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

Farida Achemoukh et Rachid Ahmed-Ouamer

*Laboratoire LARI, Université Mouloud Mammeri
15000 Tizi-Ouzou, Algérie
{achemoukh.farida, ahm_r}@yahoo.fr*

RÉSUMÉ. Pour satisfaire les besoins d'informations d'un utilisateur et la personnalisation de l'information, l'emploi du profil utilisateur a été adopté. Ceci engendre le problème d'évolution de ce profil au cours du temps. Dans ce papier est proposé un modèle d'évolution du profil utilisateur en se basant sur les réseaux Bayésiens temporels. Ce modèle est le résultat d'une répétition dans le temps, sur une séquence de longueur T , d'un réseau de base représentant une activité de recherche caractérisée par la requête soumise à un instant t et le centre d'intérêt correspondant et les termes d'indexation. Les différents réseaux résultants sont alors reliés par des liens causaux d'un pas de temps à l'autre ce qui permet la dynamique d'évolution de profil utilisateur repérée par la modification de centres d'intérêt tout au long des différentes requêtes utilisateur soumises sur une séquence de temps prédéfinie.

ABSTRACT. To meet user needs of information and the personalization of the information the use of the user profile has been adopted. This generates the problem of evolution of this profile over time. In this paper is proposed a model of evolution of the user profile based on the temporal Bayesian networks. This model is the result of a repetition over time on a sequence of T length, of a basic network representing a search activity characterized by the request submitted at t time and the user interest corresponding with the relevant documents and the terms of the index. The resulting networks are connected by causal links between each two successive steps of time, which allows the dynamic evolution of user profile indicated by the change of the user interests throughout the various submissions of user queries at the sequence of time

MOTS-CLÉS : Réseaux bayésiens temporels, profil utilisateur, recherche d'information personnalisée

KEYWORDS: Temporal Bayesian networks, user profile, personalization of information retrieval

1. Introduction

Avec l'explosion de l'information sur le web, la satisfaction des besoins en informations d'un utilisateur demeure un but très important à atteindre pour les systèmes de recherche d'information actuels. Dans ce contexte, l'emploi de profil utilisateur ou groupe d'utilisateurs dans le processus de recherche d'information a été adopté, ce que l'on nomme la personnalisation de l'information [Kostadinov 2007] [Chevalier et al 2008].

L'objectif de la personnalisation de l'information consiste en la manière de modéliser l'utilisateur et l'intégration du profil résultant dans le processus d'accès à l'information. Différents travaux portent sur les approches et techniques de définition du profil utilisateur [Gonzalez et al 2002]. Cependant, la personnalisation de l'information engendre le problème de l'évolution de profil d'utilisateur au cours du temps. Dans la plupart des systèmes d'accès personnalisé, l'évolution du profil est exprimée par l'ajout de nouvelles informations [Berisha Bohe et al 2007]. Elle consiste à adapter le contenu du profil aux variations des besoins utilisateur en information exprimés généralement par les changements de ses centres d'intérêt [Zemirli 2008].

Dans ce contexte, nous proposons un modèle d'évolution du profil utilisateur basé sur les réseaux bayésiens classiques pour l'inférence d'un centre d'intérêt lors d'une activité de recherche. L'évolution de ce profil sur une séquence prédéfinie du temps, s'appuie sur la dynamique de ce réseau, ce qui est appelé, un réseau bayésien temporel. Le choix de ce type de réseau est dû au cadre théorique qu'il offre. Les réseaux bayésiens tirent leur puissance de leur aspect graphique pour représenter les relations de causalité entre les variables du domaine spécifié et de l'aspect mathématique qui fait d'eux une approche quantitative [Pearl 1988], [Wong et al 1995], ce qui les rend adaptés pour inférer et faire évoluer les centres d'intérêt de l'utilisateur, à partir de ses interactions avec un système de recherche d'information.

Ce papier s'intéresse à la problématique d'intégration et d'évolution du profil utilisateur dans le processus de recherche d'information. Dans la section 2 sont abordées la modélisation utilisateur et l'évolution de son profil. La section 3 présente une définition des réseaux bayésiens classiques et temporels. La section 4 décrit le modèle d'intégration et d'évolution de profil utilisateur selon l'inférence bayésienne. La dernière section fait la synthèse de cette étude.

2. Modélisation utilisateur et évolution de profil

Sans modèle utilisateur, un système de recherche d'information se comportera exactement de la même manière avec tous les utilisateurs, mais ces derniers sont différents : ils ont des connaissances différentes, des préférences différentes, ainsi que des buts et des centres d'intérêt. Toutes ces variations, peuvent se regrouper sous le terme de profil utilisateur.

Différentes définitions du profil utilisateur ont été proposées ; comme par exemple : un profil utilisateur est défini comme une structure qui permet de

modéliser et de stocker les centres d'intérêt, les préférences et les besoins en information de l'utilisateur [Bouzeghoud et al. 2005]. L'objectif de la personnalisation de l'information consiste alors en la manière de modéliser l'utilisateur et l'intégration du profil résultant dans le processus d'accès à l'information.

