Une nouvelle approche pour la modélisation du profil de l'utilisateur dans les systèmes de filtrage d'information: le modèle de filtre détecteur de nouveauté

Randa Kassab, Jean-Charles Lamirel, Emmanuel Nauer

Laboratoire LORIA
Campus scientifique, BP 239
54506 Vandoeuvre-lès-Nancy Cedex
France

{kassabr, lamirel, nauer}@loria.fr

RÉSUMÉ. *Cet article présente un mécanisme original pour la modélisation du profil de l'utilisateur dans les systèmes de filtrage d'information basés sur le contenu. Ce mécanisme repose sur un modèle de filtre basé sur la détection de nouveauté. En exploitant les bouclages de pertinence positif et négatif, ce modèle permet à la fois de construire incrémentalement une représentation synthétique, ou profil, du besoin de l'utilisateur et d'adapter ce profil selon le changement de ses centres d'intérêt. Une expérimentation du fonctionnement de ce modèle, menée sur un corpus de plusieurs milliers de sites web référencés dans une des catégories principales de l'annuaire ouvert DMOZ, est également décrite. Cette expérimentation permet de démontrer l'efficacité du modèle proposé pour l'analyse du besoin de l'utilisateur et pour l'identification des alternatives utiles à la représentation de ce besoin.*

ABSTRACT. *In this paper, we present an original mechanism based on a novelty detector filter for modelling user's profile in the context of content-based filtering systems. This filter learns possibly evolving user's need by means of positive and negative user's relevance feedback. An experiment on the behaviour of this filter that has been conduced on a corpus of several thousands of web sites taken from one of the main categories of the open directory DMOZ is also described. This experiment demonstrates the accuracy of our filter both for analyzing user's need and for highlighting useful alternatives for representing this need.*

MOTS-CLÉS : *détection de nouveauté, modèle utilisateur, bouclage de pertinence, bouclage négatif, filtrage d'information basé sur le contenu, web*

KEYWORDS: *novelty detection, user model, relevance feedback, negative relevance feedback, content-based information filtering, web.*

1. Introduction

La fourniture d'information à la demande peut être vue comme un processus consistant à adapter de manière dynamique la distribution d'information en tenant compte à la fois de l'évolution des besoins des utilisateurs et des nouvelles informations à distribuer. La modélisation du profil de l'utilisateur est une tâche centrale dans ce processus qui conditionne largement son efficacité. Cette tâche nécessite, d'une part, de représenter les centres d'intérêts de l'utilisateur dans le système, et, d'autre part, d'adapter cette représentation aux changements des centres d'intérêts de l'utilisateur au cours du temps.

De nombreuses méthodes de modélisation du profil de l'utilisateur ont été proposées dans les systèmes de filtrage d'information à la demande. Trois modes différents d'acquisition du profil peuvent être distingués. Le mode manuel dans lequel le profil est complètement acquis par intervention directe de l'utilisateur, le mode automatique qui utilise les informations fournies au cours d'un processus de bouclage de pertinence pour créer, puis pour affiner, le profil, et le mode mixte, dans lequel l'utilisateur conserve le contrôle sur le profil généré par le système. Les expériences comparatives menées entre ces différents modes tendent à prouver que le mode automatique est le plus avantageux. Une des raisons est qu'il ne souffre pas du risque le plus important inhérent au mode manuel qui est que les utilisateurs ne parviennent pas à construire leur profil, ou soient lents à le mettre à jour. Dans le mode mixte, les utilisateurs ne semblent pas parvenir à améliorer de manière sensible la qualité d'un profil appris par le système (Annika, 2004). Le comportement optimal du mode automatique reste cependant directement dépendant de celui du bouclage de pertinence. Dans le contexte du modèle vectoriel (Salton et al., 1975) où il est le plus utilisé, le bouclage de pertinence présente cependant de nombreux défauts. Il s'agit souvent d'un processus sans mémoire qui ne tient compte des décisions de l'utilisateur que de manière ponctuelle pour construire son profil. Le mécanisme de bouclage vectoriel le plus courant, proposé par (Rocchio, 1971), présente également un déséquilibre significatif dans la gestion des bouclages positif (consistant à traiter les votes positifs sur les documents, ou choix) et négatif (consistant à traiter les votes négatifs sur les documents, ou rejets). En outre, il a été montré que, dans certaines conditions, le bouclage négatif présente un comportement paradoxal qui rend ainsi incohérent le comportement général du processus de bouclage (Dunlop, 1991). Les modifications apportées à ce mécanisme (Ide, 1971) ne corrigent pas ces problèmes, et tendent même, dans certains cas, à les amplifier (Lamirel, 1995).

