Atelier 2

La recherche d’information personnalisée sur le Web

Marie-Aude Aufaure, Hajer Baazaoui Zghal,

Yves Lechevallier et Christophe Claramunt

Atelier Recherche d'Information Personnalisée sur le WEB

Marie-Aude Aufaure**, Hajer Baazaoui*, Christophe Claramunt****
Yves Lechevallier***, Henda Ben Ghezala*

* Laboratoire Riadi-GDL, ENSI Tunis, Campus la Manouba, La Manouba, 2010, Tunisie
hajer.baazaouizghal@riadi.rnu.tn / henda.bg@cck.rnu.tn
**Ecole Centrale Paris, Laboratoire MAS Chaire SAP BusinessObjects Grande Voie des Vignes 92 295 Chatenay-Malabry
Marie-aude.aufaure@ecp.fr
*** INRIA-Rocquencourt, Domaine de Voluceau. 78 153 Le Chesnay Cedex, France
Yves.Lechevallier@inria.fr
**** Ecole Navale, Lanvoc-Poulmic, BP 600, Brest naval, France
christophe.claramunt@ecole-navale.fr

La croissance très importante des informations disponibles sur Internet nécessite des outils de recherche de plus en plus performants et permettant de discerner efficacement les informations pertinentes parmi des centaines voire des milliers de documents. Cette structuration des documents peut être notamment effectuée par une analyse de l'usage quand il s'agit d'informations spatiales et/ou temporelles. La personnalisation des résultats rendus à l'utilisateur est un des soucis des SRI et des moteurs de recherche. La modélisation utilisateur est incontournable quand il s'agit de fournir des résultats en tenant compte du profil et des préférences des utilisateurs, sachant que ces profils sont fonction de plusieurs composants tels que les types de requêtes spatiales et temporelles utilisées, la nature et les échelles de données manipulées, les choix de visualisation effectués, et les processus de navigation effectués sur le Web.

La présence d'information spatiale est devenue incontournable sur le Web, ce qui impose aux mécanismes de recherche et d'extraction d'information à prendre en compte cette dimension spatiale. Les données spatiales peuvent exister dans les documents Web eux même ou encore au niveau des requêtes des utilisateurs. L'objectif de cet atelier consiste à faire le point sur les recherches en cours dans ce domaine, et de présenter les résultats obtenus dans le cadre de projets de recherche menés autour de cette thématique. Il se veut un lieu de rencontre et d'échange et restera ouvert, permettant de présenter un travail abouti comme des réflexions sur le domaine, des travaux préliminaires ou encore des démonstrations de logiciels et prototypes développés. Les articles sélectionnés pour l'atelier couvrent une large variété de domaines et de problématiques et sont regroupés autour des thématiques suivantes :

Approches théoriques et ontologies

- F. Dammak et H. Kammoun, Apprentissage de fonctions d'ordonnancement d'alternatives avec approche actif
- K. Kamoun, M. Harzallah, P. Kuntz et S. Ben Yahian, Evolution d'ontologies : revue et critiques

- X. Aimé, F. Fürst, P. Kuntz et F. Trichet, Enrichissement sémantique de requêtes au moyen d'ontologies de domaine personnalisées

Analyse et personnalisation de données Web et spatiales

- M. Hadjouni, B. Hajer, M.-A. Aufaure et H. Ben Ghezala, Système de personnalisation Web basé sur la construction d'un réseau d'utilisateurs
- R. Haddad, H. Baazaoui, M.-A. Aufaure, C. Claramunt, Y. Lechevallier et H. Ben Ghezala, Proposition d'une architecture pour la personnalisation de l'information spatiale sur le Web
- M. Charrad, Y. Lechevallier, M. B. Ahmed et G. Saporta, WCUM pour l'analyse d'un site WebB
- H. Essid, I. R. Farah,V. Barra et H. Ben Ghezala, Analyse de la variation spatio-temporelle des objets dans les images satellitaires à base de modèle de Markov caché couplé

Apprentissage de Fonctions d'Ordonnancement d'Alternatives avec Approche Actif

Faïza Dammak *, Hager Kammoun **

MIRACL : Route de Tunis km10 - B.P. n° 242 - 3021 Sfax
miracl.isims@gmail.com
*faiza.dammak@gmail.com
**hager.kammoun@isd.rnu.tn

Résumé. Dans cet article, nous proposons une méthode d'apprentissage actif de fonctions d'ordonnancement d'alternatives. La motivation principale qui nous a conduit à l'élaboration de cette proposition est que pour trouver une fonction d'ordonnancement efficace il est nécessaire d'avoir une base d'apprentissage qui demande souvent l'étiquetage manuel des alternatives sur plusieurs exemples. Le but est de trouver les meilleures entrées à étiqueter pour réduire au maximum le nombre de données étiquetées. Pour une utilisation efficace de l'apprentissage actif sur des grandes collections, nous proposons d'apporter une modification de l'algorithme d'ordonnancement d'instances supervisé RankBoost. Nous présentons le modèle proposé et décrivons ses fonctionnalités ainsi que les choix d'implémentation.

1 Introduction

L'apprentissage de fonctions d'ordonnancement a récemment suscité de nombreux travaux (Amini, 2007). Il correspond à apprendre des fonctions qui ordonnent des éléments entre eux, plutôt que de les classer (Usunier, 2006).

Dans le cas de la recherche documentaire (RD), il s'agit d'apprendre une correspondance entre un ensemble de requêtes et un ensemble de documents étiquetés, capable d'ordonner ces derniers par rapport à une requête donnée. Le but est d'inférer un ordre partiel sur l'ensemble des documents de façon à ce que les documents pertinents par rapport à la requête soient mieux ordonnés que les documents non pertinents. Cet ordre sera généralisé pour déterminer un ordre total. Le principal inconvénient de ce paradigme est que l'étiquetage nécessite l'intervention d'un expert qui doit examiner manuellement une grande quantité de données. L'intérêt s'est donc porté à deux cadres d'ordonnancement. Le premier avec un apprentissage semi-supervisé (Amini, 2007) où il s'agit de prendre en compte les données étiquetées et non-étiquetées dans le processus d'apprentissage (Amini *et al,* 2008). Le deuxième avec un apprentissage actif. Ce dernier permet de sélectionner incrémentalement les exemples d'apprentissage les plus utiles à un instant donné pour effectuer l'apprentissage (Settles, 2009).

Notre approche consiste à appliquer une méthode d'apprentissage actif afin de considérablement réduire la base nécessaire à l'élaboration d'un modèle pertinent. En fait, un modèle interagit avec un oracle en lui présentant des données non-étiquetées à étiqueter et dont l'ajout (des données et de leur étiquette) permet de concevoir un modèle plus performant à l'étape suivante (Truong, 2009). Dans cet article, nous expliquons comment le principe de l'apprentissage actif est appliqué à l'ordonnancement d'alternatives dans un système de recherche d'information SRI.

Pour une utilisation efficace de l'apprentissage actif sur des grandes collections, nous adaptons une modification de l'algorithme d'ordonnancement supervisé RankBoost sur des

alternatives. Cet algorithme sera intégré dans un apprentissage actif à fin d'obtenir de bonnes performances avec le moins possible d'exemples d'apprentissage.

Dans la section 2, nous introduisons le principe d'ordonnancement ainsi que son intérêt dans la recherche d'information RI. Nous détaillons également le problème d'ordonnancement d'alternatives, l'algorithme RankBoost et le principe d'ordonnancement actif. Dans la section 3, nous proposons un algorithme d'ordonnancement inductif pour une RD qui induit une fonction d'ordonnancement à base d'apprentissage actif, en adaptant l'algorithme RankBoost sur des données partiellement étiquetées. Dans la section 4, nous présentons la collection de données LETOR (Liu *et al.*, 2007) ainsi que les mesures d'évaluation à utiliser pour évaluer l'algorithme proposé. En conclusion, nous présentons le travail à suivre.

2 Ordonnancement et Recherche d'Information

L'ordonnancement est une technique couramment utilisée en RI. Les systèmes de recherche documentaire (SRD) prennent en entrée une requête utilisateur et renvoient une liste ordonnée de documents, qui doit présenter en premier les documents pertinents pour la requête. La liste ordonnée est déterminée par une fonction d'ordonnancement. L'apprentissage de cette fonction peut être considéré comme l'apprentissage d'une fonction de score : une fonction à valeurs réelles, qui prend en entrée un élément d'un ensemble à ordonner. L'ordre est ensuite prédit en triant les éléments selon les scores croissants ou décroissants.

En apprentissage, l'ordonnancement désigne la capacité d'apprendre à créer des listes ordonnées d'objets pour une requête de l'utilisateur. Les fonctions en ordonnancement ne cherchent donc plus à prédire une sortie par rapport à une entrée, mais à comparer les entrées entre elles et à les retourner sous forme de liste ordonnée.

Nous avons fait le choix de couvrir juste une petite partie de la littérature en ordonnancement. En effet, ce terme peut cacher plusieurs tâches différentes (ordonnancement d'instances ou d'alternatives, régression ordinale, catégorisation, apprentissage de préférence, etc.). Il suscite de plus un intérêt de plus en plus croissant dans la communauté d'apprentissage (Truong 2009).

Dans la suite, nous présentons le principe d'ordonnancement d'alternatives, l'algorithme d'ordonnancement supervisé RankBoost ainsi que le principe d'ordonnancement actif.

2.1 Ordonnancement d'alternatives

L'ordonnancement d'alternatives est le type d'ordonnancement le plus répandu en RI, il englobe des tâches comme la RD ou le résumé automatique de textes (Amini, 2007) (Usunier, 2006). Il s'agit ici d'ordonner les éléments (appelés alternatives) d'une collection donnée par rapport à chaque observation en entrée de telle façon que l'ordre prédit reflète le critère de pertinence pour chacune des observations. En RD, une observation est une requête et le but est d'ordonner les documents (alternatives) d'une collection donnée de façon à ce que les documents pertinents soient ordonnés au-dessus des documents non-pertinents. Formellement, cela revient à déterminer un sous-ensemble de la collection initiale des documents en rapport avec une requête donnée. Si X l'ensemble des observations de la base d'apprentissage et Y l'ensemble des sorties réelles, pour chaque observation $x_i \in X$ est associée un vecteur désiré de taille variable m_i, $y_i = (y_i^1, ..., y_i^{m_i})$, avec $y_i^k \in IR$, l'ensemble de ses alternatives candidates. Ce vecteur définit l'ordre que l'on cherche à prédire sur les alter-

natives dans *Y*. La fonction de score *F* que doit prédire cet ordre prend en entrée un couple (x_i , *k*) où x_i est une observation et *k* un indice d'alternative candidate pour x_i , et renvoie un score réel reflétant la similarité entre une observation et une alternative i.e. $F : X \times Y \rightarrow IR$.

2.2 Algorithme RankBoost

L'algorithme RankBoost est l'un des premiers algorithmes d'ordonnancement d'instances supervisé, son but est d'estimer une fonction de score f_t pour chaque document (Freund et al, 2003). Il est conçu pour des problèmes d'ordonnancement. Comme tous les algorithmes de la famille Boosting, RankBoost apprend à chaque itération une fonction de base (fonction de score) { f_t } et les poids { α_t } avec t∈ {1..T}, et il construit itérativement une combinaison linéaire de ces fonctions, en adaptant à chaque itération une distribution de probabilité D_t sur l'ensemble des paires composées d'exemples (pertinent, non-pertinent) nommées paires critiques.

L'idée nouvelle introduite par RankBoost est qu'il propose un algorithme de sélection des fonctions de base efficace lorsque le nombre de ces fonctions est fini. Chaque fonction de base f_t est uniquement définie par une fonction caractéristique d'entrée φ_{jt} avec *jt*∈ {1...d} et un seuil θ_t :

$f_t(x) = \begin{cases} 1, \textit{if } \varphi_{jt}(x) > \theta_t \\ 0, \textit{si non} \end{cases}$ où $\varphi_{jt}(x)$ est la j^{ème} fonction caractéristique fournie pour x.

Notre intérêt est porté par la suite sur les méthodes d'apprentissage de fonctions d'ordonnancement actif, et qui ont pour objectif commun d'apprendre en présence d'une petite quantité de données étiquetées, simultanément avec une grande quantité de données non-étiquetées.

3 Apprentissage Actif

De nombreuses tâches reposent sur la disponibilité de données étiquetées. Celles-ci sont souvent trop coûteuses à produire, d'où l'idée d'étiqueter uniquement de petits ensembles de données et d'utiliser cette connaissance pour extraire l'information nécessaire présente sur des grosses masses de données non-étiquetées disponibles via le web par exemple. Ces méthodes sont devenues centrales pour de très nombreux domaines d'applications. Contrairement à l'apprentissage semi-supervisé qui développe des algorithmes pour exploiter conjointement des données étiquetées et non étiquetées. L'apprentissage actif propose à l'utilisateur des stratégies optimales pour étiqueter un ensemble réduit de données (Settles, 2009).

La notion d'apprentissage actif fait référence à une méthode où l'algorithme d'apprentissage sélectionne à chaque itération les instances à étiqueter et les inclut dans la série d'exemples d'apprentissage. Ceci permet très souvent de réduire de façon importante la quantité de données nécessaires pour apprendre un modèle d'apprentissage supervisé. Au lieu d'étiqueter des instances aléatoirement pour obtenir les données d'apprentissage, le module d'apprentissage actif suggère d'étiqueter les instances dont il espère que le bénéfice sur l'apprentissage soit maximum.

Nous trouvons dans la littérature deux déclinaisons de l'ordonnancement actif. La première consiste à sélectionner une entrée et à étiqueter l'ensemble des alternatives associées.

L'étiquetage consiste à résumer un document. La deuxième déclinaison cherche à sélectionner uniquement une paire entrée-alternative (Truong ,2009). L'utilisateur indique si l'alternative est pertinente ou non par rapport à cette entrée (Settles, 2009).

Nous nous intéressons à la deuxième approche (Cohn et al., 1994) qui permet de résoudre ce problème en se focalisant uniquement sur les données non-étiquetées disponibles. Ces exemples sont supposés provenir de la même source que les exemples étiquetés. Cette approche est illustrée par la figure ci-dessous et peut se résumer de la manière suivante : l'algorithme d'apprentissage dispose au départ d'un petit nombre d'exemples étiquetés, puis (a) demande une étiquette pour quelques exemples choisis judicieusement et (2) incorpore les informations obtenues pour choisir des nouveaux exemples (étape (1)) et ainsi de suite.

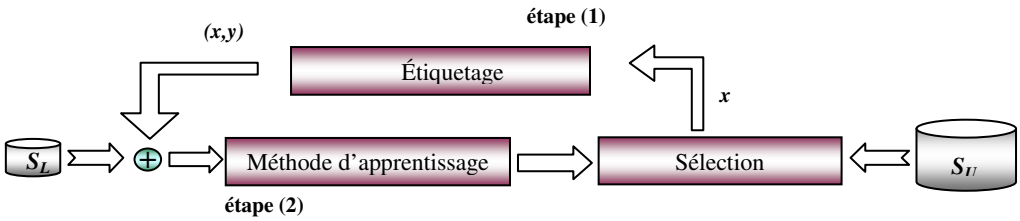


Fig. 1 – Illustration de l'approche sélective en apprentissage actif

L'approche sélective en apprentissage actif cherche à sélectionner uniquement une paire entrée-alternative. L'utilisateur indique si l'alternative est pertinente ou non par rapport à cette entrée.

Dans (Truong ,2009), Truong propose plusieurs stratégies de sélection d'instances pour l'ordonnancement d'instances. Ces méthodes peuvent être vues comme des extensions des approches proposées en classification (Settles, 2009). Elles sont de trois types : sélection de l'exemple dont le score (ou le rang) est le moins sûr, sélection de l'exemple qui pourrait modifier le plus le modèle, et sélection de l'exemple qui contribuerait le plus à l'erreur moyenne sur les exemples non-étiquetés.

4 Proposition d'une méthode d'ordonnancement d'alternatives actif

Nous proposons une approche qui consiste à apprendre une fonction d'ordonnancement d'alternatives à partir de l'adaptation de l'algorithme RankBoost sur une base d'exemples étiquetés et une autre base d'exemples non-étiquetés en utilisant une approche d'apprentissage actif. Les données non-étiquetées vont être étiquetées par cet apprentissage. En effet, l'apprentissage actif est une stratégie bien connue pour apprendre progressivement à partir des instances non-étiquetées. Généralement, le processus est itéré jusqu'à ce qu'une condition d'arrêt soit satisfaite.

RankBoost est un algorithme d'ordonnancement supervisé à caractère inductif. En effet, il est capable d'ordonner un ensemble d'exemples non vus durant la phase d'apprentissage en inférant un ordre sur cet ensemble (Freund et al, 2003).

Nous pouvons résumer notre approche par les étapes suivantes :

- sélectionner les entrées de la base étiquetée;
- apprendre une nouvelle fonction de score sur cette base en utilisant RankBoost ;
- sélectionner une ou plusieurs alternatives d'un ensemble non-étiqueté et interroger un oracle pour obtenir leur étiquette ;

- ajouter ces alternatives à l'ensemble d'apprentissage pour apprendre une nouvelle fonction de score ;

Ces étapes vont être répétées jusqu'à ce qu'une condition d'arrêt soit satisfaite.

Nous disposons donc d'une base d'apprentissage étiquetée S_L constituée de m couples (observation, sortie réelle) $S_L = \{(x_1, y_1), ..., (x_m, y_m)\}$, où chaque y_i est un vecteur de taille variable m_i et m_i désigne le nombre d'alternatives candidates pour x_i. Aussi, $y_i = (y_i^1, ..., y_i^{m_i})$, avec $y_i^k \in IR$. Nous disposons également d'une base d'apprentissage non-étiquetée S_U, cette base va être étiquetée par la méthode d'apprentissage sélectif actif.

Si X l'ensemble des observations de la base S_L et Y l'ensemble des sorties réelles. $Y = Y_+ \cup Y_-$ où Y_+ et Y_- les ensembles respectifs des alternatives pertinentes et non-pertinentes de S_L. Le but est de trouver une fonction de score F en utilisant l'information contenue dans les deux bases S_L et S_U.

Dans ce qui suit, nous détaillons le fonctionnement de l'algorithme RankBoost appliqué à notre contexte.

4.1 Adaptation de RankBoost à l'ordonnancement d'alternatives

L'adaptation de RankBoost est donnée dans l'algorithme 1 : A chaque itération t, l'algorithme maintient une distribution λ_t sur les exemples de la base d'apprentissage B, une distribution ν_t^i sur les alternatives associées à l'exemple x_i et une distribution D_t^i sur les paires critiques d'alternatives, représentée par une distribution sur les couples (k, l) tels que $y_i^k \in Y_+$ et $y_i^l \in Y_-$ pour chaque exemple x_i. La distribution D_t^i est définie à base des deux autres: $\forall$ i$\in \{1,..,m\}, \forall (k, l) \in \{1,..,\ m_i\}^2, y_i^k \in Y_+, y_i^l \in Y_- : D_t^i\ (k, l) = \lambda_t^i\ \nu_t^i(k)\ \nu_t^i(l)$.

Ces distributions sont mises à jour grâce à la fonction de base f_t, sélectionnée à partir de l'algorithme de recherche des fonctions de score qui renvoie la valeur du seuil θ_{res} résultante associée à chaque caractéristique et les valeurs possibles qui peuvent être associées à f_t, tel que :

$$f_t(x_i,k) = \begin{cases} 1 & si\ \varphi_j(x_i,k) > \theta_{res} \\ 0 & si\ \varphi_j(x_i,k) \le \theta_{res} \end{cases}$$

avec x_i est l'observation d'indice i, k est l'indice de l'alternative associée à cet x_i.

Le poids α_t est défini par :

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1+r_t}{1-r_t}\right) \quad \text{où} \quad r_t = \sum_{k,l} D_t^i(k,l)\left(f_t(x_i,k) - f_t(x_i,l)\right)$$

D'après (Freund, 2003) et (Usunier, 2006), le critère d'apprentissage (ou erreur d'ordonnancement) a pour borne supérieure :

$$R_m^{OA}(F,S) \le \prod_t^T Z_t \text{ où } R_m^{OA}(F,S) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i p_i}\sum_{k:y_i^k \in Y_+}\sum_{l:y_i^k \in Y_-}[\![f(x_i,k) - f(x_i,l) \le 0]\!]$$

A chaque itération t, α_t est choisi de façon à minimiser Z_t. L'algorithme détaillé fait donc décroître itérativement la borne supérieure sur la fonction $R_m^{OA}(F,S)$.

