
Présentation et évaluation d'un modèle d'accès personnalisé à l'information basé sur les diagrammes d'influence

W. Nesrine Zemirli — Lynda Tamine-Lechani — Mohand Boughanem

*IRIT (Institut de Recherche en Informatique de Toulouse) – UMR 5505
118 route de Narbonne
31062 Toulouse cedex 9
{nzemirli, lechani,bougha}@irit.fr*

RÉSUMÉ. L'objet de cet article est de décrire un modèle formel capable d'intégrer l'utilisateur dans le processus d'accès à l'information. Nous avons orienté nos travaux vers l'utilisation des diagrammes d'influence comme support théorique nous permettant de formaliser l'utilité des décisions associées à la pertinence des documents compte tenu de la requête et du profil de l'utilisateur. L'idée de base est de substituer à la fonction de pertinence classique qui mesure le degré d'appariement requête-document une fonction indexée par l'utilisateur. Dans notre approche, le profil utilisateur est représenté par la dimension de ses centres d'intérêt caractérisant ses besoins récurrents. N'ayant pas de cadre standard d'évaluation, nous proposons un cadre d'évaluation adapté à l'accès personnalisé à l'information en augmentant les collections de la campagne TREC par des profils utilisateurs simulés. Nous validons notre contribution par comparaison au modèle de recherche bayésien classique.

ABSTRACT. The goal of this paper is to describe a formal model able to integrate the user in the process of information access. We have directed our works towards the use of influence diagrams as a theoretical support allowing us to formalize the utility of the decisions associated to the relevance of the documents taking into account the query and the user profile. The basic idea is to substitute to the traditional relevance function which measures the degree of matching document-query, a function indexed by the user. In our approach, the user profile is represented by his long term interests. As there does not exist a standard evaluation protocol, we propose one by integrating simulated users to the standard TREC collections. We validate experimentally our contribution by comparison to the basic bayesian retrieval.

MOTS-CLÉS : recherche d'information personnalisée, profil utilisateur, diagramme d'influence

KEYWORDS: personalized information retrieval, user profile, influence diagramm

Table des matières

1	Introduction	2
2	Accès personnalisé à l'information : Problématique et aperçu d'état de l'art	3
2.1	Problématique	3
2.2	Synthèse des travaux	4
3	Le modèle d'accès personnalisé à l'information	5
3.1	Le cadre formel : les réseaux Bayésiens	6
3.2	Description du modèle	6
3.2.1	La composante qualitative	6
3.2.2	La composante quantitative	8
3.2.3	Evaluation de la requête	10
4	Evaluation expérimentale	11
4.1	Simulation des centres d'intérêt	12
4.2	Protocole d'évaluation	12
4.3	Expérimentations et résultats	13
5	Conclusion	14
6	Bibliographie	15

1. Introduction

Face au phénomène actuel d'accroissement incessant d'informations hétérogènes, la nécessité de systèmes de recherche d'information (SRI) efficaces se fait de plus en plus sentir. Le but de ces systèmes est d'aider l'utilisateur à trouver l'information désirée parmi la masse de données disponibles. Ainsi, le problème n'est pas tant la disponibilité de l'information mais sa pertinence relativement à un contexte d'utilisation particulier. Les travaux [ALL 02, COO 01] 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. Dès lors l'accès à l'information tend vers une nouvelle définition [ALL 02] : *Com-*

bine 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.

Dans ce cadre, notre objectif consiste à proposer un modèle formel capable d'intégrer l'utilisateur dans le processus d'accès à l'information. A cet effet, on s'oriente vers l'utilisation des diagrammes d'influence [ROC 71, JEN 01] comme support théorique nous permettant de formaliser l'utilité des décisions associées à la pertinence des documents compte tenu de la requête et du profil de l'utilisateur. L'idée de base est de substituer à la fonction de pertinence classique qui mesure le degré d'appariement requête-document, une fonction indexée par l'utilisateur. Dans notre approche, le profil utilisateur est représenté par la dimension de ses centres d'intérêts caractérisant ses besoins récurrents. Cette dimension est déterminée et évolue à partir des interactions de l'utilisateur avec le SRI [TAM 06b].

Le présent article est organisé comme suit : la section 2 aborde la problématique à l'origine de la personnalisation de l'accès à l'information puis rapporte une synthèse des travaux du domaine. La section 3 présente la description de notre modèle d'accès à l'information. Le modèle étant basé sur les diagrammes d'influence, nous en décrivons alors en premier le cadre formel puis la topologie du diagramme à travers les aspects qualitatifs et quantitatifs du modèle. La section 4 détaille notre protocole d'évaluation et présente les résultats expérimentaux. La section 5 résume notre contribution et présente les perspectives envisagées.

2. Accès personnalisé à l'information : Problématique et aperçu d'état de l'art

2.1. Problématique

On estime actuellement que 63% à 66% des 85% des utilisateurs sont insatisfaits de la qualité des réponses fournies par les SRI [KOB 00]. Ceci est dû à plusieurs facteurs dont celui lié au degré de prise en compte effective de l'utilisateur dans le modèle de recherche. En effet, force est de constater qu'il n'y a généralement pas de mécanisme explicite qui représente et intègre l'utilisateur dans le processus d'accès à l'information [BUD 00]. Ainsi, les modèles de recherche d'information classiques supposent que l'utilisateur est complètement représenté par sa requête : les résultats retournés pour une même requête sont identiques même si elle est exprimée par des utilisateurs différents. Pour cause, l'évaluation des requêtes s'effectue sans prise en compte des centres d'intérêts récurrents de l'utilisateur, ni de ses préférences en qualité, ni du contexte de recherche. Les problèmes engendrés par une telle représentation sont essentiellement l'impossibilité de sélectionner des sources opportunes et l'intelligibilité des résultats, d'autant que la majorité des requêtes exprimées sont courtes (2.29 mots en moyenne par requête) et que les sources d'information sont volumineuses et hétérogènes [BUD 00].

