Modélisation dynamique et temporelle de l'utilisateur pour un filtrage personnalisé de documents textuels

Rachid AREZKI *, Abdenour MOKRANE*
Pascal PONCELET*, Gérard DRAY*
David Wiliams PEARSON**

*Centre LGI2P EMA, Site EERIE Parc Scientifique Georges Besse 30035 Nimes Cedex 1, France
{rachid.arezki, abdenour.mokrane, gerard.dray, pascal.poncelet}@ema.fr,
**IUT de Roanne, 20 Avenue de paris 42334 Roanne, France
david.pearson@univ-st-etienne.fr

Résumé. L'apprentissage efficace du profil utilisateur est un challenge car il évolue sans cesse. Dans cet article nous proposons une nouvelle approche pour l'apprentissage du profil long-terme de l'utilisateur pour le filtrage de documents textuels. Dans ce cadre les documents consultés sont classés de manière dynamique et nous analysons la répartition dans le temps des classes de documents afin de déterminer le mieux possible les classes d'intérêts de l'utilisateur. L'étude empirique confirme la pertinence de notre approche pour une meilleure personnalisation de documents.

1 Introduction

Avec le développement d'Internet et les nouveaux moyens de stockage de données, les serveurs de documents en ligne regorgent d'énormes quantités de documents de différentes thématiques. Les moteurs de recherche sont d'une grande utilité pour la recherche de documents pertinents mais ils nécessitent de formuler de nouvelles requêtes à chaque fois que l'utilisateur a besoin de nouveaux documents. Récemment, des systèmes capables de proposer des documents adaptés à l'utilisateur, sans que ce dernier ne formule de requêtes ont été développés. Cependant ils ne prennent pas en compte l'évolution dans le temps des classes de documents consultés par l'utilisateur. Dans ce papier nous proposons une nouvelle approche d'apprentissage du profil long-terme de l'utilisateur pour le filtrage de documents textuels. Cette approche est basée sur l'analyse de l'évolution dans le temps des classes de documents consultés par l'utilisateur. Dans ce cadre, les documents consultés sont classés de manière dynamique et nous analysons ensuite la répartition dans le temps de ces classes de documents. Le but de notre approche est de déterminer le mieux possible les classes d'intérêts de l'utilisateur, cela en donnant plus d'importance aux classes de documents régulièrement consultées qu'à celles concentrées sur de courtes périodes. Notre approche ne requiert donc pas que l'utilisateur fournisse de manière explicite des informations au système. Nous avons développé *LUCI*, un système permettant la personnalisation de documents en ligne via notre approche [Arezki *et al.*, 2003]. *LUCI* apprend le profil de l'utilisateur via les documents consultés par celui-ci et lui propose des documents de manière dynamique, i.e. à chaque fois que l'utilisateur modifie son profil (consulte un nouveau document), le système lui propose une collection de documents adaptés à son profil.

L'article est organisé de la manière suivante. Dans la section 2 nous présentons les problématiques de la modélisation des intérêts long-terme de l'utilisateur pour le filtrage personnalisé de documents textuels. La section 3 détaille notre approche pour la modélisation de l'intérêt long-terme de l'utilisateur. Nous présentons à la section 4 une série d'expériences sur un corpus documentaire de référence, nous montrons que notre approche permet de détecter et de proposer des documents d'intérêt régulier. Un bref état de l'art sur la modélisation de l'utilisateur est proposé à la section 5. Enfin, dans la section 6, nous concluons en résumant les avantages et présentons les perspectives associées au modèle proposé.

2 Problématique

L'objectif de notre proposition est l'apprentissage du profil long-terme de l'utilisateur pour une personnalisation efficace de documents textuels. Pour atteindre cet objectif, nous classons les documents consultés par l'utilisateur et nous analysons l'évolution de ces classes dans le temps. L'idée générale est de considérer qu'une classe de documents consultée régulièrement par l'utilisateur a plus d'intérêt à long-terme pour ce dernier qu'une classe de documents concentrée sur une courte période. Par exemple, un utilisateur ayant des actions en bourse consulte tous les jours un ou deux documents sur la thématique *finance*. Ce même utilisateur, ayant voulu faire un exposé sur les dinosaures, a consulté plus de 100 documents sur une période de deux jours sur la thématique *dinosaure*. On considère qu'à long-terme (un mois par exemple) la thématique *finance* a plus d'intérêt pour l'utilisateur que la thématique *dinosaure*, alors que le nombre de documents consultés de la thématique *dinosaure* est bien plus important que le nombre de documents consultés de la thématique *finance*. La thématique *finance* représente un intérêt régulier pour l'utilisateur alors que la thématique *dinosaure* est d'un intérêt spontané. A long-terme nous devons être capable de proposer à l'utilisateur des documents de la thématique *finance* car elle est d'un intérêt régulier, et de ne pas proposer des documents de la thématique *dinosaure*, car elle ne représente qu'un intérêt spontané. La problématique de la personnalisation de documents à long-terme consiste à trouver un modèle capable de prendre en compte l'évolution, dans le temps, des classes de documents consultés par l'utilisateur et de détecter au mieux les classes d'intérêts réguliers pour une meilleur personnalisation de documents.

