée par l'environnement de consultation.

II.4.4. Le langage naturel

Le langage naturel recouvre les modes d'interrogation où l'utilisateur saisit sa requête sans contrainte et où la technique d'appariement employée traite de toute la complexité lexicologique, morphologique, syntaxique, sémantique et pragmatique de la langue. Notons que dans ce cadre, l'interprétation des questions en langage naturel nécessite d'articuler les aspects de contenu et les aspects relatifs à la structure du document.

Par exemple, pour interpréter la requête : « les décrets qui depuis 1984 légifèrent les changements de résidence » il faudra rechercher tous les documents dont l'élément type de texte est décret dont l'élément date est postérieur au premier janvier 1984 et dont les contenus d'article légifèrent les changements de résidence [KAHN00].

II.5. Les Modèles de Recherche d'Information

Un modèle de recherche est l'ensemble formé par le modèle de document, le modèle de requête et la fonction de correspondance entre ces deux modèles.

Nous distinguons les modèles suivants :

II.5.1. Les Modèles classique de la Recherche d'Information

Pour les documents textuels, il existe des représentations efficaces. Soit $D = \{d_1, d_2, \dots, d_i, \dots, d_m\}$ un ensemble de documents (le corpus) et $T = \{t_1, t_2, \dots, t_j, \dots, t_n\}$ un ensemble de mot-clés décrivant (indexant) ces documents.

Exemple :

$T = \{\text{Communication, Téléphonie, Multimédia, Média, Radio, Télévision}\}$

$D = \{d_1, d_2, d_3\}$, $Td_1 = \{\text{Communication}\}$, $Td_2 = \{\text{Radio}\}$ et $Td_3 = \{\text{Téléphone, Radio}\}$.

Différents modèles de comparaison dans les systèmes de recherche d'information ont été proposés.

Ceux-ci peuvent être regroupés en trois grandes catégories :

- le modèle booléen qui utilise un calcul d'appariement exact entre la représentation de la requête et celle des documents ;
- le modèle vectoriel qui évalue une ressemblance entre la représentation vectorielle de la requête et celle des documents ;
- le modèle probabiliste qui procède au calcul de la probabilité de la pertinence d'un document par rapport à la représentation de la requête.

II.5.1.1. Modèle booléen

Dans ce modèle, un document D est représenté par un ensemble de termes (ti). Une requête R est une expression logique composée de termes assemblés par les opérateurs logiques **ET**, **OU** et **SAUF**. L'approche booléenne se définit par:

a) une indexation des documents par un ou plusieurs descripteurs non pondérés.

b) une formulation de requête qui se base sur les trois opérateurs booléens:

- ↳ b1) la conjonction (connecteur **ET**) exige que les termes soient présents simultanément dans la description d'un document.
- ↳ b2) la disjonction (connecteur **OU**) exige qu'au moins un des termes soit présent dans la description des documents retrouvés.
- ↳ b3) la négation (connecteur **SAUF**) permet d'éliminer les documents possédant un terme particulier.

c) les documents sont retrouvés suivant la présence ou l'absence des termes utilisés dans l'équation de recherche.

L'approche booléenne utilise le mode d'appariement exact qui consiste à ne restituer que les documents répondant exactement à la requête.

L'évaluation de la fonction de correspondance P entre un document d et les différentes formes de requêtes est donnée par les équations suivantes :

- $P(D, t_1 \text{ ET } t_2) = P(D, t_1) \cdot P(D, t_2)$
- $P(D, t_1 \text{ OU } t_2) = P(D, t_1) + P(D, t_2) - (P(D, t_1) \cdot P(D, t_2))$
- $P(D, t_1 \text{ SAUF } t_2) = P(D, t_1) \cdot (1 - P(D, t_2))$
- $P(D, t) = \begin{cases} 1 & \text{si } t \text{ appartient au document } D \\ 0 & \text{sinon} \end{cases}$

Exemple : Si l'on cherche les documents qui contiennent les mots Téléphonie ou Radio sans Multimédia, alors on a :

$d_1 = \text{Communication}$, $d_2 = \text{Radio}$, $d_3 = \text{Téléphonie} \wedge \text{Radio}$,

Requête $R = \text{Téléphonie} \vee (\text{Radio} \wedge \neg \text{Multimédia})$ et

d_1 ne répond à la requête, d_2 répond à la requête, d_3 répond à la requête

La réponse à la requête R est $\{d_2, d_3\}$.

Avantages

- Simplicité de mise en œuvre (fichiers inverses)
- Structuration possible des requêtes

Inconvénients

- Ce modèle ne permet pas de pondérer les mots dans le document. Un document est soit pertinent, soit non pertinent, par conséquent les réponses ne sont pas ordonnées et leur nombre est difficile à contrôler.
- Les documents importants, mais qui ne sont pas indexés par les termes de la requête, ne sont pas sélectionnés.

Pour remédier aux inconvénients du modèle booléen, Salton (1983) a mis au point une technique dite "booléenne étendue". Le principe de cette approche est d'attribuer aux termes de l'équation booléenne des poids et de considérer les opérateurs de l'équation comme des distances entre les requêtes et les documents. La plupart des systèmes de recherches opérationnels sont actuellement basés sur l'approche booléenne traditionnelle [IHAD99].

II.5.1.2. Modèle vectoriel [IHAD99]

Dans ce modèle, les termes d'indexation forment une base de l'espace vectoriel dans lequel chaque document est un vecteur représenté par une combinaison linéaire de ces termes:

$$d_i = a_{i1}t_1 + a_{i2}t_2 + \dots + a_{in}t_n$$

Les coefficients a_{in} sont les poids du terme t_j pour le document D .

Une requête R est aussi formalisée par un vecteur de termes de recherche pondérés:

$$R = b_{i1}t_1 + b_{i2}t_2 + \dots + b_{in}t_n$$

b étant le poids du terme t dans la requête.

Durant le processus d'évaluation de la pertinence d'un document en regard de la question **R**, le système sélectionne les documents **D** dont la similarité avec la requête est la plus grande. Pour calculer cette similarité, on peut utiliser diverses méthodes comme celle du cosinus, de Jaccard, de Dice ou le produit d'Inner. La fonction **cosinus** consiste à mesurer le cosinus de l'angle des deux vecteurs **R** ET **D**.

Les documents sont alors retrouvés en fonction de leur proximité avec la question. Les réponses sont classées et restituées en fonction de cette mesure. Les travaux sur l'utilisation de ces méthodes montrent que la mesure de similarité qui correspond au **cosinus** donne les meilleurs résultats en termes de précision et de rappel.

Cette possibilité de jugement de pertinence est utilisée pour reformuler automatiquement les questions en prenant en compte les documents pertinents et non pertinents. Le prototype SMART est un exemple du modèle vectoriel.

Il y a plusieurs façons de calculer la similarité entre deux vecteurs. En voici quelques unes: [NIE04]

$$Sim0(D, R) = \sum_i (a_i * b_i) \text{ (produit interne)}$$

$$Sim1(D, R) = \sum_i (a_i * b_i) / [\sum_i (a_i)^2 * \sum_i (b_i)^2]^{1/2} \quad (\text{cosinus})$$

$$Sim2(D, R) = 2 \sum_i (a_i * b_i) / [\sum_i (a_i)^2 + \sum_i (b_i)^2]$$

$$Sim3(D, R) = \sum_i (a_i * b_i) / [\sum_i (a_i)^2 + \sum_i (b_i)^2 - \sum_i (a_i * b_i)]$$

Exemple:

d1, d2, d3 trois documents

R est une requête

Le document d2 est le document
le mieux classé

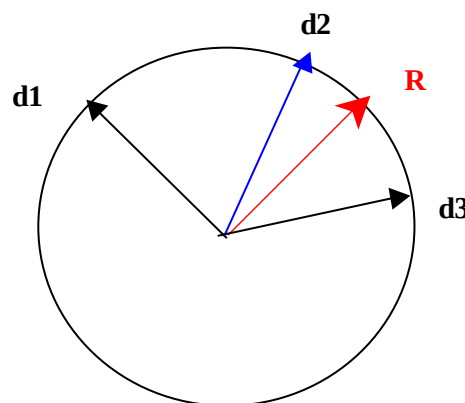


Figure 4: Exemple de la Représentation vectorielle des documents et requêtes

Avantages et inconvénients du modèle vectoriel:

La puissance du modèle vectoriel réside dans sa simplicité conceptuelle et de mise en oeuvre. Il permet de représenter, de manière uniforme, un vecteur document et une requête et intègre simplement des pondérations sur les mots-clés attachés tant à un document qu'à une requête. Il offre aussi des moyens simples pour classer les résultats d'une recherche, en plaçant en tête les documents jugés les plus pertinents par le système (à travers une mesure de similarité) [SMAI94].

Ce modèle a été critiqué essentiellement par rapport à l'orthogonalité implicite de l'espace de représentation supposant une indépendance entre les termes. Une autre source de critique est, outre la limite du pouvoir d'expression, la non unicité du choix et le manque de justification théorique de certains composants du modèle, telle la fonction de similarité utilisée [SMAI94].

II.5.1.3. Modèle probabiliste [IHAD99]

Ce modèle est basé sur le fondement de la théorie de la décision où le but est de calculer la probabilité que le document soit pertinent par rapport à une demande formulée par un usager. Etant donné une requête **R**, elle découpe la banque de données en deux ensembles: les documents pertinents et ceux qui ne le sont pas. La probabilité que la réponse soit pertinente, étant donné **D**, est exprimée sous une forme conditionnelle : **$P(\text{pertinent}/D) = P1$**

D'après le théorème de Bayes:

$$P1 = \frac{Pp(D) \times P(\text{pertinent})}{P(\text{pertinent}) \times Pp(D) + P(nper) \times Pn.p(D)}$$

$Pp(D)$ et **$Pn.p(D)$** représentent respectivement la probabilité pour **D** d'être un document pertinent ou non pertinent.

$P(\text{pertinent})$ et **$P(nper)$** sont respectivement la probabilité de trouver à priori des documents pertinents et non pertinents.

Pour un corpus donné, le calcul de **$P1$** est réduit à celle de **$Pp(D)$** et **$Pn.p(D)$** . Un moyen de les estimer consiste à établir les probabilités de pertinence et de non-pertinence d'un document en fonction des probabilités des termes contenus. On estime cette probabilité selon la pondération des mots de la question et des termes d'indexation.

Le poids des termes est souvent déterminé en fonction :

- De la fréquence relative du terme **t** dans le document,
- De la fréquence relative du terme **t** dans la collection,
- De la taille du document.

Avantages

- Précision des résultats

Inconvénients

- Paramétrage très lourd
- Dépendant du fonds original (gestion de l'évolution très peu souple)

II.5.2. Autres Modèles de recherche d'Information

Des modèles de représentation des documents ont été proposés pour tenir compte de la dépendance entre les termes (d'indexation). Les réseaux inférentiels Bayésiens s'appuient sur des calculs probabilistes et statistiques pour calculer la probabilité de pertinence des documents, sachant que certains événements ont été rencontrés.

D'autres techniques, issues de l'intelligence artificielle, ont été adaptées aux besoins des SRI. En particulier, les réseaux sémantiques, utilisés en compréhension du langage naturel, ont été utilisés pour enrichir la modélisation des documents. Les systèmes experts ont également été utilisés en RI.

II.5.2.1. Réseaux inférentiels Bayésiens [MOTH97]

Les réseaux inférentiels Bayésiens complètent le modèle probabiliste en tenant compte de la dépendance entre les termes d'indexation.

II.5.2.1.1. Concepts de base

Un réseau inférentiel Bayésien est un graphe de dépendances, acyclique et orienté. Il est composé de nœuds qui représentent les variables de proposition ou des constantes, et de liens qui représentent les relations de dépendances entre les propositions. Si la proposition représentée par le nœud p implique la proposition représentée par le nœud q , on aura un lien de p vers q . Le nœud q contient alors une matrice de liens qui spécifie, pour toutes les valeurs possibles de p et de q , la force de l'implication (calculée par la probabilité conditionnelle $P(q/p)$).

Si le nœud q à des parents multiples (chacun représentant une des causes potentielles de q), la matrice spécifie la dépendance du nœud q sur l'ensemble des parents.

A partir d'un ensemble de probabilités préalables pour les racines des graphes, on peut calculer la probabilité ou degré de croyance associé à tous les nœuds restants.

II.5.2.1.2. Utilisation des réseaux inférentiels en RI

Les réseaux inférentiels Bayésiens sont utilisés en RI pour représenter les dépendances entre termes et entre termes et documents. Ils reprennent des éléments du modèle probabiliste, en particulier le calcul de probabilité conditionnelle. Par rapport à ce modèle, les réseaux inférentiels permettent d'utiliser diverses sources d'évidence pour calculer les probabilités (représentation du contenu des documents).

Ce modèle comporte deux réseaux : le réseau de documents qui est construit définitivement et le réseau de requête qui est différent à chaque requête et peut évoluer au cours de l'interrogation.

Le réseau de documents comporte trois types de nœuds interconnectés (figure 5) :

- les nœuds documents **di** représentent les documents de la collection,
- les nœuds texte **ti** correspondent à la représentation des documents,
- les nœuds représentation d'un concept **ri** correspondent aux concepts décrivant un document. Ces nœuds peuvent être partagés par plusieurs documents et un document peut être caractérisé par plusieurs nœuds représentation.

La probabilité initiale d'observer un document est donnée par $P(di) = 1/N$. Chaque nœud texte contient la spécification de sa dépendance avec son père ($P(ti/di) = 1$) et chaque nœud représentation contient la spécification des probabilités conditionnelles des dépendances avec ses parents.

Le réseau de requête possède une seule feuille correspondant au fait qu'une information recherchée est rencontrée. Il possède plusieurs racines correspondant aux concepts d'interrogation. Un nœud concept d'interrogation contient une spécification de sa dépendance avec les nœuds représentations du réseau des documents. Pour simplifier, on suppose qu'un nœud interrogation n'a qu'un seul père.

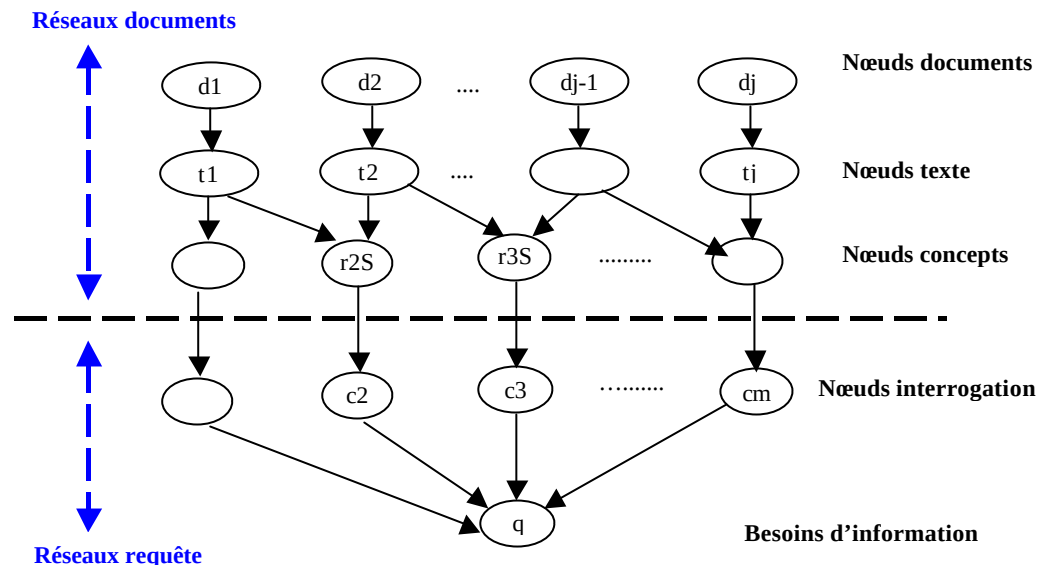


Figure 5: Réseau inférentiel Bayésien [MOTH97]

Le principe de recherche de l'information est le suivant : pour une requête donnée, on crée le réseau de requête et on le rattache au réseau de documents, on peut alors calculer la probabilité qu'une information pertinente soit rencontrée sachant qu'un document donné a été observé ; cela pour chaque document de la base. On peut alors ordonner chacun des documents en fonction de la probabilité trouvée.

Les réseaux inférentiels permettent une recherche qui prend en compte la dépendance des termes entre eux ainsi que leur dépendance avec les documents [MOTH97].

II.5.2.2. Réseaux sémantiques

Les modèles sémantiques ont été introduits afin de tenir compte du contenu sémantique des documents et des requêtes, en intégrant des connaissances sur le sens des termes d'indexation et sur leurs liaisons sémantiques.

II.5.2.2.1. Concepts des réseaux sémantiques [MOTH97]

Le concept de réseau sémantique a été introduit comme un mécanisme d'association général pour encoder le sens des mots. Un réseau sémantique permet de décrire le sens d'un texte en langage naturel de façon formelle : la connaissance est représentée par l'intermédiaire de nœuds et de liens. Les nœuds représentent des objets, des concepts, des faits, etc. ; les liens représentent des relations entre ces nœuds.

Pour simplifier l'interprétation du graphe, ces liens typés peuvent posséder un label (AGENT, OBJET, INSTRUMENT, SOURCE ou DESTINATION), Salton distingue deux sortes de liens :

- ✓ les liens ayant une valeur booléenne VRAI, FAUX (réseaux logiques) : “x est le père de y” (prédicat où x et y sont des variables), “Paul est le père de Jeanne” (proposition : prédicat où les variables ont été remplacées par des entités),
- ✓ les liens flexibles (graphes conceptuels): “les objets physiques ont un poids”, “Paul pèse 85kg”.

Ces systèmes se différencient par les méthodes de déduction utilisées.

Dans un **réseau logique**, le graphe est interprété comme un énoncé logique de propriétés des différentes entités. Les primitives de base d'un réseau logique sont les prédicats définis pour les entités sans discussion. Des propositions distinctes peuvent être combinées en utilisant les opérateurs logiques (et, ou, non).

La théorie des **graphes conceptuels** a été conçue dans le but de définir un modèle général de représentation des connaissances, quelle que soit leur forme. Ce modèle est adapté pour traiter des données exprimées en langage naturel.

Un graphe conceptuel se compose de concepts (**nœuds**) et de relations conceptuelles. Un concept possède un type (exemple : PERSONNE) et un référent qui précise exactement l'instance représentée par le concept (exemple : PAUL). Le référent peut représenter plusieurs individus du même type (quantifié ou non). Le référent du concept peut ne pas être précisé, c'est alors un concept générique (représentant tout élément du type correspondant). Les relations conceptuelles sont les liens qui existent entre les concepts (Exemple : [PERSONNE : 'PAUL'] <-- (AGT) <-- [MANGER]) spécifiant les rapports entre concepts. Les relations sont orientées, peuvent avoir un nombre quelconque d'arguments et sont typées (exemple (localisation spatiale), (AGT)). Elles peuvent posséder plusieurs niveaux, par exemple, pour deux niveaux, le premier indique la relation (exemple : localisation temporelle), le second niveau précise le sens donné au premier niveau (exemple : (début, fin)). On peut ainsi créer une hiérarchie de relations.

De la même façon, il existe une hiérarchie de concepts : un concept a un type qui peut lui-même avoir un type (exemple : CHIEN de type MAMMIFERE de type ANIMAL). Les graphes de départ expriment les contraintes sémantiques du domaine représenté.

II.5.2.2.2. Utilisation en recherche d'information [MOTH97]

Les réseaux sémantiques ont été utilisés dans différentes applications, pour la plupart dans le domaine du traitement du langage naturel. Plus récemment, ils ont également été utilisés pour les SRI. En particulier, les thésaurus on-line pour les systèmes de recherche, ont été représentés en utilisant la structure des réseaux sémantiques. La prise en compte de la sémantique se fait en représentant les termes d'indexation sous forme de graphe. Les relations sémantiques classiques sont : la synonymie, la généralisation/spécialisation et le voisinage sémantique.

Dans certains modèles les auteurs proposent en plus des relations de généralité et de spécificité (lien NT/BT) entre termes d'indexation, un lien UTILISE qui relie des termes d'indexations et d'autres termes. La prise en compte de la sémantique peut se faire à un second niveau, par l'enrichissement de la notion de terme pour exprimer la connaissance du domaine.

II.5.2.3. Systèmes experts

II.5.2.3.1. Composants d'un système expert [MOTH97]

Un système expert comporte essentiellement trois composants :

- un langage d'expression des connaissances fournies par les experts,
- une base de connaissances pour représenter la connaissance du domaine,
- un moteur d'inférences pour exploiter les connaissances

La base de connaissances comprend d'une part des connaissances assertionnelles, ou faits, qui décrivent des situations et d'autre part des connaissances opératoires, ou règles, qui représentent le savoir faire sur le domaine.

Les faits sont souvent représentés sous une forme structurée du type <attribut><objet><valeur>, des prédicats ou des frames (objets), reliés entre eux dans des réseaux sémantiques ou des graphes.

Les règles ont généralement la forme « si condition alors action (règle de production) ». Le moteur d'inférence puise parmi les règles, les interprète, les enchaîne jusqu'à satisfaire des conditions d'arrêt.

II.5.2.3.2. Utilisation en Recherche d'Informations [MOTH97]

Le principe de recherche d'informations consiste à réécrire le document jusqu'à ce qu'il corresponde à la requête. Plus il est changé par rapport au document initial, moins il est considéré comme susceptible de répondre à la requête.

La base de connaissances contient également des règles d'inférences. Ces règles permettent au système de déduire la typologie de l'utilisateur en fonction de la formulation de sa requête et personnalise ainsi les réponses. Par exemple, si les termes utilisés dans la requête sont très spécifiques, le système déduira que l'utilisateur est un spécialiste et ne devra pas restituer des documents trop généraux.

II.5.2.4. L'approche connexionniste [MOTH97], [BOUG92]

Cette approche est le fruit de recherche en intelligence artificielle. Elle a pour but d'imiter certaines fonctions du cerveau humain. L'application de cette approche à la recherche documentaire consiste à représenter une collection d'information sous forme d'un réseau de neurones. Cette représentation doit permettre l'évolution de la connaissance par auto-apprentissage.

Les principes de l'approche connexionniste au domaine de la RI., sont les suivants :

- ✓ le **stimulus externe** correspond à la requête de l'utilisateur,
- ✓ le **problème à résoudre** par le réseau est la recherche des documents susceptibles d'être pertinents par rapport à la requête,
- ✓ la **réponse correspond** aux documents jugés pertinents par le réseau,
- ✓ l'**apprentissage** doit permettre une optimisation du processus de recherche par la limitation du bruit et du silence.

II.5.2.4.1. Neurones formels [MOTH97]

Le fonctionnement d'un neurone formel (automate) est inspiré du fonctionnement connu des neurones réels.

D'une façon générale, un neurone formel *i* reçoit des entrées des neurones auxquels il est connecté en tant que successeur ; ce neurone calcule alors une somme pondérée de ces entrées qui correspond à son activation. Il calcule ensuite sa sortie, en fonction de son activité et éventuellement d'un seuil propre. Cette sortie est transmise aux neurones auxquels *i* est connecté en tant que prédécesseur (figure 6).

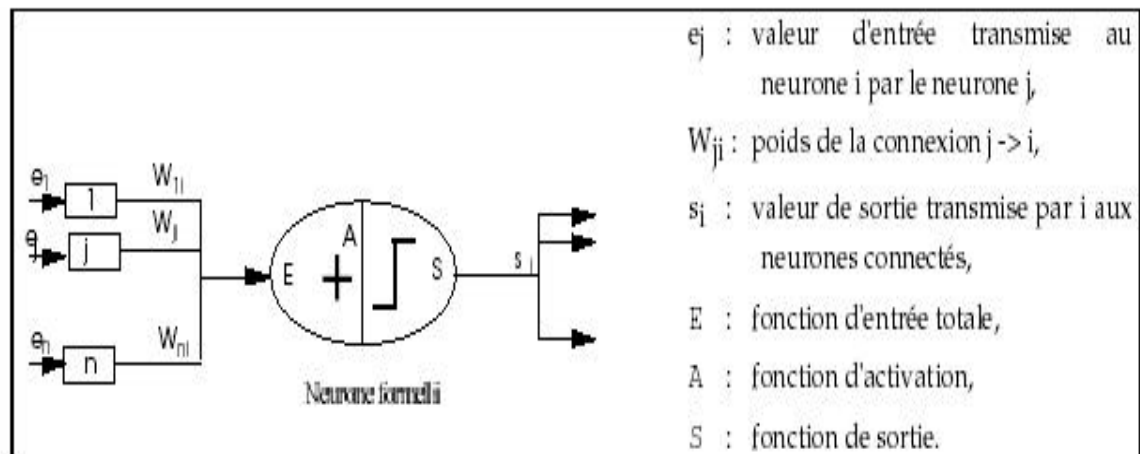


Figure 6: Modélisation d'un neurone [MOTH97]

II.5.2.4.2. Modélisation des connexions entre neurones

Les réseaux de neurones biologiques sont des réseaux partiellement connectés et sont composés de neurones de divers types. Ces connexions sont assurées au niveau des synapses qui assurent la transmission de l'influx nerveux entre neurones.

Les réseaux neuromimétiques -ou réseaux de neurones formels- sont constitués d'automates d'un même type. Plusieurs types de réseau ont été définis (figure 7) :

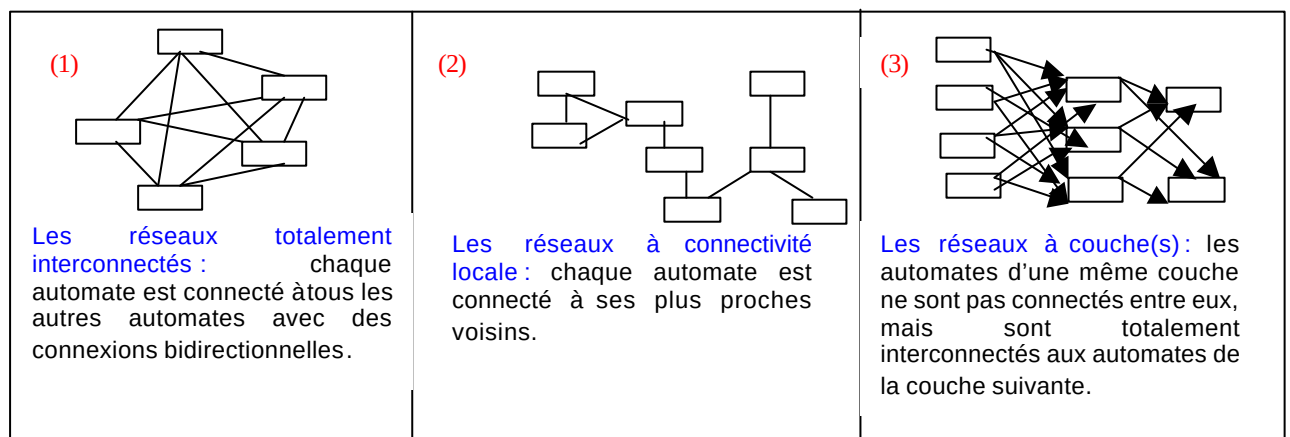


Figure 7: Les différents types de connectivité dans les réseaux neuromimétiques

Dans les réseaux à couches (type 3), les deux couches externes correspondent à celle recevant les entrées du milieu extérieur et à celle fournissant les résultats ; les autres, si elles existent, sont appelées couches cachées. Les connexions sont orientées de la couche d'entrée vers la couche de sortie. Dans les autres types de réseau, la sortie correspond à un état global à un instant donné.

Dans tous les cas, les connexions sont pondérées, généralement par une valeur réelle. Ce poids traduit le caractère excitateur ou inhibiteur de ces connexions. Il rend donc compte de l'importance de l'influence d'un automate sur un autre. Lorsque le nœud i a un effet excitateur sur le nœud j , l'activation du nœud i tend à accroître celle du nœud j [MOTH97].

II.5.2.4.3. Modèles connexionnistes pour la Recherche d'Informations

Les Systèmes de Recherche d'Information, utilisant l'approche connexionniste peuvent être divisés en deux catégories, en fonction du type d'architecture de réseau connexionniste qu'ils utilisent [MOTH97].

II.5.2.4.3.1. Modèles à auto-organisation [MOTH97], [BOUG92]

Les SRI, basés sur un modèle connexionniste à auto-organisation, utilisent la représentation du contenu sous forme de "carte", ces systèmes basent sur le modèle des cartes topologiques de Kohonen. Dans ces systèmes, un vecteur de poids est associé à chacun des neurones composant le réseau. Chaque document est présenté au réseau comme un vecteur entrée ; les coordonnées correspondent aux termes représentatifs des documents. A partir de ce vecteur, le réseau recherche le neurone gagnant (neurone le plus actif) et renforce ses poids ainsi que les poids des neurones les plus proches. Le document, correspondant au vecteur d'entrée, est rattaché au neurone gagnant.

L'opération est répétée, pour l'ensemble des documents de la collection, jusqu'à convergence du réseau. Cet algorithme permet donc de classer les documents en fonction de leur ressemblance par rapport aux caractéristiques d'entrée (termes d'indexation).

L'auto-organisation du réseau permet de conserver la topologie de l'ensemble des documents, c'est-à-dire que les documents similaires seront représentés par le même neurone et que les documents proches seront représentés par des neurones physiquement proches sur la carte. De plus, lorsqu'un grand nombre de documents sont proches, plusieurs neurones pourront être dédiés à leur représentation. Les neurones composent alors la "carte sémantique" de la base (figure 8).

La consultation de la base est réalisée à partir de la carte construite : l'utilisateur choisit une zone (un *noeud*) de la carte ; les documents rattachés à cette zone lui sont restitués.

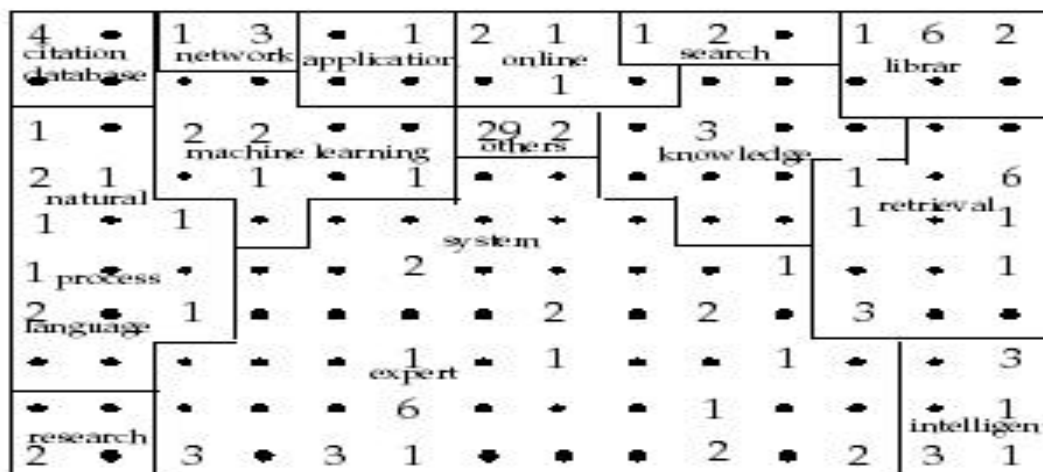


Figure 8: Carte sémantique

Les domaines sur la carte sont construits automatiquement, les nombres sur la carte correspondent aux nombres de documents associés à chaque *noeud*.

II.5.2.4.3.2. Modèles à couches

Les SRI, basés sur un modèle connexionniste à couches, sont en général modélisés par un réseau pondéré non aléatoire, construit sur une représentation initiale de la connaissance. Les neurones (ou *noeuds*) correspondent aux informations à gérer (documents, images ...etc) et aux items descriptifs qui leur sont associés. Ces items sont soit des attributs, au sens classique, comme par exemple le

nom des auteurs, les caractéristiques de publication pour des informations textuelles,... etc, soit des mots-clés descriptifs, plus ambiguës ou imprécis. Le mécanisme de recherche est basé sur une activation initiale des neurones termes, induite par une requête qui se propage vers les documents à travers les connexions du réseau. Généralement, la construction d'un réseau est basée sur des modèles utilisés par ailleurs (modèle vectoriel, modèle probabiliste).