La modélisation de l'utilisateur est un processus à différentes étapes à savoir, la représentation du profil utilisateur qui peut être :

- Vectorielle : le profil est constitué d'un ou de plusieurs vecteurs définis dans un espace de termes d'indexation [Gowan 2003].
- Hiérarchique : les caractéristiques d'un utilisateur sont organisées dans une structure hiérarchique de concepts représentant les domaines d'intérêt [Gonzalez et al 2002],
- Multidimensionnelle : le profil est représenté par un modèle structuré de dimensions prédéfinies (données personnelle, domaine d'intérêt, préférences de livraison, etc.) [Kostadinov 2003].

Une fois effectué le choix de la représentation, l'étape de construction du profil consiste en la collecte des informations qui le représentent et cela d'une manière explicite, en se basant sur les informations fournies directement par l'utilisateur [Maghoul et al 2005], par exemple, lorsque l'utilisateur visualise un document, il indique son opinion sur le degré de pertinence du document par rapport à sa requête, ou implicitement, à partir des documents consultés et de comportement de l'utilisateur (temps de lecture d'un document, sauvegarde, impression, etc.) [Gauch et al 2003].

La personnalisation d'accès à l'information consiste alors à intégrer le profil utilisateur dans au moins l'une des phases du processus de recherche. Son intégration dans la phase de la reformulation de requête consiste à augmenter la requête initiale par des termes issus du profil utilisateur [Sieg et al 2004] [Koutrika et al 2005]. Dans la phase d'appariement requête-document le profil est intégré dans le calcul de pertinence du document [Walker et al 1995]. Pour la phase d'ordonnancement des résultats de recherche cela consiste à associer le profil utilisateur dans les résultats finaux de la recherche. Le score final est la combinaison de score de similarité entre le document et le profil avec le score d'appariement initial du document [Liu et al 2004] [Gowan 2003].

La personnalisation de l'information a engendré le problème de l'évolution de profil utilisateur au cours du temps. Dans la plupart des systèmes d'accès personnalisé, l'évolution de profil consiste à adapter son contenu aux variations des besoins utilisateur en information :

- Dans le cas de systèmes à représentation ensembliste du profil, son évolution est exprimée par l'ajout de nouveaux vecteurs de termes extraits des documents correspondants aux centres d'intérêt détectés de l'utilisateur [Billsus et al 1999].
- Dans le cas d'une représentation hiérarchique, le système fait évoluer le contenu du profil en associant de nouveaux documents collectés aux classes similaires appropriées. L'adaptation de la structure du profil aux nouvelles classes s'effectue en mettant à jour les relations entre ces classes [Moukas 1997]. Elle traduit dans [Chen et al 2002] la notion de cycle de vie d'un centre d'intérêt, où on associe aux classes

du profil utilisateur une valeur d'énergie, traduisant le degré d'importance d'un centre d'intérêt par rapport à un autre. Basée sur les valeurs d'énergie des catégories, la structure de profil peut être modulée pendant que les centres d'intérêt des utilisateurs changent.

- Dans [Zemirli 2008] l'évolution de profil consiste en une évolution de la dimension centre d'intérêt basée sur une mesure de corrélation des rangs qui évalue le degré de changement entre contextes d'usage associés à des périodes successives.

Avant d'aborder le modèle d'intégration et d'évolution de profil utilisateur que nous avons développé, il est utile d'introduire la notion de réseaux bayésiens.

3. Réseaux bayésiens classiques et temporels

Un réseau bayésien classique est un graphe acyclique orienté $G = (V, E)$ où V comprend des nœuds qui représentent des variables aléatoires, et E un ensemble d'arcs qui représentent des liens de causalités entre nœuds parents et nœuds fils [Turtle et al 1991]. Pour chaque variable V_i une distribution de probabilité est associée ; elle spécifie les probabilités de ses états conditionnellement aux états des variables qui l'influencent (ses parents).

L'inconvénient des réseaux bayésiens statiques est qu'ils ne prennent pas en compte la contrainte de temps ce qui peut être suffisant pour de nombreuses applications, mais peut être un handicap pour d'autres.

Les réseaux bayésiens temporels sont une extension des réseaux bayésiens classiques, permettant de représenter l'évolution temporelle de variables aléatoires. Ces modèles ont été introduits dans [Dean et al 1989] [Nicholson 1992]. Ils ont comme objectif la modélisation de distributions de probabilités d'une suite de variables aléatoires, évoluant sur une séquence de temps de longueur $T \in \mathbb{N}^*$:

$$(V_{i,t})_{1 \leq t \leq T} = (V_{1,t}, \dots, V_{N,t})$$

Dans ce papier nous utilisons pour notre modélisation un type particulier de réseaux temporels, à savoir les réseaux bayésiens à 2 tranches de temps [Kevin 2002]. Ils sont définis par un couple de réseaux bayésiens (B_1, B_{\rightarrow}) .