Algorithme 1 : Algorithme RankBoost adapté à l'ordonnancement d'alternatives

Entrées : Une base d'apprentissage $S_L = \{(x_i, y_i); i \in \{1,..,m\}\}$, pour chaque exemple x_i, il y a m_i alternatives candidates

Initialisation : $\forall$ i$\in \{1, ..., m\}$, $\lambda_1^i = \frac{1}{m}$ $\quad$ $v_1^i(k) = \begin{cases} \frac{1}{p_i} \; si \; y_i^k \in Y_+ \\ \frac{1}{n_i} \; si \; y_i^k \in Y_- \end{cases}$

Début

pour t :=1..T faire

- Sélectionner la fonction de base f_t à partir de D_t
- Calculer α_t en utilisant la formule définie ci-haut
- $\forall$ i$\in \{1, .., m\}$, $\forall$ $(k, l) \in \{1, ..., m_i\}^2$ t.q. $y_i^k \in Y_+$, $y_i^l \in Y_-$,
- mettre à jour $D_{t+1}^i(k, l)$: $D_{t+1}^i(k, l) = \lambda_{t+1} v_{t+1}^i(k) \; v_{t+1}^i(l)$
- $\forall$ i$\in \{1, .., m\}$, $\lambda_{t+1}^i = \frac{\lambda_t^i Z_t^{-1i} Z_t^{1i}}{Z_t}$

$$v_{t+1}^i(k) = \begin{cases} \frac{v_t^i(k)\exp(-\alpha_t f_t(x_i, k))}{Z_t^{1i}} \; si \; y_i^k \in Y_+ \\ \frac{v_t^i(k)\exp(\alpha_t f_t(x_i, k))}{Z_t^{-1i}} \; si \; y_i^k \in Y_- \end{cases}$$

où Z_t^{1i}, Z_t^{-1i} et Z_t sont défini par: $Z_t^{1i} = \sum_{k: y_i^k \in Y_+} v_t^i(k)\exp(-\alpha_t f_t(x_i, k))$

$$Z_t^{-1i} = \sum_{l: y_i^l \in Y_-} v_t^i(l)\exp(\alpha_t f_t(x_i, l)), \; Z_t = \sum_{i=1}^{m} \lambda_t^i Z_t^{-1i} Z_t^{1i}$$

Fin

Sortie : La fonction de score finale $F = \sum_{t=1}^{T} \alpha_t f_t$

Nous proposons ainsi un algorithme d'apprentissage actif pour l'ordonnancement d'alternatives en intégrant l'algorithme RankBoost adapté à l'ordonnancement d'alternatives.

Algorithme 2. Algorithme l'ordonnancement d'alternatives actif

Entrée :

Une base d'apprentissage S_L et une base d'apprentissage non-étiquetée S_U
L'algorithme d'ordonnancement supervisé RankBoost adapté à l'ordonnancement d'alternatives (algorithme 1)
k : nombre de partitions de S_L
Pour t : = 1,…, k **faire**

- Apprendre une fonction d'ordonnancement avec RankBoost sur S_L
- étiqueter les alternatives associées en se focalisant uniquement sur les données non-étiquetées disponibles S_U

- Les retirer de S_U et les ajouter dans S_L avec les étiquettes sur les alternatives

Fin

Dans ce qui suit, nous proposons la préparation de la partie expérimentale pour l'évaluation de l'algorithme proposé dans le processus de RI.

5 Préparation de la partie expérimentale

Pour valider notre approche, nous choisissons d'utiliser la collection de données OHSUMED issue du benchmark standard LETOR (LEarning TO Rank) (Liu, 2007) qui a été réalisé pour l'apprentissage de fonctions d'ordonnancement dans le domaine de la RI. LETOR propose un ensemble de résultats issus de plusieurs algorithmes d'ordonnancement utilisant plusieurs collections tels que RankBoost (Freund, 2003). Pour l'évaluation des résultats, nous choisissons de comparer notre algorithme avec RankBoost.

OHSUMED définit des paires requête-document, chacune est constituée d'un vecteur de caractéristiques et d'un jugement de pertinence correspondant. Les jugements de pertinence sont à trois niveaux dans le sous-ensemble OHSUMED: {0, 1, 2} correspondant respectivement à « non pertinent », « partiellement pertinent », et « certainement pertinent ».

Nous choisissons la version courante de LETOR (version 4.0), utilisant 46 caractéristiques extraites à partir d'OHSUMED. Nous nous servons, dans la partie expérimentale, de ces caractéristiques pour déterminer les valeurs des caractéristiques φ_j afin de compléter par la suite les valeurs des fonctions de base.

Chaque sous-ensemble de OHSUMED a été divisé en cinq parties, dénotées S1, S2, S3, S4 et S5. L'ensemble d'apprentissage est utilisé pour apprendre le modèle d'ordonnancement proposé. L'ensemble de validation est employé pour corriger les paramètres du modèle d'ordonnancement, tels que le nombre d'itérations T de notre algorithme. L'ensemble de test est employé pour vérifier la performance du modèle d'ordonnancement. Il est à noter que puisque nous conduisons la validation sur la base de ces cinq ensembles, le calcul effectué est réellement la moyenne à travers ces différents ensembles.

Afin de comparer la performance et évaluer l'efficacité de l'algorithme proposé, nous utilisons trois mesures d'évaluation largement utilisées pour comparer des SRI et prouvées par LETOR : Precision at position n (P@n), Mean Average Precision (MAP) et Normalized Discounted Cumulative Gain (NDCG). Et sont définie comme suit :

$$P@n = \frac{\#docs\ pertinents\ dans\ les\ n\ ins\tan ces\ les\ mieux\ ordonnées}{\mathrm{n}},$$

$$MAP = \frac{\sum_{n=1}^{N}(P@n * rel(n))}{\#total\ docs\ pertinents\ pour\ cette\ requête},\ N(n) = Z_n \sum_{j=1}^{n} \frac{2^{r(j)}-1}{\log(1+j)}.$$

6 Conclusion

Dans cet article, nous avons présenté une stratégie d'apprentissage actif de fonctions d'ordonnancement d'alternatives. L'apprentissage actif permet de créer un modèle pertinent avec des données d'apprentissage peu nombreuses car bien choisies. Cette approche permet l'élaboration d'une base étiquetée à faible coût. Notre objectif est d'améliorer la qualité de la

RD en proposant une adaptation de l'algorithme inductif RankBoost pour la génération d'une fonction d'ordonnancement d'alternatives.

Nous nous proposons dans l'étape à suivre de compléter la partie expérimentale et d'intégrer d'autres méthodes d'apprentissage.

Références

Amini, M.-R. (2007). *Apprentissage de Fonctions de Classification et d'Ordonnancement avec des Données Partiellement Étiquetées*, Habilitation, Labo d'Informatique de Paris 6.

Usunier, N, (2006) *Apprentissage de fonctions d'ordonnancement: une étude théorique de la réduction à la classification et deux applications à la Recherche d'Information.* Thèse.

Amini, M.-R., V. Truong et C. Goutte, (2008) A Boosting Algorithm for Learning Bipartite Ranking Functions with Partially Labeled Data, In Proceedings of SIGIR 2008 Workshop on Learning to Rank for Information Retrieval.

Freund, Y., R. Iyer., R. E Schapire et Y. Singer, (2003). An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research : 933-969.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

Cohn, D., Atlas, L., and Ladner, R. (1994). Improving generalization with active learning. Machine Learning, 15(2) :201-221.

Truong, V., (2009). *Apprentissage de Fonctions d'Ordonnancement avec peu d'Exemples Étiquetés.* Thèse de Doctorat.

Liu, T.-Y., J. Xu, T. Qin, W.-Y. Xiong et H. Li. (2007) LETOR : Benchmark dataset for research on learning to rank for information retrieval. In SIGIR.

Summary

In this paper, we propose an active learning method of ranking functions of alternatives. The principal motivation which led us to the development of this proposal is that to find an effective ranking function it is necessary to have a base of learning which often requires the manual labelling of the alternatives on several examples. The goal is to find the best entered to label to reduce to the maximum the number of labelled data. For an effective use of the active learning on large collections, we adapt a modification of the supervised ranking RankBoost algorithm. We present the model suggested and describe its functionalities as well as the choices of implementation.

Evolution des ontologies : un panorama

K. Kamoun[1,2], M. Harzallah[2], P. Kuntz[2], S. Ben Yahia[1]
[1]LINA, Polytech'Nantes - Site de la Chantrerie
rue Christian Pauc, BP 50609, 44306 Nantes cedex 3
{karim.kamoun@gmail.com},
[2]Département des Sciences de l'Informatique, Faculté de Sciences de Tunis
Campus Universitaire, 1060 Tunis, Tunisie
sadok.benyahia,@fst.rnu.tn

Résumé. Durant cette décennie, les ontologies ont été l'un des modèles d'ingénierie des connaissances les plus utilisés dans de nombreux domaines applicatifs. Après la phase de construction, leur utilisation opérationnelle renvoie à un nouveau problème : celui de la gestion de leur évolution eu égard aux changements inhérents aux domaines dans lesquelles elles sont utilisées et aux applications qui les intègrent. Différentes approches ont été récemment développées dans la littérature. L'objectif de cette communication est de dresser un panorama qui tente de mettre en évidence les avantages et lacunes des technologies actuelles de gestion de l'évolution des ontologies.

1 Introduction

Depuis les années 90, l'importance des ontologies est reconnue dans divers domaines de recherche comme la représentation des connaissances, l'ingénierie des connaissances, la conception de bases de données, l'intégration d'information, les systèmes d'information etc. Les ontologies sont aussi centrales pour le Web sémantique qui, d'une part, cherche à s'appuyer sur des modélisations de ressources du Web à partir de représentations conceptuelles des domaines concernés et, d'autre part, a pour objectif de permettre à des programmes de faire des inférences dessus. Cependant, le Web sémantique est un environnement dynamique, multi-acteurs et distribué. Les ontologies du Web sémantique ne doivent pas être considérées comme des modèles stables qui ne subissent aucune mise à jour au cours du temps. Par conséquent, nous nous retrouvons actuellement devant la problématique de la gestion de leur évolution pour qu'elles restent utilisables devant les changements fréquents de leur domaine ou des applications qui les utilisent (Berners-Lee et al (2001), Wach et al(2001)).
Dans cet article, nous étudions et comparons les méthodologies existantes d'évolution des ontologies et nous proposons une description synthétique des différents outils déployés actuellement afin d'identifier leur avantages et leur lacunes par rapport aux nouveaux besoins de l'évolution.

2 Evolutions des ontologies : Définitions

D'une façon générale, la notion d'évolution des ontologies est une notion complexe qui fait référence à des facteurs structurels, fonctionnels et également logiciels dans leur

implémentation. Plusieurs tentatives de définitions ont été proposées dans la littérature. Maedche et al (2003), et Stojanovic (2004) pensent l'évolution comme « l*a modification appropriée d'une ontologie et la propagation des changements dans les autres ontologies qui en dépendent.* »
En parallèle, Klein, (2004), Noy, (2004), et par Noy et al (2003) ont proposé une autre définition plus précise qui dit que « l*'évolution d'une ontologie est* l*a capacité de gérer les changements apportés lors de l'évolution en créant et en maintenant différentes versions d'une ontologie. Cette capacité consiste à identifier et à différencier les versions, à modifier les versions, à spécifier des relations qui rendent explicites les changements effectués entre les versions.* »
Contrairement à la première définition, la seconde définition présuppose qu'une ontologie peut avoir plusieurs états (appelés versions) qui peuvent être explicitement différenciés.

3 Méthodologies existantes

Nous avons pu dégager, dans la littérature, deux méthodologies principales sur l'évolution d'ontologies : la méthodologie AIFB (Institute of Applied Informatics and Formal Description Methods), et la méthodologie de IMSE (Information Management and Software Engineering). Dans cette section, nous allons décrire ces deux méthodologies ainsi que deux autres qui s'en inspirent.

3.1 Méthodologie AIFB

Les auteurs de la méthodologie AIFB de l'université de Karlsruhe, proposent, pour supporter le processus d'évolution d'ontologie, une méthodologie qui repose sur six étapes décrites par Stojanovic et al (2002). Ces étapes sont les suivantes : détection des changements, représentation des changements, sémantique des changements, implémentation des changements, propagation des changements et enfin validation.

3.1.1 Détection des changements

La détection des changements se base sur les besoins explicites (ajout, suppression d'une entité ontologique) exprimés par les développeurs d'ontologies et sur l'application de méthodes heuristiques sur trois niveaux différents : la structure, les instances et les usages. Au niveau structurel, il s'agit d'exploiter un ensemble d'heuristiques pour améliorer une ontologie en se basant sur l'analyse de sa structure. Au niveau des données, il faut détecter les changements en basant l'analyse sur les instances des entités. Au niveau des usages, on peut suivre puis analyser la trace des utilisateurs qui ont utilisé l'ontologie au cours de son évolution.

3.1.2 Représentation des changements

Pour intégrer les changements détectés, ceux-ci doivent être représentés dans un format adéquat. Cette étape vise l'édition des changements élémentaires, composites ou complexes de façon à les adapter au langage ontologique utilisé.

3.1.3 Sémantique des changements

L'ontologie doit évoluer d'un état consistant vers un autre état consistant, c'est-à-dire un état où les contraintes structurelles et logiques sont respectées. Afin de résoudre les inconsistances introduites par certains changements, d'autres changements peuvent être

nécessaires. Cette étape vise alors à permettre la résolution de tous les changements additionnels d'une manière systématique en proposant à l'utilisateur des stratégies d'évolution.

3.1.4 Implémentation des changements

L'implémentation du changement, ou l'application physique du changement à l'ontologie, consiste à informer le développeur d'ontologie de toutes les conséquences d'une demande de changement (notification du changement), à appliquer tous les changements nécessaires et dérivés (application du changement) et à garder trace des changements effectués (notation de tous les changements).

3.1.5 Propagation des changements

La mise à jour d'une ontologie peut affecter négativement d'autres ontologies dépendantes. La phase de propagation du changement consiste à assurer la consistance des objets dépendants après qu'une mise à jour de l'ontologie ait été effectuée. Ces objets peuvent inclure des ontologies et des applications fonctionnant avec l'ontologie.

3.1.6 Validation des changements

Dans cette phase les utilisateurs évaluent le résultat de l'évolution et testent l'ontologie évoluée. Si la qualité de l'ontologie est préservée ou améliorée, c'est-à-dire si l'ontologie répond correctement aux requêtes des utilisateurs, alors les changements peuvent être validés ; sinon les annuler ou reprendre les phases d'évolution selon un processus cyclique. La quantification de la qualité d'une ontologie est un problème délicat encore largement ouvert.

3.2 Méthodologie IMSE

Les auteurs de la méthodologie IMSE (Information Management and Software Engineering) proposent un processus qui permet d'identifier et de différencier les versions, de modifier les versions, de spécifier des relations qui rendent explicites les changements effectués entre les versions et d'utiliser des mécanismes d'accès pour les artefacts dépendants[1]. Ce processus se base sur deux modules fondamentaux pour une méthodologie de '*versionnage*' des ontologies (Noy et Musen (2004), Klein (2002) et Klein et Fensel (2001)).

3.2.1 Module d'analyse de relation entre deux versions d'ontologies

Ce module permet les fonctionnalités suivantes : *(i)* mise en évidence des changements effectués dans la définition des entités ontologiques, *(ii)* spécification de la relation sémantique entre les entités ontologiques qui ont subi un changement dans leurs définitions, *(iii)* description des changements effectués par un ensemble de méta-données qui décrivent l'auteur, la date et le but de chaque changement.

[1] Artefacts dépendant : il s'agit objets qui dépendent de l'ontologie à évoluer comme les ontologies, les applications etc.

3.2.2 Module d'identification des versions d'ontologies sur le Web

Ce module permet de rendre opérationnelle la nouvelle version de l'ontologie après évolution. Il se base sur deux principes : *(i)* Un changement dans la définition d'une entité ontologique produit une nouvelle version, ayant un nouveau URI (*Uniform Resource Identifier*) alors qu'un changement dans l'annotation textuelle d'une entité produit seulement un nouveau fichier avec un nouveau URL (*Uniform Resource Locator*), *(ii)* la forme d'URI indique si la version est compatible avec la version antérieure.

3.2.3 Limites et autres méthodologies

Selon Rogozan (2008), les auteurs de la méthodologie AIFB ne proposent aucune étape d'analyse des effets des changements sur la relation entre l'ontologie évoluée et les artefacts dépendants. De plus, l'étape de propagation des changements est essentiellement unidirectionnelle car elle vise uniquement la modification des ontologies dépendantes afin de préserver leur consistance structurelle avec l'ontologie de base et ne touche pas ainsi le référencement sémantique des ressources. Djdidi et al. (2007) considèrent que l'étape d'évaluation de la qualité doit être intégrée dans le processus avant la mise en opération de l'ontologie après évolution. Concernant IMSE, Rogozan considère que les auteurs de cette méthodologie proposent une approche pour supporter la gestion des versions d'une ontologie après son évolution et non pas pour supporter l'évolution proprement dite des ontologies. En effet, ils fournissent un modèle d'analyse de la relation entre les versions de l'ontologie, mais ne se préoccupent pas de la gestion de l'accès aux artefacts dépendants. Sur la base de ces limites d'autres méthodologies ont été proposées, parmi lesquelles nous allons décrire brièvement celles de Rogozan (2008) et de Djdidi (2007).

La méthodologie proposée par Rogozan (2008) s'inspire des deux méthodologies AIFB et IMSE. Cette méthodologie suit un processus composé de huit étapes. Les six premières étapes correspondent à celles de la méthodologie AIFB suivies par l'étape d'analyse des changements qui permet de fournir une analyse des effets de changements afin d'identifier ceux susceptibles de provoquer des problèmes de dysfonctionnement de systèmes, des inconsistances d'ontologies dépendantes ou encore des pertes d'accès aux ressources référencées sémantiquement. La dernière étape du processus consiste en la mise en opération de la version VN+1, avec la prise en compte de la propagation des changements vers les artefacts dépendants. L'objectif de cette étape est de préserver les rôles de l'ontologie pour la nouvelle version VN+1 soit en utilisant seulement la nouvelle version, soit des versions multiples présentant des liens de correspondances. Le choix de l'une de situation se fait selon la possibilité des changements exécutés pour passer de VN àVN+I. Cette méthodologie ignore la vérification de la qualité de l'ontologie après évolution et avant la mise en opération de la nouvelle version.

La méthodologie proposée par Djdidi (2007) s'inspire largement de celle d'AIFB. Cette méthodologie se base sur un processus en six étapes qui sont : identification du changement, représentation, implémentation, validation, évaluation de la qualité de l'ontologie et annotation. Les deux premières étapes sont identiques à celles d'AIFB. L'étape d'implémentation consiste à appliquer le changement identifié et ses changements dérivés en propageant automatiquement le changement vers les entités qui en dépendent. L'application du changement se fait tout en préservant la trace des modifications effectuées et les métadonnées associées. A l'issue de cette phase, l'ontologie sera « provisoirement » mise à

jour. L'intégration finale du changement est contrainte par la vérification de la consistance et par la préservation de la qualité de l'ontologie. Une fois les changements acceptés, la phase d'annotation permet de garder la trace des opérations de changements et l'historique de l'évolution de l'ontologie. La méthodologie de Djdidi complète des aspects qui ont été ignorés dans les autres approches ; en particulier elle met l'accent sur l'évaluation de l'ontologie qui évolue. Cependant, cette méthodologie ne comporte aucune étape qui décrit la propagation des changements vers les artefacts dépendants.

4 Outils et approches existants

Dans cette section, nous présentons différents outils qui nous semblent parmi les plus performants et nous tentons de les situer selon les différentes étapes du processus d'évolution sur lesquelles ils se focalisent. Un des premiers outils qui intègre le processus d'évolution d'ontologies a été le système KAON[2] sigle correspondant à KArlsruhe ONtology. Cet outil a essayé de suivre le processus d'évolution de la méthodologie AIFB, en mettant l'accent surtout, sur la représentation des changements avec un langage interne à l'outil KAON et sur la vérification de la consistance structurelle et logique en proposant à l'utilisateur des stratégies de résolution des inconsistances déduites après évolution. Cependant, l'outil KAON ne traite pas des changements complexes (fusion, éclatement,…) et a ignoré une étape très importante du processus d'évolution qui est la propagation des changements. Pour intégrer les changements complexes**,** Rogozan (2008) a récemment proposé un « ontoanalyseur ». Cet outil est fondé sur un modèle uniformisé de représentation des changements, mais il est aussi, riche sémantiquement car il est basé sur une ontologie des changements. Ontoanalyseur se focalise sur la propagation des changements vers les artefacts dépendants et plus précisément vers les référencements sémantiques des ressources de l'ontologie. Pour améliorer la propagation des changements vers le référencement sémantique des ressources Luong et Dieng-Kuntz (2007) ont développé l'outil CoSWEM (Corporation Semantic Web Evolution Management). Ce système permet à l'utilisateur, grâce à ses interfaces graphiques, de comparer les différences entre deux versions de l'ontologie.