3 Modélisation de l'intérêt long-terme

Dans cette section nous décrivons notre approche pour la modélisation de l'intérêt long-terme. Notre motivation derrière cette modélisation est de capturer le mieux possible l'intérêt général de l'utilisateur. Ainsi, nous considérons qu'une classe régulièrement consultée (intérêt permanent) a plus d'intérêt qu'une classe concentrée sur une courte période (intérêt spontané).
Notre approche est basée sur : (1) Une classification dynamique des documents consultés

par l'utilisateur (le detail de cet algorithme est disponible dans [Arezki *et al.*, 2003]), (2)L'association à chaque classe d'un vecteur, ce dernier correspond à la somme des vecteurs de documents de la classe, (3)Un poids associé à chaque classe, celui-ci détermine la répartition dans le temps des documents de la classe, (4)Un vecteur nommé LTV *(*Long-Term Vector), déterminant l'intérêt long-terme de l'utilisateur est calculé à partir des vecteurs de classes et de leurs répartitions, (5)Un calcul de similarité entre les documents du corpus et le vecteur LTV. Les documents les plus similaires sont proposés à l'utilisateur.

3.1 Structure du modèle long-terme

Le modèle utilisateur est défini par le tuple $X =< id, S >$ où : id identifie de manière unique l'utilisateur et S est l'ensemble des classes de documents consultés par l'utilisateur id, i.e $S = \{C_1, ...C_n\}$ où n est le nombre de classes. Chaque classe C_i, i=1 à n, est définie par le tuple $C_i = \{V, Rep_{C_i}, V_{C_i}\}$, tel que :

1. $V = \{(V_1^{C_i}, Pos_1)..(V_{\|C_i\|}^{C_i}, Pos_{\|C_i\|})\}$: est un ensemble de tuples contenant les documents consultés de la classe C_i (vecteurs) et leurs positions (dans le temps). où $\|C_i\|$ représente le nombre de documents de la classe C_i, $V_j^{C_i}$ est le j^{eme} document de la classe C_i, et Pos_i est la position d'un document, représentant son ordre d'apparition dans le temps par rapport aux documents consultés.
2. Rep_{C_i} : répartition dans le temps des documents de la classe C_i (voir section 3.2)
3. V_{C_i} : le vecteur de la classe C_i, avec : $V_{C_i} = \sum_{j=1}^{j=\|C_i\|} V_j^{C_i}$

3.2 Calcul de la répartition des classes de documents

Pour le calcul de la répartition dans le temps des classes de documents consultés par l'utilisateur, on associe à chaque classe de documents consultés par l'utilisateur un nuage de points représenté dans un espace à deux dimensions. Chaque point représente un document de la classe. A chaque document on associe deux coordonnées :

1. position du document dans l'ensemble des documents, c'est à dire l'ordre de consultation de ce document par l'utilisateur,
2. position du document par rapport aux documents de sa classe, c'est à dire l'ordre d'ajout de ce document à sa classe.

Une droite de regression (Δ) des moindres carrés du nuage de points est calculée. La répartition Rep_{C_i} d'une classe de documents C_i est donnée par la formule :

$$Rep_{C_i} = \frac{Pos_{\|C_i\|} - Pos_1 + 1}{N} * \frac{1}{1 + \sum_{j=1}^{j=\|C_i\|} Distance(D_j, \Delta)}$$

$Pos_{\|C_i\|}$: position du dernier document consulté appartenant à la classe C_i,
Pos_1 : position du premier document consulté, appartenant à la classe C_i,
Δ : droite de régression des moindres carrés du nuage de points de la classe,
D_j : coordonnées du j^{eme} document de la classe C_i,
N : nombre de documents consultés par l'utilisateur,
$Distance(D_j, \Delta)$: distance entre le point D_j et la droite Δ.