Il s'agit donc d'un modèle incrémental, en ce sens que les résultats, obtenus avec les différents modèles de recherche, ne sont pas remis en cause mais permettent de définir de nouveaux modèles plus performants. Ces systèmes diffèrent par des caractéristiques inhérentes aux SRI et par des caractéristiques propres aux systèmes connexionnistes [MOTH97].

II.6. Bilan sur l'approche classique des SRI

La majorité des SRI commercialisés, gérant des données textuelles, sont basés sur l'approche dite *classique*. Elle comprend une analyse, manuelle ou plus souvent automatique, du contenu de chaque document. Cette analyse permet d'associer à chaque document un ensemble de mots-clés, éventuellement issus d'un lexique préétabli. Ces mots-clés ou termes d'indexation sont considérés comme représentatifs du contenu et constituent le langage d'indexation ou thésaurus. Ils sont utilisés pour la phase de recherche et sont souvent organisés en fichier inverse pour faciliter le mécanisme de recherche. Lorsque le système supporte l'interrogation en langage naturel, la requête subit un processus d'analyse similaire dont le résultat est une représentation utilisant les termes du langage d'indexation. Les représentations des documents et de la requête sont mises en correspondance afin de déterminer le degré de ressemblance des différents documents avec la requête. Dans ce cas, l'organisation en fichier inverse permet de ne prendre en compte que les documents indexés par au moins un des termes de la représentation de la requête. Les documents les plus ressemblants sont alors restitués à l'utilisateur. Les interfaces de communication proposées sont en général assez conviviales et simples d'utilisation.

Les performances de ces systèmes sont mesurées en terme de documents pertinents et de documents restitués pour une requête. Ces performances dépendent du mécanisme d'indexation et d'analyse de la requête et du mécanisme de calcul de similarité entre les documents et les requêtes [MOTH97].

II.7. Méthodes d'évaluation d'un SRI

On distingue habituellement deux critères d'évaluation différents :

- Les critères quantitatifs : combien de documents peuvent être indexés, quel est le temps de réponse maximum à une requête ?
- Les critères qualitatifs : quelle est la pertinence des réponses ?

Les critères quantitatifs ne posent pas de problème, ils sont directement quantifiables.

Nombre de Kilo-octets maximum autorisé, nombre de secondes ou millisecondes pour la réponse à une question de "taille" moyenne, ou le rapport temps de réponse en fonction du nombre de documents...

L'évaluation de la **qualité** d'un système de recherche d'information peut se faire selon des critères subjectifs (réponses pertinentes, peu pertinentes, inadaptées, aberrantes...), mais, si l'on veut être le plus objectif possible, il faut trouver des métriques pour quantifier la pertinence des réponses.

Le principe consiste à répartir les documents en trois ensembles pour une requête donnée :

- Les documents correspondant réellement à la requête (pertinents)
- Les documents ne correspondant réellement pas à la requête (non pertinents) (ces deux premiers ensembles étant disjoints)
- Les documents retournés par le SRI. (retournés)

Ces trois ensembles peuvent être schématisés ainsi : [POUL02]

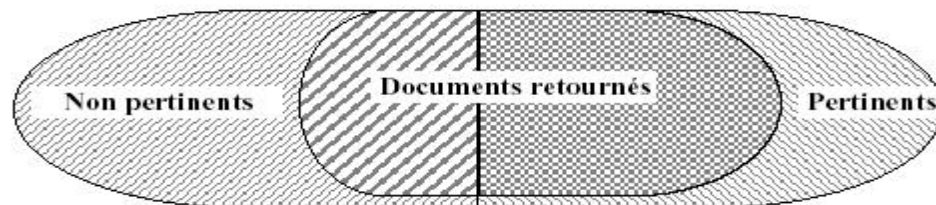


Figure 9: Evaluation, Classification des documents

On peut ensuite quantifier les résultats en termes de bruit et de silence :

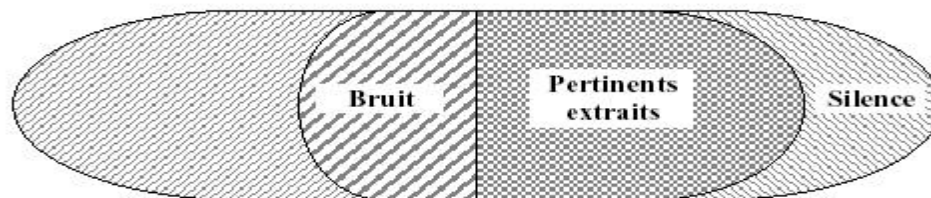


Figure 10: Evaluation, bruit et silence

Le bruit représente les documents extraits mais non pertinents

$$\text{Bruit} = \frac{\text{Nombre de documents retournés et non pertinents}}{\text{Nombre de documents extraits}}$$

Le silence représente les documents pertinents non extraits

$$\text{Silence} = \frac{\text{Nombre de documents non retournés et pertinents}}{\text{Nombre de documents extraits}}$$

Un SRI sera d'autant meilleur que le bruit et le silence seront faibles. On pourra représenter le bruit et le silence en proportion du nombre de documents extraits.

Deux autres mesures sont souvent utilisées, qui sont en fait les compléments des deux précédentes :

La précision qui représente le nombre de documents pertinents extraits par rapport au nombre de documents extraits

$$\text{Précision} = \frac{\text{Nombre de documents retournés et pertinents}}{\text{Nombre de documents extraits}}$$

Le rappel, qui représente le nombre de documents pertinents extraits par rapport au nombre de documents pertinents

$$\text{Rappel} = \frac{\text{Nombre de documents retournés et pertinents}}{\text{Nombre de documents pertinents}}$$

Conclusion

Les principaux modèles de recherche d'information pour des documents textuels sont le modèle booléen, le modèle vectoriel et le modèle probabiliste. Tous ont donné lieu à des systèmes utilisés, à grande échelle notamment, via les serveurs d'information et certains sont actuellement à la base de moteurs de recherche disponibles sur le Web.

Cependant, ils ne répondent que partiellement aux contraintes imposées par l'évolution des corpus, qui peuvent maintenant contenir des textes, des images, des vidéos, des sons, ou encore des documents structurés, combinant ces divers médias. Ces modèles sont mal adaptés au traitement d'information multimédia structurée, car principalement dédiés aux documents textuels non structurés.

Dans ce type de systèmes, un document est considéré comme une entité atomique qui est représentée et retrouvée comme un tout par le système et donc c'est le document entier qui est présenté à l'utilisateur comme réponse à sa requête.

Cependant, les documents affichent souvent une structure définie par l'auteur. Par exemple, un document peut avoir plusieurs sections; chacune avec plusieurs sous sections et ainsi de suite.

Avec un document structuré, l'unité de recherche et de présentation devrait être les composants du document au lieu du document, car souvent seules des parties du document sont appropriées à un besoin d'information, et par conséquent, devront être retournées à l'utilisateur. D'ailleurs, la représentation d'un document structuré doit permettre au processus de recherche de retourner certains composants, par exemple, une section, un ensemble de requête au lieu de fournir le document entier.

Bien que les systèmes de recherches d'informations traditionnels soient d'un grand apport dans le domaine de recherche d'information, ces derniers restent insuffisants quant aux besoins et exigences actuels des utilisateurs, imposés par le développement remarquable de la technologie, d'une part; et une gestion efficace de l'information qui nécessite plus que jamais sa représentation dans un format adéquat, préservant son indépendance vis-à-vis de toute plate-forme et/ou architecture, d'autre part.

Par conséquent, produire des facilités de recherche, selon des aspects structurels, est un problème crucial pour les domaines des bases de données et de la recherche d'informations structurées. C'est pourquoi une nouvelle génération de SRI est entrain de voir le jour, ce sont les **SRI structurées**.

III. La Recherche d'Information Structurée

La récente explosion du Web donne accès à une quantité d'information énorme. Une gestion efficace de l'information nécessite de nouveaux standards de représentation des documents en liaison avec le Web et avec les bases de données. En particulier, les représentations structurées se développent avec la proposition de standards issus du langage XML. L'utilisation de descriptions structurées pour les documents s'est rapidement développée et est entrain de s'imposer.

La recherche d'information apporte peu d'outils pour le traitement de ces documents structurés : les représentations des documents en RI ont été conçues pour des documents « plats » et homogènes, ils ne sont pas adaptés au traitement simultané de la structure et du contenu.

Différents travaux, concernant l'utilisation de l'information de structure pour la RI, ont déjà été développés. Un texte sera représenté par une information plus complète que son simple contenu textuel. Il sera accessible à différents niveaux de description et pourra contenir différents types d'information (e.g. texte et meta-données). L'accès simultané à la structure des documents et à leur contenu permet d'envisager de nouveaux modes d'exploitation de l'information textuelle. Celle-ci nécessite la création de nouveaux outils et de nouveaux modèles capables d'exploiter de telles données [BORD00].

Avant de définir la recherche d'information structurée, il faut savoir : ce qu'est un document structuré ?

III.1. Documents structurés

Un document se présente rarement sous la forme d'un "bloc" textuel car il est utile de le structurer pour établir une décomposition des thèmes abordés et une hiérarchie entre les différentes sous parties. L'auteur structure le document pour faciliter la compréhension de l'information qu'il cherche à communiquer [GERY02].

III.1.1. Définition [GERY02]

Classiquement, un document structuré est composé d'un *ensemble de parties* (le contenu), organisées de manière *hiérarchique* (la structure logique). De plus, on peut définir des *attributs externes* associés à chacune des parties. Un auteur rédige un document dans le but de communiquer une information aux lecteurs, avec une cohésion entre les différentes parties.

Nous allons maintenant présenter la notion de document structuré, à travers les définitions de ses principaux composants : contenu, structure, attributs externes et sens de lecture.

a) Contenu

Le contenu d'un document structuré désigne le contenu, textuel ou multimédia, représenté sous la forme d'un ensemble de fragments insécables et non structurés, comme par exemple des paragraphes, des figures ou des images. Le contenu est alors l'atome de description des documents.

b) Structure

La création et l'échange de documents sur des plateformes hétérogènes ont conduit à la définition de normes de représentation de documents structurés, telles ODA (Office Document Architecture), SGML. Ces normes distinguent deux types de structures, la structure physique et la structure logique, qui sont définies de la manière suivante :

Structure physique : correspond à l'organisation d'affichage des données qui composent le document c'est-à-dire la présentation. Généralement, un document est composé d'un ensemble de pages, elles-mêmes composées d'une en-tête, de lignes, de notes de bas de pages, de figures, etc. (figure 12). La structure physique dépend de l'environnement de présentation du document, comme le format du papier ou l'écran d'un ordinateur.

Structure logique : correspond à l'organisation hiérarchique des données du document. La structure logique est spécifiée par l'auteur. Elle lui permet de décomposer et d'organiser le document pour mieux exprimer ses idées, généralement à l'aide d'abstractions représentant des parties du document. Par exemple, un document se compose d'un titre, d'une ou plusieurs sections, elles-mêmes composées d'un titre, d'une ou plusieurs sous-sections, etc. (figure 11).

La structure logique propose un “**sens de lecture**” implicite, qui correspond à l'enchaînement des idées de l'auteur. La décomposition d'un document structuré reste la même quel que soit l'environnement de présentation : la structure logique est donc indépendante de celui-ci.

Par contre, la structure physique est dépendante de cet environnement, qui n'est généralement pas connu du SRI. On ne peut donc pas en tirer d'information relative au document : par exemple, les pages d'un document au format A4 ne présenteront pas les mêmes informations que celles du même document au format A3. La structure physique n'est d'aucune utilité pour la RI, sauf si l'utilisateur précise le mode de présentation dans la requête.

La structure logique est définie à la création du document. Elle est porteuse d'une partie du message que l'auteur cherche à exprimer. Par exemple, nous avons choisi de séparer en deux sections intitulées “Documents structurés” et “Hypertextes” la question des documents structurés et celle des systèmes hypertextes, car il s'agit de deux domaines distincts.

La structure logique est donc susceptible d'être utilisée au cours d'un processus de recherche d'information, comme information supplémentaire pour améliorer les résultats de la recherche, mais aussi comme critère de recherche mis à la disposition de l'utilisateur.

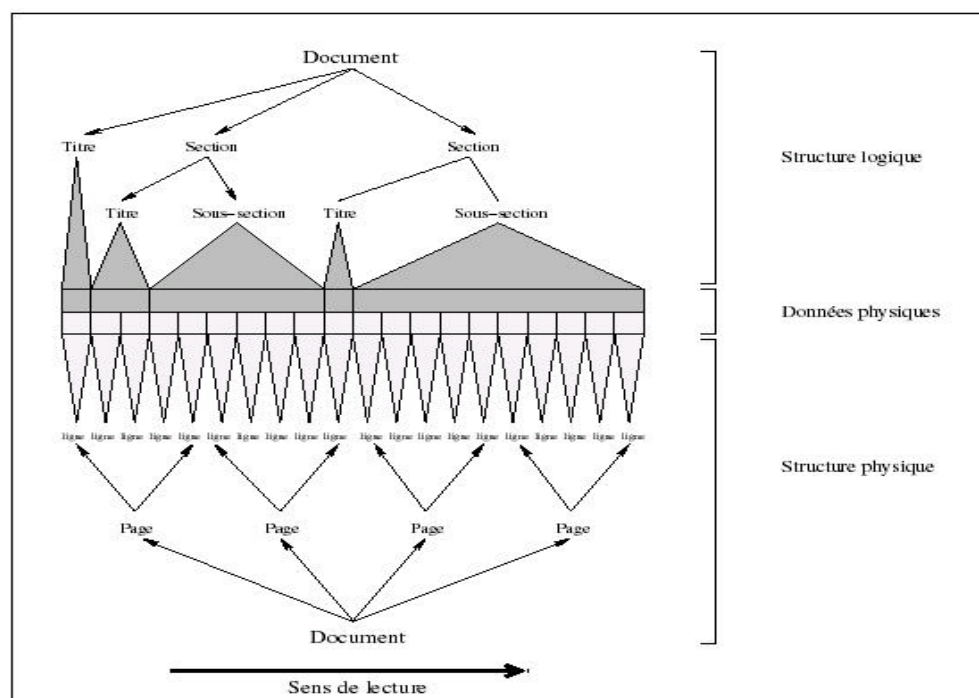


Figure 11: Structure logique et structure physique d'un document [GERY02]

c) Attributs externes

Les attributs externes sont des éléments de description de l'information attachés aux éléments d'un document structuré. Il s'agit de méta-information, c'est-à-dire d'*information à propos de l'information*, comme par exemple le titre, l'auteur ou la date de création d'un document. Ils apportent une information supplémentaire qui ne concerne pas l'information décrite dans le document, mais le document lui-même. Comme la structure logique, les attributs externes représentent une information supplémentaire qui peut être utilisée lors d'une recherche d'information.

d) Sens de lecture

Comme avec un livre, la lecture d'un document structuré consiste à consulter l'introduction et à enchaîner la lecture des chapitres successifs, jusqu'à la conclusion. Dans le cas particulier d'un document structuré "pur", il n'est pas possible de suivre un autre sens de lecture sans nuire à la compréhension de l'ensemble.

Définition : *La lecture d'un document structuré consiste à suivre un chemin de lecture linéaire et unique qui parcourt la totalité du document. Il est imposé par l'auteur, et le lecteur consulte les informations de manière séquentielle.*

2.2. A quoi sert un document structuré ?

Outre la pérennité de l'archivage et la multiplication des modes de restitution du document, les préoccupations propres à l'auteur ou au diffuseur, l'utilisation du document structuré va permettre au **lecteur**, lui-même, une utilisation beaucoup plus riche des données mises en ligne.

A une diffusion de documents, par voie électronique, on associe un signalement électronique, lui aussi structuré en différents champs. Dans le cadre d'une recherche documentaire, l'utilisation d'une structure permettra ainsi de mieux cibler, à *priori*, les interrogations faites sur les corpus de documents.

Imaginons, par exemple, le cas d'une recherche portant sur "Anna Becker" dans un corpus de documents structurés. Les documents que nous trouverons seront très différents d'une interrogation à l'autre.

Recherche de documents	
Requêtes	Réponses
Auteur = Anna Becker	Ensemble de l'œuvre de Becker
Titre = Anna Becker	Ensemble des ouvrages écrits sur Becker
Directeur de Recherche = Anna Becker	Ensemble des travaux de recherches dirigés par Becker
Jury = Anna Becker	Ensemble des travaux de recherches validés par jury auquel participait Becker

- Un bibliothécaire cherchera la liste des oeuvres de Becker,
- Un étudiant en psychologie cherchera l'ensemble des ouvrages traitant de l'oeuvre de Becker,
- Un chercheur voudra, lui, trouver l'ensemble des travaux auxquels Becker a participé en temps qu'encadrant ou de membre de jury pour une thèse, etc.

Une fois le document trouvé, la consultation d'un document structuré est aussi plus souple et plus riche : la structure du document étant codée dans le document, elle est visible, facilitant ainsi la navigation dans le document grâce à des outils tels que les tables des matières arborescentes, listes des tableaux ou des figures, etc.

Considérons maintenant la présence d'un mot dans un document. Un des atouts du document - logiquement - structuré que nous n'avons pas encore mentionné, est son apport en terme de recherche ; le rôle que joue, dans le document, un élément de texte donné peut en effet être une source d'information très pertinente.

Dans un document non structuré, par une simple recherche, on ne peut que savoir si un mot est présent ou non ; dans un document structuré, on peut connaître avec une précision relativement fine le degré d'importance de chaque occurrence du mot dans le texte.

Imaginons par exemple le cas d'une recherche sur Emir Abdelkader, la présence de son nom dans un document peut avoir un sens et une valeur très différents selon la nature de son rôle dans le texte: s'agit-il d'une citation d'Emir Abdelkader? Son nom apparaît-il dans un titre? etc... Ce type d'information est perçu naturellement lors de la lecture, mais dans le cadre d'une recherche, il permet une localisation plus rapide et efficace des différentes zones d'intérêt dans le document.

Recherche dans un document	
Requêtes	Réponses
Emir Abdelkader in <DocAuteur>	Le document consulté a été écrit par Emir Abdelkader
Emir Abdelkader in <DocTitre>	Le titre de l'ouvrage consulté contient l'Emir Abdelkader
Emir Abdelkader in <Citation>	Emir Abdelkader est cité dans cet ouvrage
Emir Abdelkader in <RefBiblio>	Des ouvrages d' Emir Abdelkader sont référencés dans la bibliographie de l'ouvrage

Grâce à l'ensemble des outils liés à la structuration logique, le lecteur va donc pouvoir consulter des documents plus pertinents pour sa recherche par :

- l'accès plus efficace aux documents susceptibles de l'intéresser,
- la recherche et la sélection des parties les plus pertinentes dans le document.

III.2. La recherche d'information dans les documents structurés

L'émergence de documents structurés, nécessite de proposer des modèles de recherche d'information et de navigation capables de fusionner des informations provenant de différents médias.

L'objectif est de permettre à un utilisateur de retrouver, dans une base de documents structurés, les documents ou les parties de documents qui répondent à son besoin d'information. Cette fonctionnalité devient nécessaire dans la mesure où il est courant de trouver des environnements de production de documents qui s'articulent autour de leur structure logique.

Le développement et l'utilisation de normes SGML et XML en est la preuve. Le processus d'interrogation permet d'exprimer ce besoin en utilisant la structure, les descriptions externes (date, auteur, etc.) la nature du média, et le contenu sémantique du document.

Un modèle de documents structurés permet d'accueillir des documents composés de textes et d'images fixes. Les relations, prises en compte, sont la relation de composition (agrégation des parties de documents), la relation de séquence (définissant le sens de lecture d'un document), et la relation de référence (à l'intérieur d'un document ou vers d'autres documents).

Le modèle de document structuré fournit un squelette sur lequel des opérateurs de recherche structurelle sont définis pour le processus d'interrogation.

L'indexation structurelle permet, selon la nature des attributs (attributs de forme ou attributs de contenu), de rendre explicites toutes les informations nécessaires à la recherche d'information. Il s'agit de définir la portée d'un attribut et ses caractéristiques afin de contrôler la propagation de l'information via les relations de structure.

Le processus d'interrogation et de traitement de requêtes utilise les dépendances entre valeurs d'attributs pour retrouver, de manière efficace, le ou les éléments de structure pertinents des documents structurés. Ce travail se situe à l'intersection des SGBD et des SRI.

III.3. XML et la Recherche d'Information [GAUC02]

Un document en texte brut n'est pas agréable à lire, car on ne peut identifier immédiatement les différentes catégories d'information présentes. Lorsque ce document est mis en forme et respecte une mise en page précise et claire, l'œil humain distingue le titre, l'auteur, etc.

Mais cette mise en forme « superficielle » n'est toujours pas exploitable par la machine. **XML** apporte la solution en structurant l'information. Cette structure exprime une sémantique ouvrant la possibilité de recherche très fine.

Au-delà de la recherche plein-texte par exemple : Quels documents contiennent les mots « XML » et « database » ? on peut adjoindre un critère de recherche structurel : Quels documents contiennent « XML » et « database », dans le titre ou le résumé ?

La structure peut également être exploitée afin de procéder à une extraction fine d'information. Par exemple : récupérer le troisième chapitre du document X, ou encore, combiné à une recherche : obtenir les résumés des documents contenant « XML » dans le titre, etc.

XML est un format auto-descriptif, ce qui assure la pérennité des données. Compréhensible par l'homme comme par la machine, il est simple d'utilisation. Mais au-delà de ces aspects, il apporte deux types d'avancées : d'une part, il étend les horizons de la gestion de documents ; et enrichit d'autre part, les possibilités de la gestion de données classique.

La figure suivante représente un corpus de documents structurés XML organisé de manière hiérarchique.

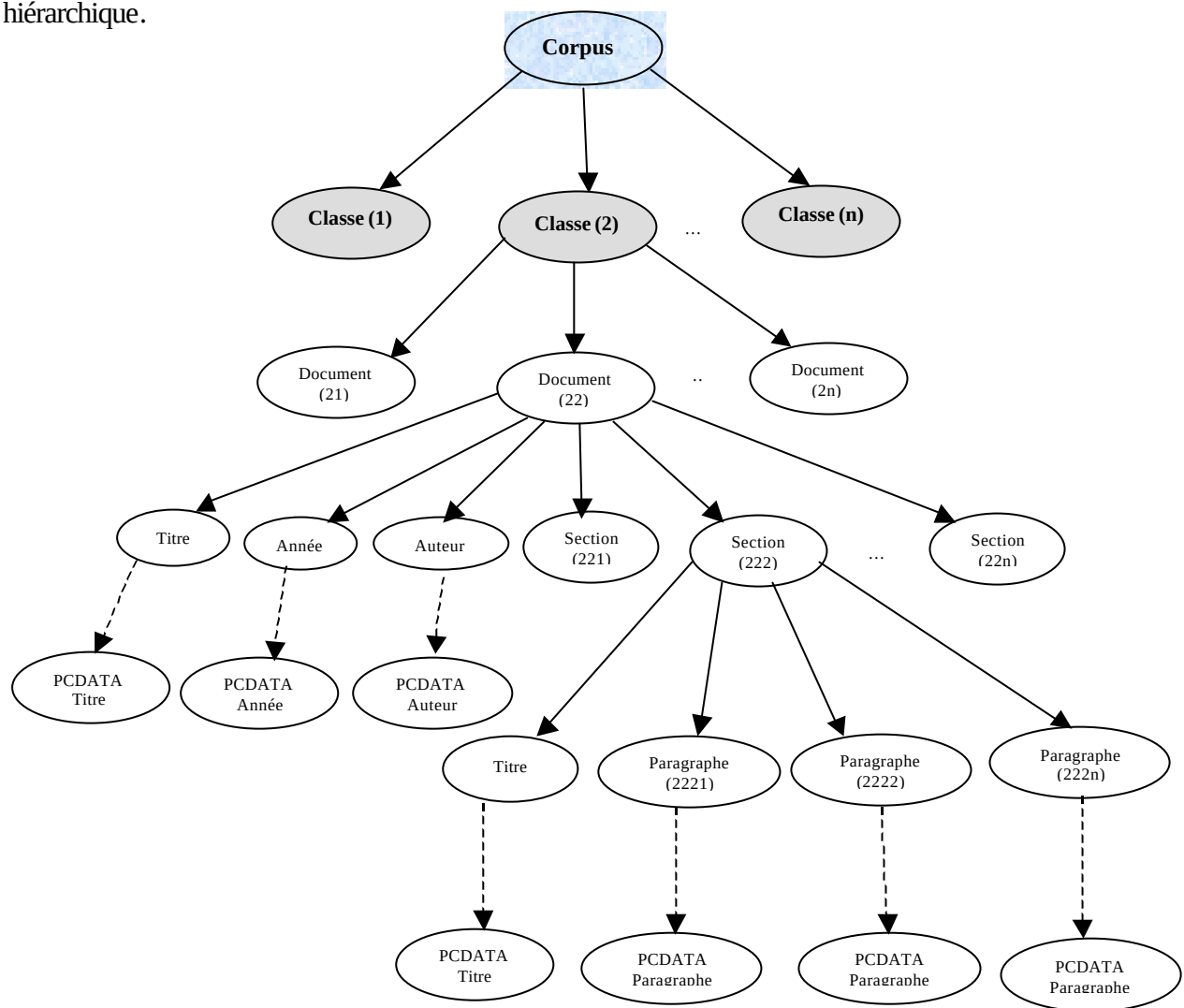


Figure 12: corpus de documents structurés XML

III.3.1. Définition d'un corpus de documents structurés XML

Le corpus de la figure (12) est organisé de manière hiérarchique, à la façon d'un système de fichiers arborescent composé de classes ou catégories et de documents.

Chaque document est défini suivant la vue logique d'un document XML : il s'agit d'une structure hiérarchique où les éléments structurels sont liés. A chaque élément structurel sont associés éventuellement des données textuelles et des attributs.

III.3.2. La structure

Une structure est, tout simplement, un ensemble de nœuds, liés par des arcs, qui définit les différentes relations structurelles.

A partir de ces quelques définitions, nous définissons les relations structurelles (enfant, parent, descendant, etc.).

III.3.3. Les relations structurelles

Nous allons définir un certain nombre de *relations structurelles* permettant de désigner des éléments proches dans la structure. La figure (13), ci-dessus, donne une vue graphique de ces différentes relations. Les relations seront comme suit:

- **pa(e)** désignera le(s) parent(s). Pour toute structure arborescente, en général, et pour tout document XML, en particulier, **pa(e)** est réduit à un élément au plus.
- **anc(e)** désignera le(s) ancêtre(s)
- **enf(e)** désignera le(s) enfant(s)
- **desc(e)** désignera le(s) descendant(s)
- **fr(e)** désignera l'ensemble des frères (c'est à dire, les enfants des parents)
- **chemin (e1; e2)** désignera l'ensemble des chemins possibles entre deux noeuds e1 et e2.

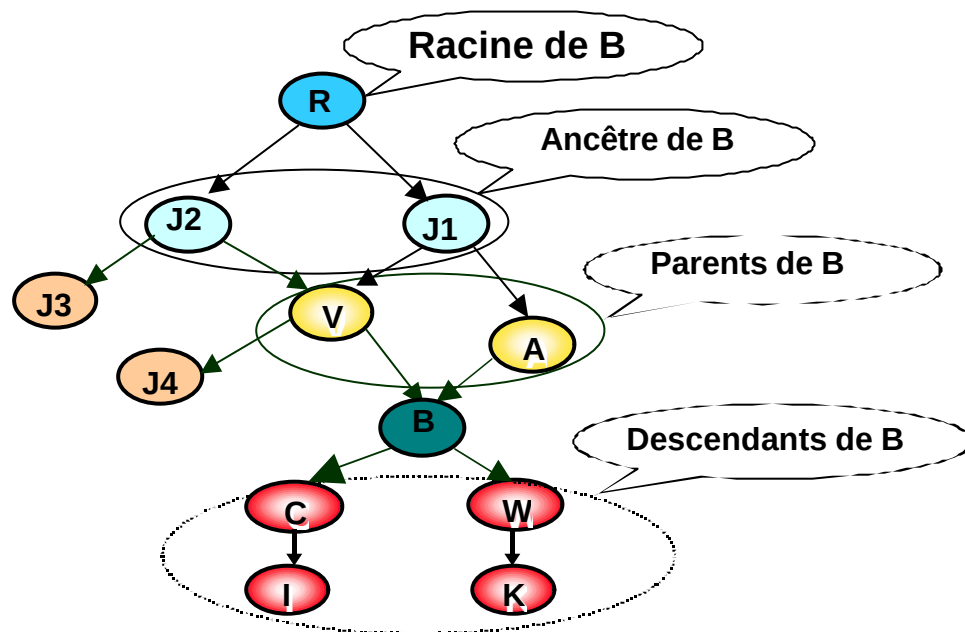


Figure 13: les Relations structurelles (les ensembles correspondent aux différentes relations)

Conclusion

Dans cette première partie, nous avons présenté les principes, les concepts de bases et les différentes approches utilisés dans les Systèmes de Recherche d'Informations. Nous avons montré que dans ces approches, les termes sont considérés comme indépendants.

Le travail, que nous avons présenté, évoque les limites des systèmes de recherche d'information traditionnels pour gérer de manière pertinente les données structurées et montre la nouvelle dimension que prend la recherche d'information, en conjuguant les termes de recherche avec leur emplacement dans la structure logique du document (chapitres, sections, paragraphes, citations, etc.).

Nous avons montré que la connaissance de la structure des documents est une ressource additionnelle qui améliore la pertinence du système de recherche d'information par un accès plus efficace aux documents susceptibles d'intéresser l'utilisateur.

Pour réaliser ces objectifs de dynamisme des représentations des documents et de dynamisme des dépendances entre termes, les réseaux bayésiens nous sont parus appropriés.

Avant de décrire plus en détail les applications possibles de la théorie des réseaux bayésiens en RI, nous allons présenter dans la deuxième partie de manière détaillée les réseaux bayésiens, leur fonctionnement et leurs intérêts dans la recherche d'information.

Introduction

Les modèles graphiques sont le mariage entre la théorie des probabilités et celles des graphes. Ils fournissent des outils intuitifs et naturels pour traiter des problèmes dans lesquelles l'incertitude et la complexité des données joue un rôle important. L'idée fondamentale des modèles graphiques est la modularité: un système complexe est construit en combinant des parties plus simples. La théorie des probabilités combine alors ces parties assurant une cohérence à l'ensemble du système. La théorie des graphes apporte une interface intuitive pour la modélisation des connaissances, permettant de structurer facilement les données [POPU00].

Les réseaux bayésiens font partie des modèles graphiques. Ils combinent une représentation de la connaissance sous forme graphique (des relations de dépendance directe) et probabiliste (une incertitude sur la connaissance).

Les Réseaux Bayésiens sont les successeurs et héritiers naturels des approches symboliques, connexionnistes et statistiques de l'Intelligence Artificielle et du Data Mining. Ils allient la rigueur d'un formalisme mathématique puissant et stable à l'efficacité d'une représentation « distribuée » de connaissances et à la lisibilité des modèles à base de règles. Particulièrement adaptés à la prise en compte de l'incertitude, ils peuvent aussi bien être décrits manuellement, par les experts du domaine, que généré automatiquement par apprentissage.

Reconnus par les géants de l'informatique comme une des conditions de leur succès dans les prochaines décennies, les Réseaux Bayésiens constituent le socle incontournable des solutions décisionnelles capables de relever les défis actuels [2].

Un Réseau Bayésien est une représentation graphique des dépendances entre des variables aléatoires, auxquelles on associe les distributions de probabilités de ces variables aléatoires.

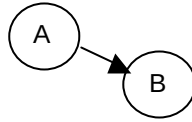
Dans cette partie, nous présentons, de façon intuitive, les fondements des réseaux bayésiens. Nous les situons par rapport aux autres techniques de modélisation, nous décrivons leurs applications potentielles, et les principaux algorithmes de l'inférence et apprentissage.