Les outils proposés précédemment nécessitent l'intervention de l'utilisateur pour indiquer le changement à effectuer et la partie de l'ontologie concernée par ce changement. Cette opération s'avère une tache très complexe même pour les spécialistes du domaine. Par conséquent, un axe de recherche actuel concerne l'automatisation de l'étape de détection et de placement du changement.

L'approche de Tho et al (2008) entre dans ce contexte. L'auteur propose une technique d'évolution d'ontologies qui peut enrichir une ontologie construite manuellement avec des connaissances supplémentaires découvertes d'une manière automatique. Les connaissances extraites, sont formalisées en des entités liées formant une petite ontologie qui sera intégrée, grâce à des mesures de similarité floues, dans l'ontologie à évoluer.

L'approche proposée par Chrisment et al. (2006) permet de mettre à jour une ontologie légère du domaine en analysant de nouveaux documents appartenant au domaine. La mise à jour de l'ontologie se réalise à partir de l'analyse d'un corpus et de la gestion de types abstraits (concepts de haut niveau d'abstraction). La détection de nouveaux termes est

[2] KAON (KArlsruhe ONtology and Semantic Web infrastructure): http://kaon.semanticweb.org!.

effectuée grâce à deux règles. La première permet d'extraire les termes généraux qui n'existent pas dans l'ontologie, la deuxième permet d'extraire des termes spécifiques du corpus. Les termes extraits sont ensuite intégrés dans l'ontologie de base moyennant deux règles. La première vise à intégrer les nouveaux termes dans la hiérarchie des concepts existants et la seconde permet de créer de nouvelles relations associatives (relations sémantiques) entre les concepts existants.
L'outil de Trousse et al (2008) se focalise précisément sur la détection des changements au niveau des usages. Il intègre des algorithmes de fouille de données qui permettent d'extraire des connaissances concernant le profil et les préférences des utilisateurs grâce à l'analyse de leur trace.
Le système Evolva, (Zablith (2009)) s'appuie sur des heuristiques d'extraction de connaissances qui permettent d'établir, grâce à des bases de connaissances, des relations entre les connaissances extraites et les concepts de l'ontologie initiale.

Notons que dans les outils évoqués ci-dessous, l'évaluation de la qualité des ontologies n'est pas abordée. Pour palier à ce manque, différentes pistes sont en cours d'investigation. Yinglin (2008) a récemment proposé une approche qui étudie l'impact des changements et plus précisément de deux types de changements (le range de la propriété et la propriété Split) sur l'ontologie et leur propagation sur les applications dépendantes.
L'outil ONTO-Evo[a]l (Ontology evolution evaluation) de Djdidi et al (2009) s'appuie sur une modélisation à l'aide de trois types de patrons (patrons de changements, patrons des incohérences et patrons des alternatives) pour la résolution des incohérences. Une fois la consistance de l'ontologie vérifiée l'outil passe à la phase évaluation dans laquelle il mesure l'impact du changement sur la qualité de l'ontologie. ONTO-Evo[a] utilise différentes métriques d'évaluations classées sous deux aspects : l'aspect structurel et l'aspect usage. Dans l'aspect structurel les métriques utilisées exploitent les critères suivants : la complexité, la cohésion, la modularité, la taxonomie et l'abstraction et, pour l'aspect associé aux usages, trois critères sont distingués : la complétude, la modularité et la compréhension.

5 Discussion

La plupart des approches et des outils que nous avons décrits reposent sur une méthodologie qui suit un processus d'évolution d'ontologies bien déterminé. Un tel processus doit être complété par une phase d'évaluation de la qualité de l'ontologie avant la mise en opération de la nouvelle ontologie. Les travaux proposés par Djdidi et Yinglin se situent dans ce cadre. Cependant, Il faut signaler qu'une ontologie est avant tout un outil. C'est donc en situation qu'il convient d'évaluer sa qualité et son intérêt. Cette étape sert à juger de la valeur ajoutée (compréhension, utilisabilité. . .) et de la qualité de l'ontologie du point de vue de l'utilisateur. Aucun des outils décrits ne présentent une méthode pour évaluer la qualité de l'ontologie évoluée au niveau des usages d'une façon générale avec n'importe quelle application interagissant avec l'ontologie en question On remarque aussi, que presque tous les outils qui utilisent des techniques pour une détection automatique des changements effectuent seulement l'enrichissement de l'ontologie en évolution et ignorent par conséquent les opérations de modification et de suppression des éléments existants. De plus, les propagations des changements vers les artefacts dépendants (à savoir les ontologies, les référencements sémantiques, les applications) sont rarement abordées et mal traitées. En effet, il n'y a que les outils OntoAnalyseur et CoSWEM qui assurent seulement une propagation vers les référencements sémantiques des ressources.. Un autre aspect qui, à notre

connaissance, n'a jamais été abordé dans le processus de la gestion d'évolution d'ontologie est la prise en compte de l'aspect temporel : sur quelle échelle de temps doit-on prendre en compte les changements ?

Références

Berners-Lee T., Hendler J., Lassila O. (2001), *The semantic Web* , Scientific American.

Chrisment.C, Hernandez.N , Hubert.G, et Mothe.J (2006), Mise à jour d'une ontologie de domaine à partir de l'analyse de nouveaux documents du domaine pour l'indexation de documents, In information - Interaction - Intelligence (I3), Cépaduès Editions, *Numéro spécial Textes et ressources terminologiques et/ou ontologiques: évolution et maintenance*.

Djdidi,R., Abboute,H., Aufaure M.A.(2007), Evolution d'ontologie : Validation des changements basée sur l'évaluation. *actesde premières journées francophones sur les ontologie (JFO).*

Djdidi.R, Aufaure.M.A(2009)., Patrons de gestion de changements. *ingénieries des connaissances.*

Klein, M.(2004), *Change Management for Distributed Ontologies*. PhD thesis, Vrije University of Amsterdam.

Klein, M. et Fensel, D(2001). Ontology versioning for the Semantic Web. *Journal of the First International Semantic Web Working Symposium (SWWS)*, pages 75-91.

Klein, M(2002). Supporting evolving ontologies on the internet; *Journal of the EDBT 2002 PhD Workshop.*

Luong, P-H. et Dieng-Kuntz, R. (2007). A Rule-based Approach for Semantic Annotation Evolution. *Journal of The Computational Intelligence*, 23(3):320-338. Blackwell Publishing, Malden, MA 02148..

Maedche, A., Motik, B., et Stojanovic,(2003), Managing Multiple and Distributed Ontologies in the Semantic Web. *VLDE Journal -Special Issue on Semantic Web*, 12,286-302,

Noy, N. (2004), Tools for Mapping and Merging Ontologies. In S. Staab et R. Studer (Eds.), *Handbook on Ontologies:* Springer Verlag,

Noy, N., Kunnatur, S., Klein, M., et Musen, M(2003), *Tracking Changes During Ontology Evolution. Journal of 3rd International Semantic Web Conference (ISWC2004).*

Noy, N.F., Musen, M.A. (2004). Ontology Versioning in an Ontology Management Framework. IEEE Intelligent Systems, Vol. 19, No. 4.

Rogozan,D. C (2008)., *Gestion de l'évolution des ontologies : méthodes et outils pour un référencement sémantiques évolutif fondé sur une analyse des changements entre versions d'ontologie.* Rapport de thèse de recherche doctorale en informatique cognitive (DIC 9410). Télé-Université du Québec.

Stojanovic, L(2004). *Methods and Tools for Ontology Evolution*. PhD thesis, University of Karlsruhe.

Stojanovic, L., Maedche, A., Stojanovic, N. et Motik, B (2002).User-driven ontology evolution management. *Journal of the 13th European Conference on Knowledge Engineering and Knowledge Management (EKAW-2002),* pages 285-300. Springer-Verlag.

Tho T.Q and Thai D.N (2008) : Ontology Evolution for Customer Services. *Journal of The Knowledge Representation Ontology Workshop (KROW 2008*), Sydney, Australia. Conferences in Research and Practice in Information Technology, Vol. 90

Trousse, B., Aufore, M.A., Le grand, B., Lechevakier, Y., and Masseglia, F. in *Web Usage Mining for ontology management* in Data mining Mining with ontologies imlementations, finding, and frameworks, chapter 3, page 37-64

Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., et al.(2001), Ontology-based integration of information -a survey of existing approaches. *Journal of IJCAI on Ontologies and Information Sharing.*

Yinglin.W, Xijuan.L, Rongwei.Y (2008), Ontology Evolution Issues in Adaptable Information Management Systems. *Journal of IEEE International Conference on e-Business Engineering* pp.753-758.

Zablith, F. (2009) Evolva: A Comprehensive Approach to Ontology Evolution, European Semantic Web Conference (ESWC) PhD Symposium, Crete, Greece, *Journal of the 6th European Semantic Web Conference*, LNCS 5554, eds. L. Aroyo et al., pp. 944-948, Springer-Verlag, Berlin, Heidelberg

SUMMARY

In these recent years, we have witnessed a substantial increasingly important ontologies. Several ontologies are further developed and actually we find ourselves standing against the problem of how we can manage their development in a way so that they can remain available to frequent changes in their fields or even in the applications that use. Following a well-defined methodology, a variety of tools and approaches have been suggested. In this article we are presenting a whole review of these tools and the different existing methodologies and we classify it.

Enrichissement sémantique de requêtes au moyen d'ontologies de domaine personnalisées.

Xavier Aimé* ***, Frédéric Fürst**, Pascale Kuntz*, Francky Trichet*

*LINA - Laboratoire d'Informatique de Nantes Atlantique (UMR-CNRS 6241)
Équipe COD - Connaissances & Décision
2 rue de la Houssinière BP 92208 - 44322 Nantes Cedex 03
pascale.kuntz@univ-nantes.fr, francky.trichet@univ-nantes.fr
**MIS - Laboratoire Modélisation, Information et Système
UPJV, 33 rue Saint Leu - 80039 Amiens Cedex 01
frederic.furst@u-picardie.fr
***Société TENNAXIA
37 rue de Châteaudun - 75009 Paris
xaime@tennaxia.com

Résumé. Les approches classiques de recherche d'information sont fondées essentiellement sur la détection de présence de mot-clés dans des documents. Nos travaux visent à étendre les requêtes saisies par un utilisateur, en élargissant le champ de recherche au moyen d'ontologies personnalisées. Par le biais des gradients de prototypicalités (représentées, pour la prototypicalité conceptuelle, par des pondérations sur les liens hiérarchiques et les propriétés, et, pour la prototypicalité lexicale, par des pondérations sur les termes), nous personnalisons pour un utilisateur donné tant l'extension de la requête saisie que la quantité de résultats fournis. Ce nouveau processus a pour effet (1) d'augmenter le rappel (nous récupérons davantage de documents) et la précision (nous limitons le nombre de documents pertinents non retournés), et (2) de fournir pour une même requête des résultats adaptés au profil des utilisateurs.

1 Introduction

Selon Zaher et al. (2007), dans de nombreux cas, la recherche d'information s'inscrit dans la problématique de l'analyse d'une situation complexe dont on cherche à percevoir les contours sans savoir à l'avance s'il existe des ressources documentaires susceptibles de répondre au besoin. La recherche d'information syntaxique et lexicale, basée sur la recherche de *mot-clés / termes* dans des documents, est très proche de la reconnaissance de formes. Ce n'est que par opérations cognitives successives de l'utilisateur qu'il y a assimilation entre une suite de symboles (la syntaxe), le terme recherché (le lexique), et sa signification (la sémantique). La recherche d'information sémantique a pour objet, par rapport à la recherche d'information syntaxique et lexicale, d'augmenter (1) le rappel et (2) la précision. Pour augmenter le rappel, l'extension de requête va permettre de récupérer davantage de documents et donc de limiter

le nombre de documents pertinents non retournés. Pour augmenter la précision, effectuer des requêtes sur des concepts et non sur des termes permet de limiter le retour de documents non pertinents. Étendre une requête va donc consister à compléter les mot-clés par une liste de termes dénotant le même concept (ainsi que sa descendance). Plusieurs types d'extension de requêtes sont possibles : synonymes, méronymes, hyponymes, hyperonymes, co-occurences et autres relations sémantiques [Guelfi et al. (2007); Messai et al. (2006)]. Ces extensions peuvent être interactives (l'utilisateur choisit dans l'ontologie les concepts sur lesquels il souhaite étendre sa recherche) ou automatiques. Le processus d'extension d'une requête, dès lors, peut-être décomposé en deux problématiques : (1) comment ajouter les termes et (2) quels sont les termes à ajouter. Dans un modèle booléen, s'il y a une relation forte entre le mot-clé A et un terme B (synonymie par exemple) alors la requête initiale A sera remplacée par $A \vee B$. Les nouveaux termes peuvent être pondérés, auquel cas il peut se poser la question d'une part de leur ordonnancement dans la requête, et d'autre part de leur sélection ou non. Derrière la question de savoir quels termes ajouter, se cache en fait l'idée du type de ressources à adopter pour enrichir la requête. La première idée est d'utiliser un dictionnaire de synonymes ou encore un thésaurus. Une autre solution, plus élaborée, va consister à utiliser une ontologie de domaine, en exploitant toute la richesse des relations sémantiques offertes. Notre approche vise à prendre en compte cette richesse, en s'adaptant à l'utilisateur au moyen d'une ontologie personnalisée.

La suite de cet article est structurée comme suit. La section 2 introduit brièvement notre approche de la personnalisation des ontologies, les différents types de prototypicalité utilisées, ainsi qu'une mesure de similarité définie dans ce contexte. La section 3 décrit en détail les différents cas d'enrichissement sémantique de requêtes au moyen d'une ontologie de domaine personnalisée.

2 Contexte de nos travaux

Les systèmes d'information (SI) exploitent depuis des années les ontologies, définies comme des représentations conceptuelles des connaissances d'un domaine donné et reposant sur un consensus partagé par un endogroupe[1]. Classiquement, une ontologie est composée d'ensembles hiérarchisés de concepts et de propriétés[2], enrichis à l'aide d'axiomes affinant la représentation de la sémantique du domaine. Cependant, une telle ontologie ne capture pas la totalité des connaissances que les membres de l'endogroupe possèdent sur le domaine. Ainsi, une ontologie ne dit rien quant à la représentativité d'un concept par rapport à son (ou ses) sur-concept(s). Cette notion, connue sous le nom de *prototypicalité* en psychologie cognitive, est pourtant sous-jacente à toute catégorisation conceptuelle [Rosch (1975)]. Par exemple, en Europe, si les perroquets, les poules et les moineaux sont tous considérés comme des sortes d'oiseaux, le concept de moineau est cependant plus proche conceptuellement de celui d'oiseau, que ne le sont ceux de poule ou de perroquet. En d'autres termes, penser à un oiseau nous conduira bien plus volontiers à penser à un moineau qu'à un perroquet ou une poule.

[1]Endogroupe, terme utilisé en science cognitive pour désigner un groupe d'individus partageant des connnaissances communes, est ici utilisé pour désigner l'ensemble des personnes qui partagent la conceptualisation exprimée par l'ontologie, et non uniquement les personnes qui ont participé à sa construction. D'un point de vue social, un endogroupe peut être assimilé à un réseau épisthémique.

[2]Le terme propriété est pris au sens large et inclut les relations unaires (attributs) et n-aires.

La prototypicalité, comme toute connaissance, est subjective, et peut varier d'un individu à l'autre. Il est cependant possible de bâtir une ontologie au sein d'un endogroupe où il existe un consensus, non seulement sur les hiérarchies de concepts et les propriétés, mais également sur les prototypicalités entre concepts. Nous proposons d'exploiter cette notion de prototypicalité pour la personnalisation des ontologies, en considérant que le consensus sur lequel est basée l'ontologie ne porte que sur les concepts, les propriétés, les liens hiérarchiques et les connaissances axiomatiques. Au sein de l'endogroupe, les prototypicalités peuvent donc varier d'un individu à l'autre, ce qui va permettre d'adapter l'ontologie à chaque utilisateur, ou groupe d'utilisateurs. Dans le cadre d'une recherche d'information, par exemple, ces prototypicalités pourront servir à l'extension de requête (la requête est étendue aux concepts les plus prototypiques de ceux qui y apparaissent déjà) ou la personnalisation de la présentation des résultats (les résultats les plus prototypiques sont présentés en premier).

Nous proposons donc de faire de ces ontologies elles-mêmes le support de la personnalisation du SI, en ce sens qu'elles représentent un fond cognitif commun à tous les utilisateurs potentiels du système, et qu'il est possible de les moduler en y ajoutant des connaissances supplémentaires, variables selon les utilisateurs. Nous proposons d'utiliser comme connaissances additionnelles les degrés de prototypicalité entre deux entités cognitives, c'est-à-dire des degrés de représentativité d'une entité par rapport à l'autre [Aimé et al. (2009a)]. Notre approche sémiotique permet de combiner les trois dimensions d'une conceptualisation : (1) le *signifié*, *i.e.* le concept défini en intension (structure formelle), (2) le *signifiant*, *i.e.* les termes désignant le concept (contenus dans un corpus de textes relatifs au domaine couvert par l'ontologie), et (3) le *référent*, *i.e.* le concept défini en extension (population d'instances des concepts de l'ontologie). Nous introduisons les prototypicalités, d'une part, entre deux concepts liés hiérarchiquement (*prototypicalité conceptuelle*[3]) et, d'autre part, entre un concept et un terme le dénotant (*prototypicalité lexicale*), ce qui nous permet de personnaliser l'ontologie sur le plan conceptuel et sur le plan terminologique. Ces prototypicalités sont représentées, pour la prototypicalité conceptuelle, par des pondérations sur les liens hiérarchiques et les propriétés, et, pour la prototypicalité lexicale, par des pondérations sur l'ensemble des termes utilisés pour dénoter les conepts. Dans cet article, nous utilisons également SEMIOSEM, une mesure de similarité définie dans ce même cadre sémiotique [Aimé et al. (2009b)]. SEMIOSEM est une mesure issue de l'agrégation et l'enrichissement de travaux existants, avec pour particularité d'être indépendante de la structure de la hiérarchie de subsomption.

3 Une méthode de RI fondée sur une ontologie personnalisée

L'objectif est de concevoir un système de recherche d'information sémantique fondée sur une ontologie personnalisée. À partir d'une indexation lexicale des documents d'un corpus, cette application offre à l'utilisateur une interrogation, non plus par mots-clés uniquement, mais par concepts. Cette extension, guidée par les valeurs de gradients de prototypicalité, se fait tant sur les concepts (parents, descendants, ...) que sur les termes dénotant les concepts. Nous focalisons nos travaux sur des requêtes portant sur un ou deux concepts. Le processus de recherche d'information va s'articuler autour (1) de l'ontologie et des gradients de prototy-

[3]Comme le montre la figure 1, nous distinguons deux types de prototypicalité conceptuelle : une ascendante (calculée sur chaque sur-concept pour un concept donné) et une descendante (calculée sur chaque sous-concept pour un concept donné).

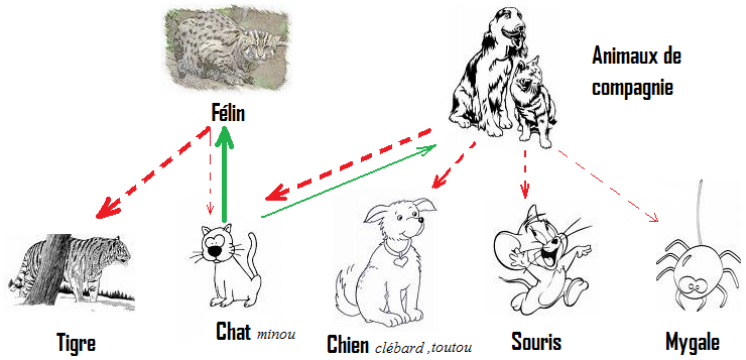


FIG. 1 – *D'un point de vue prototypicalité conceptuelle descendante (flèches en pointillés), pour un individu donné, les chats sont considérés comme les animaux de compagnie les plus prototypiques. D'un point prototypicalité conceptuelle ascendante (flèches pleines), pour un individu donné, les chats sont davantage considérés comme des félins que des animaux de compagnie. D'un point de vue prototypicalité lexicale, pour un individu donné, le terme Chat est plus prototypique pour désigner ce concept que le terme Minou.*

picalité conceptuelle qui vont nous guider sur le focus de recherche, (2) de la prototypicalité lexicale qui va faciliter la reformulation des requêtes.