En réponse à ce constat, des travaux s'orientent actuellement vers la révision de la chaîne d'accès à l'information dans la perspective d'intégrer l'utilisateur dans le pro-

cessus d'accès à l'information. Ces travaux s'inscrivent dans le cadre de la recherche personnalisée [GAU 03, LIU 04, KIM 03]. Les systèmes d'accès personnalisé à l'information se basent sur la représentation et l'exploitation des préférences, les intérêts et les habitudes de l'utilisateur, dans ce que l'on nomme *profil utilisateur*, dans le but de retrouver les informations les plus pertinentes en adéquation avec le profil.

Plusieurs systèmes d'accès personnalisé furent développés [PAZ 96, CHE 98] mais ils diffèrent par le type de service d'accès qu'ils offrent. Cela est principalement dû à la manière d'exploiter le profil utilisateur. En effet, c'est en fonction de l'intégration du profil dans une ou toutes les phases du cycle de vie de la requête, que l'on peut spécifier le type de services offerts par le système.

Une étude préalable des systèmes de personnalisation existants [ZEM 05] a montré que les problèmes majeurs de mise en œuvre concernent la modélisation et l'exploitation du profil utilisateur. En effet, la difficulté à ce niveau porte sur plusieurs points concernant l'étape de construction du profil utilisateur qui doit fournir une représentation fidèle des besoins et centres d'intérêt récurrents de l'utilisateur, ainsi que l'évolution de ce profil. D'autre part l'étape, d'exploitation du profil, dans le processus d'accès à l'information (reformulation de la requête, exécution de la requête, réordonnement des résultats etc.), doit être en mesure d'intégrer les informations adéquates du profil permettant d'améliorer les résultats de recherche.

Le paragraphe suivant présente une synthèse des travaux ayant abordé la problématique de l'accès contextuel à l'information.

2.2. Synthèse des travaux

La personnalisation vise à satisfaire les utilisateurs en leur présentant les informations pertinentes et valables pour leurs besoins spécifiques, et ainsi optimiser l'interaction pour une efficacité maximale de la recherche [VAS 02]. La mise en œuvre d'un tel système s'effectue donc en premier par la modélisation du profil de l'utilisateur, puis par son intégration dans le processus d'accès à l'information. L'étape de modélisation consiste à proposer un profil utilisateur contenant l'ensemble des caractéristiques informationnelles de l'utilisateur, correspondant généralement à ses centres d'intérêt récurrents. Ces derniers sont représentés selon trois principales approches : ensembliste [BUD 00], sémantique [GAU 03, CHA 04, LIN 05] et multidimensionnelle [AMA 99, BOU 05].

Lors de la phase d'évaluation de la requête, les paramètres modélisant le profil utilisateur sont inclus dans la fonction de calcul de pertinence des documents. Dans ce cadre, on trouve dans [FAN 04] une proposition pour l'adaptation des paramètres de la fonction de pertinence au contexte de l'utilisateur, en utilisant les techniques de programmation génétique. Jeh et Widom [JEH 03] ont proposé une variante personnalisée de l'algorithme PageRank en l'occurrence PPV (Personalized PageRank Vector). L'algorithme adapte le principe de calcul de l'Authority d'une page, donné par l'al-

gorithme PageRank, en utilisant une distribution de probabilités tenant compte de la présence de liens entrants et liens sortants vers les pages favorites de l'utilisateur.

Dans [SPE 04], l'approche de personnalisation consiste, via un calcul de similarité, à réordonner les résultats retrouvés en combinant l'ordre produit par le processus de sélection et celui donné par le contexte de utilisateur, représenté par une liste de mots-clés issus des documents sélectionnés. En outre, il est possible d'effectuer une reformulation de la requête avant de lancer le processus d'appariement, ceci en augmentant la requête des informations issues du profil utilisateur ; ce qui mène à une meilleure interprétation de la requête [COO 02].

Une autre approche [SHE 05] se base sur un modèle probabiliste d'accès à l'information utilisant des sources de croyance liées aux contextes. Plus précisément, ceci consiste à exploiter, en premier lieu, l'historique des interactions de l'utilisateur avec le SRI pour étendre le modèle de langage de la requête ; en second lieu à appliquer la méthode classique de Kulback-Leiber pour calculer le taux de divergence du modèle de langage des requêtes et documents, traduisant ainsi le score de pertinence.

3. Le modèle d'accès personnalisé à l'information

Le modèle d'accès à l'information que nous proposons peut être défini par un quadruplet $M = \langle I, U, E^t(U), P(I, U) \rangle$ où I est le modèle de représentation de l'information (documents et requêtes), U est le modèle de représentation de l'utilisateur, $E^t(U)$ est la fonction d'évolution de l'utilisateur en fonction du temps et $P(I, U)$ est la fonction de pertinence d'une information relativement à un utilisateur [TAM 06a]. Dans notre approche, le profil utilisateur est représenté par la dimension de ses centres d'intérêt caractérisant ses besoins récurrents. Cette dimension est déterminée et évolue à partir des interactions de l'utilisateur avec le SRI.