Le mécanisme de traitement automatique que nous proposons est donc essentiellement un modèle à long terme utilisant, de manière homogène, des bouclages positif et négatif de pertinence. Ce mécanisme est basé sur le modèle du filtre détecteur de nouveauté inspiré des travaux réalisés par (Kohonen, 1984) et par (Lamirel, 1995). Il permet d'accéder à la synthèse incrémentale des caractéristiques des votes de l'utilisateur et à la description automatique d'alternatives au besoin

exprimé par ce dernier. Il permet également d'évaluer le comportement de l'utilisateur en mesurant son degré de contradiction. Notre expérimentation sur le fonctionnement du filtre détecteur de nouveauté a permis de montrer l'efficacité de ce filtre pour l'analyse des votes de l'utilisateur et, conjointement, pour l'apprentissage du profil de ce dernier. Elle nous a également permis d'illustrer comment le bouclage négatif pouvait être utilisé pour construire et ajuster le profil de l'utilisateur de manière à améliorer la performance générale du système de filtrage.

Cet article est organisé comme suit : la prochaine section est consacrée à la description des mécanismes de base qui gouvernent la construction du profil de l'utilisateur à partir du modèle du filtre détecteur de nouveauté. Nous décrivons par la suite notre expérimentation pour l'évaluation du modèle. Nous présentons enfin nos conclusions et nos perspectives concernant ce modèle.

2. Modélisation du profil de l'utilisateur

2.1. Le modèle de filtre détecteur de nouveauté

Le rôle initial du filtre détecteur de nouveauté est de caractériser, après apprentissage, les propriétés nouvelles d'une donnée relativement à un ensemble de données de référence (les données d'apprentissage) ; les propriétés sont dites nouvelles si elles ne sont pas représentées dans les données de référence. Conjointement à son rôle initial, ce type de filtre permet de caractériser les propriétés communes de cette même donnée par rapport aux données de référence (Kassab, 2004). Le filtre est construit à partir d'un ensemble de données de référence représentées sous forme vectorielle. Il définit simultanément le sous-espace vectoriel de l'espace de description qui synthétise la nouveauté par rapport aux données de référence, et le sous-espace complémentaire qui synthétise l'habituation par rapport à ces mêmes données. La construction du filtre se base sur le théorème de Greville (Kohonen, 1984). Ce théorème s'exprime sous forme simplifiée de la manière suivante :

$$\phi_k = \phi_{k-1} - \frac{x'_k x'^T_k}{\left\| x'_k \right\|^2}$$

où

$X_k = [x_1, x_2, ..., x_k]$ représente les k données de référence utilisés pour l'apprentissage du filtre ; $x'_k = \phi_{k-1} x_k$ représente la projection orthogonale du vecteur x_k dans le sous-espace de nouveauté ϕ_{k-1} qui est orthogonal au espace défini par les k-1 premières données de référence ; $\|x\|$ représente la norme du vecteur x ; et la récursivité commence par $\phi_0 = I$.

2.1.1. Proportion de nouveauté

La proportion de nouveauté qu'il est possible d'associer à une donnée u, relativement à un ensemble de données de référence, peut être obtenue à partir de la norme du vecteur de nouveauté $\phi.u$ associé à cette donnée.

$$N_u = \frac{\|\phi \cdot u\|}{\|u\|}$$

Étant donné que l'espace de nouveauté et l'espace d'habituation sont orthogonaux, il est aussi possible de calculer la proportion complémentaire, à savoir la proportion d'habituation (voir figure 1):

$$H_u = 1 - N_u$$

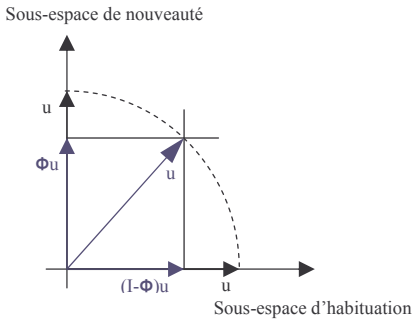

Figure 1. *Interprétation des proportions de nouveauté et d'habituation d'une donnée.*

2.1.2. Concept de saturation

La saturation d'un filtre détecteur de nouveauté peut être comprise comme l'inaptitude de ce filtre à extraire des caractéristiques nouvelles par rapport à celles des données de référence. Elle correspond en fait à l'apprentissage par le filtre de l'ensemble des descripteurs de l'espace de représentation des données. Il n'est plus possible de générer un filtre détecteur de nouveauté qui, quelque soit la donnée, produise un vecteur de nouveauté qui ne soit pas nul. Ce cas peut survenir si le nombre de données de référence est tel qu'il engendre un sous-espace dont la dimension est égale à la dimension de l'espace de représentation des données (voir figure 2).