B_1 correspond à un réseau bayésien classique (statique), il a pour but de décrire la loi initiale des variables. Chaque variable est indexée par deux indices i, t : $(V_{i,t})$ tel que, i représente l'indice de la variable et t son indice temporel, avec $1 \leq t \leq T$:

$$P(V_{i,t}) = \prod_{i=1}^N P(V_{i,t} / \text{parents}(V_{i,t})).$$

D'autre part, B_{\rightarrow} est un réseau décrivant la loi de transition d'un état courant à l'instant $(t+1)$ conditionnellement à un état précédent correspondant à l'instant (t) :

$$P(V_{i,t})_{1 \leq t \leq T} = P(V_{i,t}) \prod_{t=2}^T P(V_{i,t} / V_{i,t-1}) =$$

$$\prod_{i=1}^N P(V_{i,(t=1)}/parents V_{i,(t=1)}) \prod_{t=2}^T \prod_{i=1}^N P(V_{i,(t)}/parents V_{i,(t-1)})$$

4. Modélisation d'évolution de profil utilisateur proposée

Pour la modélisation d'évolution d'un profil utilisateur, nous utilisons une technique qui s'appuie sur les réseaux bayésiens temporels, précisément les réseaux bayésiens à 2 tranches de temps.

Comme nous l'avons vu auparavant, plusieurs dimensions permettent la définition de profil utilisateur. Nous nous intéressons dans cette étude à la dimension domaine d'intérêt qui exprime les caractéristiques générales d'information que l'utilisateur souhaite obtenir. Cette dimension couvre les données représentant l'intérêt de l'utilisateur que l'on décrit par les documents choisis de l'utilisateur. L'évolution d'un profil utilisateur désigne alors son adaptation à la variation de ses centres d'intérêt au cours de diverses sessions de recherche définies lors des différentes activités de recherche. Les notions de profil utilisateur, d'activité de recherche et de session de recherche sont précisées ci-après.

-- *Profil utilisateur* : le profil utilisateur correspond ici à l'ensemble de ses centres d'intérêt. La détermination de la librairie des centres d'intérêt suppose M requêtes utilisateur pour lesquelles sont définis les documents pertinents correspondants. A chaque requête, est alors associé un centre d'intérêt ck représenté par un vecteur de termes pondérés ; $ck = \{(t_1, w_1), (t_2, w_2) \dots (t_b, w_b)\}$, selon la formule BM25 [Robertson et al. 1997] définit comme suit :

$$w(t_i, ck) = \log \frac{(r + 0.5)}{(n - r + 0.5)} / \frac{(R - r + 0.5)}{(N - R - n + r + 0.5)}$$

N est le nombre total de documents de la collection ; n le nombre de documents de la collection contenant le terme t_i ; R le nombre de documents pertinents associé à une requête utilisateur et r le nombre de documents pertinents contenant le terme t_i .

-- *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 de plusieurs activités de recherche correspondant au même centre d'intérêt.

4.1. Modélisation d'une activité de recherche

La structure du modèle consiste en une répétition dans le temps, sur une séquence de longueur T , d'un réseau bayésien représenté par la figure 1 suivante. Celle-ci représente une activité de recherche qui inclue les nœuds de la collection des

documents d_j , le nœud requête Q , les nœuds termes d'indexation t_i et les nœuds de la librairie des centres d'intérêt c_k .

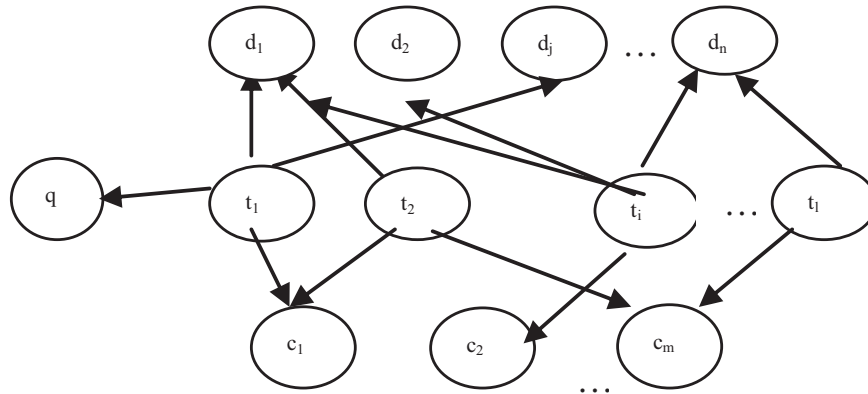


Figure 1. Réseau bayésien représentant une activité de recherche

Chaque nœud document d_j de l'ensemble des documents de la collection $D = \{d_1, d_2, \dots, d_n\}$, représente une variable aléatoire binaire dont le domaine est $\text{Dom}(d_j) = \{d_j, \bar{d}_j\}$, où d_j et \bar{d}_j désignent respectivement que « le document d_j est instancié » et que « le document d_j n'est pas instancié ». Un seul document est instancié positivement à la fois.