3.1 Hypothèses de départ

Un certain nombre d'hypothèses de départ sont fixées quant à la recherche d'information sémantique :

- nous entendons par « *recherche sur un concept c* » une recherche lexicale sur l'ensemble des termes dénotant ce concept (triés par ordre décroissant de valeur de prototypicalité lexicale) ;
- tous les termes saisis dans une requête appartiennent au dialecte de l'endogroupe et sont sans ambiguïté (chaque terme ne désigne qu'un seul et unique concept dans l'ontologie considérée[4]) ;
- tout concept recherché est différent du concept universel et appartient à la liste des concepts de l'ontologie ;
- l'ontologie est complète et validée par les membres de l'endogroupe ;
- afin de personnaliser la recherche d'information, nous fixons des valeurs seuils respectifs pour les gradients de prototypicalité conceptuelle et lexical (valeur seuil $\lambda \in [0, 1]$, en dessous desquels les concepts et termes ne sont pas pris en compte).

[4] Si tel est le cas, nous présupposons que l'utilisateur fixe le concept parmi les éventuels prétendants via une session d'interaction.

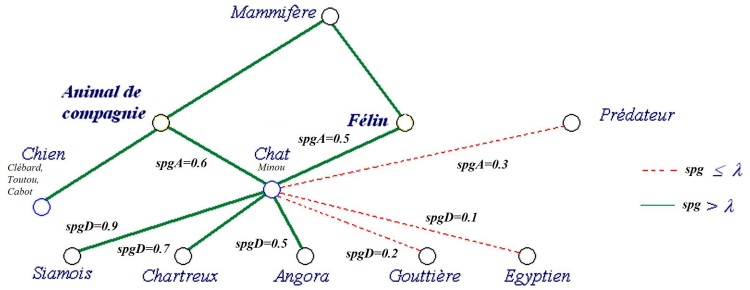


FIG. 2 – *Hiérarchie de concepts*

Le parcours dans la descendance du concept traité est un parcours de graphe en profondeur d'abord et guidé par les valeurs de gradients de prototypicalité[5].

Le processus de recherche d'information s'effectue en plusieurs étapes : (1) identification du parcours par rapport à la requête initiale en fonction de la valeur des gradients de prototypicalité conceptuelle, (2) identification et évaluation des documents pertinents par rapport à chaque concept étudié lors du parcours, et (3) tri et restitution des résultats en fonction de la valeur des gradients de prototypicalité conceptuelle (plus un sous-concept est prototypique, plus la quantité de documents afférents sera importante).

3.2 Recherche sur un seul concept

L'extension de la requête va consister non seulement à effectuer une recherche sur ce concept, mais également sur toute sa descendance, ainsi que sur le (ou les) concept(s) père(s). Si nous prenons l'exemple illustré par l'extrait de la hiérarchie conceptuelle de la figure 2, une requête sur le concept "*Chat*" sera étendue aux concepts "*Siamois*", puis "*Chartreux*" puis "*Angora*" (espèces de chats), classés par ordre de prototypicalité conceptuelle descendante (spg_D) décroissante suivant l'ontologie de l'endogroupe. Elle sera ensuite étendue aux concepts de "*Animaux de compagnie*" et de "*Félins*" (les chats sont une sous-catégorie d'animaux de compagnie, et aussi une sous-catégorie de félins), par ordre de prototypicalité conceptuelle ascendante (spg_A) décroissante.

D'un point de vue formel, toute requête portant sur le concept c_1 tel que $c_1 \in C$ et $c_1 \neq universel$ est traduit par une recherche :

- sur le concept c_1 ;
- puis sur toute la descendance du concept c_1, tel que $spg_D(c_p, c_f) > \lambda_1$, avec $\leq^C (c_p, c_f)$, c_p, c_f appartenant à la descendance de c_1 ;
- sur tout concept c_{pi} père de c_1 et tel que $spg_A(c_1, c_{pi}) > \lambda_2$.

Chaque sous-concept c_f du concept c_p est pris par ordre décroissant de la valeur de spg.

[5]Il n'est possible, à partir d'un concept, d'atteindre un autre concept que si la valeur du gradient de prototypicalité conceptuelle est supérieure au seuil fixé.

3.3 Recherche sur deux concepts ascendants

Deux cas de figure peuvent se présenter, il peut s'agir : (1) soit de désambiguïser le concept le plus spécifique en précisant sa catégorie, (2) soit de rechercher un domaine général, dont un cas particulier. Si nous prenons l'exemple illustré de la figure 2, une requête sur les concepts "*Chat*" et "*Félin*" sera étendue aux concepts "*Siamois*", puis "*Chartreux*" puis "*Angora*" (espèces de chats, classés par ordre de prototypicalité conceptuelle descendante décroissante suivant l'ontologie de l'endogroupe), enfin une recherche sur le concept "*Félin*" sera effectuée.

D'un point de vue formel, toute requête portant sur les concept $(c_1, c_2) \in \mathcal{C} \times \mathcal{C}$ tel que $c_i \neq universel$ et $\leq^C (c_1, c_2)$ est traduit par une recherche :

- sur le concept c_2 ;
- puis sur toute la descendance du concept c_2, tel que $spg_D(c_p, c_f) > \lambda$, avec $\leq^C (c_p, c_f)$, c_p, c_f appartenant à la descendance de c_2 ;
- puis sur le concept c_1 parent de c_2.

Chaque sous-concept c_f du concept c_p est pris par ordre décroissant de la valeur de spg.

Plusieurs points méritent d'être étudiés. Tout d'abord, le concept c_1 n'est pas forcément père du concept c_2 ; il peut exister une chaîne de longueur supérieure strictement à 1 entre ces deux concepts. Si la valeur du gradient de prototypicalité conceptuelle descendant (spg_D) entre c_1 et c_2 est inférieure à un seuil fixé, alors nous pouvons estimer qu'il n'y a pas de véritable représentativité entre ces deux concepts. En ce cas, nous pouvons soit considérer ces deux concepts comme distincts (cf. section 3.6), soit privilégier le concept le plus spécifique et le traiter comme un concept seul (cf. section 3.2). Dans le cas contraire (*i.e.* si la valeur du spg_D est supérieure à ce seuil), il ne paraît pas forcément pertinent de prendre en compte tous les concepts présents sur le chemin le plus court entre les deux concepts, sous peine de tomber dans un excès d'information néfaste à l'utilisateur. Enfin, il peut être pertinent de s'interroger sur les volontés de l'utilisateur : s'agit-il d'une volonté de spécialisation ou de généralisation. Dans notre approche, nous faisons le pari de cette seconde hypothèse en privilégiant le concept le plus spécifique. Dans le cas de la première, cela reviendrait à privilégier le concept situé le plus haut dans la hiérarchie et à ne prendre que les concepts situés sur la chaîne entre les deux concepts.

3.4 Recherche sur deux concepts frères

Il s'agit d'une recherche sur deux concepts - et leur(s) domaine(s) commun(s) - qui peut être sujette à plusieurs interprétations possibles. Si nous prenons l'exemple illustré de la figure 2, une requête sur les concepts "*Animal de compagnie*" et "*Félin*" peut être étendue au moins de deux manières. Soit une extension au concept de "*Chat*" (seul *concept fils commun* de ces deux concepts) puis aux concepts de "*Siamois*", puis "*Chartreux*" puis "*Angora*" (espèces de chats, classés par ordre de prototypicalité conceptuelle décroissante suivant l'ontologie de l'endogroupe), mais également au concept de "*Mammifère*" (seul *concept père commun* de ces deux concepts). Soit nous supposons que l'utilisateur énonce deux exemples de ce qu'il cherche (et qu'il veut avoir dans ses résultats tous les fils de cette catégorie), auquel cas nous faisons une extension avec l'ensemble des sous-concepts de félins, et une deuxième avec l'ensemble des sous-concepts d'animaux de compagnie.

Pour la première hypothèse et d'un point de vue formel, toute requête portant sur les concept $(c_1, c_2) \in \mathcal{C} \times \mathcal{C}$ tel que (1) $c_i \neq universel$ et (2) il existe au moins un concept $c_0 \in \mathcal{C}$ avec $\leq^C (c_0, c_1)$ et $\leq^C (c_0, c_2)$, est traduit par une recherche :

- sur le(s) concept(s) c_0 (père(s) commun(s) aux deux concept(s)) ;
- sur les concepts c_1 et c_2 ;
- puis sur tous les concepts fils communs à c_1 et c_2 et leur descendance ;
- puis la descendance propre à c_1, puis la descendance propre à c_2.

Pour la seconde hypothèse et d'un point de vue formel, toute requête portant sur les concepts $(c_1, c2) \in \mathcal{C} \times \mathcal{C}$ tel que (1) $c_i \neq universel$ et (2) il existe au moins un concept $c_0 \in \mathcal{C}$ avec $\leq^C (c_0, c_1)$ et $\leq^C (c_0, c_2)$, est traduit par une recherche sur le(s) concept(s) c_0 (père(s) commun(s) aux deux concept(s)), puis sur les concepts c_1, c_2 et autres frères issus de leur(s) père(s) commun(s) avec leur descendance.

3.5 Recherche sur deux concepts parents non ascendants et non frères

La recherche s'effectue sur deux concepts qui peuvent être cousins ou oncle-neveu (à des degrés de parenté diverses). Il peut être dès lors intéressant d'analyser leur position par rapport au plus petit concept père commun[6] (ppcpc), rechercher s'ils sont quasi-frères ou quasi-père ; c'est ce que nous appellerons la recherche de similarité sémantique équilibrée. Cette recherche s'effectue sur les concepts $(c_1, c2) \in \mathcal{C} \times \mathcal{C}$ tels que il existe un concept c différent de universel, et tel que $c = ppcpc(c_1, c_2)$. Nous allons calculer ainsi deux valeurs de la mesure de similarité sémiotique SEMIOSEM : $SemioSem(c_1, c)$ la similarité sémantique entre c_1 et c, et $SemioSem(c_2, c)$ la similarité sémantique entre c_2 et c. Trois cas peuvent se présenter : (1) soit $SemioSem(c_1, c) \approx SemioSem(c_2, c)$, en ce cas nous nous ramenons au cas où nous avons quasiment deux concepts frères (*cf.* section 3.4), les deux concepts ressemblent tous deux à leur ancêtre ; (2) soit $SemioSem(c_i, c) \approx 0$ et $SemioSem(c_j, c) \approx 1$ (*i.e.* un des concepts est très proche du plus petit concept père commun, l'autre éloigné), auquel cas nous sommes ramené dans la situation où nous avons presque un concept père de l'autre, un quasi *is-a* (*cf.* section 3.3) ; (3) soit, enfin, aucun de ces deux cas n'est rencontré et nous considérons ces concepts comme distincts (*cf.* section 3.6). Dans le cas où nous avons deux concepts quasi frères, nous étendons la requête à la fratrie de chaque concept et au plus petit concept père.

3.6 Recherche sur deux concepts distincts

Le plus petit père commun est le concept universel, alors nous considérons chaque concept séparément, indépendamment.

4 Conclusion

Notre mécanisme d'extension de requête exploite toute la richesse des relations sémantiques offertes par les ontologies. Son adaptation à l'utilisateur par le biais des prototypicalités (représentées, pour la prototypicalité conceptuelle, par des pondérations des liens hiérar-

[6] Le parcours de graphe pour la recherche du plus petit père commun se fait uniquement sur des arcs tels que $spg_D(c_p, c_f) > \lambda$, afin de simplifier de manière non négligeable le treillis.

chiques et des propriétés, et, pour la prototypicalité terminologique, par des pondérations sur les termes) permet de personnaliser tant l'extension de la requête saisie que la quantité de résultats fournis. Ce nouveau processus a pour effet d'augmenter le rappel (nous récupérons davantage de documents) et la précision (nous limitons le nombre de documents pertinents non retournés), et de fournir des résultats différents pour des utilisateurs distincts ayant soumis une même requête.

Références

Aimé, X., F. Fürst, P. Kuntz, et F. Trichet (2009a). Gradients de prototypicalité appliqués à la personnalisation d'ontologies. In F. Gandon (Ed.), *IC 2009 : Actes des 20es Journées Francophones d'Ingénierie des Connaissances IC 2009*, pp. 241–252. PUG. ISBN 978-2-7061-1538-7. Papier primé.

Aimé, X., F. Furst, P. Kuntz, et F. Trichet (2009b). Semiosem : A semiotic-based similarity measure. In R. Meersman, P. Herrero, et T. Dillon (Eds.), *On the Move to Meaningful Internet Systems : OTM 2009 Workshops*, Volume 5872 of *Lecture Notes in Computer Science*. Springer. ISBN 978-3-642-05289-7.

Guelfi, N., C. Pruski, et C. Reynaud (2007). Les ontologies pour la recherche ciblée d'information sur le web : une utilisation et extension d'owl pour l'expansion de requêtes. In *18èmes journées francophones d'Ingénierie des Connaissances, IC'2007, Plate Forme de l'AFIA, Grenoble*.

Messai, N., M. Devignes, A. Napoli, et M. Smail-Tabbone (2006). Treillis de concepts et ontologies pour interroger l'annuaire de sources de données biologiques bioregistry. *Ingénierie des Systèmes d'Information : Systèmes d'information spécialisés 11*(1), 39–60.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology* (7), 532–547.

Zaher, L., A. Bénel, J. Cahier, R. E. Sawda, et M. Zacklad (2007). Digital identities and management of identifiers for a socio-semantic web. In M. S. Bouhlel et B. Solaiman (Eds.), *Proceedings of the 4th international conference on Sciences of Electronics, Technologies of Information an Telecommunication, SETIT*. ISBN 978-9973-61-475-9.

Summary

Traditional approaches of information retrieval are primarily founded on the detection of presence of keywords in documents. Our work aims at extending the requests from a end-user, by widening the research field with personalized domain ontologies. With the gradients of prototypicalities (conceptual prototypicality and lexical prototypicality), we personalize for a user the extension of the request and the quantity of provided results. This new process (1) increases the recall and the precision, and (2) provides different results for distinct users having the same request.

Système de personnalisation web basé sur la construction d'un réseau d'utilisateurs

Hadjouni Myriam*, Baazaoui Hajer*, Aufaure Marie Aude**, Ben Ghezala Henda*
* Laboratoire RIADI-GDL, Ecole Nationale des Sciences de l'Informatique,
Université de La Manouba, 2010 la Manouba, Tunisie
{myriam.hadjouni, hajer.baazaouizghal, henda.benghezala}@riadi.rnu.tn
** Laboratoire MAS, SAP BusinessObjects Chair, Ecole Centrale Paris,
Grande Voie des Vignes, 92 295 Chatenay-Malabry, France
marie-aude.aufaure@ecp.fr

Résumé. Cet article présente un système de personnalisation web basé sur la construction d'un réseau de modèles utilisateurs. Ce réseau est construit en fonction de distances spatiales et sémantiques pouvant exister entre les modèles utilisateurs. L'idée de base est d'utiliser le réseau de modèles en le combinant aux modèles individuels des utilisateurs pour améliorer les résultats fournis à ces derniers. Dans notre approche, le modèle utilisateur est représenté par quatre dimensions corrélées et le réseau est construit itérativement.

1 Introduction

Compte tenu de la croissance continue du nombre et du type de 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 [Zemirli et al.2005]. La personnalisation a donc pour objectif d'améliorer la performance de l'extraction des informations selon la perception et les intérêts de l'utilisateur final ainsi que ses préférences. Dans cette vision, cet article présente un système de recherche d'information (SRI) pour la personnalisation sur le Web qui est basé sur la modélisation de l'utilisateur. Ce système effectue une modélisation multidimensionnelle de l'utilisateur et lui fourni un résultat personnalisé qui satisfait au mieux ses besoins (1) en fonction de ses préférences et (2) en considérant les préférences d'autres utilisateurs de ce même système. En effet, nous construisons un réseau d'utilisateurs reliés entre eux par leurs centres d'intérêts, leurs recherches (sémantiques) et leurs intérêts spatiaux/géographiques dans l'objectif de faire profiter de manière dynamique l'utilisateur en cours des recherches et intérêts des utilisateurs du système. En effet, le réseau proposé est construit au fur et à mesure des navigations des utilisateurs et calcule les distances sémantique et spatiale pouvant exister entre l'utilisateur courant et les autres nœuds du réseau.

L'article est organisé comme suit: la seconde section concerne un aperçu de l'état de l'art. La section suivante présente le système proposé ainsi que ses différents composants. Dans la section 4, nous présentons le prototype développé ainsi que l'expérimentation. Nous concluons ensuite notre travail dans la section suivante.

2 La personnalisation en recherche d'informations

La personnalisation Web inclut de plus en plus de domaines de recherche tels que les réseaux sociaux, le data mining du Web, le Web usage mining et la modélisation de

l'utilisateur [Kelly and Teevan2003]. Nous détaillons ci-dessous quelques techniques de personnalisation relevant du domaine de la RI.

2.1 Méthodes de personnalisation

Une des principales méthodes de personnalisation citées dans la littérature est la reformulation de la requête : génération d'une nouvelle requête avec les préférences de l'utilisateur. Une autre méthode est l'enrichissement de la requête qui consiste à doter cette dernière de termes prédéfinis et existant dans le profil. [Gauch et al.2003] ajoutent à la requête les préférences implicites de l'utilisateur. [Messai et al.2005] enrichissent les requêtes en leur ajoutant de nouvelles propriétés à partir des ontologies de domaine disponibles. Le point fort de l'enrichissement de la requête est de traiter l'information au niveau sémantique mais elle présente un problème de temps d'exécution à grande échelle. D'autres méthodes de personnalisation se basent sur des techniques de filtrage. En effet, nous avons en premier lieu le filtrage collaboratif ou social qui est basé sur les relations existantes entre le profil de l'utilisateur en cours et les profils d'autres utilisateurs [Mobasher2005]. Son point fort est que le profil du groupe est construit à partir des documents consultés, jugés et annotés. Mais, à grande échelle, ce type de filtrage manque de performance. Ensuite, vient le filtrage explicite : les informations utilisées sont celles fournies par l'utilisateur. A cours terme, la recherche est proche des vœux de l'utilisateur mais avec le temps, le profil tend à être statique. Et le filtrage implicite où le profil dynamique de l'utilisateur est utilisé. Ce profil est constitué de l'analyse du comportement et des activités utilisateur. Le recours au profil dynamique décharge l'utilisateur d'une formulation explicite de ses préférences, mais les éventuelles évolutions du comportement de celui-ci doivent être prises en compte.

Ces méthodes de personnalisation ont toutes le même objectif: améliorer la qualité des résultats de recherche fournis à l'utilisateur. Cet objectif implique la prise en compte de ce dernier dans le processus d'extraction de l'information d'où la nécessité d'en avoir une connaissance et de la modéliser. La section suivante présente la modélisation de l'utilisateur.

2.2 La modélisation de l'utilisateur

En 1996, Hook [Hook1996], a présenté le modèle utilisateur comme étant "une connaissance à propos de l'utilisateur explicitement ou implicitement codée, utilisée par le système afin d'améliorer son interaction". L'utilisateur peut être modélisé selon des points de vue différents ; la caractéristique la plus utilisée est la connaissance de ce dernier. La façon la plus simple de gérer les connaissances est de mémoriser ce que l'utilisateur connaît ou ne connaît pas. Pour cela, on utilise souvent soit un stéréotype (modèle de groupe) soit un modèle de recouvrement (modèle individuel). [Fink and Kobsa2002] ont proposé un système générique de modélisation de l'utilisateur pour la personnalisation des services du tourisme qui analyse les comportements de l'utilisateur et effectue des prédictions quant aux actions futures de ce dernier. Par ailleurs, [Razmerita2003] a proposé un système pour la modélisation des utilisateurs à base d'ontologie qui acquiert les données par l'intermédiaire d'un éditeur de profil utilisateur (de façon explicite), ou en utilisant différentes techniques de modélisation de l'utilisateur (techniques à bases d'heuristiques et de logique floue). [Bohnert2008], quant à eux, incluent dans leur système de visite de musées des approches de modélisation collaborative et basées sur le contenu afin de modéliser les intérêts utilisateur. L'utilisateur peut aussi être modélisé de manière ensembliste : le profil est formalisé en vecteurs de

termes pondérés ou en classes de vecteurs. Cette modélisation est simple à mettre en œuvre mais manque de structuration et ne facilite ni la prise en compte ni l'interprétation des différents niveaux de généralités caractérisant l'utilisateur. Une mise en évidence des relations sémantiques entre les informations du profil est alors nécessaire, ce que propose de faire la modélisation sémantique qui est principalement basée sur l'utilisation des ontologies [Blanco-Fernandez et al.2008]. Elle contourne les ambigüités sémantiques mais ne se sert pas des caractéristiques de la structure hiérarchique pour capturer la dynamique des changements. Un troisième type de modélisation se propose de structurer le profil selon un ensemble de dimensions représentées selon divers formalismes: c'est la modélisation multidimensionnelle [Kostadinov et al.2007]. Ses points forts sont une meilleure interprétation de la sémantique du profil et son indépendance de tout type d'application. Quant au profil, un mélange entre ses données d'évolution et ses données persistantes a été constaté, d'où un manque d'interprétation du rôle de chaque dimension.