La proportion d'habituation permet d'obtenir une bonne estimation de la saturation. Cette estimation correspond à la moyenne des proportions d'habituation de l'ensemble des descripteurs de l'espace de représentation.

$$S = \frac{\sum_{t \in U} H_t}{|U|}$$

où H_t est la proportion d'habituation associée au descripteur t dans l'espace de représentation U, et $|U|$ est le nombre de descripteurs de l'espace U.

Dans le cas où le filtre est utilisé pour le bouclage de pertinence, l'estimation de la saturation s'avère très importante. En effet, plus la saturation est importante, plus le comportement de l'utilisateur peut être considéré comme imprécis. En fonction du contexte, la saturation permet donc, soit de moduler l'analyse des votes, soit de tirer des conclusions sur le type de recherche mené par l'utilisateur (précise ou exploratoire).

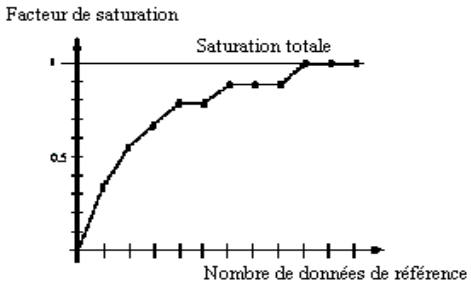

Figure 2. *Évolution de la saturation en fonction du nombre de données de référence apprisses.*

2.1.3. Trace du filtre détecteur de nouveauté

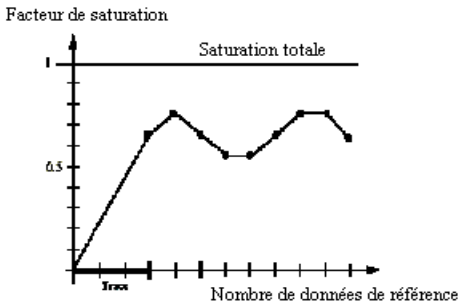

Figure 3. *Évolution de la saturation en fonction du nombre de données de référence apprisses, pour une trace courte.*

Nous définissons la trace du filtre détecteur de nouveauté comme le nombre maximal de données qui peuvent être prises en compte lors de l'apprentissage avant

d'éliminer les données les plus anciennement apprises. Dans le cadre du bouclage de pertinence, le concept de trace est également important puisqu'il conditionne la faculté d'évolution du profil de l'utilisateur sans nécessairement tendre vers une saturation totale. Contrairement au cas de la figure 2 où le facteur de saturation ne fait que croître jusqu'à atteindre sa valeur limite, si aucune trace n'est prise en compte (trace infinie), le cas de la figure 3 illustre la régulation du facteur de saturation générée par le choix d'une trace courte.

2.2. Utilisation des filtres détecteurs de nouveauté pour la modélisation du profil de l'utilisateur

Le principe de la détection de nouveauté est particulièrement intéressant à appliquer pour analyser les votes de l'utilisateur afin de construire son profil de manière automatique. Dans les mécanismes de bouclage de pertinence, les votes de l'utilisateur sont classés en deux principaux types : choix et rejets. Cela amène naturellement à structurer le composant destiné à modéliser le besoin de l'utilisateur sous la forme de deux filtres détecteurs de nouveauté, chacun étant associé à un type de vote. Le rôle de chaque filtre est d'assurer le traitement des votes associés à son type, de manière à en extraire les caractéristiques synthétiques en terme d'habituation et de nouveauté. Pour simplifier le discours, nous emploierons par la suite le terme détecteur de choix pour désigner le filtre détecteur de nouveauté associé au traitement des choix, et, détecteur de rejet pour désigner le filtre détecteur de nouveauté associé au traitement des rejets.