I. Fondations

I.1. Une représentation graphique de la causalité

I.1.1. Introduction [BECK99]

La représentation la plus commune et la plus intuitive de l'influence d'un événement, d'un fait, ou d'une variable sur une autre, est probablement de relier la cause à l'effet par une flèche orientée.



Supposons que **A** et **B** sont des faits, qui peuvent être vrais ou faux.

Du point de vue du sens commun, ce graphe peut se lire comme “la connaissance que j’ai de **A** détermine la connaissance que j’ai de **B**”.

Cette détermination peut être stricte, c’est-à-dire que, sachant avec certitude que **A** est vrai, je peux en déduire **B** avec certitude. Il peut aussi s’agir d’une simple influence. Dans ce cas, cela signifie que si je connais **A** avec certitude, *mon opinion* sur **B** est modifiée, sans que je puisse toutefois affirmer si **B** est vrai ou faux.

Avant d’aller plus loin, il est important de comprendre que, bien que la flèche soit orientée de **A** vers **B**, elle peut cependant fonctionner dans les deux sens, et ce, même si la relation causale est stricte.

Supposons par exemple que la relation causale soit **A** $\rightarrow$ **B**. Cela signifie que si **A** est vrai, **B** l’est également. Si **A** est faux, **B** peut être vrai ou faux.

A	B
V	V
F	V
F	F

La table ci-dessus nous permet d’affirmer que si **B** est faux, **A** l’est également.

Du point de vue de la logique, il s’agit simplement de la contraposée de **A** $\rightarrow$ **B**. Du point de vue de la causalité, cela montre qu’une relation causale, même orientée, est inversible, de l’effet vers la cause, même si elle ne l’est que partiellement.

En d’autres termes :

*S’il existe une relation causale de **A** vers **B**, toute information sur **A** peut modifier la connaissance que j’ai de **B**, et, réciproquement, toute information sur **B** peut modifier la connaissance que j’ai de **A**.*

En présence d’un graphe plus complexe, il est donc essentiel de conserver à l’esprit que l’information ne circule pas seulement dans le sens des flèches.

I.1.2. Circulation de l'information dans un graphe causal

Nous allons à présent étudier, de plus près, comment l'information circule au sein d'un graphe causal.

Pour l'instant, nous continuons de baser cette discussion sur une notion très intuitive de ce qu'est un graphe causal : il s'agit simplement de relier des "causes" et des "effets" par des flèches orientées.

I.1.2.1. Exemple [BECK99]

Pour cela, nous allons utiliser un exemple, extrêmement classique dans la littérature sur les réseaux bayésiens :

Ce matin-là, alors que le temps est clair et sec, M. Holmes sort de sa maison, et s'aperçoit que la pelouse de son jardin est humide. Il se demande alors s'il a plu pendant la nuit, ou s'il a simplement oublié de débrancher son arroseur automatique. Il jette alors un coup d'oeil à la pelouse de son voisin, M. Watson, et s'aperçoit qu'elle est également humide. Il en déduit alors qu'il a probablement plu, et décide de partir au travail sans vérifier son arroseur automatique [BECK99].

La représentation graphique du modèle causal utilisé par M. Holmes est la suivante:

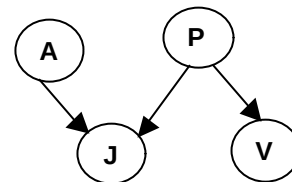
Avec les notations suivantes :

A : j'ai oublié de débrancher mon arroseur automatique

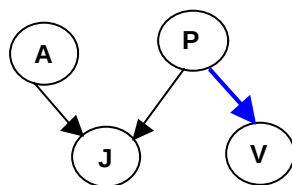
P : il a plu pendant cette nuit

J : l'herbe de mon jardin est humide

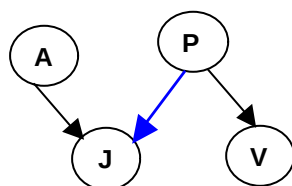
V : l'herbe du jardin de mon voisin est humide



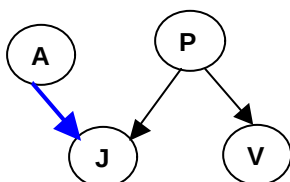
Les causalités représentées sont bien conformes à l'intuition :



S'il a plu pendant la nuit, l'herbe du jardin de mon voisin est également humide.



S'il a plu pendant la nuit, l'herbe de mon jardin est humide.



Si j'ai oublié de débrancher mon arroseur automatique, l'herbe de mon jardin est humide.

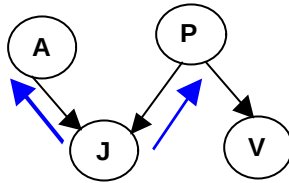
Comment ce graphe est-il utilisé ici pour raisonner ? Autrement dit, comment l'information **J**, dont on sait qu'elle est vraie, est-elle utilisée ?

Tout d'abord, le modèle nous indique que **J** a du être causé soit par **A**, soit par **P**. Faute d'information complémentaire, les deux causes sont a priori également plausibles.

Le fait que **V** soit également vrai renforce la croyance en **P**, quoique le voisin pourrait lui aussi avoir oublié de débrancher son arroseur automatique.

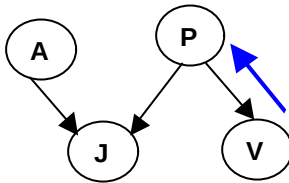
Dans cet exemple simple, on voit que l'information a circulé uniquement dans le sens **Effet**→**Cause**.

(1)



La connaissance de J renforce la croyance en l'une des deux causes A, ou P.

(2)



La connaissance de V augmente la croyance en la cause P. Par rapport à l'état (1) précédent, la cause A devient moins

Conclusion

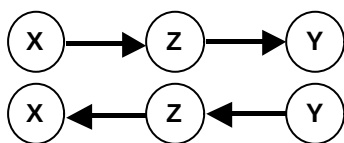
Pour prendre un raccourci, M. Holmes a déduit que son arroseur automatique était à l'arrêt à partir du fait que la pelouse de son voisin était humide !

Cet exemple simple, sur lequel nous n'avons utilisé que du raisonnement de sens commun, nous montre bien que l'information peut suivre des chemins à première vue lorsqu'elle se propage dans un réseau de causalité.

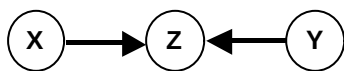
I.1.2.2. Le cas général

Nous allons maintenant étudier la circulation de l'information dans un graphe causal du point de vue général. Dans l'exemple ci-dessus, nous avons vu qu'une information *certaine* se propage dans un graphe en modifiant les croyances que nous avons de certains faits.

Nous allons étudier quels chemins cette information peut prendre à l'intérieur du graphe. En considérant les 3 cas suivants.



Connexion en Série, X cause Z, Z cause Y
(ou le cas symétrique)



Connexion convergente, X et Y causent Z



Connexion divergente, Z cause X et Y

Pour chacun de ces cas, nous allons essayer de déterminer si l'information peut circuler de **X** à **Y**, en considérant un petit exemple connus.

<i>Graphe</i>	<i>Propriété</i>	<i>Exemple</i>
Connexion Convergente  <pre> graph LR X((X)) --> Z((Z)) Y((Y)) --> Z((Z)) </pre>	L'information ne peut circuler de X à Y que si Z est connu.	X = tremblement de terre Y = voleur Z = alarme Le fait qu'il y ait eu un tremblement de terre dans le voisinage n'a aucun lien a priori avec que le fait que ma maison ait été cambriolée. En revanche, si mon alarme s'est déclenchée, j'ai tendance à croire que je viens d'être cambriolé. Si maintenant j'apprends qu'il vient d'y avoir un tremblement de terre dans le voisinage, je m'apprends sur l'éventualité d'un cambriolage.
Connexion en Série  <pre> graph LR X((X)) --> Z((Z)) Z((Z)) --> Y((Y)) </pre>	L'information ne peut circuler de X à Y que si Z n'est pas connu.	X = ensoleillement Y = récolte Z = prix du blé Si la saison a été ensoleillée, la récolte sera abondante. Si la récolte est abondante, le prix du blé est bas. Si je sais déjà que la récolte a été abondante, le fait de connaître l'ensoleillement ne m'apprend plus rien sur le prix du blé.
Connexion Divergente  <pre> graph LR Z((Z)) --> X((X)) Z((Z)) --> Y((Y)) </pre>	L'information ne peut circuler de X à Y que si Z n'est pas connu.	X = la pelouse de mon jardin est humide Y = la pelouse de mon voisin est humide Z = il a plu cette nuit Si la pelouse de mon jardin est humide (X), j'ai tendance à croire qu'il a plu cette nuit (Z), et donc que la pelouse de mon voisin sera aussi humide (Y). Si en revanche je sais qu'il a plu cette nuit (Z), je peux affirmer que la pelouse du jardin de mon voisin sera humide (Y), et l'information que je peux avoir sur l'état de ma propre pelouse (X) n'y change rien.

I.1.3. D-séparation (blocage)

Nous savons maintenant exactement dans quelles conditions une information peut circuler à l'intérieur d'un graphe. On voit qu'il ne s'agit pas de suivre le sens des flèches !

Supposons que nous disposions d'un graphe relativement complexe, pour lequel nous disposons déjà d'un certain nombre d'informations (i.e., certaines variables sont déjà connues).

Si nous apprenons maintenant une autre information, devons-nous réviser notre opinion sur l'ensemble des autres noeuds de ce graphe ?

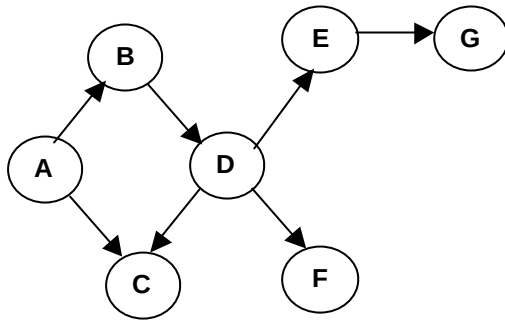
Pour répondre à cette question, nous pouvons essayer de synthétiser l'étude des circuits d'informations dans la figure n°1 ci-dessus en une règle appelée d-séparation, qui décrit dans quelles conditions l'information entre un noeud X et un noeud Y est bloquée.

Règle : On dira que X et Y sont d-séparés (bloqués) par Z si pour tous les chemins entre X et Y , l'une au moins des deux conditions suivantes est vérifiée :

- Le chemin converge en un noeud W , tel que $W \rightarrow Z$, et W n'est pas une cause directe de Z .
- Le chemin passe par Z , et est soit divergent, soit en série au noeud Z .

X est d-séparé de Y par Z est noté par : $\hat{A}X|Z|Y\hat{n}$.

Exemple :



(A|B|D):

Le chemin A-B-D est en série en B (B)

Le chemin A-C-D est convergent en C (C)

(A|D|E):

Tous les chemins de A à E passent par D.

Le chemin A-B-D-E est en série en D (B D E)

Le chemin A-C-D-E est divergent en D (C D E)

Figure 14: Circulation de l'information dans un graphe

Supposons que Z soit la seule information connue dans le graphe.

Si X et Y sont d-séparés par Z , que se passe-t-il ?

Considérons un chemin entre X et Y .

Soit ce chemin converge en un point W ($W \rightarrow Z$), tel que $W \rightarrow Z$, et W n'est pas une cause directe de Z . Comme par hypothèse Z est la seule information connue dans le graphe, aucune information n'est disponible sur W . Ce chemin est donc bloqué.

Sinon, ce chemin passe par Z , et on a soit $X \rightarrow Z$, soit $Z \rightarrow Y$. D'après la règle de d-séparation, comme Z est connu, l'information ne peut circuler à travers Z . Tous ces chemins sont donc bloqués.

Donc si X et Y sont d-séparés par Z , et si Z est la seule information connue dans le graphe, une nouvelle information sur X ne modifie en rien mon opinion sur Y .

En résumé : Un algorithme simple a été proposé pour déterminer les d-séparations, le "Bayes Ball" (Figure15). Le principe est simple ; sur la première ligne figurent les variables dont la valeur n'est pas connue, sur la seconde celles dont la valeur est connue [PIWO03].

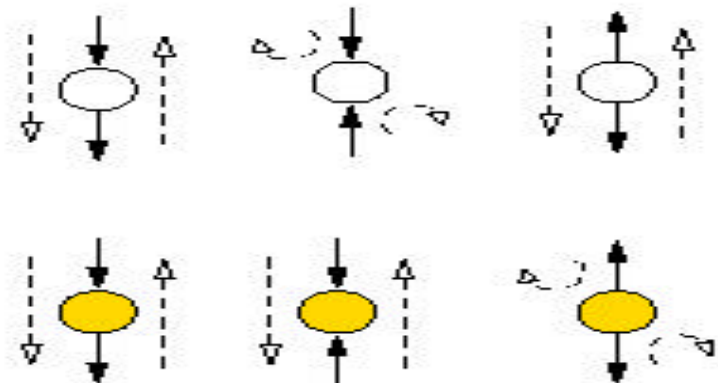


Figure 15: "Bayes ball" et d-séparation. Les variables observées ont un fond coloré [PIWO03]

Pour savoir si deux variables sont d-séparées dans ce formalisme, il suffit d'imaginer le lancement d'une balle à travers tous les chemins possibles entre ces deux variables. Si la balle arrive à un noeud du type convergent pour une variable dont la valeur n'est pas connue (haut milieu) ou du type divergent pour une variable dont la valeur est connue (bas droite), alors elle rebondit. Dans tous les autres cas, elle passe. En utilisant ces règles, les deux variables sont d-séparées si et seulement si il n'existe aucun chemin valide pour la balle [PIWO03].

Conclusion

A partir d'une représentation graphique de la causalité, nous avons déduit que :

- (1) l'information ne circule pas uniquement dans le sens *Cause* *Effet*,
- (2) la circulation de l'information suit cependant des règles bien précises,
- (3) une combinaison de ces règles permet de restreindre l'impact d'une information à l'intérieur du graphe.

I.2. Une représentation probabiliste associée.

I.2.1. Définition des variables

Si le graphe causal initial contient les noeuds $\{A, B\}$ pouvant prendre chacun la valeur Vrai ou Faux, on définit l'espace probabilisé E constitué des couples suivants :

$E = \{(A=V, B=V), (A=V, B=F), (A=F, B=V), (A=F, B=F)\}$. Chaque couple est appelé un évènement.

La variable A est alors une variable aléatoire sur E , définie de la façon suivante :

$$A((A=V, B=V)) = 1$$

$$A((A=V, B=F)) = 1$$

$$A((A=F, B=V)) = 0$$

$$A((A=F, B=F)) = 0$$

L'ensemble $\{(A=V, B=V), (A=V, B=F)\}$, qui est l'image réciproque de 1 par l'application A est noté simplement $A=V$. La variable aléatoire B est définie de façon similaire.

I.2.2. Définition des probabilités

Pour compléter la transposition d'un graphe causal en espace probabilisé, nous devons également fournir:

- ✓ Si A n'a aucune cause directe nous devons définir $P(A)$, c'est-à-dire les deux nombres $P(A=V)$ et $P(A=F)$.
- ✓ Si B a une seule cause directe A , nous devons définir $P(B/A)$, c'est-à-dire les quatre nombres : $p(A=V/B=V)$, $p(A=V/B=F)$, $p(A=F/B=V)$, $p(A=F/B=F)$.
- ✓ Si C a deux causes directes A et B nous devons définir $P(C/A, B)$, c'est-à-dire les huit nombres: $p(C=V/A=V, B=V)$, $p(C=V/A=V, B=F)$, $p(C=V/A=F, B=V)$, ...etc.

I.3. Réseau bayésien (RB)

Les réseaux bayésiens font partie de la famille des modèles graphiques. Ils regroupent au sein d'un même formalisme la théorie des graphes et celle des probabilités afin de fournir des outils autant efficaces qu'intuitifs pour représenter une distribution de probabilités jointe sur un ensemble de variables aléatoires. Ce formalisme, très puissant, permet une représentation intuitive de la connaissance sur un domaine d'application donné et facilite la mise en place de modèles performants et clairs. La représentation de la connaissance se base sur la description, par des graphes, des relations de causalité existant entre des variables décrivant le domaine d'étude. A chaque variable est associée une distribution de probabilité locale quantifiant la relation causale [BELL02].

I.3.1. Histoire des réseaux bayésiens

A la fin des années 1970, les experts en Intelligence Artificielle (IA) ont développé des outils qui utilisaient des systèmes à base de règles, du type si-alors. Mais ces systèmes mettaient du temps à être réalisés, et de plus, il arrivait souvent qu'on ne puisse répondre de manière claire aux questions posées par ce même système.

En effet, ces règles attendaient des réponses du type oui/non, ou alors une valeur spécifique à une question particulière. Mais dans le monde réel, les réponses aux questions ne sont souvent pas aussi catégoriques.

Puis sont apparus les réseaux de neurones. Cet outil permettait, à l'aide d'un grand nombre de données, d'effectuer des classifications. Par exemple, en introduisant les caractéristiques d'individus et en indiquant à chaque fois le sexe des individus, le système nous donne le sexe de l'individu correspondant à des caractéristiques données. Ces techniques ont continué à être utilisées, elles fonctionnèrent très bien dans de nombreuses applications, telles que la vérification de transactions bancaires par cartes bleues.

Mais, à la fin des années 1980, grâce aux travaux de Judea Pearl et ceux d'une équipe de chercheurs danois de l'université d'Aalborg, les experts en intelligence artificielle découvrirent, parallèlement la logique non monotone et possibiliste, que les réseaux bayésiens offraient un moyen assez efficace de contourner l'obstacle qu'était le manque ou l'ambiguïté de certaines informations [POPU00].

I.3.2. Origine des réseaux bayésiens

Les Réseaux Bayésiens sont les successeurs et héritiers naturels des approches symboliques, connexionnistes et statistiques de l'Intelligence Artificielle et du Data Mining. Ils allient la rigueur d'un formalisme mathématique puissant et stable à l'efficacité d'une représentation « distribuée » de connaissances et à la lisibilité des modèles à base de règles.

Particulièrement adaptés à la prise en compte de l'incertitude, ils peuvent aussi bien être décrits manuellement par les experts du domaine et généré automatiquement par apprentissage.

Le nom **réseau bayésien** provient du révérend Bayes et plus précisément de son fameux Théorème, mais il n'y a pas toujours de rapprochement entre un modèle de type "réseau bayésien" et la "statistique bayésiennes". Le terme **diagramme d'influence**, plus général et dénué de connotation est donc parfois préféré à réseau bayésien.

Les réseaux bayésiens sont aussi appelés réseaux de croyances, réseaux probabilistes, réseaux causals encore carte de connaissance.

I.3.3. Définitions

- Les réseaux bayésiens constituent une **technique d'acquisition, de représentation et de manipulation de connaissances** [BECK99].
- Les réseaux bayésiens font partie des modèles graphiques. Ils combinent une représentation de la connaissance sous forme graphique (des relations de dépendance directe) et probabiliste (une incertitude sur la connaissance) [BUCH02].
- Les RB sont des graphes acycliques orientés. Les noeuds de ces graphes représentent des variables aléatoires qui peuvent avoir un nombre discret d'états possibles ou qui peuvent être continues selon la loi de Gauss. Les arcs orientés du réseau représentent des relations de causalité et donc de dépendance entre les variables associées aux noeuds ainsi reliés [BOUC02].
- Un réseau bayésien (ou probabiliste ou de croyance) est un graphe dont les noeuds définissent les variables du système et dont les arcs définissent l'existence de relations entre variables. Un réseau bayésien est en quelque sorte un réseau causal, auquel on ajoute des éléments de probabilités issus de la théorie bayésienne... Ce qui est important avec les réseaux bayésiens, c'est le fait que les relations causales ne sont pas absolues, mais sont associées à une probabilité indiquant le degré de croyance que l'on a dans l'évènement [POPU00].

La figure 16, ci-après, est l'illustration des réseaux Bayésiens. Les noeuds X_i sont associés à des variables aléatoires définies par des observations (un segment, une partie d'un objet ou plus généralement une primitive) et les noeuds H_i représentent des hypothèses (un objet, l'orientation ou la position d'un objet).

L'utilisation de tels réseaux nécessite la spécification complète du modèle probabiliste. Pour chaque relation de dépendance entre plusieurs noeuds, il faut, en effet, définir les tables de probabilités conditionnelles (TPC).

Sur la Figure16, suivante, les valeurs p , q , r et s désignent des probabilités: les probabilités d'avoir un état pour la variable H_1 en fonction des états des noeuds parents.

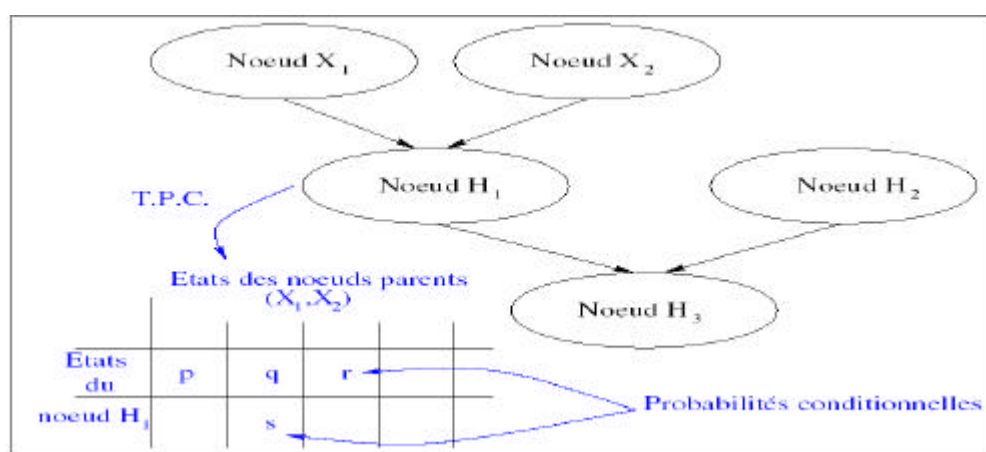


Figure 16: Exemple de réseaux bayésiens [BELL02]

Considérons maintenant un noeud associé à la variable X qui a pour unique enfant un noeud associé à la variable H . Il faut alors définir pour chaque état de la variable H , la probabilité à priori d'avoir un état de H connaissant l'état de X ce qui revient à préciser la distribution de H conditionnée par X . D'autre part, il faut préciser les dépendances avec un parent. C'est l'instanciation du réseau.

Cette connaissance a priori est ensuite utilisée pour la propagation des informations selon le mécanisme d'inférence défini par la loi de **Bayes** [USTL04].

$$P(H / X) = \frac{P(H, X)}{P(X)} = \frac{P(X / H)P(H)}{P(X)}$$

Avec

$$P(X) = P(X / H)P(H) + P(X / \neg H)P(\neg H)$$

- **$P(H/X)$** représente la probabilité à posteriori d'un état de l'hypothèse **H** connaissant l'état de l'observation **X** .
- **$P(X=H)$** représente la vraisemblance d'avoir un état pour l'observation **X** sachant l'état de l'hypothèse **H** ,
- **$P(H)$** et **$P(X)$** représentent les probabilités, à priori, d'avoir un état pour l'hypothèse et une certaine observation **X** .

Le fait d'avoir de nombreux parents et de nombreux états possibles risque de faire exploser la combinatoire dans les TPCs: si chaque parent **X_i** a **n_i** états possibles et **H** à **N** états, alors le nombre de valeurs à spécifier est égal au produit **X_i** .

$$nb \text{ probabilité} = N \prod_{n_{iparent}} X_i$$

I.3.4. Fonctionnement des réseaux bayésiens [BALM02]

L'étude d'un modèle de réseau bayésien nécessite, comme tout traitement statistique, une base de données et cherche à fournir à cette base une modélisation sous forme de graphe caractérisant les dépendances conditionnelles des différentes variables. Elle se déroule en deux phases :

- la phase de construction du réseau qui peut solliciter experts et techniques d'apprentissage ;
- la phase d'utilisation qui fait appel à une capacité très intéressante des réseaux bayésiens, l'inférence.

I.3.4.1. Phase de construction

Les réseaux bayésiens permettent de combiner la connaissance d'experts avec la connaissance extraite à partir de données.

Les experts peuvent, par exemple, déterminer les dépendances entre les variables; alors qu'un apprentissage automatique permettra de déterminer la distribution de probabilité associée à chaque variable.

Il ne s'agit ici que d'un exemple car il est tout à fait envisageable que des experts définissent entièrement un réseau bayésien, graphe et distribution de probabilité. Au contraire, ces deux éléments peuvent être construits automatiquement par apprentissage du système [BALM02].

I.3.4.2. Phase d'utilisation

L'utilisation des réseaux bayésiens repose sur la propagation de l'information au sein du réseau, c'est à dire des calculs de probabilités, c'est ce que l'on nomme l'inférence. Après avoir fait une observation sur une variable comment cette information va t-elle se représenter sur l'état des autres variables ? [BALM02]

I.3.5. Propriétés des Réseaux Bayésiens

Un réseau bayésien est donc un graphe causal auquel on a associé une représentation probabiliste sous-jacente.

Comme nous l'avons vu, cette représentation permet de rendre quantitatifs les raisonnements sur les causalités que l'on peut faire à l'intérieur du graphe.

Nous avons également évoqué, très rapidement, le lien entre d-séparation et indépendance. En réalité, un résultat très important existe qui affirme que "si X et Y sont d-séparés par Z alors X et Y sont indépendants sachant Z ." :

$$\langle X/Z/Y \rangle \Rightarrow P(X/Y, Z) = P(X/Z)$$

Ce résultat est très important car il permet de limiter les calculs de probabilités, grâce à des propriétés du graphe. En effet, supposons que X et Y soient d-séparés par Z , et que Z soit connu. Supposons, par ailleurs, que je vient de calculer $P(X/Z)$. Si une nouvelle information sur Y est alors connue, ce résultat me permet de conserver mon calcul de $P(X/Z)$ comme valeur de $p(X/Z, Y)$. Autrement dit, le résultat sur la d-séparation et le blocage d'informations que nous avons décrit intuitivement sur les graphes de causalité est valable également dans la représentation quantitative probabiliste sous-jacente !

Combiné avec un autre résultat, qui établit qu'un noeud est d-séparé du reste du graphe par l'ensemble constitué de ses parents, de ses enfants, et des autres parents de ses enfants, cette propriété permet de rendre locaux tous les calculs de probabilités dans un graphe causal [4].

I.3.6. Utilisation et difficultés des Réseaux Bayésiens

L'utilisation essentielle des réseaux bayésiens est donc de calculer des probabilités conditionnelles d'événements, reliés les uns aux autres par des relations de cause à effet. Cette utilisation s'appelle *inférence*.

La correspondance qui existe entre la structure graphique et la structure probabiliste associée va permettre de ramener l'ensemble des problèmes de l'inférence à des problèmes de théorie des graphes. Cependant, ces problèmes restent relativement complexes et constituent des sujets de recherche très actifs. L'autre difficulté essentielle des réseaux bayésiens se situe précisément dans l'opération de transposition du graphe causal à une représentation probabiliste. Même si les seules tables de probabilités nécessaires pour définir entièrement la distribution de probabilité sont celles d'un noeud conditionné par rapport à ses parents, il reste que la définition de ces tables n'est pas toujours facile pour un expert [4].

Nous allons donc maintenant aborder ces deux problèmes du point de vue technique.

II. Techniques & Algorithmes

II.1. Inférence [LERa00], [LERb00], [LERc00], [BECK99]

Tout calcul portant sur la distribution de probabilité associée à un réseau bayésien relève de l'inférence.

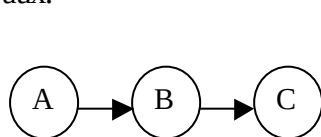
D'un point de vue intuitif, l'inférence, dans un réseau de causalités, consiste à propager une ou plusieurs informations certaines, au sein de ce réseau, pour en déduire comment les croyances, concernant les autres nœuds, sont modifiées.

Supposons que nous disposions d'un réseau bayésien défini par un graphe et la distribution de probabilités associée (G, P) . Supposons que le graphe soit constitué de n nœuds, notés $X = \{X_1, X_2, X_3, \dots, X_n\}$. Le problème général de l'inférence est de calculer $P(X_i/Y)$, où $Y \perp X$, $X_i \perp Y$.

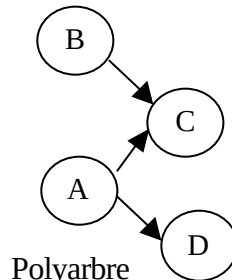
On voit bien que la complexité de ce problème dépend de la structure du réseau. Nous allons tout d'abord étudier le problème de l'inférence de façon empirique, en montrant que la méthode "intuitive" qui consiste à propager l'information le long des arcs, conduit à des conclusions erronées dans le cas général.

II.1.1 Approche intuitive

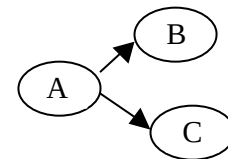
Supposons que nous disposions d'un réseau bayésien (G, P) , par exemple l'un des quatre réseaux présentés ci-dessous, où toutes les variables sont binaires, et peuvent prendre les valeurs Vrai, ou Faux.



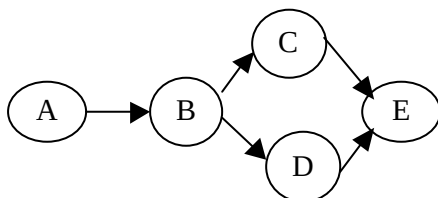
Chaîne $P(C/A)$?



Polyarbre
 $P(D/B)$?



Arbre $P(C/B)$?



Réseau avec boucles $P(E/A)$?

Par exemple, dans le cas du réseau en forme de chaîne, supposons que nous disposions d'une information sur le nœud A , par exemple que $A = \text{Vrai}$. Comment propager cette information dans le réseau, i.e., comment calculer $P(C/A = \text{Vrai})$?

D'après la structure de ce graphe, nous savons que (définition d'un réseau bayésien) :

$$P(A, B, C) = P(C/B) \cdot P(B/A) \cdot P(A)$$

Comme de plus (définition de la probabilité conditionnelle) :

$$P(A, B, C) = P(C/A, B) \cdot P(B/A) \cdot P(A),$$

On a: $P(C/A, B) = P(C/B)$.

De plus (Théorème de Bayes) : $P(C/A) = \sum P(C/A, B) \cdot P(B/A)$.

Et donc: $P(C/A) = \sum P(C/B) \cdot P(B/A)$.

Finalement: $P(C=V/A=V)=P(C=V/B=V).P(B=V/A=V)+P(C=V/B=F).P(B=F/A=V).$

Et de même: $P(C=F/A=V)=P(C=F/B=V).P(B=V/A=V)+P(C=F/B=F).P(B=F/A=V).$

On voit donc que ce calcul se résume à calculer de proche en proche la probabilité de chaque noeud, en propageant les probabilités conditionnelles connues.

II.1.1.1 Chaînes

Considérons une chaîne de longueur n , et calculons $P(X_i/X_j)$

Si le noeud X_i est situé en aval du noeud X_j , mais n'est pas le descendant direct de X_j ($j < i-1$),



On peut écrire :

$$P(X_i / X_j) = \sum_{X_{i-1}} P(X_i / X_{i-1}).P(X_{i-1} / X_j)$$

Si le noeud X_{i-1} est un descendant direct de X_j , on a terminé, sinon, il suffit de décomposer $P(X_{i-1}/X_j)$ de la même façon, jusqu'à arriver au descendant direct de X_j .

Dans le cas où le noeud X_i est situé en amont du noeud X_j , c'est un peu plus compliqué.



Il faut d'abord utiliser la propagation avant à partir du début de la chaîne, pour connaître pour chaque noeud sa probabilité inconditionnelle $P(X_k)$ pour $(1 \leq k \leq j)$.

On peut utiliser la propriété d'inversion de la probabilité conditionnelle :

$$P(X_i / X_{i+1}) = \frac{P(X_{i+1} / X_i).P(X_i)}{P(X_{i+1})}$$

De même, si X_i est l'ascendant direct de X_j , on a alors terminé, sinon, il suffit également de continuer de proche en proche.

II.1.1.2 Arbres

Le cas d'un arbre se traite de la même façon que les chaînes, par exemple en considérant qu'un noeud situé à un point de jonction peut être doublé, pour obtenir deux chaînes.