Dans ce papier, nous focalisons sur le modèle décisionnel d'accès personnalisé à l'information (composante $P(I, U)$ du modèle). Le calcul du score de pertinence d'un document relativement à une requête émise par l'utilisateur, est fondé sur une fonction permettant d'estimer l'utilité de présenter à l'utilisateur ce document et ce, compte tenu de sa requête et de ses centres d'intérêt spécifiques. Dans le but de formaliser cette fonction de pertinence, notée $P(I, U)$, on s'oriente vers l'utilisation d'une extension des réseaux bayésiens [PEA 88, JEN 01] en l'occurrence les diagrammes d'influence [SHA 88]. L'idée de base est de substituer à la fonction de pertinence classique qui mesure le degré d'appariement requête-document $RSV(Q, D) = p(Q/D)$, une fonction indexée par l'utilisateur $RSV_u(Q, D) = p(D/Q, U)$ où $p(A/B)$ est la probabilité conditionnelle de l'événement A sachant l'événement B et U représente l'utilisateur décrit par ses centres d'intérêt.

Suite à la présentation des principes théoriques des réseaux bayésiens et des diagrammes d'influence, nous en illustrons l'application à notre problème précis : mettre en œuvre un appariement requête-document-utilisateur.

3.1. *Le cadre formel : les réseaux Bayésiens*

Les Réseaux Bayésiens (RBs) sont utilisés dans divers domaines et constituent un outil puissant pour la représentation des connaissances et le traitement de l'incertitude. 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 [TUR 90]. 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. C'est pour cette raison que différents travaux [RIB 96] ont tenté d'exploiter l'apport des RBs pour définir des modèles de recherche d'information.

Un réseau bayésien (RB) (dit « réseau de croyance » ou « réseau probabiliste ») est un modèle graphique orienté sans cycle. Les modèles graphiques sont issus du mariage entre la théorie des graphes et la théorie des probabilités. Un graphe est appréhendé selon un aspect qualitatif et un aspect quantitatif. L'aspect qualitatif est l'ensemble des nœuds du graphe représentant les variables du domaine traité (les documents de la collection, les termes d'indexation des documents ou de la requête et du besoin utilisateur), ainsi que les relations de dépendance entre ces variables. Ces dépendances permettent d'effectuer des inférences, offrant ainsi un support à la prise de décision. L'aspect quantitatif permet d'évaluer les arcs reliant toute paire de nœuds au moyen d'un calcul de probabilités.

3.2. *Description du modèle*

Les diagrammes d'influence (DI) constituent une extension des RB à un problème de décision. Notre choix pour l'utilisation des DI est motivé par le besoin d'utiliser un cadre théorique pour la formalisation du problème décisionnel lié à la présentation d'un document à l'utilisateur compte tenu de son utilité vis-à-vis de la requête émise et des centres d'intérêt spécifiques de l'utilisateur. Dans notre approche, nous considérons uniquement la sous-catégorie des « centres d'intérêt » de l'utilisateur comme composantes explicites représentées sous forme de listes de termes pondérés. Après avoir présenté les différentes composantes du modèle, nous passerons à son exploitation du point de vue de l'évaluation de la requête.

3.2.1. *La composante qualitative*

La composante qualitative est représentée par le graphe acyclique $G = (V, E)$ où V comprend les nœuds représentant des variables aléatoires X_1, X_2, \dots, X_n . A chaque variable X_i est associée un ensemble de valeurs mutuellement exclusives définies dans $dom(X_i)$. L'ensemble E comprend les arcs existants entre nœuds qui traduisent des relations de causalité décrites par des probabilités conditionnelles attachées à chaque nœud. La topologie du diagramme d'accès personnalisé à l'information est illustrée par l'exemple suivant :

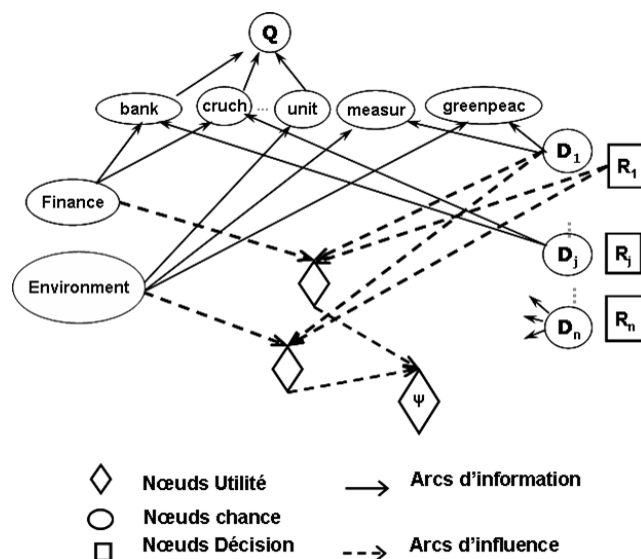


Figure 1 – Exemple du Diagramme d’Influence pour les centres d’intérêt *Finance* et *Environment* du disque 1, 2 de TREC

◇ **Les variables** : L’ensemble des variables V est composé de trois sous-ensembles de nœuds :

