République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mouloud MAMMERI de Tizi-Ouzou
Faculté génie électrique et informatique
Département Informatique

Mémoire

En vue de l'obtention du diplôme de Master II en Informatique

Thème

Modélisation d'évolution du profil utilisateur en recherche d'information personnalisée

Proposé et dirigé par :
Mme F.ACHEMOUKH

Réalisé par :
SEHIL Sadia

Promotion 2012/2013

Remerciements

Qu'il me soit permis ici de remercier, tout d'abord, le bon dieu pour le courage et la patience qu'il ma donné afin de mener ce projet a terme.

Je tiens à remercier vivement, peut être jamais assez, ma promotrice madame Farida ACHEMOUKH, Je la remercie pour la confiance qu'elle m'a accordé en me proposant le sujet de mon mémoire, pour avoir encadré et dirigé mes recherches , pour ses remarques objectives, ses encouragements ininterrompus et sa disponibilité.
Qu'elle trouve ici l'expression de ma considération profonde.

J'aimerais remercier très chaleureusement Mr HAMMACHE pour son aide
Qu'il trouve ici les marques de mes sincères reconnaissances et de mon grand respect.

Je remercie aussi les membres de jury qui m'ont fait l'honneur en acceptant de juger mon travail.

Enfin, je remercie du fond du cœur et avec un grand amour mes parents.

Dédicaces

Je dédie ce modeste travail

A Mes parents En témoignage de leurs sacrifices et de mon amour.

A mes frères Ghiles, Noureddine, kouceila, a ma sœur zahra En témoignage de mon amour en leur souhaitant beaucoup de chance et de réussite.

A toute ma famille ;

A tous mes ami(e)s ;

A toute la promotion Informatique 2012/2013.

SOMMAIRE

Chapitre I : la Recherche d'Information Classique.

Chapitre II : La Recherche d'Information Personnalisée.

Chapitre III : Les réseaux bayésiens en Recherche d'Information

Chapitre IV : Conception et Réalisation

LISTE DES FIGURES

Chapitre I : la Recherche d'Information Classique.

Chapitre II : la Recherche d'Information Personnalisée.

Chapitre III : les réseaux bayésiens en Recherche d'Information.

LISTE DES FIGURES

Chapitre IV : Conception et réalisation.

Introduction générale

Depuis l'apparition de l'informatique, les connaissances stockées sur support numérique n'ont cessé de s'accumuler, et le nombre des documents qui les stockent s'accroît 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. Devant le nombre important de documents disponibles, la recherche séquentielle1 est bien sur très limitée et l'accès `a l'information basé sur une requête semble plus efficace. Ainsi, la Recherche d'Informations (RI) devient davantage cruciale et les Systèmes de Recherche d'Information (SRI) deviennent une aide inestimable pour rechercher une information.

La Recherche d'Information (RI) est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage et la recherche des informations. Elle propose des outils, appelés Systèmes de Recherche d'Information (SRI), dont l'objectif est de capitaliser un volume important d'information et d'offrir des moyens permettant de localiser les informations pertinentes relatives au besoin d'un utilisateur exprimé à travers une requête.

En fait, un SRI est un système qui gère une collection d'informations organisées sous forme d'une représentation intermédiaire reflétant aussi fidèlement que possible le contenu des documents grâce à un processus préalable d'indexation, manuelle ou automatique. La recherche d'information désigne alors le processus qui permet, à partir d'une expression des besoins d'information d'un utilisateur, de retrouver l'ensemble des documents contenant l'information recherchée **[Abbadeni et al, 98]** et ce par la mise en œuvre d'un mécanisme d'appariement entre la requête de l'utilisateur et les documents ou plus exactement entre la représentation de la requête et la représentation des documents. La notion de document est prise ici au sens large et peut représenter une combinaison multimédia (documents hétérogènes intégrant du texte, du son, des graphiques et de la vidéo).

Certes, les systèmes de recherche d'information sont des outils qui ont permis d'améliorer sans cesse la qualité des services d'accès a l'information, grâce a la capitalisation des théories issus de nombreux travaux de recherche ;cependant en raison de la surabondance de l'information d'une part et de sa large accessibilité à travers le web ,d'autre part ,leur mise en œuvre est confronté a de nouveaux problémes.la situation est actuellement paradoxal : la masse d'information est telle que l'accès a une information pertinente ,adapté au besoin d'un utilisateur donné devient une nécessité.

En claire, le problème n'est pas tant la disponibilité de l'information mais sa pertinence relativement à un contexte utilisateur spécifique.

Afin d'effectuer une recherche pertinente, le SRI ne doit plus se contenter d'une analyse simple de la collection de documents et d'une mise en correspondance directe entre les requêtes et les documents. Dans le but d'améliorer la qualité de la recherche, des techniques plus élaborées sont introduites, ces techniques sont en rapport avec la manière d'intégrer de la façon la plus efficace possible l'utilisateur dans le processus de recherche. Dès lors l'accès à l'information tend vers une nouvelle dfinition **[Allan et al, 02]** : "*Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs*" (la personnalisation de l'information).L'interaction entre l'utilisateur et le SRI permet à l'utilisateur de transmettre ses jugements en terme de pertinence, ce qui peut résoudre en partie le problème de la complexité de la requête. Grâce à ce mécanisme, il est possible au système d'acquérir des connaissances liées aux utilisateurs et de construire des profils permettant de représenter leurs centres d'intérêt, et d'effectuer un filtrage et un routage d'information.

En s'intéressant à la RI personnalisée, le contexte cognitif de l'utilisateur est modélisé dans une structure informationnelle, appelée profil. Les approches développées dans ce domaine se distinguent par la technique de modélisation du profil de l'utilisateur et de son exploitation dans la chaine d'accès à l'information. La modélisation du profil utilisateur repose sur des techniques et des outils permettant non seulement de capturer et de le représenter mais aussi de gérer son évolution de manière dynamique au cours du temps.

Dans ce cadre, notre objectif est de montrer l'impacte de l'intégration du profil utilisateur dans les systèmes de recherche d'information et cela en augmentant la collection de test TREC par des centres d'intérêt qui sont construit a partir des documents pertinents retournés par le système de recherche d'information classique terrier-3.0

.

Le présent mémoire est organisé en quatre chapitres :

Le premier chapitre présente la recherche d'information classique. Ainsi, nous commençons par définir les notions de base de la RI. Puis, nous intéressons à l'architecture des Systèmes de recherche d'information, et au processus de recherche qui montre la mise en correspondance entre la requête et les documents afin de sélectionner les documents pertinents, et une phase de reformulation de la requête dont le but est de combler le fossé existant entre la pertinence liée à l'évaluation de l'utilisateur et la pertinence jugée par le

système. Aussi on s'intéresse aux modèles de Recherche d'Information (RI). En étudions les modèles les plus connus de la RI Nous présentons à la fin de ce chapitre les techniques utilisées pour l'évaluation des SRI.

Le chapitre 2 introduit la thématique de la personnalisation de l'information, il présente donc la problématique générale de la personnalisation, c'est quoi la personnalisation et quel est son objectif. Il traite aussi l'architecture fonctionnelle d'un SRIP et la notion de profil utilisateur en s'intéressant à ces principales approches de représentation, à sa demarche de construction et à son évolution. A la fin il introduit l'évaluation des SRIP.

Une des grandes problématiques de notre époque est de traiter la grande quantité des données qui est mise à notre disposition (notamment grâce à l'informatique) pour en extraire de l'information. Il serait donc intéressant d'avoir un (ou plusieurs) modèle(s) effectuant le lien entre les observations et la réalité pour un objectif précis, et cela, même lorsque les observations sont incomplètes et/ou imprécises. Pour cela, dans le Chapitre 3, nous abordons les Réseaux Bayésiens (RBs), leur utilisation et difficulté, nous décrivons ensuite les modèles de RI basées sur les RBs.

Dans le chapitre 4 nous présentons notre approche ainsi que les outils de développement utilisés. Et nous terminons par une conclusion générale et les perspectives envisagés.

Le mémoire contient également deux annexes A et B qui présentent respectivement la plateforme de recherche terrier et un rappel sur les probabilités.

Introduction :

Le développement rapide des nouvelles technologies de l'information et de la communication ainsi que l'essor du web, nous a confrontés à une très grande masse d'informations hétérogènes. Les masses d'informations accessibles n'ont cessé d'augmenter, et les volumes de documents qui les stockent s'accroissent très rapidement. En mars 2002, le plus grand moteur de recherche a contenu approximativement 968 millions de pages classées dans sa base de données. A titre d'exemple, aujourd'hui le moteur de recherche Google inclut 3.083.324.652 sites web **[neuhold, 03]**.

Trouver dans cette accumulation ce que l'on recherche précisément, devient de plus en plus ardu. Le problème n'est plus la disponibilité de l'information mais la capacité de sélection de l'information répondant aux besoins précis d'un utilisateur.
La conception et la mise en œuvre d'outils efficaces, permettant notamment à l'utilisateur de n'avoir accès qu'à l'information qu'il juge pertinente, devient une nécessité absolue.
Les efforts continus des chercheurs ont permis jusqu'à présent d'améliorer sans cesse les performances et la qualité des services d'accès à l'information. Les premiers travaux en RI sont qualifiés d'approche classique.

On présente, dans ce qui suit, les concepts généraux, mais fondamentaux, de la recherche d'informations classiques tout en incluant une description de processus de recherche et les différents modèles permettant de mesurer la similitude requête-document.

I.1. historique de la recherche d'information :

Le nom de « recherche d'information » (information retrieval) fut donné par Calvin N. Mooers en 1948 pour la première fois quand il travaillait sur son mémoire de maîtrise **[Mooers, 48].**

Le domaine de recherche d'information remonte au début des années 1950, peu après l'invention des ordinateurs, Les pionniers de l'époque étaient enthousiastes à utiliser l'ordinateur pour automatiser la recherche des informations, qui dépassaient la capacité humaine. A cette année on commençait de petites expérimentations en utilisant de petites collections de documents (références bibliographiques). le modèle utilisé était le modèle booléen.

Dans les années 1960 et 1970, des expérimentations plus large ont été menées et des corpus de test ont été conçus pour évaluer des systèmes différents .Le système qui a le plus

d'impact sur le domaine est le système SMART (salton's Magical Autoatic Retreiver of text), développé a la fin des années 1960, les travaux sur ce système ont été dirigé par Gérard salton, professeur a l'université de cornell(USA). Dans ce projet une série d'expérimentations a été menée, portant sur divers sujets comme :

- La comparaison entre l'indexation manuelle et l'indexation automatique ;
- Le problème de recherche d'information interactive et la rétroaction de pertinence (relevance feedback);
- L'architecture de système de RI ;
- L'utilisation du modèle vectoriel ;
- Le regroupement de documents (ou clustering) ;
- etc.

Le système SMART fut réécrit dans les années 1970 et 1980 par E. Fox et C. Buckley. Ce système a été, et est encore, utilisé par de nombreux chercheurs pour des expérimentations en RI **[Smail, 09]**.

Dans les années 1980, avec le développent de l'intelligence artificielle (IA), on a alors tenté d'intégrer des techniques de l'IA en RI (ex : réalisation de systèmes experts pour la RI).

Les années 1990 (surtout a partir de 1995) sont les années de l'internet .cela a pour effet d'élargir la problématique de la RI, en traitant plus souvent des documents multimédia qu'avant. Du coté des modèles il y a eu beaucoup de développent notamment le modèle probabiliste qui a connu de nombreuses améliorations et a été adapter a différents type d'information.

I.2. Principes généraux de la recherche d'informations :

La recherche d'informations est l'ensemble des méthodes et techniques pour l'acquisition, L'organisation, le stockage, la recherche et la sélection d'information pertinente pour un utilisateur **[Boughanem, 12].** Elle est apparue dans les années soixante, et a une vision orientée système, en ce sens où la recherche des informations pertinentes se base uniquement sur l'appariement des documents avec la requête soumise par l'utilisateur. Toutefois, cette vision de l'accès à l'information suppose que l'utilisateur est extérieur au système de recherche.

Elle propose des outils, appelés systèmes de recherche d'information (SRI).

I.2.1. Les systèmes de recherche d'informations :

I.2.1.1. Définition :

Un système de recherche d'information (**SRI**) est un système informatique qui permet de retrouver une information pertinente par rapport à une requête suivant un processus de sélection a partir d'une ou plusieurs collections de documents **[Rozknop]**.

Il inclut un ensemble de procédures et d'opérations qui permettent la gestion, le stockage, l'interrogation, la recherche, la sélection et la représentation de cette masse d'informations.

Dans cette définition on distingue trois notions clés : document, requête, pertinence.

I.2.1.2 Notions de document-requête et de pertinence :

- **Document :**

Le document représente le conteneur élémentaire d'information, exploitable et accessible par le SRI, Un document peut être un texte, un morceau de texte, une page WEB, une image, une bande vidéo, etc. On peut appeler document toute unité qui peut constituer une réponse à un besoin en information exprimé par un utilisateur **[zemirli, 03/04]**.

Un document pertinent est un document qui doit contenir l'information que l'utilisateur recherche.

- **Requête :**

Une requête exprime le besoin d'information de l'utilisateur, elle représente l'interface entre le SRI et l'utilisateur. Plusieurs systèmes utilisent des langages différents pour décrire la requête:

- par une liste de mots clés : cas des systèmes SMART **[Salton, 71]** et Okapi **[Robertson, 99]**,
- en langage naturel : cas des systèmes SMART **[Salton, 71]** et SPIRIT **[Fluhr 85]**,
- en langage booléen : cas du système DIALOG **[Bourne, 79]**,
- en langage graphique : cas du système NEURODOC **[Lelu, 92]**.

- **Pertinence :**

Un document est dit pertinent s'il contient les informations d'ont l'utilisateur a besoin

C'est sur cette base que le système doit juger si un document doit être donné a l'utilisateur comme réponse.

Le problème fondamentale de la recherche de l'information est la formalisation du concept de pertinence d'un document face a une requête .les critère de la pertinence sont très difficile a identifier par l'utilisateur donné dans un sujet particulier .c'est souvent plus facile pour lui de juger si un document spécifique est pertinent que de déterminer les critères de pertinence.

La notion de pertinence d'un document face à une requête ne dépend que de l'utilisateur. Elle n'est cependant pas une mesure objective, généralisable à tous les utilisateurs. Elle se définit par un ensemble de critères et de préférences personnalisables spécifiques à chaque utilisateur ou communauté d'utilisateurs.

I.3. Architecture générale d'un SRI :

L'architecture générale d'un SRI est représentée dans la figure1.

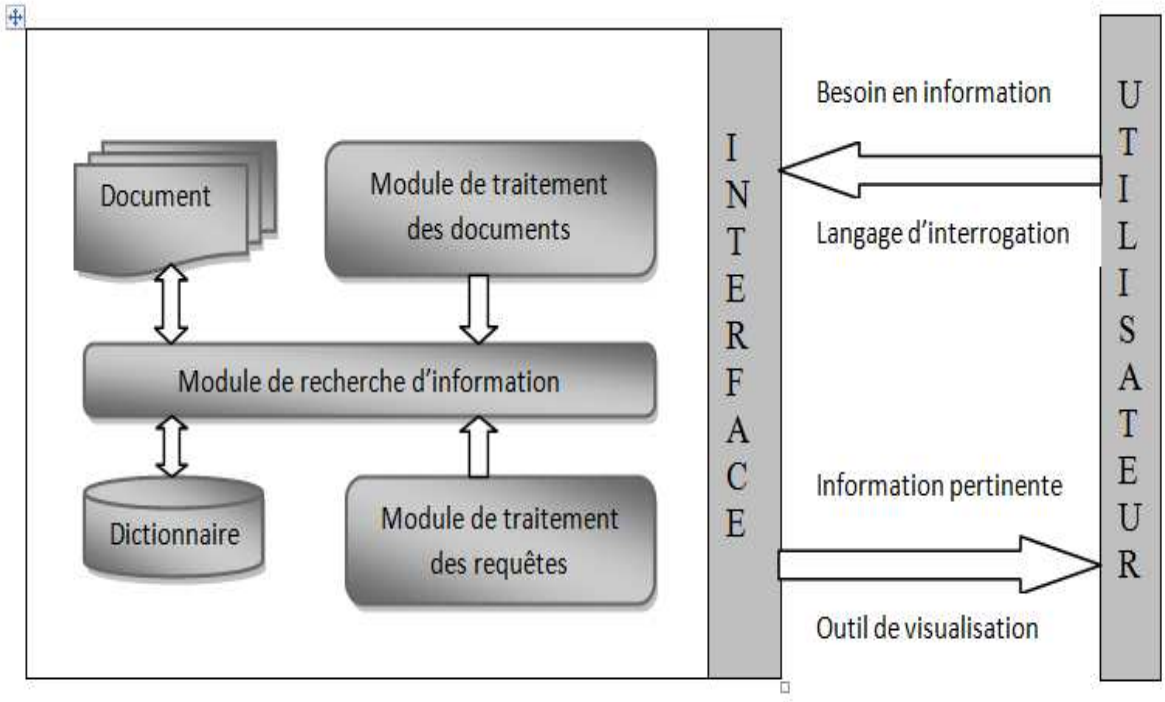


Figure I.1: Architecture générale d'un SRI [Tamine, 98].

Dans cette figure on distingue les éléments de base suivant :

- **L'interface :**

Assure la communication entre la base documentaire et l'utilisateur. Elle doit être ergonomique et conviviale pour faciliter l'accès à l'information.

La mise en œuvre de langages d'interrogation et outils de visualisation pour l'expression des requêtes d'une part et pour la visualisation de l'information pertinente d'autre part est nécessaire pour la communication entre le SRI et l'utilisateur.

- **Langage d'interrogation** :

  Plusieurs langages ont été mis au point dans les SI, les plus répondus sont les suivant :

  _ **Langage booléen** :

  L'utilisateur exprime sa requête sous forme de termes reliés par des operateurs de la logique booléenne. Comme le cas des systèmes LEXIS, STAIRS.

  Ce type d'interrogation est assez strict imposant une syntaxe difficilement accessible à un large public, et les requêtes sont de plus en plus complexe avec le nombre d'operateur utilisés .les résultats de la recherche dépend de l'ordre des operateurs dans la requête, ainsi seul les documents répondant a la requête sont restitues.

  _ **Langage naturel** :

  L'utilisateur exprime sa requête en langage libre ce qui permet une utilisation généralisé des SRI, ça n'exige pas la connaissance d'une syntaxe pour formuler la requête comme est le cas avec les langages booléen.

  Cependant le traitement de ces requêtes ambiguës pour le système nécessite la mise en œuvre de mécanismes élaborés pour les traduire en mot clés sans perte de signification.

  _ **Langage graphique** :

  Avec ce langage, une interface d'aide a la formulation des requêtes est proposé a l'utilisateur .en effet une vue d'ensemble de la base d'information est donné a l'utilisateur pour lui faciliter la formulation de sa requête.

  Dans ce cas l'utilisateur assiste dans la formulation de la requête.

- **Outil de visualisation :**

  Les outils de visualisation dans les SRI, offrent la possibilité de consultation de l'intégralité des documents, les documents retournés sont présenté à l'utilisateur

sous une forme qui lui permet de consulter l'information pertinente. On distingue différentes forme de présentation des résultats, dont les principales sont :

- **Présentation du document intégrale :** le système présente les documents intégraux ordonnés par ordre décroissant de ressemblance avec la requête.

- **Présentation de passage de document :** le système présente des unités d'information au lieu de documents entiers.

- **Présentation d'un identifiant de document :** le système retourne une liste d'identifiant de document a l'utilisateur qui peut alors visualiser le contenu de chaque documentent en sélectionnant son identifiant

- **Module de traitement des documents et requêtes :**

  Le module de traitement des documents s'occupe de la représentation interne des documents, leur organisation et leur stockage.

  Le module de traitement des requêtes permet de représenter les requêtes sous un formalisme prédisposant à la recherche.

- **Module de recherche d'information :**

  Ce module calcule le degré de correspondance des deux représentations internes du document et de la requête et retourne les documents jugés pertinents**.**

- **La base documentaire :**

  Contient un nombre important de documents, son contenu diffère d'une base a une autre selon le domaine d'application du SRI. Principalement on distingue deux types de bases documentaires :

  - **Les référothéques :** constituées d'un ensemble d'enregistrements faisant référence au document dans lequel se trouve l'information intégrale.

_ **Les bibliothèques :** composés de textes intégraux de document.

- **Dictionnaire :**

  Comprend les mots clé du domaine de la base documentaire et les mots nécessaire au traitement des requêtes.

I.4. Processus de recherche d'information :

le système de recherche regroupe un ensemble de méthodes et procédures permettant la gestion des collections de documents stockés sous forme d'une représentation intermédiaire permettant de refléter aussi fidèlement que possible leurs contenus sémantiques. L'interrogation de la collection de documents a l'aide d'une requête nécessite la représentation de cette dernière sous une forme unifié compatible avec celle des documents.ces fonctionnalités sont représentées a l'aide du processus globale de la RI, communément nommé processus en U.

Les différentes étapes du processus de RI, sont représentées schématiquement dans la **Figure2** qui illustre particulièrement :

_ Les notions de documents et de requêtes qui sont des conteneurs d'informations,

_ Les opérations d'analyse, d'indexation et d'appariement qui permettent globalement de traiter la requête dans le but de sélectionner des documents à présenter à l'utilisateur.

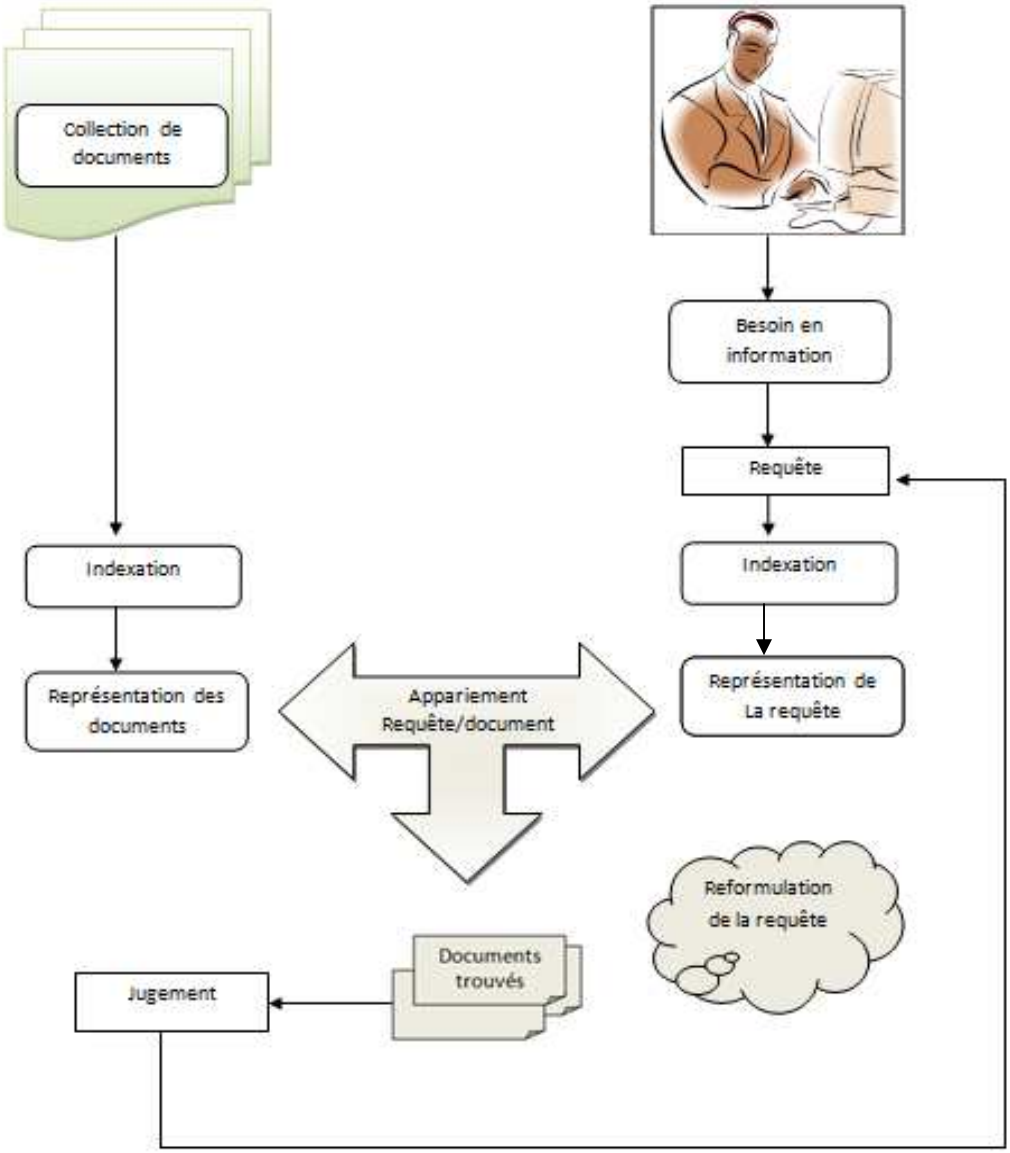


Figure I.2 : Processus en U de recherche d'informations [BADR, 07].

I.4.1. Collection de document :

Constitue l'ensemble d'information exploitable par l'utilisateur.

I.4.2. Besoin en information :

Le besoin en information est une expression mentale correspondant à ce que l'utilisateur attend, cette expression est exprimée à travers une requête en utilisant un langage d'interrogation qui dépend du système de recherche d'information.

Une requête est une instanciation d'un besoin en information.

Le déroulement de ce processus induit deux principales phases : indexation et appariement requête/document.

Comme illustré dans la Figure 2, un SRI intègre deux fonctions principales, représentées schématiquement par le processus U de recherche d'information **[Belkin 92]** : indexation document (resp. requête) et appariement document-requête.

I.4.3. L'indexation :

L'indexation est un traitement pour construire une représentation des documents et une représentation des requêtes efficace pour la recherche, cela consiste à déterminer et extraire des termes ou groupe de termes représentatifs du contenu d'un document ou d'une requête, qui couvrent au mieux leur contenu sémantique. La liste de ces termes constitue, ce que l'on nomme le **descripteur** du document ou de requête.

Les descripteurs des documents (mots, groupe de mots) sont rangés dans une structure appelée dictionnaire constituant le **langage d'indexation**.
L'efficacité d'une analyse d'un contenu de document, ou d'un mécanisme d'indexation, est mesurée par deux paramètres, l'exhaustivité de l'indexation et la spécificité des termes d'indexation.