II.1.1.3 Poly-arbres

On appelle poly-arbre un réseau sans boucle. Dans la pratique, cela signifie que chaque noeud peut avoir plusieurs parents.

La propagation de l'information dans un poly-arbre est plus complexe, car l'information peut circuler d'un parent à un autre.

Cependant, le cas des poly-arbres peut se traiter de la même façon que arbre, c'est-à-dire en utilisant une propagation locale.

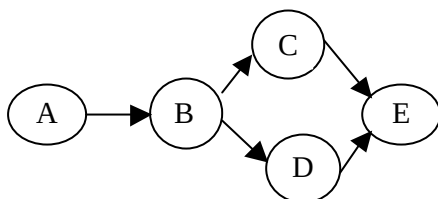
II.1.2 Cas général

Essayons à présent d'appliquer un raisonnement local à un graphe présentant des boucles.

Supposons que nous cherchions à représenter, dans un réseau bayésien, les règles logiques suivantes:

$$A \Rightarrow B, B \Leftrightarrow C, B \Leftrightarrow \neg D, E = \text{XOR}(C, D)$$

Cette connaissance peut être représentée par le graphe suivant :



Nous devons également utiliser les tables de probabilité suivantes :

	A = V	A = F
B = V	1	1/2
B = F	0	1/2

	B = V	B = F
C = V	1	0
C = F	0	1

	B = V	B = F
D = V	0	1
D = F	1	0

	C = V		C = F	
	D = V	D = F	D = V	D = F
E = V	0	1	1	0
E = F	1	0	0	1

Supposons que **A** soit faux, et essayons de conclure sur **E**.

Essayons d'abord le raisonnement basé sur la logique. Comme **A** est faux, je ne peux pas utiliser la règle **A** $\Rightarrow$ **B**, et donc je ne peux rien dire sur **B**.

Toutefois **B** est certainement, soit vrai, soit faux.

Supposons que **B** soit vrai. Dans ce cas, **C** est vrai, et **D** est faux, et **E** est donc vrai.

Supposons que **B** soit faux. Dans ce cas, **C** est faux, et **D** est vrai, et **E** est donc vrai.

Donc si **A** est faux, **E** est vrai.

Essayons maintenant la propagation "locale" des probabilités.

Comme **A** est faux, la probabilité que **B** soit vrai (respectivement faux) est **1/2**.

Donc la probabilité que **C** soit vrai est également **1/2**, et de même pour **D**.

Finalement, on conclut que la probabilité que **E** soit vrai est également de **1/2**.

Dans le cas d'un réseau comportant des boucles, la propagation locale des probabilités ne fonctionne pas.

II.1.2.1 Conditionnement

Essayons de comprendre intuitivement pourquoi le raisonnement probabiliste que nous avons fait ci-dessus ne fonctionne pas. En propageant l'information de **B** à **C**, puis de **B** à **D**, nous avons fait comme si **C** et **D** étaient indépendants. Or, nous savons que dans une connexion divergente, **C** et **D** ne sont indépendants que si **B** est connu (si **B** n'est pas connu, l'information peut circuler de **C** à **D**). Dans notre exemple, comme **A** était faux, **B** n'était pas connu et donc nous avons fait un calcul erroné.

Dans le cas général, il n'est donc pas possible d'effectuer une propagation locale des informations.

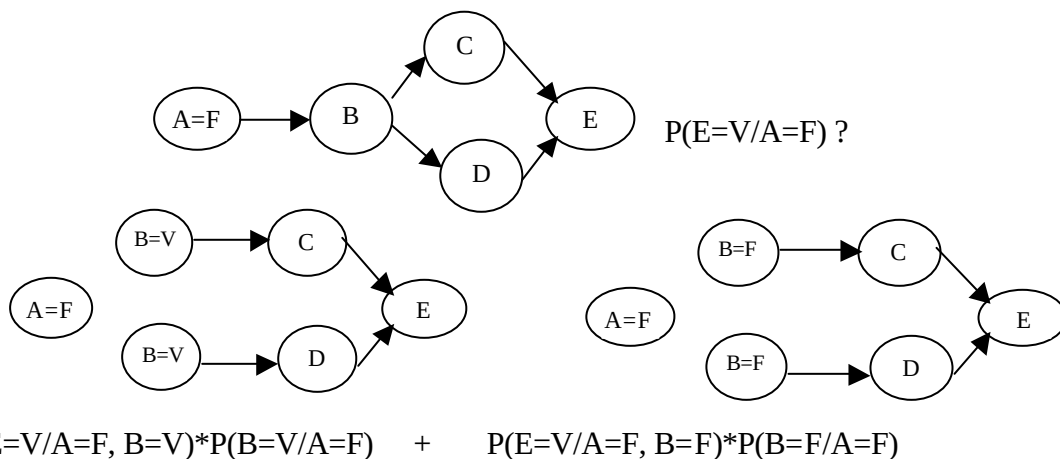
L'une des méthodes employées s'appelle conditionnement et consiste, simplement, à exécuter les étapes suivantes :

- Identifier un ensemble de noeuds tel que, si tous les arcs partant de ces noeuds étaient supprimés du réseau, le réseau n'aurait plus aucune boucle (**B** joue ce rôle dans l'exemple précédent),
- Considérer l'ensemble des hypothèses possibles sur les valeurs de chacun de ces noeuds,
- Dans le cadre de chacune de ces hypothèses, effectuer les propagations "locales" dans le réseau sans boucle correspondant, et en déduire la probabilité conditionnelle recherchée,
- Sommer les probabilités obtenues, dans chaque hypothèse, pondérées par la probabilité de chaque hypothèse [4].

Dans l'exemple précédent, il suffit d'écrire :

$$P(E/A) = \sum_B P(E/B, A) \cdot P(B/A)$$

On voit que dans ce type d'approche, il est important de bien choisir l'ensemble des N noeuds qui suppriment toutes les boucles. En effet, en supposant que chaque noeud a k états possibles, le nombre de propagations complètes à effectuer est égal à k^N .



II.1.2.2. Arbre de jonction

Il existe une autre méthode plus technique appelée la construction de l'arbre de jonction. A partir d'un réseau quelconque, on peut construire un réseau qui est un arbre dont les noeuds correspondent à des combinaisons de noeuds du réseau initial. Il est alors possible de transformer le problème de l'inférence dans le réseau initial en un problème de propagation d'informations plus complexes dans le réseau transformé.

Cette méthode est, aujourd'hui, la meilleure connue en termes de complexité algorithmique. Cependant, il a été démontré que le problème général de l'inférence dans un réseau bayésien est NP-complet. Dans certains cas, l'utilisation de méthodes heuristiques ou approchées est nécessaire pour des raisons de grande taille [4].

II.1.2.3. Méthodes approchées

L'une des méthodes, peut-être les plus simples et les plus efficaces pour des réseaux complexes, est basée sur la simulation.

Considérons à nouveau l'exemple précédent. Supposons qu'on réalise un ensemble de simulations, grâce à un générateur de nombres aléatoires, de la façon suivante :

- Pour chaque noeud, dont on connaît l'état à priori, ne rien faire (ici, A)
- Pour chaque noeud sans parent, effectuer un tirage conforme à ses probabilités inconditionnelles, et placer ce noeud dans l'état obtenu (ici, aucun noeud n'est concerné).
- Dès qu'un noeud voit tous ses parents affectés d'un état déterminé, effectuer un tirage de l'état de ce noeud, conformément à la table de probabilités qui le conditionne à l'état de ses parents (ici, B, C, D, E).
- Dès que tous les noeuds ont un état déterminé, ranger la configuration obtenue dans une table.

Effectuer ce tirage un grand nombre de fois.

Une fois ce travail fait, nous pouvons disposer d'une estimation de la probabilité de n'importe quel noeud, sous l'hypothèse initiale (ici $A=F$). Ceci permet donc de lire directement une valeur estimée de $P(E/A)$ [4].

II.1.3. Les principaux algorithmes d'Inférence [HIGR99], [BELL02]

Les réseaux bayésiens ont été développés, au début des années 1980, pour tenter de résoudre certains problèmes de prédiction et d'abduction, courants en Intelligence Artificielle (IA). Dans ce type de tâche, il est nécessaire de trouver une interprétation cohérente des observations avec les données connues *a priori*. L'inférence probabiliste signifie donc le calcul de $P(Y/X)$ où X est un ensemble d'observations et Y un ensemble de variables décrivant le problème et qui sont jugées importantes pour la prédiction ou le diagnostic.

Les premiers algorithmes d'inférence, pour les réseaux bayésiens qui ont été proposés s'agissaient d'une architecture à passage de messages et ils étaient limités aux arbres. Dans cette technique, à chaque noeud est associé un *processeur*, qui peut envoyer des messages, de façon asynchrone à ses voisins, jusqu'à ce qu'un équilibre soit atteint, en un nombre fini d'étapes. Cette méthode a été, depuis, étendue aux réseaux quelconques pour donner l'algorithme de Jensen Lauritzen Olsen (JLO) servant comme moteur d'inférence exacte. Cette méthode est aussi appelée *algorithme de l'arbre de jonction*.

L'autre méthode, développée s'appelle le *cut-set conditioning* : elle consiste à instancier un certain nombre de variables de manière à ce que le graphe restant forme un arbre. On procède à une propagation par messages sur cet arbre. Puis une nouvelle instantiation est choisie. On réitère ce processus jusqu'à ce que toutes les instantiations possibles aient été utilisées.

Le principal avantage de cette méthode est que son besoin en mémoire est minimal (linéaire sur la taille du réseau); alors que la méthode de l'arbre de jonction, plus rapide, a une complexité en espace exponentielle [BELL02].

II.2. Apprentissage [LERa00], [LERb00], [LERc00]

Nous avons montré que la représentation intuitive d'un graphe de causalités pouvait être rendue quantitative par l'utilisation de probabilités.

Ensuite, nous avons montré que les propriétés du graphe de causalité permettaient de faciliter les calculs à l'intérieur de ce graphe, et nous avons décrit les principales méthodes d'inférence.

La dernière question qui se pose, et elle est importante, est : où trouver ces probabilités ? Il est en effet assez peu réaliste de penser qu'un expert pourra fournir de façon numérique l'ensemble des paramètres nécessaires à l'inférence dans un graphe. Même si certaines études ont montré que la sensibilité des conclusions aux paramètres étaient relativement faible (c'est-à-dire que l'on a surtout besoin d'ordres de grandeur plutôt que de probabilités réelles), il peut être intéressant, dans certains cas, de déterminer ces paramètres à partir d'une base d'exemples. Il s'agit donc d'apprentissage, dans la mesure où nous cherchons à trouver le jeu de paramètres, tel que *la structure du réseau étant connue*, celui-ci prenne en compte, de la meilleure façon possible, la base d'exemples dont nous disposons [BECK99].

Dans les réseaux bayésiens, l'apprentissage peut porter sur la structure ou sur les paramètres des distributions conditionnelles. Nous allons, par la suite, présenter les principes des méthodes utilisées dans ces deux types d'apprentissage :

- . **Apprentissage de paramètres** : la structure d'un réseau étant donnée, rechercher le meilleur jeu de paramètres (c'est-à-dire, rappelons-le, les différentes probabilités conditionnelles utilisées dans le graphe) pour rendre compte des données observées,
- . **Apprentissage de structure** : sans aucune hypothèse sur la structure du graphe de causalités, rechercher celle, qui, une fois munie de ses meilleurs paramètres, rende compte, le mieux possible, des données observées.

II.2.1. Les principaux algorithmes d'Apprentissage

II.2.1.1. L'algorithme PC, recherche de causalité

L'algorithme PC a été introduit par Spirtes, Glymour et Scheines en 1993. Cette méthode utilise un test statistique pour évaluer s'il y a indépendance conditionnelle entre deux variables. A partir de l'ensemble des tables d'indépendances conditionnelles, il est alors possible de reconstruire la structure du réseau bayésien (aussi appelé graphe causal). En pratique, un graphe complètement connecté sert de point de départ, et lorsqu'une indépendance conditionnelle est détectée, l'arc correspondant est retiré [FRAN01].

II.2.1.2 Structures de Bayes naïves

Le classifieur de Bayes naïf peut être vu comme un réseau bayésien dont la structure contient uniquement des arcs du noeud classe C vers les autres noeuds X_i (variables observées).

Cette structure est liée à la simplification de la loi jointe suivante :

$$P(C, X_1, \dots, X_n) = P(C) \cdot P(X_1 / C) \dots P(X_n / C)$$

La structure de Bayes naïve suppose que les observations sont indépendantes conditionnellement à la classe. Cette hypothèse peut être levée en utilisant un classifieur de Bayes naïf *augmenté* où les observations peuvent être reliées selon une structure plus ou moins compliquée. Nous utilisons, plus précisément, comme structure augmentée, le meilleur arbre reliant les observations [FRAN01].

II.2.1.3. L'algorithme K2

L'algorithme le plus connu est K2, qui ajoute progressivement des arcs, en ne conservant un arc qui vient d'être ajouté que s'il améliore la performance du réseau suivant une métrique donnée.

Voici comment cet algorithme construit le graphe $X_1 \circledast X_2 \circledast X_3$. Après avoir commencé avec un réseau sans arc, **K2** essaie d'ajouter l'arc $X_1 \circledast X_2$. Comme cet arc améliore la performance, il est conservé. Ensuite, **K2**, essaie d'ajouter l'arc $X_1 \circledast X_3$, puis l'arc $X_2 \circledast X_3$. C'est ce dernier qui obtient le meilleur score, par rapport à la métrique donnée. Et ainsi de suite [BECK99].

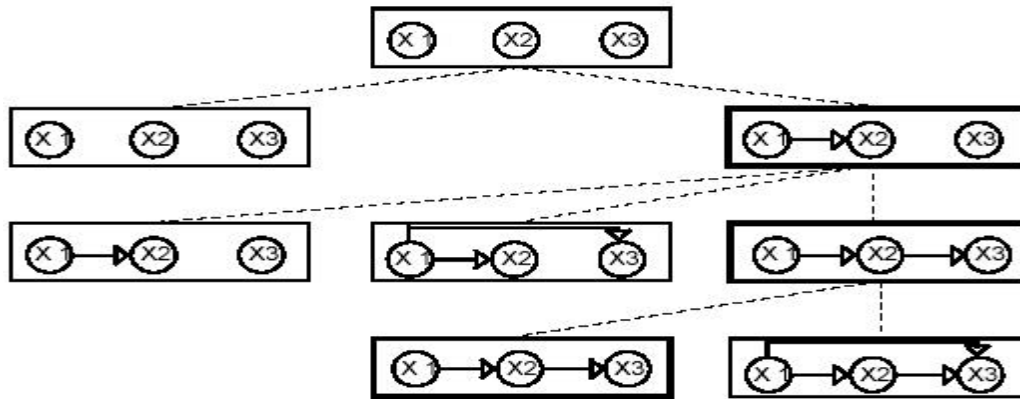


Figure 17 : Exemple d'algorithme K2 [BECK99]

II.2.1.4. L'algorithme Structural EM [HIGR99]

L'algorithme *Expectation-Maximisation* (**EM**) permet de trouver le paramètre qui maximise la vraisemblance dans les problèmes où certaines variables sont cachées ou non observées.

L'algorithme EM estime le paramètre, de manière itérative, en partant d'une estimation initiale. Chaque itération consiste en une étape appelée *Expectation* (**E**), calculant la distribution de probabilité des variables cachées sachant les variables visibles et le paramètre courant; et d'une étape appelée *Maximisation* (**M**), consistant à réestimer le paramètre de manière à ce qu'il maximise la vraisemblance des observations sous la condition que la distribution estimée à l'étape E soit exacte. On peut montrer que chaque itération augmente la vraisemblance ou la laisse inchangée si un minimum local est atteint.

L'algorithme **EM** est particulièrement bien adapté aux réseaux bayésiens, et ce pour au moins deux raisons. La première est que l'étape **E** n'est autre qu'un problème d'inférence, et cette étape prend donc une forme générique quelque soit le type de réseaux. Deuxièmement, l'étape **M** est relativement simple à effectuer car elle peut s'effectuer localement pour chaque variable. En effet, si on note P_i la distribution conditionnelle de chaque variable X_i on a :

$$P_i = P(X_i / pa(X_i), \mathbf{q}_i)$$

Avec $\mathbf{q} = \{\mathbf{q}_1, \dots, \mathbf{q}_N\}$ l'ensemble de tous les paramètres locaux $\mathbf{q}_i$.

III. Mise en œuvre des réseaux bayésiens

Nous abordons maintenant la mise en oeuvre des réseaux bayésiens dans des applications pratiques. Dans ce chapitre, nous présentons essentiellement des aspects méthodologiques, en essayant de répondre aux trois questions suivantes : pourquoi, où (dans quelles applications) et comment utiliser des réseaux bayésiens ?

III.1. Pourquoi utiliser des réseaux bayésiens ? [NAIM04]

Selon le type d'application, l'utilisation pratique d'un réseau bayésien peut être envisagée au même titre que celle d'autres modèles : réseaux neuronaux, systèmes experts, arbres de décision, modèles d'analyses de données (régressions linéaires), arbres de défaillances, modèles logiques, etc.

Naturellement, le choix de la méthode fait intervenir différents critères, comme la facilité, le coût et le délai de mise en oeuvre d'une solution. En dehors de toute considération théorique, les aspects suivants des réseaux bayésiens les rendent, dans de nombreux cas, préférables à d'autres modèles :

- **Acquisition des connaissances.** La possibilité de rassembler et de fusionner des connaissances de **diverses natures** dans un même modèle : retour d'expérience (données historiques ou empiriques), expertise (exprimée sous forme de règles logiques, d'équations, de statistiques ou de probabilités subjectives), observations. Dans le monde industriel par exemple, chacune de ces sources d'information, quoique présente, est souvent insuffisante individuellement pour fournir une représentation précise et réaliste du système analysé.
- **Représentation des connaissances.** La représentation graphique d'un réseau bayésien est explicite, intuitive et compréhensible, par un non-spécialiste, ce qui facilite à la fois la validation du modèle, ses évolutions éventuelles et surtout son utilisation. Typiquement, un décideur est beaucoup plus enclin à s'appuyer sur un modèle dont il comprend le fonctionnement qu'à faire confiance à une « boîte noire ».
- **Utilisation des connaissances.** Un réseau bayésien est polyvalent : on peut se servir du même modèle pour évaluer, prévoir, diagnostiquer, ou optimiser des décisions, ce qui contribue à « rentabiliser » l'effort de construction du réseau bayésien.
- **Qualité de l'offre en matière de logiciels :** il existe aujourd'hui de nombreux logiciels pour saisir et traiter des réseaux bayésiens. Ces outils présentent des fonctionnalités plus ou moins évoluées : apprentissage des probabilités, apprentissage de la structure du réseau bayésien, possibilité d'intégrer des variables continues, des variables d'utilité et de décision, etc.

Nous allons présenter, plus en détail, ces différentes motivations de l'utilisation des réseaux bayésiens.

III.1.1 Acquisition des connaissances

➤ Un recueil d'expertise facilité

La représentation des connaissances utilisées dans les réseaux bayésiens est la plus intuitive possible : relier simplement des causes et des effets par des flèches. Pratiquement toute représentation graphique d'un domaine de connaissances peut être présentée sous cette forme.

De nombreuses expériences montrent qu'il est souvent plus facile pour un expert de formaliser ses connaissances sous forme de graphe causal que sous forme de système à base de règles, en particulier parce que la formulation de règles sous la forme SI... ALORS est très contraignante, et peut être facilement mise en défaut.

Certains auteurs considèrent qu'il existe une différence de nature entre les deux processus d'acquisition de connaissances. Lorsqu'on essaie de mettre au point un système expert, par exemple pour une application de diagnostic, l'expert doit décrire le processus de raisonnement qui le conduit de ses observations à une conclusion. En revanche, un modèle fondé sur un graphe causal décrit la perception de l'expert du fonctionnement du système. Effectuer un diagnostic n'est alors qu'une résultante de cette modélisation.

➤ Un ensemble complet de méthodes d'apprentissage

Les algorithmes actuels utilisés pour les réseaux bayésiens permettent d'envisager l'apprentissage de façon très complète :

- En l'absence totale de connaissances, on peut rechercher à la fois la structure du réseau la plus adaptée, c'est-à-dire les relations de dépendance et d'indépendance entre les différentes variables, et les paramètres, ou probabilités, c'est-à-dire, la quantification de ces relations.
- Si l'on dispose de connaissances, à priori, sur la structure des causalités, et d'une base d'exemples représentative, la détermination des matrices de probabilités conditionnelles, qui sont les paramètres du réseau, peut être effectuée par simple calcul de fréquences, par détermination du maximum de vraisemblance, ou par des méthodes bayésiennes.

Ces méthodes peuvent être étendues dans le cadre de bases de données incomplètes. Dans l'optique de rechercher un compromis entre apprentissage et généralisation, il est également possible d'effectuer des apprentissages en contraignant la structure du réseau.

III.1.2. Représentation des connaissances

➤ Un formalisme unificateur

La plupart des applications qui relèvent des réseaux bayésiens sont des applications d'aide à la décision. Par nature, ces applications intègrent un certain degré d'incertitude, qui est très bien pris en compte par le formalisme probabiliste des réseaux bayésiens.

Par exemple, les applications de data mining sont, en général, construites sur le schéma suivant. On utilise une base de données pour mettre au point un modèle prédictif. Par définition, une prévision comporte une part d'incertitude. Or la décision, elle, doit souvent être binaire : dans une application de scoring, on doit par exemple accorder ou refuser le crédit. La façon la plus naturelle d'interpréter un score est donc une probabilité.

Les techniques disponibles pour traiter ce genre de problème (modèles de régression, réseaux de neurones, arbres de décision) ne sont pas construites sur un formalisme de probabilités. C'est, à posteriori, qu'on attribue en général une interprétation en termes de probabilités de la prévision d'un réseau neuronal ou d'un arbre de décision.

Les réseaux bayésiens ne sont qu'une représentation d'une distribution de probabilité. C'est une telle distribution que l'on représente à partir de connaissances explicites ou qu'on approche à partir d'une base de données, et c'est à partir de la distribution approchée que l'on effectue des inférences.

Toute prévision, issue d'un réseau bayésien, est donc, par construction, une probabilité.

De plus, les réseaux bayésiens permettent de considérer, dans un même formalisme, la représentation de modèles de causalités et les statistiques multivariées. Il en est de même des techniques les plus utilisées pour le data-mining comme les arbres de décision ou les réseaux de neurones, qui peuvent également être représentés au sein de ce formalisme.

➤ Une représentation des connaissances lisible

Les deux propriétés fondamentales des réseaux bayésiens sont, d'abord, d'être des graphes orientés, c'est-à-dire, de représenter des causalités et non de simples corrélations et, ensuite, de garantir une correspondance entre la distribution de probabilité, sous-jacente, et le graphe associé.

D'après le théorème d'indépendance graphique, les relations de causalité et d'indépendance qui peuvent être lues sur le graphe sont également vraies dans la distribution sous-jacente.

Considérons le cas d'une application de data-mining, où l'on cherche à comprendre les interrelations entre des variables contenues dans une base de données de clients. Si nous se trouvons dans le cas où le réseau est entièrement mis au point à partir des données (cas de l'apprentissage de la structure et des paramètres), cela signifie que nous va disposer d'une visualisation graphique de ces interrelations. Avant même d'utiliser ce réseau pour effectuer des inférences, nous va disposer d'une visualisation de la connaissance, directement lisible et interprétable par des experts du domaine.

III.1.3 Utilisation de connaissances

➤ Une gamme de requêtes très complète

L'utilisation première d'un réseau bayésien est le calcul de la probabilité d'une hypothèse, connaissant certaines observations. Cependant, les possibilités offertes par les algorithmes d'inférence permettent d'envisager une gamme de requêtes très complète, et qui peut être extrêmement intéressante dans certains types d'applications.

Tout d'abord, il n'y a aucune réelle contrainte sur les informations nécessaires pour être en mesure de calculer la probabilité d'un fait : on peut connaître exactement la valeur d'une variable, savoir qu'elle est égale à l'une à l'autre de deux valeurs, ou encore savoir, avec certitude, qu'une de ses valeurs possibles est exclue. Dans tous les cas, l'inférence est possible, et la nouvelle information permet d'affiner les conclusions.

Il n'y a pas d'entrées ni de sorties dans un réseau bayésien (ou de variables indépendantes et dépendantes). Le réseau peut donc être utilisé pour déterminer la valeur la plus probable d'un noeud en fonction d'informations données (prévoir, ou sens entrées sorties), mais également pour connaître la cause la plus probable d'une information donnée (expliquer, ou sens sorties entrées). En termes d'inférences, cette dernière requête s'appelle explication la plus probable et revient, l'état de certaines variables étant observé, à rechercher l'état des autres variables pour lequel ce qui a été observé, était le plus probable. Parmi les autres requêtes importantes, l'analyse de sensibilité à une information mesure comment la probabilité d'une hypothèse s'accroît quand on a fait une observation. Certaines observations peuvent ainsi être considérées comme inutiles, suffisantes ou cruciales, par rapport à une hypothèse donnée.

Le mécanisme de propagation peut être, également, utilisé pour déterminer l'action la plus appropriée à effectuer, ou l'information la plus pertinente à rechercher. Considérons par exemple un problème de diagnostic, dans lequel manquent plusieurs des données qui permettraient de conclure.

Le mécanisme de propagation, dans un réseau, permet de connaître la donnée dont la connaissance apporterait le maximum d'informations. Dans le cas où la recherche de chaque donnée a un coût, il est possible de rechercher la solution optimale, en tenant compte de ce coût. De plus, il est possible de chercher également une séquence optimale d'actions ou de requêtes.

➤ Optimisation d'une fonction d'utilité

Imaginons un problème de classification, par exemple, un problème de détection de fraudes, sur des cartes bancaires, ou dans l'utilisation de services de télécommunications. Rechercher le système qui donne, avec la meilleure fiabilité possible, la probabilité de fraude n'est peut être pas l'objectif réel de ce type d'application. En effet, ce qu'on cherche ici à optimiser est une utilité économique. Sachant que les fausses alarmes aussi bien que les fraudes manquées ont un coût, l'objectif est bien de minimiser le coût global. Une version spécifique des réseaux bayésiens, appelée diagramme d'influences, permet de les adapter à ce type de problème.

Dans les diagrammes d'influence, deux autres types de noeuds sont ajoutés aux noeuds qui représentent des variables :

- les noeuds de décision, figurés par des carrés,
- un noeud d'utilité, figuré par un losange.

Le graphe de la figure 18 ci-après représente un diagramme d'influence pour un problème de détection de fraude sur une carte bancaire. Les variables représentées sont les suivantes :

- La variable F est binaire et représente le fait qu'il y a ou non fraude.
- La variable B représente le résultat d'une vérification effectuée sur une base de données. Cette variable a trois modalités : le contrôle est négatif, positif, ou non effectué.
- La variable P a également trois modalités, et représente le résultat d'un contrôle d'identité du porteur.
- Le noeud de décision D représente la décision d'effectuer les contrôles complémentaires B et P. Ce noeud a donc également trois modalités : n'effectuer aucun test, effectuer le test B, ou effectuer les deux tests B et P.
- Le noeud de décision A représente la décision d'autoriser la transaction, et est donc binaire.
- Le noeud d'utilité V est une fonction de l'ensemble des variables précédentes, représentant le coût de la situation.

En outre, on suppose connus, le montant de la transaction et le coût de chaque contrôle, et les tables de probabilités conditionnelles, reliant les variables entre elles. L'objectif est de prendre les bonnes décisions D et A; autrement dit, de prendre les décisions qui minimisent l'espérance mathématique de V.

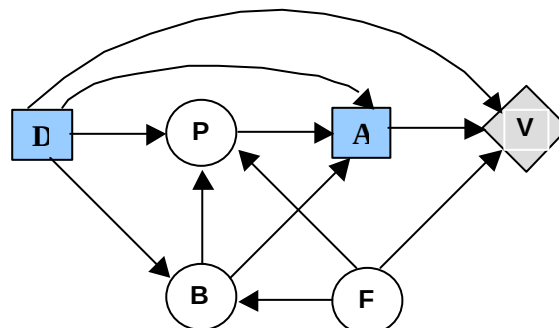


Figure 18: Un diagramme d'influence pour la fraude sur carte bancaire

III.1.4. Limites des réseaux bayésiens [NAIM04]

➤ Un recul encore insuffisant pour l'apprentissage

Dans la mesure où elle s'est surtout développée dans le cadre des systèmes experts, la technique des réseaux bayésiens n'a pas immédiatement intégré l'ensemble de la problématique de l'apprentissage, comme cela avait été le cas des réseaux neuronaux. Aujourd'hui, l'essentiel de la littérature sur l'apprentissage avec des réseaux bayésiens ignore le problème de la capacité de généralisation d'un modèle, et des précautions que cela implique au moment de la construction du modèle. La prise en compte de ce problème peut s'effectuer par le choix du critère de recherche ou de distance des distributions de probabilité. En effet, l'apprentissage de réseaux bayésiens revient à rechercher, parmi un ensemble de distributions, la distribution la plus proche possible, en un certain sens, de la distribution, représentée par les données. En limitant l'ensemble de recherche, on peut éviter le problème de l'apprentissage, qui revient dans ce cas à calculer exactement la distribution représentée par les exemples.

➤ Utilisation des probabilités

L'utilisation des graphes de causalités est, nous l'avons dit, une approche très intuitive. L'utilisation des probabilités, pour rendre ces modèles quantitatifs, était justifiée. Il reste, cependant, que la notion de probabilité est, au contraire, assez peu intuitive. Il est en effet assez facile de construire des paradoxes fondés sur des raisonnements probabilistes.

Les modèles déterministes, formulés en termes d'entrées et de sorties, comme les modèles de régression, les réseaux de neurones, ou les arbres de décision, même s'ils peuvent être réinterprétés dans le cadre d'un formalisme probabiliste, restent d'un abord plus facile.

➤ Lisibilité des graphes

En effet, même si la connaissance manipulée dans les réseaux bayésiens, ou extraites des données, par les algorithmes d'apprentissage associé, est lisible, puisque représentée sous forme de graphes. Elle reste moins lisible que celle représentée par un arbre de décision, par exemple, surtout si ce graphe présente un grand nombre de noeuds. Notons aussi que l'information, représentée par le graphe est la structure des causalités. Les probabilités ne sont pas représentables, et nous n'avons donc pas idée, à la simple lecture du graphe, quel arc est « important ».

➤ Les variables continues

L'essentiel des algorithmes, développés pour l'inférence et l'apprentissage dans les réseaux bayésiens, aussi bien que les outils, disponibles sur le marché pour mettre en oeuvre ces algorithmes, utilisent des variables discrètes.

En effet, la machinerie des algorithmes d'inférence est, essentiellement, fondée sur une algèbre de tables de probabilités. De même, les algorithmes d'apprentissage modélisent, en général, les distributions de probabilité des paramètres contenus dans les tables du réseau, c'est-à-dire de probabilités discrètes.

Même s'il est théoriquement possible de généraliser les techniques développées aux variables continues, il semble que la communauté de recherche, travaillant sur les réseaux bayésiens n'a pas encore vraiment intégré ces problèmes. Cela pénalise cette technologie, en particulier, pour des applications de data mining où variables continues et discrètes cohabitent.

➤ La complexité des algorithmes

La généralité du formalisme des réseaux bayésiens, aussi bien en termes de représentation que d'utilisation, les rend difficiles à manipuler à partir d'une certaine taille. La complexité des réseaux bayésiens ne se traduit pas seulement en termes de compréhension par les utilisateurs. Les problèmes sous-jacents sont pratiquement tous de complexité non polynomiale, et conduisent à développer des algorithmes approchés, dont le comportement n'est pas garanti pour des problèmes de grande taille.

III.1.5. Comparaison avec d'autres techniques [NAIM04]

Du point de vue des applications, les avantages et inconvénients des réseaux bayésiens, par rapport à quelques-unes des techniques concurrentes, peuvent se résumer sur le tableau ci-dessous.