Chaque nœud centre d'intérêt c_k de l'ensemble des centres d'intérêt $C = \{c_1, c_2, \dots, c_m\}$, représente une variable aléatoire binaire dont le domaine est $\text{Dom}(c_k) = \{c_k, \bar{c}_k\}$, où c_k et \bar{c}_k désignent respectivement que « le centre d'intérêt c_k est instancié » et que « le centre d'intérêt c_k n'est pas instancié ». Un seul centre d'intérêt est tiré à la fois du profil de l'utilisateur.

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

Chaque nœud terme t_i de l'ensemble des termes indexant les documents de la collection $T = \{t_1, t_2, \dots, t_l\}$, représente une variable aléatoire binaire dont le domaine est $\text{Dom}(t_i) = \{t_i, \bar{t}_i\}$, où t_i désigne que le terme est présent dans un document d_j ou la requête q ou un centre d'intérêt c_k et \bar{t}_i désigne que le terme n'est pas présent dans un document d_j , ou dans la requête q , ou dans un centre d'intérêt c_k .

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

4.2 Calcul de pertinence

Le calcul de pertinence revient ici à instancier chaque document de la collection et un à un chacun des centres d'intérêt utilisateur. La propagation d'information dans le modèle est déclenchée par la réception de la requête utilisateur et consiste à

calculer la probabilité que la requête soit rencontrée étant donné le document d_j et le centre d'intérêt c_k instanciés. Pour chaque document, ce processus est réitéré pour tous les centres d'intérêt de l'ensemble C .

Ainsi, les probabilités pour chaque nœud du modèle sont évaluées.

Initialement des probabilités a priori sont associées à tous les nœuds racines qui représentent ici les termes d'indexation. Elles indiquent la probabilité d'importance et de non importance d'un terme t_i .

La probabilité a priori pour chaque nœud terme est alors : $P(t_i) = 0,5$ et $P(\bar{t}_i) = 0,5$

Une probabilité conditionnelle est ensuite associée à chaque nœud document, centre d'intérêt et requête, et elle représente la probabilité de contenir un terme donné.

-- Pour estimer la probabilité conditionnelle d'un document d_j , $P(d_j | \text{parents}(d_j))$, nous utilisons une fonction de probabilité définie comme suit:

$$P(d_j | \text{parents}(d_j)) = \sum_{t_i \in \text{parents}(d_j)} w(t_i, d_j)$$

Où
$$w(t_i, d_j) = \frac{(tf_{ij} \cdot idf)^2}{\sqrt{\sum_{t_i \in \text{parents}(d_j)} (tf_{ij} \cdot idf)^2}}$$

$0 \leq w(t_i, d_j) \leq 1$. Représente le poids d'un terme t_i dans un document d_j ; estimé selon une variante de pondération $tf \cdot idf$ qui repose sur la loi de Zipf [Salton et al. 1988].

$$tf \cdot idf = \frac{tf_{ij}}{\max_{t_k \in d_j} tf_{kj}} \times \frac{\log(N/n_i)}{\log(N)}$$

tf_{ij} est la fréquence d'apparition d'un terme t_i dans un document d_j , n_i est le nombre de documents contenant le terme t_i , N est le nombre total de documents de la collection et $\max_{t_k \in d_j} tf_{kj}$ est la plus grande valeur de tf dans un document d_j ;

A partir de la topologie du modèle proposé; les concepts requête, document et centre d'intérêt sont indépendants (d-séparés par les nœuds termes). La probabilité $P(q, d_j, c_k)$ est alors donnée comme suit :

$$P(q, d_j, c_k) = \sum_{u \in U} P(q/u) \times P(d_j/u) \times P(c_k/u) \times P(u) \quad [1]$$

U est l'ensemble des u configurations possibles de l termes sur T . $P(u)$ est la probabilité a priori associée à une configuration u de termes.

En appliquant la forme canonique ET-combinaison définie dans [Turtle and al. 1991] nous simplifions l'espace de termes, à la configuration u couverte par la

requête q , cela donne : $P(q/u)=1$ si $q=u$ (on prend seule la configuration u des termes qui composent la requête q). Ce qui réduit la formule [1] en :

$$P(q, d_j, c_k) = P(d_j / q) \times P(c_k / q) \times P(q) \quad [2]$$

Comme le modèle de réseau bayésien généralise tous les modèles classique ; les probabilités $P(d_j, /q)$ et $P(c_k/q)$, correspondent respectivement à la mesure de cosinus vectorielle entre le document et la requête d'un côté et entre le centre d'intérêt et la requête de l'autre.