2.3 Synthèse

Le processus de personnalisation dans les systèmes de recherche d'information est principalement confronté à la question de la définition des informations nécessaires concernant l'utilisateur. En effet, la question de savoir comment représenter ce qui caractérise l'utilisateur, et comment l'utiliser dans le processus de l'extraction de l'information est toujours d'appoint. L'introduction de l'utilisateur dans ce processus nécessite une modélisation de ce dernier et une fiabilité de son profil [Kobsa2005]. En effet, on constate que l'une des principales raisons du manque de performances des techniques de personnalisation est typiquement l'application d'un profil utilisateur hors contexte [Gauch et al.2007]. Les utilisateurs peuvent avoir des préférences générales, récurrentes et stables. Cependant, l'ensemble des informations contenues dans le profil n'est pas forcément approprié à toutes les situations de recherche. Le plus souvent, les systèmes n'utilisent seulement qu'un sous-ensemble de ces informations, qu'ils supposent pertinents pour la recherche en cours.

3 Système de personnalisation web basé sur la construction d'un réseau d'utilisateurs

Nous proposons dans cette section le système de personnalisation web basé sur la construction d'un réseau d'utilisateurs. Ce réseau a pour objectif de faire profiter l'utilisateur courant des résultats de recherche et des modèles des autres utilisateurs. Ce système intègre :

- La construction itérative d'un réseau d'utilisateurs
- Une modélisation multidimensionnelle de l'utilisateur. Ainsi que la modélisation d'un réseau d'utilisateurs construit à partir de modèles utilisateurs.
- La combinaison de trois modes de recherche : écriture d'une requête textuelle, positionnement sur une carte géographique ou sélection d'images représentatives des centres d'intérêt de l'utilisateur.

Les résultats affichés à l'utilisateur sont fournis en fonction: des ses anciennes navigations, des ses déplacements spatiaux, de son modèle utilisateur ainsi que des informations sur les autres utilisateurs. Le résultat fourni à l'utilisateur est une liste "ordonnée" d'objets, positionnée sur la carte avec une mise en valeur des degrés de pertinence des informations fournies. Nous mettons en place un système de retour d'information (feedback) qui nous permet

d'évaluer la pertinence des résultats fournis. Cette évaluation est faite de manière implicite par un calcul d'intérêt sur les résultats sélectionnés [Hadjouni et al.2009a].

3.1 Construction du réseau d'utilisateurs

Le système repose sur la construction d'un réseau basé sur des modèles des utilisateurs. Notre hypothèse est que, lors de la recherche d'un document pertinent, le système devrait, en plus des besoins spécifiques des utilisateurs et de leurs recherches antérieures, exploiter les connaissances extraites des autres modèles utilisateurs existants. Les nœuds du réseau représentent des modèles d'utilisateurs qui sont interconnectés par des distances spatiales et sémantiques (cf. fig 1). Un tel réseau permet aux utilisateurs (a) d'avoir des résultats correspondant à leurs propres préférences (implicites) et (b) de bénéficier des résultats du voisin le mieux correspondant sémantiquement (via les critères de recherche) et spatialement (à travers les différentes positions spatiales de recherche).

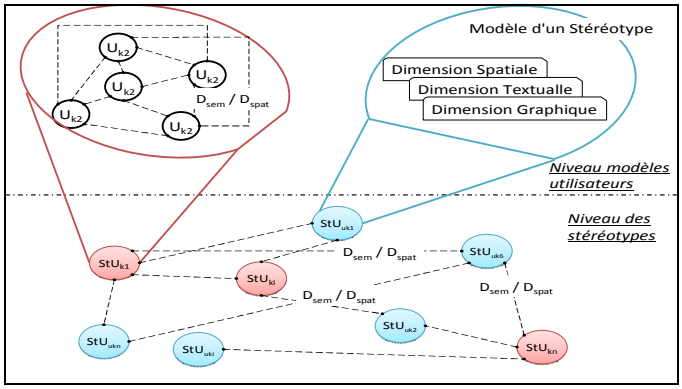


FIG. 1 Le réseau d'utilisateurs. Sur cette figure, les deux niveaux sont mis de manière superposée pour montrer la relation de contenance existant entre eux.

La contribution de notre proposition réside dans la construction d'un réseau sur deux couches: une première couche donnant accès aux modèles des utilisateurs connus, et une seconde couche pour les modèles des stéréotypes des utilisateurs inconnus. Les arcs séparant les nœuds de ce réseau sont calculés en fonction de distances sémantique et spatiales [Hadjouni et al.2009b]. Le choix de ces deux distances est basé sur le fait qu'un utilisateur effectue généralement une recherche textuelle et peu avoir besoin d'informations d'ordre spatial.

3.2 Construction du modèle utilisateur

La construction du modèle de l'utilisateur est basée sur la prise en compte implicite et interactive des données enregistrées à partir des navigations de l'utilisateur à travers le système. Ces données sont dites implicites car l'utilisateur n'est pas amené à fournir des informations sur ses préférences, et interactives car nous utilisons l'historique de la navigation pour mesurer son intérêt pour une entité donnée. Ces mesures sont basées sur:

- Les similarités pouvant exister entre les attributs et les entités d'intérêt : Pour les similarités entre les valeurs des attributs, nous distinguons les différents types qui peuvent exister (les valeurs numériques, des intervalles numériques, les ensembles ...). Pour le

calcul de la similarité entre entités, nous considérons l'agrégation des degrés de similarités existant entre les attributs de ces entités.

- La déduction des intérêts des utilisateurs à partir de l'ensemble de leurs navigations : nous considérons pour cela la fréquence et les durées des visites effectuées, que nous avons appelés indicateurs d'intérêt [Hadjouni et al.2009a].
- Le calcul de la pertinence d'un éventuel déplacement spatial : Dans notre approche, nous considérons également que si un utilisateur demande ou sélectionne une zone spatiale, c'est qu'il a l'envie de s'y déplacer. Cette déduction est effectuée en corrélation avec la déduction des intérêts des utilisateurs à partir de l'ensemble de leurs navigations.

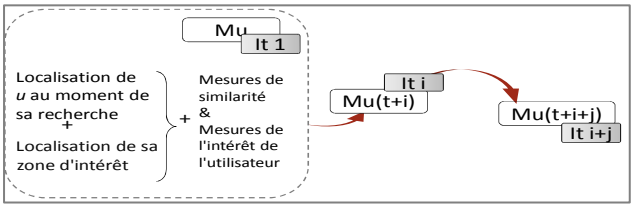


FIG. 2 Construction du modèle de l'utilisateur

La construction du modèle utilisateur est effectuée de manière itérative. Les itérations débutent avec la première interaction de l'utilisateur et la construction du modèle évolue avec les recherches et les navigations au cours de la session de recherche (figure 2). Deux cas de figures, pour un utilisateur, se présentent : soit qu'il est connecté au système à travers ses nom d'utilisateur est mot de passe, et de ce fait il est reconnu, soit qu'il commence sa session de travail de manière anonyme, nous disons alors qu'il est inconnu du système. Pour le premier cas, le niveau modèle utilisateur du système active le modèle de l'utilisateur correspondant et en effectue la mise à jour au cours de la navigation. Si l'on considère M_u $(t_0 = 0)$ ce modèle utilisateur au moment de l'activation ($t_0 = 0$), nous aurons à $(t_1 = t_0+\lambda)$:

$$M_u(t_1=t_0+\lambda)=M_u(t_0)+Z_i+Sim(x)+I_e(u,x) \qquad (1)$$

Avec:

Z_i	La zone de recherche cible de l'utilisateur.
$Sim(x)$	La similarité pouvant exister entre les attributs des résultats de la recherche et les entités d'intérêt x [Hadjouni et al.2009b].
$I_e(u,x)_i$	La déduction des intérêts de l'utilisateur.

4 Expérimentation

La phase d'expérimentation présentée porte sur la construction des stéréotypes utilisateurs. L'évaluation a été effectuée sur la base de la navigation des utilisateurs à travers la liste des résultats affichés et sur les différentes localisations spatiales choisies. L'analyse effectuée pour l'étude des relations entre ces utilisateurs est une analyse conceptuelle à travers les treillis de gallois [Barbut and Monjardet1970]. Notre objectif est de visualiser et de comprendre la composition des stéréotypes construits par le biais du calcul d'*empreintes conceptuelles* à partir de treillis de Galois. Ces empreintes vont nous aider à comprendre la structure et les propriétés des données extraites des requêtes des utilisateurs étudiés.

Soit (O, A, I) le contexte correspondant à un treillis de gallois. Selon la terminologie de [Wille1992], O est l'ensemble des objets, A l'ensemble des propriétés de O et I est la relation

binaire ente O et A; *($I \subseteq OxA$).* Les notions d'objet et de propriété dans notre contexte représentent respectivement les utilisateurs et leurs préférences. Le treillis de Galois est donc constitué de concepts comprenant des ensembles de personnes décrits par leurs préférences.

Pour cette expérimentation, nous avons utilisé un échantillon de 83 utilisateurs. Leurs préférences ont été traduites en relations binaires, avec:

O = {utilisateur1,…, utilisateur83}

A = {Hôtel, Restaurant, Auberge, Sport Nautique, Magasin, Banque, Tunis, Madrid, Sousse, Paris, Avion, 5*, musées, internet, Proche aéroport, 5*, Sauna, Zone touristique, internet}

La figure suivante montre le treillis construit:

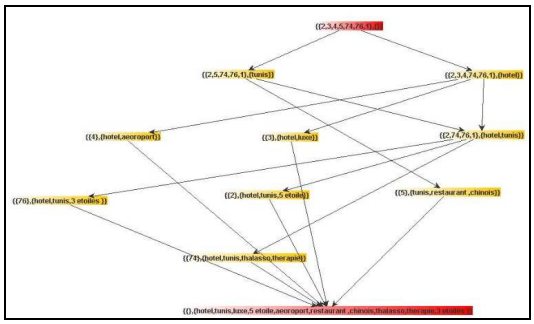


FIG. 3 Visualisation du treillis

L'objet $o \in O$ est caractérisé par deux paramètres appelés *Relatedness* (lien) et *Closeness* (proximité) [Le Grand et al.2009]. Dans notre contexte, la *Relatedness* indique si l'utilisateur considéré possède des préférences communes avec d'autres utilisateurs du système. Ce paramètre sert donc à exprimer le degré de relation d'un utilisateur avec les autres utilisateurs, et ceci d'un point de vue quantitatif. Quant à la *Closeness*, elle sert à nous informer sur la quantité de propriétés communes ou non, c'est un paramètre qui exprime d'un point de vue qualitatif la connectivité de l'objet *o* avec les autres objets. De ce fait, un utilisateur ayant une *Relatedness* élevée a plusieurs préférences en commun avec d'autres utilisateurs du système. S'il a aussi la valeur de la *Closeness* élevée, c'est qu'il possède de nombreuses propriétés en commun avec les utilisateurs du concept auquel il a été attribué.

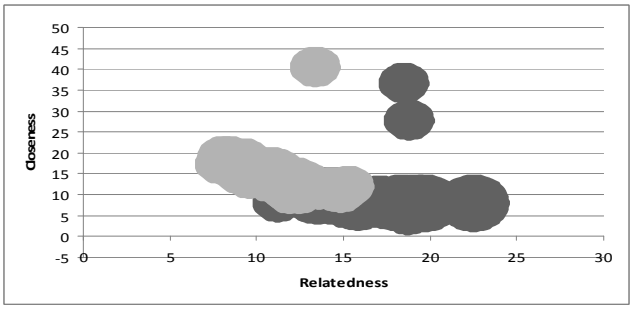


FIG. 4 Evaluation des Closeness et Relatedness

La distribution des utilisateurs est représentée dans la figure 4 où chaque utilisateur est représenté par une bulle. La *Relatedness* est représentée dans l'axe des abscisses et la *Closeness* dans l'axe des ordonnées. Ces bulles sont homogènes et regroupées; cela dénote d'un rapprochement entre les utilisateurs du système. La distribution obtenue fait ressortir deux

stéréotypes assez proches et trois utilisateurs indépendants, ils forment ainsi cinq nœuds du réseau. Ces résultats ayant été obtenus en ayant considéré les préférences implicites, les différentes sélections spatiales, et la requête courante des utilisateurs du jeu de tests. Cette distribution confirme la possibilité de construire un réseau en se basant sur des informations de la recherche actuelle de l'utilisateur en cours ainsi que de données implicites.

5 Conclusion

Dans cet article, nous avons commencé par présenter une synthèse de l'état de l'art dans le domaine de la recherche d'information personnalisée et de la modélisation de l'utilisateur. Ceci a conduit à la proposition d'un système de RI pour la personnalisation sur le Web basé sur la construction d'un réseau d'utilisateurs ainsi que sur la modélisation de l'utilisateur. En effet, le système proposé intègre la construction de la modélisation utilisateur ainsi que la constitution d'un réseau de modèles utilisateurs dans l'objectif de fournir aux utilisateurs des résultats de recherche personnalisés. Un processus de construction itérative du modèle utilisateur basé sur des évaluations des résultats a été ensuite détaillé. Enfin, nous avons procédé à une première phase d'expérimentation qui consiste en la construction des stéréotypes d'utilisateurs. Cette première phase a permis, en fonction des modèles des utilisateurs, de faire ressortir les nœuds du réseau: nœud connus et nœuds de stéréotypes. Ce regroupement nous montre que l'utilisation d'un réseau construit dynamiquement et itérativement améliore les résultats retournés à l'utilisateur.

Actuellement, nous élargissons le champ des tests en déployant le système sur le Web afin d'avoir un réseau d'utilisateurs plus large et de terminer avec des tests en temps réel. Ces expérimentations permettront de calculer les mesures sur la satisfaction des utilisateurs du web et de travailler sur la construction des modèles utilisateurs à grande échelle.

Références

Barbut, M. and Monjardet, B. (1970). Ordre et classification, algèbre et combinatoire, T.2.

Blanco-Fernandez, Y., Pazos-Arias, J., Gil-Solla, A., Ramos-Cabrer, M., and M., L.-N. (2008). Semantic reasoning: A path to new possibilities of personalization. Proceedings of the 5th European Semantic Web Conference.

Bohnert, F. (2008). Constraint-aware user modelling and personalisation in physical environments. Adjunct Proceedings of the 6th Int. Conf. on Pervasive Computing, pp167–172.

Fink, J. and Kobsa, A. (2002). User modeling for personalized city tours. Artificial Intelligence Review, 18(1) 4: pp 33–7.

Gauch, S., Chaffee, J., and Pretschner, A. (2003). Ontology based personalized search and browsing. Web Intelligence and Agent Systems, 1,pp 219–234.

Gauch, S., Speretta, M., Chandramouli, A., and Micarelli, A. (2007). User profiles for personalized information access. In The Adaptive Web: Methods and Strategies of Web Personalization, v. 4321 of Lecture Notes in Computer Science, ch 2, pp 54–89.

Hadjouni, M., Haddad, M. R., Baazaoui, H., Aufaure, M. A., and Ben Ghezala, H. (2009a). Personalized information retrieval approach. In Web Information Systems Modeling, in conjunction with the 21st Int. Conf. on Advanced Information Systems: CAiSE 2009.

Hadjouni, M., Haddad, M. R., Baazaoui, H., Aufaure, M. A., and Ben Ghezala, H. (2009b). A spatially enhanced web personalization approach. Second International Conference on Web and Information Technologies, in cooperation with ACM SIGAPP.

Hook, K. (1996). A Glass Box Approach to Adaptive Hypermedia. PhD thesis, Stockholm university.

Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference: a bibliography. SIGIR Forum 37(2), pp 18-28.

Kobsa, A. (2005). User modeling and user-adapted interaction. User Model. User-Adapt. Interact., 15(1-2):185–190.

Kostadinov, D., Bouzeghoub, M., and Lopes, S. (2007). Accès personnalisé à des sources de données multiples: évaluation de deux approches de reformulation de requêtes. In INFORSID, pp 73–88.

Le Grand, B., Aufaure, M.-A., and Soto, M. (2009). Empreintes conceptuelles et spatiales pour la caractérisation des réseaux sociaux. In Extraction et gestion des connaissances (EGC'2009), Revue des Nouvelles Technologies de l'Information, V(15) : 349–354.

Messai, N., Devignes, M., Napoli, A., and Smaïl-Tabbone, M. (2005). Méthode sémantique pour la classification et l'interrogation des sources de données génomiques.Ateliers EGC 2005, Extraction des connaissances : Etat et perspectives. Ch1 : 43–47.

Mobasher, B. (2005). Web usage mining and personalization. Chapter in Practical Handbook of Internet Computing.

Razmerita, L. (2003). User Model and User Modeling in Knowledge Management Systems: An Ontology-based Approach. PhD thesis, University of Toulouse, France.

Wille, R. (1992). Concept lattices and conceptual knowledge systems. Computers and Mathematics with Applications, 23:493–515.

Zemirli, N., Tamine, L., and Boughanem, M. (2005). Accès personnalisé à l'information : Proposition d'un profil utilisateur multidimensionnel. 7th International Symposium On Programming Systems' Algiers (ISPS).

Summary

This paper aims to present a network based Web personalization system based on user modeling that provides personalized search results to the user. The basic idea is to use the users' network with the individual users' models to improve results provided to them. In our approach, the user model is represented by four interrelated dimensions and the network is built iteratively.

Proposition d'une architecture pour la personnalisation de l'information spatiale sur le Web

Mohamed Ramzi Haddad*, Hajer Baazaoui*
Marie Aude Aufaure**,***, Christophe Claramunt****
Yves Lechevallier**, Henda Ben Ghezala*

*Laboratoire RIADI-GDL, ENSI, Campus la Manouba, La Manouba, 2010, Tunisia
haddad.medramzi@gmail.com, hajer.baazaouizghal@riadi.rnu.tn, henda.benghezala@riadi.rnu.tn,
**INRIA-Rocquencourt, Domaine de Voluceau. 78 153 Le Chesnay Cedex, France
Yves.Lechevallier@inria.fr, marie-aude.aufaure@inria.fr
***Ecole Centrale Paris, MAS Laboratory.
Chaire SAP BusinessObjects Grande Voie des Vignes 92 295 Chatenay-Malabry, France
marie-aude.aufaure@ecp.fr
****Institut de Recherche de l'École navale, Lanvéoc-Poulmic, BP 600, Brest naval, France
christophe.claramunt@ecole-navale.fr

Résumé. Le contenu mis à disposition par les systèmes d'information devient difficile à explorer par les utilisateurs en quête d'informations s'alignant à leurs préférences personnelles. L'émergence de l'information spatiale, en présence d'un contenu sémantiquement complexe, a montré l'insuffisance des approches de personnalisation classiques puisqu'elles ne considèrent pas l'aspect spatial l'information. Dans ce contexte, l'application présentée dans cet article repose sur une architecture pour la personnalisation de l'information spatiale sur le Web. D'une part, l'approche utilisée introduit un modèle de données et un ensemble d'operateurs traitant de la sémantique de l'information. D'autre part, elle permet la modélisation et la prédiction des préférences de l'utilisateur. Enfin, la prise en compte de l'aspect spatial est assurée par l'injection de mesures d'accessibilité personnalisées dans le processus de recommandation. L'expérimentation de l'architecture proposée a été menée sur des données spatiales du domaine du tourisme.

1 Introduction

La personnalisation dans les systèmes d'information fait l'objet de recherches visant à fournir aux utilisateurs une réponse adaptative et intelligente améliorant ainsi leur utilisabilité. Mais, le contenu informationnel fourni devient de plus en plus géoréférencé et avec une sémantique complexe. A la différence des autres types de contenus, l'information spatiale manipulée révèle en plus de la sémantique, des associations spatiales qu'un processus de personnalisation efficace devrait prendre en considération pour simuler le processus de prise de décision de l'utilisateur et ainsi approcher le niveau de pertinence souhaité. Dans ce contexte, la notion

d'accessibilité est à la base définie et utilisée pour étudier et caractériser la distribution spatiale des objets géographiques dans l'espace ainsi que leurs propriétés. En effet, l'accessibilité d'un lieu est un facteur qui influence le processus de planification et de déplacement d'un utilisateur (Haddad et al., 2009a). Elles permettent de mesurer l'impact de la distribution des localisations (services, commerces, etc...) sur les déplacements des usagers en déterminant des modèles ou des patrons de voyages.