– **Les nœuds chance**. Nous distinguons quatre types de nœuds chance $V^{info} = Q \cup D \cup T \cup C$. Le nœud unique Q correspondant à la requête de l’utilisateur est représenté par une variable aléatoire binaire définie dans l’ensemble $dom(Q) = \{q, \bar{q}\}$, où q désigne que la requête Q est *satisfaite* et \bar{q} désigne que la requête *n’est pas satisfaite* ; dans notre cas, on ne s’intéressera qu’à l’instanciation positive q . L’ensemble $D = \{D_1, D_2, \dots, D_n\}$ correspond aux documents de la collection. Chaque nœud document D_j , représente une variable aléatoire binaire prenant des valeurs dans l’ensemble $dom(D_j) = \{d_j, \bar{d}_j\}$, où d_j désigne que *le document D_j est observé*¹, et \bar{d}_j désigne que *le document D_j n’est pas observé*. L’ensemble $T = \{T_1, T_2, \dots, T_m\}$ correspond aux termes d’indexation. Chaque nœud terme T_i représente une variable aléatoire binaire prenant des valeurs dans l’ensemble $dom(T_i) = \{t_i, \bar{t}_i\}$, où t_i désigne que *le terme T_i est pertinent pour la requête Q* et \bar{t}_i désigne que *le terme T_i n’est pas pertinent pour Q* . La pertinence d’un terme signifie sa présence éventuelle dans un document observé. L’ensemble $C = \{C_1, C_2, \dots, C_u\}$ correspond aux centres d’intérêt associés à l’utilisateur. Chaque nœud centre d’intérêt C_k représente une variable aléatoire binaire prenant des valeurs dans l’ensemble $dom(C_k) = \{c_k, \bar{c}_k\}$, où

1. Un document observé, dans la théorie de Turtle [TUR 90], traduit un document tiré aléatoirement de la collection et dont on calcule le degré de pertinence

c_k désigne que *le centre d'intérêt C_k est observé*² et \bar{c}_k désigne que *le centre d'intérêt C_k n'est pas observé*. La pertinence d'un centre d'intérêt traduit le fait qu'il couvre l'objet de la requête.

– **Les nœuds décision.** On associe à chaque document D_j de la collection, un nœud décision R_j prenant ses valeurs dans $dom(R_j) = \{r_j, \bar{r}_j\}$ ce qui correspond respectivement aux décisions de présenter ou pas le document D_j à l'utilisateur et ce, compte tenu de sa requête et de son profil décrit par ses centres d'intérêt ;

– **Les nœuds utilité.** Un *nœud utilité* exprime *l'utilité de la décision de présenter un document compte tenu des centres d'intérêt* de l'utilisateur. De ce fait, on associe un nœud utilité à chaque document D_j et chaque centre d'intérêt C_k , ce nœud correspondant à l'évaluation de la pertinence du document D_j vis-à-vis de la requête, et au regard du centre d'intérêt C_k . Les valeurs restituées par le sous-ensemble de ces nœuds concernant un document D_j de la collection sont utilisées par un nœud utilité particulier qui les intègre dans le calcul de *l'utilité globale de la décision de restituer ce document D_j en considérant tous les centres d'intérêt* de l'utilisateur.

◇ **Les arcs :** On distingue deux types d'arcs :

– **Les arcs d'information,** qui relie chacun des nœuds termes T_i de l'index de D_j noté $\tau(D_j)$ aux nœuds documents D_j qu'ils indexent, ainsi qu'aux nœuds centre C_k qu'ils représentent. Il existe également des arcs d'information qui relie chaque terme T_i avec le nœud requête Q .

– **Les arcs d'influence,** qui traduisent le degré d'influence des variables associées à la décision prise. Dans le cas de notre modèle, des arcs d'influence relient les nœuds décision, centres d'intérêt et documents en utilisant un opérateur d'agrégation.

3.2.2. La composante quantitative

La composante quantitative comprend des distributions de probabilités conditionnelles où pour chaque variable $x_i \in V$, est attachée une classe de probabilités $P(X) = P(x_i/pa(x_i))$ qui est fonction de toutes les configurations possibles de ses nœuds parents $pa(x_i)$ dans G , notée θ . Par exemple, soit la requête Q composée des deux termes T_1 et T_2 , $Q = \{T_1, T_2\}$; alors l'ensemble des configurations possibles des parents de la requête, tel que leur domaine est binaire, est $\theta = \{\{t_1, t_2\}, \{t_1, \bar{t}_2\}, \{\bar{t}_1, t_2\}, \{\bar{t}_1, \bar{t}_2\}\}$. L'instance θ_1^1 du terme T_1 dans la première configuration de θ , $\theta_1^1 = \{t_1, t_2\}$, est $\theta_1^1 = t_1$.

◇ **Distribution de probabilités :**

La quantification des distributions de probabilités dans le réseau, consiste à donner une sémantique (basée sur les probabilités) des arcs reliant les différents types de nœuds du réseau, et à estimer l'utilité des décisions de présentation des documents pertinents compte tenu de tous les centres d'intérêt de l'utilisateur.

2. Un centre d'intérêt observé, est par analogie au document, un centre tiré aléatoirement du profil de l'utilisateur et dont on calcule le degré de pertinence relativement à la requête.

– *Nœud requête*. La probabilité $P(Q/pa(Q))$ traduit le degré de correspondance entre la configuration de termes donnée par $pa(Q)$ avec la configuration initiale de la requête. Le calcul de $P(Q/pa(Q))$ est effectué sur la base d'une agrégation Ou-Flou proposée dans [PEA 88], de la manière suivante :

$$P(Q/pa(Q)) = 1 - \prod_{t_i \in R(pa(Q))} (1 - nidf(t_i)) \quad [1]$$

Où $R(pa(Q))$ est l'ensemble des configurations instanciées positivement parmi les configurations possibles $pa(Q)$ des termes parents du noeud Q . $nidf(t_i)$ est la fréquence inverse normalisé du terme t_i dans la collections.