- **L'exhaustivité de l'indexation :**

Exprime la qualité de la représentation obtenue, une indexation est dite exhaustive quand elle génère un terme d'indexation pour chaque concept dans le document, et elle est dite non exhaustive quand elle génère seulement des termes représentant les concepts principaux de document.

- **La spécificité de l'indexation :**

Exprime le degré de généricité ou de spécialité des termes d'indexations, quand un terme représente un concept très vaste, un grand nombre de documents sont restitués a l'utilisateur comme réponse a une requête comprenant ce terme, donc les termes généraux sont incapable de différencier les documents .par contre, les termes spécifiques génèrent un ensemble réduit de document dont la plupart sont pertinents. par exemple le terme information est un terme générale qui apparait dans la plupart des documents, par contre le terme recherche d'information est un terme spécifique qui génère un ensemble réduit de documents.

L'indexation peut être caractérisée par son mode et fonction de pondération.

I.4.3.1. Mode d'indexation :

Peut être manuelle, automatique ou semi-automatique.

- ✓ **Manuelle :**

Chaque document est analysé par un spécialiste du domaine ou un documentaliste, il détermine les unités syntaxiques qui lui semblent le plus significatifs pour représenter le contenu du document.

Le processus d'indexation s'appui généralement sur un vocabulaire contrôlé qui permet la recherche par concepts et le regroupement de documents par sujet ou par thème.

- ✓ **Automatique :**

Chaque document est analysé à l'aide d'un processus entièrement automatisé. Peut se faire selon deux approches : statistique et linguistique.

- _ **Approche statistique :**

Se base sur la distribution statistique des éléments linguistiques (mot) dans le document. En appliquant des méthodes quantitatives.

- _ **Approche linguistique :**

Se base sur les techniques de traitement du langage naturel, pour extraire les concepts les plus discriminants dans un document.

- ✓ **Semi-automatique (mixte) :**

C'est une combinaison des deux méthodes précédentes un premier processus automatique permet d'extraire les termes du document. Cependant, le choix final reste au spécialiste du domaine ou au documentaliste pour établir les relations entre les mots clés et choisir les termes significatifs.

I.4.3.2. Fonction de pondération :

La pondération permet d'affecter à chaque terme d'indexation une valeur qui mesure son importance dans le document où il apparaît.

Le pouvoir de discrimination des termes pour décrire le contenu des documents n'est pas identique pour tous les termes. Pour trouver les termes du document qui représentent le mieux son contenu sémantique, **[Roberston, 76]** a défini la **fonction de pondération** d'un terme dans un document connue sous la forme de ***Tf.Idf***, qui est reprise dans différentes versions par la majorité des SRI.

- **Tf (term frequency) :**

  Cette mesure est proportionnelle à la fréquence du terme dans le document. L'idée sous-jacente est que plus un terme est fréquent dans un document, plus il est important dans la description de ce document.

  Le *Tf* est souvent exprimé selon l'une des déclinaisons suivantes :

  1. Tf : utilisation brute,
  2. $0.5+0.5\,\frac{\mathrm{Tf}}{\max(\mathrm{Tf})}$

- ***Idf* (*Inverse of Document Frequency*) :**

  Mesure l'importance d'un terme dans toute la collection. L'idée sous-jacente est que les termes qui apparaissent dans peu de documents de la collection sont plus représentatifs du contenu de ces documents que ceux qui apparaissent dans tous les documents de la collection. Cette mesure est exprimée selon l'une des déclinaisons suivantes :

  1. $\mathrm{Idf} = \log\left(\frac{\mathrm{N}}{\mathrm{df}}\right)$.
  2. $\mathrm{Idf} = \log\left(\frac{\mathrm{N}-\mathrm{df}}{\mathrm{df}}\right)$.

  Où *df* est la proportion de documents contenant le terme et N le nombre total de documents dans la collection.

La fonction de pondération de la forme ***Tf.Idf*** consiste à multiplier les deux mesures Tf et Idf. Une formule largement utilisée est la suivante:

$$\mathrm{Tf} * \mathrm{Idf} = \left(0.5 + 0.5\frac{\mathrm{Tf}}{\max(\mathrm{Tf})}\right) * \log\left(\frac{\mathrm{N}}{\mathrm{df}}\right)$$

Une normalisation de la mesure du *Tf.Idf* par rapport à la longueur des documents a été proposée.

$$\text{Tf*Idf} = \frac{\text{Tf}+\log\left(\frac{N-df+0.5}{df+0.5}\right)}{2.\left(0.25+0.75.\frac{dl}{\Delta d}\right)}$$

dl est la longueur du document en nombre de termes et Δd la longueur moyenne des documents de la collection.

En effet, lors des compagnes d'évaluation internationales, la mesure a eu des performances très limitées dans des corpus de taille très variable. Le problème posé est que les termes appartenant aux documents longs apparaissent très fréquemment et emportent le poids sur les termes appartenant à des documents moins longs. Les documents longs auront alors plus de chance d'être sélectionnés **[DeClaris, 94].**

I.4.4. reformulation de la requête :

La reformulation de requête est un processus ayant pour objectif d'affiner et améliorer automatiquement la requête initiale de l'utilisateur, en rajoutant de nouveaux termes et/ou supprimant des termes inutiles. Cette reformulation permet de coordonner le langage de recherche utilisé par l'utilisateur dans sa requête et le langage d'indexation des documents.
On distingue principalement deux approches de reformulation de requêtes : une approche basée sur un processus automatique et une autre, basée sur un processus interactif nous allons les présenter dans ce qui suit :

I.4.4.1. Utilisation de ressources linguistiques :

L'approche d'expansion de requête (utilisation de ressources linguistiques) permet d'étendre automatiquement les requêtes sans l'intervention de l'utilisateur via des ressources linguistiques tels que les thesaurus qui représente des relations entres différents termes.
Elle est basé sur l'utilisation de liens sémantiques établis entre les termes .Ces liens permettent de remplacer les termes de la requête par un groupe de termes équivalent au niveau sémantique, mais permettant de trouver plus de documents.

I. 4.4.2. Réinjection de pertinence (relevance feedback):

l'approche interactive (ou par réinjection de pertinence) exploite uniquement un sous ensemble de documents sélectionnés parmi les premiers résultats obtenus de l'exécution de la

requête initiale, cette technique prend en considération les jugements de pertinence de l'utilisateur ,en effet l'utilisateur examine les documents retournés de la première recherche pour sa requête et détermine les document pertinents et non pertinents ,le SRI alors modifie sa requête a partir de ses jugements ,soit pour repondérer les termes de la requête initiale , soit pour y ajouter (resp. supprimer) d'autre termes contenus dans les documents pertinents (resp. non pertinents) . La nouvelle requête ainsi obtenue à chaque itération de *feedback*, permet de corriger la direction de la recherche dans le sens des documents pertinents.

I.4.5. Appariement requête-document :

Le processus d'appariement permet de mesurer la pertinence d'un document vis-à-vis d'une requête.

A cet effet, une mesure de similitude (correspondance) entre la requête indexée et les descripteurs des documents de la collection est calculée. Seuls les documents dont la similitude dépasse un seuil prédéfini sont sélectionnés par le SRI. Le processus d'appariement est étroitement lié au processus d'indexation et de pondération des termes. Il existe deux méthodes d'appariement :

- **Appariement exact « exact match retrieval »:**

  Le résultat est une liste de documents respectant exactement la requête spécifiée avec des critères précis. Les documents retournés ne sont pas triés.

- **Appariement approché « best match retrieval » :**

  Le résultat est une liste de documents sensés être pertinents pour la requête. Les documents retournés triés selon un ordre de mesure. Cet ordre reflète le degré de pertinence document/requête.

I.5. les principaux modèles de la recherche d'information :

Le système de recherche d'information définit une méthode d'appariement entre la représentation des documents (après le processus d'indexation) et la représentation de la requête afin de déterminer le degré de correspondance (similarité), cette méthode correspond au modèle de recherche .Le modèle de recherche détermine alors le comportement clé d'un SRI.

On peut distinguer trois grandes classes de modèles regroupés selon les fondements mathématiques sur lesquels ils se basent.

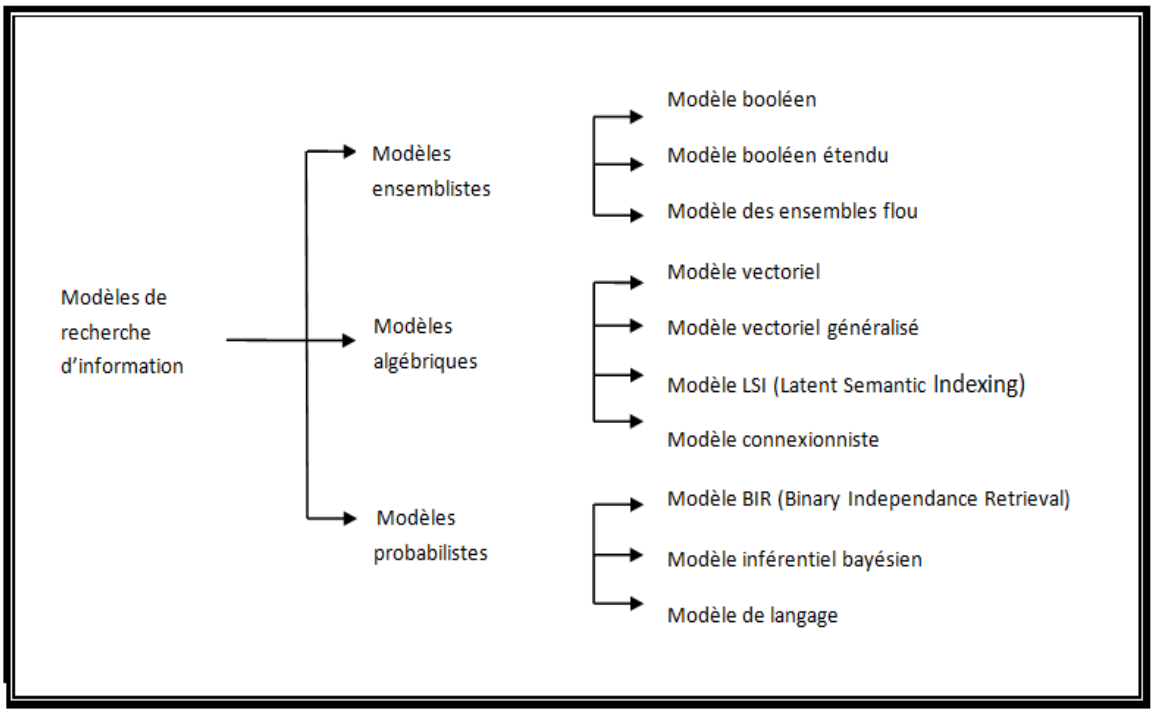


Figure I. 3 : les trois principaux modèles de la RI [Baeza-Yates 99]

Nous présentons ici les modèles les plus couramment utilisés pour la RI.

I.5.1. Le modèle booléen :

C'est le premier modèle utilisé en RI **[Salton, 71]** , il est basé sur la théorie des ensembles et l'algèbre de Boole, les documents sont vus comme un ensemble de termes associés a une variable booléenne, celle-ci étant positionnée a "vrai" pour un document si le terme y est présent.

Les requêtes sont des suites de termes séparés par des operateurs logiques *ET*, *OU* et *NON*, et les documents pour lesquels cette expression est vérifiée sont retournés à l'utilisateur, ce modèle ne renvoie aucun document répondant partiellement a la demande pour cela il est dit modèle exact.

Les inconvénients du modèle booléen résident dans :

- La difficulté de la formulation de la requête par l'utilisateur (ambigüité ET/OU) ;
- La sélection des documents est basée sur une décision binaire ;
- Pas d'ordre des documents sélectionnés ;
- Pas de pondération des termes.

Cependant des extensions de ce modèle ont été effectuées pour remédier a ces inconvénients, intégrant des poids dans l'expression de la requête et des documents. La sélection des documents s'effectuera donc sur la base d'un appariement rapproché et non plus exact.

I.5.2. Le modèle vectoriel :

Après le modèle booléen le modèle qui a plus influencé la recherche d'information est le modèle vectoriel.

Il a été proposé par Gerard salton [**Salton, 71]** dans le projet SMART (*Salton's Magical Automatic Retriever of Text*), repose sur des bases mathématiques des espaces vectoriels.

Dans ce modèle, les documents et les requêtes sont représentés dans un espace vectoriel engendré par l'ensemble des termes d'indexation t_1, t_2,…, t_i,…t_N ou N est le nombre total de termes issus de l'indexation de la collection de documents.

Chaque document est représenté par un vecteur : $D_j = (d_{1j}, d_{2j}, \ldots, d_{ij}, \ldots, d_{Nj})$.

Chaque requête est représentée par un vecteur : $Q = (q_1, q_2, \ldots, q_N)$.

Avec :

d_{ij} est le poids du terme t_i dans le document D_j.

qi est le poids du terme t_i dans la requête Q.

La pertinence d'un document par rapport a la requête est évaluée par le degré de similarité entre le vecteur du document et celui de la requête RSV (*Retrieval Status Value)* appelé, cette similarité étant calculée par exemple par :

Un produit scalaire : $RSV(Q, D_j) = \sum_{i=1}^{N} (q_i * d_{ij})$.

Une mesure de **Jaccard :** $RSV(Q, D_j) = \frac{\sum_{i=1}^{N} q_i * d_{ij}}{\sum_{i=1}^{N} {q_i}^2 + \sum_{i=1}^{N} {d_{ij}}^2 - \sum_{i=1}^{N} q_i * d_{ij}}$

Une mesure de **Dice :** $RSV(Q, D_j) = \frac{2 * \sum_{i=1}^{N} q_i * d_{ij}}{\sum_{i=1}^{N} {q_i}^2 + \sum_{i=1}^{N} {d_{ij}}^2}$

Ou, plus couramment, la mesure de **cosinus** de l'angle entre les deux vecteurs.

La mesure de **cosinus :** RSV (Q, D_j)= $\frac{\sum_{i=1}^{N} q_i * d_{ij}}{\left(\sum_{i=1}^{N} {q_i}^2\right)^{1/2} * \left(\sum_{i=1}^{N} {d_{ij}}^2\right)^{1/2}}$

Ce modèle permet des requêtes moins expressives (sans operateurs) que le modèle booléen.
Dans ce modèle les documents sélectionnés sont triés selon leur ordre de pertinences, mais le fait que l'association entre les termes d'indexation n'est pas considérée. Il est impossible de représenter des phrases ou des mots multi termes. On considère effectivement que les termes sont indépendants.

I.5.3. Le modèle probabiliste :

Il a été proposé par Robertson et Sparck Jones **[Robertson 76]**, il aborde la notion de probabilité de pertinence et de non-pertinence d'un document par rapport à une requête. La pertinence entre un document et une requête est mesurée par le rapport entre la probabilité qu'un document *D* donné soit pertinent pour une requête *Q*, notée *P* (*D/pert*), et la probabilité qu'il soit non pertinent, notée *P* (*D/Npert*). Ces probabilités sont estimées par les probabilités qu'un terme de la requête soit dans un document pertinent et non pertinent.la pertinence globale est alors calculée par :

$$RSV(D,Q) = \frac{\mathrm{P(D/pert)}}{\mathrm{P(D/Npert)}} = \sum_{i=1}^{n} \log \frac{\mathrm{P(1-q)}}{\mathrm{q(1-p)}}$$

Avec :

- p=P (t_i/pert) : la probabilité que le terme ti apparaisse dans un document D sachant sa pertinence pour la requête.
- q=P (t_i/Nonpert) : la probabilité que le terme ti apparaisse dans un document D sachant sa non pertinence pour la requête.
- n : le nombre de termes dans la requête.

Dans ce modèle les documents jugés pertinent sont restitué dans l'ordre de leurs pertinences, mais il ne tient pas compte des dépendances entre les termes.

I.6.Evaluation des SRI:

L'évaluation des SRI trouve son origine, a la fois théorique et méthodologique dans les projets d'évaluation des systèmes d'indexation menée a crandfield (Royaume-Uni) sous la direction de C.Cleverdon en 1957 et 1967 , et dans le projet d'évaluation de MEDLARS(Medical Literatur Analysis and Retrieval System) entre aout 1966 et juillet 1967.

La démarche de validation en RI se base sur l'évaluation expérimentale des performances du modèle ou du système proposé. L'évaluation des performances d'un modèle de RI, permet de paramétrer le modèle, d'estimer l'impact de chacune de ses caractéristiques et de fournir des éléments de comparaison entre modèles, et porte sur plusieurs critères : **[hadjouni, 09]**

. La pertinence.
. Le temps de réponse du système.
. La qualité et la présentation des résultats.

Néanmoins la capacité du système à satisfaire le besoin en information de l'utilisateur (la pertinence des résultats retournés), reste le plus important. Elle s'exprime à travers les notions de rappel, de précision, de bruit et de silence.

Lors de recherche d'information, le système retourne comme réponse à une requête des documents qui peuvent être classifiés en : document retournés par le système de recherche d'information, documents pertinents et document non pertinents.

I.6.1. Les mesures de rappel précision et silence bruit :

Le rappel : mesure la capacité de sélectionner tous les documents pertinents de la collection répondant à une requête.

$$\textbf{le rappel} = \frac{\text{le nombre de documents pertinent retournés}}{\text{le nombre total de documents pertinent dans la collection}}$$

La précision : mesure la capacité de rejeter tous les documents non pertinents.

$$\textbf{la precision} = \frac{\text{le nombre de documents pertinents retournés}}{\text{le nombre total de documents retournés}}$$

Le bruit : représente les documents retournés, mais non pertinents, Bruit =1-Précision.

$$\textbf{le bruit} = \frac{\text{nombre de documents retournés et non pertinents}}{\text{nombre de documents retournés}}$$

Le silence : représente les documents pertinents non retournés, Silence=1-Rappel.

$$\textbf{le silence} = \frac{\text{le nombre de documents non retournés et pertinent}}{\text{nombre de documents pertinents}}$$

La performance d'un SRI est appréciée si, les taux de précision et de rappel sont élevés et, les taux de bruit et de silence sont bas.

Un taux de rappel égale a 1 indique que tous les documents pertinents sont retournés .une précision égale a 1 indique que les documents retrouvés sont pertinents

I.6.2. La courbe précision-rappel :

Les mesures de précision-rappel ne sont pas indépendantes, en effet en réponse à une requête on à un taux de rappel égal à 1, mais une précision faible, voir de même, si on augmente la précision en restreignant le nombre de documents retournés, dans ce cas le rappel pouvant diminuer. Dans les SRI on cherche à améliorer le couple rappel et précision.

Ces deux métriques ne sont pas statiques non plus (c'est-à-dire qu'un système n'a pas qu'une mesure de précision et de rappel). Le comportement d'un système varie en fonction de précision et de rappel donc de la liste ou du rang du document dans la liste. Ainsi, la courbe de la Figure suivante montre la forme générale que peut prendre la variation de rappel précision pour un système.

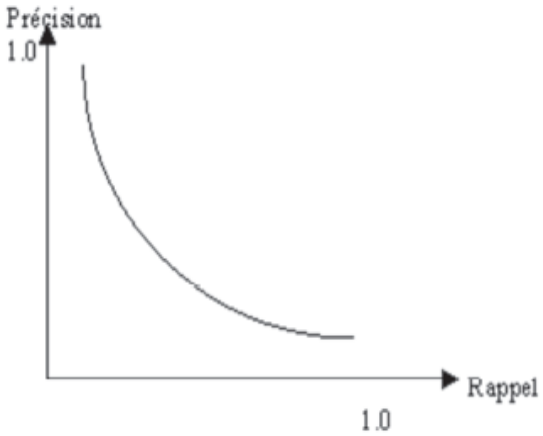


Figure I.4 : Courbe de précision _ rappel [Ihab, 11].

Le calcule de ces valeurs est rendu possible grâce aux collections de tests .cette dernière comprend un ensemble de documents (collection de documents) à indexer et sur lesquels le

système sera évaluer, une liste de requêtes prédéfinies, les jugements de pertinence manuellement établis, pour chaque requête.

L'initiative la plus importante actuellement pour la construction de collections de tests est sans conteste TREC (Text REtrieval Conference, http://trec.nist.gov). TREC est plus qu'une collection de tests, c'est un programme d'évaluation des SRI, initié par le NIST (National Institute of Standards and Technology) aux USA. TREC fournit une plate-forme comportant des collections de tests, des tâches spécifiques et des protocoles d'évaluation pour chaque tâche, pour l'évaluation et la comparaison d'expérimentations sur des collections volumineuses de textes.

Conclusion :

Nous avons présenté dans ce chapitre les concepts liés à la recherche d'information classique. Nous y avons abordé les problématiques reliées aux termes, requête et document, la définition et l'architecture générale d'un SRI et les principaux modèles de recherche .les mesures de performance sont aussi présentées comme le rappel, la précision, le silence et le bruit.

Les performances d'un SRI, ne dépendent pas seulement de l'efficacité et de la qualité d'indexation et d'appariement, mais aussi de la capacité du système à prendre en considération les besoins de l'utilisateur. D'où l'apparition d'une nouvel tendance en RI, qui consiste en la personnalisation de l'information. Le chapitre suivant porte sur la RI personnalisé.

Introduction

Compte tenu de la croissance continue du nombre et du type des documents disponibles dans le Web, il devient de plus en plus difficile pour un utilisateur de trouver les ressources pertinentes qui répondent à sa requête. Les moteurs de recherche disponibles renvoient habituellement plus de 1.500 résultats par question **[Zemirli, 03/04],** chiffre qui ne correspond généralement pas au nombre de documents pertinents et répondant au besoin de l'utilisateur. Plusieurs travaux se sont focalisés sur l'amélioration de la qualité d'accès à l'information.

La recherche d'information classique suppose que l'utilisateur est complètement représenté par sa requête et que les résultats retournés pour une même requête sont identique même si elle est exprimée par des utilisateurs différents, Cela dit que la recherche des informations pertinentes se base uniquement sur l'appariement des documents avec la requête soumise par l'utilisateur. De plus, la difficulté qu'à l'utilisateur à exprimer son besoin en information par une requête, ainsi que la différence de vocabulaire entre les termes choisis par l'utilisateur pour formuler sa requête et les termes utilisés pour représenter les documents engendrent un défaut d'appariement. Ce défaut d'appariement est à l'origine d'une dégradation des performances de recherche et de fait, l'insatisfaction de l'utilisateur. Cette problématique est encore plus accentuée avec les requêtes courtes et l'accroissement continuel des sources d'information hétérogènes et la diversité des utilisateurs **[Tamine, 05] [Zemirli, 08]**.

Les premières approches suivies pour améliorer les performances des systèmes sont dites adaptatives, se sont particulièrement axés sur l'amélioration de l'efficacité du processus de recherche notamment lors de la phase d'exécution de la requête. Les techniques développées ont eu pour but de désambiguïser le sens des mots de la requête utilisateur afin de mieux cerner le but de sa recherche. Plus particulièrement, la RI *adaptative* s'articule autour de la reformulation de requêtes, et expansion de requêtes en utilisant des techniques de désambiguïsation.

Un autre facteur de l'insatisfaction des utilisateurs, mis à part leur inexpérience et le manque de connaissances de leur besoin informationnel effectif est, que la majorité des systèmes de recherche disposent de peu d'informations sur les utilisateurs pouvant améliorer le processus de recherche. De ce fait, les travaux s'orientent actuellement vers la révision de la chaîne d'accès à l'information dans la perspective d'intégrer l'utilisateur dans l'ensemble des phases de recherche et ce, dans le but de lui délivrer l'information pertinente adaptée à son contexte

et ses préférences, répondant à ses besoins précis. Ces travaux s'inscrivent dans le cadre précis de la « personnalisation de l'information ».

Donc la personnalisation de l'information constitue un enjeu majeur pour l'industrie informatique. Que ce soit dans le contexte des systèmes d'information d'entreprise, du commerce électronique, de l'accès au savoir et aux connaissances ou même des loisirs, la pertinence de l'information délivrée, son intelligibilité et son adaptation aux usages et préférences des clients constituent des facteurs clés du succès ou du rejet de ces systèmes.

II.1. Personnalisation de l'information :

II.1.1. Problématique :

Dans les SRI classique l'évaluation des requêtes se fait sans tenir compte du contexte de l'utilisateur qui la émise. Ils supposent que l'utilisateur est complètement représenté par sa requête et que les résultats retournés pour une même requête sont identiques même si cette dernière est exprimée par des utilisateurs différents. Considérons a titre illustratif la requête « virus » émise par deux utilisateurs, un informaticien qui cherche des virus informatique et un biologiste qui cherche des virus biologique. Vu la non considération du contexte de l'utilisateur en cour du processus de recherche et face aux phénomènes actuels d'accroissement incessant d'informations (**volume**), ainsi qu'à leur **hétérogénéité** et leur **disparité**, cela pose des problèmes de l'ambiguïté du sens des mots, l'impossibilité de sélectionner des sources opportunes et l'inintelligibilité des résultats retournés. Ceci a pour corollaire la non pertinence des résultats de recherche et de fait, l'insatisfaction de l'utilisateur.

Volume : le volume d'information ne se mesure plus actuellement en giga-octets mais en téra-octets voire en péta-octets et exa-octets. Malheureusement, tous les algorithmes en RI ne sont pas de complexité linéaire en fonction du volume des informations ceci fait émerger le problème de passage a l'échelle qui engendre des dégradations des performances des processus de recherche **[Chevallet et al, 04]**.

Hétérogénéité : le web est caractérisé par une forte hétérogénéité des sources d'information. Cette hétérogénéité porte sur divers aspects : langue (plus de cent langue actuellement sur le web), média (texte, image, vidéo) etc.

Disparité : est une caractéristique qui traduit l'occurrence disséminée de l'information dans de large collections de documents, généralement interconnecté .les outils de navigation hypertextes sont, a ce titre, destinés a matérialisé la proximité des informations autour d'un besoin particulier. Cependant copte tenu du volume important d'informations disponible, les utilisateurs sont vite submergés par le nombre considérable de lien proposés, ce qui engendre les phénomènes fort connus de désorientation de l'utilisateur et de surcharge informationnelle.

Ces problèmes font émerger la personnalisation comme approche essentielle aux systèmes d'accès à l'information.