Cet article présente dans une première partie l'approche de personnalisation proposée et dont l'objectif est l'aider à la navigation et à la planification de voyages. La manière avec laquelle les préférences de l'utilisateur "géolocalisé" sont déterminées ainsi que les contraintes spatiales qui influencent ses déplacements sera développée. Dans la deuxième partie, le cas d'étude pris en considération ainsi que le prototype développé seront présentés.

2 Une architecture pour la personnalisation de l'information spatiale sur le Web

L'approche de personnalisation proposée repose sur trois composantes complémentaires, à savoir, la base de données, la composante de personnalisation et l'interface utilisateur par laquelle l'utilisateur interagit avec le contenu disponible. La figure 1 présente les différents éléments de l'architecture proposée.

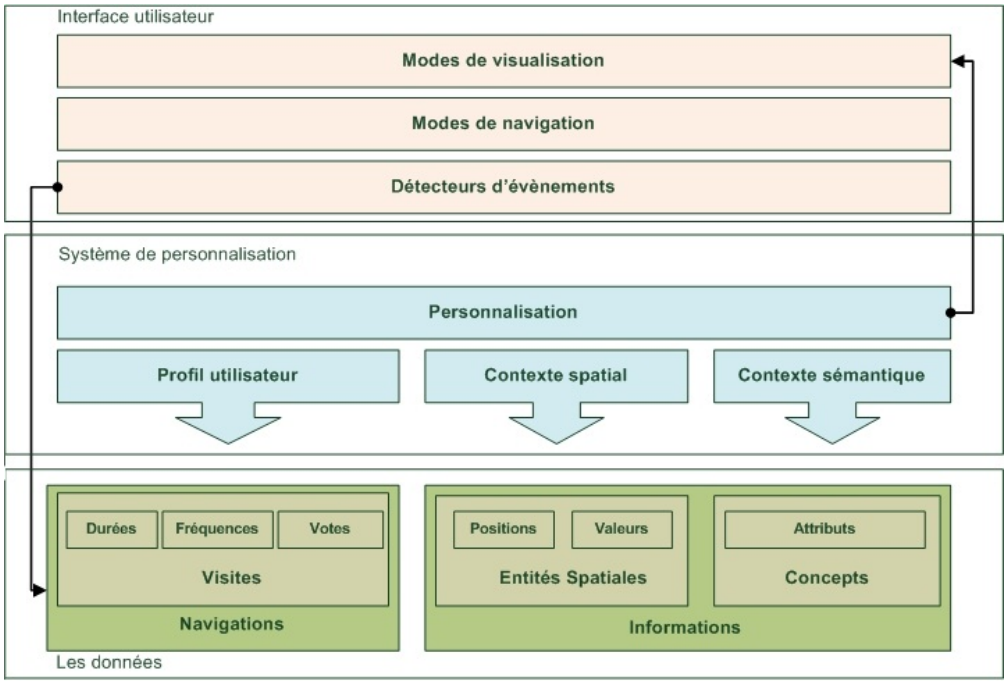


FIG. 1 – *Architecture du système proposé*

2.1 Interface utilisateur

C'est la couche frontale avec laquelle interagit l'utilisateur. Elle propose différents modes de visualisation et d'exploitation des données selon un ensemble de modes de navigation et d'interaction avec les données (liens hypertexte, spatio-carte, recherche textuelle, recherche par critères, navigation par concept, etc...). Le rôle principal de cette composante est l'alimentation du processus de personnalisation par les informations concernant le comportement de l'utilisateur lors de l'utilisation du système d'information. Les entrées de cette couche sont les interactions utilisateur/interface. En effet, elle regroupe les capteurs d'événements qui reflètent l'intérêt de l'utilisateur vis-à-vis du contenu comme, par exemple, les consultations de données, les recherches, les mouvements de souris, de barre de défilement, les votes, les commentaires, l'ajout d'un item aux favoris ou la sauvegarde de ses informations. Les sorties de cette couche sont les actions codées et datées de l'utilisateur permettant de retracer sa navigation. Ces informations alimentent la couche de données avec les connaissances nécessaires pour la détermination ultérieure des préférences de l'utilisateur par le processus de personnalisation.

2.2 Les composants de personnalisation

Plusieurs aspects doivent être pris en considération pour pouvoir bien mener le processus de personnalisation.

- Profil utilisateur : Le profil est une structure de données qui décrit les intérêts et les préférences de l'utilisateur par l'analyse de ses navigations. Ceci permet de faire la personnalisation sur la base des préférences propres de l'utilisateur (Hadjouni et al., 2009b).
- Contexte spatial : Modélise les contraintes spatiales qui régissent les déplacements de l'utilisateur dans l'espace. Ce composant regroupe des méthodes de calcul du degré de satisfaction des contraintes spatiales ainsi que des opérateurs spatiaux de sélection et de filtrage de données évaluant, par exemple, la proximité, le voisinages ou la densité spatiale. Ces notions sont intégrées dans une mesure d'accessibilité spatiale orientée utilisateur (Haddad et al., 2009b).
- Contexte sémantique : Modélise les relations entre les objets et leurs descriptions. Cette composante définit des mesures de similarités liées à la sémantique des données utilisées et permettant de mesurer la similarité entre les différentes propriétés des objets. Le rôle principal de ces mesures est la maîtrise de la sémantique de l'information et la compréhension des similitudes qui peuvent exister entre les entités (Hadjouni et al., 2009a).
- Personnalisation : Désigne le processus de personnalisation utilisant les mesures de similarité, le profil utilisateur et les opérateurs spatiaux pour prédire la navigation de l'utilisateur et faire des recommandations pertinentes par rapport aux préférences et au contexte spatial de celui-ci. Ce processus est basé sur une mesure d'accessibilité intégrant à la fois les notions de pertinence et d'analyse spatiale (Haddad et al., 2009c).

2.3 Les données

Les données requises pour la tâche de personnalisation peuvent être classées en deux catégories :

– Contenu informationnel : regroupe l'information mise à disposition de l'utilisateur ainsi que l'ensemble de métadonnées décrivant sa sémantique.
– Navigations : C'est l'ensemble de données décrivant de façon détaillée les navigations des utilisateurs. Ces informations sont issues de l'interaction de l'utilisateur avec le contenu et sont utilisées lors de l'analyse de ses navigations et de détermination de ses préférences.

3 Prototype développé

Afin d'étudier les contributions de l'approche proposée et évaluer son apport par rapport aux approches classiques, une étape de mise en IJuvre et d'expérimentation du prototype a été effectuée. Cette section commence par la présentation du jeu de données utilisé. Ensuite, la logique métier est détaillée. Enfin, les interfaces utilisateur les plus significatives sont décrites.

3.1 Jeu de donnée utilisé

Pour l'évaluation de l'approche de personnalisation proposée, un ensemble de données réelles issues d'un système d'information en ligne [1] a été utilisé. Le service principal de ce système d'information est la présentation des destinations touristiques potentielles dans la région de Nièvre en France. Les informations mises à disposition sont :

1. Un ensemble d'entités spatiales (700 points) caractérisées par :
   (a) Une position spatiale.
   (b) Un ensemble de métadonnées décrivant la sémantique du domaine (catégorie spatiale, attributs, caractéristiques).
2. Un ensemble d'utilisateurs ayant chacun visité un ensemble d'entités.

3.2 Fonctionnement du système

Le prototype développé est un site web d'e-tourisme présentant un catalogue de 700 localisations touristiques classées en un ensemble de catégories. Ce système permet la consultation des méta-données relatives aux entités spatiales, de les noter en fonction de ses intérêts, ainsi d'en sélectionner les plus pertinentes en vue d'un futur déplacement à planifier. Le système gère les utilisateurs en déterminant et en sauvegardant leurs navigations, leurs intérêts et leurs contraintes de déplacement de manière à disposer de toutes les informations requises par le système de recommandation. Le cycle de fonctionnement du système est le suivant :

1. Après l'identification, l'utilisateur est soit reconnu comme inscrit par le biais de ses identifiants, soit comme anonyme.
2. Une session de navigation est ouverte au niveau du navigateur et contient toutes les informations nécessaires concernant.
3. Le profil de l'utilisateur est construit ; il est composé d'un profil à court terme vide et qui servira à décrire la navigation courante de celui-ci. Dans le cas où l'utilisateur est déjà

1. http ://www.nievre-tourisme.com/

inscrit, le profil à court terme est complété par autre à long terme déduit depuis son historique de navigation déjà sauvegardé dans la base de données au cours des navigations précédentes.

4. À chaque sélection d'une entité, les données nécessaires sont collectées pour mettre à jour le profil et ainsi recalculer les degrés d'intérêt pour les entités spatiales existantes.
5. Les entités spatiales sont projetées dans l'espace et la mesure d'accessibilité orientée utilisateur est appliquée pour déterminer les destinations potentielles et qui respectent les contraintes de déplacement de l'utilisateur.
6. Collecte de l'information à fournir à l'utilisateur ; à savoir les résultats de sa requête et les recommandations du système.

3.3 Aperçus du prototype développé

Le prototype développé propose aux utilisateurs un ensemble d'interfaces pour interagir avec les services mis à disposition. Celles-ci sont présentées dans cette section.

3.3.1 Navigation dans une catégorie spatiale

La figure 2 présente l'interface qui permet à l'utilisateur de naviguer parmi les entités d'une catégorie spatiale sélectionnée. Les entités sont présentées par groupes de quinze images (au milieu). L'utilisateur dans ce cas n'est pas identifié, c'est pour cela une seule colonne d'entités recommandées est présente (à droite).

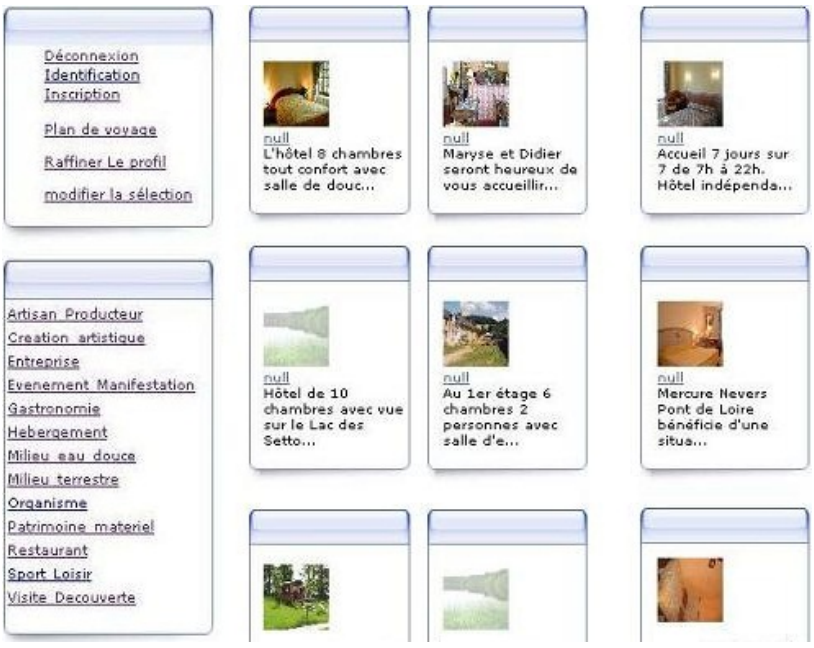


FIG. 2 – *Aperçu des entités d'une catégorie spatiale*

3.3.2 Sélection d'une entité

Lorsque l'utilisateur sélectionne une entité pour visualiser sa description détaillée, l'interface de la figure 3 se présente. Cette page comporte deux colonnes de recommandations (à droite) relatives aux profils long et court termes puisque l'utilisateur dans ce cas est identifié. Outre la description de l'entité, celui-ci dispose d'un composant de vote qui lui permet de noter son intérêt par rapport à l'information présentée. Enfin l'utilisateur dispose, pour la navigation, d'une spatio-carte affichant son emplacement (en rouge), les entités recommandées (en bleu et en vert) ainsi que les autres entités non pertinentes (en gris). la carte permet l'agrandissement, le déplacement et l'interaction directe avec les entités représentées.

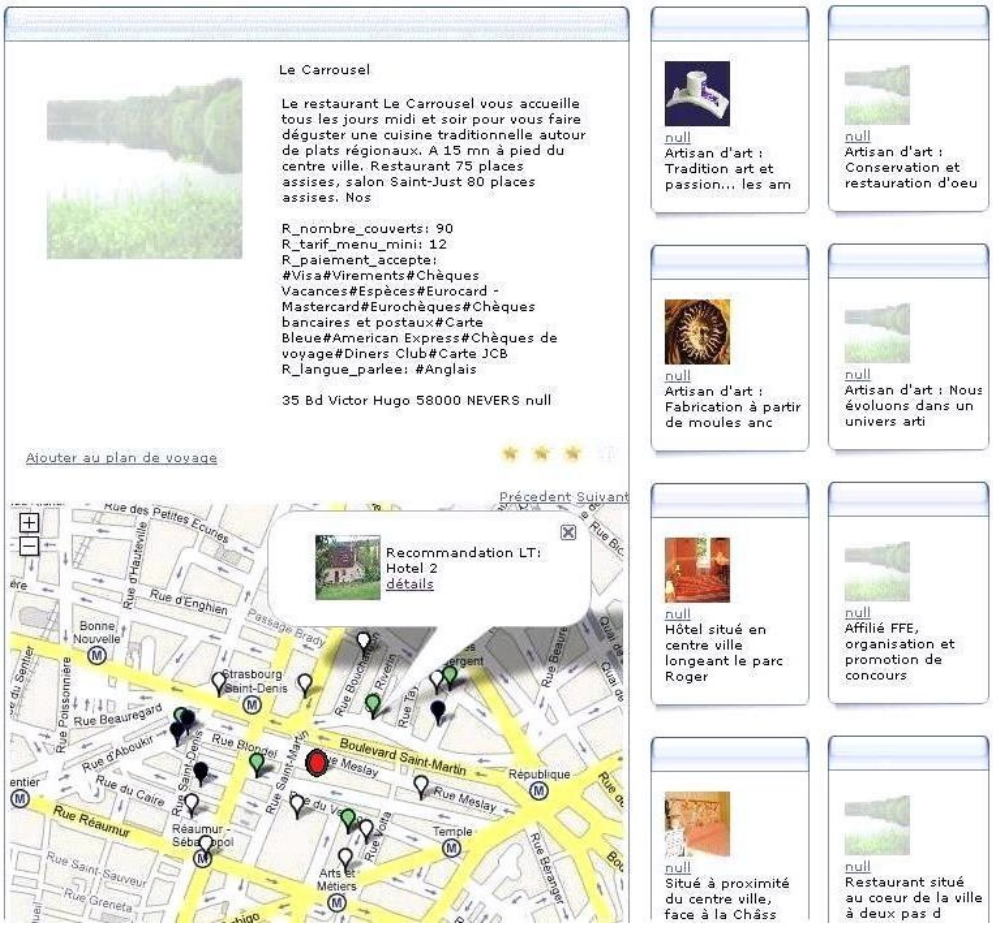


FIG. 3 – *Consultation de l'information relative à une entité spatiale*

3.3.3 Gestion du plan de voyage

En consultant la description dune entité, l'utilisateur a la possibilité de l'ajouter dans son plan de voyage. Ce dernier peut être consulté comme le montre la figure 4. Ici l'utilisateur peut supprimer une destination ou consulter la description d'une des entités déjà ajoutées.

FIG. 4 – *Gestion du plan de voyage*

4 Conclusion

L'approche proposée, permet de faire converger les notions de personnalisation et d'analyse spatiale pour améliorer la qualité de l'information fournie à l'utilisateur par la prise en compte de ses centres d'intérêts et de ses contraintes spatiales. Elle présente une solution aux limites des approches de personnalisation et des mesures d'accessibilité spatiale existantes. Cette approche peut être utilisée lors de l'étude de l'accessibilité individuelle dans un espace donné ou intégrée dans les systèmes d'information proposant des services personnalisés liés à la mobilité comme la planification de voyages ou l'aide au déplacement. Nous travaillons actuellement sur le développement d'une plateforme d'évaluation de la qualité de la person-

nalisation et de son impact sur l'utilisabilité des systèmes d'information en ligne intégrant des données géographiques.

Références

Haddad, M. R., H. Baazaoui, M.-A. Aufaure, C. Claramunt, Y. Lechevallier, et B. G. Henda (2009b). A user-oriented accessibility measure for spatial web personalization. In *Proc. European Conference on Information Retrieval (ECIR'09)*.

Haddad, M. R., H. Baazaoui, M.-A. Aufaure, et B. G. Henda (2009a). Proposition d'une mesure d'accessibilité spatiale orientée utilisateur. In *Proc. COGIST'09 conference*.

Haddad, M. R., H. B. Zghal, M.-A. Aufaure, C. Claramunt, Y. Lechevallier, et H. H. B. Ghézala (2009c). Towards an integration of space and accessibility in web personalization. In *W2GIS*, pp. 39–55.

Hadjouni, M., M. R. Haddad, H. Baazaoui, M.-A. Aufaure, et H. B. Ghezala (2009a). Personalized information retrieval approach. In *Proc. WISM'09 conference*.

Hadjouni, M., M. R. Haddad, H. Baazaoui, M.-A. Aufaure, et H. B. Ghezala (2009b). A spatially enhancedweb personalization approach. In *Proc. ICWIT'09 conference*.

Summary

The Web content made available by Information systems became difficult to explore by users searching only for relevant information respecting their personal preferences and tastes. The emergence of spatial information, in addition to the increasing semantic complexity of the available content has shown the inadequacy of the classic personalization approaches since they do not consider the spatial aspect of the data. In this context, this article presents an application that proposes an architecture for spatial Web personalization. The proposed approach is based on a data model enhanced with several operators dealing with information semantics. Moreover, a user preferences modelling process is developed. Finally, the spatial aspect of the data was considered by using a personalized accessibility measures within the approach. An experimentation of the proposed architecture was conducted on a set of spatial data from the tourism field.

WCUM pour l'analyse d'un site web

Malika Charrad*,*** Yves Lechevallier**
Mohamed Ben Ahmed*, Gilbert Saporta***

*Ecole Nationale des Sciences de l'Informatique
malika.charrad@riadi.rnu.tn,
**INRIA-Rocquencourt
yves.lechevallier@inria.fr
***Conservatoire National des Arts et Métiers
gilbert.saporta@cnam.fr

Résumé. Dans ce papier, nous proposons une approche WCUM (Web Content and Usage Mining) permettant de relier l'analyse du contenu d'un site Web à l'analyse de l'usage afin de mieux comprendre les comportements de navigation sur le site. L'apport de ce travail réside d'une part dans la proposition d'une approche reliant l'analyse du contenu à l'analyse de l'usage et d'autre part à l'extension de l'application des méthodes de block clustering, appliquées généralement en bioinformatique, au contexte Web mining afin de profiter de leur pouvoir classificatoire dans la découverte de biclasses homogènes à partir d'une partition des instances et une partition des attributs recherchées simultanément.

1 Introduction

La caractérisation des internautes fréquentant un site Web est un problème incontournable pour assister l'internaute et prédire son comportement. Ces considérations ont motivé d'importants efforts dans l'analyse des traces des internautes sur les sites Web. D'autres efforts ont été concentrés sur l'analyse du contenu des pages Web. Sachant que le comportement des utilisateurs sur un site web dépend fortement du contenu des pages du site et inversement le contenu du site devrait répondre aux attentes des usagers du site, nous proposons de faire la liaison entre le contenu et l'usage d'un site web. Notre idée est d'exploiter les différentes informations relatives au contenu d'un site Web et de son usage en vue de l'analyser. Le point de départ de cette approche est le contenu textuel du site et les fichiers logs contenant les traces des utilisateurs.

2 Approche WCUM

L'approche WCUM relie l'analyse du contenu à l'analyse de l'usage d'un site Web (fig. 1). Elle se déroule en deux principales étapes. La première consiste à l'analyse textuelle d'un site Web afin de découvrir les thèmes des pages. La seconde étape consiste à introduire ces thèmes dans l'analyse de l'usage du site. L'application de cette approche nécessite d'une part

l'aspiration du site afin de transformer ses pages en fichiers texte, et d'autre part la collecte des fichiers Logs contenant la trace des utilisateurs sur le site.