– *Nœud terme*. La probabilité $P(t_i/d_j, c_k)$ traduit la représentativité du terme T_i dans la représentation du document D_j et le centre C_k . En supposant que les documents et les centres sont indépendants, l'estimation de la probabilité correspond donc à :

$$P(t_i/d_j, c_k) = P(t_i/d_j) * P(t_i/c_k) \quad [2]$$

où

$$P(t_i/d_j) = \delta + (1 - \delta) * Wtd(i, j), \delta \in]0, 1[\quad [3]$$

$$P(t_i/c_k) = \gamma + (1 - \gamma) * Wtc(i, k), \gamma \in]0, 1[\quad [4]$$

Avec $Wtd(i, j)$ et $Wtc(i, k)$ représentant respectivement les degrés d'importance du terme T_i dans le document D_j et le centre C_k . La formule de pondération $Wtc(i, k)$, détaillée et expérimentée, sera présentée en sections 4. $Wtd(i, j)$ est donné par la formule BM25 [ROB 97] :

$$Wtd(i, j) = 0,5 * \frac{tf_{ij} \log\left(\frac{N-n_i+0,5}{n_i+0,5}\right)}{2 * \left(0,25 + \frac{0,75 * dl_j}{avg - dl}\right) + tf_{ij}} \quad [5]$$

où : n_i : le nombre de documents contenant t_i , N : le nombre de documents pertinents dans la collection, dl : la longueur du document d_j , $avg - dl$: la longueur moyenne des documents de la collection, tf_{ij} : la fréquence d'apparition du terme t_i dans le document d_j

◇ Valeur d'utilité :

Le degré d'utilité de la décision de présenter ou non un document dépend à la fois du contenu du document et des centres d'intérêt relativement à la requête en cours d'évaluation [TAM 06a]. En effet, l'utilité globale $EU(R_j/Q)$ de la décision de restituer un document observé D_j en réponse à la requête Q de l'utilisateur, intègre les utilités calculées pour le noeud document D_j et chacun des noeuds centres d'intérêts $C_k \in C$.

Cette integration est effectuée en se basant sur un opérateur d'agrégation basique en l'occurrence la somme.

$$EU(R_j/Q) = \sum_{c \in C(1..u)} u(R_j/c_k) * P(Q/d_j, c_k) * P(c_k) \quad [6]$$

$P(Q/d_j, c_k)$ est la probabilité que la requête Q soit satisfaite compte tenu du document d_j et du centre d'intérêt c_k .

La valeur de l'utilité élémentaire exprime le degré de concordance entre le centre d'intérêt instancié et le document observé. On propose la formulation suivante :

$$u(R_j/C_k) = \frac{1 + \sum_{T_i \in D_j} nidf_i}{1 + \sum_{T_i \in D_j - C_k} nidf_i} \in [1, 1 + \sum_{T_i \in D_j} nidf_i] \quad [7]$$

Cette valeur sera égale à 1 lorsqu'il n'y aura aucune correspondance entre le centre d'intérêt C_k et le document D_j observés. Plus il y aura de termes communs entre le document et le centre d'intérêt, plus cette valeur d'utilité sera importante.

3.2.3. Evaluation de la requête

Dans le cadre de notre approche, l'évaluation de requête RSV_u revient à propager des nouvelles croyances (document et centre d'intérêt observés) selon la structure du diagramme d'influence. Plus précisément, le processus d'évaluation est enclenché comme dans un problème décisionnel en maximisant la mesure d'utilité :

$$RSV_u(Q/D_j) = EU(R_j/Q) = \sum_{C_k \in C(1..u)} u(R_j/C_k) * P(Q/D_j, C_k) * P(C_k) \quad [8]$$

La propagation dans ce modèle consiste à calculer, pour chaque noeud, la probabilité *a posteriori* étant donné les probabilités conditionnelles et marginales *a priori*. La propagation tente de calculer la probabilité que la requête soit satisfaite étant donné un document instancié positivement (d_j) en observant un par un les centres d'intérêt (c_k) de l'utilisateur. Pour chaque document, ce processus est réitéré pour tous les centres d'intérêt ; puis l'algorithme reprend pour considérer tous les documents candidats à l'évaluation.

La quantification revient à estimer chaque membre de la formule(8). Partant du principe de marginalisation et en tenant compte de l'hypothèse d'indépendance entre les nœuds documents et les nœuds centres d'intérêt, on obtient par application de la loi jointe, le développement de la probabilité suivante :

$$P(Q/d_j, c_k) = \sum_{\theta^s \in \theta} [P(Q/\theta^s) * \prod_{T_i \in Q \cap (D_j \cup C_k)} P(\theta_i^s/d_j) * P(\theta_i^s/c_k)] \quad [9]$$

Où θ représente l'ensemble des configurations possibles des termes d'indexation de $pa(Q)$, θ^s la configuration d'ordre s associée, et θ_i^s la configuration d'ordre s associée au terme $t_i \in pa(Q)$

Des probabilités *a priori* sont affectées aux documents de la collection et aux centres d'intérêt de l'utilisateur. Etant donné que ces probabilités sont équiprobables pour l'ensemble des variables du modèle, elles n'affectent aucunement le calcul final de la fonction de pertinence. En appliquant cette dernière simplification ainsi que celle de la formule (8), la formule du calcul de pertinence (9) devient :

$$RSV_u(Q/D_j) = \sum_{C_k \in C_{(1..u)}} u(R_j/C_k) * \sum_{\theta^s \in \theta} [P(Q/\theta^s) * \prod_{T_i \in Q \cap (D_j \cup C_k)} P(\theta_i^s/d_j) * P(\theta_i^s/c_k)] \quad [10]$$