La personnalisation est considérée dès lors comme un aspect dominant dans plusieurs secteurs. Elle peut cibler les deux aspects d'un système (interface ou contenu) de manière spécifique ou simultanément :

Au niveau de la présentation : personnalisé des aspects de l'interface utilisateur, y compris les couleurs, les polices, le positionnement et l'affichage des données. Cet aspect correspond à ce que l'on nomme la customisation des systèmes. Les aspects de mobilité et d'environnement géographique des utilisateurs sont également pris en considération.

Au niveau du contenu : cibler la recherche en fonction des besoins et des centres d'intérêts des différents utilisateurs. Cet aspect correspond plus à ce que l'on nomme «*personnalisation*».

II.1.2. Définition :

La personnalisation est une dimension qui permet la mise en œuvre de systèmes centrés utilisateurs, non dans le sens d'un utilisateur générique mais d'un utilisateur spécifique et ce, en vue d'adapter son fonctionnement a son contexte précis **[Zemirli, 08]**. Elle implique une modélisation des besoin d'un utilisateur sous forme de profil décrivant ses centres d'intérêt, ses préférences et ses déférents contextes d'utilisation pour répondre de façon adaptée a un même type de requête émis par des utilisateurs de profils différent **[Bouzeghoub]**.

La personnalisation de l'information consiste à fournir à un utilisateur une information pertinente correspondant à ses préférences et à ses besoins **[Abdou et al, 06].**

Pour Kostadinov **[Kostadinov, 03]**, la personnalisation de l'information se définit par un ensemble de préférences individuelles, par des ordonnancements de critères ou par des règles sémantiques spécifiques à chaque utilisateur ou communauté d'utilisateurs. Ces modes de spécification servent à décrire le centre d'intérêt de l'utilisateur, le niveau de qualité des données qu'il désire ou des modalités de présentation de ces données.
Le Gartner Group **[Janowski et al, 01]** définit la personnalisation comme « toute interaction avec l'utilisateur dans laquelle le message, l'offre ou le contenu a été taillé sur mesure pour un utilisateur ou groupe d'utilisateur spécifiques ».

II.1.3. Objectif de la personnalisation :

La personnalisation a pour objectif d'intégrer l'utilisateur dans tout le processus de recherche, donc de faciliter l'expression du besoin utilisateur et de rechercher des informations sur un sujet en écartant l'information non pertinente et donc réduire considérablement l'espace de recherche d'une part et , d'autre part ,de rendre cette information sélectionnée intelligible a l'usagers et exploitable. Pour se faire il est nécessaire, au minimum, de présenter les résultats en fonction des attentes de chacun.

II.1.4. Domaine de la personnalisation :

On peut aborder la personnalisation selon différents point de vue : soit a travers des applications qui en ont besoin, soit a travers les technologies de base qui permettent de développer ces applications. Nous liston ci-dessous les deux points de vue et nous donnons un certain nombre d'exemples pour monter a quel point la personnalisation de l'information est devenue aujourd'hui un enjeu industriel et a quel point elle pénètre dans plusieurs domaines de recherche.

- **Domaines d'application :**

Les domaines d'application qui ont recours à la personnalisation sont nombreux, mais leurs besoins ne sont pas toujours identiques. Selon les domaines, la personnalisation de l'information consiste en l'une ou plusieurs des taches suivantes : filtrer un flux d'informations entrant pour éliminer le bruit **[Belkin 92],** guider la navigation dans un espace d'information trop vaste, recommander un ensemble d'information a l'utilisateur (nouvelles offres par exemple) **[Trousse, 01],** ajuster le résultat d'une requête selon le profil. Parmi les domaines qui ont le plus souvent recours à la personnalisation, on peut citer les suivants : Commerce électronique (e-commerce), dissémination sélective d'information (ce domaine

concerne la diffusion d'information culturelles ou d'actualité : informations journalistiques, recherche d'emploi ,forums, etc. la personnalisation permet en générale de filtrer le flux d'information en tenant compte d'un profil traduisant non seulement les centres d'intérêt de l'utilisateur, mais aussi sa langue, sa position géographique), apprentissage assisté par ordinateur (e-learning), accès aux bibliothèques électroniques (digital libraries) et systèmes d'information mobiles **[Bouzeghoub]**.

- **Domaines technologiques :**

L'offre en personnalisation de l'information est faite au sein des technologies informatiques qui permettent de développer les applications précédentes. Parmi ces technologies on distingue entre autre, les systèmes de bases de données, les moteurs de recherche d'information, les interfaces homme-machine et les intergiciels (middleware) **[Bouzeghoub]**.

II.2. les systèmes de recherche d'informations personnalisées :

II.2.1. Définition :

Un système de recherche d'information personnalisée (SRIP) est un SRI qui intègre totalement l'utilisateur tout au long du processus de recherche.

Il vise à augmenter le processus de recherche initié explicitement par la requête de l'utilisateur avec des caractéristiques informationnelles extraites explicitement/implicitement de l'utilisateur, dans le but d'améliorer ses différents besoins. . Il répond ainsi de manière personnelle aux besoins en informations de chaque utilisateur.

Toute information sur l'utilisateur, comme ses préférences, ses centres d'intérêt, ses besoins en information et son environnement de recherche sont de ce fait supposés pertinents et exploitables par le système de personnalisation. L'ensemble de ces informations va correspondre à ce que l'on nomme le **contexte de l'utilisateur** ou dans un cadre plus spécifique **profil utilisateur [Zemirli, 08].** Ces deux notions sont introduites dans ce qui suit**.**

II.2.2. contexte de recherche :

Le contexte de l'utilisateur peut être assimilé a l'ensemble de facteurs qui permettent de décrire ses intentions et perceptions de ce qui l'entour.ces facteurs peuvent couvrir divers aspects : psychologiques, sociaux, culturels, professionnels etc.…

D'après N.Fuhr **[Furh, 2000]**, le contexte possède trois principales dimensions : social, application et temps.

- **La dimension Sociale :**

  Définit le composant d'appartenance de l'utilisateur : individuel, groupe ou communauté.

- **La dimension Application** :

  Définit le contexte applicatif du besoin exprimé.

- **La dimension Temps :**

  Permet de décrire la circonscription temporelle du besoin exprimé : temps passé (batch), instant courant ou a court terme, intention ou long terme. Sous l'angle de cette dimension on distingue deux types de contextes avec des démarches de personnalisation appropriés :

  - **Le contexte courant ou a court terme :**

    Décrit les besoins et la préférence de l'utilisateur lors d'une session de recherche.

  - **Le contexte a long terme :**

    Décrit les besoins de l'utilisateur sur divers sessions de recherche.la personnalisation a long terme introduit des lors des mécanismes d'adaptation de contexte de l'utilisateur en fonction de la variation des ses besoins inscrit sur une longue période.

II.2.3. profil utilisateur :

L'utilisateur est un élément clé dans la personnalisation en RI. Il représente le noyau central d'un SRIP : il est la source, le déclencheur d'une RI et le seul à valider le résultat de cette recherche.

Le concept de profil utilisateur a été introduit pour l'accès à l'information en premier dans les travaux de filtrage d'information **[Beklin, 92],** pour décrire une structure représentative de l'utilisateur, en l'occurrence ses centres d'intérêts. Cette notion a ensuite été réexploitée en RI personnalisée pour former les composantes du contexte directement

dépendantes de l'utilisateur : centres d'intérêts, préférences, domaines professionnels, expertise, etc.

On appelle profil utilisateur toute structure qui permet de modéliser et de stocker les données caractérisant l'utilisateur. Ces données représentent les centres d'intérêts, les préférences et les besoins en informations de l'utilisateur ou un groupe d'utilisateurs **[BK, 05], [ZTB, 05].**

- **Le centre d'intérêt :**

  Exprime le domaine d'expertise de l'utilisateur ou son périmètre d'exploration ciblé au cours de sa session ou ses sessions de recherche. Il peut être définit par un ensemble de mots clés ou un ensemble d'expression logique (requête) **[Zemirli, 03/04]**.

- **Les préférences :**

  Peuvent être de différents niveaux tels que préférence de forme (style de la page, etc.) et préférence de domaine permettant permettent de cibler le centre d'intérêt de l'utilisateur **[Tamine, 05]**.

- **Besoin en information :**

  Un besoin en information est exprimé à travers une requête en utilisant un langage d'interrogation qui dépend du système de recherche d'information **[Achemoukh, 06].**

  La notion de 'besoin d'information' est centrale dans le domaine de la recherche d'information puisque elle est définie comme une interaction entre « un individu qui a besoin d'information » et « un document qui contient ou non la réponse à ce besoin » **[Mizzaro, 98].**

  L'utilisateur doit donc formuler une requête, c'est-à-dire exprimer son besoin en information sous forme de descripteurs ou mots clés plus au moins liés, dont la relation est exprimée par la présence d'opérateurs entre eux

Selon **[Gaussier, 03]**, « t*outes les variations qui caractérisent un utilisateur ou un groupe d'utilisateurs, peuvent se regrouper sous le terme de profil de l'utilisateur* ».

Un profil regroupe l'ensemble des connaissances nécessaire à une évaluation efficace des requêtes et à une production d'une information pertinente adaptée à chaque utilisateur. Il convient donc de distinguer la notion de profil de la notion de requête. Un profil est défini

comme une mise en équation du centre d'intérêt et des préférences de l'utilisateur, alors qu'une requête est l'expression d'un besoin circonstancié que l'utilisateur souhaite voir satisfait en tenant compte de son profil.

Un profil d'utilisateur peut être statique, quand l'information qu'il contient n'est jamais ou rarement changé (par exemple, l'information démographique), ou dynamique quand le profil d'utilisateur les données changent fréquemment, (par exemple, toutes les pages visitées peuvent être considérées comme intérêts d'utilisateur à de divers degrés).Une telle information est obtenue explicitement, en utilisant les fiches et les questionnaires en ligne ayant pour résultat des profils d'utilisateur statiques, ou implicitement, en enregistrant le comportement de navigation et les préférences de chaque utilisateur, ayant pour résultat profils d'utilisateur dynamiques **[Asfari, 11].**

II.2.4. Architecture fonctionnelle d'un SRIP :

Un système de recherche d'information personnalisé (*SRIP*) est un système qui intègre l'utilisateur, en tant que structure informationnelle, tout au long de la chaîne d'accès à l'information. Il inclut :

- Des modèles et algorithmes pour capturer et modéliser le but, les préférences et les centres d'intérêts de l'utilisateur ou un groupe d'utilisateur. Un modèle de profil est alors décrit et instancié.
- Une procédure de mise à jour du profil qui traduit son évolution dans le temps.
- Des mécanismes et algorithmes pour intégrer le profil de l'utilisateur dans le processus d'accès et retourner l'information pertinente en fonction de ce profil.

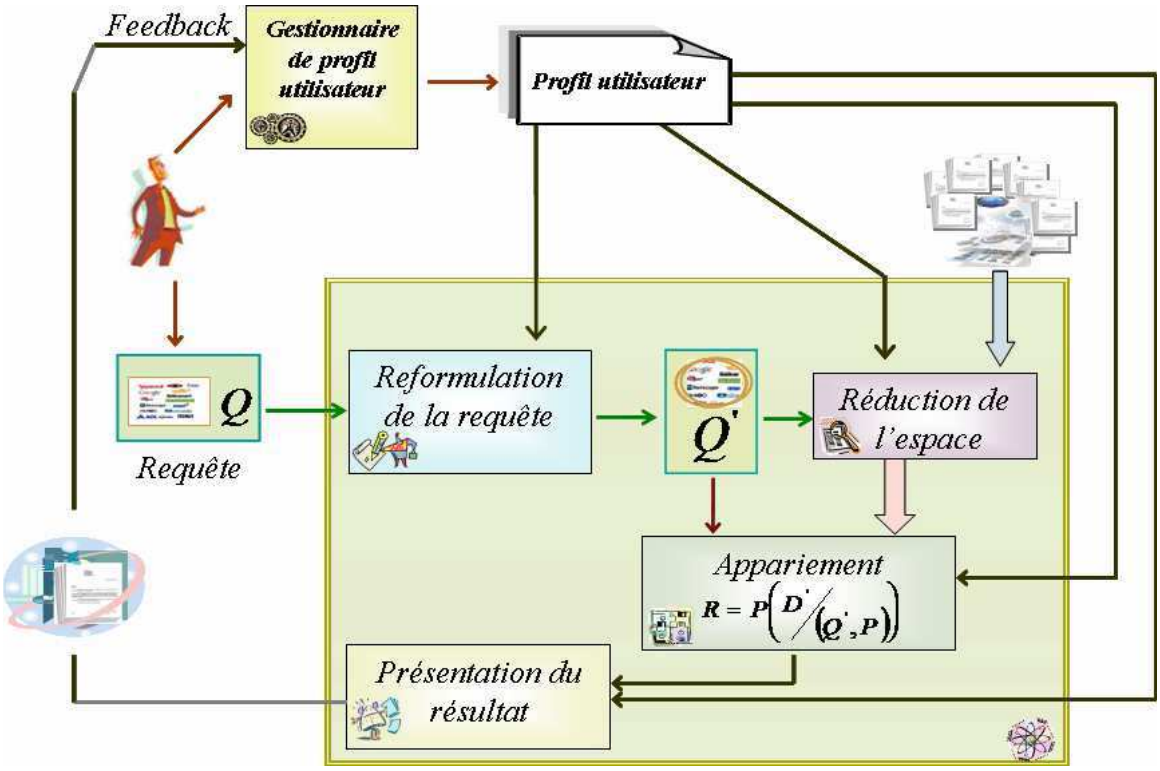


Figure II.1 : architecture fonctionnelle d'un SRIP [Zemirli, 08].

Cette figure met en évidence différents niveaux du processus de recherche la ou le profil utilisateur est intégré :

- ☞ Lors de la phase de reformulation de la requête afin de mieux cibler le contexte de la recherche de l'utilisateur,
- ☞ Lors de la phase de réduction de l'espace de recherche pour restreindre l'espace de recherche aux documents qui ciblent les besoins de l'utilisateur,
- ☞ Lors de la phase d'appariement pour calculer la pertinence des documents en fonction des caractéristiques spécifiques de l'utilisateur,
- ☞ Lors de la phase de présentation des résultats pour restituer les informations selon le contexte et les préférences de l'utilisateur.

Sur la base de cette architecture, nous dégageons principalement deux fonctions fondamentales qui sont la gestion des profils et la sélection de l'information.

II.3. Gestion des profils :

La représentation de l'utilisateur a travers la notion de profil permet de cibler ses besoins spécifiques dans le but d'améliorer les performances de recherche. **[Dan, 1986]** définit deux classes de modèles de profils utilisateur :

- Les modèles quantitatifs et empiriques : leur but est de modéliser le comportement externe de l'utilisateur (l'information qu'il consulte, celle qu'il sauvegarde, le temps passé pour lire l'information…).
- Les modèles analytiques et cognitif : leurs but est de comprendre le comportement interne de l'utilisateur (ses préférences de recherche, son but, ses connaissances …)

La gestion de profil permet de représenter, construire et évoluer le profil des utilisateurs.

II.3.1. Représentation du profil utilisateur :

La représentation de l'utilisateur a travers la notion de profil permet de mieux comprendre ses mécanismes cognitifs, notamment ceux permettant de percevoir le concept subjectif de la pertinence et au delà, cibler ses besoins spécifiques dans le but d'améliorer les performances de recherche. Dans le cadre de la RI, l'unité élémentaire utilisée pour représenter les paquets d'informations qui constitue le profil utilisateur est le terme pondéré **[Zemirli, 08]**. Un modèle de représentation permet d'organiser ces éléments afin de faciliter leur exploitation dans le processus d'accès à l'information. On distingue les principales approches de représentation : vectorielle, hiérarchique et multidimensionnelle.

II.3.1.1. La représentation vectorielle :

Ce type de représentation s'appuie généralement sur le modèle vectoriel **[Salton, 71].** Le contenu du profil est constitué d'un ou de plusieurs vecteurs définis dans un espace de termes. Ces termes sont obtenus à partir de plusieurs sources d'informations concernant l'utilisateur. Les coordonnées des vecteurs correspondent aux poids associés aux termes retenus dans le profil. L'utilisation de plusieurs vecteurs correspond à deux préoccupations : pouvoir prendre en compte des centres d'intérêt multiples et gérer leur évolution dans le temps. On peut citer comme exemple des systèmes : Alipes, WebMate et Surfagent **[Somlo, 03].**

Cette représentation est simple a mettre en œuvre et peut prendre en considération des centres d'intérêt multiple en utilisant plusieurs vecteurs, mais elle manque de structuration, et ne facilite pas l'interprétation et la prise en compte des différents niveaux de généralités caractérisant l'utilisateur **[Bottraud, 04].**

II.3.1.2. La représentation hiérarchique :

Dans cette approche, la modélisation de l'utilisateur est fondée sur l'élaboration d'une ontologie personnelle. L'ensemble des caractéristiques de l'utilisateur est organisé dans une

structure hiérarchique de concepts (catégories) où chaque catégorie représente la connaissance d'un domaine d'intérêt de l'utilisateur.

Le premier à avoir utilisé une telle structure fut Pretschner **[Pretschner, 99]** dans le système OBIWAN. On peut citer aussi le système SmartPush **[Kurki, 99] [Gauch, 03].**

II.3.1.3. Représentation multidimensionnelle :

La représentation multidimensionnelle permet de capturer puis catégoriser l'ensemble des informations caractérisant le profil utilisateur .dans cette approche le profil est structuré selon un ensemble de dimensions.

Différents travaux ont abordé cet aspect. Ainsi, les propositions de standards P3P pour la sécurisation des profils ont défini des classes distinguant les attributs démographiques des utilisateurs (identité, données personnelles), les attributs professionnels (employeur, adresse, type) et les attributs de comportement (trace de navigation).

Amato **[Amato, 99]** complète cette classification, dans le cadre d'un projet sur les librairies électroniques en ajoutant trois autres catégories : données collectées (contenu, structure et provenance des documents), données de livraison (moment ou moyens de livraison), et les données de sécurité (conditions d'accès aux informations du profil). Dans ce même cadre, Kostadinov **[Kostadinov, 03]** a poursuivi cette classification en proposant un ensemble de dimensions ouvertes, pouvant contenir la plupart des informations susceptibles de caractériser l'utilisateur. Dans sa représentation il distingue principalement huit dimensions décrites brièvement dans ce qui suit :

_ **Les données personnelles** :

Les données personnelles sont la partie statique du profil. Elles comprennent l'identité de l'utilisateur (nom, prénom, numéro de sécurité sociale, etc.), des données démographiques (âge, genre, adresse, situation familiale, nombre d'enfants, etc.), les contacts personnels et professionnels de l'utilisateur et d'autres informations comme le numéro de la carte bancaire ou de la carte Vitale.

_ **Le centre d'intérêt :**

Exprime le domaine d'expertise de l'utilisateur ou son périmètre d'exploration.il délimite intentionnellement l'espace de recherche des requêtes futur.

_ **L'ontologie du domaine :**

Complète la définition du centre d'intérêts en explicitant la sémantique de certains termes ou de certains opérateurs employés par l'utilisateur dans son profil ou dans ses requêtes. Par exemple on peut explicitement définir que 'BD' signifie base de données et non pas bande dessinée**.**

_ **La qualité attendue** :

Elle permet d'exprimer des préférences extrinsèques comme l'origine de l'information, sa précision, sa fraîcheur, sa durée de validité, le temps nécessaire pour la produire ou la crédibilité de sa source.

_ **La customisation** :

Concerne tout ce qui est lié aux modalités de présentation des résultats en fonction de la plateforme, de la nature et du volume des informations délivrées, des préférences esthétiques ou visuelles de l'utilisateur.

_ **La sécurité :**

La sécurité du processus exprime la volonté de l'utilisateur de cacher un traitement qu'il effectue.

_ **Le retour de préférences :**

Cette dimension regroupe l'ensemble des informations collectées sur l'utilisateur soit explicitement fournit par lui soit délivrés a son insu.(nombre de click effectué ,nombre de requêtes émises...).

_ **Les informations diverses :**

Regroupe différentes informations non classable dans les dimensions précédentes, mais pouvant aussi caractériser un profil par exemple la bande passante attribuée au gestionnaire du profil.

II.3.2. Construction de profil utilisateur :

La construction de profil traduit un processus qui permet d'instancier sa représentation L'approche de construction dépend fortement de la représentation choisie pour le profil utilisateur: les techniques utilisées par les systèmes différents selon qu'ils représentent le

profil par un (des) vecteur(s) de termes ou par des classes (hiérarchiques ou pas). Cependant la démarche de construction commune à tous les systèmes s'effectue donc en deux étapes : l'acquisition et la collecte des données utilisateur ; puis la construction proprement dite du profil.

II.3.2.1. La collecte des données utilisateur :

Cette phase consiste à collecter les informations pertinentes pour instancier le profil de l'utilisateur. Ce processus peut être explicite ou implicite :

- **Le processus explicite (feedback explicite) :**

Basé sur la collecte d'information directement fournies par l'utilisateur via l'interface du système en lui demandant par exemple de remplir des formulaires pour collecter les données personnelles et démographiques tels que sa date de naissance, son statut marital, son activité professionnelle et ses centres d'intérêt.

L'inconvénient de cette méthode, malgré la simplicité, est que l'utilisateur est parfois incapable d'exprimer ses besoins et ses préférences et de les définir de façon formelle. En plus, le processus de saisie de tous les paramètres du profil est souvent long et ennuyant. Cette méthode induit donc un désintéressement et l'abandon de l'utilisateur, ce qui en résulte une détérioration de l'efficacité du système de recherche.

La construction du profil dépend fortement du degré d'implication de l'utilisateur, s'il ne fournit pas volontairement les informations, aucun profil ne sera construit.

- **Le processus implicite (feedback implicite) :**

Repose sur un procédé d'inférence du contexte et préférences de l'utilisateur via son comportement (durée de lecture des documents, dernières pages visitées). Cette approche ne nécessite aucune implication directe de l'utilisateur.

Le procédé d'acquisition implicite du profil d'utilisateur peut être classifié comme suit :

. **Un modèle peu profond :** basé sur l'observation du comportement d'interaction relativement à court terme avec un système ; il ne tient pas compte des interactions de l'utilisateur avec le système durant les sessions précédentes.

. **Un modèle profond :** qui observe le comportement de l'utilisateur durant son interaction à long terme avec un système. Il tient compte des interactions de l'utilisateur avec le système durant les sessions précédentes.

II.3.2.2. la construction du profil :

Consiste à organiser et extraire les éléments qui constituent le profil à partir des données de l'utilisateur collectées lors de l'étape précédente. Après avoir regroupé ces sources de données, le système doit être capable de les analyser et d'en déduire le profil de l'utilisateur et également de les stocker dans la structure spécifiée qu'il utilise.

La construction s'appuie sur différentes techniques selon la représentation de profil utilisateur. On distingue trois principales approches :

- **Analyse statistique des termes :**

Basée sur des techniques d'extraction de mots clés, le rajout d'un poids exprime l'importance de chaque terme et est souvent associé à la fréquence d'apparition du terme. Cette approche ne peut être appliquée que sur des éléments textuels et les mots sont analysés en isolation avec le reste du document ce qui entraîne une perte d'information contextuelle pouvant dégrader l'exactitude des données du profil.

Il existe différentes structures de stockage et de représentation des mots clés : utilisation d'un vecteur simple en mettant directement les mots clés dedans, ou l'utilisation d'un vecteur de valeurs booléennes dans lequel chaque valeur correspond à la présence ou l'absence d'un mot donné. En plus dans certains cas, on peut rajouter un poids qui exprime l'importance de chaque terme et qui est souvent associé à la fréquence d'apparition du terme.

Une autre méthode de représentation plus riche en sémantique est présentée par **[Sorensen, 95]** .cette représentation consiste a stocker les termes sous forme de graphe orienté, dans lequel les sommets sont les mots et les arcs expriment les relations entre ces mots. Ces relations sont caractérisées par la probabilité qu'un mot apparaisse après un autre dans le texte (s'il y a un arc entre « langage » et «naturel » dont le poids est 0.8, cela signifie que dans 80% des cas le mot naturel est précédé par le mot langage) **[kostadinov, 03].**

- **Techniques d'apprentissage :**

Utilisent généralement des algorithmes de classification qui extraient les termes à partir de ces différentes sources et les regroupent en des classes indiquant les domaines des centres d'intérêts de l'utilisateur. L'ajustement du poids des termes fait de plus en plus appel à des techniques d'apprentissage, comme des réseaux de neurones, des probabilités Bayésiennes, des algorithmes à base de règles.

L'avantage de l'approche est la fraîcheur et l'exactitude des données dérivées. L'inconvénient se trouve dans la complexité des algorithmes utilisés qui nécessitent beaucoup de temps **[Zemirli, 03/04]**.

- **Concept de la vie artificielle :**

Le profil d'utilisateur, est décrit par un ensemble d'éléments représentant ces centres d'intérêts. Chaque élément est décrit par un vecteur Vi de mot-clé et une valeur Ei d'énergie traduisant le degré d'importance du centre correspondant.

Les centres pour lesquelles l'énergie diminue auront tendance à disparaitre, dans le cas contraire d'autre vecteur sont alors générés décrivant ainsi les intérêts de l'utilisateur dans un certain niveau de détail. L'utilisation de la théorie de la vie artificielle est une approche assez novatrice pour construire et aussi mettre à jour le contenu du Profil **[Chen, 02]**.

II.3.3. Evolution du profil utilisateur :

L'évolution des profils réside dans la capacité du système de personnalisation de détecter les changements des centres d'intérêt pour effectuer la mise à jour du profil .elle se fait par l'addition de nouvelles informations ,ce qui implique des changement au niveau des centres d'intérêt qui conduisent a la suppression de quelque domaines ou a l'émergence d'autre domaines .

L'évolution consiste à adapter la structure et/ou le contenu du profil aux changements des centres d'intérêt et aux variations des besoins en information de l'utilisateur.

Dans le cas d'une représentation ensembliste, le profil utilisateur évolue en ajoutant de nouveaux vecteurs de termes extraits des documents correspondant aux centres d'intérêt détectés de l'utilisateur. Comme il n'y a souvent pas de dépendance entre les vecteurs, l'ajout d'un nouveau vecteur ne fait qu'augmenter le nombre des centres d'intérêts et non le degré d'importance du domaine.