L'interaction entre le système et l'utilisateur peut être vue comme une suite d'étapes. À chaque étape, le système fournit à l'utilisateur un ensemble de documents qui doivent être jugés par ce dernier. L'ensemble des documents ayant fait l'objet d'un même jugement par l'utilisateur (choix ou rejet) peut ainsi être considéré comme un nouvel ensemble de données de référence destinées à alimenter un détecteur de nouveauté spécifique à la catégorie de jugement concernée. Les détecteurs de nouveauté (détecteur de choix et détecteur de rejet) sont construits par un algorithme d'apprentissage spécifique basé sur le théorème de Greville. Cet algorithme est détaillé dans (Kassab, 2004). A l'issue de cet apprentissage, le contenu de chaque détecteur sera représentatif des caractéristiques de l'ensemble des documents ayant bénéficié d'un même jugement par l'utilisateur. Nous illustrons ci-après comment construire un profil représentatif du besoin de l'utilisateur en utilisant ces détecteurs.

2.2.1. Synthèse élémentaire de l'apprentissage

La construction d'un filtre détecteur de nouveauté basé sur les votes émis par l'utilisateur sur les documents ne permet pas d'obtenir directement une synthèse de ces votes en terme de descripteurs appris (cf. définition ci-dessous). Cette synthèse peut être construite par une projection des descripteurs de l'espace de représentation

sur le sous-espace d'habituation (I^{-}=I - $\varnothing$). Les valeurs obtenues pour cette projection — qui correspondent aux proportions d'habituation des descripteurs — permettent de partitionner l'ensemble des descripteurs en trois sous-ensembles distincts :

- le sous-ensemble des descripteurs habitués, ou appris, T_H, constitué des descripteurs pour lesquels la proportion d'habituation est supérieure à un seuil s_H.
- le sous-ensemble des descripteurs nouveaux T_N, constitué des descripteurs pour lesquels la proportion d'habituation est inférieure à un seuil s_N.
- le sous-ensemble des descripteurs neutres T_n, constitué des descripteurs pour lesquels la proportion d'habituation est comprise entre les seuils s_H et s_N.

La répartition des descripteurs en trois sous-ensembles peut être représentée graphiquement, comme le montre la figure 4.

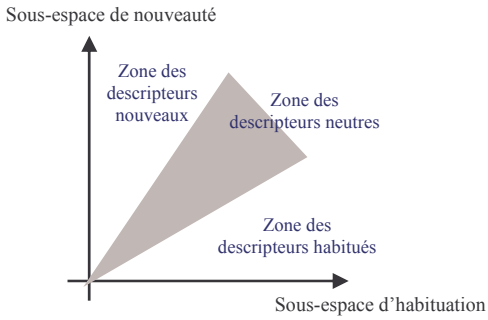

Figure 4. *Répartition des descripteurs en trois sous-ensembles*

2.2.2. Construction du profil

La partition des descripteurs en trois sous-ensembles permet de former directement un vecteur représentatif d'habituation et un vecteur représentatif de nouveauté par rapport aux données de référence. Le vecteur d'habituation s'exprime sous la forme :

$$P_H = \sum_{t \in T_H} H_t \quad \vec{t}$$

où $\vec{t}$ représente le vecteur directeur unité associé au descripteur t.

Le vecteur d'habituation engendré par le détecteur de choix fournit une représentation potentielle du profil de l'utilisateur. Il peut alors être utilisé comme profil dès lors que les votes de l'utilisateur ne consistent qu'en des documents choisis (bouclage positif de pertinence).

Le vecteur de nouveauté s'exprime sous la forme :

$$P_N = \sum_{t \in T_N} N_t \quad \vec{t}$$

Le vecteur de nouveauté engendré par le détecteur de rejet peut être utilisé comme profil dès lors que les votes de l'utilisateur ne consistent qu'en des documents rejetés (bouclage négatif de pertinence).