4. Evaluation expérimentale

Actuellement, il n'existe pas, à notre connaissance, de cadre d'évaluation standard pour une tâche d'accès personnalisé à l'information. Dans le but d'y remédier d'une part et d'exploiter au mieux les ressources de la campagne de référence TREC (Text Retrieval Conference, <http://trec.nist.gov>) d'autre part, nous proposons un cadre d'évaluation par augmentation des collections TREC de centres d'intérêt simulés. Plus précisément, nous simulerons les profils des utilisateurs en créant des centres d'intérêt à partir de requêtes issues des disques 1, 2 de la tâche ad hoc de la campagne d'évaluation TREC. Le choix de cette collection a été motivé par le fait que ces requêtes sont décrites par un champ particulier qui spécifie leurs domaines respectifs et qui nous permet de simuler les centres d'intérêt de l'utilisateur. L'ensemble des domaines de la collection est illustré dans la Figure 2.

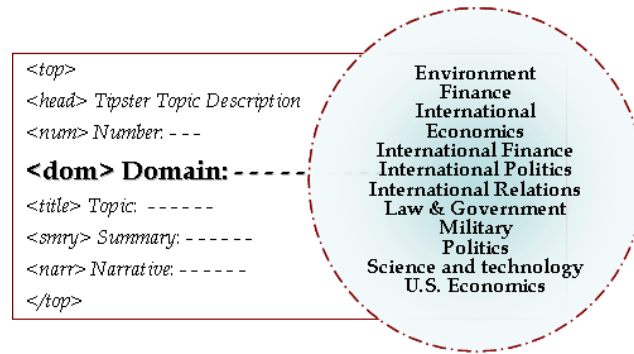


Figure 2 – Liste des domaines associés aux requêtes de la collection

Nous présentons dans ce qui suit la méthode utilisée pour la simulation des centres d'intérêt des utilisateurs puis nous présenterons le protocole d'évaluation adapté suivi des résultats expérimentaux obtenus.

4.1. Simulation des centres d'intérêt

Notre hypothèse de travail suppose qu'un domaine de requête correspond à un centre d'intérêt plausible pour l'utilisateur. A chaque domaine est associé un nombre déterminé de requêtes nous permettant d'inférer les centres d'intérêt. Pour cela, nous procédons comme suit :

- Pour chaque domaine traité par les documents de la collection, nous sélectionnons, parmi les requêtes qui lui sont associées, un sous-ensemble qui constitue l'ensemble d'apprentissage des centres d'intérêt ;
- A partir de cet ensemble d'apprentissage, un processus automatique se charge de récupérer, la liste des vecteurs³ documents pertinents et non pertinents associés à chaque requête.

Partant de ces vecteurs documents, un centre d'intérêt est construit comme un vecteur de termes pondérés selon l'algorithme d'apprentissage de Rocchio [ROC 71] :

$$C_k = \frac{\alpha}{|R|} \sum_{i=1}^R DR_i - \frac{\beta}{|NR|} \sum_{j=1}^{NR} DNR_j, \alpha = 0.75, \beta = 0.25 \quad [11]$$

où : DR : les documents pertinents associés à la requête, DNR : les documents non pertinents associés à la requête, R : le nombre total de documents pertinents de la requête, NR : le nombre total de documents non pertinents de la requête.

4.2. Protocole d'évaluation

Pour évaluer le modèle que nous proposons nous avons besoin d'un modèle de référence permettant de quantifier l'apport du profil utilisateur dans le processus d'accès à l'information. Nous pouvons comparer notre modèle à n'importe quel modèle classique de recherche d'information ne tenant pas compte des centres d'intérêt de l'utilisateur. Cependant, notre modèle étant une extension des réseaux bayésiens, il est plus significatif de considérer comme référence les résultats obtenus avec un tel modèle. De ce fait, lors de nos expérimentations, nous avons implémenté un modèle de recherche bayésien classique qui évalue la fonction de pertinence $RSV(Q, D)$, en mesurant le degré d'appariement requête-document selon la formule suivante :

$$P(Q/D_j) = \sum_{\theta^s \in \theta} [P(Q/\theta^s) * \prod_{T_i \in (Q \cap D_j)} P(\theta_i^s/D_j)] \quad [12]$$

Dans le but de mesurer l'impact de l'intégration du profil utilisateur dans le processus d'accès à l'information nous optons pour un cadre d'évaluation TREC avec un

3. Le vecteur document est composé de poids associés à chacun des termes d'indexation

scénario qui se base sur la méthode de la *validation croisée* et ce, pour ne pas biaiser les résultats avec un seul jeu de test. La validation croisée [MIT 97] ou la *k-fold cross validation* est une méthode d'évaluation qui consiste à diviser la collection de test en k sous ensembles de tailles égales (approximativement), d'utiliser $k - 1$ sous ensembles pour l'apprentissage des centres d'intérêt dans notre cas, et le $k^{ième}$ sous ensemble pour le test. On réitère ensuite le processus k fois pour chacun des centres d'intérêt évalué.