Dans le cas d'une représentation en classes (hiérarchiques), pour chaque nouveau document Le système recalcule, sa similitude avec les classes déjà existantes, ce document est assigné a une classe de profil si il appartient cette classe, le centre d'intérêt est ainsi met a jour, si non une nouvelle classe est criée par le système traduisant un nouveau centre. L'adaptation de la structure du profil aux nouvelles classes s'effectue en mettant à jour les relations entre ces classes **[Zemirli, 08]**.

II.4. Sélection de l'information :

Ce processus consiste à intégrer le profil utilisateur préalablement construit dans le processus de recherche d'information proprement dit .en ce sens les informations contenues dans le profil courant sont exploitées pour identifier éventuellement le profil parmi ceux qui sont en cour de construction, réécrire puis exécuter la requête, enfin présenter les résultats de la recherche **[Tamine, 05]** :

- **Identification du profil :**

Consiste à apparier la structure instanciée du profil avec ceux définis préalablement dans le système.

- **Exécution des requêtes :**

Traduit la succession éventuelle des opérations de sélection de sources d'information, reformulation et calcule de score de pertinence.

- **Présentation des résultats :**

La présentation des résultats est la phase ultime du processus d'accès a l'information, cette phase peut également considérer le profil utilisateur en réordonnant les résultats fournis par le processus de sélection, l'ordre finale des documents fournis a l'utilisateur est généré a partir de la combinaison de l'ordre des résultats produit par le processus de sélection et celui donné par le contexte utilisateur via un calcule de similarité ou jugement explicite de pertinence.

II.5. Mise en œuvre d'un SRIP :

L'introduction de la dimension utilisateur dans le processus d'accès à l'information nécessite la modélisation de l'entité utilisateur.

II.5.1. Modélisation de l'utilisateur :

La modélisation de l'utilisateur est une discipline de recherche datant des années 70 et évoquant en premier lieu les travaux d'Allen, Cohen et Perrault **[All, 79] ; [CP, 79].**dans le but d'améliorer la qualité des interactions homme-machine.

Les systèmes qui adaptent leur comportement aux besoins de l'utilisateur individuel ont souvent une structure explicite de représentation qui contient des informations sur leurs utilisateurs ; cette structure s'appelle généralement un modèle d'utilisateur.

Le modèle utilisateur est une source de connaissances, une base de données sur un utilisateur. Plus précisément il représente un ensemble de données persistantes qui caractérisent un utilisateur ou un groupe d'utilisateurs particuliers en contenant des caractéristiques sur les préférences, les connaissances, les objectifs, les centres d'intérêt, etc. de l'utilisateur.

En général, dans la littérature, les termes « modèle utilisateur » ou « profil utilisateur » signifient la même appellation. Mais d'après Koch **[KOC, 00],** une différence existe entre le profil utilisateur et le modèle. Il définit le profil utilisateur comme une version simple du modèle utilisateur, Le profil utilisateur peut être vu comme une collection d'informations personnelles caractérisant un utilisateur telles que les buts, besoins, compétences, préférences, contexte d'utilisation, etc. Ces informations sont stockées sans aucune interprétation ou description. Ces informations sont généralement représentées par des couples (attribut, valeur). Ces valeurs sont soit stables soit évolutives. Le profil utilisateur est utilisé pour construire le modèle utilisateur. Donc ce dernier est vu comme une représentation des croyances du système sur un utilisateur donné.

[KOC, 00] définit l'application du modèle utilisateur comme suit:

« Users are different: they have different background, different knowledge about a subject, different preferences, goals and interests. To individualise, personalise or customise actions a user model is needed that allows for selection of individualised responses to the user."

Donc le modèle utilisateur et le profil utilisateur sont nécessaires dans le processus de personnalisation.

Dans ce qui suit on va donner un aperçu des différentes approches et technique de modélisation.

II.5.1.1. Approches :

On distingue trois principales approches **[Gow, 03]** : l'approche canonique, l'approche explicite et l'approche automatique.

- **Approche canonique :**

Cette approche préconise l'intégration de modèles utilisateur typique lors de la conception du système. Les interactions permettent de cataloguer l'utilisateur courant par rapport a un modèle prédéfini dans le système. Cette approche a été peu performante ; notamment en raison de l'inadéquation des langages des concepteurs et des utilisateurs pour décrire les situations permettant d'apparier l'utilisateur a un modèle canonique **[GRE, 84].**

- **Approche explicite :**

Dans le cas de cette approche, le système maintient un panel de modèles canoniques caractérisé par une parie flexible contrôlée par l'utilisateur lors de ces interactions avec le système. Cette approche remédie aux inconvénients de l'approche décrite précédemment en réduisant l'erreur de catalogage due a une description incertaines des situations, mais elle induit une surcharge cognitive pour l'utilisateur et une complexité dans la conception du système.

- **Approche automatique :**

Cette approche préconise d'inférer le modèle de l'utilisateur a partir des informations collectées implicitement lors de ses sessions d'utilisation du système et non pas a partir de ses interaction explicite avec le système, et ce, dans le but de pallier au problème de surcharge cognitive et l'incertitude engendré par les deux approche précédentes. Deux principales classes de techniques sont issues de cette approche : les techniques collaboratives et les techniques statiques.

II.5.1.2. Techniques :

- **Les techniques collaboratives :**

Sont basées sur l'idée de prédire le modèle individuelle d'un utilisateur courant sur la base d'un comportement assimilable à celui d'un groupe d'utilisateurs. Les utilisateurs du système participent collectivement a alimenter des stéréotypes qui sont affectés a des groupes d'intérêts communs puis utilisés pour prédire les préférences inconnues de nouveaux utilisateurs. L'approche reste peu performante pour un nouvel utilisateur avec peu d'informations collectées à partir d'un groupe et d nouveaux centres d'intérêt.

- **Les techniques statistiques :**

Ces techniques sont basées sur des modèles théoriques issus de la statistique, soutenus par des heuristiques et algorithmes appropriés.les principaux modèles sont :

- **Le modèle linéaire :**

  L'hypothèse de base est que la valeur de prédiction présumé et inconnu d'un objet cible du système est une combinaison linéaire des valeurs calculées a partir d'un comportement passé de l'utilisateur. Le modèle linéaire peut être combiné avec des technique collaboratives ou les valeurs connues sont issues de l'appréciation des membres du groupe associes a l'utilisateur courant.

- **Le modèle markovien :**

  Est basée sur l'hypothèse markovienne qui permet de présenter une séquence d'événement ultérieur sur la base d'un nombre fixe d'événement antérieurs.la théorie markovienne offre des éléments pour calculer la probabilité d'occurrence des événements futurs.

- **Réseaux de neurones :**

  Sont destinés a résoudre des problèmes de décision non linéaires. Dans le cas précis du vaste domaine d'application da la modélisation de l'utilisateur, l'entrée représente une situation ou faits observables a partir de l'utilisateur, les sorties représentent des objets cibles du système avec des valeurs d'activation qui traduisent le degré de prédiction.

- **Classification :**

  Cette méthode permet de partitionner un espace d'objets en classe de manière a réduire sa dimension .les objets d'une même classe ont des propriétés partageable. De point de vue de la modélisation utilisateur, cette méthode permet généralement d'identifier la classe de caractéristiques de l'utilisateur courant à partir d'information drivées de son comportement.

- **Les réseaux bayésiens :**

  Les réseaux bayesiens **[Pearl, 88]** sont des graphes acycliques orientés ou les nœuds correspondent à des variables aléatoires, les liens orientés représentent les liens de causalité entre les nœuds parents et nœuds fils.ils permettent de représenter

explicitement les relations de causalité entre faits et d'émettre des prédictions sur de nombreux paramètres du système.

II.6. Evaluation des SRIP :

L'évaluation des SRI est depuis le début des travaux sur la RI un des piliers de l'évolution de ce domaine. La démarche de validation en RI se base sur l'évaluation expérimentale des performances du modèle ou du système proposé, selon le modèle de Cranfield **[Cle, 67].** L'évaluation des performances d'un modèle de RI, permet de paramétrer le modèle et d'estimer l'impact de chacune de ses caractéristiques.

Les méthodes d'évaluation largement adoptées en RI, sont empirique (évaluation par observation expérimental).Elle sont souvent basée sur une évaluation d'avantage quantitative que qualitative. Ce type d'évaluation est orienté vers une approche comparative de plusieurs systèmes reposant sur le principe d'évaluation des collections de test. Bien qu'adopté par des compagnes d'évaluation tels que TREC. Ces méthodes sont contestées en raison de non prise en considération du contexte de recherche et de la perception de pertinence utilisateur dans ce même contexte. Elles ne sont donc pas adaptées à la recherche d'information personnalisé.

L'évaluation orientée vers l'utilisateur est une composante primordiale dans le cadre de l'accès personnalisé à l'information. En effet, les objectifs d'une telle évaluation sont de mesurer l'adéquation des profils utilisateur construits par le système avec les centres d'intérêt effectifs de l'utilisateur ; ainsi que l'impact de l'intégration de ce profil, dans le processus d'accès, sur les performances de recherche. L'introduction de la dimension utilisateur dans le processus d'accès à l'information, accentue d'avantage la difficulté d'évaluation. en effet, en plus de l'absence de collections de tests standard pour évaluer l'efficacité de l'accès personnalisé à l'information, la recherche dans ce domaine est confrontée à l'inexistence de méthodologies formelles, de mesures standards d'évaluation de l'adéquation des profils appris aux centres d'intérêt de l'utilisateur, ni l'existence de système référentiel. Il est d'autant plus difficile de réaliser des scénarios d'évaluations objectifs en intégrant la dimension de l'utilisateur dans le processus d'accès pour les principales raisons suivante :

Si l'évaluation est effectuée par différents utilisateurs alors les différences personnelles (l'intelligence, la capacité de raisonnement, l'expérience) exogènes a l'expérimentation, ont un impact sur la perception de la pertinence, ainsi si un utilisateur est impliqué dans les différents scenarios d'évaluation alors son expérience passée avec le système peu influer sur sa perception de la pertinence.

Les conditions de déroulement des expérimentations ont aussi un impact non négligeable sur les résultats tels que la configuration des machines (un processus lent conduit a la fatigue de l'utilisateur), interfaces peu ergonomique …). De ce fait pour une évaluation acceptable des SRIP, il faut respecter certaines conditions pour éviter les erreurs de mesure de pertinence **[Chi, 01].** Ces conditions sont résumées comme suit :

- ☞ Définir un nombre suffisant de groupes d'utilisateurs avec des effectifs adéquats.
- ☞ Isoler au mieux les utilisateurs.
- ☞ S'assurer de l'ergonomie des applications.
- ☞ Préparer un canevas unique qui décrit le protocole d'expérimentation et l'adresser a l'ensemble des utilisateurs.
- ☞ Les utilisateurs ne doivent pas être informés des facteurs à évaluer dans le système ; leurs appréciation doit porter sur des aspects perceptibles, liés a son fonctionnement.
- ☞ Les variables exogènes doivent être identifiées explicitement et leur influence mesurée pour être prise en compte dans le processus global d'évaluation.

Conclusion :

La personnalisation de l'information immerge comme une approche capitale dans le développement des systèmes du futur. Elle a pour but de fournir une information pertinente, qui correspond exactement aux besoins de l'utilisateur.

Répondre aux besoins en information des utilisateurs d'une manière personnelle, ne peut se faire sans inclure l'utilisateur dans le processus de RI. Inclure l'utilisateur dans le processus de RI implique la représentation de ce dernier dans un modèle ou par une structure qui permet son exploitation par le SRI.

Introduction :

Les réseaux bayésien sont la combinaison des approches probabilistes et de la théorie des graphes .Autrement dit, ce sont des modèles qui permettent de représenter des situations de raisonnement probabiliste à partir de connaissance incertaines. Ils constituent une technique d'acquisition, de représentation et de manipulation de connaissance et on les utilise, surtout, pour leur capacité d'effectuer des inférences dans un contexte d'incertitude .Ils sont utilisés pour prévoir, contrôler et simuler le comportement d'un système, à diagnostiquer les causes d'un phénomène observé, à analyser des données et à prendre des décisions. D'ailleurs les domaines d'application sont variés.

La première utilisation des RBs en recherche d'information est apparue dans les années 80 mais s'est largement développée par les travaux de **[Turtle, 91], [turtle et al, 91].** Elle montre qu'un modèle de RI basé sur un réseau bayésien est plus générale et peut englober d'autres modèles comme le modèle probabiliste, booléen ainsi que la pondération tf-id du modèle vectoriel.

Le principal avantage apporté est de pouvoir combiner des informations provenant de différentes entités (requête, termes et documents) pour restituer les documents qui seraient les plus pertinents étant donnée une requête. **[Ricardo, 99]** indique que « le Réseau Bayésien fournit un formalisme clair pour combiner les différentes sources d'évidences (requêtes passées, cycle de rétroaction, formulation de requêtes) pour le calcul de la correspondance entre la requête et les documents ». C'est pour cette raison que différents travaux **[RIB, 96], [Brini, 05]** ont tenté d'exploiter l'apport des RBs pour définir des modèles de recherche d'information.

Les notions et les théorèmes principaux de la théorie des probabilités nécessaires à la compréhension du réseau Bayésien sont présentés dans l'annexe.

III.1. Définition d'un réseau bayésien :

Les réseaux bayésiens (dit « réseau de croyance » ou « réseau probabiliste ») sont des modèles qui permettent de présenter des situations de raisonnement probabiliste basé sur le théorème de **bayes [Fienberg, 05].** C'est un résultat de base en théorie des probabilités, issu des travaux du Thomas Bayes (1702-1761), présenté à titre posthume en 1764. Cette théorème dit que, si l'on a deux événements A et B, alors :

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

Ainsi ,les RBs associent une partie qualitative que sont les graphes et une partie quantitatives représentant les probabilité conditionnelles associées à chaque nœud du graphe relativement au parent.la partie qualitative exprime des indépendances conditionnelles entre variable et des liens de causalités et ce grâce a un graphe orienté acyclique dont les nœuds correspondent a des variables aléatoire.la partie qualitative est constituée de tables de probabilités .un réseau bayésien est donc définit **[Pearl, 88]** par un graphe dirigé ,un espace probabiliste et un ensemble de variables aléatoires. Le graphe est sans circuit G = (V, E) où V comprend des nœuds qui représentent des variables aléatoires du domaine, et E un ensemble d'arc qui représentent des liens de causalité entre nœud parent et nœud fils, a chaque nœud est associé une table de probabilités de cette variable en fonction des valeurs parents.

Pour construire un Réseau Bayésien:

1. Choisir un ensemble de variables pertinentes ordonné X1, X2, ..., Xm.
2. Pour i=1 à m
3. Ajouter Xi au graphe
4. Parents(Xi) = sous-ensemble minimal de {X1, ..., Xi-1} tel que il y ait indépendance conditionnelle de Xi et des éléments de {X1, ..., Xi-1} étant donné Parents(Xi)
5. Définir la table de probabilités P (Xi=k|valeurs affectées à Parents(Xi).

Dans le contexte de la recherche d'information les nœuds et les arcs sont définis comme suit :

– Les nœuds : représentent des concepts, des groupes de termes ou des documents.

– Les arcs : représentent les dépendances entre termes et entre termes et documents.

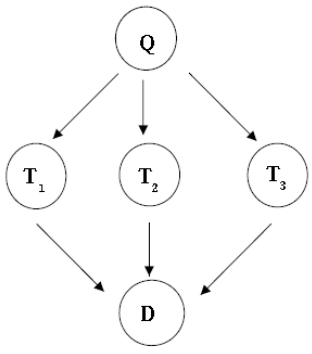


Figure III.1 : Modèle de réseau bayésien simple [Nassr, 02].

Le réseau de la figure illustre le réseau de Turtle **[Turtle et al, 90]** de pertinence d'un document vis à vis d'une requête composée de trois termes.

III.2. Utilisations et difficultés :

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 donnent lieu à de nombreuses recherches. Les RBs sont caractérisé par les aspects suivant qui les rendent préférable par rapport à d'autre modèle tel que les réseaux de neurones, système expert, arbre de discision, etc **[Nguyen, 05]** :

- **Acquisition des connaissances :** La possibilité de rassembler et de fusionner des connaissances de diverses natures dans un même modèle.

- **Représentation des connaissances :** La représentation graphique d'un réseau bayésien est explicite, intuitive et compréhensible par un non spécialiste, ce qui facilite la validation du modèle, ses évolutions et son utilisation.

- **Utilisation des connaissances :** Un RB 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 logciels :** 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.

Une 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

nœud conditionné par rapport à ses parents, il reste que la définition de ces tables n'est pas toujours facile pour un expert.

II. 3. Différents modèles graphique des réseaux bayésiens :

Il existe plusieurs variantes des réseaux bayésien tel que **[Smail, 04]** : les RB multi agents, les RB de niveau deux, les RB orienté objet, les diagrammes d'influence, les RB dynamiques (temporels) **[Eduardo, 05],** les RB multi entité, les filtres bayésiens qui sont les RB dynamiques particulier et les RB adaptés a la classification tels que ; les réseaux naïf, les RB naïf augmenté, etc.

III. 4. Construction des réseaux bayésiens :

La construction d'un réseau bayésien s'effectue en trois étapes essentielles :

- Identification des variables et de leurs espaces d'état
- Définition de la structure du réseau bayésien
- Définition de la loi de probabilité conjointe des variables.

III.4.1. Identification des variables et de leurs espaces d'état :

Dans la première étape, l'utilisateur définit l'ensemble des variables du système, en précisant l'espace d'états de chaque variable, à savoir l'ensemble de ses valeurs possibles. En général, les variables considérées sont aléatoires mais il est possible d'introduire des valeurs déterministes liées a des observations particulières ou simplement pour la compréhension globale du système.

III.4.2. Définition de la structure du réseau bayésien :

Dans la structure du réseau, l'utilisateur identifie les liens entre les différentes variables. Il s'agit simplement de relier des causes et des effets par des flèches orientées. Cependant s'il existe une relation causale de A vers B, toute information sur A peut modifier la connaissance sur B et réciproquement. L'ensemble des nœuds et des flèches forme la structure du réseau bayesien : c'est donc la représentation qualitative de la connaissance.

III.4.3. Définition de la loi de probabilité conjointe des variables :

La dernière étape de construction consiste à affecter à chaque nœud une table de probabilité qui représente la distribution locale de probabilité. Suivant le type de nœuds, deux cas de figure se présentent :

- si la variable n'a pas de cause, d'où l'obligation de préciser la loi de probabilité marginale associée.
- Si la variable possède différentes causes, d'où la nécessité de préciser la dépendance en fonction de ces causes par une table de probabilités conditionnelles, ou par des relations déterministes qui conduisent au calcul de cette table, Cette dernière étape constitue la représentation quantitative de la connaissance.

III.5. Principe du Réseau Bayésien

Les Réseaux Bayésiens (RB) sont des modèles probabilistes qui s'appuient sur des graphes traduisant par des nœuds les variables du système et par des arcs l'existence de liaisons directes entre ces variables.

L'étude d'un modèle de Réseau Bayésien nécessite 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 **[Halloulis, 04]** :

- **Apprentissage ou constitution du réseau :** Il s'agit ici de trouver la structure et les probabilités associées du réseau, à partir des données de la base et de traitements principalement statistiques.

- **Inférence Bayésien :** A partir des résultats de la première phase, le réseau permet la propagation d'information à l'intérieur de la structure, permettant toute interrogation sur la base et peut fournir pour chaque état partiel ou complet de la base (instanciation partielle ou complète des variables de celle-ci) des probabilités d'occurrence de toutes les valeurs possibles de toutes les variables.

III.6. Relations de dépendance :

Les réseaux sont utiles pour calculer de façon locale l'impact de la modification d'une information d'une variable sur les états des autres variables. Le changement d'un état d'une variable suite à la réception d'une information dans le réseau dépend de la topologie du graphe, et trois situations principales sont possibles. Un exemple est donné dans l'annexe.

- **Connexion en série :** Soit la situation de la figure 1. A à une influence sur B qui a une influence sur C. L'information peut circuler de A vers C ou de C vers A `a travers

B dans les deux cas. Par contre, si B est connue ou instanciée, la voie est bloquée et A et C deviennent indépendants. On dit dans ce cas, que A et C sont d-séparés étant donnée B, lorsque B est instanciée.

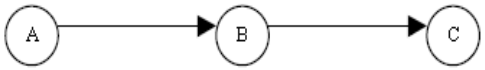


Figure III.2 : Connexion en série.

_ **Connexion divergente :** L'information peut passer entre les enfants de A lorsque la variable A est non instanciée. Dans la figure 2, les enfants B, C, D sont dits d-séparés par A.

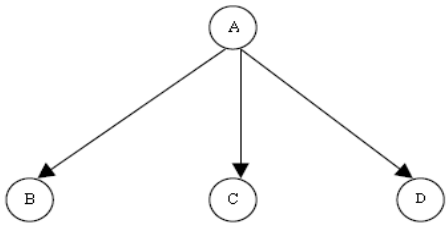


Figure III.3 : Connexion divergente.

_ **Connexion convergente :** Dans ce type de connexion telle que décrite dans la figure 3, lorsque aucune information n'est donnée sur le nœud fils mis à part l'information apportée par les parents, les parents sont dits dans ce cas indépendants. Par contre, si l'´etat du fils est connu alors la cause, c'est-`a-dire un des états des parents va pouvoir donner de l'information sur les états des autres parents. L'information peut circuler dans une connexion convergente uniquement lorsque la variable de la connexion ou un de ses descendants a reçu de l'information.

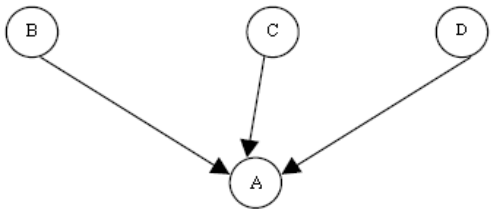


Figure III.4 : Connexion convergente.

III.6.1. La d-séparation : Les situations décrites ci-dessus recouvrent les manières possibles de transmettre l'information à travers un réseau. Deux variables distinctes A, B d'un réseau

sont d-séparées, si pour tout chemin entre A et B, il existe une variable intermédiaire C, distincte de A et de B telle que :

– soit la connexion est en série ou divergente et C est instanciée ;

– ou la connexion est convergente et ni C, ni ses descendants ne sont instancies.

Ainsi, si deux variables A et B sont d-séparées, alors tout changement d'´etat dans A n'aura pas d'impact sur l'´etat de B.

III.7. Probabilités conditionnelles :

Les réseaux Bayesiens sont des modèles graphiques probabilistes permettant de représenter les influences entre des événements. Un réseau Bayesien est défini par un graphe acyclique orient´e G = (V, L). Dans ce graphe, V représente l'ensemble des nœuds du graphe et L l'ensemble des arcs reliant des paires de nœuds. Chaque nœud Vi représente une variable aléatoire associée à une distribution de probabilité, et chaque arc définit une influence du nœud de départ sur le nœud d'arrivée. La distribution de probabilité associée à une variable spécifie les probabilités de ses états conditionnellement aux états des variables qui l'influencent. On note P (Vi / Parents(Vi)) ou Parents(Vi) représentent l'ensemble des parents de la variable Vi.

III.8. Modèle Bayésien en RI :

Des travaux récents ont permis d'exploiter l'apport des Réseaux Bayésiens (RBs) pour définir des modèles de RI. L'avantage apporté par l'utilisation de ces réseaux a été principalement de pouvoir combiner des informations provenant de différentes sources pour restituer les documents qui seraient les plus pertinents étant donnée une requête.

III.8.1 Architecture générale du modèle Bayésien :

La figure présente l'architecture générale du modèle de RI basé sur les réseaux Bayésiens. Les nœuds du réseau dans un modèle BNR (modèle RI basé sur les réseaux Bayésiens) **[De Campos et al, 02] [De Campos et al, 03]** ont été décomposés en deux ensembles de variables *T* et *D :*

- L'ensemble des termes *T* = (T1, T2,……, TM), où M est le nombre de termes dans la collection ;

- L'ensemble des documents de la collection *D* = (D1, D2,……, DN), où N est le nombre de documents dans la collection.

Les domaines des nœuds sont binaires {vrai, faux} signifiant que le nœud est instancié ou non.

T est l'ensemble des nœuds termes; une variable *Ti* associée à un terme prend ses valeurs dans le domaine dom(*Ti*) = {*ti*, $\bar{t}i$}, ou *ti* désigne le fait que le terme *Ti* est non pertinent et *ti* désigne le fait qu'il est pertinent. Un terme est considéré comme pertinent si tous les documents qui le contiennent sont jugés pertinents par l'utilisateur et non pertinent sinon.

D est l'ensemble des nœuds documents, une variable *Dj* prend ses valeurs dans le domaine dom(*Dj*) = {dj, $\bar{d}j$}, *ou dj* signifie « le document *Dj* n'est pas pertinent » et *dj* signifie « le document *Dj* est pertinent ». Un document est pertinent s'il répond au besoin utilisateur.

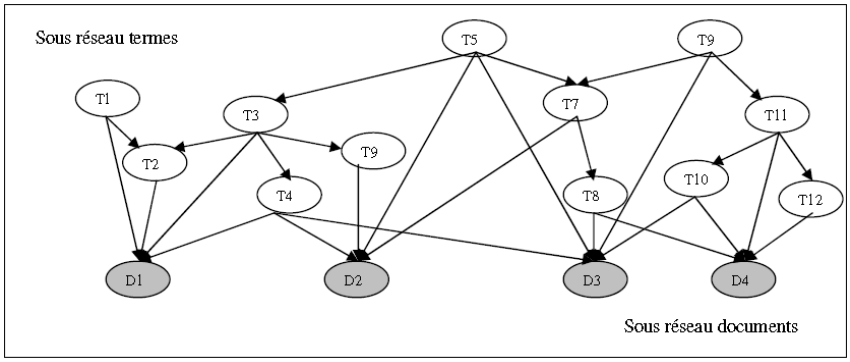


Figure III.5 : Architecture générale du modèle Bayésien [Elayeb, 09].

Les RBs fournissent un formalisme pour fusionner des informations provenant de différentes sources, et ont permis de combiner différentes approches de RI. Les modèles les plus connus en RI utilisant les RBs sont les Réseaux d'Inférence **[Turtle et al, 90]** et les Réseaux de Croyance.

III.8.2. Le Réseaux Bayésiens d'Inférence :

La naissance du modèle d'inférence est le résultat de l'extension de deux idées :

- la proposition d'utiliser des logiques non classiques pour déterminer le degré auquel un document implique ou correspond à une requête **[Van 89].**
- la notion d'inférence plausible et la possibilité de combiner plusieurs sources pour inférer la probabilité de pertinence d'un document étant donné une requête **[Croft et al, 87**].