3.3 Calcul du vecteur long-terme LTV et Filtrage de documents

A partir du modèle long-terme, le vecteur LTV définissant l'intérêt long-terme de l'utilisateur est calculé comme suit :

$$LTV = \sum_{i=1}^{i=\|S\|} Rep_{Ci} * V_{Ci} \quad où\, \|S\|\, est\, le\, nombre\, de\, classes.$$

A chaque fois que l'utilisateur consulte un document, son profil est modifié, ainsi le système lui propose une nouvelle collection de documents. L'algorithme ci-dessous décrit le processus de filtrage documentaire associé à la consultation d'un nouveau document par l'utilisateur.

Algorithm 1: Algorithme de filtrage documentaire
Input:
Document V_D, la position Pos_{V_D} du document V_D, modèle utilisateur $X =< id, S >$,
α : constante représentant le seuil de similarité
Output: proposition d'un ensemble de documents à l'utilisateur
begin

1. Associer le document V_D à une classe de documents (voir [Arezki *et al.*, 2003]) ;
2. Calculer le vecteur de la classe où se trouve V_D ;
3. Calculer la répartition de l'ensemble des classes ;
4. Calculer le vecteur LTV ;
5. Calculer la similarité entre le vecteur LTV et l'ensemble des documents du corpus ;
6. Proposer à l'utilisateur les documents dont la similarité est supérieur au seuil α ;

end

4 Expérimentation

Une évaluation a été faite pour mesurer la capacité d'apprentissage du système *LUCI*. L'objectif principal est de mesurer la capacité du système à proposer des documents d'intérêts réguliers. Nous avons choisi de comparer *LUCI* à l'algorithme *LUCI-NEG*, une implementation de *LUCI* sans prise en compte de la répartition dans le temps des classes de documents. Les documents utilisés pour notre étude sont des articles de presse, collectés de 5 journaux en ligne différents sur des périodes différentes. Notre corpus documentaire contient des documents de 6 thématiques différentes de 200 documents chacune. Les thématiques choisies sont : *Économie, Football, Crise-Irakienne, Politique, Cinéma, Informatique*. Pour montrer la capacité du système *LUCI* à proposer des documents d'intérêts réguliers, nous considérons un utilisateur consultant en premier un ensemble de documents de la thématique *informatique* (12 documents), ensuite un ensemble de documents de la thématique *économie* (30 documents), suivi d'un ensemble de documents de la thématique *football* (24 documents). Ainsi qu'une consultation régulière de documents de la thématique *cinéma* (11 documents). Cette dernière représente un intérêt régulier pour l'utilisateur. Notre motivation est de voir la capacité du système à proposer des documents de la classe *cinéma*.

La figures 1 montre les pourcentages de documents par thématique proposés à l'utilisateur, respectivement par le système *LUCI* et l'algorithme *LUIC-NEG*. On remarque qu'à partir de l'itération 29 le système *LUCI* propose des documents de la thématique *Cinéma* (thématique régulièrement consultée), alors que *LUCI-NEG* n'en

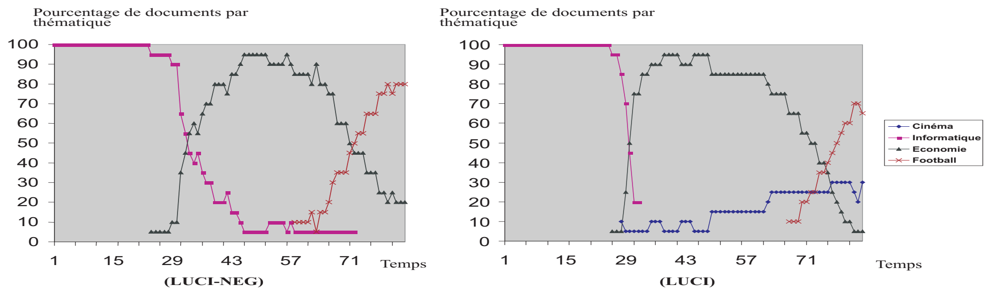

FIG. 1 – *LUCI-NEG, LUCI* : Pourcentage de documents par thématique proposés à l'utilisateur

propose pas. Cela est dû à la capacité du système *LUCI* de détecter les classes de documents d'intérêts réguliers. La thématique *Informatique* est d'un intérêt spontané pour l'utilisateur. En effet l'utilisateur ne s'intéresse plus à cette thématique à partir de l'itération 14. On remarque que *LUCI-NEG* continue à proposer d'une manière importante des documents de cette thématique jusqu'à l'itération 72, alors que le nombre de documents de la thématique *Informatique* proposé par *LUCI* décroît très rapidement pour s'annuler à l'itération 28. Les mêmes constatations sont faites pour la thématique *Économie*. Cela est dû à la capacité de *LUCI* de détecter les classes de documents d'intérêts spontanés et d'en diminuer constamment et rapidement le nombre de documents proposés appartenant à ces classes.