Connaissances	Analyse de données	Réseaux neuronaux	Arbre de décision	Système expert	Réseaux bayésiens
ACQUISITION					
Expertise seulement					
Données seulement	+		+		+
Mixte	+	+	+		
Incrémental		+			
Généralisation	+		+		+
Données incomplètes		+			
REPRESENTATION					
Incertitude				+	
Lisibilité	+		+	+	
Facilité		+			
Homogénéité					
UTILISATION					
Requêtes élaborées	+			+	
Utilité économique	+	+			
Performance	+				

Tableau 1: Avantages comparatifs des réseaux bayésiens

Nous avons regroupé avantages et inconvénients selon les trois rubriques utilisées précédemment, l'acquisition, la représentation et l'utilisation des connaissances. La représentation adoptée est la suivante :

- A chaque ligne correspond une caractéristique, qui peut être un avantage ou la prise en compte d'un problème spécifique.
- Si la technique considérée permet de prendre en compte ce problème ou présente cet avantage, un signe + est placé dans la case correspondante.
- Un signe est placé dans la case de la meilleure technique du point de vue de la caractéristique considérée.

III.2. Où utiliser des réseaux bayésiens ?

Toute application, mettant en œuvre des connaissances, peut donc relever de l'utilisation des réseaux bayésiens, qu'il s'agisse de formaliser la connaissance d'experts, d'extraire la connaissance contenue dans des bases de données, ou d'utiliser, le plus rationnellement possible, l'un ou l'autre type de connaissance.

Un réseau bayésien est un moyen de représenter la connaissance d'un système. Une telle représentation n'est, bien entendu, pas une fin en soi ; elle s'effectue, selon les contextes, dans le but de :

- **prévoir** le comportement du système
- **diagnostiquer** les causes d'un phénomène, observées, dans le système
- **contrôler** le comportement du système
- **simuler** le comportement du système
- **analyser** des **données** relatives au système
- prendre des **décisions** concernant le système

Les réseaux bayésiens ont pour objectif d'acquérir, représenter et utiliser la connaissance. Leur champ d'application est donc vaste, d'autant que le terme «système» s'entend ici dans son sens le plus large. Il peut s'agir, pour donner quelques exemples, du contenu du caddie d'un client de supermarché, d'un navire de la Marine, du patient d'une consultation médicale, du moteur d'une automobile, d'un réseau électrique ou de l'utilisateur d'un logiciel. Ajoutons que la communauté de chercheurs, qui développent la théorie et les applications des réseaux bayésiens, rassemble plusieurs disciplines scientifiques : l'intelligence artificielle, les probabilités et statistiques, la théorie de la décision, l'informatique et aussi les sciences cognitives. Ce facteur contribue à la diffusion et donc à la multiplicité des applications des réseaux bayésiens [NAIM04].

III.2.1. Classification des applications [BECK99]

III.2.1.1. Par type

Les applications de prédilection des réseaux bayésiens sont essentiellement :

- Systèmes experts de diagnostic,
- Systèmes de classification et de scoring,
- Systèmes de data-mining.

Les systèmes de diagnostic sont ceux qui utilisent le plus complètement les possibilités des réseaux bayésiens, en particulier en ce qui concerne les capacités d'explication, de simulation, etc.

Les systèmes de classification mettent en général en œuvre des architectures simplifiées de réseaux (arbres ou poly-arbres). Des études et des applications récentes montrent que les classifieurs, basés sur des arbres bayésiens, donnent des résultats en général significativement meilleurs que les algorithmes de classification de type arbre de décision.

Les méthodes d'apprentissage de structure dans les réseaux bayésiens permettront de donner tout son sens au terme de *data-mining*. S'il s'agit, en effet, de rechercher des relations entre des variables sans à priori, ou les réseaux de neurones, ou les arbres de décision ne sont adaptés à ce type de problème.

Dans les applications de modélisation numérique comme la prévision, le contrôle, ou l'estimation, il nous semble en revanche que le formalisme global des réseaux bayésiens est trop lourd pour être utilisé tel quel. Cela ne signifie pas pour autant qu'il ne puisse pas contribuer à de telles applications. A terme, l'unification des algorithmes d'apprentissage permettra sans doute d'intégrer un modèle neuronal de prévision à l'intérieur d'un système plus global, où pourront être optimisés, simultanément, soit des paramètres d'apprentissage, soit des paramètres de la décision, basée sur la prévision.

III.2.1.2. Par domaine

➤ Santé

Les premières applications des réseaux bayésiens ont été développées dans le domaine du diagnostic médical.

Dans le domaine de la santé en général, une application récente des algorithmes, issus des réseaux bayésiens, a permis d'améliorer considérablement la recherche de la localisation de certains gènes (projet Human Genome).

➤ Industrie

Les applications dans le domaine industriel sont :

- Diagnostic,
- Fusion de données

➤ Banque/Finance

Bien que les applications dans le domaine de la banque et de la finance soient encore très rares (si elles existent), les réseaux bayésiens sont, à terme, un concurrent sérieux des autres techniques dans les applications de :

- Analyse financière,
- Scoring,
- Évaluation de risques, détection de fraudes.

➤ Informatique

Les réseaux bayésiens peuvent être utilisés pour des applications d'accès intelligent à l'information, d'adaptation intelligente des systèmes aux utilisateurs.

Un projet européen cherche à promouvoir les applications des réseaux bayésiens dans la vérification de logiciels pour lesquels des contraintes de sécurité importantes existent, en conformité avec les normes européennes.

III.3. Comment utiliser des réseaux bayésiens ? [NAIM04]

Malgré la diversité des applications, la construction d'un réseau bayésien se réalise, schématiquement, en trois étapes essentielles:

- ✓ Identification des variables et de leur espace d'état,
- ✓ Définition de la structure du réseau bayésien,
- ✓ Définition de la loi de probabilité conjointe des variables.

III.3.1. Identification des variables et de leur espace d'états

La première étape de construction du réseau bayésien est la seule pour laquelle l'intervention humaine est absolument indispensable. Il s'agit de déterminer l'ensemble des **variables** X_i , catégorielles ou numériques, qui caractérisent le système. Comme dans tout travail de modélisation, un compromis entre la précision de la représentation et la faisabilité de la construction du modèle doit être trouvé, au moyen d'une discussion entre les experts et le modélisateur.

Lorsque les variables sont identifiées, il est ensuite nécessaire de préciser l'espace d'états de chaque variable X_i , c'est-à-dire l'ensemble de ses valeurs possibles.

La majorité des logiciels de réseaux bayésiens ne traite que des modèles à variables discrètes, ayant un nombre fini de modalités. Si tel est le cas, il est impératif de discrétiser les plages de variation des variables continues.

Cette limitation est parfois gênante en pratique, car des discrétisations trop fines peuvent conduire à des tables de probabilités de grande taille, de nature à saturer la mémoire de l'ordinateur.

III.3.2. Définition de la structure du réseau bayésien

La deuxième étape consiste à identifier les **liens** entre variables, c'est-à-dire à répondre à la question: pour quels couples $(i; j)$ la variable X_i influence-t-elle la variable X_j ?

Dans la grande majorité des applications, cette étape s'effectue par l'interrogation d'experts. Dans ce cas, des itérations sont souvent nécessaires pour aboutir à une description consensuelle des interactions entre les variables X_i . L'expérience montre cependant que la représentation graphique du réseau bayésien, est dans cette étape, un support de dialogue extrêmement précieux.

Un réseau bayésien ne doit pas comporter de circuit orienté ou «boucle » (figure 20). Cependant, le nombre et la complexité des dépendances, identifiées par les experts, laissent parfois supposer que la modélisation par un graphe acyclique est impossible. Il est alors important de garder à l'esprit que, quelles que soient les dépendances stochastiques entre des variables aléatoires discrètes, il existe toujours une représentation par réseau bayésien de leur loi conjointe. Ce résultat théorique est fondamental et montre bien la puissance de modélisation des réseaux bayésiens.

Lorsque nous dispose d'une quantité suffisante de données de retour d'expérience, concernant les variables X_i , la structure du réseau bayésien peut également être apprise automatiquement par le réseau bayésien, à condition, bien sûr, que le logiciel utilisé soit doté de la fonctionnalité adéquate.

III.3.3. Loi de probabilité conjointe des variables

La dernière étape de construction du réseau bayésien consiste à renseigner les **tables de probabilités** associées aux différentes variables.

Dans un premier temps, la connaissance des **experts**, concernant les lois de probabilité des variables, est intégrée au modèle.

Concrètement, deux cas se présentent selon la position d'une variable X_i dans le réseau bayésien:

- La variable X_i n'a pas de variable parente : les experts doivent préciser la loi de probabilité marginale de X_i .
- La variable X_i possède des variables parentes : les experts doivent exprimer la dépendance de X_i en fonction des variables parentes, soit au moyen de probabilités conditionnelles, soit par une équation déterministe (que le logiciel convertira ensuite en probabilités).

Le recueil de lois de probabilités, auprès d'experts, est une étape délicate du processus de construction du réseau bayésien. Typiquement, les experts se montrent réticents à chiffrer la plausibilité d'un événement qu'ils n'ont jamais observé.

La quasi-totalité des logiciels commerciaux de réseaux bayésiens permet l'apprentissage automatique des tables de probabilités à partir de **données**.

Par conséquent, dans un second temps, les éventuelles observations des X_i peuvent être « incorporées » au modèle, afin d'affiner les probabilités introduites par les experts.

Il est rare, en pratique, que les données soient suffisamment nombreuses et fiables pour caractériser, de manière satisfaisante, la loi de probabilité conjointe des variables X_i . Cependant, si tel est le cas, l'apprentissage automatique des probabilités rend inutile la phase de renseignement du modèle par des probabilités « expertes » ; on peut alors se contenter, dans la phase initiale, d'attribuer à chaque variable une loi de probabilité uniforme.

Conclusion

Le développement théorique des réseaux bayésiens date des années 1970, mais à cette époque, le manque d'algorithmes efficaces et d'outils pratiques a empêché le développement d'applications. Aujourd'hui, grâce à des outils puissants et conviviaux, les réseaux bayésiens sont une technique en pleine expansion.

Les réseaux bayésiens sont actuellement une des techniques les plus intéressantes de l'intelligence artificielle car ils permettent la représentation de la connaissance par un graphe causal intuitif et compréhensible. De plus, comme ils sont basés sur des probabilités, ils intègrent l'incertitude dans le raisonnement.

Les réseaux bayésiens apportent des solutions rapides et intuitives à différentes sortes de problématiques d'évaluation, de décision, de prévision et de diagnostic, etc.

Dans cette partie nous avons présenté les réseaux bayésiens de manière générale (les principaux concepts et définitions) ainsi que les principaux algorithmes de l'inférence et ceux de l'apprentissage et nous avons terminé par la mise en œuvre des réseaux bayésiens.

Les réseaux bayésiens représentent les dépendances entre les variables et donnent une spécification concise de la distribution de probabilités conjointe, en rendant explicites les relations d'indépendance conditionnelle.

La flexibilité des réseaux bayésiens et la possibilité d'effectuer de nombreuses inférences et apprentissages nous ont conduit à l'utilisation des réseaux bayésiens dans le domaine de la recherche d'information structurée.

Dans le prochain chapitre nous allons proposer un prototype de Recherche d'Information Structurée, utilisant les Réseaux Bayésiens.

1. Introduction

Les systèmes développés aujourd'hui pour la Recherche d'Information Structurée (RIS) dans la communauté recherche d'information sont soit des extensions des modèles classiques de la Recherche d'Information (RI), comme les modèles vectoriels ou probabilistes, soit des modèles conçus spécifiquement pour la prise en compte de la structure des documents.

Les modèles graphiques, tels que les Réseaux Bayésiens (RB), permettent de représenter des données structurées, de réaliser de l'apprentissage et de faire des inférences plus complexes que la plupart des modèles d'apprentissage qui se contentent de produire une probabilité à posteriori ou une estimation de densité. Les modèles graphiques permettent de traiter, de façon unifiée le problème de la représentation des documents et des corpus structurés, ainsi que la réponse à des demandes d'information simples et/ou structurées. Ce sont les principales raisons pour lesquelles nous nous sommes basés sur ces modèles pour la construction de notre prototype de recherche d'information structurée.

Après l'adoption des réseaux bayésiens, de nombreuses possibilités restent pour modéliser des tâches de RI et en particulier de RIS. En RIS les possibilités sont encore plus nombreuses.

L'idée, que nous avons utilisé, a été directement inspirée de la structure des documents et de celle corpus pour définir la structure du réseau bayésien.

D'après l'étude que nous avons faite sur les réseaux bayésiens, nous avons remarqué une forte limitation de la complexité des modèles, liée à la taille des données manipulées (les espaces de description des données textuelles, noeuds ou documents, ont plusieurs milliers de dimensions) et à la taille des corpus (plusieurs millions de noeuds). Par conséquent, nous avons adopté des modèles simples permettant de faire effectivement les calculs nécessaires à l'inférence.

La particularité dans la modélisation de notre prototype est d'associer un réseau bayésien à chaque document. Chaque document, qui a sa propre structure, sera de même pour le réseau bayésien qui le modélise.

Une autre particularité de notre prototype est d'associer à chaque noeud de réseaux une matrice qui contient les mots qui représentent l'élément structurel du document et leurs poids. C'est la nouveauté de notre prototype, par rapport aux systèmes de recherches d'information utilisant les réseaux bayésiens.

Notre prototype est capable de traiter des demandes d'information mêlant structure et contenu. Nous utilisons les RB pour modéliser le processus de recherche dans un corpus structuré. Nous nous intéressons à des unités d'informations, de plus en plus fines, ce qui nous conduit directement à prendre en compte le contexte d'un noeud.

2. Modélisation du prototype de RIS utilisant les RB

Avant de présenter le modèle de notre prototype de recherche d'information structurée, il est nécessaire de préciser les points suivants :

- Quels sont les objets sur lesquels travaillons nous?
- Quelle unité d'information faut-il considérer ?

Nous supposons que le corpus manipulé est organisé de manière hiérarchique, à la façon d'un système de fichiers arborescent, composé de classes ou catégories, et de documents. Cette hypothèse n'a pas d'incidence réelle sur le modèle mais simplifiera plutôt la présentation. Les documents sont définis suivant la vue logique d'un document XML : il s'agit d'une structure hiérarchique où les éléments structurels sont liés. A chaque élément structurel sont associés éventuellement des données textuelles et des attributs.

Le fonctionnement des RB se déroule en deux phases :

- la construction du réseau sollicitant experts et techniques d'apprentissage ;
- l'utilisation faisant appel à une capacité très intéressante des RB: l'inférence.

2.1. Construction du RB de notre prototype de RIS

Rappelons que la phase de construction d'un réseau bayésien se réalise schématiquement en trois étapes essentielles :

2.1.1. Identification des variables et de leur espace d'état

La construction du RB consiste à déterminer l'ensemble des **variables** V_n qui caractérisent le domaine d'application.

Nous avons utilisé la structure d'un document XML et celle de corpus pour définir la structure du RB. Par conséquent, les **nœuds** ou variables de notre RB sont alors les éléments structurels (corpus, classes, documents, titre, année, auteur, section, paragraphe...etc.) (figure20).

L'espace d'état de chaque variable X_i , dans notre cas, est que chaque **nœud** ou variable est un ensemble de mots qui le représente avec ces poids.

2.1.2. Définition de la structure du RB

La deuxième étape consiste à identifier les **liens** entre variables, c'est-à-dire à répondre à la question: pour quels couples (i; j) la variable X_i influence-t-elle la variable X_j ?

Les liens, dans notre réseau bayésien, sont inhérent à la structure des documents et de corpus et sont, par conséquent, définis par les relations suivantes (figure19): liens entre **nœud** racine et parents, les **nœuds** parents et enfants, **nœuds** enfants et **nœuds** descendants ...etc.

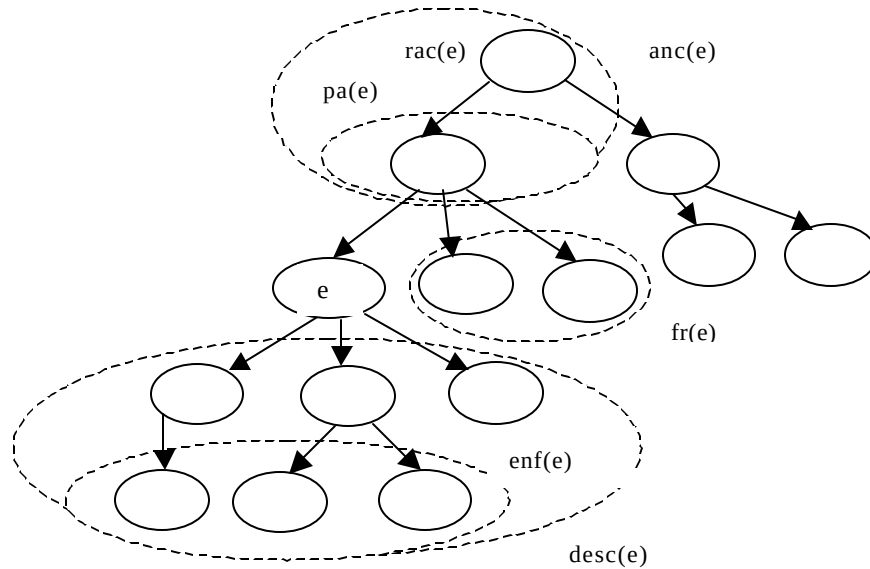


Figure 19: Différentes relations entre noeuds

La structure de notre réseau bayésien organisée, de manière hiérarchique, à la façon d'un système de fichiers arborescent, composé de classes ou catégorie et de documents structurés XML, est illustré dans la figure20 suivante :

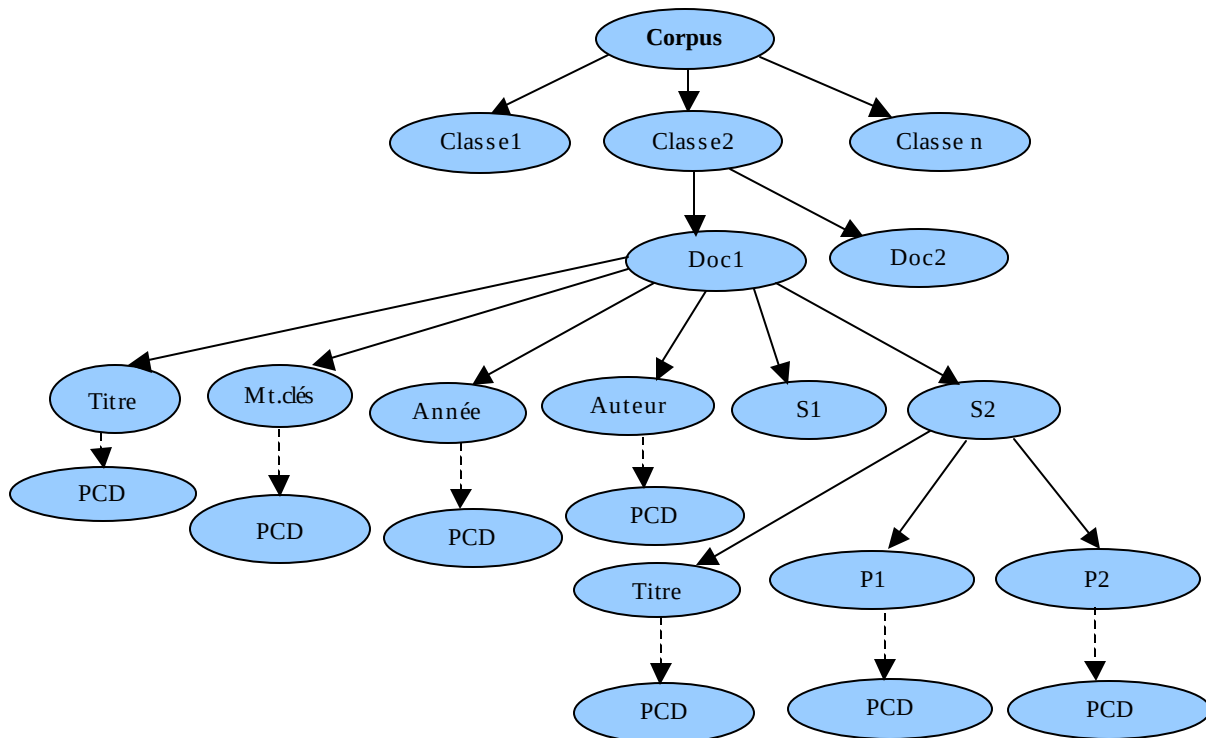


Figure 20 : la structure de Réseau Bayésien

2.1.3. Loi de probabilité conjointe des variables (inférence)

La dernière étape de construction du réseau bayésien consiste à construire les tables de probabilités associées aux différentes variables.

Dans la structure du réseau bayésien que nous avons proposé, les **nœuds** ou variables aléatoires possèdent des variables parentes. Nous allons donc exprimer les dépendances entre **nœuds** ou variables en fonction des **nœuds** ou variables parentes, au moyen de probabilités conditionnelles comme suit :

Pour calculer la probabilité $P(X)$ d'une observation, nous utilisons les indépendances décrites par le RB comme suit:

$$P(X) = \sum_{v \in X} \prod P(v_n / v_{pa(n)})$$

Tel que :

V_n : variable aléatoire (**nœud**)

$V_{pa(n)}$: parent du noeud

2.2. Utilisation du Réseau Bayésien

L'utilisation du RB repose sur la propagation de l'information au sein du réseau, c'est-à-dire, des calculs de probabilités c'est ce qu'on appelle l'inférence. Comment cette information va t-elle se répercuter sur l'état des autres variables après avoir fait une observation sur une variable ?

La modélisation de notre prototype de recherche d'information structurée, utilisant les réseaux bayésiens, se base sur deux hypothèses, à savoir :

1^{ère} Hypothèse: le corpus, utilisé pour la recherche documentaire, est supposé classé.

2^{ème} Hypothèse: la requête est composée d'un ensemble de termes.

Rappelons que les deux tâches principales d'un SRI sont:

↳ l'indexation qui est l'extraction et le stockage du contenu sémantique des documents du corpus,

↳ l'interrogation qui n'est autre que la recherche dans le corpus et la présentation des résultats.

2.2.1. L'indexation (extraction de termes)

L'opération d'indexation consiste à analyser le corpus documentaire, puis à extraire un ensemble de termes susceptible de représenter les concepts contenus dans ce dernier.

Nous avons associé à chaque terme extrait du **noeud** un poids. Ce poids mesure l'importance ou le degré de signification d'un mot dans la représentation du contenu d'un **noeud**.

Pour prendre en considération l'importance de la structure nous avons optés par la pondération suivante [13]:

Si le mot apparaît dans le **noeud** mot clé , la pondération du mot est égale à 12 points,

Si le mot apparaît dans le **noeud** titre de document, la pondération du mot est égale à 08 points,

Si le mot apparaît dans le **noeud** titre de chapitre, la pondération du mot est égale à 06 points,

Si le mot apparaît dans le **noeud** titre de section, la pondération du mot est égale à 04 points,

Si le mot apparaît dans le **noeud** introduction de document, la pondération du mot est égale à 06 points,

Si le mot apparaît dans le **noeud** introduction de chapitre, la pondération du mot est égale à 04 points,

Si le mot apparaît dans le **noeud** introduction de section, la pondération du mot est égale à 02 points,

Si le mot apparaît dans le **noeud** titre paragraphe, la pondération du mot est égale à 02 points,

Si le mot apparaît dans le **noeud** paragraphe , la pondération du mot est égale à 01 points, ...etc.

Remarque : la pondération se fait au niveau des **noeuds** descendants seulement. Pour déterminer la fonction de poids, nous nous sommes basés sur la fonction suivante :

$$poids = \frac{pti \times n_{occ} ti}{Nt}$$

Tel que :

pti : pondération du terme ti (selon la structure)

$n_{occ} ti$: nombre d'occurrence du terme ti dans le **noeud**

Nt : nombre total des termes dans le noeud

Nous avons proposé un algorithme d'indexation pour chaque **noeud**. La méthode choisie est ascendante et itérative comme suit :

1. commençons par les **noeuds** descendants. Pour chaque **noeud**, extraire les termes et calculer leur poids, en fixant un seuil introduisant les termes qui ont le maximum de poids dans la matrice.
2. construire les termes pour chaque **noeud** enfant à partir des termes de leur **noeud** descendant par l'union des mots de ces derniers. Recalculer le poids de chaque terme,
3. répéter l'étape 2 pour chaque **noeud** des ensembles parents et ancêtres.

2.2.2. La recherche d'information

L'interrogation, ou la recherche, consiste à utiliser la structure du RB, associé au prototype, et de faire de l'inférence et de l'apprentissage sur ce réseau.

2.2.2.1. L'inférence :

L'utilisation des RB repose sur la propagation de l'information au sein du réseau, c'est-à-dire, des calculs de probabilités.

Nous allons calculer les probabilités conditionnelles de chaque nœud, connaissant son parent :

$$P(v_n / v_{pa(n)})$$

D'après la définition de probabilité conditionnelle, nous avons [15] :

Et :

$$P(A / B) = \frac{|A \cap B|}{|B|}$$

$$P(A / B) = \frac{\text{nombre de réalisations possibles de } A \text{ et } B \text{ en même temps}}{\text{nombre de réalisations de } B}$$

Si $A \subset B$: si A est réalisé, alors B aussi.

Alors $P(A \cap B) = P(A)$

$$P(A / B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)}$$

Dans notre cas chaque nœud est un composant de son parent.

Donc :

$$P(v_n / v_{pa(n)}) = \frac{P(v_n)}{P(v_{pa(n)})}$$

La probabilité conditionnelle se calcule par :

$$P(v_n / v_{pa(n)}) = \frac{\text{card}(v_n)}{\text{card}(v_{pa(n)})}$$

2.2.2.2. Apprentissage

Nous avons défini deux démarches de recherches, basées sur le contenu et structure :

➤ Recherche par le contenu :

Nous proposons un algorithme pour faire une recherche plus riche qui donne plus d'informations sur l'élément qui correspond à la requête, donc, nous diminuons le bruit dès le départ.

L'algorithme que nous proposons pour la recherche est le suivant :

- 1- commencer la recherche par le **nœud** racine,
Si le terme recherché figure dans la matrice qui est associée au **nœud** racine, alors continuez la recherche, sinon arrêtez la.
- 2- passer aux **nœuds** ancêtres et lancer la recherche,
Si le terme figure dans les matrices, associées à chaque **nœud**, alors sélectionner le/les **nœud** (s) répondant (s) mieux à la requête, (sélectionner le/les **nœud** (s) qui ont le maximum de probabilité conditionnelle), en fixant un seuil.
- 3- répéter 2 pour le reste des **nœuds** du réseau jusqu'à l'atteinte du but.

Notre prototype a la capacité de trier et classer le résultat de la recherche, selon leurs pertinences par apport à la requête par ordre croissant.

➤ Recherche par structure :

La particularité de cette recherche est que, dans la structure du réseau bayésien, la requête peut conditionner n'importe quel **nœud** de notre réseau bayésien.

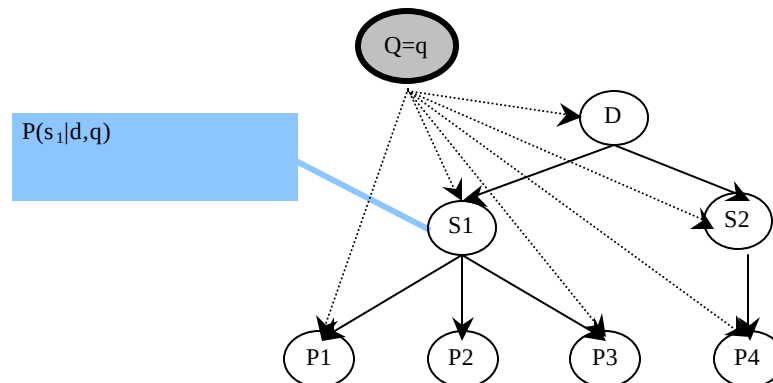


Figure 22: Exemple de recherche par structure

L'algorithme de recherche que nous proposons est comme suit :

- définissez les **nœuds** , liés à l'élément structurel recherché ;
- faire la recherche du terme dans les matrices associées à chaque **nœud** .

Si le terme recherché figure dans les matrices, associées à chaque **nœud**, alors sélectionner les **nœuds** répondant mieux à la requête. Le résultat est un ensemble de **nœuds** , triés selon la probabilité conditionnelle. Sélectionner les **nœuds** qui ont le maximum de probabilité , connaissant la requête, sinon aucun **nœud** ne répond à la requête .

Nous illustrons notre modélisation par l'exemple suivant :

3. Exemple:

<Document>

<Titre Document> Utilisation des Réseaux Bayésiens pour la recherche d'information structurée </Titre Document>

<Chapitre>

<Titre chapitre> Système de recherche d'information </Titre chapitre>

<Introduction> Devant la masse croissante d'information et de données accessibles de nombreuses recherches ont été mise en place au niveau de la recherche automatique d'information. Le but de ces recherches étant de mettre à disposition des demandeurs d'information, des outils et méthodes capables de trouver à partir d'un lot important de documents présents dans une ou plusieurs bases de données, quelques documents pertinents traitant d'un sujet donné. C'est abouti à la création de logiciels et d'outils de recherche documentaire. L'objectif des logiciels documentaires, regroupés sous la dénomination plus générique de Système de Recherche d'Information (SRI), est de proposer des outils et des mécanismes assurant la recherche automatique des informations documentaires. Un système de recherche d'information doit permettre à un utilisateur de retrouver les informations qui l'intéressent parmi un ensemble d'informations. Un tel système doit donc posséder des fonctions de mémorisation et d'organisation des informations ainsi que des mécanismes de recherche et de consultation de ces informations. Les systèmes de recherches d'informations assurent donc le stockage et permettent la recherche et consultation d'information bibliographiques ou textuelles </Introduction>

<Paragraphe>

<Titre paragraphe> Définition d'un Système de recherche d'information </titre paragraphe>

<PCDATA> Un système de recherche d'information est un système qui permet de retrouver les documents pertinents à une requête d'utilisateur, à partir d'une base de documents volumineuse. Dans cette définition, il y a trois notions clés: documents, requête, pertinence. Document: Un document peut être un texte, un morceau de texte, une page Web, une image, une bande vidéo, etc. On appelle document toute unité qui peut constituer une réponse à une requête d'utilisateur. Dans notre étude, nous traitons seulement des documents textuels.

Requête: Une requête exprime le besoin d'information d'un utilisateur. Elle est en général de la forme suivante: "Trouvez les documents qui ...".

Pertinence: Le but de la recherche d'information est de trouver seulement les documents pertinents. La notion de pertinence est très complexe. De façon générale, dans document pertinent, l'utilisateur doit pouvoir trouver les informations dont il a besoin. C'est sur cette notion de pertinence que le système doit juger si un document doit être donné à l'utilisateur comme réponse </PCDATA>

</Paragraphe>

<Paragraphe>

<Titre paragraphe> L'indexation </Titre paragraphe>

<PCDATA> L'indexation est le processus destiné à représenter par les éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou d'une question. On désigne également ainsi le résultat de cette opération. L'indexation consiste donc, d'une part, à identifier les informations caractéristiques du contenu d'un ou plusieurs textes. D'autre part, il s'agit de représenter ces informations sous une forme compacte, homogène et manipulable, c'est-à-dire utilisable par un système de recherche d'information. On distingue deux types d'indexation: l'indexation manuelle, l'indexation automatique </PCDATA>

</Paragraphe>

<Paragraphe>

<Titre paragraphe> L'interrogation </Titre paragraphe>

<PCDATA> Nous introduirons deux dimensions de l'interrogation :

- La possibilité d'énoncer des requêtes à partir de la structure des documents. Il s'agit ici d'exprimer des requêtes telles que : quels sont les documents qui contiennent l'élément x? Combien de documents contiennent l'élément x ? et des combinaisons complexes de ces requêtes ;
- La possibilité d'énoncer des requêtes à partir des mots clés des documents.