Un réseau d'inférence en RI est matérialisé par un graphe orienté sans cycle. Les nœuds du graphe correspondent à des concepts, à des groupes de mots ou à des documents Un nœud particulier va représenter la requête. Les arcs du graphe représentent des relations sémantiques entre les nœuds. A ces nœuds sont associés des probabilités de croyance. Ce modèle repose

sur le théorème de Bayes pour l'expression de la probabilité conditionnelle et sur la stratégie d'activation propagation.

La recherche peut être donc considérée comme un processus de raisonnement incertain pour estimer la probabilité qu'un document satisfasse la requête.

III.8.2.1. Architecture générale :

Les Réseaux d'Inférence sont utilisés dans le système INQUERY [**Turtle et al, 90] [Turtle, 91] [Turtle et al, 91]** et ses performances sont liées à sa capacité à représenter différentes approches de la RI et à les combiner dans un seul modèle.

Le réseau d'inférence est composé de deux réseaux : le réseau document ainsi que ses termes d'indexation et le réseau requête. Le réseau document représente les documents de la collection et contient différents schémas de représentation (résumés, textes, etc.). Les nœuds du réseau requête représentent les concepts de la requête et le besoin utilisateur. Les réseaux document et requête sont liés par l'intermédiaire des nœuds termes d'indexation.

La figure suivante décrit le modèle de base proposé.

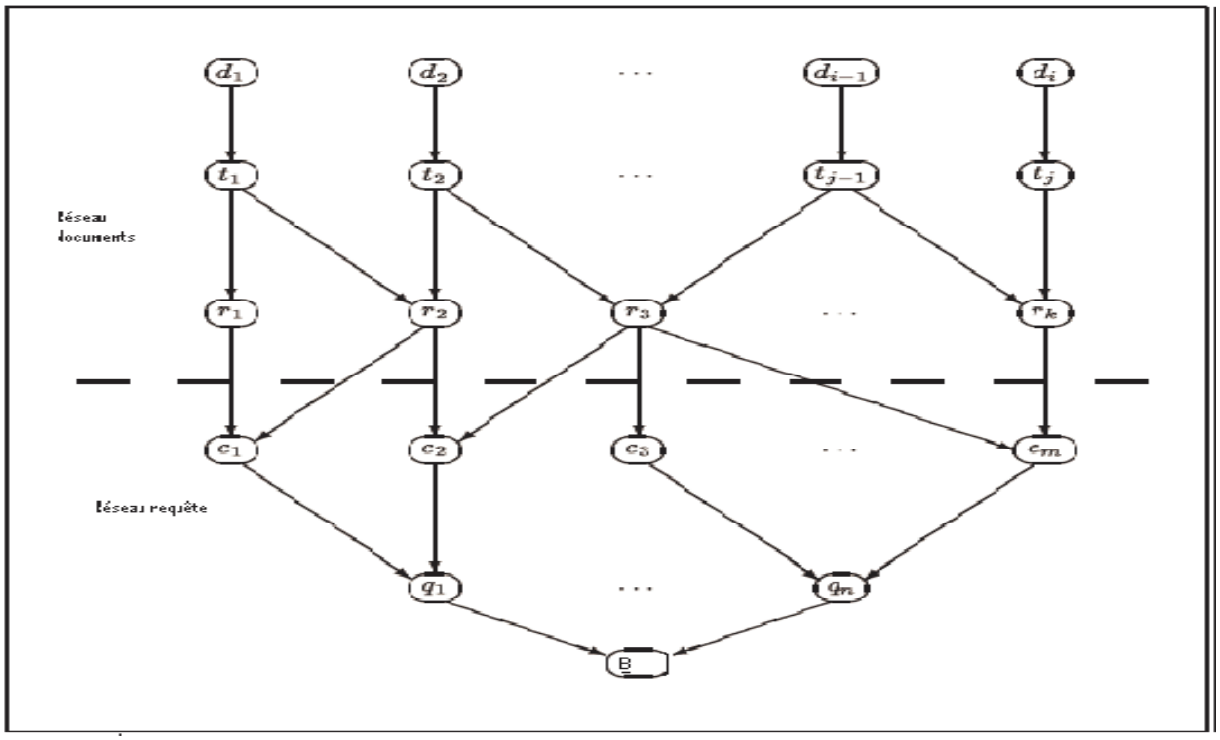


Figure III.6 : Architecture générale [Brini, 05].

Le réseau document représente les nœuds documents (dj) de la collection, les nœuds de représentation des textes (ti), les nœuds de représentation des concepts (rk). Un nœud document correspond a l'événement qu'un document donné de la collection est observé. Un nœud de représentation de texte correspond au texte indexant le document et correspond à l'événement qu'un texte d'un document est rencontré. Les auteurs ont considéré dans leur approche de base les documents de type textes bruts mais suggèrent l'extension aux figures, images, aux documents multimédias etc. Il existe une correspondance entre chaque représentation de texte et le document auquel il se réfère.

La dépendance entre un document et un texte est symbolisée par un arc entre les nœuds document et texte. Les nœuds de représentation de concepts correspondent a différentes techniques d'indexation utilisées pour obtenir les concepts d'indexation des documents comme par exemple une indexation automatique et une indexation manuelle. Un même concept peut ainsi être généré par les deux techniques d'indexation, et l'arc reliant dans ce contexte le concept au document aura deux sens différents.

Les domaines de tous les nœuds sont binaires {vrai, faux} signifiant que le nœud est instancié ou non. Par exemple, un nœud représentant le texte aura pour instanciation vrai uniquement lorsque son nœud parent, document, est aussi instancie à vrai.

Le réseau requête est un graphe acyclique oriente représentant le besoin utilisateur (B) et des nœuds racines qui représentent les concepts de ce besoin (ci). Plusieurs expressions de la requête peuvent être utilisées et représentées dans ce réseau (qk). Les auteurs suggèrent de simplifier cette représentation en supprimant ces nœuds et de répercuter leur signification sur le nœud global B.

La valeur d'instanciation du nœud B est vraie lorsqu'elle désigne qu'un besoin utilisateur est rencontré dans un document. Le domaine des nœuds qk est vrai pour designer que la représentation de la requête est satisfaite. L'apport le plus important de ce modèle a été de pouvoir combiner l'information provenant de représentations différentes de documents ainsi que de combiner différentes formulations de la requête.
Une simplification du réseau a été proposé **[Turtle, 91]** et exposée dans la figure Dans cette topologie, les nœuds documents sont des nœuds racine et il existe une relation entre les termes

d'indexation et les documents ou la requête pour designer que l'objet (document ou requête) contient le terme.

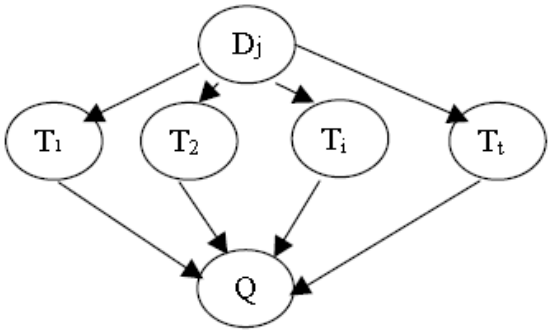


Figure III.7 : Architecture simplifiée.

Soit Dj un document dont le domaine est dom(Dj) = {dj,$\bar{d}$j}.

Un terme d'indexation est référencé par Ti dans la figure, et le domaine des termes, noté dom(Ti), est dom(Ti) = {ti,$\bar{t}$i}.

Le domaine de la requête est dom(Q) = {q,$\bar{q}$}.

III.8.2.2. Calcul de la pertinence :

Le calcul de la pertinence revient dans ce modèle à instancier chaque document de la collection et à calculer la croyance de satisfaire la requête étant donné le document instancié. Le réseau pris dans sa globalité représente les dépendances qui existent entre une requête et les documents de la collection.

Un seul document est instancié positivement (Dj = dj) à la fois. La propagation de l'information est déclenchée par cette instanciation.

La propagation dans ce modèle consiste à calculer pour chaque nœud la probabilité a posteriori étant données les probabilités a priori conditionnelles et marginales. La propagation tente de calculer la probabilité que l'information a été rencontrée étant donné un document instancié à Dj = dj. Ce processus est réitéré pour tous les documents de la collection. Une liste des documents ordonnés par ordre décroissant de pertinence est restituée.

La probabilité conditionnelle d'un nœud est fonction de toutes les configurations possibles de ses nœuds parents. Soit θ l'ensemble des configurations possibles des parents de Q, et θ_i^j une instance d'un nœud particulier Ti telle que dans la configuration de θj de θ. Par exemple, soit

la requête Q composée des deux termes T1 et T2, Q = {T1, T2} ; alors l'ensemble des configurations possibles des parents de la requête, tel que leur domaine est binaire, est θ = {{t1, t2}, {$\bar{t1}$, t2}, {t1,$\bar{t2}$ }, {$\bar{t1}$,$\bar{t2}$}}. L'instance θ_1^1 du terme T1 dans la première configuration de θ, θ1 = {t1, t2}, est θ_1^1 = t1.

La propagation dans le réseau dont la topologie est donnée dans la figure III.8 est :

$$P(Q \mid d_j) = \sum_{\forall\, \theta^k \in \theta} (P(Q \mid \theta^k) \cdot \prod_{T_i \in Q \wedge D_j} P(\theta_i^k \mid d_j) \cdot P(d_j))$$

La quantification totale de la pertinence revient à quantifier chaque membre de cette formule. Des probabilités a priori sont affectées aux documents de la collection, égales à P (Dj = dj) =$\frac{1}{N}$, mais elles sont supprimées du calcul de la propagation globale parce que ce terme est considéré comme un coefficient uniforme appliqué à tous les documents de la collection.
La section suivante décrit le traitement de la requête dans ce modèle.
Il s'agit particulièrement des diverses possibilités utilisées pour connecter ses termes.

III.8.2.3. Agrégation de la requête :

Turtle a proposé cinq formes canoniques pouvant répondre à tout type de recherche. La requête peut être agrégée par les opérateurs booléens (ET, OU, et NON). D'autre part, l'utilisation des réseaux permet d'agréger la requête par la somme probabiliste ou une de ses variations la somme pondérée. Pour évaluer les probabilités conditionnelles P (Q | θ) d'un nœud Q ayant n parents, {θ1, .., θn},
et, P (θ1 = t1) = p1, .., P (θn = tn) = pn les agrégations suivantes sont définies :
P_{Ou} (Q | θ) =1 − (1 − p1) −... − (1 − pn).
P_{Et} (Q | θ) =p1 ×... × pn.
P_{Non} (Q | θ1) =1 − p1.

$P_{Somme}(Q \mid \theta) = \frac{p1 + ..+ pn}{n}$.

$P_{SommePonderee}(Q \mid \theta) = \frac{(w1p1 + \ ...+ wnpn)wq}{w1 + ..+ wn}$.

Lorsque la négation d'un terme est spécifiée dans la requête, la quantification de sa présence dans le document est obtenue par 1 − p_i . Ici, la négation du terme n'est pas son absence de la représentation du document lorsqu'il est spécifié dans la requête. Les termes de la requête absents des représentations des documents ne sont pas considérés dans le calcul de la

pertinence d'un document en réponse à une requête. La somme probabiliste tient compte du nombre de parents instancies positivement dans la configuration des parents ($|\theta_j = t_j|$) et la somme pondérée mesure la configuration positive en fonction du poids de chaque parent instancie positivement, ainsi que du poids de la requête w_q. Le poids utilisé peut être le facteur de discrimination idf ou une de ses variantes ou un poids attribué par l'utilisateur.

Ces deux dernières techniques d'agrégation permettent un gain de temps lors des calculs de la pertinence puisque uniquement les termes présents dans une configuration (documents) sont considérés.

III.8.3. Le Réseaux Bayésiens de croyance :

Un Réseaux de Croyance **[Ribeiro-Neto et al, 96]** est basé sur la définition préalable d'un espace d'échantillonnage qui permet de séparer clairement les portions de documents des portions de requêtes et donc de calculer d'une manière efficace les degrés de croyance. Le modèle basé sur les réseaux de croyance, est supposé plus général par sa capacité à subsumer le réseau d'inférence ; Cette hypothèse est théoriquement prouvée par Baeza-Yates **[Baeza-Yates, 99]**, ainsi il permet de généraliser tous les modèles classiques de RI (booléens, probabilistes et vectoriels).

Les relations de dépendance définies dans ce modèle différent de celles de Turlte. Dans ce modèle le processus de recherche est déclenché par la réception de la requête.

La figure suivante montre le réseau de croyance da Baeza.

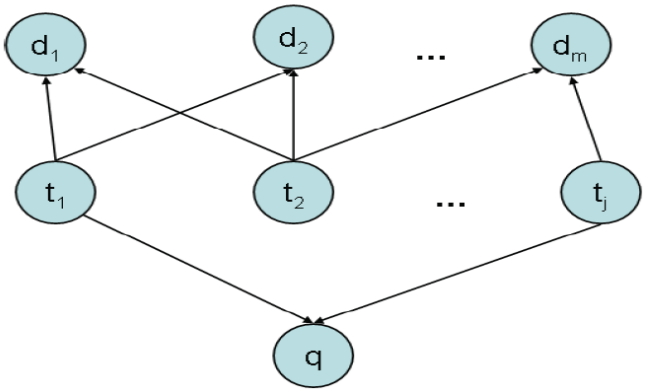


Figure III.8 : le réseau de croyance de Baeza [Thi, 09].

L'univers de discours est donné par l'ensemble des termes d'indexation utilisés dans le système, noté U, et U = {T1, .., TT} ou T est le nombre de termes manipulés dans le système

(pour représenter les documents ou la requête). θ est l'ensemble des configurations possibles sur U.

La définition des domaines d'un nœud terme donné est similaire à celle du modèle de Turtle. Un terme appartient ou non a un concept. Un concept peut être une requête ou un document. Les termes d'indexation pointent vers les documents et la requête qu'ils indexent.

Un document Dj, est une variable aléatoire de domaine binaire, dom(Dj) = {dj,$\bar{d}$j}. Un document Dj de la collection est instancié à Dj = dj pour indiquer que le document couvre complètement U.

Une variable aléatoire binaire est associée a une requête Q de domaine dom(Q) = {q,$\bar{d}$j}. Q=q signifie que la requête couvre complètement l'espace des termes.

La couverture de l'espace U par un concept (document ou requête) est la conformité du concept avec chaque élément de l'espace U.

III.8.3.1. Calcul de la pertinence :

l'instanciation de la requête permet de calculer la probabilité de pertinence d'un document étant donnée une requête, P(Dj | Q), donnée par la formule :

$$P(D_j \mid Q) = \frac{P(D_j \wedge Q)}{P(Q)}$$

La probabilité P(Q) est calculée pour tous les documents de la collection, et est considérée comme une constante. Ainsi, une approximation possible de la Probabilité d'un document étant donnée une requête peut être :

$$P(D_j \mid Q) \propto P(D_j \wedge Q)$$

D'après la topologie du réseau, l'instanciation des termes d'indexation (ici, cas de d-séparation) rend les nœuds documents et requête indépendants. Ainsi :

$$P(D_j \mid Q) \propto \sum_{\theta} P(D_j \mid \theta) \times P(Q \mid \theta) \times P(\theta)$$

θ représente l'ensemble des configurations possibles des termes de l'univers U. Ce modèle généralise les modèles classiques de la RI. Nous donnons ici le traitement opéré sur le calcul de la propagation pour reproduire le modèle vectoriel. $\vec{d}$= {t1, .., tT } et $\vec{q}$= {t'1, .., t'T } désignent respectivement le vecteur document et le vecteur requête. Pour chaque document, une similarité par cosinus **[Salton, 88]** entre un document et une requête est calculée. La

probabilité d'un document étant donnée une configuration d'un concept est approximée par le produit entre les poids des termes du document et de la requête. Ainsi :

$$sim(\vec{d_j}, \vec{q}) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|}$$

Ou | |désigne la cardinalité. Nous présentons dans ce qui suit, les mesures proposées pour calculer les probabilités conditionnelles d'un document donné (Dj) et de la requête (Q) étant donnes leurs parents respectifs, notés ParDj et ParQ.

III.8.3.2. Probabilité des documents P(Dj | ParDj) :

L'univers de discours U définit l'ensemble des termes d'indexation du système. La probabilité P(dj) donne le degré auquel le document Dj couvre complètement l'espace des termes U. Cette couverture est calculée en contrastant chaque élément de U avec le document Dj , à travers P(Dj | θ) et en additionnant les contributions de chacun. Cette somme est pondérée par la probabilité P(θ) avec laquelle θ apparaît dans U. Cette probabilité répondrait à la croyance associée à la proposition Est-il vrai que dj couvre complètement U ? et est donnée par :

$$P(d_j) = \sum_{\theta} P(d_j \mid \theta) P(\theta)$$

$$P(\theta) = (\frac{1}{2})^T$$

De plus, toujours dans le contexte vectoriel décrit dans la section ci-dessus, nous avons :

$$P(d_j \mid \theta^Q) = \frac{\sum_{\theta_i^Q=1}^{T} w_{ij} \times w_{iq}}{\sqrt{\sum_{\theta_i^Q=1}^{T} w_{ij}^2} \times \sqrt{\sum_{\theta_i^Q=1}^{T} w_{iq}^2}}$$

$$P(\overline{d_j} \mid \theta^Q) = 1 - P(d_j \mid \theta^Q)$$

où θ^Q_i est la configuration des termes telle que donnée dans la requête Q, et wij , wiq les poids du terme ti dans le document dj et la requête Q respectivement. Les poids wij sont des variantes de la pondération par tf * idf **.**

III.8.3.3. Probabilité de la requête P (Q | ParQ) :

La probabilité P(Q) donne le degré auquel la requête couvre complètement l'espace des termes U. Cette probabilité répondrait à la croyance associée à la proposition Est-il vrai que Q couvre complètement U ?

$$P(Q) = \sum_{\theta} P(Q \mid \theta) P(\theta)$$

$$P(\theta) = (\frac{1}{2})^{T}$$

Le calcul de cette équation nécessiterait 2^T calculs, où T est le nombre de termes manipulés par le système, mais en réalité uniquement les termes indexant la requête sont considérés.

La valeur 1 est attribuée aux arcs reliant les termes d'indexation à la requête lorsque tous les termes présents dans la requête sont instanciés positivement dans une configuration donnée des parents. Ainsi :

$$P(Q \mid \theta) = \{1 \text{ si } \square \text{ Ti}, \theta^{Q}_{i} = \theta i$$

$$= 0 \text{ sinon}\}$$

$$P(\bar{Q} \mid \theta) = 1 - P(Q \mid \theta)$$

Où θ^{Q}_{i} , θi l'instanciation du terme Ti dans la requête et dans θ respectivement.

Conclusion :

Dans ce chapitre, nous avons discuté les réseaux bayésiens et nous avons présenté les différents modèle de RI basées sur ces derniers.

Les modèles basés sur un RB combinent la théorie des probabilités avec la théorie des graphes. Ce modèle a non seulement la capacité de modéliser les variables et leurs dépendances, mais aussi la capacité de modéliser les évidences et leurs influences sur les variables via un processus d'inférence. Ce dernier permet de mettre à jour les probabilités des variables dans tout le réseau .Par ailleurs, plusieurs domaines sont intéressés par ce type de représentation.

Introduction :

L'objectif fondamental de l'accès personnalisé à l'information est de répondre au mieux aux besoins en information de l'utilisateur en intégrant le profil utilisateur défini par des facteurs dépendants directement de ce même, de sa requête ou de l'environnement de recherche dans le processus d'accès a l'information.

Comme nous l'avons cité dans le chapitre 2, plusieurs dimensions permettent la définition du profil utilisateur. Nous considérons dans ce travail, qu'un utilisateur est représenté par un centre d'intérêt exprimant les caractéristiques générales d'information qu'il souhaite obtenir. L'évolution du profil utilisateur désigne alors son adaptation à la variation des ses centres d'intérêt au cours d'une ou plusieurs session de recherche définies lors des différentes activités de recherche.

. Pour cela, ce présent chapitre est structuré comme suit : en premier lieu une problématique, quelques concepts clés de notre approche, d'autre part nous exposons la modélisation de notre approche.

IV.1. Problématique :

Dans le but de montrer l'impact de l'intégration du profil utilisateur ainsi que son évolution dans les systèmes de recherche d'information nous augmentons la collection de test TREC par les centres d'intérêt d'utilisateurs simulés.

TREC fournit des collections de test contenant un ensemble de requêtes, un ensemble de documents et des jugements de pertinence associés pour chaque requête, mais aucun élément caractérisant l'utilisateur. Notre objectif est d'exploiter ces associations de pertinence pour construire des centres d'intérêt pour des utilisateurs simulés. Dans ce sens, nous proposons d'utiliser des sous ensembles de requêtes d'un même domaine, pour construire un profil d'utilisateur a partir des centres d'intérêt de chaque requête, qui sont construits a partir des documents pertinents de chaque requête.

IV.2. Concepts clés de notre approche :

Les notions de profil utilisateur, d'activité de recherche et de session de recherche sont précisées ci-après.

- Un centre d'intérêt : est l'ensemble des besoins en information récurrents de l'utilisateur
- Profil utilisateur : correspond ici à l'ensemble de ses centres d'intérêt.
- Activité de recherche : c'est l'association d'une requête et le centre d'intérêt correspondant avec l'ensemble des documents jugés pertinents pour cette requête et les termes qui les indexent.
- Session de recherche : c'est l'association d'une ou plusieurs activités de recherche correspondant au même centre d'intérêt.

IV.3. Modélisation d'une activité de recherche :

Ce modèle est représenté par un graph acyclique orienté G= (V, E), il inclue des nœuds qui englobent : requête q, les nœuds termes d'indexation t_i et les nœuds de la librairie des centres d'intérêt C_k et les nœuds documents.

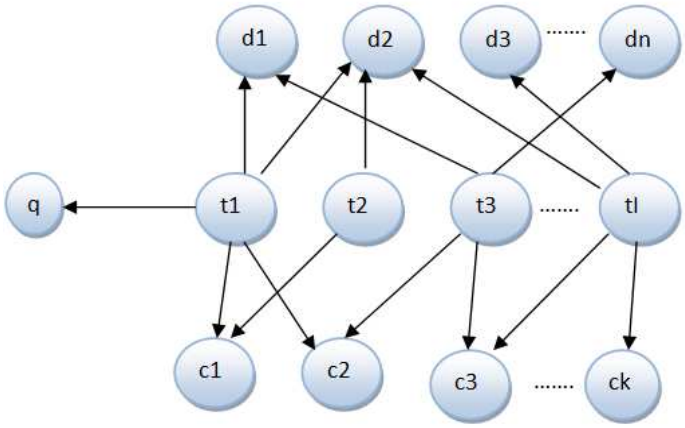


Figure IV.1 : Réseau bayésien d'une activité de recherche [Ach et al, 02].

Le nœud requête utilisateur q représente une variable aléatoire binaire dont le domaine est Dom(q)= $\{q,\bar{q}\}$, ou q désigne que la requête est instanciée et $\bar{q}$ désigne que la requête q n'est pas instanciée, on ne s'intéresse ici qu'a l'instanciation positive de q.

L'ensemble des termes indexant les documents de la collection T={t1,t2,t3,…,tI}, chaque nœud ti représente une variable aléatoire binaire dont le domaine est Dom(t_i)={t_i,$\bar{t}_i$},ou t_i désigne que le terme t_i est présent dans un document d_j ou la requête ou le centre d'intérêt ck, et le terme $\bar{t}_i$ désigne que le terme $\bar{t}_i$ n'est présent dans le document d_j ou la requête ou le centre d'intérêt ck .

L'ensemble des documents de la collection D= {d1, d2, d3,…,dn},chaque nœud document dj représente une variable aléatoire binaire dont le domaine est Dom(d_j)={d_j,$\bar{d}_j$},ou d_j désigne que le document d_j est instancié et $\bar{d}_j$ et désigne que le document $\bar{d}_j$ n'est pas instancié. Un seul document est instancié positivement à la fois.

L'ensemble des centre d'intérêt C={c1,c2,c3,…,ck},chaque nœud centre d'intérêt ck est une variable aléatoire binaire dont le domaine est Dom(ck)={ck,$\overline{ck}$},ou *ck et* $\overline{ck}$ désigne respectivement que le centre d'intérêt *ck* est instancié et que le centre d'intérêt *ck n'est* pas instancié. Un seul centre d'intérêt est instancié à la fois.

Les relations de dépendance entre nœuds sont traduites par des arcs orientés des nœuds termes vers les nœuds documents, les nœuds centre d'intérêt et le nœud requête pour designer qu'un terme appartient respectivement, a un document, a un centre et a une requête.

IV.4. Architecture de système personnalisé :

La personnalisation d'accès a l'information consiste a intégrer le profil utilisateur dans au moins l'une des phase de processus de recherche ; son intégration dans la phase de la reformulation de la requête consiste a augmenter la requête initiale par des termes issus du profil utilisateur , dans la phase d'appariement requête-document le profil est intégré dans le calcule de pertinence du document et pour la phase d'ordonnancement des résultats de recherche cela consiste a associé le profil utilisateur dans les résultats finaux de la recherche . La figure suivante montre un modèle de RI intégrant le profil utilisateur dans la phase d'appariement.