5 Travaux antérieurs

Récemment beaucoup de systèmes de filtrage d'informations et de documents ont été développés. *NewsWeeder*, un système de filtrage de dépêches d'informations sur le Net, où différentes alternatives pour la modélisation statique des intérêts de l'utilisateur sont décrites [Lang, 1995]. *WebMate* utilise plusieurs vecteurs TF-IDF pour l'apprentissage des intérêts de l'utilisateur [Chen et Sycara, 1998]. *Fab* est un système adaptatif pour la recommandation de pages Web, où l'utilisateur est modélisé par un seul vecteur [Balbanovic, 1997]. *Alipes* utilise trois vecteurs TF-IDF pour l'apprentissage du profil long-terme et court-terme de l'utilisateur [Widyantoro *et al.*, 1999].
[Kohrs et Merialdo, 2001] utilisent l'entropie et de la variance pour la prévision des besoins de l'utilisateur. Toutes ces recherches ne prennent pas en considération la repartition dans le temps des classes de documents consultés par l'utilisateur.

6 Conclusion et Perspectives

Partant du constat que les systèmes de personnalisation d'informations actuels ne prennent pas en considération le facteur temporel dans la proposition de documents aux utilisateurs, nous proposons dans cet article une approche prenant en compte l'historique des actions de l'utilisateur et leur évolution dans le temps. Cela en avantageant les classes de documents régulièrement consultées par l'utilisateur, par rapport à celles concentrées sur de courtes périodes. L'outil *LUCI* permet de détecter ces classes et de les avantager. Les premiers résultats obtenus ont montré l'intérêt de notre approche.

Une première perspective de recherche est de considérer la représentation des documents et notamment d'utiliser des approches comme *LSI* [Deerwester *et al.*, 1990] ou *TF-IDF* [Salton et Gill, 1983] pour l'assignation des poids. Une autre perspective de recherche consisterait à proposer une variante prenant en compte les réactions positives et négatives de l'utilisateur sur la pertinence des documents.

Références

[Arezki *et al.*, 2003] R. Arezki, A. Mokrane, G. Dray, P. Poncelet, et D.W. Pearson. Luci : A personnalization documentary system based on the analysis of the user's actions. *Rapport de recherche interne, Centre LGI2P*, 2003.

[Balbanovic, 1997] M. Balbanovic. An adaptative web page recommendation service. *In Proceeding of the First International Conference on Autonomous Agents*, pages 378–385, 1997.

[Chen et Sycara, 1998] L. Chen et K. Sycara. Webmate : Personal agent for browsing and searching. *In Proceeding of the Second International Conference on Autonomous Agents*, pages 132–139, 1998.

[Deerwester *et al.*, 1990] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, et R. Harsman. Indexing by latent semantic analysis. *Journal of the Society for Information Science, vol. 41*, pages 391–407, 1990.

[Kohrs et Merialdo, 2001] A. Kohrs et B. Merialdo. Filtering for new users by smart object selection. *Proceedings on Media Features (ICMF)*, 2001.

[Lang, 1995] K. Lang. Newsweeder : Learning to filter netnews. *Proceedings of Machine Learning Conference*, pages 331–339, 1995.

[Salton et Gill, 1983] G. Salton et M.J Mc Gill. Introduction to modern information retrieval. *New York : McGraw-Hill*, 1983.

[Widyantoro *et al.*, 1999] D.H. Widyantoro, T.R Ioerger, et J. Yen. An adaptative algorithm for learning changes in user interests. *Proceedings of the Eighth International Conference on Information and Knowledge Management*, pages 405–412, 1999.

Summary

The effective learning of a user profile is a challenge because it changes continuously. In this paper, we propose a new approach for learning the user long-term profile for textual document filtering . In this framework, the documents consulted by the user are classified in a dynamic way, then we analyze the distribution in the time of the document classes. The approach aim is to determine, as well as possible, the user interests in terms of document classes. An empirical study confirms the relevance of our approach.