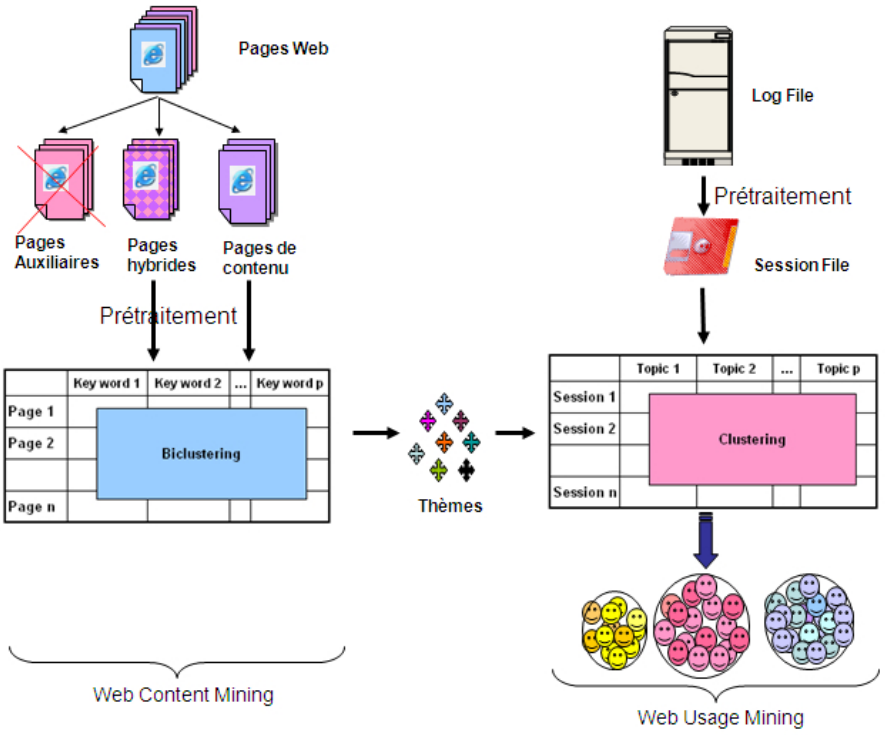


FIG. 1 – *Approche WCUM*

2.1 WCM : Analyse textuelle

2.1.1 Prétraitement des pages Web

L'analyse textuelle d'un site Web consiste en premier lieu à distinguer les pages de navigation des pages de contenu. Les pages de navigation, ou pages auxiliaires, servent à faciliter la navigation sur le site alors que les pages de contenu présentent une information éventuellement utile aux utilisateurs. Certaines pages de contenu contiennent plusieurs hyperliens permettant de naviguer sur le site. Ces pages présentant les caractéristiques communes de pages de contenu et de pages auxiliaires sont appelées "pages hybrides" . Cette étape a pour objectif d'exclure les pages auxiliaires de l'analyse et de limiter le travail de prétraitement aux pages de contenu. La classification des pages est basée sur un ensemble de variables tels que le nombre de liens entrants et sortants et la taille des documents...etc (Charrad et al., 2008). La deuxième étape consiste au prétraitement linguistique et à la sélection de descripteurs afin d'aboutir à une représentation matricielle du site. Le prétraitement nécessite tout d'abord la conversion

des pages Web en fichiers Textes, et le remplacement des images qu'ils contiennent par leurs légendes. L'étiquetage et la lemmatisation à l'aide de TreeTagger permettent de remplacer les verbes par leur forme infinitive, les noms par leur forme au singulier et certaines formes des verbes tels que les participes présents et les participes passés par leurs racines. Afin de réduire la dimension de l'espace vectoriel des vecteurs représentant les textes, il s'avère nécessaire de supprimer :

- Les formes de ponctuation,
- Les mots vides tels que les prépositions, les déterminants, les numéros, les conjonctions, les pronoms et les abréviations,
- Les mots inutiles à la classification tels que les adverbes et les adjectifs. Ainsi, seuls les noms et les verbes sont conservés dans la base des descripteurs.
- Les mots très fréquents : Nous avons adopté la méthode proposée par Stricker (2000). En effet, le rapport $R(m,t) = TF(m,t)/CF(m)$ tel que $TF(m,t)$ est l'occurrence du mot m dans un texte t et $CF(m)$ est l'occurrence du mot m dans l'ensemble des documents, permet de classer les mots par ordre décroissant. Plus le mot m est fréquent, plus le ratio est faible et, inversement, plus un mot est rare, plus le ratio est élevé. Dans le cas limite où un mot n'apparait qu'une seule fois dans l'ensemble de documents, ce ratio vaut 1 et le mot est classé en tête de liste.
- Les mots très peu frequents : ce sont les mots dont le nombre de documents dans lesquels ils apparaissent est inférieur à un certain seuil. Dans notre cas, nous supprimons les mots qui apparaissent dans une seule page du site Web.

Le prétraitement des textes aboutit à la construction d'une matrice croisant les descripteurs aux pages avec le nombre d'occurrences du descripteur dans une page du site comme poids. Un algorithme de classification croisée, CROKI2 (classification CROisée optimisant le Khi2 du tableau de contingence) proposé par Govaert (1983), est ensuite appliqué à la matrice pour découvrir des biclasses de pages et de descripteurs permettant d'attribuer un thème à chaque groupe de pages.

2.1.2 CROKI2 pour la classification croisée

Dans la littérature, la majorité des travaux sur la classification des documents appliquent des méthodes de classification simple sur l'une des deux dimensions (documents ou termes). Dans ce cas, un document appartenant à une classe L est décrit par tous les termes et chaque terme appartenant à une classe K caractérise tous les documents. Ainsi, en faisant porter la structure sur un seul ensemble, la détermination des liens entre les deux partitions est difficile. Dans notre cas, nous cherchons à identifier des classes de documents qui sont mieux décrits par un sous-ensemble de descripteurs, ce qui nécessite de découvrir dans les données des blocs de pages et de termes qui sont fortement correlés. Par conséquent, les algorithmes de classification simultanée sur les lignes et les colonnes sont plus adaptés à ce type de problème. L'algorithme CROKI2 proposé pour la classification croisée d'un tableau de contingence permet la découverte de blocs homogènes à partir d'une partition des descripteurs et une partition des pages recherchées simultanément. Il repose sur l'optimisation du critère du χ^2 de contingence. Disposant d'un tableau de contingence défini sur deux ensembles I et J, il s'agit de trouver une partition P de I en K classes et une partition Q de J en L classes telles que le χ^2 de contingence du nouveau tableau de contingence construit en regroupant les lignes et les colonnes suivant les partitions P et Q soit maximum. L'algorithme proposé construit une suite de couples de

partitions (P^n,Q^n) optimisant le χ^2 du tableau de contingence en appliquant alternativement sur I et sur J une variante de la méthode des Nuées Dynamiques de Diday (1971).

2.2 WUM : Analyse de l'usage

La première étape dans un processus de Web Usage Mining, une fois les données collectées, est le prétraitement des fichiers Logs qui consiste à nettoyer et transformer les données. La deuxième étape est la fouille des données permettant de découvrir des règles d'association, un enchainement de pages Web apparaissant souvent dans les visites et des "clusters" d'utilisateurs ayant des comportements similaires en terme de contenu visité. La dernière étape dans le processus est celle d'analyse et d'interprétation. Nous proposons dans ce papier d'exploiter les résultats de l'analyse textuelle pour l'analyse de l'usage.

2.2.1 Prétraitement des fichiers Logs

Le prétraitement des fichiers logs a comme objectif la structuration et l'amélioration de la qualité des données provenant de ces fichiers pour les préparer à une analyse des usages. Les objets à reconstruire ou à identifier dans un processus de prétraitement de fichiers logs web sont les clics, les utilisateurs, les robots web, les sessions, les navigations et parfois les épisodes. Le prétraitement des données se décompose en deux phases principales : une phase de nettoyage des données et une phase de transformation. Le nettoyage consiste à filtrer les données inutiles à travers la suppression des requêtes ne faisant pas l'objet de l'analyse, telles que les requêtes aux images ou aux fichiers multimédia, et celles provenant des robots Web. La transformation des données regroupe plusieurs tâches telles que l'identification des utilisateurs et la construction des sessions et des visites. L'identification des internautes est effectuée à l'aide des adresses IP, des Cookies ou du couple (pseudonyme, mot de passe). La reconstitution des sessions se fait en regroupant les requêtes émises par cet utilisateur. Chaque session est décomposée en visites en se basant sur le critère empirique de Kimball et Merz (2000).

2.2.2 Fouille de données

Cette étape consiste à appliquer des techniques de fouille des données sur le fichier de sessions afin d'extraire des connaissances sur les comportements des utilisateurs du site. Ces techniques varient entre les méthodes factorielles, les méthodes de classification automatique telles que les règles d'association pour la découverte de motifs fréquents de navigation (Marascu et Masseglia, 2006), les cartes de Kohonen pour la classification des utilisateurs (Charrad, 2005), (Lechevallier et al., 2003), (Fu et al., 2000), (Benedek et Trousse, 2003), (Srivastava et al., 2000) et les méthodes de classification supervisée telles que les arbres de décision, les réseaux de neurones et le raisonnement à base de mémoire.

3 Expérimentations et résultats

Nous proposons d'appliquer l'approche WCUM à un site Web de tourisme. Le prétraitement des textes aboutit à la construction d'une matrice croisant 418 descripteurs à 125 pages. Chaque cellule dans la matrice correspond au nombre d'occurrences du descripteur dans la page.

3.1 Résultats de l'analyse textuelle

L'application de l'algorithme CROKI2 à cette matrice aboutit à un ensemble de biclasses. La sélection des meilleures nécessite le recours aux critères suivants :

– **Pertinence de la biclasse :** la pertinence P de la biclasse est mesurée à travers la part de l'inertie conservée par la biclasse, notée B_{kl}, dans l'inertie totale B.

$$P = B_{kl}/B$$

avec

$$B_{kl} = f_{k.}f_{.l}(\frac{f_{kl}}{f_{k.}f_{.l}-1})^2$$

et

$$B = \sum_{k,l} B_{kl}$$

– **Homogénéité de la biclasse :** l'homogénéité, H, de la biclasse est mesurée par la part d'inertie B_{kl}, conservée par la classe par rapport à l'inertie initiale T_{kl} des points de la classe. La valeur obtenue comprise entre 0 et 1 est d'autant plus grande que la biclasse est homogène.

$$H = (B_{kl}/T_{kl})$$

avec

$$T_{kl} = \sum_{i \in P_k} \sum_{j \in Q_l} f_{i.}f_{.j}(\frac{f_{ij}}{f_{i.}f_{.j}-1})^2$$

A chaque classe de descripteurs, un thème est attribué en fonction des termes qui le composent (tab. 1).

A chaque classe des pages, une classe de descripteurs est associée pour former la biclasse. Chaque classe de pages appartient à au moins une biclasse. A titre d'exemple, la classe 6 de pages appartient à la fois aux biclasses (3,6) et (6,6). Deux thèmes sont donc associés aux pages composant cette classe. Le thème principal est identifié en se basant sur l'homogénéité et la pertinence des biclasses. Le tableau ci-dssous (tab. 2) croise quelques classes de pages aux thèmes qui leur sont associés.

D'autre part, l'URL de chaque page est organisé de façon hiérarchique sous forme de rubriques et de sous-rubriques représentant, selon le point de vue du concepteur, le contenu de la page. En examinant la structure arborescente des pages, la décomposition de l'URL de chaque page en rubriques et la comparaison de ces rubriques avec les résultats de la classification des pages basée sur le contenu permet de vérifier si les rubriques reflètent le contenu des pages.

3.2 Résultats de l'analyse de l'usage

En considérant les fichiers Logs du site dont on a analysé le contenu, nous procédons au nettoyage des requêtes non valides (dont le statut est inférieur à 200 ou supérieur à 399), les requêtes provenant des robots Web, les requêtes dont la méthode est différente de "GET" et les scripts. L'identification des sessions est effectuée en utilisant le couple (IP, User-Agent). Par suite, deux requêtes provenant de la même adresse IP mais de deux user-agents différents

Classes de descripteurs	Mots-clé	Thèmes
Classe 6	Bergerie, Brasserie, Centre, Distance, Fax, Hôtel, Magasin, Nord, Port, Zone, Restaurant, Sud, Technopôle, Tél, Village	Hôtels et Restaurants
Classe 5	Activité, Fête, Bal, Football, lieu, Manifestation, Occasion, Réunion	Activités et Manifestations
Classe 1	Amande, Crème, Eau, Flamber, Fruit, Gastronomie, Glacer, Lait, Mirabelle, Oeuf, Purée, Recette, Sucre,Hôspitalité	Recettes de cuisine
Classe 2	Arme, Art, Artiste, balade, Château, Découverte, Eglise, Exposition, Galerie, Guerre, Habit, Histoire, Illustrer, Maréchal, Monument, Moyen-Age, Palais, Pasteur, Peintre, Peinture, Promeneur, Renaissance, République, Saint, Siècle, Spectacle, Trésor	Histoire et Monuments
Classe 3	Cathédrale, Bibliothèques-médiathèques Boulevard, Capitale, Direction, Edifices, Gare, Guide, Hôpital, Information	Autres Adresses

TAB. 1 – *Exemples de thèmes*

	Th 1	Th 2	Th 3	Th 5	Th 6	Th. principal
Classe 2	X	X	X			Thème 2 : Histoire et Monuments
Classe 4	X		X			Thème 1 : Spécialités de cuisine
Classe 6			X		X	Thème 6 : Hôtels et Restaurants

TAB. 2 – *Exemples de Biclasses*

appartiennent à deux sessions différentes. Chaque session est décomposée en visites. Une visite est composée d'une suite de requêtes séquentiellement ordonnées, effectuées pendant la même session et ne présentant pas de rupture de séquence de plus de 30 minutes. Suite au nettoyage et transformations des requêtes provenant des fichiers logs, une matrice croisant les sessions aux pages est construite. Or d'après l'analyse du contenu, chaque page est affectée à une biclasse donc à un thème. Par suite, il est possible de croiser les visites (ou navigations) aux thèmes.

	Thème 1	Thème 2	...	Thème m
Navigation 1	20	0	...	2
Navigation 2	0	11	...	0
...	...	...	...	...
Navigation n	0	43	...	10

TAB. 3 – *Matrice des pages et des descripteurs*

Chaque cellule de la matrice correspond au nombre total de visites effectuées aux pages appartenant au thème j au cours de la navigation i. Un algorithme de classification simple est appliqué à la matrice $Navigations \times Themes$ pour découvrir des classes d'utilisateurs ayant un comportement similaire sur le site.

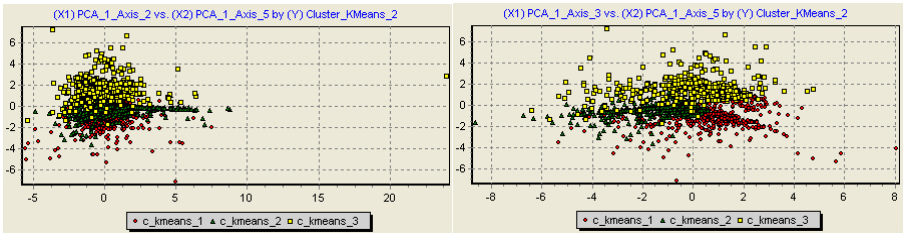


FIG. 2 – *Projection des classes d'utilisateurs*

L'application du Kmeans à la matrice croisant les navigations aux thèmes permet d'identifier trois classes d'utilisateurs. La classe C3 (fig. 2) est composée d'utilisateurs intéressés par les pages traitant du thème 4 (Informations utiles). Ils sont alors à la recherche des informations sur les horaires d'ouverture et de fermetures de certains établissements, des tarifs et des calendriers. Les internautes de la classe C2 sont par contre intéressés par les pages ayant pour thème "Recettes de cuisine", "histoire et monuments" et "Hôtels et Restaurants" (càd Thème1, Thème2 et Thème6). La classe C1 regroupe les visiteurs dont le motif est la recherche des adresses utiles (Thème 3), des manifestations et des activités culturelles (thème 5) et des informations utiles (thème4). Comme la classe majoritaire est C2 (70% des visiteurs), on déduit que ce sont les thèmes 1,2 et 6 qui intéressent le plus les visiteurs. Il s'en suit que les pages traitant de ces thèmes devraient être accessibles facilement et reliés par des hyperliens pour faciliter la navigation sur le site.

4 Conclusion

Dans ce papier, nous avons proposé une approche reliant l'analyse du contenu à l'analyse de l'usage. L'apport de cette approche est qu'elle permet d'identifier les thèmes qui intéressent les visiteurs et de tester si le contenu du site répond à leurs attentes. D'autre part, elle permet de réorganiser les pages de manière à faciliter le parcours du site.

Références

Benedek, A. et B. Trousse (2003). Adaptation of self-organizing maps for case indexing. *In 27th Annual Conference of the Gesellschaft fur Klassifikation, Germany*, 31–45.

Charrad, M. (2005). *Techniques d'extraction des connaissances appliquées aux données du Web*. Mémoire de mastère, Ecole Nationale des Sciences de l'Informatique de Tunis.

Charrad, M., Y. Lechevallier, G. Saporta, et M. B. Ahmed (2008). Web content data mining : la classification croisée pour l'analyse textuelle d'un site web. *Revue des Nouvelles Technologies de l'Information (Cépaduès) 1*, 43–54.

Diday, E. (1971). Une nouvelle méthode de classification automatique et reconnaissance des formes : la méthode des nuées dynamiques. *Revue de Statistique Appliquée 19 2*, 19–33.

Fu, Y., K. Sandhu, et M. Shih (2000). A generalization-based approach to clustering of web usage sessions. *In Proceedings of the 1999 KDD Workshop on Web Mining, San Diego, Springer*, 21–38.

Govaert, G. (1983). *Classification croisée*. Thèse de doctorat, Université Paris 6.

Kimball, R. et R. Merz (2000). Le data webhouse : Analyser des comportements clients sur le web. *Editions Eyrolles, Paris*.

Lechevallier, Y., D. Tonasa, B. Trousse, et R. Verde (2003). Classification automatique : Applications au web mining. *In Yadolah Dodge and Giuseppe Melfi, editor, Méthodes et Perspectives en Classification, Presse Académiques Neuchâtel*, 157–160.

Marascu, A. et F. Masseglia (2006). Extraction de motifs séquentiels dans les flots de données d'usage du web. *Extraction et Gestion des Connaissances (EGC'06), Lille*, 627–638.

Srivastava, J., R. Cooley, M. Deshpande, et P.-N. Tan (2000). Web usage mining : Discovery and applications of usage patterns from web data. *SIGKKD Explorationsy*, 12–23.

Summary

The Web Content and Usage based (WCUM) Approach proposed in this paper deals with tha analysis of both content and usage of the web site to better understand the behaviour of web site users. Our main contribution consists in associating the content analysis with usage analysis and adaptating block clustering algorithms, traditionally used in bioinformatics, to web mining problems in order to discover homogeneous blocs of instances and attributes.

Analyse de la variation spatio-temporelle des objets dans les images satellitaires à base de modèle de Markov caché couplé

Houcine ESSID[1,2], Imed Riadh FARAH[1,3]
Vincent BARRA[2], Henda BEN GHEZALA[1]

[1]Laboratoire de Recherche en Informatique Arabisée et Documentique Intégrée - Génies Documentiel et Logiciel - Ecole Nationale des sciences de l'informatique. Campus Universitaire de Manouba, 2010 Manouba, Tunis, Tunisie
[2]Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes UMR CNRS 6158 - Campus Scientifique des Cézeaux - Office B139 63177 AUBIERE CEDEX, France -
[3]TELECOM-Bretagne, Département ITI Technopôle Brest Iroise CS 83818, 29238 Brest Cedex France
houcine.essid@isima.fr
riadh.farah@ensi.rnu.tn
vincent.barra@isima.fr
www.isima.fr/vbarra
henda.bg@cck.rnu.tn

Résumé. Cet article est consacré à l'étude et la réalisation d'un système d'interprétation des images satellitaires par le biais des réseaux Bayésiens dynamiques. La contribution de cet article est de proposer une approche basée sur le modèle du Markov caché couplé. Chaque image est segmentée en plusieurs régions et à partir de chaque région, on extrait un ensemble de descripteurs. A partir des régions et des descripteurs qui les caractérisent, on construit un modèle de Markov caché couplé qui sera transformé en réseau Bayésien dynamique pour lui appliquer par la suite les algorithmes d'inférence et d'apprentissage. A l'aide des tables de probabilités et des paramètres associés au réseau Bayésien, on dégage les variations dans les images satellitaires pour les interpréter et mesurer l'influence descripteurs sur la dynamique des scènes. Un système est réalisé pour données des résultats qui nous on permis d'envisager d'utiliser les méthodes de l'extraction et de fouille pour le choix des régions obtenues après la segmentation et le choix du modèle le plus adéquat à partir d'un ensemble de modèle comme le modèle hiérarchique et le modèle triplet.