Dans le cas général la probabilité $P(q/d_j, c_k)$ est exprimée par la formule suivante :

$$P(q / d_j, c_k) = \frac{P(q, d_j, c_k)}{P(d_j, c_k)} \quad [3]$$

D'après le modèle proposé et pour l'estimation de $P(d_j, c_k)$, les nœuds documents et centres d'intérêt seront indépendants après l'instanciation des nœuds termes (*d-séparation*). Ce qui donne :

$$P(d_j, c_k) = \sum_{u \in U} P(d_j / u) \times P(c_k / u) \times P(u) \quad [4]$$

De même, en appliquant la forme canonique ET-combinaison définit dans [Turtle and al. 1991] nous simplifions l'espace de termes, à la configuration u couverte par le centre d'intérêt c_k , cela donne : $P(c_k/u) = 1$ si $c_k = u$ (on prend seule la configuration u des termes qui composent le centre c_k).

$$P(d_j, c_k) = P(d_j / c_k) \times P(c_k) \quad [5]$$

$P(d_j/c_k)$ correspondent donc à la mesure de cosinus vectorielle entre le document d_j et le centre d'intérêt c_k .

La quantification de la formule [3] revient alors à estimer chacun de ses membres, ce qui donne après simplification :

$$P(q / d_j, c_k) = \frac{P(d_j / q) \times P(c_k / q)}{P(d_j / c_k)} \quad [6]$$

Les valeurs résultantes pour chaque requête q , correspondent aux probabilités de similitude entre chaque instanciation d'un document d_j et d'un centre d'intérêt c_k avec la requête q [Achemoukh et al 2011] ; représentant les éléments d'une matrice R de dimension (nombre de documents de la collection \times nombre de centres d'intérêt de la librairie).

4.3. Evolution de profil utilisateur

La répétition dans le temps sur une séquence de longueur T , du réseau bayésien représentant une activité de recherche (présentée dans la figure 1 précédente), engendre un ensemble de réseaux reliés par des liens causaux d'un pas de temps à l'autre ce qui permet la dynamique d'évolution de profil utilisateur repérée par le changement des centres d'intérêt tout au long des soumissions de différentes requêtes utilisateur.

Les composantes du modèle présenté dans la figure 2 suivante, consistent en une suite d'activités de recherche $Ac_{(t)}$ pour tout $1 \leq t \leq T$ sur une séquence de longueur T , caractérisée chacune par la variable $q_{i,(t)}$. Celle-ci indique la requête soumise à l'instant t et le centre d'intérêt $c_{k,(t)}$ correspondant, avec une suite $Du_{(t)}$ pour tout $1 \leq t \leq T$ représentant à chaque instant t la durée $d \in \{1 \dots D\}$ unités de temps, restantes avant le changement de l'activité de recherche $Ac_{(t)}$ vers $Ac_{(t+1)}$.

Enfin une troisième suite de variables aléatoires $S_{(t)}$ à valeurs dans $\{0,1\}$ pour tout $1 \leq t \leq T$ représentant les changements de sessions de recherche. Lorsque cette variable est activée à l'instant t ($S_{(t)} = 1$), le système change de session de recherche à cet instant précis t . En revanche, si $S_{(t)} = 0$, le système reste dans la même session de recherche même si l'activité de recherche change :

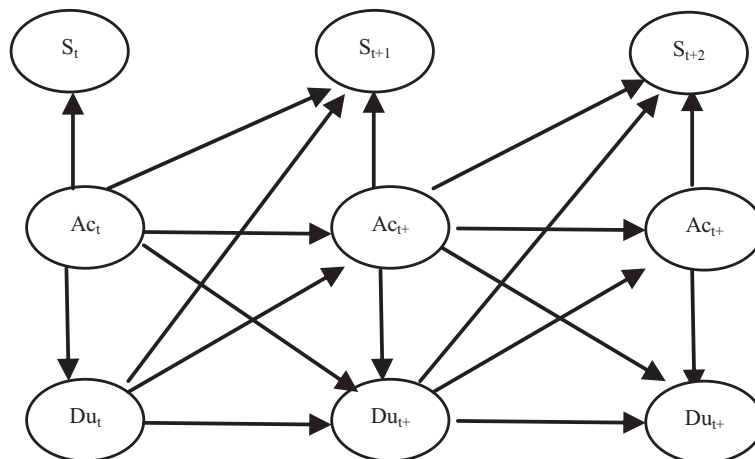


Figure 2. Modélisation de la dynamique d'activités de recherche

4.4. Probabilités d'inférences

Les probabilités d'inférences associées au modèle proposé sont les suivantes :

- Les probabilités de départ correspondant à la probabilité de l'activité de recherche initiale associant une requête initiale et le centre d'intérêt correspondant $P(Ac_{(t=1)}) = P(q_{i,(t=1)}, c_{k,(t=1)})$ et à la probabilité initiale de la durée $P(Du_{(t=1)})$, associée à l'activité de recherche initiale avec $P(S_{(t=1)})=1$ indiquant qu'initialement on démarre avec une nouvelle session de recherche ;

- Les probabilités de changement d'activités de recherche $P(Ac_{(t+1)}/Ac_{(t)})$;
- Les probabilités des durées des activités de recherche pour un instant $t > 1$;
- Les probabilités des changements de sessions de recherche.