Par ailleurs, la méthode d'évaluation est faite selon le protocole TREC. Plus précisément, pour chaque requête de la collection, les 1000 premiers documents sont restitués par le système et des précisions sont calculées à différents points (5, 10, 15, 30, 100 et 1000 premiers documents restitués), puis une moyenne de toutes ces précisions est calculée. Nous comparons ensuite les résultats obtenus en utilisant notre modèle à ceux obtenus en utilisant le modèle de référence.

4.3. Expérimentations et résultats

Notre scénario d'évaluation consiste à évaluer le système avec quatre utilisateurs. Chaque profil utilisateur ne contient qu'un seul centre d'intérêt représentant un domaine spécifique choisi parmi ceux associés aux requêtes, en l'occurrence celles présentées dans le tableau 1.

Domaines	Requêtes associées
Environment	59 77 78 83
Law & Government	70 76 85 87
Military	62 71 91 92
Political	74 80

Tableau 1 – Domaines choisis pour la construction des profils utilisateurs

Nous avons effectué plusieurs expérimentations afin de :

- 1) paramétrer les distributions de probabilités associées aux différents noeuds du modèle ;
- 2) évaluer l'impact de l'intégration du profil dans le calcul de la fonction de pertinence.

Les expérimentations préliminaires réalisées par variation des différents paramètres nous ont amenés à fixer les probabilités suivantes :

$$P(t_i/d_j) = 0,5 + 0,5 * Wtd(i, j) \quad [13]$$

$$P(t_i/c_k) = 0,1 + 0,9 * Wtc(i, k) \quad [14]$$

La seconde série d'expérimentations a permis d'évaluer l'apport de notre modèle dans les performances de recherche. Les résultats obtenus pour chacune des requêtes test sont présentés dans le tableau 2 .

Requête	Modèle bayésien simple			Notre modèle		
	P5	P10	Map	P5	P10	Map
59	0,0000	0,0000	0,0065	0,6000	0,5000	0,0395
62	0,2000	0,2000	0,1309	0,4000	0,3000	0,1552
70	0,0000	0,0000	0,0361	0,0000	0,0000	0,1073
71	0,2000	0,2000	0,0417	0,4000	0,2000	0,0431
74	0,2000	0,1000	0,0022	0,0000	0,0000	0,0012
76	0,0000	0,0000	0,0241	0,4000	0,2000	0,0909
77	0,0000	0,0000	0,0659	0,6000	0,6000	0,2802
78	0,2000	0,1000	0,0487	0,4000	0,2000	0,1308
80	0,0000	0,2000	0,0260	0,0000	0,0000	0,0207
83	0,4000	0,2000	0,0132	0,4000	0,4000	0,0226
85	0,0000	0,0000	0,0831	0,6000	0,5000	0,1332
87	0,0000	0,0000	0,0013	0,0000	0,0000	0,0042
91	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
92	0,0000	0,0000	0,0032	0,2000	0,2000	0,0311

Tableau 2 – Résultats expérimentaux

Les résultats préliminaires obtenus montrent globalement que notre modèle est à l'origine d'un accroissement significatif des précisions $P5$, $P10$ et MAP par rapport au modèle bayésien simple. Plus particulièrement, le MAP moyenne qui est un indicateur de référence est augmentée de 0,0345 à 0,0757 soit un taux d'accroissement de 119%.

Les taux d'accroissement sont cependant variables, dépendant généralement de la difficulté de la requête. On note, en outre, que les requêtes du domaine *Political* (requêtes 74 et 80) ne sont pas améliorées ; ceci peut être dû à l'insuffisance des données d'apprentissage pour la simulation des centres d'intérêt. A juste titre, l'un de nos objectifs ultérieurs est d'étendre le processus d'apprentissage à plusieurs centres d'intérêt pour un même utilisateur et par conséquent à un nombre plus élevé de requêtes, puis d'en évaluer l'impact sur la taille des données d'apprentissage d'une part, et des performances de recherche d'autre part.

5. Conclusion

Notre contribution, à travers ce papier, porte sur deux volets. Le premier porte sur la proposition des bases d'un modèle d'accès personnalisé à l'information. Le modèle proposé est basé sur une extension des réseaux bayésiens en l'occurrence les diagrammes d'influence. La composante qualitative du diagramme traduit la structure des centres d'intérêt de l'utilisateur, des documents et index de la collection. La

composante quantitative, traduit la mutuelle influence existante entre un besoin en information exprimé par une requête et un contexte associé, dans une situation liée à la prise de décision quant à la pertinence d'un document. La valeur de pertinence d'un document est exprimée à l'aide de l'utilité de la décision liée à sa présentation.

Le second volet de notre contribution consiste en la définition d'un cadre d'évaluation approprié pour l'accès personnalisé à l'information. Nous avons alors appliqué ce cadre pour évaluer précisément notre modèle et avons présenté les résultats expérimentaux obtenus. Le cadre proposé a l'intérêt de réutiliser les ressources de la campagne d'évaluation standard TREC. Les résultats obtenus sont encourageants et ouvrent des perspectives intéressantes. A court terme, notre objectif est d'explorer l'intégration de différents centres d'intérêt pour un même utilisateur et d'en évaluer l'impact sur les performances de recherche.