1 Introduction

L'analyse d'une scène dynamique, formée d'un ensemble d'images satellitaires, est un thème de recherche très important dans le domaine de la modélisation spatio-temporelle. Le

suivi et l'estimation des variations permettent de segmenter les objets dynamiques et d'estimer les variations futures. Cela permet la résolution d'un certain nombre de tâches dans plusieurs domaines tels que l'occupation du sol, la gestion de l'urbanisme, l'aménagement du territoire et la surveillance à distance. Durant les dernières années, les chercheurs en modélisation spatio-temporelle ont essayé de trouver des solutions dans les problèmes de la dynamique des scènes. Les approches généralement incorporent les Statistiques Appliquées et la Géométrie. Dans cet article, nous proposons une approche spatio-temporelle basée sur les modèles de Markov cachés couple pour la représentation et la modélisation des variations dans une séquence d'images satellitaires d'une même portion de la terre. Le but de cette modélisation est de détecter les variations connaissant les descripteurs obtenus suite à la phase de segmentation ; de mesurer l'influence des descripteurs sur la dynamique des scènes et de prévoir les variations futures. Pour réaliser ce travail, nous modélisons l'ensemble des régions et des descripteurs sous forme d'un modèle de Markov caché couplé (MMCC) qui sera transformé en un réseau Bayésien dynamique (RBD). Cette transformation a pour but d'appliquer au modèle obtenu les algorithmes d'inférence et les algorithmes d'apprentissage. Cette estimation se base soit sur une base des échantillons soit sur des données plus ou moins complètes. Cet article est structuré en quatre sections. Dans la première, nous présentons les images satellitaires et leurs caractéristiques. Dans la deuxième section, nous donnons l'état de l'art des modèles d'analyse spatio-temporelle. La troisième section est réservée à la modélisation des variations temporelles au sein d'une séquence d'images satellitaires par le MMCC. Enfin, nous terminons par une conclusion.

Le résumé du début est dans la langue de l'article. S'il est en anglais, on utilisera le titre « Abstract ». La traduction du résumé dans l'autre langue doit être donnée en fin d'article. Pour les textes de deux pages ou moins (posters), ne pas mettre le résumé du début. Donner cependant à la fin du texte un « summary » en anglais d'au plus 40 mots.

2 Les modèles d'analyse spatio-temporelle

Une image satellitaire est une illustration d'une portion de la surface terrestre observée. Plusieurs systèmes ont été développés pour l'analyse d'images satellitaires en utilisant différentes architectures et méthodes pour différentes applications pour tenir compte de l'aspect temporel et dynamique des images satellitaires. Sans vouloir être exhaustif, nous donnons ci-après les principales références qui ont servi à notre analyse [1] :

- Le système KUMAR, développé pour l'interprétation d'images qui utilise un réseau Bayésien de niveau 2 [2] ;
- Le système AMIT, conçu pour segmenter les régions intéressantes d'une image et fusionner des données issues de plusieurs capteurs. Se système utilise les réseaux Bayésiens hiérarchiques [3] ;
- Le système ANDRÉ, dédié à la déconvolution d'images satellitaires et aériennes. Ce système permet d'estimer les paramètres décrivant les propriétés de l'image que l'on cherche à reconstruire. Il se base sur un modèle Bayésien hiérarchique permettant de tenir compte des caractéristiques des images étudiées [4] ;
- Le système ASCENDER II: l'utilisation des réseaux Bayésiens hiérarchique combinée avec la théorie de l'utilité permet de manipuler les informations à partir de plusieurs images 3D ayant différentes caractéristiques [5] ;

- Le système PIECZYNSKI: dédié à la segmentation par des méthodes Bayésiennes et des modèles de Markov [6] ;
- Le système LAI utilisé pour estimer un indice de zones abandonnées. Ce système est une excellente application des réseaux Bayésiens [7] ;
- Le système IHBN : permet l'extraction de données d'une base d'images. Ce système utilise l'architecture Bayésienne combinée avec les arbres de décision pour l'apprentissage incrémental [8] ;
- Le système BOUBCHIR : conçu pour récupérer une image de bonne qualité, proche de l'image originale recueillie en sortie de tout capteur. Ce système utilise l'estimation statistique Bayésienne dans le domaine des transformées multi-échelles parcimonieuses orientées et non orientées comme solution au problème de débruitage [9] ;
- Le système TANAKA : utilise les réseaux Bayésiens et la propagation des croyances dans le traitement probabiliste des images [10] ;
- Segmentation spatio-temporelle en utilisant un front de propagation et champs de déplacement [11] ;
- Modélisation spatio-temporelle de l'attention visuelle [12] ;
- Apport de la Télédétection et du SIG pour le suivi spatio-temporel de l'occupation du sol et de l'érosion nette dans le bassin de l'Oued Tlata (Maroc) [13] ;
- Cartes de Saillance spatio-temporelle basées Contrastes de Couleur et Mouvement Relatif [14] ;
- Segmentation spatio-temporelle d'une séquence d'images par compétition de mouvements [15] ;
- Modélisation hiérarchique spatiotemporelle de données alignées d'incidence de cancers [16].

La plupart des systèmes existant ont traité la composante spatio-temporelle sans pour autant, modéliser l'influence des descripteurs sur la dynamique de la scène. Pour remédier à cette insuffisance, nous proposons ici une approche qui utilise un réseau Bayésien dynamique comme représentation des modèles de Markov cachés couplés pour l'interprétation des images satellitaires et ce en tenant compte de la composante temporelle.

3 Modélisation par les modèles de Markov cachés couplés

Dans un 1er temps des classes d'état (du genre normal, dégradé, critique), mesurer les descripteurs et faire l'inférence pour voir on est en quel état ou bien on a resté beaucoup dans tel état ou on tend à aller vers tel état.

.

Mais dans un 2ème temps, pour faire tourner cette chaîne, on a besoin des matrices de transition et d'émission) ? Pour les avoir, on a besoin de faire de l'apprentissage à partir d'une séquence d'images: mesurer à chaque fois le descripteur (géométrique et couleur pour la simulation) au temps t1, voir son état et refaire pour t2 jusqu'à tn (les états sont lu à partir d'un graphe qui contient des courbes tracées à partir de loi de probabilité ???). Cela pour la première série d'images. On refait la même chose pour la deuxième série et ainsi de suite jusqu'à la dernière. Finalement, on applique l'algorithme du Baum-Welch pour estimer les matrices de transition et d'émission.

Un MMCC peut être considéré comme une collection de MMC, un pour chaque flux de données. Les nœuds cachés au temps t sont conditionnés par les nœuds au temps t-1 de tous les MMC. Dans le MMCC (figure 1), les variables interagissent avec leurs voisins. En plus chaque nœud possède sa propre observation. Le modèle couple est donc une représentation des dépendances conditionnelles entre les nœuds et leurs voisins les plus proches. L'utilisation des MMCC peut améliorer les divers résultats de segmentations.

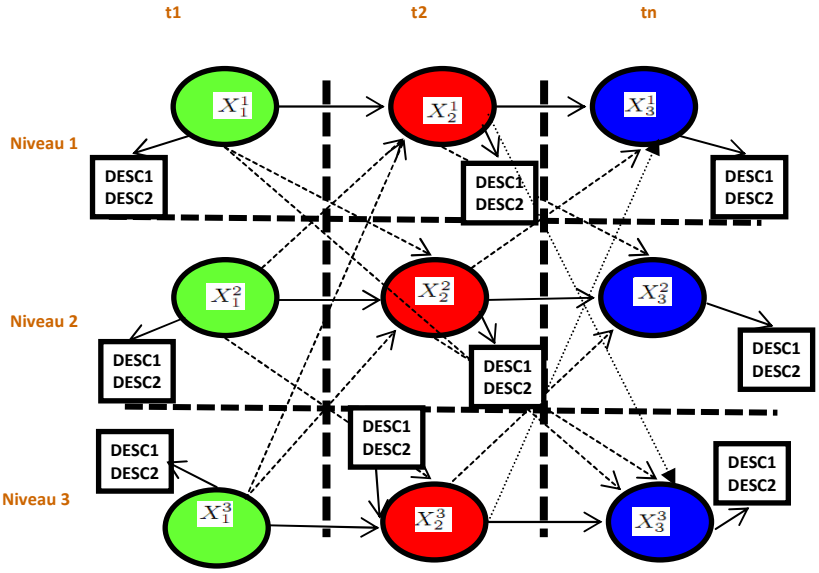


FIG. 1 – *Modèle de Markov caché couplé. Les Y_i sont les événements observables, les X_i sont les états cachés, interagissant avec leurs voisins*

Le modèle couplé présente une structuration parfaite des dépendances conditionnelles ce qui permet de représenter les différents composants d'une manière plus logique et surtout plus proche de la réalité. Les probabilités de transition, ainsi que les lois marginales sont calculables et permettent une décision Bayésienne plus exacte. Ainsi, ce graphe est une association du MMC avec une représentation des dépendances entre les voisins tout en gardant les caractéristiques du MMC toujours visibles. Après la construction du modèle, on doit définir les distributions de probabilité conditionnelle pour tous les nœuds, avec les mécanismes d'inférence et d'apprentissage. L'objectif de l'inférence est de calculer toutes les lois de probabilité marginales, et elle se base essentiellement sur l'approche Backward - Forward. Les techniques d'apprentissage dans les RBD sont une extension des techniques d'apprentissage dans les réseaux Bayésiens classiques, et concernent l'apprentissage des paramètres hors ligne (type algorithme EM) ou en ligne, et l'apprentissage de structure. Ici, nous utilisons l'approche du maximum de vraisemblance pour le premier cas. Étant donné un RBD avec un ensemble d'observations, les paramètres sont choisis tels que la probabilité des observations soit maximisée. Ces paramètres peuvent être calculés en utilisant l'algorithme EM.

4 Système d'interprétation spatio-temporelle d'images satellitaires à base de MMCC

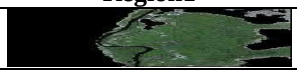
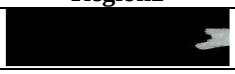
Pour modéliser le problème sous forme de MMCC (figure 2), nous allons considérer une séquence composée de trois images satellitaires décrivant la même portion terrestre observée du même endroit mais qui sont acquises dans trois instants séparés par un intervalle de temps (par exemple un mois). Considérons maintenant que chaque image parmi ces trois est segmentée en trois régions R1, R2 et R3. A partir de ces images segmentées, on peut passer d'une région à une autre suivant une loi de probabilité. Chaque région possède un ensemble de descripteurs et des invariants qui lui sont affectés suivant une distribution de probabilité que l'on peut déterminer. Nous considérons que chaque région est modélisée par un état caché et que les descripteurs qui lui sont attribués sont les observations liées à cet état. Si nous voulons déterminer les variations qui ont touché une région alors, nous devons chercher l'influence d'une composition d'un certain nombre de descripteurs sur le passage d'une région à une autre dans l'espace du temps.

5 Expérimentation

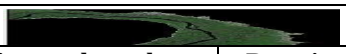
Le système réalisé permet d'insérer une séquence d'images acquises à des instants séparés par des intervalles de temps qui peuvent être constants ou non et en se basant sur le modèle de Markov caché couple qui n'a jamais été utilisé dans ce cadre de travail ou dans des cas similaires. L'interprétation dégagée par notre système peut être très efficace du fait qu'elle ne requiert pas beaucoup d'effort physique ou matériel mais elle nécessite la présence d'une ancienne base d'échantillons dans laquelle on met un ensemble d'exemples réels à partir des résultats expérimentaux. Pour tester notre système, nous avons travaillé sur une base d'images contenant trois séquences dont la première est composée de deux exemplaires qui illustrent une observation de l'Indonésie à deux instants différents, la deuxième séquence est une illustration d'une image satellitaire composée de trois régions qui ont subi des variations dans le temps. La dernière séquence est composée de deux images dont chacune est composée de trois régions.

Pour effectuer notre étude expérimentale nous avons pris comme échantillon la séquence n°1 qui se compose de deux images satellitaires prises à deux instants différentes et présentant plusieurs variations au cours du temps. Nous commençons par la segmentation de ces deux images pour faire l'extraction des régions (tableau 1) puis nous calculons les descripteurs et les invariants à partir de chaque région (tableau 2). Ensuite nous effectuons le calcul pour extraire la matrice de transition et la matrice d'observation et enfin nous déduisons la matrice de confusion (tableau 3).

Image (t1)		
		
Région1	**Région2**	**Région3**
		

Image (t2)		
		
Région1	**Région2**	**Région3**
		

TAB. 1 – *Segmentation et extraction des régions des images (t1) et (t2).*




Descripteur de couleur		**Descripteur de texture**		**Descripteur de forme**	
image 1	**Image 2**	**image 1**	**Image 2**	**image 1**	**image 2**
0.0000	0.0000	0.7656	0.7656	3.0000	5.0000
0.0001	0.0001	0.1094	0.1094	37.6691	62.8319
4.8563	1.9491	0.7956	0.7656	0	0
0.0000	0.0000	0.5646	0.6433	0	0
0.0001	0.0001	0.3763	0.4090	0.7854	0.7854
2.5970	2.0744	0.2783	0.2783	1.0000	1.0000
0.0000	0.0000	0.0000	0.0000	0.3333	0.2000
0.0001	0.0401	0.0000	0.0000	0.3333	0.2001
0.7312	0.7921	0.0001	0.0002	0.1566	0.1678

TAB. 2 – *Calcul des descripteurs à partir de la région n°1.*

matrice de transition	**matrice d'observation**	**matrice de confusion**
$\begin{bmatrix} 0.65 & 0.015 & 0.06 & 0.145 & 0.05 & 0.08 \\ 0 & 0.465 & 0 & 0.355 & 0.005 & 0.175 \\ 0.05 & 0.22 & 0.61 & 0.105 & 0.005 & 0.001 \\ 0 & 0.15 & 0 & 0.65 & 0.05 & 0.15 \\ 0.15 & 0.25 & 0.05 & 0.05 & 0.15 & 0.05 \\ 0 & 0.325 & 0 & 0.015 & 0.65 & 0.01 \end{bmatrix}$	$\begin{bmatrix} 0.25 & 0.15 & 0.15 & 0.05 & 0.15 & 0.25 \\ 0.25 & 0.25 & 0.15 & 0.15 & 0.1 & 0.1 \\ 0.15 & 0.25 & 0.15 & 0.25 & 0.115 & 0.005 \\ 0.25 & 0.15 & 0.15 & 0.15 & 0.05 & 0.25 \\ 0.25 & 0.25 & 0.05 & 0.05 & 0.25 & 0.2 \\ 0.15 & 0.1 & 0.1 & 0.25 & 0.15 & 0.25 \end{bmatrix}$	$\begin{bmatrix} 0.454 & 0.355 & 0.191 \\ 0.590 & 0.255 & 0.055 \\ 0.110 & 0.339 & 0.651 \end{bmatrix}$

TAB. 3 – *Interprétation des résultats.*

L'étude expérimentale sur cette séquence a montré que la région R1 de l'image t1 a augmenté de 30% par apport à la région R2 et de 19% par apport à la région R3 et que la Région R2 à diminué de taille de 10% par apport à R1 mais, elle a augmenté de 20% par apport à R3. La région R3 a diminué de 15% par apport à R2 et de 10% par rapport à R1.

6 Conclusion

L'interprétation des images satellitaires est un domaine qui attire de plus en plus les chercheurs en traitement et analyse d'images. Les recherches orientées dans ce domaine et travaillant sur ce sujet n'utilisent pas le modèle de Markov caché couplé. Notre contribution réside dans l'adoption de ce type de modèle pour dégager et interpréter les variations temporelles au sein d'une séquence d'images satellitaires. Le modèle de Markov caché couple permet d'exploiter l'aspect symbolique de l'image pour mieux dégager les variations tempo-

relles au sein de ces derniers. Ainsi, l'approche proposée est basée sur l'extraction des régions à partir de chaque image de la séquence observée et la détermination des vecteurs caractéristiques de ces régions qui sont les descripteurs, pour construire un modèle de Markov caché couplé répondant à nos exigences. Le développement du système nous a permis d'avoir des premiers résultats élémentaires mais prometteurs. Nos perspectives envisagent d'étendre notre outil par l'intégration d'un module d'apprentissage capable de générer les matrices de transition et d'émission. Nous envisageons aussi d'utiliser des seuils afin d'indiquer des alertes en cas de situation critique. On va essayer d'étendre ce travail sur des cas réels (déforestation, érosion, ...) ensemble ou séparés et de voir comment adapter la simulation à ces cas réels.

Références

[1] Essid H., Farah I. R., Ben Ghzala H., Barra V., Modèle hybride spatio-temporel d'analyse d'images satellitales à base de réseaux Bayésiens dynamiques, Actes COSI'09, Annaba Algérie, 2009

[2] Kumar V., Desai U., Image Interpretation Using Bayesian Networks, IEEE Transactions on PAMI, 18:74-77 1996

[3] Singhalt, A., Luos, J., Brownt, C. A., Multilevel Bayesian Network Approach to Image Sensor Fusion. Dept of Computer Sciences, Univ. Rochester, NY 14627, 2000

[4] Jalobeanu, A., Modèles, estimation Bayésienne et algorithmes pour la déconvolution d'images satellitaires et aériennes. Thèse à l'Université de Nice-Sophia Antipolis, 2001

[5] Marengoni, M., Han son, A., Zilberstein, S., Riseman, E. Decision Making and Uncertainty Management in a 3D Reconstruction System. IEEE Transactions on PAMI, 25 : 852-858, 2003.

[6] Pieczynski, W. Modèles de Markov en traitement d'images. Traitement du Signal, 20 :255-277, 2003.

[7] Kalácska, M., Sánchez-Azofeifa, G.A., Caelli, T., Rivard, B. Boerlage, B. Estimating Leaf area From Satellite Imagery. IEEE Transactions On Geoscience And Remote Sensing, 43 :1866-1873, 2005.

[8] Baice, L. et Senmiao, Y. Incremental Hybrid Bayesian Network in Content-Based Image Retrieval. IEEE, CCECE/CCGEI, Saskatoon, 2005.

[9] Boubchir L. Approches Bayésiennes pour le débruitage des images dans le domaine des transformées multi-échelles parcimonieuses orientées et non orientées. Thèse, Université de Caen/Basse-Normandie, 2005

[10] Tanaka K., Bayesian Network and Probabilistic Image Processing Statistical Aspect of Belief Propagation Method, Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan, 2005

[11] Santiago venegas-martinez, Manuel rendon mancha et Georges stamon Segmentation spatio-temporelle en utilisant un front de propagation et champs de deplacement , Paris, 2001.

[12] O. Le Meur P. Le Callet D. Barba D. Thoreau, Modélisation spatio-temporelle de l'attention visuelle, 2005.

[13] Abdelkader El garouani1, H. chen2, L. Lewis3, A. Tribak4, M. Abahrour4, Apport de la Télédétection et du SIG pour le suivi spatio-temporel de l'occupation du sol et de l'érosion

nette dans le bassin de l'Oued Tlata (Maroc), Actes des JSIRAUF, Hanoi, 6-9 novembre 2007
[14] O. Brouard1 V. Ricordel1 D. Barba1, Cartes de Saillance Spatio-Temporelle basées Contrastes de Couleur et Mouvement Relatif, 2008.

[15] Sylvain BOLTZ Laboratoire I3S, Segmentation spatio-temporelle d'une séquence d'images par compétition de mouvements, Equipe CReATIVe Laboratoire I3S, 2005.
[16] Erik A. Sauleau, Arnaud Etienne et Antoine Buemi, Modélisation hiérarchique spatio-temporelle de données alignées d'incidence de cancers, IRISA – INRIA Rennes / Équipe Vista / Rennes, 2005.

Summary

This paper is devoted to study and to realize a satellite image interpretation system using coupled hidden Markov model. Each image is segmented into several areas from which we extract their descriptors. From areas and descriptors we construct the coupled hidden Markov model that will be transformed into a dynamic Bayesian network to apply inference and learning algorithms. Using the tables of probabilities and the parameters associated to the dynamic Bayesian network, we take out the variations in the satellite images to interpret them and measure descriptor influence on scene dynamic. The results of the implementation show that we can use data mining to choose areas and model from hierarchical or triplet Markov model.