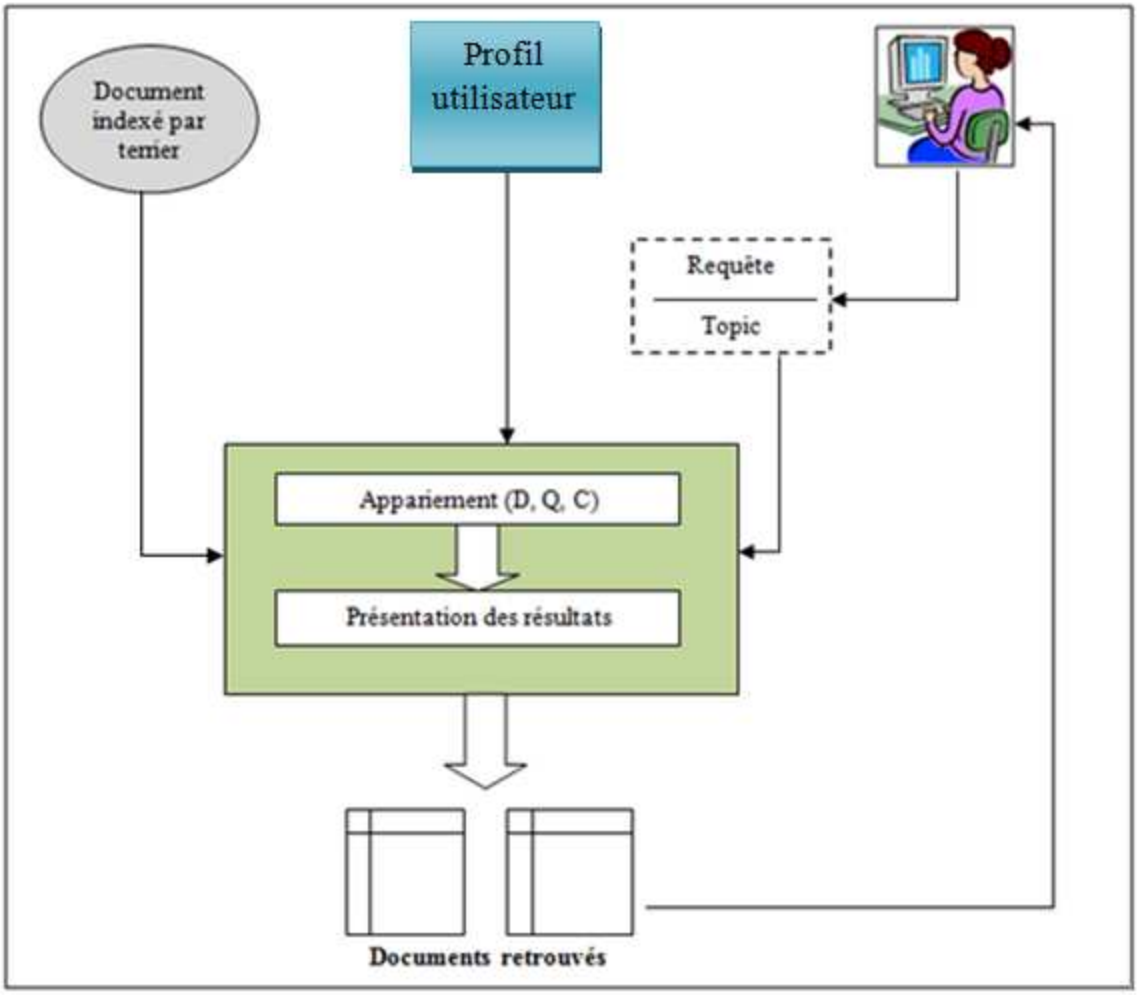


Figure IV.2. : Architecture de SRIP intégrant le profil utilisateur dans la phase d'appariement.

Appariement :

Avec l'integration du profil utilisateur dans la mesure d'appariement ,le calcul de la pertinenece revient à instancier chaque document de la collection à chacun des centre d'interet utilisateur. La propagation de l'information dans le modele est declanchée par la reception de la requete utilisateur et consiste a calculer la probabilité que la requete soit rencontrée étant données le document dj et le centre d'interet ck.

D'après la loi de bayes la probabilité P(Q/D,C) est exprimé par :

$$P(Q/D,C) = \frac{P(Q \wedge D \wedge C)}{P(C \wedge D)}$$

a partir de la topologie du modele proposé ,les conceptes requete, document et centre d'interet sont indépendant (D-séparés par les nœud termes).

La probabilité $P(q \wedge dj \wedge ck)$ est donc :

$$P(q \wedge dj \wedge ck) = \sum_{u \in U} P(q/u) \times P(dj/u) \times P(ck/u) \times P(u) \quad [1]$$

Nous simplifions l'espace des termes ,a la configuration u couverte par la requete q, dans ce cas :p(q/u)=1 si q=u (on prend seul la configaration u des termes qui composent la requete q) ce qui reduit l'equation 1 en :

P(q ∧ dj ∧ ck)= $P(dj/uq) \times P(ck/uq) \times uP(q)$. [2]

Puisque le modèle « reseau bayésien » généralise « le modèle vectoriel » ,les parametres de l'équation [2] sont estimés ainsi :

$$P(dj, q) = \frac{\sum_{i=1}^{l} W(ti, dj) \times W(ti, q)}{\sqrt{\sum_{i=1}^{l} W(ti, dj)^2} \times \sqrt{\sum_{i=1}^{l} W(ti, q)^2}}$$

$$P(ck, q) = \frac{\sum_{i=1}^{l} W(ti, ck) \times W(ti, q)}{\sqrt{\sum_{i=1}^{l} W(ti, ck)^2} \times \sqrt{\sum_{i=1}^{l} W(ti, q)^2}}$$

Ou :

$$\begin{cases} W(ti, q) = 1 & \text{si } ti \in q \\ 0 & \text{si non} \end{cases}$$

W(ti,dj)=tf*idf

W(ti,ck) est déjà calculé,das la phase de définition de centre d'interet.

L'appariement entre un document ,une requete et un centre d'interet est alors effectué en appliquant l'algorithme suivant :

Pour chaque document de la collection des documents faire

Pour chaque centre de la librairie faire

$$P(dj, q) = \frac{\sum_{i=1}^{l} W(ti,dj) \times W(ti,q)}{\sqrt{\sum_{i=1}^{l} W(ti,dj)^2} \times \sqrt{\sum_{i=1}^{l} W(ti,q)^2}}$$

$$P(ck, q) = \frac{\sum_{i=1}^{l} W(ti,ck) \times W(ti,q)}{\sqrt{\sum_{i=1}^{l} W(ti,ck)^2} \times \sqrt{\sum_{i=1}^{l} W(ti,q)^2}}$$

Fait

Fait

IV.5. Librairie de centre d'intérêt :

Pour la définition de librairie de centre d'intérêt on soumit différente requête au SRI terrier, pour chaque requête on définit son centre d'intérêt comme un vecteur de termes pondères en appliquant la formule BM25 **[Robertson 97]** suivante :

$$W(ti, Ck) = \log \frac{(r + 0.5)/(R - r + 0.5)}{(n - r + 0.5)/(N - n - R - r + 0.5)}$$

Avec :

N : le nombre total de documents de la collection ;

n : le nombre de documents de la collection contenant le terme ti ;

R : le nombre de documents pertinents associé à une requête utilisateur ;

r : le nombre de documents pertinents contenant le terme ti.

La figure suivante montre la création de la librairie de centre d'intérêt.

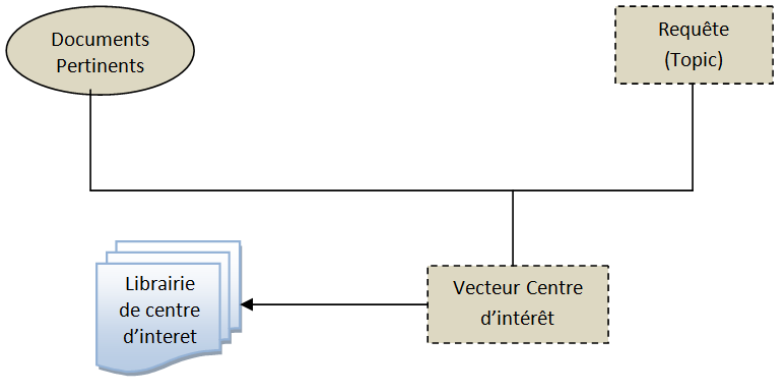


Figure IV.3. : Création de la librairie de centre d'intérêt

IV.6. Définition de profil utilisateur :

Cette section décrit le processus de définition des profils utilisateurs en se servant de la collection de test TREC. L'idée de base est d'exploiter les associations « Domaines - Requêtes - Documents » comme ressources informationnelles pour l'apprentissage des centres d'intérêt des utilisateurs. En effet, tel que nous l'avons mentionné, les requêtes de la collection sont annotées d'un domaine particulier pouvant être associé à des domaines d'intérêt. De plus,

TREC fournit, pour chaque requête de la collection de test, un ensemble de jugement de pertinence (liste des documents pertinents et non pertinents jugés par des utilisateurs). Partant de l'hypothèse qu'à chaque domaine de la collection correspond un profil utilisateur. Le processus de définition consiste à construire pour chaque domaine contenant *n* requêtes, un ensemble de centres d'intérêt. Comme suit :

- Pour chaque domaine *k* de la collection (avec *k* = (1..6)), nous sélectionnons parmi les *n* requêtes associées à ce domaine, un sous-ensemble de *m* requêtes pour constituer l'ensemble d'apprentissage.
- A partir de cet ensemble, on extrait automatiquement la liste des documents pertinents associés à chaque requête selon le système de recherche terrier-3.0.
- Partant de ces documents, un centre d'intérêt est construit comme un vecteur de termes pondérés en appliquant la formule *BM*25 définit précédemment (dans la section librairie de centre d'interet).

Le tableau suivant donne un exemple pour la construction des centres d'intérêt associés au domaine« Environment» :

Domaine	Environment
Requêtes associés	59, 77, 78,83

Le domaine contient 4 requêtes, on construite donc 4 centres d'intérêts.

Centre d'intérêt	Requête d'apprentissage
C1	59
C2	77
C3	78
C4	83

Le tableau suivant montre un extrait du vecteur de terme obtenu pour le premier centre :

Nom_terme	Poid_terme
olanta	1.0
deich	1.0
pantepec	0.9418013652150369
kopaonik	0.9418013652150369

Tableau 1: Exemple de construction de centre d'intérêt

IV.7.profil utilisateur a court terme :

Pour chaque domaine k on construit un profil utilisateur a court terme qui correspond a l'évolution du profil de l'utilisateur au cour d'une session de recherche, et sera définit par une combinaison linéaire des vecteurs centres d'intérêt associé a chaque requête. Afin de privilégier les termes récurrents dans la session. Selon ce principe, l'évolution du profil à court terme consiste à définir une fonction linéaire qui :

1. augmente le poids des concepts de la requête, récurrents dans la session via le profil à court terme,
2. atténue le poids des concepts non récurrents.

Soient ck_{q1} et $ck_{q2,}$ les centres d'intérêt correspondant aux requêtes q_1,q_2 respectivement , on calcule le nouveau poids d'un terme t_i dans le nouveaux centre comme suit :

$$W(t_i, \text{profil}) = \begin{cases} \propto \times w(ti, \text{ckq1}) + (1-\propto) \times w(ti, \text{ckq2}) \; si \; ti \in \text{ckq2} \\ \propto \times w(ti, ckq1) \qquad\qquad si \; non \end{cases}$$

$w(ti, ckq)$ est le poids du terme t_i dans le centre correspondant a la requête q.

Remarque : si on a d'autres requêtes on fait la combinaison linéaire entre le nouveau centre et le centre de la nouvelle requête.

Ainsi on construit une librairie de centres d'intérêt, chaque centre correspond au profil utilisateur a court terme (session ou domaine).

Le tableau suivant montre un extrait du vecteur de terme obtenu pour le profil du domaine «Environment» :

Nom_terme	Poid_terme
conver	0.44012671934520525
elec	0.1530349402318611
eduardu	0.014732608275434156
fuelwood	0.014732608275434156

Tableau 2: Exemple de construction de centre d'intérêt générique (profil) d'une session.

IV.8. Stratégie de test :

Dans le but de mesurer l'impact de l'intégration du profil utilisateur dans le processus d'accès à l'information nous optons pour le scénario suivant qui se base sur la méthode de la validation croisée :

Pour chaque domaine D_k avec k = (1..6)), on prend m requêtes (requêtes de test différentes des requêtes d'apprentissage)

Pour chaque requête Q_i on recherche les documents pertinents et cela en intégrant le profil utilisateur dans le processus de recherche.

La validation croisée ou la *k*-fold *cross validation* est une méthode d'évaluation qui consiste à diviser la collection de test en *k* sous ensembles de taille égale, d'utiliser *k*-1 sous ensembles pour la définition des profils utilisateur a court terme dans notre cas, et le k^{ime} sous ensemble pour le test. On réitère ensuite le processus *k* fois.

Le tableau suivant montre un exemple de test avec la validation croisé du domaine « environement » :

Profil utilisateur a court terme	Requêtes d'apprentissage	Requêtes de test
Profil1	77 78 83 59	59
Profil2	59 78 83	77
Profil3	59 77 83	78
Profil4	59 77 78	83

II. Outils de développements :

II.1. la collection de test AP88 :

On a utilisé dans nos travaux la collection de test de la campagne d'évaluation TREC AP88 (Associated Press 1988), elle est de taille moyenne (237 Mo) et contient un ensemble de document (79919) qui sont issus de différents articles de presse tels que Associate Press (AP) , un ensemble de requêtes qui se trouve dans le fichier topics et un ensemble de jugement de pertinence (liste des documents pertinents et non pertinents pour une requête donnée) ces jugement sont regroupés dans le fichier qrels.

la collection que nous avons utilisée se caractérise par les données statistiques résumées dans le tableau suivant :

Collection	AP88
Taille	237 Mo
Nombre de domaines	12
Nombre de documents	79919
Nombre de requêtes	1-100
Nombre de termes	155425

Tableau : Données statistiques de la collection de test.

II.1.1.les documents :

Le format d'un document da la collection AP88 est le suivant :

<DOC>

<DOCNO> AP880222-0001 </DOCNO>

<FILEID>AP-NR-02-22-88 2333EST</FILEID>

<FIRST>u i AM-Waldheim 02-22 0366</FIRST>

<SECOND>AM-Waldheim, 0378</SECOND>

<HEAD>Socialist Party Leadership Suggests Waldheim Must Resign</HEAD>

<DATELINE>VIENNA, Austria (AP) </DATELINE>

<TEXT>

The Socialist Party leadership on Monday strongly suggested President Kurt Waldheim should step down but stopped short of calling for his resignation…

</TEXT>

</DOC>

II.1.2.Les Requêtes (topics):

Les topics sont des textes à partir desquels les requêtes sont construites. Requêtes sont annotées d'un champ particulier noté « Domain» qui décrit un sujet d'actualité traité par la requête. Ci-dessous un exemple de requête extraite du domaine « Military» :

<top>

<head> Tipster Topic Description

<num> Number: 062

<dom> Domain: Military

<title> Topic: Military Coups D'etat
<desc> Description: Document will report a military coup d'etat,
either attempted or successful, in any country.
<smry> Summary: Document will report a military coup d'etat,
either attempted or successful, in any country.
</top>.

II.2.Terrier :

Terrier est une plate-forme dédiée à la recherche d'information. Elle implémente les différents modules intervenant dans le processus de RI classique et offre en plus un cadre pour l'évaluation des résultats de recherche pour différentes applications (Ounis et al. 2006). Le choix de cette plate-forme est dû aussi à sa capacité à traiter de grandes collections de documents telles que les collections TREC.

II.2.1.Architecture de terrier :

L'architecture de la plate-forme Terrier distingue les deux phases classiques : l'indexation et la recherche.

La figure suivante montre l'architecture générale de terrier

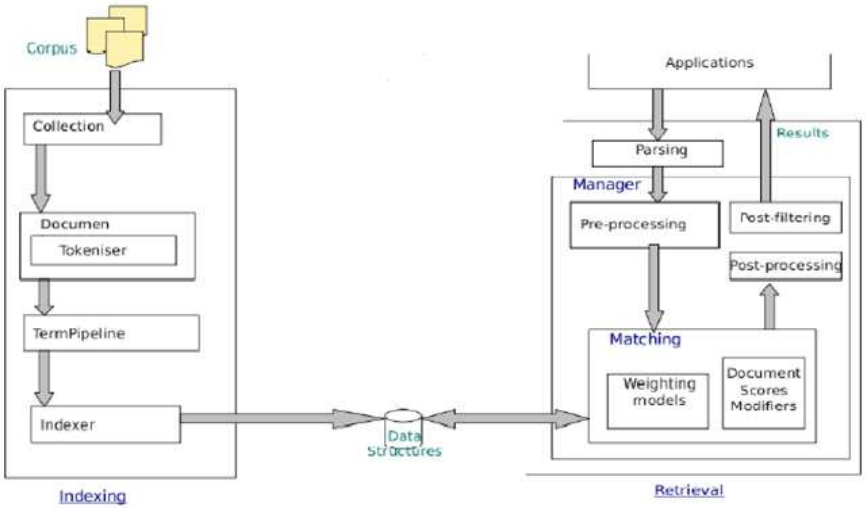


Figure IV.4 : Vue de l'architecture de terrier

II.2.1.1.API d'indexation :

Un corpus documentaire est fourni en entrée au module d'indexation. Les documents de la collection passent par un ensemble de pré-traitements tels que la tokenisation. Les tokens sont ensuite injectés dans une chaîne de traitement TermPipLines, à savoir le StopWords PipeLine pour l'élimination des mots vides de sens, ou encore les Stemming pipeline et qui dépendent de la langue en question. La phase d'indexation conduit à la construction de l'index (Data structures) suivantes :

- Lexicon : contient les informations sur chaque terme de la collection (Terme, id terme, le nombre de documents qui contiennent le terme(NT), la fréquence de terme dans la collection(TF), offset dans le fichier inverse).
- Inverted index : fichier inverse (id terme, id document, fréquence terme dans le document,#filds).
- Direct index : index (id terme, id document, fréquence terme dans le document,#filds).
- Document index (id terme, fréquence terme,#filds).

La figure suivante donne une vue d'ensemble d'interaction des composants principaux impliqués dans le processus d'indexation.

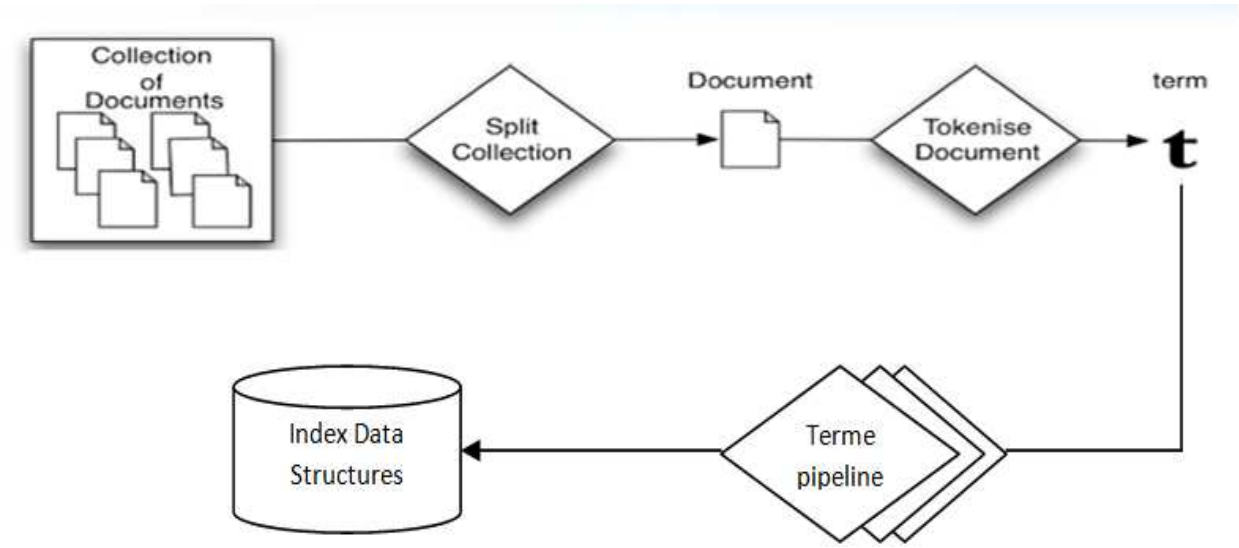


Figure IV.5 : Processus d'indexation dans terrier.

II.2.2.API de recherche :

La phase de recherche comprend le Manager, un module qui interagit avec l'application, réalise la mise en correspondance à travers les calculs des pondérations (selon le schéma de

pondération (Weighting Model) choisi : PL2, BM25, etc.) ainsi que les scores des documents. Le résultat renvoyé à l'utilisateur, est la liste des documents jugés pertinents et leurs scores respectifs.la figure suivante donne une vue d'ensemble d'interaction des composants de terrier dans la phase de recherche.

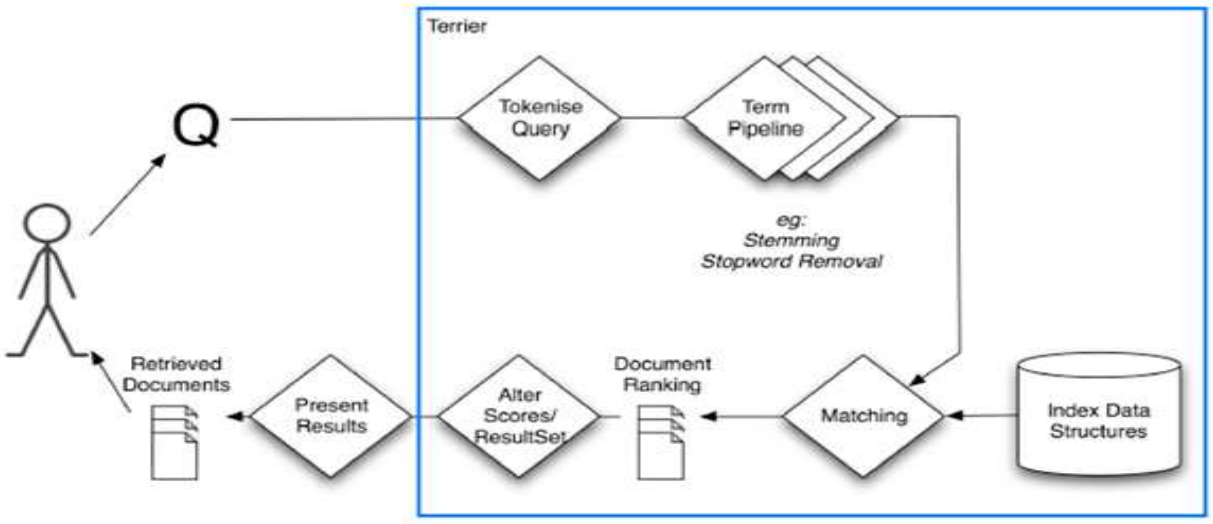


Figure IV.6 : le processus de recherche dans terrier.

L'API de recherche contient les classes suivantes :

Query : classe abstraite qui représente la requête.
Terrier supporte trois types de requêtes :
SingleTermQuery: requête qui contient un seul terme.
MultiTermQuery : représente la requête avec plus d'un terme.
FieldQuery : terme qualifié par un champ.

Manager : chargé de la gestion de la recherche, il contient les étapes suivantes :
Pre-processing : appliquer l'élimination de mots vides et troncatures.
Matching : déterminer les documents qui rependent a la requête en initialisant :

- WeightingModels : assigner un score pour chaque terme de la requête dans le document (pondération), plusieurs modèles de pondération sont implémentés : TF_IDF, BM25,…
- DocumentScoreModifiers : permet de modifier le score d'un document en fonction du langage de la requête.

Post-filtring : filtrer les documents pertinents selon un critère bien définit.

Post-pricessing : reclasser les documents pertinents s'il ya une expansion de la recherche.

Set-results : contient la liste des documents tournés classées selon leur degré de pertinence.

II.3. le langage java :

Java est une technologie développée par Sun Microsystems dans les années 1990. Elle correspond à plusieurs produits et spécifications de logiciels qui, ensemble, constituent un système pour développer et déployer des applications autonomes et portables qui s'exécutent indépendamment du système d'exploitation utilisé.

Depuis des années, Sun Microsystems appelle Java la « technologie Java » dans son ensemble. En pratique et par abus de langage, beaucoup de programmeurs utilisent le mot « Java » pour désigner le langage de programmation, tandis que la plate-forme d'exécution est appelée « JRE » (Java Runtime Environment, environnement d'exécution Java), le système de compilation : « JDK» (Java Development Kit, kit de développement Java) plutôt que « compilateur Java » et un EDI ou IDE (Integrated Development Environment, environnement de développement intégré) qui utilise le JDK.

Notre choix s'est porté sur ce langage pour les raisons suivantes :

- Java est un langage multi plateforme qui permet aux concepteurs, d'écrire un code capable de fonctionner dans tous les environnements. Pour cela, il suffit que l'environnement possède un JVM (java virtuel machine) ;
- Java est un langage orienté objet, simple qui réduit le risque d'erreurs d'incohérence ;
- Java est doté d'une riche bibliothèque de classes, comprenant la gestion des interfaces graphiques et la gestion des exceptions ;
- Le JDK, fournit gratuitement par Sun, regroupe l'ensemble des éléments permettant le développement, la mise au point et l'exécution des programmes java ;
- Java est caractérisé aussi par sa sécurité, un programme java planté ne menace pas le système d'exploitation.il ne peut pas avoir d'accès directe a la mémoire, ainsi que sa simplicité de mise en œuvre.

Dans notre travail nous avons utilisé éclipse INDIGO, l'image suivante présente l'interface de travail sous Eclipse :

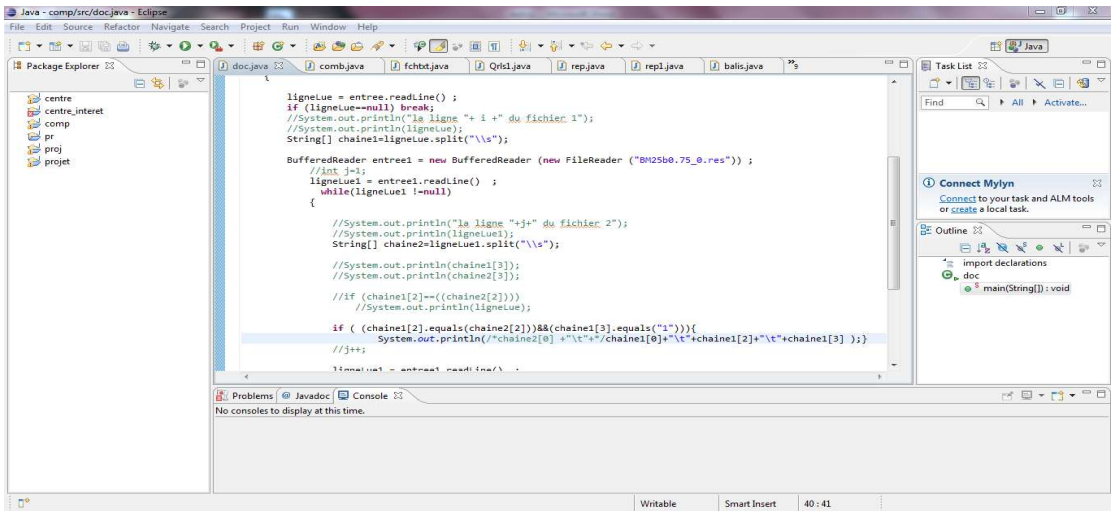


Figure: capture d'écran présentant l'interface de développement d'Eclipse.