4.4.1. *Activité de recherche*

Considérons N différentes activités de recherche ce qui permet un ensemble de N requêtes $Q = \{q_1, \dots, q_e, \dots, q_N\}$ soumises sur une séquence de longueur $t \in \{1, \dots, T\}$ et une librairie de M centres d'intérêt $C = (c_1, \dots, c_k, \dots, c_M)$ ($M=N$ si toutes les requêtes ont des centres d'intérêt différents).

Pour chaque réseau activité de recherche $Ac_t \in \{Ac_1, Ac_2, \dots, Ac_T\}$, nous nous intéressons aux nœuds dynamiques requêtes et centres d'intérêt qui changent au fil de la séquence du temps. Nous Définissons tout d'abord la probabilité associée à l'activité de recherche initiale $Ac_{(t=1)}$, associant la première requête soumise par l'utilisateur et le centre d'intérêt correspondant.

La probabilité $P(q_{e,t})$, désigne la probabilité que la requête q_e soit soumise à l'instant t . Elle est donnée par un vecteur de probabilités V_t de N éléments et cela pour chaque activité de recherche correspondante à un instant t .

Donc pour tout $q_e \in Q$, on a :

$$P(q_{e,t}) = v_{e,t} = \begin{cases} 1 & \text{si la requête est soumise} \\ 0 & \text{sinon} \end{cases}$$

4.4.1.1. *Probabilité initiale d'une activité de recherche*

Pour toute requête $q_e \in Q$ et tout $c_k \in C$, la probabilité initiale d'une activité de recherche $P(Ac_{(1)})$ est définie comme :

$$P(q_{e,(1)}, c_{k,(1)}) = P(q_{e,(1)}) P(c_{k,(1)}/q_{e,(1)})$$

Donc pour une e -ème requête soumise au temps initial $t=1$, la probabilité initiale d'une activité de recherche est donnée par un vecteur de M (nombre de centres d'intérêt) éléments situés à la e -ème ligne d'une matrice $X_{e,k}$ de N lignes (nombre de requêtes) et M colonnes (nombre de centres d'intérêt), désignant les probabilités $P(c_{k,(1)}/q_{e,(1)})$.

Pour toute requête q_i , on simplifie l'espace de termes, à la configuration u couverte par cette requête, cela donne :

$$P(c_{k,(t)}/q_{e,(t)}) = \frac{\sum_{i=1}^1 w(t_i, c_k) \times w(t_i, q_e)}{\sqrt{\sum_{i=1}^1 w(t_i, c_k)^2} \sqrt{\sum_{i=1}^1 w(t_i, q_e)^2}}$$

Avec $w(t_i, c_k)$ et $w(t_i, q_e)$ indiquant respectivement le poids d'un terme t_i dans un centre d'intérêt c_k et dans une requête q_e . Ce qui donne :

$$P(Ac_{(1)}) = P(e_{(1)}) P(c_{k,(1)} / q_{e,(1)}) = \sum_{q_e \in Q} v_{e,1} \times X_{e,k}$$

4.4.1.2. Probabilité de changement vers une nouvelle activité de recherche

Pour définir l'évolution du profil utilisateur repérée par le changement de centres d'intérêt tout au long des différentes activités de recherche, nous nous intéressons aux probabilités de transition des activités de recherche et cela pour tout $t > 1$, $P(Ac_{(t+1)} / Ac_{(t)})$, ce qui correspond d'après l'architecture du modèle proposé au calcul de :

$$P(c_{k,(t+1)} / c_{k,(t)}, Du_{(t)}=1, q_{e,(t+1)}) = P_{k,e,k}$$

Avec $c_{k,(t+1)}$ et $q_{e,(t+1)}$ sont le centre d'intérêt et la requête de l'activité de recherche courante, $c_{k,(t)}$ est le centre d'intérêt de l'activité de recherche précédente et $Du_{(t)}=1$ est l'indicateur de temps restant avant le changement d'activité de recherche (indique qu'il reste une unité de temps pour le changement d'activité vers la suivante).

$P_{k,e,k}$ est une matrice de $N \times M$ lignes et M colonnes. Elle exprime la probabilité de changement de centre d'intérêt au $c_{k,(t+1)}$ dans l'activité de recherche $Ac_{(t+1)}$, alors qu'il était au centre d'intérêt $c_{k,(t)}$ dans l'activité de recherche précédente $Ac_{(t)}$ et cela pour une requête courante $q_{e,(t+1)}$.

Donc pour toute q_e , soumise à listant $(t+1)$, P est une matrice carrée de M (nombres de centres d'intérêt), dont les éléments correspondent aux probabilités de changement de centres d'intérêt.