Idéalement, ces deux dimensions sont combinées. Nous allons donc considérer les différents langages d'interrogation au regard de ces deux dimensions. Nous distinguons quatre types de langage d'interrogation: les langages booléens, les langages graphiques, les langages libres, le langage naturel </PCDATA>

</Paragraphe>

<Paragraphe>

<Titre paragraphe> Les modèles de recherche d'information </Titre paragraphe>

<PCDATA> Différents modèles de comparaison dans les systèmes de recherche d'information ont été proposés. Ceux-ci peuvent être regroupés en trois grandes catégories :

- Le modèle booléen qui utilise un calcul d'appariement exact entre la représentation de la requête et celle des documents.
- Le modèle vectoriel qui évalue une ressemblance entre la représentation vectorielle de la requête et celle des documents.
- **Le modèle probabiliste qui procède au calcul de la probabilité de la pertinence d'un document par rapport à la représentation de la requête** </PCDATA>

</Paragraphe>

</Chapitre>

<Chapitre>

<Titre chapitre>Système de recherche d'information structurée</Titre chapitre>

<Introduction>Bien que les systèmes de recherche d'information traditionnels soient d'un grand apport dans le domaine de recherche d'information, ces derniers restent insuffisants quant aux besoins et exigences actuelles des utilisateurs, imposés par le développement remarquable de la technologie d'une part et une gestion efficace de l'information qui nécessite plus que jamais sa représentation dans un format adéquat, préservant son indépendance vis-à-vis de toute plate-forme et/ou architecture, d'autre part. Dans ce type de systèmes un document est considéré comme une entité atomique qui est représentée et retrouvée comme un tout par le système et donc c'est document entier qui est présenté à l'utilisateur comme réponse à sa requête. Cependant, les documents affichent souvent une structure définie par l'auteur. Par exemple, un document peut avoir plusieurs sections, chacune avec plusieurs sous sections et ainsi de suite. Avec un document structuré, l'unité de recherche et de présentation, devrait être les composants du document au lieu du document, car souvent seulement des parties du document sont appropriées à un besoin d'information, et par conséquent devront être retournées à l'utilisateur. D'ailleurs, la représentation d'un document structuré doit permettre au processus de recherche de retourner certains composants, par exemple, une section, un ensemble de requête au lieu de fournir de document entier. Par conséquent, produire des facilités de recherche selon des aspects structurels est un problème crucial pour les domaines des bases de données et de la recherche d'informations structurées. C'est pourquoi une nouvelle génération de SRI est entrain de voir le jour, ce sont les SRI structurées </Introduction>

<Paragraphe>

<Titre paragraphe>La Recherche d'Information Structurée (RIS)</Titre paragraphe>

<PCDATA>L'émergence de documents structurés, nécessite de proposer des modèles de recherche d'information et de navigation capables de fusionner des informations provenant de différents médias. L'objectif est de permettre à un utilisateur de retrouver dans une base de documents structurés les documents ou les parties de documents qui répondent à son besoin d'information. Cette fonctionnalité devient nécessaire dans la mesure où il est courant de trouver des environnements de production de documents qui s'articulent autour de leur structure logique. Un modèle de documents structurés permet d'accueillir des documents composés de textes et d'images fixes. Les relations prises en compte sont la relation de composition (agrégation des parties de documents), la relation de séquence (définissant le sens de lecture d'un document), et la relation de référence (à l'intérieur d'un document ou vers d'autres documents). Le modèle de document structuré fournit un squelette sur lequel des opérateurs de recherche structurelle sont définis pour le processus d'interrogation. L'indexation structurelle permet selon la nature des attributs (attributs de forme ou attributs de contenu) de rendre explicites toutes les informations nécessaires à la recherche d'information. Il s'agit de définir la portée d'un attribut et ses caractéristiques afin de contrôler la propagation de l'information via les relations de structure. Le processus d'interrogation et de traitement de requêtes utilise les dépendances entre valeurs d'attributs pour retrouver de manière efficace le ou les éléments de structure pertinents des documents structurés. Pour réaliser ces objectifs de représentations des documents structurés et de dépendances entre termes, les réseaux bayésiens nous ont paru appropriée. Nous allons présenter dans ce qui suit les réseaux bayésiens, leur fonctionnement et leurs intérêts à la recherche d'information structurée </PCDATA>

</Paragraphe>

<Paragraphe>

<Titre paragraphe>XML et la Recherche d'Information</Titre paragraphe>

<PCDATA>Un document en texte brut n'est pas agréable à lire, car on ne peut identifier immédiatement les différentes catégories d'information présentes. Lorsque ce document est mis en forme et respecte une mise en page précise et claire, l'œil humain distingue le titre, l'auteur, etc. Mais cette mise en forme « superficielle » n'est toujours pas exploitable par la machine. XML apporte la solution en structurant l'information. Cette structure exprime une sémantique ouvrant la possibilité de recherche très fine. Au-delà de la recherche plein-texte par exemple : Quels documents contiennent les mots « XML » et « database » ? on peut adjoindre un

critère de recherche structurel : Quels documents contiennent « XML » et « database », dans le titre ou le résumé ? La structure peut également être exploitée afin de procéder à une extraction fine d'information. Par exemple : récupérer le troisième chapitre du document X, ou encore, combiné à une recherche : obtenir les résumés des documents contenant « XML » dans le titre, etc. XML est un format auto-descriptif, ce qui assure la pérennité des données. Compréhensible par l'homme comme par la machine, il est simple d'utilisation. Mais au-delà de ces aspects, il apporte deux types d'avancées : d'une part, il étend les horizons de la gestion de documents, et enrichit d'autre part les possibilités de la gestion de données classique

</Paragraphe>

</Chapitre>

<Chapitre>

<Titre chapitre> Etude des Réseaux Bayésiens </Titre chapitre>

<Introduction> Les réseaux bayésiens (RB) font partie de la famille des modèles graphiques. Ils regroupent au sein d'un même formalisme la théorie des graphes et celle des probabilités afin de fournir des outils efficaces autant qu'intuitifs pour représenter une distribution de probabilités jointe sur un ensemble de variables aléatoires. Ce formalisme très puissant permet une représentation intuitive de la connaissance sur un domaine d'application donné et facilite la mise en place de modèles performants et clairs </Introduction>

<Section>

<Titre section> Définitions des Réseaux Bayésiens </Titre section>

<Paragraphe>

<Titre paragraphe> Définitions des Réseaux Bayésiens </Titre paragraphe>

<PCDATA> Les Réseaux Bayésiens sont des graphes acycliques orientés. Les noeuds de ces graphes représentent des variables aléatoires qui peuvent avoir un nombre discret d'états possibles ou qui peuvent être continues selon une loi de Gauss. Les arcs orientés du réseau représentent des relations de causalité et donc de dépendance entre les variables associées aux noeuds ainsi reliés. La figure 1 ci-après est une illustration des RB. Les noeuds X_i sont associés à des variables aléatoires définies par des observations et les noeuds H_i représentent des hypothèses. L'utilisation de tels RB nécessite la spécification complète du modèle probabiliste. Pour chaque relation de dépendance entre plusieurs noeuds, il faut en effet définir les Tables de Probabilités Conditionnelles </PCDATA>

</Paragraphe>

</Section>

<Section>

<Titre section> Mise en œuvre des réseaux bayésiens </Titre section>

<Paragraphe>

<Titre paragraphe> Pourquoi utiliser des réseaux bayésiens </Titre paragraphe>

<PCDATA> Naturellement, le choix d'une méthode fait intervenir différents critères, comme la facilité, le coût et le délai de mise en œuvre d'une solution. En dehors de toute considération théorique, les aspects suivants des RB les rendent, dans de nombreux cas, préférables à d'autres modèles : Acquisition des connaissances, Représentation des connaissances, Utilisation des connaissances, Qualité de l'offre en matière de logiciels </PCDATA>

</Paragraphe>

<Paragraphe>

<Titre paragraphe> Où utiliser des réseaux bayésiens ? </Titre paragraphe>

<PCDATA> Toute application mettant en œuvre des connaissances peut donc relever de l'utilisation des réseaux bayésiens, qu'il s'agisse de formaliser la connaissance d'experts, d'extraire la connaissance contenue dans des bases de données, ou d'utiliser le plus rationnellement possible l'un ou l'autre type de connaissance. Un réseau bayésien est un moyen de représenter la connaissance d'un système. Une telle représentation n'est bien entendu pas une fin en soi ; elle s'effectue, selon les contextes, dans le but de :

- prévoir le comportement du système
- diagnostiquer les causes d'un phénomène observées dans le système
- contrôler le comportement du système
- simuler le comportement du système
- analyser des données relatives au système
- prendre des décisions concernant le système

</PCDATA>

</Paragraphe>

<Paragraphe>

<Titre paragraphe> Comment utiliser des réseaux bayésiens ? </Titre paragraphe>

<PCDATA> Malgré la diversité des applications, la construction d'un réseau bayésien se réalise, schématiquement, en trois étapes essentielles :

-Identification des variables et de leur espace d'états

-Définition de la structure du réseau bayésien

-Loi de probabilité conjointe des variables

</PCDATA>

</Paragraphe>

</Section>

</Chapitre>

</Document>

La structure de réseau bayésien correspond à cet exemple de document XML est comme suit :

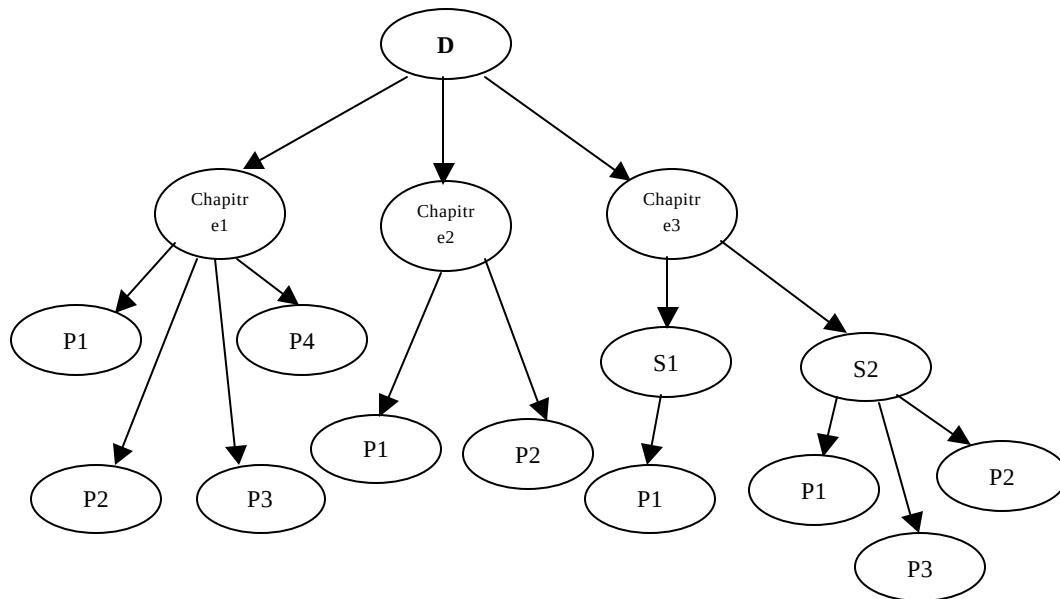


Figure 21 : Structure du réseau bayésien de l'exemple

La phase d'indexation :

L'étape d'extraction des mots donne les résultats suivants :

Pour le ~~not~~ chapitre1 nous avons le titre, l'introduction et quatre paragraphes. Nous avons fixé un seuil de cinq mots. Les mots extraits sont:

- ✓ Titre Chapitre : système, recherche, d'information.
- ✓ Introduction : recherche, d'information, information, système, documentaire.
- ✓ Titre Paragraphe1 (TP1) : définition, système, recherche, d'un, d'information.
- ✓ PCDATA paragraphe1 (P1) : utilisateur, requête, document, pertinent, système, d'information.
- ✓ Titre Paragraphe2 (TP2) : l'indexation.

- ✓ PCDATA paragraphe2 (P2): indexation, information, contenu, représente, l'indexation.
- ✓ Titre Paragraphe3 (TP3) : l'interrogation.
- ✓ PCDATA paragraphe3 (P3) : langage, requête, document, dimension, l'interrogation.
- ✓ Titre Paragraphe4 (TP4) : modèles, recherche, d'information
- ✓ PCDATA paragraphe4 (P4) : modèle, document, requête, représentation, calcul

Pour le ~~no~~d chapitre2 nous avons le titre, l'introduction et deux paragraphes. Nous avons fixé un seuil de cinq mots. Les mots extraits sont:

- ✓ Titre Chapitre : système, recherche, d'information, structurée.
- ✓ Introduction : recherche, d'information, document, structuré, système.
- ✓ Titre Paragraphe1 (TP1) : recherche, d'information, structurée.
- ✓ PCDATA paragraphe1 (P1) : information, document, structuré, attribut, relation.
- ✓ Titre Paragraphe2 (TP2) : XML, recherche, d'information.
- ✓ PCDATA paragraphe2 (P2): document, XML, recherche, peut, titre.
- ✓ Titre Paragraphe3 (TP3) : l'interrogation.

Pour le ~~no~~d chapitre3 nous avons le titre, l'introduction et quatre paragraphes. Nous avons fixé un seuil de cinq mots. Les mots extraits sont:

- ✓ Titre Chapitre : étude, réseaux, bayésiens.
- ✓ Introduction : modèle, formalisme, probabilité, réseaux, bayésiens.
- ✓ Titre Paragraphe1 (TP1) : définition, réseaux, bayésiens.
- ✓ PCDATA paragraphe1 (P1) : variable, représentent, ~~no~~ds, RB, aléatoire .
- ✓ Titre section : mise, ~~œ~~vre, réseaux, bayésiens .
- ✓ Titre Paragraphe1 (TP1) : utiliser, réseaux, bayésiens.
- ✓ PCDATA paragraphe1 (P1): connaissance, RB, représentation, utilisation, acquisition.
- ✓ Titre Paragraphe2 (TP2) :
- ✓ PCDATA paragraphe2 (P2): connaissance, système, comportement, réseaux, bayésiens.
- ✓ Titre Paragraphe3 (TP3) : utiliser, réseaux, bayésiens.
- ✓ PCDATA paragraphe3 (P3): réseau, bayésien, identification, définition, probabilité.

Les probabilités conditionnelles (poids) des termes (ou mots) de chaque nœud sont calculées comme suit :

Chapitre1 :

Probabilité Termes	P(Tit=ti)	P(Int=ti)	P(tp1=ti)	P(p1=ti)	P(tp2=ti)	P(p2=ti)	P(tp3=ti)	P(p3=ti)	P(tp4=ti)	P(p4=ti)	P(Ch1=ti)
système	8*(1/3)	4*(05/97)	2*(1/4)	03/90							33/332
recherche	8*(1/3)	4*(11/97)	2*(1/4)						2*(1/3)		56/332
d'information	8*(1/3)	4*(09/97)	2*(1/4)						2*(1/3)		48/332
définition			2*(1/4)								02/332
utilisateur		0		05/90							05/332
requête		0		05/90				04/43		03/40	12/332
document		0		11/90				04/43		03/40	18/332
pertinent		0		03/90							03/332
documentaire		4*(03/97)									12/332
information		4*(04/97)				02/51					18/332
indexation					2*(1)	03/51					05/332
contenu						02/51					02/332
représente						02/51					02/332
l'indexation						02/51					02/332
langage								06/43			06/332
dimension								03/43			03/332
l'interrogation							2*(1)	03/43			05/332
modèle									2*(1/3)	04/40	06/332
représentation										03/40	03/332
calcul										02/40	02/332

Probabilité Termes	P(Tit=ti/Ch)	P(Int=ti/Ch)	P(tp1=ti/Ch)	P(p1=ti/Ch)	P(tp2=ti/Ch)	P(p2=ti/Ch)	P(tp3=ti/Ch)	P(p3=ti/CH)	P(tp4=ti/Ch)	P(p4=ti/Ch)
système	08/33	20/33	02/33	03/33						
recherche	08/56	44/56	02/56						02/56	
d'information	08/48	36/48	02/48						02/48	
définition			02/02							
utilisateur		0		05/05						
requête		0		05/12				04/12		03/12
document		0		11/18				04/18		03/18
pertinent		0		03/03						
documentaire		12/12								
information		16/18				02/18				
indexation					02/05	03/05				
contenu						02/02				
représente						02/02				
l'indexation						02/02				
langage								06/06		
dimension								03/03		
l'interrogation							02/05	03/05		
modèle									02/06	04/06
représentation										03/03
calcul										02/02

Chapitre2 :

Probabilité Termes	P(Tit=ti)	P(Int=ti)	P(tp1=ti)	P(p1=ti)	P(tp2=ti)	P(p2=ti)	P(Ch2=ti)
système	8*(1/4)	4*(06/130)					32/400
recherche	8*(1/4)	4*(09/130)	2*(1/3)		2*(1/3)	04/118	52/400
d'information	8*(1/4)	4*(07/130)	2*(1/3)		2*(1/3)		40/400
structuré	8*(1/4)	4*(06/130)	2*(1/3)	07/142			41/400
attribut				05/142			05/400
relation				05/142			05/400
XML		4*(10/130)			2*(1/3)	05/118	47/400
document		0		14/142		07/118	21/400
peut		0				03/118	03/400
titre						03/118	03/400
information				06/142			06/400

Probabilité Termes	P(Tit=ti/Ch)	P(Int=ti/Ch)	P(tp1=ti/Ch)	P(p1=ti/Ch)	P(tp2=ti/Ch)	P(p2=ti/Ch)
système	08/32	24/32				
recherche	08/52	36/52	02/52		02/52	04/52
d'information	08/40	28/40	02/40		02/40	
structuré	08/41	24/41	02/41	07/42		
attribut				05/05		
relation				05/05		
XML		40/49			02/49	05/49
document		0		14/21		07/21
peut		0				03/03
titre						03/03
information				06/06		

Chapitre3 :

Probabilité Termes	P(Tit=ti)	P(Int=ti)	P(tp1=ti)	P(p1=ti)	P(ts=ti)	P(tp1=ti)	P(p2=ti)	P(tp3=ti)	P(p3=ti)	P(tp4=ti)	P(p4=ti)	P(Ch3=ti)
étude	8*(1/3)											08/261
réseaux	8*(1/3)	4*(01/44)	02*(1/3)		02*(1/4)	02*(1/3)		02*(1/3)	02/56	02*(1/3)	02/22	26/261
bayésiens	8*(1/3)	4*(01/44)	02*(1/3)		02*(1/4)	02*(1/3)		02*(1/3)	02/56	02*(1/3)	02/22	26/261
modèle		4*(02/44)										08/261
formalisme		4*(02/44)										08/261
probabilité		4*(02/44)									01/22	09/261
Variable				03/88								03/261
représentent				03/88								03/261
nœuds				03/88								03/261
RB				02/88			02/32					04/261
aléatoire				02/88								02/261
connaissance							03/32		05/56			08/261
représentation							01/32					01/261
utilisation							01/32					01/261
acquisition							01/32					01/261
système									07/56			07/261
comportement									03/56			03/261
identification											01/22	01/261
Définition			02*(1/3)								01/22	03/261
mise					02*(1/4)							02/261
oeuvre					02*(1/4)							02/261
utilisé						02*(1/3)		02*(1/3)		02*(1/3)		06/261

Modélisation de notre prototype de RIS utilisant les RB

Probabilité Termes	P(Tit=ti/Ch)	P(Int=ti/Ch)	P(tp1=ti/Ch)	P(p1=ti/Ch)	P(ts=ti/Ch)	P(tp1=ti/Ch)	P(p2=ti/Ch)	P(tp3=ti/Ch)	P(p3=ti/Ch)	P(tp4=ti/Ch)	P(p4=ti/Ch)
étude	08/08										
réseaux	08/10								02/10		
bayésiens	08/10								02/10		
modèle		08/08									
formalisme		08/08									
probabilité		08/08									01/08
réseau		04/16	02/16		02/16	02/16		02/16		02/16	02/16
bayésien		04/16	02/16		02/16	02/16		02/16		02/16	02/16
Variable				03/03							
représentent				03/03							
nœuds				03/03							
RB				02/04			02/04				
aléatoire				02/02							
connaissance							03/08		05/08		
représentation							01/01				
utilisation							01/01				
acquisition							01/01				
système									07/07		
comportement									03/03		
identification											01/01
Définition			02/03								01/03
mise					02/02						
oeuvre					02/02						
utilisé						02/06		02/06		02/06	

Document :

Probabilité Termes	P(Ch1=ti)	P(Ch1=ti/D)	P(Ch2=ti)	P(Ch2=ti/D)	P(Ch3=ti)	P(Ch3=ti/D)	P(D=ti)
système	33/332	33/72	32/400	32/72	07/261	07/72	72/993
recherche	56/332	56/108	52/400	52/108			108/993
d'information	48/332	48/88	40/400	40/88			88/993
définition	02/332	02/05			03/261	03/05	05/993
utilisateur	05/332	05/05					05/993
requête	12/332	12/12					12/993
document	18/332	18/39	21/400	21/39			39/993
pertinent	03/332	03/03					03/993
documentaire	12/332	12/12					12/993
information	18/332	18/24	06/400	06/24			24/993
indexation	05/332	05/05					05/993
contenu	02/332	02/02					02/993
représente	02/332	02/02					02/993
l'indexation	02/332	02/02					02/993
langage	06/332	06/06					06/993
dimension	03/332	03/03					03/993
l'interrogation	05/332	05/05					05/993
modèle	06/332	06/14			08/261	08/14	14/993
représentation	03/332	03/04			01/261	01/04	04/993
calcul	02/332	02/02					02/993
structuré			41/400	41/41			41/993
attribut			05/400	05/05			05/993
relation			05/400	05/05			05/993
XML			47/400	47/47			47/993
peut			03/400	03/03			03/993
titre			03/400	03/03			03/993
étude					08/261	08/08	08/993
réseaux					26/261	26/26	26/993
bayésiens					26/261	26/26	26/993
formalisme					08/261	08/08	08/993
probabilité					09/261	09/09	09/993
variable					03/261	03/03	03/993
représentent					03/261	03/03	03/993
noms					03/261	03/03	03/993
RB					04/261	04/04	04/993

aléatoire					02/261	02/02	02/993
connaissance					08/261	08/08	08/993
utilisation					01/261	01/01	01/993
acquisition					01/261	01/01	01/993
comportement					03/261	03/03	03/993
identification					01/261	01/01	01/993
mise					02/261	02/02	02/993
oeuvre					02/261	02/02	02/993
utilisé					06/261	06/06	06/993

La phase de recherche :

↳ Recherche par Contenu

Supposons que nous avons les requêtes suivantes :

q1 : recherche d'information structurée

q2 : représentation de connaissance

Nous cherchons les termes de la requête q1 et la requête q2 dans la matrice associée au **noeud** racine document.

Les termes de la requête q1 sont : (recherche) et (d'information) et (structurée). Ces termes existent dans la matrice associée au **noeud** document. La réponse est donc :

Titre Document : Utilisation des Réseaux Bayésiens pour la recherche d'information structurée

Pour la requête q2 les termes sont : (représentation) et (connaissance). Ces termes existent dans la matrice associée au **noeud** document. La réponse est donc :

Titre Document : Utilisation des Réseaux Bayésiens pour la recherche d'information structurée

↳ Recherche par Structure

Supposons que nous avons les requêtes suivantes :

q1 : Paragraphe = système de recherche d'information

q2 : Chapitre = réseaux bayésiens

Nous cherchons les termes de la requête q1 et la requête q2 dans la matrice associée au **noeud** racine document.

Les termes de la requête q1 sont : (système) et (recherche) et (d'information). Ces termes existent dans la matrice associée au **noeud** document. Nous cherchons maintenant le chapitre le plus pertinent. Nous avons :

	Chapitre1	Chapitre2	Chapitre3
système	33/72	32/72	07/261
recherche	56/108	52/108	0
d'information	48/88	40/88	0

Nous remarquons que le maximum de la probabilité conditionnelle est celle du chapitre1. Dans ce chapitre nous cherchons le paragraphe le plus pertinent et nous avons donc :

	Titre Paragraphe1	Paragraphe1	Titre Paragraphe4	Paragraphe4
Système	02*(1/4)	03/90		
recherche	02*(1/4)		02*(1/3)	
d'information	02*(1/4)		02/*(1/3)	

Nous remarquons que le maximum de la probabilité conditionnelle est celle du Paragraphe1. La réponse est donc :

Titre Paragraphe : Définition d'un système de recherche d'information

Pour la requête q2 les termes sont : (réseaux) et (bayésiens). Ces termes existent dans la matrice associée au ~~not~~ document. Nous cherchons maintenant le chapitre le plus pertinent. Nous avons :

	Chapitre1	Chapitre2	Chapitre3
réseaux	0	0	26/26
Bayésiens	0	0	26/26

Nous remarquons que les termes de la requête q2 ne figurent que dans le chapitre3. La réponse est donc :

Titre Chapitre : Etude des réseaux bayésiens

Conclusion

Une information peut être disponible mais inaccessible, étant donnée la diversification des disciplines et les grosses masses de documents qui augmentent chaque jour. Par conséquent, une personne, qui cherche une information, doit avoir, à sa disposition, des outils puissants qui permettent l'accès aux données désirées, dans des délais raisonnables.

Nous avons proposé un prototype de recherche d'information structurée. Pour cela, nous nous sommes basés sur une modélisation probabiliste qui est celle des réseaux bayésiens. Nous avons introduit, dans cette partie, le modèle de recherche d'information qui est au coeur de notre prototype de système de RIS. Ce formalisme permet, de façon assez naturelle, de traiter des structures, et permet de faire des inférences sur les différents éléments de la structure. Notre modèle est capable de répondre à des questions portant à la fois sur la structure et le contenu.

Dans cette partie, nous avons justifié notre choix pour ce type de modèle, décrit le modèle spécifique que nous avons utilisé pour représenter les documents structurés (et plus généralement un corpus de documents structurés), et proposé un algorithme d'indexation pour ce corpus.

Nous avons présenté un algorithme de recherche par le contenu et un autre pour la recherche par structure, et montré comment ce formalisme permet de décrire le processus de la pertinence. La recherche peut être alors de granularité plus fine.

Introduction

Dans cette partie nous présentons les principales interfaces et états de sorties de notre prototype de recherche d'information structurée utilisant les réseaux bayésiens.

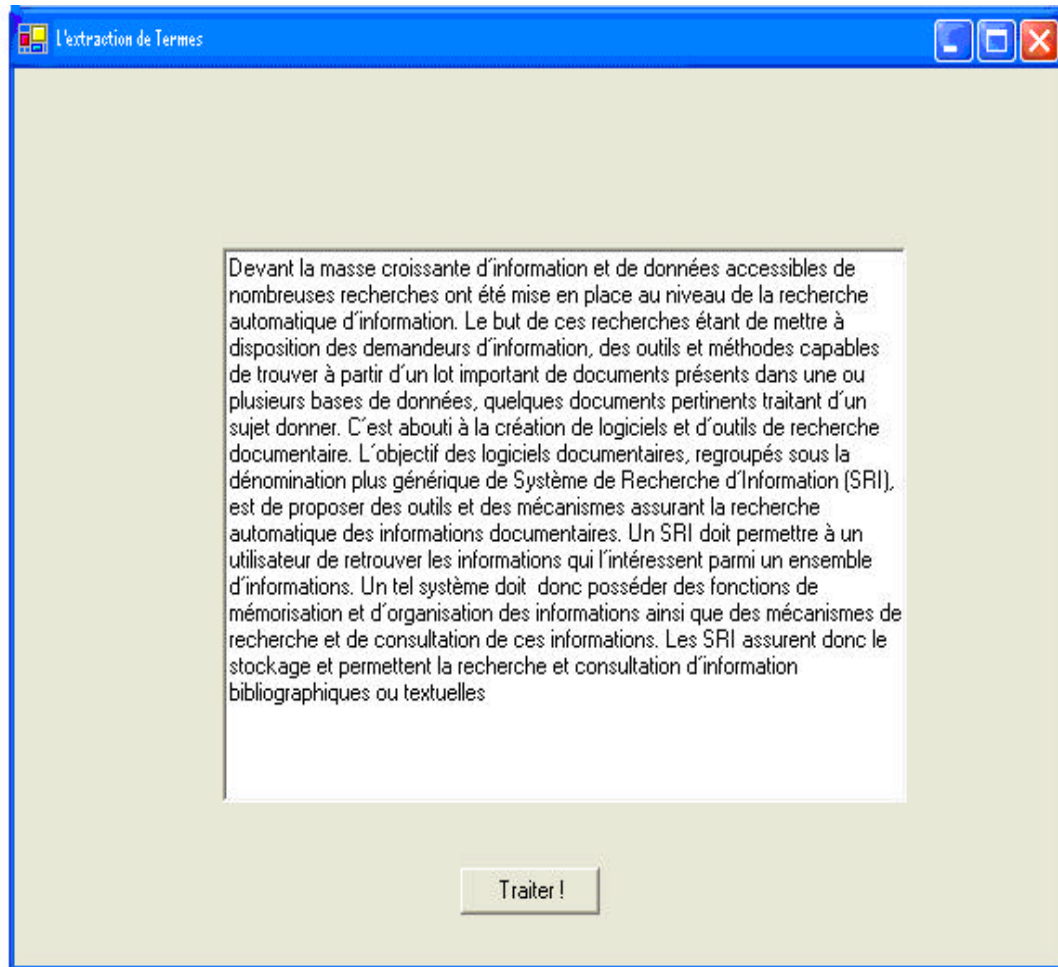
Pour l'étape de l'indexation ou l'extraction de termes, nous avons utilisé un service web qui nous a permis d'extraire les termes pour chaque nœud de notre réseau bayésien et de calculer leurs poids.

Nous avons modifié et utilisé ce service web comme une application séparée de notre prototype. Nous avons stocké le résultat de l'extraction des termes dans une base de données Access.

Notre prototype de recherche d'information structurée est capable de répondre à des questions portant sur le contenu et la structure.

Nous présentons une interface qui contient deux types de recherche : recherche par contenu (recherche simple), et recherche par structure. Ensuite nous présentons un état de sortie pour une recherche par structure, et une autre pour une recherche simple.

Cette forme représente l'interface de l'extraction de termes qui permet de saisir un texte et de stocker le résultat de l'extraction dans une base de donnée.