III. Teste et évaluation:

La collection contient au total 12 domaines annotant les requêtes. Nous avons sélectionné aléatoirement six domaines d'intérêts : *Environment, Military, Law & Government, International Relations, US Economics,International politics*. TREC attribut une numérotation unique à l'ensemble des requêtes pour toutes les collections qu'il fournit. Pour notre part, nous avons utilisé plus particulièrement les requêtes de la collection numérotées de 51 à 100 (*q*51_*q*100), par domaine. Chaque domaine contient le nombre de requêtes suivant : 4,4, 4, 5,3, 5, respectivement. Le tableau suivant donne les numéros de requêtes associées à chacun de ces domaines.

Domaines	**Requêtes associées**	**Numéro du domaine**
Environment	59, 77, 78,83	1
Military	62, 71, 91,92	2
Law & Government	70, 76, 85,87	3
International Relations	64, 67, 69, 79,100	4
US Economics	57, 72,84	5
International politics	61, 74, 80, 93,99	6

Tableau 1: Numéros de requête associée aux domaines sélectionnés.

Pour la comparaison de nos résultats avec ceux de terrier et ceux de l'approche de Gauche on a pris les quatre premiers domaines.

III.1. Comparaison par rapport à une approche classique (sans personnalisation) :

Apres avoir construit les profils utilisateur a court terme pour chaque domaine, effectuer les testes en utilisant la validation croisé et en se basant sur la métrique d'évaluation (*PX* : la précision pour les *X* premiers documents restitués) telle que la *PX* est la proportion de documents pertinents dans les *X* premiers documents retrouvés. Elle permet d'exprimer la satisfaction de l'utilisateur vis-à-vis des *X* premiers résultats pertinents. Elle constitue ainsi une mesure importante pour l'évaluation des modèles d'accès personnalisé à l'information.

Nous avons calculé la moyenne des valeurs obtenues par toutes les requêtes associées, Nous comparons ensuite les résultats obtenus en utilisant notre modèle à ceux obtenus en utilisant un modèle de référence (terrier). Le tableau suivant montre ces résultats :

	Terrier			notre approche		
Requête	P5	P10	P15	P5	P10	P15
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
77	0.4000	0.4000	0.4000	0.8000	0.8000	0.6000
78	0.6000	0.7000	0.6667	1.0000	1.0000	1.0000
83	0.2000	0.1000	0.0667	0.2000	0.1000	0.0667
62	0.4000	0.3000	0.2000	0.4000	0.3000	0.2000
71	1.0000	0.8000	0.6667	1.0000	0.9000	0.8667
91	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
92	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
70	0.8000	0.5000	0.3333	0.8000	0.6000	0.6667
76	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
85	0.0000	0.2000	0.2000	0.0000	0.2000	0.2000
87	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	0.0000	0.1000	0.0667	0.6000	0.6000	0.6667
67	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
69	0.0000	0.2000	0.2667	0.2000	0.2000	0.2667
79	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	0.2000	0.1000	0.0667	0.2000	0.1000	0.0667

Moyenne	0,211764	0,2	0,172558	0,30588	0,282352	0,2706

Tableau 1: Résultats comparatifs requête par requête entre notre approche et l'approche classique

L'histogramme ci-dessous montre la comparaison des moyennes de PX (X=5, 10,15) de l'ensemble des requêtes utilisées.

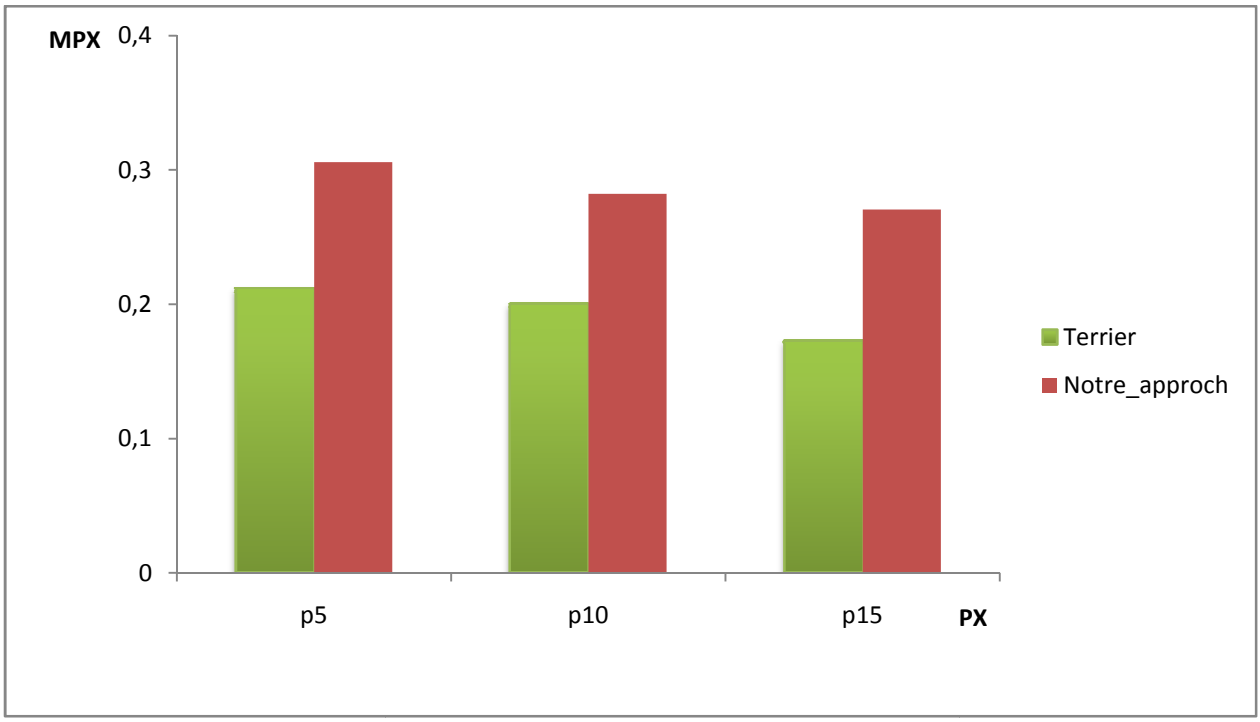


Figure IV.7 : Résultats comparatifs entre notre approche et l'approche classique

A partir du tableau (1) et du graphe (IV.7) nous constatons que :

Notre approche donne de meilleurs résultats par rapport a ceux obtenu avec terrier, et cela sur

Les points de précision P5, P10 et P15 pour l'ensemble des requêtes testées Les moyennes des précisions sont respectivement augmentées de 0.211764 à 0.3058, de 0,2 à 0.2823 et de 0.1725 à 0.2706.Ceci montre que l'intégration du profil utilisateur a un impact très positif sur la recherche d'information.

III.2. Comparaison par rapport à l'approche de Gauch (avec personnalisation) :

Dans ce qui suit nous comparons les résultats de notre approche à celle de Gauch **[Gauch, 03],** basée sur la projection directe des documents jugés pertinents par l'utilisateur sur l'ontologie de l'ODP dans une session de recherche. Cette méthode est résumée comme suit :

1) pour chaque document jugé pertinent par l'utilisateur, appliquer une mesure de similarité vectorielle avec les vecteurs représentatifs des catégories sémantiques de l'ODP,

2) classifier le document dans les premières cinq catégories ayant les scores de similarité les plus élevées avec son vecteur représentatif,

3) calculer pour chaque catégorie un poids par agrégation des scores de similarité vectorielle des documents classifiés sous cette catégorie,

4) le centre d'intérêt est ainsi représenté par les cinq premières catégories ayant les poids les plus élevés, utilisé ensuite dans le réordonnancement des résultats de recherche.

	Notre approche			Modèle de Gauch		
Requête	P5	P10	P15	P5	P10	P15
59	0.0000	0.0000	0.0000	0.2000	0.1000	0.1333
77	0.8000	0.8000	0.6000	0.8000	0.7000	0.6000
78	1.0000	1.0000	1.0000	1.0000	1.0000	0.9333
83	0.2000	0.1000	0.0667	0.0000	0.1000	0.0667
62	0.4000	0.3000	0.2000	0.2000	0.3000	0.2000
71	1.0000	0.9000	0.8667	0.8000	0.9000	0.8667
91	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
92	0.0000	0.0000	0.0000	0.0000	0.2000	0.2000
70	0.8000	0.6000	0.6667	0.6000	0.6000	0.6669
76	0.0000	0.0000	0.0000	0.6000	0.7000	0.6669
85	0.0000	0.2000	0.2000	0.8000	0.7000	0.8000
87	0.0000	0.0000	0.0000	0.2000	0.2000	0.2000
64	0.6000	0.6000	0.6667	0.4000	0.5000	0.6000
67	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
69	0.2000	0.2000	0.2667	0.2000	0.2000	0.2000
79	0.0000	0.0000	0.0000	0.2000	0.1000	0.0667
100	0.2000	0.1000	0.0667	0.6000	0.4000	0.2667
Moyenne	0,4333	0,4	0,4333	0.45	0,45	0,3833

Tableau 2 : Résultats comparatifs requête par requête entre notre approche et l'approche de Gauch.

L'histogramme ci-dessous montre la comparaison des moyennes de PX (X=5, 10,15) de l'ensemble des requêtes utilisées.

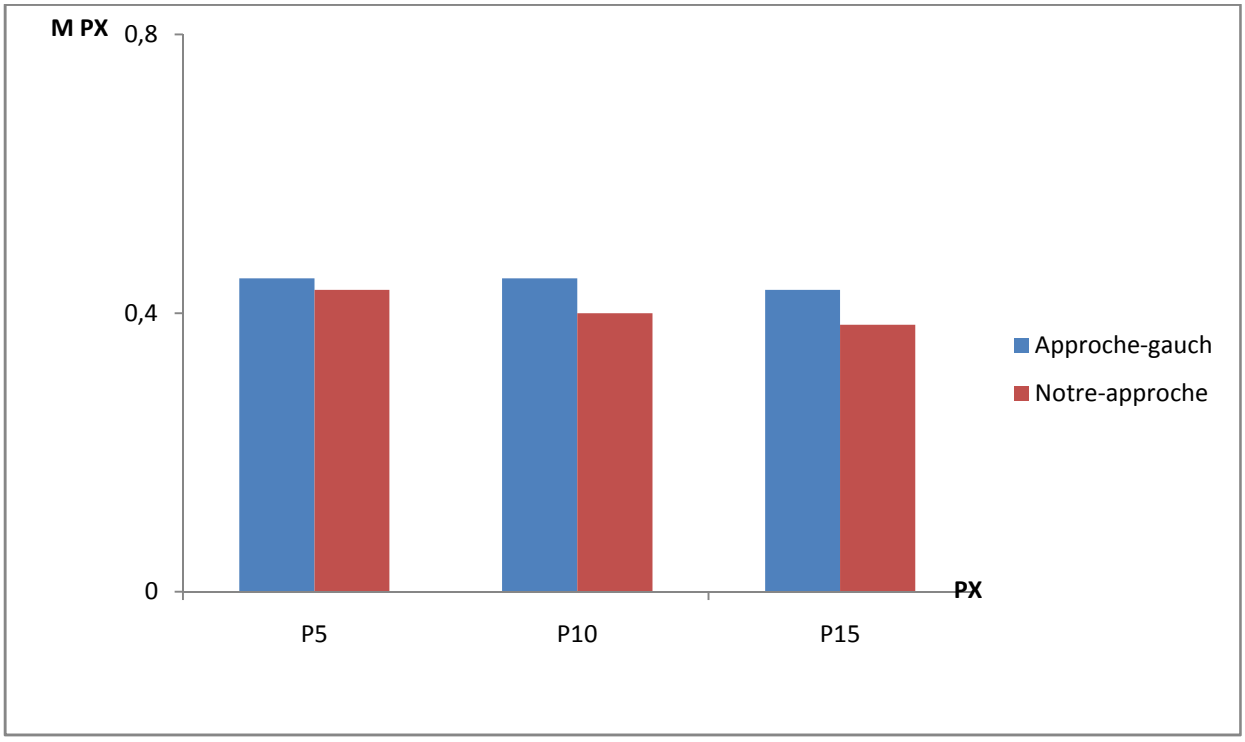


Figure IV.8: Résultats comparatifs entre notre approche et l'approche de Gauch.

A partir du tableau (2) et du graphe (IV.8) nous constatons que :

Notre approche donne des résultats légèrement inferieurs à ceux obtenu avec l'approche de Gauch, et cela sur les points de précision P5, P10 et P15 pour l'ensemble des requêtes testées. A titre d'exemple La moyenne de précisions à 5(P5) est de l'ordre 0.45 avec l'approche de Gauch et de 0.43 avec notre approche.

Ceci montre que la construction d'un bon profil utilisateur peut se faire sans l'utilisation d'une ressource externe comme l'ontologie telle que réaliser dans l'approche de Gauch.

Conclusion :

Dans ce chapitre nous avons abordé l'intégration du profil utilisateur dans le processus de recherche d'information, nous avons donc, présenté la démarche suivie pour la définition des profils utilisateurs et aussi l'intégration de ces derniers dans le processus de recherche.

Les résultats obtenus ont montré que l'approche proposée et très prometteuse.

Conclusion générale

Dans ce travail nous avons discuté le problème lié à la personnalisation de l'information qui a pour objectif d'intégrer l'utilisateur dans le processus de recherche d'information, dans le but d'améliorer les résultats de recherche. Nos contributions présentées dans ce mémoire ont porté sur :

- la définition du profil utilisateur à partir de la collection de test TREC et du système de recherche d'information classique Terrier, en construisant ainsi ces profils à partir des documents pertinent retourné par terrier.
- l'intégration de ce profil dans la phase d'appariement
- l'évaluation et la comparaison de notre approche par rapport a l'approche classique et l'approche de Gauch en se basant sur la métrique d'évaluation PX (précision a X).

Les résultats obtenus après l'évaluation de notre approche ont montré l'impact positif de l'intégration de l'utilisateur dans le processus de recherche d'information et que pour la construction d'un profil utilisateur on n'a pas besoin d'utiliser des ressources externes comme les ontologies. On peut juste se contenter des documents jugés pertinents par l'utilisateur, retournés lors de chaque activité de rechercher.

Cependant ce travail peut être amélioré on divers points :
Effectuer des tests sur des collections plus volumineuses.
Construire des profils à partir des comportements utilisateurs réels.

Introduction :

Terrier (TERabyte RetrIEveR) est un projet développé par le département de science de calcul de l'université de Glasgow en 2000, dans le but de fournir une plateforme flexible pour le développement rapide des applications de recherche d'information. Terrier est un produit open source écrit en Java.il a été avec succès employé pour la recherche ad-hoc et la recherche web.

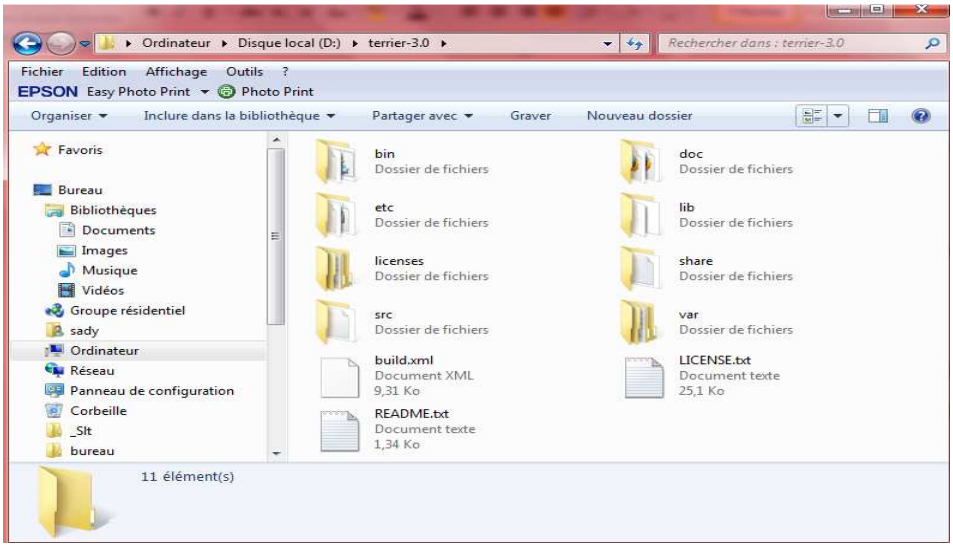
Terrier offre plusieurs modèles de pondération de documents et expansion de requêtes basées sur le Framework DFR (Divergence From Randomness).comme tous les moteurs de recherche terrier possède les principales facettes suivantes :

Indexation : permet l'extraction des termes des différents documents du corpus (basic index unit).

Recherche : permet de générer les résultats aux requêtes des utilisateurs.

1. lancement de terrier sous éclipse :

1. Extraire le contenue du fichier **terrier-3.0.zip** téléchargé a partir du site de la plateforme terrier : http://www.terrier.org .la figure suivante montre le contenu du fichier terrier après la décompression.



- **bin /** pour exécuter terrier
- **doc**/ documentation relative a terrier
- **etc**/ fichier de configuration de terrier
- **lib**/ classes compilés de terrier et les differentes librairies externes utilisées par terrier

- **share** / liste des mots vides (stopword-list.txt) et des exemples de document a tester sur terrier
- **src**/ code source java de terrier
- **var**/ contient les répertoires index et result
    . **Index**/ structures de données après indexation
    . **Result**/ résultat de la recherche et l'évaluation

2. Crier un nouveau projet dans éclipse dans le workspace

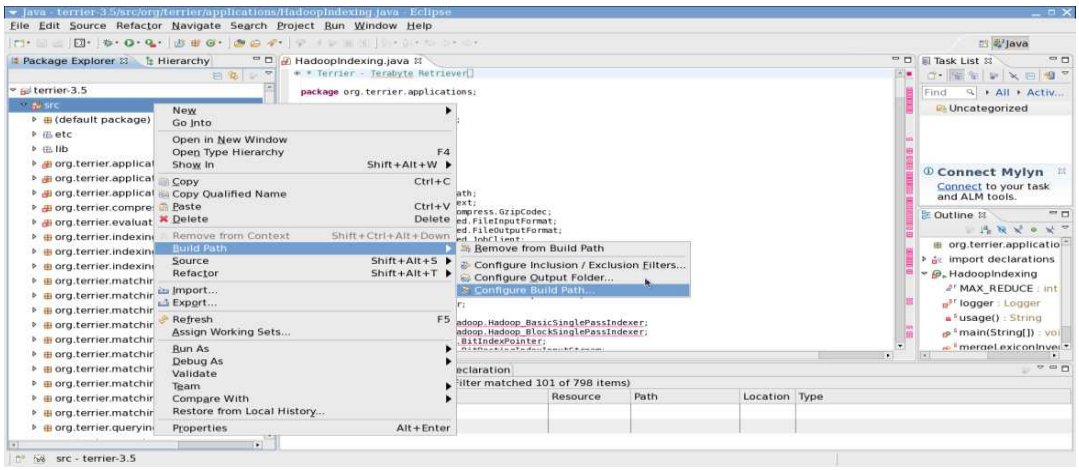
3. Copier les répertoires src ,etc ,lib ,var ,share du fichier terrier 3.0 dans le nouveau projet

4. Importer les librairies : clic droit sur le nouveau projet dans le Package Explorer

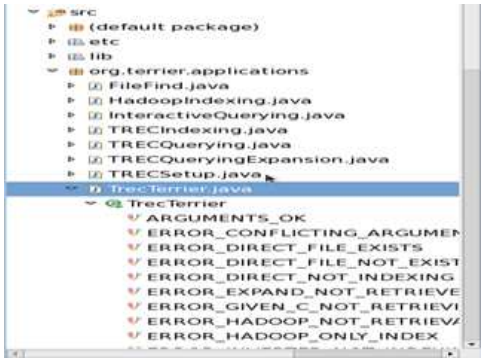
. Select Build path → Configure build path

. Select the tab 'Libraries' → Add External Jar

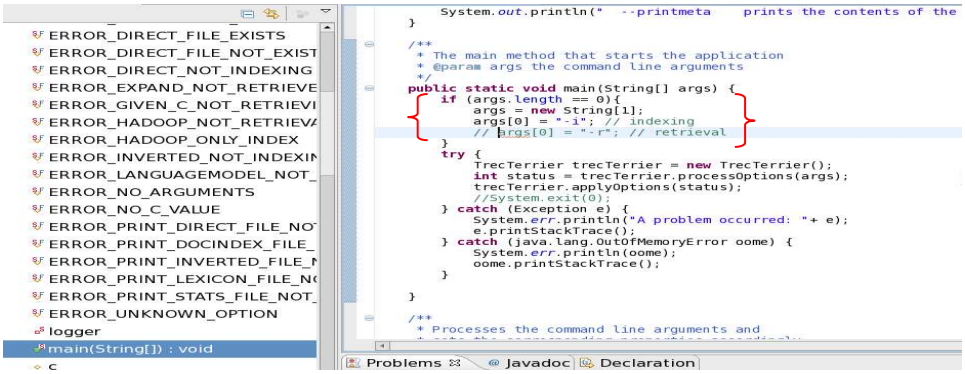
. Select all the .jar files in the 'lib' folder



5. Localiser la classe principale TrecTerrier.java dans le package org.terrier.applications



6. Ouvrir la classe TrecTerrier.java et la modifier comme suit :



7. Configurer terrier :

_ modifier dans le fichier terrier.properties les propriétés suivante :

terrier.home=…/nouveau projet/

terrier.index.path=…/ nouveau projet /var/index

terrier.results=…/ nouveau projet /var/results

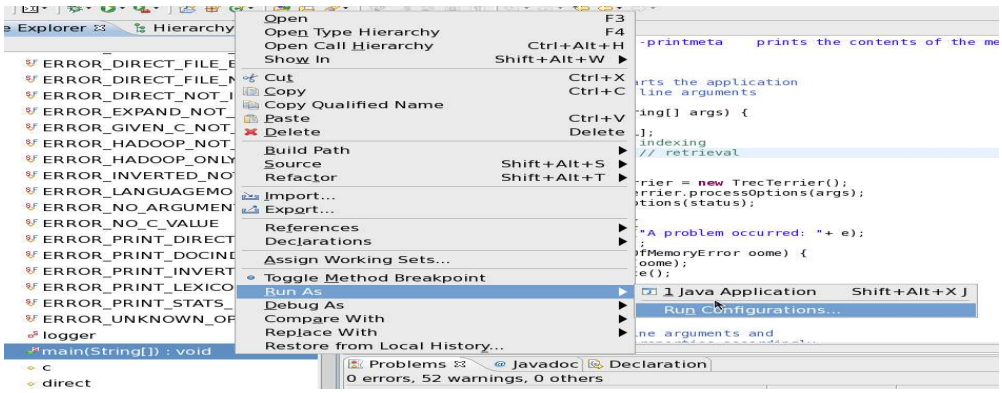
etc…

_ spécifier les inputs de terrier :

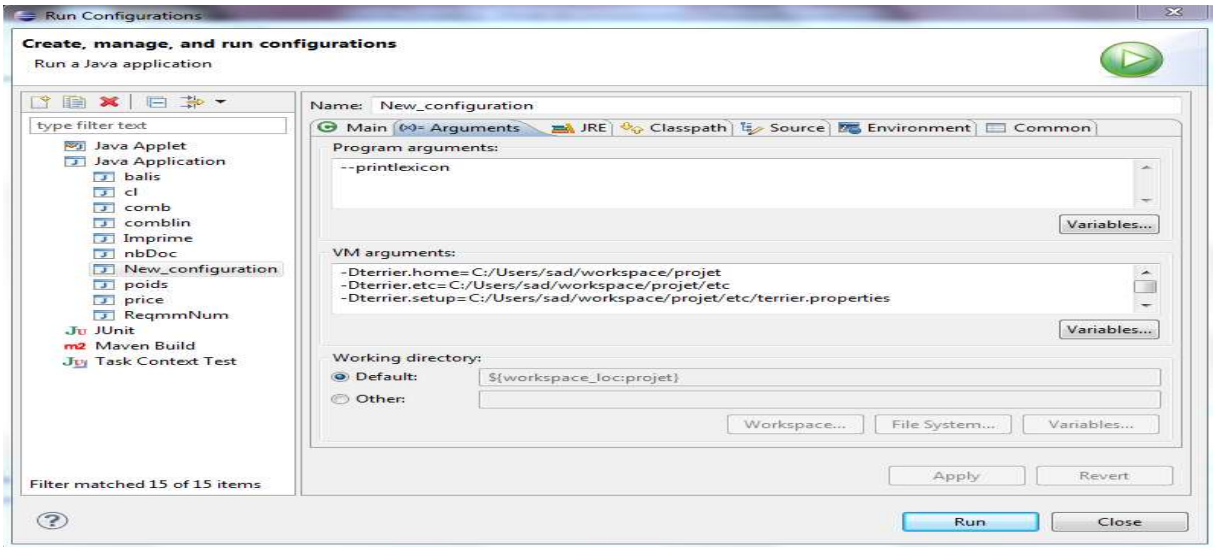
. Document collection
. Topics (queries)
. Qrel files
. Etc…

8. Lancer TrecTerrier.java :

Run as ➔ Run configuration ➔ Java application



9. spécifier les VM arguments :



1. Rappel de probabilités :

1.1.Probabilité conditionnelle :

p désigne une probabilité sur un univers fini Ω. A et B étant deux événements de Ω, B étant de probabilité no nulle. On appelle **probabilité conditionnelle** de l'événement A sachant que B est réalisé le réel noté

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

Exemple :

Disposant d'un dé parfait, on se demande quelle est la probabilité d'observation d'un 1 sachant que le résultat est impair. Tout à fait intuitivement, on dira, à juste titre, que cette probabilité est 1/3 : 1 événement élémentaire favorable (le 1) sur 3 événements conduisant à l'événement « impair ». De façon plus formelle, considérons deux événements A et B, sous ensembles de l'ensemble des événements élémentaires d'une épreuve E. Nous cherchons à définir la probabilité de l'événement A sachant que B est réalisé. Il faut se limiter au sous-ensemble B et chercher la « masse » relative de la partie de A contenue dans B par rapport à la masse de B. Ceci conduit naturellement à la définition de la probabilité conditionnelle

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

1.2.Théorème de probabilités totales :

Soient A1, A2, …, An une **partition** de l'univers Ω constituée d'événements de probabilités non nulles et B un événement quelconque contenu dans Ω.