6. Bibliographie

- [ALL 02] ALLAN J., AL, « Challenges in information retrieval and langage modelling », *Workshop held at the center for intelligent information retrieval*, Septembre 2002.
- [AMA 99] AMATO G., STARACCIA U., « User profile modelling and applications to digital libraries », *Proceedings of the 3rd European Conference on Research and avanced technology for digital libraries*, 1999, p. 184-187.
- [BOU 05] BOUZEGHOUB M., KOSTADINOV D., « Personnalisation de l'information : Aperçu de l'état de l'art et définition d'un modèle flexible de définition de profils », *Actes de la 2nde Conférence en Recherche d'Information et Applications CORIA*, 2005, p. 201-218.
- [BUD 00] BUDZIK J., HAMMOND K., « User interactions with every applications as context for just-in-time information access », *Proceedings of the 5th international conference on intelligent user interfaces*, Mars 2000, p. 44-51.
- [CHA 04] CHALLAM V., « Contextual information retrieval using ontology based user profiles », *Master of science in computer science*, Jawaharlal Nehru Technological University, 2004.
- [CHE 98] CHEN L.AND SYCARA K., « WebMate : A personal agent for browsing and searching », *Proceedings of the 2nd international conference on autonomous agents and multi agent systems, Minneapolis*, 1998, p. 10-13.
- [COO 01] COOL C., « The concept of situation in information science », *Annual review of information science and technology*, vol. 35, 2001, p. 5-42.
- [COO 02] COOL C., SPINK A., « Issues of context in information retrieval : an introduction to the special issue », *In Journal of Information Processsing and Management (IPM)*, vol. 38, n° 55, 2002, p. 605-611.
- [FAN 04] FAN W., D M. G., PATHAK P., « Discovery of Context-Specific Ranking Functions for Effective Information Retrieval Using Genetic Programming », *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, n° 4, 2004, p. 523-527, IEEE Educational Activities Department.
- [GAU 03] GAUCH S., CHAFFE J., PRETSCHNER P., « Ontology based user profiles for search and browsing », vol. Special issue on user modelling for Web and hypermedia information retrieval, 2003.

- [JEH 03] JEH G., WIDOM J., « Scaling personalized web search », *WWW '03 : Proceedings of the 12th international conference on World Wide Web*, New York, NY, USA, 2003, ACM Press, p. 271–279.
- [JEN 01] JENSEN F., *Bayesian networks and decision graphs*, Springer, 2001.
- [KIM 03] KIM R., CHAN K., « Learning implicit user interest hierarchy for context in personalization », *Proceedings of the 8th international conference on Intelligent user interfaces*, ACM Press, 2003, p. 101 - 108.
- [KOB 00] KOBAYASHI M., TAKEDA K., « Information retrieval on the web », *ACM Computing Surveys*, vol. 32, n° 2, 2000, p. 144–173.
- [LIN 05] LIN C., XUE G., ZENG H., YU Y., « Using probabilistic latent semantic analysis for personalised Web search », *Proceedings of the APWeb Conference*, 2005, p. 707-711.
- [LIU 04] LIU F., YU C., « Personalized Web search for improving retrieval effectiveness », *IEEE Transactions on knowledge Data Engineering*, vol. 16, 2004, p. 28-40.
- [MIT 97] MITCHELL T. M., *Machine Learning*, McGraw-Hill Higher Education, 1997.
- [PAZ 96] PAZANNI M., MURAMATSU J., BILLSUS D., « Syskill and Webert : Identifying interesting Web sites », *Proceedings of the 13th National Conference on Artificial intelligence*, 1996, p. 54-61.
- [PEA 88] PEARL J., *Probabilistic reasoning in intelligent systems : networks of plausible inference*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.
- [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.
- [ROB 97] ROBERTSON S. E., WALKER S., « On relevance weights with little relevance information », *SIGIR '97 : Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1997, ACM Press, p. 16–24.
- [ROC 71] ROCCHIO J., « Relevance feedback in Information retrieval », Englewood Cliffs, 1971, Prentice Hall.
- [SHA 88] SHACHTER R., « Probabilistic inference and influence diagrams », vol. 36, In *Operating Research*, 1988, p. 589–604.
- [SHE 05] SHEN X., TAN B., ZHAI C., « Context-sensitive information retrieval using implicit feedback », 2005.
- [SPE 04] SPERETTA S., GAUCH S., « Personalizing search based user search histories », *In Proceedings of the 13th International Conference on Information Knowledge and Management, CIKM*, 2004, p. 238–239.
- [TAM 06a] TAMINE L., BOUGHANEM M., CHRISMENT C., « Accès personnalisé à l'information : vers un modèle basé sur les diagrammes d'influence », *Information - Interaction - Intelligence (I3)*, vol. 6, n° 1, 2006, p. 69–90, Cépaduès Editions.
- [TAM 06b] TAMINE L., BOUGHANEM M., ZEMIRLI W. N., « Inferring the user's interests using the search history », SCHAAF M., ALTHOFF K.-D., Eds., *Workshop on information retrieval, Learning, Knowledge and Adaptability (LWA)*, Hildesheim, Germany, 09/11/06-11/11/06, 2006, p. 108–110.
- [TUR 90] TURTLE H., CROFT W. B., « Inference networks for document retrieval », *SIGIR 90 : Proceedings of the 13th annual international ACM SIGIR conference on Research and*

development in information retrieval, New York, NY, USA, 1990, ACM Press, p. 1–24.

[VAS 02] VASSILIOU C. STAMOULIS D. M. D., « The process of personalizing web content : techniques, workflow and evaluation. », *Proceedings of the SSGRR International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*, 2002.

[ZEM 05] ZEMIRLI W. N., TAMINE L., BOUGHANEM M., « Accès personnalisé à l'information : vers la définition d'un profil utilisateur multidimensionnel », *International Symposium On Programming Systems (ISPS)*, 2005.