Le résultat de l'extraction des termes pour chaque **noël** est stocké dans une base de donnée Access. Ce résultat est représenté par : le terme extrait, terme lemmatisé et son poids (nombre d'occurrences sur le nombre total de termes dans le **noël**)

Fichier Edition Affichage Insertion Format Enregistrements Outils Fenêtre ?			
mot_lem	fléchié	poids	
sujet	sujet	1,10E-02	
donner	donner	1,10E-02	
C'est	C'est	1,10E-02	
abouti	abouti	1,10E-02	
création	création	1,10E-02	
logiciel	logiciels	2,20E-02	
d'outil	d'outils	1,10E-02	
recherche	recherche	7,69E-02	
documentaire	documentaire	3,30E-02	
L'objectif	L'objectif	1,10E-02	
logiciel	logiciels	2,20E-02	
documentaire	documentaires	3,30E-02	
regroupé	regroupés	1,10E-02	
dénomination	dénomination	1,10E-02	
générique	générique	1,10E-02	
Système	Système	1,10E-02	
Recherche	Recherche	1,10E-02	
d'Information	d'Information	1,10E-02	
SRI	SRI	3,30E-02	
proposer	proposer	1,10E-02	
outil	outils	2,20E-02	
mécanisme	mécanismes	2,20E-02	
assurant	assurant	1,10E-02	
recherche	recherche	7,69E-02	

Enr : 4 sur 90

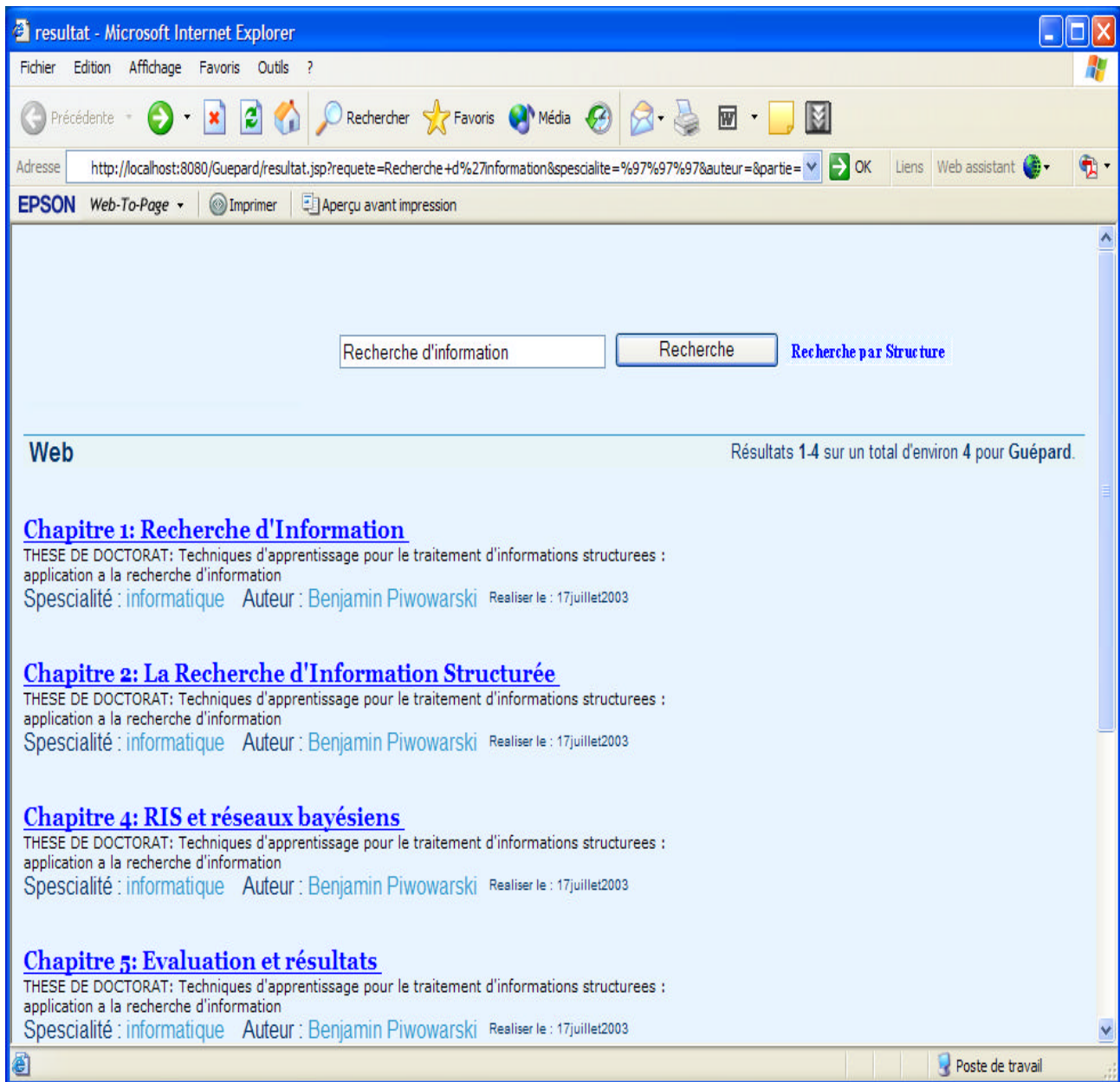
Cette interface représente les deux types de recherches : recherche par le contenu (recherche simple) et recherche par structure.

The screenshot shows a Microsoft Internet Explorer window titled "recherche_simple - Microsoft Internet Explorer". The address bar displays "D:\Documents and Settings\sam\Bureau\la file\recherche_simple.html". The main content area has a light blue background and contains a search interface with the following elements:

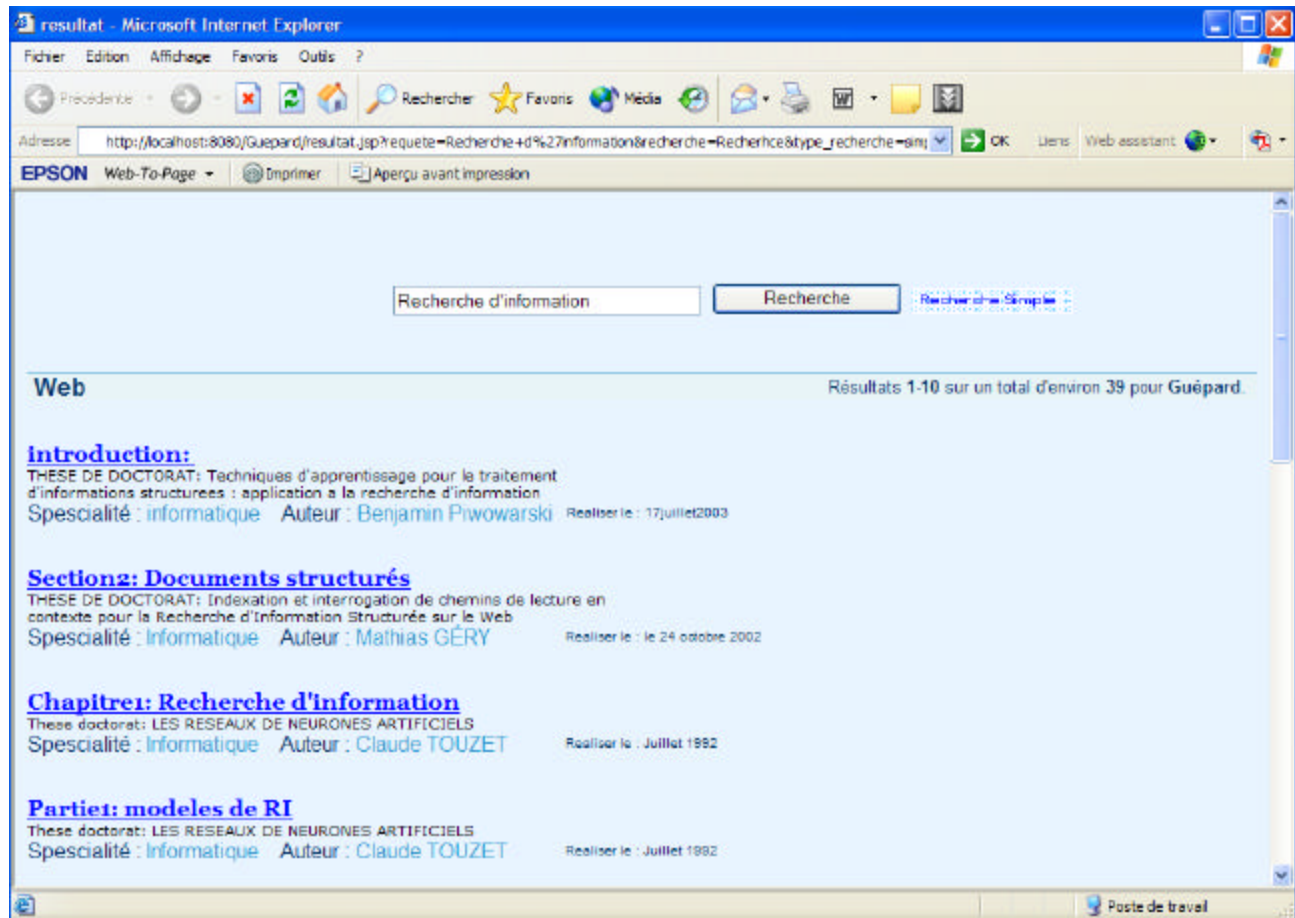
- A top section with two buttons: "Recherche d'information" and "Recherche Simple".
- A section labeled "Saisie la Requete:" with a text input field containing "Recherche d'information".
- A section labeled "Selectionnée la structure :" with a dropdown menu showing "Chapitre".
- A button labeled "Recherche par Structure".
- A bottom section with a button labeled "Administrateur".

The taskbar at the bottom shows the "Poste de travail" icon and the text "Poste de travail".

Cette interface représente le résultat d'une recherche par structure. Dans cette forme la structure demandée est le chapitre.



Cette interface représente le résultat d'une recherche par le contenu (recherche simple).



Conclusion Générale

Dans ce travail, nous nous sommes intéressés au problème de la recherche d'information pour les documents structurés, en particulier les documents XML. L'objectif de notre travail est de permettre à un utilisateur de retrouver dans une base de documents structurés, les documents ou les parties de documents qui répondent à son besoin d'information. Cette fonctionnalité devient nécessaire, dans la mesure où il est courant de trouver des environnements de production de documents qui s'articulent autour de leur structure logique.

Afin d'atteindre cet objectif, nous avons proposé un prototype de recherche d'information structurée. Pour cela, nous nous sommes basés sur une modélisation probabiliste qui est celle des réseaux bayésiens. Nous avons introduit le modèle de recherche d'information qui est au coeur de notre prototype. Ce formalisme permet, de façon assez naturelle de traiter des structures, et permet de faire des inférences sur les différents éléments de la structure. Notre modèle est capable de répondre à des questions portant à la fois sur la structure et le contenu. La recherche peut être alors de granularité plus fine.

Les perspectives des modèles de recherches d'informations structurées sont nombreuses. Nous décrivons certains aspects, qu'il serait intéressant de considérer, afin d'améliorer notre modèle :

- Nous indiquons des pistes qui pourraient être employées avec un modèle comme le notre, telles que la pondération des termes au sein de la requête et la réduction du taux de silence ; pour cela nous proposons l'utilisation d'une méthode formelle pour l'extraction des termes en tenant compte de l'aspect sémantique.
- Notre modèle de réseau bayésien peut prendre en compte tous les types d'informations écrit sous forme de fichier XML, tel que les documents multimédias.
- Le modèle que nous avons proposé est basé sur la probabilité conditionnelle. Nous pouvons explorer le raisonnement possibiliste dans les réseaux bayésiens et l'appliquer à un modèles de recherche d'information.

1. Annexe A

1. Les Descripteurs

Ils représentent l'information atomique d'un index. Ils sont censés indiquer de quoi parle le document. On parle aussi d'unités élémentaires (en anglais *"tokens"*).

Le but étant de les choisir de manière à ce que l'index (qui réduit la représentation) perde le moins d'information sémantique possible.

Habituellement les descripteurs sont :

Les **mots** du document : toute chaîne de caractères comprise entre deux séparateurs (espace, virgule...)

Les **lemmes** : un processus supplémentaire appelé lemmatiseur, convertit les documents préalablement pour extraire des mots de référence (ainsi les mots "CŒURS" et "CŒUR" aboutiront au même descripteur).

La racinisation (*"stemming"*), quant à elle, est un processus linguistiquement plus simple. Elle consiste, dans sa version la plus simple, à enlever aux mots, les derniers caractères (considérés comme décrivant les flexions de mots), par exemple : Enlever le "s" des mots au pluriel.

Certaines racinisations utilisent des connaissances morphologiques plus complètes (suffixes, préfixes...). Au lieu d'indexer un texte par des mots, nous l'indexons alors par le lemme correspondant.

Les **concepts**, termes ou mots-clés : il s'agit d'expressions (pouvant contenir un ou plusieurs mots). Ces concepts sont, le plus souvent, entrés manuellement. Ces concepts peuvent être écrits de manière libre par un utilisateur, ou doivent être choisis parmi une liste de concepts. Cette liste de concepts sera, le plus souvent, décrite dans un thésaurus.

Plus rarement :

Les **N-grammes** : Il s'agit d'une représentation originale d'un texte en séquences de *N* caractères consécutifs. On trouve des utilisations de bigrammes et trigrammes dans la recherche documentaire (ils permettent de reconnaître des mots, de manière approximative, et ainsi de corriger des flexions de mots ou même des fautes de frappe ou d'orthographe). Ils sont aussi fréquemment utilisés dans la reconnaissance de la langue d'un texte.

Les **contextes** : les documents et leurs mots sont représentés sur d'autres dimensions, où les mots, apparaissant dans un même contexte, sont proches. Cette indexation est le résultat d'une analyse des cooccurrences des mots dans un corpus.

Voici quelques exemples de résultats d'indexation sur un document ne contenant que la phrase "Les accidents vasculaires cérébraux" :

Par **mots** : "Les", "accidents", "vasculaires", "cérébraux"

Par **lemmes** : "le", "accident", "vasculaire", "cérébral"

Par **racines** : "l", "accident", "vascul", "cérébr"

Par **concepts** : "A.V.C."

Par **bigrammes** : "_l", "le", "es", "s_", "_a", "ac", "cc", "ci", "id", "de", "en", "nt", "ts", "s_", "_v", "va", "as" ... "au", "ux", "x_" [POUL02].

2. Le Thésaurus

Le thesaurus est un vocabulaire, contrôlé et dynamique, de termes ayant entre eux des relations sémantiques et génériques et qui s'applique à un domaine particulier de la connaissance (Définition de l'ISO).

Du point de vue de sa fonction, il s'agit d'un instrument de contrôle de la terminologie, basé sur les théories de la cognition et de la linguistique, et utilisé pour transposer, en un langage plus strict, le langage naturel, employé dans les documents (ex. contrôle des synonymes).

Les relations, gérées par un thesaurus, servent à isoler le ou les termes adéquats pour indexer un document ou formuler une recherche. Elles sont de trois types:

- Hiérarchique entre un descripteur et son terme spécifique (ex : "clavier" est un terme spécifique de "matériel informatique") ;
- D'équivalence entre un descripteur et ses synonymes (ex : "éducation" et "enseignement").
- D'association entre 2 termes voisins mais appartenant à des domaines différents (ex : "enseignement" et "formation professionnelle").

De la pertinence du thesaurus et de son adéquation au domaine couvert dépend l'efficacité des recherches. Afin de garantir cette pertinence, le suivi du thesaurus doit être assuré afin d'effectuer les mises à jour, si nécessaires :

- Edition de listes de descripteurs avec compteur (fréquence d'utilisation à l'indexation).
- Suppression d'un terme du thesaurus sur les fiches indexées.
- Remplacement d'un terme du thesaurus par un autre.
- Comptage du nombre d'utilisations de chaque descripteur en recherche.

3. Les différentes approches pour la Recherche d'Informations

La plupart des systèmes de recherche d'informations commercialisés sont basés sur l'approche classique. Ils évaluent la similarité des représentations des informations, celles-ci étant sous forme de mots-clés. Pour être efficace, cette approche suppose que la requête et les documents pertinents pour cette requête aient une représentation similaire, c'est-à-dire, un grand nombre de termes descriptifs de leur contenu commun. Ces systèmes ont des performances parfois insatisfaisantes. En particulier, lorsque un même concept est décrit par des termes différents dans la requête et dans les documents, des documents pertinents ne seront pas restitués. Si par contre la requête utilise des termes trop génériques ou indexant beaucoup de documents, la taille de la réponse est trop importante pour être exploitable par l'utilisateur ou la réponse contient des documents non pertinents.

D'autres approches ont donc été développées pour résoudre ces problèmes.

Un premier type de mécanisme tente de rapprocher le processus de consultation du processus mental de l'utilisateur : la consultation se fait par associations d'idées.

L'utilisateur peut :

- consulter l'ensemble des documents qui ont une certaine similitude entre eux (mécanismes de classification),
- naviguer dans la base à partir d'un document pertinent et en suivant des liens de référence vers d'autres documents (mécanismes de navigation). Généralement, ce type de processus est assuré par des associations préétablies entre les documents.

Une autre approche consiste à reformuler la requête exprimée par l'utilisateur en une requête qui assurera une recherche plus efficace. La reformulation est basée à la fois sur le besoin d'information de l'utilisateur et sur le contenu de la base. On cherchera en particulier à fournir à l'utilisateur des documents ressemblant à certains documents qu'il juge pertinents (mécanisme de réinjection de la pertinence) ou à modifier la requête en fonction de connaissances sur les termes utilisés dans la requête (utilisation de combinaisons de termes).

Parallèlement, des modèles assurent une meilleure représentation des documents et des requêtes. La plupart de ces modèles représentent l'information sous forme de réseaux, assurant ainsi l'association d'unités sémantiquement proches. En particulier, des techniques issues de l'Intelligence Artificielle ont été adaptées aux besoins des SRI. Il s'agit, en particulier, des réseaux sémantiques ou des graphes conceptuels, utilisés en compréhension de textes, ou des réseaux inférentiels et des réseaux de neurones [MOTH97].

3.1. Associations de documents

Dans les modèles que nous avons décrits, il n'y a pas d'association explicite entre les documents. Il n'est donc pas possible, à partir d'un document intéressant pour l'utilisateur, d'accéder facilement à des documents traitant des mêmes concepts ou thèmes. Ce problème peut être résolu, dans le modèle vectoriel, par un mécanisme de calcul de ressemblance entre documents ; plus généralement, il peut être résolu en créant des associations entre documents. Le but de ce type d'association est de permettre à l'utilisateur d'accéder, à partir d'un document, à d'autres documents susceptibles de l'intéresser.

On distingue deux techniques assurant les associations entre documents : la classification et le concept hypertexte.

Les techniques de classification regroupent les documents dans des classes, en fonction de leur ressemblance. Etant donné un document, on peut alors facilement retrouver les documents similaires: on peut étendre la réponse à toute la classe à laquelle ce document appartient. On obtient ainsi des documents qui n'auraient pas été retrouvés par une simple comparaison avec la requête.

L'approche hypertexte permet d'associer des documents de façon différente : les documents sont reliés par plusieurs types de liens : des liens de références (par exemple une référence bibliographique vers un autre document), des liens structurels (correspondant à la structure du document d'origine et résultant du découpage de l'information en unités documentaires), etc... La consultation correspond alors à un parcours de la base à partir d'un point d'entrée (document) vers d'autres documents en suivant ces liens pré-établis [MOTH97].

3.1.1. Classification des documents d'une collection

Généralement, les SRI manipulent une masse importante de données : le nombre de documents peut être grand tout comme le nombre de termes du "langage", nécessaire à la représentation de ces documents. Des mécanismes sont donc nécessaires pour simplifier l'accès et la manipulation des données.

Le principe de classification repose sur la ressemblance des documents selon certains critères ; deux documents, suffisamment ressemblants, seront rangés dans une même classe.

On peut distinguer deux types de classification complémentaire :

- le regroupement des documents en base selon des critères factuels (type de document, année d'édition, discipline...)
- le regroupement des documents d'une base selon la ressemblance de leur contenu

Lorsqu'il s'agit du regroupement des documents en base, la classification permet d'une part, d'augmenter la pertinence des réponses fournies, en utilisant un espace de consultation ou éventuellement un vocabulaire ciblé et d'autre part, de diminuer le temps de réponse en restreignant l'espace de consultation. Avant d'interroger le système, on doit alors se positionner sur un groupe de documents.

Le regroupement des documents d'une base en classes permet d'étendre la réponse à des documents proches des documents pertinents ; il limite donc le silence. La classification est également utilisée pour accélérer la recherche : au lieu de comparer la requête à chacun des documents, on la compare uniquement aux classes de documents [MOTH97].

➤ *Regroupement des documents en bases*

Le regroupement des documents en bases peut être considéré comme un premier type de classification. Le rattachement d'un document à une base est alors soit manuel, soit induit par les *données factuelles* qui le caractérisent : le type, la fréquence ou la date de parution, etc.. On pourra par exemple, avoir une base contenant tous les périodiques hebdomadaires et une autre contenant toutes les thèses soutenues après 1980. Dans ce cas, le rattachement d'un document à un groupe est non ambigu, mais il nécessite une connaissance préalable pour l'interrogation. En effet, l'utilisateur doit connaître la base dans laquelle se trouve l'information recherchée. Lorsque la classification en base thématique est manuelle - *basée sur le contenu sémantique* des documents-, la mise en œuvre possède les mêmes inconvénients que l'indexation manuelle en ce qui concerne la mise en œuvre : difficulté du choix des thèmes possibles, subjectivité du choix du rangement d'un document, selon un thème, ...

Les bases, ainsi constituées, n'ont pas de document en commun, ce sont des partitions au sens mathématique. Le thésaurus et l'ensemble des mots vides (ou le lexique) peuvent alors être spécifiques à chaque base.

Cela permet, en particulier, de préserver l'usage spécifique d'un terme dans un contexte. L'avantage principal de ce type de regroupement est que le choix d'une base, par l'utilisateur, correspond à un premier mécanisme de sélection de documents susceptibles d'être pertinents.

Cette présélection, si elle est correcte, permet de limiter le bruit sans pour autant diminuer le nombre de documents pertinents qui seront restitués (le taux de rappel est stable). Cependant, pour pouvoir interroger le système, de façon efficace, l'utilisateur doit connaître le schéma utilisé dans la classification.

➤ *Techniques de classification dans une base*

La classification réalisée sur une base par rapport au contenu des documents est complémentaire à la partition en bases. Ce type de classification est, généralement, réalisé, automatiquement, en tenant compte de la similarité entre les documents. L'interrogation sera, alors, effectuée sur l'ensemble de la base ou sur une partie de celle-ci.

La classification automatique des documents a pour objectif de regrouper les documents d'une base qui ont des contenus similaires. En fait, ce mécanisme est basé sur le calcul de la ressemblance des représentations des documents. Ce calcul tient compte du nombre de termes d'indexation en commun, éventuellement du poids sémantique de ces termes.

Plusieurs mécanismes sont utilisés pour regrouper automatiquement les documents d'une base en classes :

- ✓ rattachement des documents à des pôles attracteurs ou initiateurs de groupes,
- ✓ regroupement des documents en fonction de leur similarité relative,
- ✓ regroupement des documents qui sont pertinents pour une même requête.

Nous allons voir en détail ces différentes techniques.

1ère technique de classification : choix d'initiateurs de groupes

Un premier mécanisme de classification consiste à choisir un initiateur ou attracteur ("seed") pour chaque groupe, par exemple, un document. Un document à classer est alors rattaché au groupe (ou aux groupes) dont il est le plus similaire.

Le choix des attracteurs est primordial pour une bonne classification. Un attracteur correct doit attirer des documents un peu éloignés autour de lui et en même temps, il doit être, le plus possible, séparé des autres attracteurs. Ces attracteurs peuvent être choisis de façon prédéterminée, de façon aléatoire ou selon des algorithmes prenant en compte le contenu de la base [MOTH97].

2ème technique de classification : calcul d'une matrice de similarité des paires de documents Cette technique utilise la théorie des graphes : il s'agit, dans un premier temps, de calculer le coefficient de similarité de toutes les paires de documents, ensuite d'appliquer un seuil de similarité pour déterminer les groupes correspondant aux documents qui sont proches.

Dans le modèle vectoriel, la similitude de deux documents sera, par exemple, calculée par la mesure du cosinus entre les vecteurs correspondant aux documents.

$$\text{Sim}(D_i, D_k) = \frac{\sum_{j=1}^t d_{ij} \cdot d_{kj}}{\left(\sum_{j=1}^t d_{ij}^2 \right)^{1/2} \cdot \left(\sum_{j=1}^t d_{kj}^2 \right)^{1/2}}$$

où :

D_i et D_k sont des documents,

d_{ij} est le poids du terme t_j dans le document D_i ,

t est le nombre de termes d'indexation

Pour un modèle organisé en fichier inverse, un algorithme a été proposé permettant de calculer rapidement les coefficients de similarité des documents (afin de classer ces derniers).

Supposons que le document i soit indexé par les termes t_1 , t_2 et t_3 . Si les listes inverses associées à ces trois termes sont :

	d1	d2	d3	d4	d5	d6	d7	
t_1 :	1	1	0	1	0	1	1	
t_2 :	0	1	0	0	0	1	1	
t_3 :	0	1	0	0	0	0	1	

L'addition de ces trois vecteurs donne

V :	1	3	0	1	0	2	3	
-----	---	---	---	---	---	---	---	------

Ce vecteur permet de déduire que le document i précèdent a un terme en commun avec le document 1, trois termes avec le document 2, zéro avec les documents 3 et 5, etc...

A partir de ces coefficients, on peut appliquer un seuil de similarité, indiquant que deux documents, ayant un coefficient de similarité supérieur au seuil, appartiendront à la même classe (lien simple) ou que tous les documents, appartenant à une même classe, auront, entre eux, des coefficients supérieurs au seuil.

Par exemple, supposons que les coefficients de similarité (normalisés) soient les suivants :

	d1	d2	d3	d4	d5	d6	d7
d1	1	0,5	0,8	0,2	0	0	0
d2	0,5	1	0,35	0,2	0,1	0,15	0,2
d3	0,8	0,35	1	0,6	0,15	0,25	0,15
d4	0,2	0,2	0,6	1	0,1	0,2	0,25
d5	0	0,1	0,15	0,1	1	0,2	0,4
d6	0	0,15	0,25	0,2	0,2	1	0,75
d7	0	0,2	0,15	0,25	0,4	0,75	1

Figure 23: Exemple de coefficients de similarité entre documents

Les classes de documents obtenues seront (figure 28) :

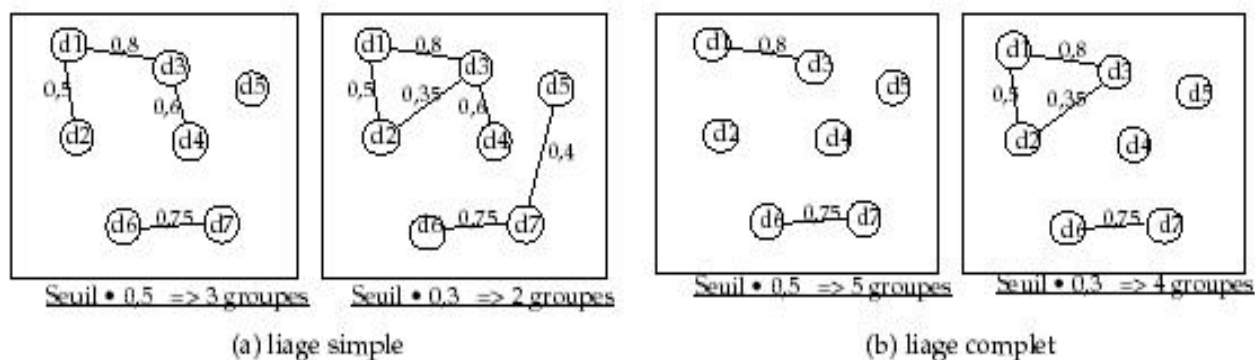


Figure 24: Regroupements de documents par seuillage

Notons que le seuil de similarité permet de décider du nombre de classes ; ce seuil est un paramètre du système qui dépend du contenu de la base. A l'ajout ou à la suppression de documents, la maintenance de la classification, obtenue par liage unique, n'est pas très coûteuse.

Par contre; pour le liage complet, il est nécessaire de connaître les coefficients de similarité avec tous les anciens documents. De plus, il y a un risque de dégénérescence de la classification pour une augmentation (de 25 à 50%) du nombre de documents de la base ; cela est dû au fait que le paramètre seuil est calculé sur la base initiale.

Une classification complète nécessite un algorithme d'ordre N^2 mais le résultat n'est meilleur que dans le cas précédent, puisque tout document est, ici, considéré comme un attracteur potentiel [MOTH97].

3ième technique de classification : prise en compte de la pertinence des documents par rapport à une requête.

La méthode de classification adaptative a été introduite dans certains modèles. Initialement, à chaque document est associée une coordonnée aléatoire sur un axe (-, +). Pour chaque requête, les coordonnées des documents pertinents sont modifiées afin de les rapprocher les uns des autres sur l'axe. Pour s'assurer que tous les documents ne sont pas concentrés, le centroïde de ces documents est éloigné du centroïde de la base complète.

Cette approche a plusieurs inconvénients : les groupes dépendent des requêtes et de leur ordre de traitement. De plus, la définition des groupes, c'est-à-dire, la séparation entre les éléments des différents groupes, nécessite un paramètre qui dépend de la base. Enfin, la convergence peut être longue, voire impossible, en particulier, lorsqu'il y a des jugements de pertinence contradictoires.

Toutes ces techniques, qui permettent d'obtenir une classification à un niveau (classification plane), possèdent un certain nombre d'inconvénients :

- certaines classes comprendront beaucoup de documents et d'autres peu (déséquilibre des classes),
- les paramètres de classification permettent de construire, automatiquement, un grand nombre de (petits) groupes ou alors un petit nombre de (grands) groupes, sans que l'on sache quelle est la solution optimale.

Ces problèmes peuvent, cependant, être résolus par une hiérarchisation de la classification [MOTH97].

➤ **Création de hiérarchie de groupes de documents** [MOTH97]

Le principe de hiérarchisation consiste à créer des sous-groupes de documents, regroupant d'autant plus de documents que leur niveau, dans la hiérarchie, est élevé.

Deux mécanismes sont utilisés pour créer des hiérarchies de classes :

- l'éclatement,
- le regroupement.

✓ **Hiérarchisation par éclatement de groupes**

A partir d'une classification plane des documents, les groupes sont éclatés en sous-groupes lorsque certaines contraintes ne sont plus vérifiées. Les contraintes pourront être par exemples : "un groupe ne doit pas contenir plus de 10 éléments, documents ou sous-groupes", "un document ne doit pas être trop éloigné du centre de gravité de la classe".

La hiérarchie de classes, ainsi définie, suppose qu'une classe contienne l'ensemble des documents, contenus dans les classes hiérarchiquement inférieures ; un document apparaît dans plusieurs classes feuilles (comme dans SMART) ou dans une seule classe feuille.

✓ **Hiérarchisation par regroupements successifs**

Les algorithmes de regroupement hiérarchique par liage unique (single-linkage) fonctionnent de façon différente : ils procèdent par regroupements successifs.

Chaque paire de groupes de documents, construits auparavant, (ou chaque paire de documents à la première itération) est considérée et la paire, qui possède le maximum de similarité, est fusionnée en un seul groupe de documents qui devient un **nœud** de la hiérarchie. Dans ce cas, la similarité est la similarité maximum qui existe entre deux documents, pris dans chacun des deux groupes considérés.

Le calcul de similarité correspond, généralement, à celui de la mesure cosinus, utilisée dans le modèle vectoriel. D'autres méthodes considèrent la similarité minimum (liage complet), la similarité moyenne (liage de moyenne), etc...

Une des difficultés rencontrée, lors de la classification, concerne le choix du nombre de documents par classe et le nombre de classes. Il n'y a aucune théorie permettant de trouver les nombres optimums.

➤ **Utilisation de la classification dans la phase de recherche**

La recherche utilisant la classification peut se faire de plusieurs façons. La manière la plus intuitive est de comparer la requête avec les centroïdes des groupes du haut de la hiérarchie ; et, pour les groupes les plus proches de la requête, de calculer la ressemblance des centroïdes des composantes, avec la requête.

Ce processus est répété en considérant, successivement, des groupes plus petits jusqu'à ce qu'éventuellement les documents individuels, de certains groupes feuilles, soient comparés avec la requête. On procède donc par descente dans la hiérarchie jusqu'au plus bas niveau.

Une autre solution est de comparer, d'abord, la requête aux différents centroïdes des feuilles. Pour les classes dont les centroïdes sont fortement semblables à la requête, celle-ci est comparée aux différents documents qui composent ces classes. On a l'avantage de limiter le nombre de comparaisons nécessaires, tout en permettant, en plus, de localiser, à partir d'un document pertinent, les documents ressemblants et donc susceptibles d'être également pertinents. En fait, soit les classes les plus ressemblantes seront restituées en intégralité, soit seuls les documents les plus ressemblants seront restitués. Dans ce second cas, le système offre la possibilité d'étendre la requête à la classe à laquelle appartient un document qu'il juge pertinent. Ce mécanisme limite donc le temps de recherche par diminution du nombre de comparaisons nécessaires. Il permet également de limiter le silence par l'inclusion à la réponse des documents vraisemblablement pertinents.

Ces documents sont choisis pour leur ressemblance à des documents pertinents (certains n'auraient pas été sélectionnés en utilisant des techniques classiques).

La classification offre, également, la possibilité, à l'utilisateur, de se positionner sur un sous ensemble de la base avant de soumettre sa requête. Ce positionnement sur une classe (éventuellement racine d'une sous hiérarchie de classes) permet de limiter l'espace de consultation, c'est-à-dire le nombre de documents à étudier, tout en ciblant cet espace. Le positionnement dans la hiérarchie de classes a donc deux effets : il diminue le temps de recherche et permet d'augmenter la pertinence des documents [MOTH97].