Alors: P (B) = p (B ∩ A1) + p (B ∩ A2) + … + p (B ∩ An) = $\sum_i^n P(B \cap A_i)$

Donc: **P (B) = $\sum_i^n P(B/A_i)$ P (A_i).**

1.3. Indépendance :

A et B sont 2 événements de probabilité non nulle.

A et B sont **indépendants** lorsque la réalisation (ou non) de l'un n'a pas d'influence sur la probabilité de réalisation de l'autre.

A et B sont **indépendants** si et seulement si :

P (A, B) =P (A) × P (B)

P (A/B) = P (A)

P (B/A) = P (B).

Exemple:

On lance deux dés et on désigne par *A* l'événement "le premier dé amène un nombre pair", par *B* l'événement "le deuxième dé amène un nombre impair" et par *C* l'événement "les deux dés amènent un nombre pair".
On a: $P(A) = 1/2$; $P(B) = 1/2$; $P(C) = 1/4$; $P(A \cap B) = 1/4$; $P(A \cap C) = 1/2$; $P(B \cap C)=0$.
On conclut : *A* et *B* sont indépendants ; *A* et *C* sont dépendants ; *B* et *C* sont dépendants.

1.4.Indépendance conditionnelle :

Etant données trois variables aléatoires A, B et C, A et B sont dites indépendantes conditionnellement à C si et seulement si

P (A, B/C) =P (A/C) P (B/C).

Il en découle avec le même raisonnement que pour l'indépendance classique que lorsque P (B/C) est non nul, une définition équivalente est :

P (A/B, C) = P (A/C).

Exemple :

Pour illustration, prenons l'exemple du foyer épidémique de la grippe porcine de 2009, dite A (H1N1), situé au Mexique et d'un patient présentant les symptômes de la grippe. Considérons les trois variables binaires suivantes :

– Mexique (Oui / Non) : le patient a t'il fait un voyage au Mexique ?

– A(H1N1) (Oui / Non) : le patient est t'il porteur du virus de la grippe porcine ?

– Test (Oui / Non) : le test de présence du virus est-il positif ?

Les variables Mexique et Test ne sont pas indépendante. Avoir fait un voyage au Mexique augmente la probabilité d'avoir été en contact avec le virus et donc que le test s'avère positif.

Maintenant supposons que le patient soit effectivement porteur de la grippe porcine. La probabilité que le test soit positif n'est plus liée à la question du voyage au Mexique. Elle ne dépend que des caractéristiques du test.

P (Test/A (H1N1)) = P (Test/A (H1N1), Mexique).

1.5. La règle de chainage :

Soit V = (A1, A2, .., An) un ensemble de variables. La probabilité jointe P (A1, A2, .., An) permet de calculer P(Ai) et P (Ai/c), telle que c est une information donnée. Le nombre et le temps de calculs effectués pour obtenir la probabilité P(V) augmentent d'une manière exponentielle par rapport au nombre de variables contenues dans V. La règle de chainage permet de calculer P(V) d'une manière plus rapide lorsqu'il y a des dépendances entre les variables. Ainsi, la probabilité jointe est donnée par :

$$P(V) = \prod_{i=1}^{n} P(Ai \mid \text{PARENTS } Ai)$$

Ou PARENTS_{Ai} constitue l'ensemble des parents de Ai.

2. Théorème de bayes :

Considérons un événement B dont la réalisation dépend de l'intervention de l'une des causes : $A_1, ..., A_i, ..., A_n$.

Soit $p(B/A_i)$ la probabilité conditionnelle de B, si c'est la cause A_i qui intervient.

Et soit $p(A_i)$ la probabilité d'intervention de A_i, appelée probabilité a priori de A_i.

Le théorème de Bayes, appelé aussi théorème de la probabilité des causes, calcule la probabilité

$P(A_i / B)$ qui est la probabilité pour que ce soit la cause A_i qui ait entraîné la réalisation de B. Cette dernière probabilité est appelée la probabilité a posteriori de A_i.

Ce théorème qui date de plus de 2 siècles et qui était tombé en désuétude, a repris de l'intérêt pendant les dernières décennies. Il est utilisé dans de nombreux domaines.

La définition des probabilités conditionnelles permet d'écrire que :

$P(Ai \cap B) = p(Ai) \times p(B/Ai) = p(B) \times p(Ai/B)$

Et le théorème des probabilités totales que :

$P(B) = \sum_i^n P(B/A_i) \times P(A_i)$.

D'où le théorème :

$$\mathbf{P(Ai/B) = \frac{p(Ai) \times p(B/Ai)}{\sum_i^n P(B/Ai) \times P(Ai)}}$$

Exemple :

Il ya 4% d'absentéisme chez les employés travaillant le jour ,8% chez ceux qui travail le soir et 22% chez ceux qui travail la nuit, sachant qu'il ya 80% des employés qui

travaillent le jour ,10% qui travaillent le soir et 10% qui travaillent la nuit, déterminer la probabilité qu'un employé donné travaillant le jour sachant qu'il était absent du travail.

Solution :

B1 :l'employé travail le jour.

B2 : l'employé travail le soir.

B3 : l'employé travail la nuit.

A : l'employé est absent.

Nous savons que

P(A/B1)=4%,P(A/B2)=8% ;P(A/B3)=22%,P(B1)=80%,P(B2)=10%,P(B3)=10%

$$\mathbf{P(B1/A)} = \frac{\mathbf{p(B1)} \times \mathbf{p\,(A\,/B1)}}{\mathbf{P(A/B1)} \times \mathbf{P(B1)} + \mathbf{P(A/B2)} \times \mathbf{P(B2)} + \mathbf{P(A/B3)} \times \mathbf{P(B3)}}$$

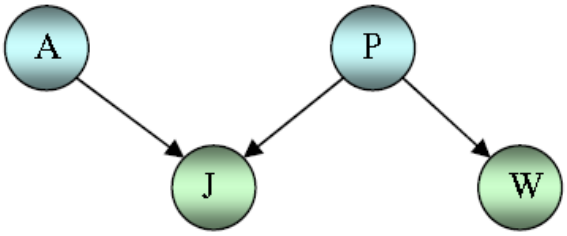
$$= \frac{\mathbf{0.004} \times \mathbf{0.80}}{\mathbf{0.004} \times \mathbf{0.80} + \mathbf{0.80} \times \mathbf{0.10} + \mathbf{0.22} \times \mathbf{0.10}} = \mathbf{0.51613}$$

3. Circulation de l'information:

Nous allons à présent étudier de plus près comment l'information circule au sein d'un graphe causal. 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. Il 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'œil 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 il décide de partir au travail sans vérifier son arroseur automatique.

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



Dans ce graphe le nœud A représente l'événement que M. Holmes a oublié de débrancher son arroseur automatique, le nœud P représente l'événement qu'il a plu cette nuit, le nœud J représente l'événement que l'herbe de son jardin est humide, le nœud W représente l'événement que l'herbe du jardin de M. Watson est humide.

L'information peut circuler des causes vers les effets, d'ailleurs on remarque M. Holmes a décider d'aller au bureau sans vérifier son arroseur après avoir su que l'herbe de jardin de M. Watson est humide aussi .c'est a dire la connaissance de W peut modifier la connaissance de P ,ou autrement dit l'information peut circuler dans la direction inverse .

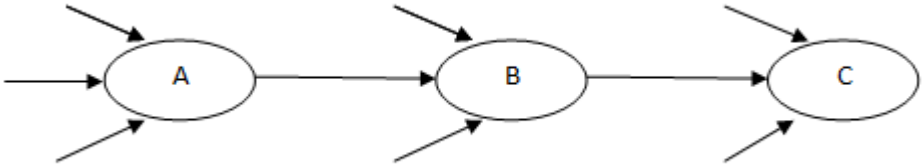
4. D-séparation :

La d-séparation est une information graphique qui renseigne sur la circulation de l'information dans un réseau causal. Elle explicite les conditions dans lesquelles l'information peut circuler entre deux sous ensemble de variables de manière complète .la d-séparation permet de déterminer si deux variables quelconque sont indépendantes conditionnellement a un ensemble de variable instanciées.

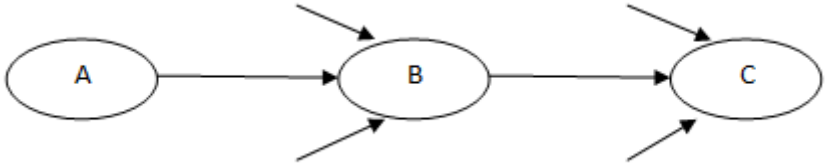
4.1.D-séparation cas des variables en série :

Soit le cas A, B et C sont en série P(A) peut être fixé ou bien dépendre d'autres valeurs.

P(A) dépend d'autres valeurs :



P(A) est fixé :



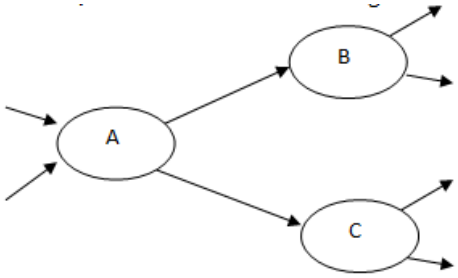
Dans les deux cas, P(C) dépend de P(A) via B.

Si on ne sait rien sur P(B), P(C) dépend toujours de P(A).

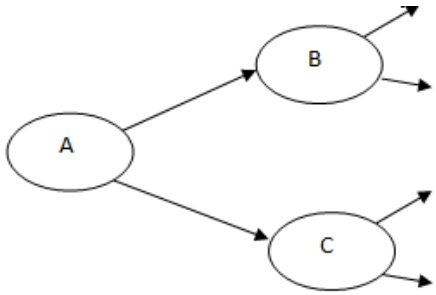
Par contre si P(B) est connu alors P(C) ne dépend plus de P(A).

4.2. D-séparation cas des variables divergentes :

A, B et C est en relation divergente



P(A) est variable



P(A) est fixé

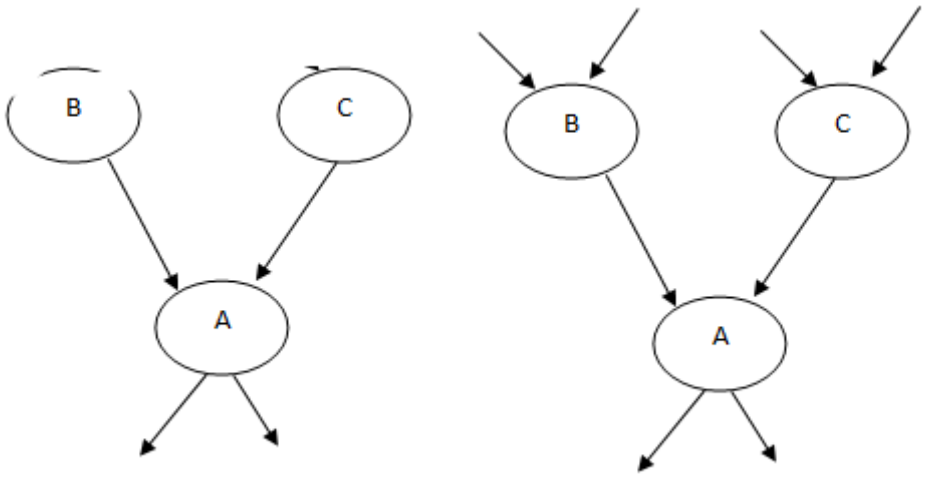
P(B) et P(C) dépendent l'une de l'autre via A.

Si P(A) est variable ou inconnu, alors P(B) et P(C) dépendent l'un de l'autre.

Si P (A) est fixée alors P (B) et P (C).ne ne dépendent pas l'un de l'autre.

4.3. D-séparation cas des variables convergentes :

A, B et C sont en relation convergente



P(B) et P(C) sont fixés P(B) et P(C) sont variables

P(B) et P(C) dépendent l'une de l'autre via A.

Si A est connu ou fixé alors P(B) et P(C) dépendent l'une de l'autre.

Si A est inconnu alors P(B) et P(C) ne dépendent pas l'une de l'autre.

BIBLIOGRAPHIE

[Abdou et al, 06] ABDOUROIHAMANE ANLI, MOURAD ABED PerSyst : Un Système de Personnalisation de l'information transport multimodale. Laboratoire d'Automatique, de Mécanique et d'Informatique industrielles et Humaines UVHC, Le Mont Houy, F-59300 Valenciennes Cedex 9, France, 2006.

[Abbadeni et al, 98] Abbadeni N., Ziou D., et Wang S., "Recherche d'images basée sur leur contenu", Rapport de recherche, université de Sherbrooke, Canada, 1998.

[Ach et al, 02] F.Achemoukh et R. Ahmed-Ouamer ,modelisation de profil utilisateur en recherche d'information personnalisé, Laboratoire LARI, Université Mouloud Mammeri, Tizi_ouzou , mars 2012.

[Achemoukh, 06] F.Achemoukh ,modele de langage pour la recherche d'information, these de magister, université Mouloud MAMMERI, Tizi_ouzou,2006.

[All, 79] Allen J.F,A plan based approch to speech act recognition.Technical report, dept of computer science,University of Toronto,Canada,79.

[Allan et al, 02] J. Allan, J. Aslam, N. Belkin, C. Buckley, J. Callan, B. Croft, S. Dumais, N. Fuhr, D. Harman, D. Harper, D. Hiemstra, T. Hofmann, E. Hovy, W. Kraaij, J. Lafferty, V. Lavrenko, D. Lewis, L. Liddy, R. Manmatha, A. McCallum, J. Ponte, J. Prager, D. Radev, P. Resnik, S. Robertson, R. Rosenfeld, S. Roukos, M. Sanderson, R. Schwartz, A. Singhal, A. Smeaton, H. Turtle, E. Voorhees, R. Weischedel, J. Xu, and C. Zhai. Challenges in information retrieval and language modeling : report of a workshop held at the center for intelligent information retrieval, university of Massachusetts amherst, september 2002.

[Amato, 99] G. Amato, U. Straccia, *User Profile Modeling and Applications to Digital Libraries*, Proc. 3rd European Conf. Research and Advanced Technology for Digital Libraries, ECDL, 1999.

[Asfari, 11] Ounas ASFARI Personalized Access to Contextual Information by using an Assistant for Query Reformulation, Thèse de Doctorat, université paris sud 11,2011.

[BADR, 07] Georges BADR, « Recherche d'information sur le web : Impact de la structure des documents sur la pertinence des résultats ».Maitrise, IRIT, Université Paul Sabatier&Institut National Polytechnique de Toulouse, 2007.

[Baeza-Yates, 99] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. Modern Information Retrieval. ACM Press / Addison-Wesley, 1999.

[Beklin, 92] N. Belkin and W. Croft. Information filtering and information retrieval: Two sides of the same coin? Communication of the ACM, 1992.

[BK, 05] M. Bouzeghoub and D. Kostadinov. Personnalisation de l'information : Aperçu del'état de l'art et définition d'un modèle flexible de définition de profils. In *Actes de la seconde édition de la Conférence en Recherche d'Information et Applications (CORIA)*, pages 201.218, Grenoble, France, 2005.

[Bottraud, 04] J.C. Bottraud , G. Bisson , M.F. Bruandet, *Apprentissage de profils pour un agent de recherche d'information*. Coria'04, IRIT Toulouse France, 2004.

[Boughanem, 12] M.Boughanem , cour Introduction a la recherche d'information, Université Paul Sabatier de Toulouse Laboratoire IRIT,2012.

[Bourne, 79] C. Bourne and B. Anderson. Dialog labworkbook. PaloAlto, Californie, USA, 1979.Second edition, Looked Information Systems.

[Bouzeghoub] M.Bouzeghoub, «Action spécifique sur la personnalisation de l'information », CNRS-AS98/RTP9, laboratoire PRiSM, Université de Versailles.

[Brini, 05] A. Brini. Un Modele de Recherche d'Information base sur les Reseaux Possibilistes. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 2005.

[Chen, 02] C. Chen, M. Chen, and Y. Sun. Pva : A self-adaptive personal view agent. *Journal of Intelligent Information Systems*, 18(2-3) :173.194, Mars 2002.

[Chevallet et al, 04] J.P Chevallet, J.Martinez, M.Boughanem, L.LechaniTamine, S.Calabretto, Rapport final de l'AS Passage a l'échelle dans la taille des corpus, janvier 2004.

[Chi, 01] David N.Chin, Empirical evaluation of user models and user adapted systems,user modeling and user-adapted interaction ,kluwer academic publishers,2001.

[Cle, 67] C. Cleverdon. The cranfield test on index language devices. *Aslib* , 1967.

[CP, 79] cohen P.R Perrault,C,R :Elements of a plan based theory of speech acts cognitive science 3,1979.

[Croft et al, 87] Croft, W., and Bruce, W. Approaches to intelligent information retrieval. Information Processing and Management: (1987), 249–254.

[Dan, 1986] J.P Daniels, cognitive models in information retrieval-An evaluation review,Journal of documentation,272:304,1986.

[De Campos et al, 02] De Campos L., Fernandez-Luna J., et Huete J., "A layered bayesian network model for document retrieval", In Proc. Of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval, pp. 169 – 182, 2002.

[De Campos et al, 03] De Campos L. M., Fernandez-Luna J. M., et Huete J. F., "The BNR Model: foundations and performance of Bayesian Network-based retrieval model", JASIST,302-313, 2003.

[DeClaris, 94] N. DeClaris, J. James, A. Nerode, W. Kohn, Intelligent integration of medical models, Proc. IEEE Conference on Systems, Man, and Cybernetics, San Antonio,1994.

[Eduardo, 05] Eduardo Sanchez Soto : Réseaux bayesiens dynamiques pour vérification du locuteur, thèse doctorat, 2005.

[Elayeb, 09] Elayeb Bilel Système multi-Agent de Recherche Intelligente Possibiliste de Documents Web, Thèse de Doctorat, Institut National Polytechnique de Toulouse, juin 2009.

[Fienberg, 05] Fienberg, S. E. When did bayesian inference become "bayesian»? Bayesian Analysis, 2005.

[Flurh, 85] C. Fluhr and F. Debili. Interrogation en langue naturelle de données textuelles et factuelles. In Intelligent Multimédia Information System and Management, Grenoble, France, 1985.

[Furh, 2000] N.Furh, Information retrieval: introduction and survey, Post-Graduate course en information retrieval, University of Duisburg-Essen, Germany, 2000.

[Gauch, 03] S. Gauch, J. Chaffe, A. Pretschner, *Ontology-Based User Profiles for Search and Browsing, To appear in J. User Modeling and User-Adapted Interaction*, the Journal of Personalization Research , Special Issue on User Modeling for Web and Hypermedia Information Retrieval, 2003.

[Gaussier, 03] Gaussier E., Stefanini MH., *Assistance intelligente à la recherche d'informations*, Hermes Science, ISBN 2-7462-0726-5, 2003

[Gow, 03] J.P Mc Gowan :Amultiple model approach to personalized information acces ,thesis of maste in computer science ,Faculty of science ,university college Dublin Fenrurary 2003.

[GRE, 84] S.Greenberg,User modeling in interactive computer systems.M.Sc thesis ,Dpt of computer science,University of Calgary ,1984.

[Hadjouni, 09] M.Hadjouni, H.Baazaoui, M.Aude Aufaure, H.Ben Ghezala , « Vers un système d'information pour la personnalisation sur le Web basé sur la modélisation de l'utilisateur »,2009 .

[Hallouli, 04] Hallouli K., "Reconnaissance de caractères par méthodes markoviennes et réseaux Bayésiens", Thèse de Doctorat spécialité Signal et Images, Ecole Nationale Supérieur des Télécommunications, Télécom Paris, Mai 2004.

[Ihab, 11] M.Ihab, De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information, Thèse de doctorat, Université Paul Sabatier, Toulouse III, juillet 2011.

[Janowski et al, 01] W. Janowski, A. Sarner. Five Opportunities for Personalization. Gartner Group, pp. 1, 05/2001.

[Koc, 00] Koch N., Software Engineering for Adaptative Hypermedia Systems – Reference Model ,Modeling techniques and development process, Ph.D Thesis, Fakultat der Mathematik und Informatik, Ludwig-Maximilians-Universität München, December 2000.

[Kostadinov, 03] D. Kostadinov. Personnalisation de l'information et gestion des profils utilisateurs, Rapport de DEA, Université de Versailles, France, 2003.

[Kurki, 99] T. Kurki, S. Jokela, R. Sulonen, M. Tirpeinen, Agents in delivering personalized content based on semantic metadata, In Proc, AAAI S*ping Symposium*, 1999.

[Lelu, 92] A. Lelu and C. François. Automatic generation of hypertext links in information retrieval systems. In *Communication of colloque ECHT'92*, New York, 1992. ACM Press.

[Mizzaro, 98] MIZZARO**, S.** "*How many relevance's in information retrieval?"*, Italie : Departement of Mathematics and Computer Science, University of Udine, 1998.

[Mooers, 48] Mooers C.N. Application of Random Codes to the Gathering of Statistical Information, MIT Master's Thesis, January 1948.

[Nassr, 02] N.Nassr Croisement de langues en recherche d'information : traduction et désambiguïsation de requêtes, thèse de Doctorat, Université Paul Sabatier de Toulouse, 2002.

[Neuhold, 03] J. E. Neuhold, Personalization and User profiling & Recommender Systems. Department of Computer Science and Business Informatics, University of Vienna ,WI/IM, Information management Proseminar 2003.

[Nguyen, 05] NGUYEN Trung Thanh Réseaux Bayésiens, Rapport du Travail d'Intérêt Personnel Encadré, 2005.

[Ouanis et al, 2006] Ounis I., Amati G., Plachouras V., He B., Macdonald C., Lioma C., « Terrier : A High Performance and Scalable Information Retrieval Platform », Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006), 2006.

[Pearl, 88] Judea Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[Pretschner, 99] Alexander Pretschner, Susan Gauch. *Ontology Based Personalized Search*. In Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), November 1999.

[RIB, 96] RIBEIRO B. A. N., MUNTZ R., « A belief network model for IR », SIGIR '96 : Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 1996, ACM Press, p. 253–260.

[Ribeiro-Neto et al, 96] Ribeiro-Neto B., Silva I., et Muntz R., "A Belief Network Model for IR", Proc. of the 19th ACM-SIGIR Conf. on Research and Development in Information Retrieval, 1996.

[Ricardo, 99] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. Modern Information Retrieval. ACM Press / Addison-Wesley, 1999.

[Roberston, 76] S. Robertson and K. Sparck Jones. Relevance weighting for search terms. Journal of The American Society for Information Science,1976.

[Roberston, 99] S. Robertson, S. Walker, and M. Beaulieu. Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive track, 1999.

[rozknop] A.Rozenknop, Recherche et extraction d'information, cour Master MICR, paris 13.

[Salton, 71] Salton, G. "The SMART Retrieval System- Experiments in Automatic Document Processing", Prentice-Hall, 1971.

[Salton, 88] Salton, G. Syntactic approaches to automatic book indexing. InProc. of the annual meeting on Association for Computational Linguis-tics (ACL) (1988), Department of Computer Science, Cornell University,Ithaca, New York, pp. 204–210.

[Smail, 04] L.Smail : algorithmes pour les réseaux bayésiens et leurs extensions, thèse doctorat de l'université de polytech Nantes, 2004.

[Smail, 09] N.Smail , Contribution à l'analyse et à la recherche d'information en texte intégral. Thèse de Doctorat ,Université Paris-Est,2009.

[Somlo, 03] G L. Somlo, A. E. Howe, *Using Web Helper Agent Profiles in Query Generation International Conference on Auonomous Agents.* Proceedings of the second international joint

conference on Autonomous agents and multi agent systems Melbourne, Australia Web technologies. 2003.

[Sorensen, 95] PSUN : A Profiling System for Usenet News, H. Sorensen, M. Mc Elligott, CIKM'95 Intelligent Information Agents Workshop, Baltimore, December 1995.

[Tamine, 05] L.Tamine, M.Boughanem « accès personnalisé a l'information, Approches et techniques », rapport interne, Institut de recherche en informatique de Toulouse, janvier 2005.

[Tamine, 98] L.Tamine « les systèmes de recherche d'information : reformulation de requête et apprentissage bases sur les algorithmes génétiques », these de magister, université Mouloud MAMMERI, Tizi_ouzou ,1998.

[Thi, 09] Thi Hoang Diem LE, Utilisation de ressources externes dans un modèle Bayésien de Recherche d'Information. Application à la recherche d'information médicale multilingue avec UMLS. Thèse de doctorat, Université Joseph Fourier - Grenoble I, 2009.

[Trousse, 01] B. Trousse. Recommandations personnalisées pour l'aide à la recherche d'informations Web basées sur l'analyse et l'utilisation du comportement des utilisateurs, ActionAxIS, INRIA Sophia Antipolis, 2001.

[Turtle, 91] H.R. Turtle and Bruce W.Croft. Inference Networks for Document Retrieval .PhD thesis, 1991.

[Turtle et al, 90] H. Turtle, W. B. Croft, Inference networks for document retrieval. Proceedings of ACM SIGIR 90, pages: 1-24, 1990.

[Turtle et al, 91] H.Turtle and Bruce W.Croft.Evaluation of an inference network-based retrieval model,ACM Transaction of Information Systems,p.187_222,1991.

[Van, 89] van Rijsbergen, C. J. Towards an information logic. In In Proc. Of the International ACM-SIGIR Conference (1989), pp. 77–86.

[Zemirli, 03/04] Zemirli W.N, Vers le développement d'un système de recherche d'information personnalisé intégrant le profil utilisateur, Université Paul Sabatier – Toulouse III, Equipe SIG/RI ,2003/2004.

[Zemirli, 08] Zemirli W.N, « Modèle d'accès personnalisé a l'information basé sur les diagrammes d'influence intégrant un profil multidimensionnel », Thèse de doctorat, Université Paul Sabatier, Toulouse, France, juin 2008.

[ZTB, 05] W. N. Zemirli, L. Tamine, and M. Boughanem. Accès personnalisé à l'information : vers la définition d'un profil utilisateur multidimensionnel. In *International Symposium On Programming Systems*, pages 20.28. USTHB, 2005.