En revanche tant qu'il ya pas de changement d'activité de recherche, le modèle reste dans l'activité de l'instant précédente, ce qui donne le non changement de centre d'intérêt, d'où :

$$P(c_{k,(t+1)} / c_{k,(t)}, Du=d, q_{e,(t+1)}) = \begin{cases} 1 & \text{si } k = k \text{ et } d > 1 \\ 0 & \text{sinon} \end{cases}$$

4.4.2. Durée d'une activité de recherche

Il s'agit de définir la probabilité de rester dans l'activité pendant d unités de temps. On suppose que la variable $Du_{(t)}$ est à valeur $d \in \{1, \dots, D\}$ unités de temps, en choisissant une valeur de D suffisamment grande pour une activité de recherche.

4.4.2.1. Durée de l'activité de recherche initiale

Pour une activité de recherche initiale caractérisée par une requête $q_{e,(1)}$ et un centre d'intérêt $c_{k,(1)}$, on a :

$$P(Du_{(1)}=d / c_{k,(1)}, q_{e,(1)}) = \sum_{q \in Q} v_{e,1} \times L_{e,k,d}$$

Où $L_{e,k,d}$ est une matrice de $N \times M$ lignes et D colonnes. Elle donne la probabilité de rester d unités de temps dans l'activité initiale regroupant le centre d'intérêt initial $c_{k,(1)}$ et la requête $q_{e,(1)}$.

4.4.2.2. Durée de l'activité de recherche prochaine

Les probabilités associées aux $Du_{(t)}$; pour tout $t > 1$; doivent maintenir la durée restante avant la prochaine activité de recherche. En effet tant qu'il reste du temps avant un changement d'une activité de recherche à l'instant précédent, le système reste dans la même activité précédente, donc :

$$P(Du_{(t+1)}=d / c_{k,(t)}, Du_{(t)}=d', c_{k,(t+1)}, q_{e,(t+1)}) = \begin{cases} 1 \text{ si } d=d'-1 \text{ avec } d' > 1 \text{ et } k=k' \\ 0 \text{ sinon} \end{cases}$$

Avec $Du_{(t+1)}=d$ est la durée restante courante et $Du_{(t)}=d'$ est la durée restante précédente.

D'autre part, lorsque le temps restant dans l'activité précédente $Ac_{(t)}$ est écoulé, l'activité de recherche change à l'instant $(t+1)$ et une nouvelle durée $d \in D$ lui est attribuée selon :

$$P(Du_{(t+1)}=d / c_{k,(t)}, Du_{(t)}=1, c_{k,(t+1)}, q_{e,(t+1)}) = L_{k,e,k,d}$$

Où, $L_{k,e,k,d}$ est une matrice de $M^2 \times N$ lignes et D colonnes dont chaque élément désigne la probabilité de rester d unités de temps dans l'activité courante caractérisée par le centre d'intérêt $c_{k,(t+1)}$ et la requête $q_{e,(t+1)}$ alors qu'on était dans l'activité précédente caractérisée par le centre d'intérêt $c_{k,(t)}$.

4.4.3. Probabilité de changement de session de recherche

Le rôle de la variable binaire $S_{(t)}$ consiste à contrôler explicitement les changements de sessions de recherche durant la séquence de temps T . Autrement dit, quand $S_{(t)} = 1$, un changement de session est déclenché à l'instant même t . En revanche, tant que $S_{(t)} = 0$, la session de recherche est toujours la même que sa précédente. Un changement se produit uniquement si la durée de l'activité précédente est écoulée. Les probabilités $S_{(t)}$ pour chaque session de recherche correspondante à un instant t , tel que $1 \leq t \leq T$ sont alors regroupées dans un vecteur de probabilité W_t . Sachant que $P(S_{(t=1)})=1$; pour tout $t > 1$, on a :

$$P(S_{(t+1)}/c_{k,(t)}, Du_{(t)}=1, c_{k,(t+1)}) = w_t = \begin{cases} 1 & \text{si } c_{k,(t+1)} < > c_{k,(t)} \text{ et } d=1 \\ 0 & \text{sinon} \end{cases}$$

Le vecteur W_t nous permet ainsi de repérer sur combien de pas de temps s'étend une session de recherche regroupant les activités correspondantes au même centre d'intérêt.

5. Conclusion

Dans ce papier est décrit notre modèle d'évolution du profil utilisateur dans le processus de recherche d'information personnalisée. La structure du modèle proposé consiste en une répétition dans le temps, sur une séquence de longueur T , d'un réseau bayésien qui représente une activité de recherche. Elle inclue les nœuds documents, le nœud requête et les nœuds termes d'indexation avec les nœuds centres d'intérêt, qui constitue la dimension principale dans la personnalisation de l'information. Les relations de dépendance sont orientées des termes, qui constituent l'espace d'indexation, vers les documents, requête et centres d'intérêts au sein du réseau de base. L'ensemble de réseaux résultant sont reliés par des liens causaux d'un pas de temps à l'autre ce qui permet la dynamique d'évolution du profil utilisateur repérée par le changement des centres d'intérêts tout au long des différentes soumissions de requêtes utilisateur. Au cours du processus de recherche d'information, les probabilités associées à l'évolution du profil utilisateur sont explicitées. La prochaine étape consiste à mettre en œuvre ce modèle, le tester et le comparer aux autres modèles de recherche d'information personnalisée.