➤ **But des techniques de classification**

En résumé, les techniques de classification ont deux buts essentiels : limiter, à la fois, le nombre de documents non pertinents restitués -bruit- et le nombre de documents pertinents non restitués -silence-. Le bruit est atténué car le nombre de documents sur lequel s'effectue la recherche est moindre (l'utilisateur choisit un sous-ensemble de documents). Le silence est limité par l'utilisation du mécanisme de voisinage rendu possible grâce à la classification : pour chaque document retrouvé, on inclut les documents proches, qui auraient été oubliés par une approche classique. Ce principe de restitution de documents par associations d'idées est également utilisé dans l'approche hypertexte, via des outils de navigation dans la base [MOTH97].

3.1.2. Techniques de navigation - Modèle hypertexte

A l'origine, l'objectif de l'approche hypertexte est d'étendre les fonctionnalités des systèmes de gestion de fichiers texte traditionnels. De façon classique, les informations textuelles sont considérées comme des informations linéaires et doivent être consultées séquentiellement. Or, les informations textuelles sont généralement des informations non linéaires (les définitions dans un dictionnaire, par exemple, font référence à d'autres définitions).

L'approche hypertexte permet d'exploiter cette caractéristique. La nature associative des informations permet leur structuration qui peut ensuite être exploitée au sein d'un hypertexte [MOTH97].

➤ **Concepts hypertextes**

Dans un système hypertexte, la base d'informations est décomposée en un ensemble de noeuds, représentant chacun un élément d'information, et de liens permettant d'associer ces éléments. Ces liens constituent la caractéristique essentielle des systèmes hypertextes : ce sont eux qui confèrent la structure non linéaire aux documents. Le réseau ainsi formé constitue l'hypergraphe de la base. Les liens sont établis entre les objets de la base, à la fois graphiquement, par l'intermédiaire de zones sensibles et dans la base de données, grâce à des pointeurs. Une fenêtre écran est associée à chaque noeud de la base et possède un titre qui sera visualisé en même temps que la fenêtre. Une fenêtre, outre le titre et le texte lui-même, contiennent un certain nombre de zones sensibles qui représentent des pointeurs vers d'autres noeuds de la base. Ces zones contiennent un libellé renseignant sur le noeud pointé. Le fait de cliquer sur une zone sensible provoque l'activation du lien, c'est-à-dire provoque l'ouverture d'une nouvelle fenêtre correspondant au noeud pointé.

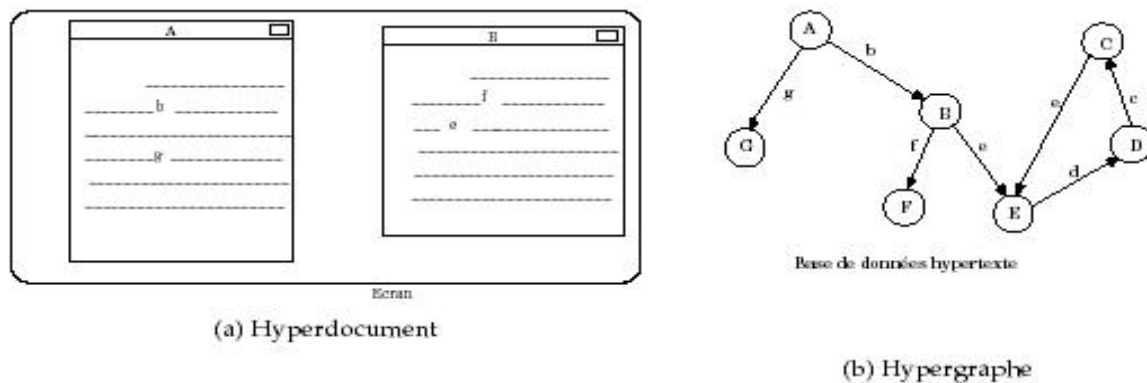


Figure 25: Correspondance entre les fenêtres sur l'écran et les noeuds et liens de la base de données

La consultation d'un hyperdocument peut se faire de plusieurs manières:

- ✓ en suivant les liens et en ouvrant successivement les fenêtres pour lire leur contenu,
- ✓ en formulant une requête par mots clés ou valeurs d'attributs, la réponse donnant les points d'entrée à l'hyperdocument,
- ✓ en utilisant un "browser". Le browser est un outil qui affiche dans une fenêtre la structure de tout ou partie de l'hyperdocument, sous la forme d'un graphe (figure 28) ; il permet la consultation par navigation. Il s'avère intéressant lorsque l'hyperdocument est de taille importante.

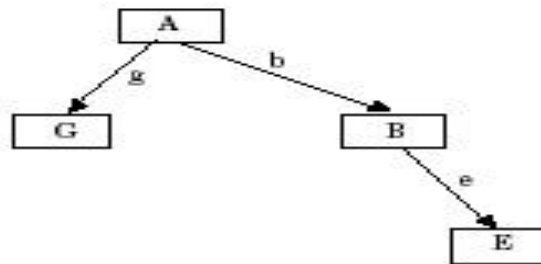


Figure 26: Exemple de "browser" correspondant à la base de données précédente.

Dans certains systèmes, les liens contiennent des indications permettant, en particulier le partage d'informations, la gestion des versions, ou la définition de parcours de consultation permettant de guider l'utilisateur. Le typage des liens est également un élément important puisqu'il facilite le mécanisme de navigation pour l'utilisateur. En plus de liens de références (citation, renvoi) communs à la plupart des systèmes hypertextes, d'autres types ont été utilisés, comme les liens d'inclusion et de références croisées dans Document EXaminer, les liens structurels qui structurent hiérarchiquement, l'hypertexte ou les liens implicites [MOTH97].

➤ Systèmes hypertextes

Les systèmes, basés sur l'approche hypertexte, offrent des interfaces assez conviviales. L'utilisateur accède aux données à partir de l'hypergraphe, affiche le contenu des noeuds et suit les liens structurels ou associatifs qui les relient. Il opère ainsi par association d'idées et oriente sa recherche au fur et à mesure de sa progression sur l'hypergraphe.

Cependant, lorsque l'hypergraphe est d'une complexité importante, il est facile de s'égarer dans le processus de consultation malgré la visualisation de l'hypergraphe par niveaux d'abstraction ou de création de structures personnalisées.

Les travaux actuels concernent l'introduction d'outils permettant une navigation efficace dans les hypergraphes. L'accent est donc mis sur la conception d'interfaces plus adaptées aux besoins des utilisateurs.

Par ailleurs, il est important de permettre à l'utilisateur de comprendre la double dimension de navigation qui lui est offerte : une navigation linéaire de parcours des documents et une navigation par sauts vers d'autres documents, etc.

Le problème majeur de ces systèmes concerne la création des hypergraphes. Les liens référentiels peuvent être créés automatiquement, spécifiés par le concepteur lors de la construction de la base, ou créés par l'utilisateur lors d'une session d'interrogation.

Les outils, basés sur les associations de documents ou d'unités d'information, permettent une consultation de la base de façon assez intuitive. Ces outils sont donc adaptés à des utilisateurs non spécialistes, même si des progrès restent à faire, en particulier pour que l'utilisateur arrive à situer sa position par rapport à la base. La consultation par association d'idées et par étapes successives (surtout pour les outils de navigation) est un atout important. L'inconvénient, cependant, est que les liens pré-établis, qu'ils soient établis, grâce à des mécanismes de classification ou grâce à des techniques de navigation, sont statiques. Les systèmes hypertextes possèdent parfois des modules d'ajout et de suppression de liens hypertextes permettant de créer des cheminements personnalisés. Par contre, il est rare qu'un document préalablement rattaché à un groupe (dans la classification) soit supprimé de celui-ci et associé à un autre.

Il est donc primordial que les liens, qui sont ainsi établis soient des liens stables, c'est-à-dire n'évoluant pas dans le temps. En particulier, dans une base non figée, l'ajout de nouveaux documents ne doit pas remettre en cause les liens existants. Ces mécanismes limitent le bruit dans les réponses puisque la consultation est interactive : L'utilisateur peut cibler l'espace de consultation par le choix d'un groupe de documents ou par le choix des liens qu'il active lors de la navigation [MOTH97]

3.2. Mécanismes de reformulation de requêtes

La reformulation de requête est proposée comme une méthode élaborée pour la recherche d'information, s'inscrivant dans la voie de conception de SRI adaptatifs aux besoins des utilisateurs ; elle met en œuvre un algorithme de modification de la requête en termes, en poids ou les deux simultanément, moyennant des critères de choix de termes d'expansion et règle de calcul de nouveaux poids [TAMI01].

Les mécanismes de reformulation de requêtes ont été introduits afin de limiter les problèmes liés à un mauvais choix de termes dans la requête et de contourner les lacunes du processus d'indexation.

La reformulation est basée soit sur la pertinence des documents déjà restitués, soit sur des connaissances supplémentaires sur les termes de la requête. Elle consiste à modifier l'importance des termes de la requête initiale ou à ajouter de nouveaux termes qui sont reliés aux termes de la requête origine [MOTH97].

Nous distinguons principalement deux types de reformulation de requête : la reformulation directe et la reformulation par injection de pertinence.

3.2.1. Reformulation directe

La reformulation directe de requête induit un processus d'expansion et /ou repondération de la requête initiale, en utilisant des critères de choix définis sans intervention de l'utilisateur. Ce type de reformulation peut être défini dans un contexte global, basé sur le thésaurus; ou alors local, basé sur les résultats de recherche en cours.

La reformulation, basée sur le contexte global, est essentiellement basée sur l'ajout et repondération de termes issus d'un thésaurus manuel ou construit automatiquement, en utilisant des calculs de poids de similarité, cooccurrence et relations contextuelles entre termes ou une combinaison de divers types de relations.

La reformulation, basée sur le contexte local, utilise des informations issues de la recherche en cours : documents retrouvés, termes et poids associés.

Les deux types de reformulation diffèrent essentiellement par la source d'information utilisée quant à la dérivation de l'association sémantique entre termes ou entre termes et documents, mais utilisent toutes les deux des fonctions caractéristiques pour la sélection et repondération des nouveaux termes de la requête [TAMI01].

3.2.2. Reformulation par injection de pertinence

La reformulation par injection de pertinence est une forme de recherche évolutive ; elle procède à la modification de la requête initiale en termes et poids, sur la base des jugements de pertinence de l'utilisateur sur les documents restitués par le SRI.

Son principe fondamental est d'utiliser la requête initiale pour amorcer la recherche d'information, puis exploiter itérativement les jugements de pertinence de l'utilisateur afin d'ajuster la requête par expansion repondération, ou combinaison des deux procédures, en direction des documents pertinents. Une étude synthétique des travaux afférents nous a permis de dégager trois principaux paramètres de performances : le nombre de termes ajoutés à la requête, la méthode de sélection des termes et la longueur moyenne de la requête initiale [TAMI01].

2. Annexe B

1. Espace probabilisé

Nous rappelons ci-dessous les axiomes de la théorie de la probabilité, ainsi que le théorème de Bayes, et la définition probabiliste de l'indépendance et de l'indépendance conditionnelle. Ces définitions sont données dans le cadre plus simple des espaces probabilisés finis.

Définition 1 : Espace probabilisé fini

Si :

Ω est un ensemble fini

$Z \subset P(\Omega)$ est un ensemble de parties de Ω

p est une application de $Z \rightarrow [0,1] \subset \mathbb{R}$,

On dit que (Ω, Z, P) est un espace probabilisé fini, si, et seulement si :

Z vérifie les propriétés suivantes :

- $$\begin{array}{l} (1) \Omega \in Z \\ (2) A \in Z, B \in Z \Rightarrow A \cup B \in Z \\ (3) A \in Z \Rightarrow \Omega \setminus A \in Z \end{array}$$

p vérifie les propriétés

suivantes :

$$(4) \forall A \in Z, p(A) \geq 0$$

$$(5) p(\Omega) = 1$$

Des axiomes (2) et (3) il résulte :

$$(6) \forall A, B \in Z \times Z, A \cap B = \emptyset \Rightarrow p(A \cup B) = p(A) + p(B)$$

Définition 2 : probabilité conditionnelle

Si (Ω, Z, P) est un espace probabilisé fini, soient A et B deux éléments de Z , tels que $P(B) > 0$, on appelle probabilité de A conditionnellement à B , et on note $p(A/B)$, la quantité :

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

Théorème 1 : Théorème de Bayes.

Si (Ω, Z, P) est un espace probabilisé fini, et $\{A_1, A_2, \dots, A_n\}$ un ensemble d'éléments de Z , tels que :

$$\forall (i, j), i \neq j, A_i \cap A_j = \emptyset$$

$$\bigcup_{i=1}^n A_i = E$$

Alors :

$$\forall B \in Z, p(B) = \sum_{i=1}^n p(B|A_i) \cdot p(A_i)$$

De ce théorème et de la définition 2, nous déduisons

$$p(A_k|B) = \frac{p(B|A_k) \cdot p(A_k)}{\sum_{i=1}^n p(B|A_i) \cdot p(A_i)}$$

Et :

$$\forall B, C \in Z^2, p(B|C) = \sum_{i=1}^n p(B|A_i \cap C) \cdot p(A_i|C)$$

Définition 3 : Indépendance

Si (Ω, Z, P) est un espace probabilisé fini, soient A et B deux éléments de Z . On dit que A et B sont indépendants, si, et seulement si :

$$p(A \cap B) = p(A) \cdot p(B)$$

Dans ce cas, et si $p(B) > 0$, on a également :

$$p(A|B) = p(A)$$

Définition 4 : indépendance conditionnelle

Si (Ω, Z, P) est un espace probabilisé fini, soient A, B et C trois éléments de Z . On dit que A et B sont indépendants conditionnellement à C , si, et seulement si :

$$p(A \cap B|C) = p(A|C) \cdot p(B|C)$$

De cette définition, et de la définition 2, nous déduisons également deux autres écritures équivalentes de l'indépendance de A et B conditionnellement à C :

$$p(A|B \cap C) = p(A|C)$$

$$p(B|A \cap C) = p(B|C)$$

Par la définition 2, on obtient également la **formule de chaînage** :

$$p(X_1 \cap X_2 \cap \dots \cap X_n) = p(X_1 | X_2 \cap \dots \cap X_n) \cdot p(X_2 | X_3 \cap \dots \cap X_n) \cdot p(X_n)$$

2. Variables aléatoires

Si $(\Omega, \mathcal{Z}, P)$ est un espace probabilisé fini, on définit une variable aléatoire X comme une application de Ω dans R .

Notation

Si $(\Omega, \mathcal{Z}, P)$ est un espace probabilisé fini, on note $X=k$, l'image réciproque de k par l'application X . $X=k$ est une partie de Ω

Explication

Si on cherche à modéliser un problème réel avec n variables "réelles", $\{X_1, X_2, \dots, X_n\}$, pouvant prendre chacune k_i valeurs finies :

$$X_i \in \{x_i^1, x_i^2, \dots, x_i^{k_i}\}$$

On va définir l'espace probabilisé $(\Omega, \mathcal{Z}, P)$ de la façon suivante :

- $\Omega = \left\{ (x_1^{j_1}, x_2^{j_2}, \dots, x_n^{j_n}) \mid 1 \leq j_i \leq k_i \right\}$ est l'ensemble de tous les n-uplets possibles de $\{X_1, X_2, \dots, X_n\}$
- $\mathcal{Z} = P(\Omega)$ est l'ensemble des parties de Ω
- p est une application vérifiant les axiomes des probabilités.

On peut alors définir des variables aléatoires

$$\{X_1, X_2, \dots, X_n\}$$

Telles que :

$$p_{X_k} \left((x_1^{j_1}, x_2^{j_2}, \dots, x_n^{j_n}) \right) = x_k^{j_k}$$

En d'autres termes, X_k associe à un n-uplet la valeur "réelle" de la variable X_k .

On a ainsi $\overset{p}{X}_k = x_k^{jk}$, qui par définition est l'image réciproque de x_k^{jk} par $\overset{p}{X}_k$:

$$(\overset{p}{X}_k = x_k^{jk}) = \left\{ (x_1^{i_1}, x_2^{i_2}, \dots, x_n^{i_n}) \mid x_k^{i_k} = x_k^{jk} \right\}$$

c'est-à-dire l'ensemble des n-uplets pour lesquels la valeur de $\overset{p}{X}_k$ est x_k^{jk} .

En confondant X_k et $\overset{p}{X}_k$, on peut donc utiliser les notations "événementielles" classiques, $p(A = V \mid B = V)$,

tout en ayant une sémantique ensembliste conforme avec les axiomes ci-dessus.

3. Théorie des Graphes

Définition 1 : graphe

Soit V un ensemble fini. Soit E une partie de $V \times V$.

Si $\forall x \in V, (x, x) \notin E$, alors $G = (V, E)$ est appelé un graphe orienté sur V .

Définition 2 : chemin

Soit $G = (V, E)$ un graphe orienté. Une séquence de noeuds $[v_0, v_1, \dots, v_n]$ est un chemin dans G si et seulement si :

$$\forall i, 1 \leq i \leq n, (v_{i-1}, v_i) \in E$$

Définition 3 : cycle

Soit $G = (V, E)$ un graphe orienté. Un cycle est un chemin $[v_0, v_1, \dots, v_{n-1}, v_0]$.

Définition 4 : graphe acyclique orienté (DAG)

Soit $G = (V, E)$ un graphe orienté. G est acyclique si G ne contient aucun cycle.

Définition 5 : parent, enfant

Soit $G = (V, E)$ un graphe orienté. Si $(u, v) \in E$, alors u est un parent de v , et v est un enfant de u .

3. Annexe C

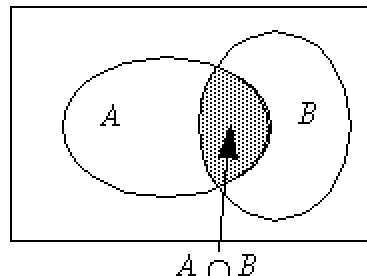
Probabilité conditionnelle

Soient A et B deux événements quelconques d'un ensemble fondamental S muni d'une loi de probabilité P . On s'intéresse à ce que devient la probabilité de A lorsqu'on apprend que B est déjà réalisé, c'est-à-dire lorsqu'on restreint l'ensemble des résultats possibles S à B .

La probabilité conditionnelle de A , sachant que l'événement B est réalisé, est notée $P(A/B)$ et est définie par la relation suivante :

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Dans cette équation, les probabilités des événements A et B doivent être calculées sur tout l'ensemble fondamental S , comme si on ne savait pas que B s'est déjà réalisé. Sinon, on obtient évidemment $P(B) = 1$



Cette relation générale pour tout espace probabilisé s'interprète facilement dans le cas où S est un espace équiprobable (mais cette relation est vraie pour un espace non équiprobable). En notant $|A|$ le nombre d'éléments de A :

$$P(A \cap B) = \frac{|A \cap B|}{|S|}, P(B) = \frac{|B|}{|S|}, P(A/B) = \frac{|A \cap B|}{|B|}$$

$P(A/B)$ traduit le rapport de la surface de $A \cap B$ sur la surface de B dans la figure précédente.

Toujours dans le cas où S est équiprobable, on a

$$P(A/B) = \frac{\text{nombre de réalisations possibles de } A \text{ et } B \text{ en même temps}}{\text{nombre de réalisations de } B}$$

Cette interprétation de la probabilité conditionnelle, facile à appréhender dans le cas d'équiprobabilité, est la définition générale de la probabilité conditionnelle qu'on doit utiliser telle quelle, sans chercher une interprétation fréquentielle dans tous les cas.

Références Bibliographiques

- [BALM02]** : Gilles Balmise. Les réseaux bayésiens. Article Septembre 2002.
- [BECK99]**: Anna Becker & Patrick Naï m. *Les réseaux bayésiens Modèles graphiques de connaissances*. Editions Eyrolles.07/1999.
- [BELL02]** : Bellot David. Fusion de données avec des réseaux bayésiens pour la modélisation des systèmes dynamiques et son application en télémédecine. Thèse pour obtention du doctorat de l'université Henri Poincaré .26 novembre 2002.
- [BIAR02]**: Biard Olivier, Deschamps Celine, Maillet Sandrine. Méthodes Bayésiennes. Article le 31 Janvier 2002.
- [BONN04]** : Stéphane BONNEVAY. Réseaux Bayésiens. Université Lyon 1. Article le 12/02/2004.
- [BORD00]** : Gloria Bordogna(*), Gabriella Pasi (°). Personalised indexing and retrieval of heterogeneous structured documents. (*) CNR-IDPA, (°)CNR-ITC Milano
- [BOUC02]**: Remco R. BOUCKAERT Low level information extraction a Bayesian network based approach. Xtal Mountain Information Technology & Computer Science Department, University of Waikato, New Zealand. Article 30/05/2002.
- [BOUG03]** : Laurent BOUGRAIN. Réseaux Bayésiens et inférence probabiliste. Projet COREX, LORIA/INRIA – Lorraine. bougrain@loria.fr. Article 12/11/2003.
- [BOUG92]** : Mohamed BOUGHANEM. Les systèmes de Recherche d'Informations d'un modèle classique à un modèle connexionniste. Thèse pour l'obtention du Doctorat de l'Université PAUL SABATIER. 17 Décembre 1992
- [BUCH02]** : Cédric Buche. Les réseaux bayésiens représentation, raisonnement et apprentissage. Laboratoire d'Informatique Industrielle .Li², EA 2215 UBO/ENIB, BREST. Article le 23 Décembre 2002.
- [CHIA04]** : Yves Chiaramella, Jean-Pierre Chevallet, Jian-Yun Nie. Vers la recherche à base de termes. Département d'informatique et recherche opérationnelle, Université de Montréal. Article 24/01/2004.
- [COMB97]** : V. COMBET. Indexation Automatique : un état de l'art. Mémoire de maîtrise de sciences du langage. Université Paris-X Nanterre Septembre 1997.

- [CORM94]:** T. Cormen, C. Leiserson, and R. Rivest. *Introduction à l'algorithmique*. Dunod , Article 1994.
- [CORS03] :** Franck CORSET. Aide à l'optimisation de maintenance à partir de réseaux bayésiens et fiabilité dans un contexte doublement censuré. Thèse pour l'obtention de grade Docteur de l'Université Joseph Fourier. 22 Janvier 2003.
- [DASN01]:** Fernando Das-Neves. A Tri-Valued Belief Network Model for Information Retrieval. Computer Science Dept. Virginia Polytechnic Institute and State University Blacksburg, VA 24060. Article décembre 2001.
- [FORM98] :** Franck Formel. Une méthode d'indexation pour les documents structurés. Equipe modélisation et recherche d'information multimédia. Thèse pour obtenir le grade de docteur de l'Université Joseph Fourier, Grenoble. Février 1998.
- [FRAN01] :** Olivier François, Philippe Leray. Evaluation d'algorithmes d'apprentissage de structure pour les réseaux bayésiens. Article INSA Rouen - Laboratoire PSI – FRE CNRS 2645.
- [GAUC02] :** Carré Rive Gauche. Lucid XML Data Manager™ pour la gestion de données XML. www.lucid-it.com. Article 2002
- [GERY02] :** Mathias GÉRY. Indexation et interrogation de chemins de lecture en contexte pour la recherche d'Information Structurée sur le Web. Thèse pour obtention de diplôme docteur de UNIVERSITÉ JOSEPH FOURIER - GRENOBLE I. 24 octobre 2002.
- [HIGR99] :** Olivier HIGELIN Contrôle optimal des processus structurés : utilisation des réseaux bayésiens. Thèse pour obtenir le titre de Docteur de l'université de CAEN. Le 03 février 1999.
- [IHAD99] :** Madjid IHADJADENE. La recherche et la navigation dans un systeme de recherche d'information grand public (*Le cas des hypercatalogues sur l'Internet*). Thèse Pour l'obtention du Diplôme de Doctorat de Science de l'Information et de la Communication. Universite Claude Bernard – Lyon 1. 17 décembre 1999
- [ISNET01] :** IS-Net 15 Bases de données relationnelles objets. Introduction théorique à la gestion de données textuelles. Article de Centre de compétence. Juin 2001
- [JENS98]:** F.V. JENSEN, *AN INTRODUCTION TO BAYESIAN NETWORKS*, UCL Press, 1998, ISBN 1-85728-332-5 HB.
- [KAHN00] :** Julien Kahn. Dynamique de l'interaction Personne - Système de Recherche d'Information (Application du paradigme de la complexité et de concepts issus de la cognition distribuée à la recherche d'information au sein de documents structurés). Thèse pour obtenir le grade de Docteur du CNAM. Discipline : Ergonomie. 28 février 2000

- [LERA98] :** Philippe Leray. *Apprentissage et Diagnostic de Systèmes Complexes : Réseaux de Neurones et des Réseaux Bayésiens*. Thèse de doctorat. Université Paris 6. 1998.
- [LERa00] :** Philippe Leray. Réseaux Bayésiens - Définition et Inférence. <http://asi.insa-rouen.fr/~pleray>. Article.
- [LERb00] :** Philippe Leray. Réseaux Bayésiens – Apprentissage des Paramètres. <http://asi.insa-rouen.fr/~pleray>. Transparents.
- [LERc00]:** Philippe Leray. Réseaux Bayésiens Apprentissage de la Structure. Philippe.Leray@insa-rouen.fr. INSA Rouen - Département ASI - Laboratoire PSI (CNRS FRE 2645). Article.
- [MEMM00] :** D. Memmi. Le modèle vectoriel pour le traitement de documents. Laboratoire Leibniz-IMAG, Grenoble (France). Article Novembre 2000.
- [MOTH97]:** Josiane Mothe. *Un modèle connexionniste pour la recherche d'informations - Reformulation de requêtes et apprentissage*. Thèse pour obtention du Doctorat de l'Université P. Sabatier - IRT : le 7 octobre 1994.
- [NAIM04] :** Patrick Naim, Pierre-Henri Willemin, Philippe Leary, Olivier Pourret, Anna Becker. Les Réseaux Bayésiens. Groupe Eyrolles, 2004
- [NIE04] :** Jian-Yun Nie. Cours IFT6255 -- Hiver 2004. Recherche d'Information. Article.
- [PILL00] :** PILLET Violaine. Méthodologie d'extraction Automatique d'information à partir de la littérature scientifique en vue d'alimenter un nouveau système d'information. Thèse pour l'obtention le diplôme Docteur en sciences. 25 janvier 2000. Université de droit, d'économie et des sciences d'aix-Marseille.
- [PILL04]:** C. PELLEGRINI. Raisonnement, réseaux bayésiens et inférence probabiliste. IA 2003-2004. Article 23/03/2004.
- [PIWO02]:** Benjamin Piwowarski. Apprentissage et recherche documentaire : une approche probabiliste différentielle. *Laboratoire d'Informatique de Paris 6 (LIP6)8, rue du Capitaine Scott.75015 Paris, France*. Benjamin.Piwowarski@lip6.fr. Article 17/05/2001
- [PIWO02]:** Benjamin Piwowarski, Ludovic Denoyer, Patrick Gallinari. Un modèle pour la recherche d'information sur des documents structurés. LIP6 – 8 rue du capitaine Scott – 75015 PARIS – France. JADT 2002 : 6es Journées internationales d'Analyse statistique des Données Textuelles. Article 22/02/2002.

- [PIWO03a] :** Benjamin Piwowarski. Techniques d'apprentissage pour le traitement d'informations structurées : application à la recherche d'information. Thèse de doctorat de l'université Paris 6. 17/07/2003.
- [PIWO03b] :** Benjamin Piwowarski, Georges-Etienne Faure, Patrick Gallinari. Bayesian Networks and INEX. LIP 6, Paris, France. Article 19/02/2003.
- [PIWO03c] :** Benjamin Piwowarski, Patrick Gallinari. A Machine Learning Model for Information Retrieval with Structured Document. LIP – Université Paris, France. Article 2003.
- [PIWO04] :** Benjamin Piwowarski* — Mounia Lalmas**. Interface pour l'évaluation de systèmes de recherche sur des documents XML. * *LIP6, Université Paris 6, Paris, France*, ** *Department of Computer Science, Queen Mary University of London, England*. Article 22/02/04.
- [POPU00] :** Sébastien Populaire. *Introduction aux réseaux bayésiens*. Rapport interne du laboratoire SIME. Université de Technologie de Compiègne. Mars 2000.
- [POUL02] :** POULIQUEN Bruno. Indexation de textes médicaux par extraction de concepts, et ses utilisations. Thèse Pour obtenir le grade de Docteur de l'Université RENNES I. 07 juin 2002.
- [RADH03] :** Saï d RADHOUANI. Extraction et Indexation de Chemins de Lecture pour la recherche d'Information sur le WEB. Mémoire de DEA Systèmes d'Informations à l'Unibersité de JOSEPH FOURIER de Grenoble. 27 juin 2003.
- [SMAI94] :** Malika SMAÏ L. Raisonnement à base de cas pour une recherche évolutive d'information; Prototype Cabri-n. Vers la définition d'un cadre d'acquisition de connaissances. Thèse pour obtention du Doctorat de l'Université Henri Poincaré_Nancy I. 14 octobre1994.
- [TAMI01] :** L.Tamine, M. Boughanem. Un algorithme génétique spécifique à une reformulation multi-requête dans un système de recherche d'information. Article CEPAD 2001. IRIT – SIG, Université PAUL SABATIER.
- [THVU02] :** Trang Huyen Vu, Ludovic Denoyer, Patrick Gallinari. Un modèle statistique pour la classification de documents structurés. *LIP6 - Université Paris 6 - 8 rue du capitaine Scott 75015 PARIS*. {vu,denoyer,gallinari}@poleia.lip6.fr. Article 14/10/2002.
- [USTL04] :** USTL. Règle de Bayes. Maîtrise d'informatique 2003/2004. Intelligence Artificiel. Licence Maîtrise d'Informatique de Lille. Article avril 2004
- [ZARG02] :** Haï fa ZARGAYOUNA Contexte et sémantique pour une indexation de documents semi-structurés. *LIMSI/CNRS-Université Paris 11*.Article 23/11/2002

Sites:

- [1]: <http://www.math.univ-mlv.fr/>. Théorie: qu'est-ce qu'un réseau bayésien ? Olivier Pourret, Marc Bouissou.
- [2] : <http://www.bayesia.com/> . Les réseaux bayésiens
- [3] : <http://kmcenter.free.fr/>. Les réseaux Bayésiens. Gilles Balmissse. 02/09/2002.
- [4] : Les réseaux bayésiens - Copyright . Elseware 1997-2000. Les réseaux bayésiens
- [5] : <http://www.loria.fr/> . Les réseaux bayésiens
- [6] : <http://www.cert.fr/> . Graphes causaux et bayésiens. Patrick Fabiani. 17/12/96.
- [7] : <http://www.bayes.lip6.fr/> Un outils d'analyse: Les réseaux bayésiens – groupe de travail réseaux bayésiens- MAJ 03/06/04.
- [8] : <http://www.enib.fr/~buche>. Cédric Buche. Les Réseaux Bayésiens représentation, raisonnement et apprentissage, Laboratoire d'Informatique Industrielle. 13/12/2002
- [9]: <http://www.sciences-en-ligne.com/> Bayesian Network
- [10]: <http://www.aima.cs.berkeley.edu>. Réseaux Bayésiens. Prentice Hall. Artificial Intelligence: A Modern Approach Russell, Norvig
- [11]: <http://www.cs.berkeley.edu/~murphyk/Bayes/bayes.html>. A Brief Introduction to Graphical Models and Bayesian Networks. 10/14/2001
- [12] : <http://iacenter.free.fr/> Les réseaux Bayésiens. Gilles Balmissse. 2003.
- [13] : [HTTP://WWW.SPIP.NET/](http://WWW.SPIP.NET/) COMMENT FONCTIONNE LE MOTEUR DE RECHERCHE DE SPIP ?
- [14] : [HTTP://WWW.EDUCNET.EDUCATION.FR/](http://WWW.EDUCNET.EDUCATION.FR/)La gestion électronique de documents.
- [15] : <http://www.chups.jussieu.fr/> Probabilité conditionnelle ; indépendance et Théorème de Bayes