6. Bibliographie

- Achemoukh F, and Ahmed-Ouamer R., «Modèle de Recherche d'Information Personnalisée Basé sur les Réseaux Bayésiens », *1^{ère} Edition des Rencontres sur la Recherche en Informatique R²I 2011*. Tizi-Ouzou, Algérie, pp. 23-30, 12-14 juin 2011.
- Berisha-Bohe S., and Rumpler B., « Modèle évolutif d'un profil utilisateur », Quatrième conférence francophone en Recherche d'Information et Applications. École Nationale Supérieure des Mines de Saint-Étienne, 28 - 30 mars 2007.
- Billsus D., and Pazzani M., «A hybrid user model for news stories classification », *In Proceedings of the seventh International Conference on User Modelins*, Banff, Canada, pp. 99-108, , 1999.
- Bouzeghoub M., and Kostadinov D., «Personnalisation de l'information : Aperçu de l'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) Grenoble, France*, pp. 201–218, 2005.

- Chen C., Chen M., and Sun Y., «A self-adaptive personal view agent », *Journal of Intelligent Information Systems*, pp.173–194, Mars 2002.
- Chevalier M., Julien C., and Soulé-Dupuy C., « Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling », IGI Global ,2008.
- Dean T., Kanazawa K., « A model for reasoning about persistence and causation », *Computational Intelligence*, pp.142–150, 1989.
- Gauch S., Chaffee J., and Pretschner A., «Ontology-based personalized search and browsing. Web Intelligence and Agent Systems», 1(3-4) , pp. 219– 234, 2003
- Gonzalez G., Lopez B., and Rosa J., «The Emotional Factor: An Innovative Approach to User Modelling for Recommender Systems» *Proceedings of Second International Conference Adaptive Hypermedia and Adaptive Web-Based Systems AH*, 2002.
- Gowan J., «A multiple model approach to personalised information access», Master thesis in computer science, Faculty of science, Université de College Dublin, February 2003.
- Kevin P., Murphy,« Dynamic Bayesian Networks: Representation, Inference and Learning » PhD thesis, U.C. Berkeley, 2002.
- Kostadinov D., « La personnalisation de l'information : définition de modèle de profil utilisateur », Master's thesis, Université de Versailles, France, 2003.
- Kostadinov D., « Personnalisation de l'Information : une Approche de Gestion de Profils et de Reformulation de Requêtes», Thèse de Doctorat, L'Université de Versailles Saint-Quentin-En-Yvelines, 2007.
- Koutrika G., and Ioannidis Y., « A unified user profile framework for query disambiguation and personalization », *In Proceedings of Workshop on New Technologies for Personalized Information Access*, July 2005.
- Liu F., Yu C., and Meng W., « Personalized web search for improving retrieval effectiveness », *IEEE Transactions on Knowledge and Data Engineering*, 16(1) , pp. 28–40, 2004.
- Maghoul F., and Chang C., « contextual search at the point of inspiration », *In CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management*, New York, NY, USA, pp. 816–823, October 2005.
- Moukas A., « Information discovery and filtering using a multiagent evolving ecosystem », *Applied Artificial Intelligence*, pp. 437–457, 1997.
- Nicholson A., « The data association problem when monitoring robot vehicles using dynamic belief networks», *In ECAI 92: 10th European Conference on Artificial Intelligence*, 1992.
- Pearl J., «Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference » *Morgan Kaufmann Publishers*, 1988.
- Salton G., and Buckley C.,«Term-weighting approaches in automatic text retrieval», *Information Processing & Management (IPM)*, pp. 513–523, 1988.
- Sciarrone F., «Anatomy and empirical evaluation of an adaptive web-based information filtering system, », *User Modeling and User-Adapted Interaction*, 14(2-3), pp. 159–200, 2004.
- Sieg A., Mobasher B., Burke R., Prabu G., and Lytinen S., «Using concept hierarchies to enhance user queries in web-based information retrieval, », *In The IASTED International Conference on Artificial Intelligence and Applications* , 2004.
- Robertson S.E., and Walker S., « On relevance weights with little relevance information»,

In 20th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 16–24 , 1997.

Turtle H. R., and Croft, W. B., « Evaluation of an inference network-based retrieval model », *ACM Transaction on Information Systems* 9, 3, pp. 187–222, 1991.

Walker M., Robertson S., Sparck Jones, and al., « Okapi at trec-3. In Second Text Retrieval Conf (TREC-3) », 1995.

Wong S.K.M., and Yao Y.Y., « On Modeling Information Retrieval with Probabilistic Inference », *ACM Transactions on Information Systems (TOIS)* 13 (1), pp. 38–68, 1995.

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

Zuckerman I., and Albrecht D.W., « Predictive statistical models for user modeling, user modeling and user adapted interaction », 2001.