L'efficacité individuelle de chaque type de bouclage est primordiale, si l'on tient compte que chacun d'entre eux doit pouvoir intervenir de manière autonome. Néanmoins, la complémentarité de comportement entre ces deux types de bouclage est également primordiale, si l'on considère maintenant qu'ils interviennent la plupart du temps de manière simultanée. Dans les mécanismes classiques de bouclage de pertinence, les fonctions de reformulation reposent sur le même principe (Ide, 1971) (Rocchio, 1971) et ont un comportement incohérent, comme précisé en introduction. Contrairement aux mécanismes classiques, le modèle de filtre détecteur de nouveauté permet de proposer une fonction de reformulation dont le comportement est homogène, en combinant les résultats obtenus par les deux détecteurs : le profil de l'utilisateur est à la fois rapproché linéairement des documents choisis (vecteur d'habituation engendré par le détecteur des choix) et linéairement des alternatives des documents rejetés (vecteur de nouveauté engendré par le détecteur des rejet). Cette fonction est définie comme suit :

$$P' = P + \beta H_A + \gamma \frac{\|H_R\|}{\|N_R\|} N_R$$

où P représente un profil initial prédéfini par l'utilisateur (optionnel), H_A représente le vecteur d'habituation engendré par le détecteur des choix, N_R représente le vecteur de nouveauté engendré par le détecteur de rejet, β et γ sont des paramètres positifs qui contrôlent l'influence des bouclages positif et négatif sur la reformulation du profil.

3. Expérimentation

Dans cette section, nous décrivons en détail notre expérimentation sur le fonctionnement du filtre détecteur de nouveauté pour la représentation du besoin de l'utilisateur à travers les bouclages positif et négatif de pertinence. Le but premier de cette expérimentation est de démontrer l'efficacité de ce filtre pour la construction du profil de l'utilisateur. Le but secondaire est de montrer comment le bouclage négatif peut être employé pour construire et ajuster le profil en vue d'améliorer la performance du système de filtrage.

3.1. Corpus

Nous avons évalué le fonctionnement du filtre détecteur de nouveauté en utilisant deux corpus de sites web. Chaque corpus est constitué de deux ensembles :

un ensemble pour l'apprentissage du filtre et l'autre pour l'évaluation de son fonctionnement. Dans le premier corpus, les deux ensembles de sites sont assez similaires, alors que dans le deuxième corpus les deux ensembles de sites sont assez différents. Nous avons profité des catégories thématiques de l'annuaire DMOZ[1] pour la construction de ces corpus.

3.1.1. Indexation des catégories

Nous avons choisi de mener notre expérimentation sur l'ensemble des sous-catégories de la catégorie *Computers* qui représente elle-même une des catégories racine de la hiérarchie de DMOZ. Cette catégorie regroupe environ 150000 références de sites web. Chaque site est décrit synthétiquement par un titre et par un court résumé de son contenu. En utilisant un mécanisme d'indexation élémentaire — extraction de termes simples et composés, avec lemmatisation et utilisation d'une fréquence minimale d'apparition —, chacune des 42 sous-catégories de la catégorie *Computers* est transformée en un vecteur binaire (valeurs 0 ou 1) défini sur l'espace de description constitué par l'ensemble des index extraits de l'intégralité des descriptions des sites référencés (c'est-à-dire des intitulés des catégories et des termes extraits des descriptions synthétiques des sites associés).

Le but de l'indexation des catégories est celui de déterminer parmi celles-ci deux catégories proches et deux catégories distantes pour pouvoir évaluer le fonctionnement du filtre en considérant un large éventail de situations possibles. La distance entre les catégories a été calculée en utilisant la similarité cosinus. Les deux catégories les plus proches que nous avons déterminées selon cette mesure sont les catégories *Artificial Intelligence* (AI) et *Robotics* (RO) et les deux catégories les plus distantes sont les catégories *Organization* (OR) et Speech *Technology* (SP). Ces quatre catégories ont permis de construire deux corpus :

- le corpus AI-RO, formé par les catégories proches.
- le corpus OR-SP, formé par les catégories distantes.

Dans chacun de ces corpus constitué de deux catégories, nous avons choisi arbitrairement une catégorie de référence et une catégorie complémentaire (voir tableau 1). Les sites de la catégorie de référence sont considérés comme représentatifs des sites intéressants pour l'utilisateur (sites choisis), alors que ceux de la catégorie complémentaire sont considérés comme représentatifs des sites inintéressants pour l'utilisateur (sites rejetés).

Catégorie de référence	Catégorie complémentaire	Similarité cosinus
Artificial Intelligence	Robotics	0.37
Organization	Speech Technology	0.04

Tableau 1. *Catégories de référence et catégories complémentaires*

[1] http://www.dmoz.org/

3.1.2. Indexation des sites

Les sites associés aux différentes catégories des deux corpus sont indexés à partir de leur description synthétique (titre et résumé dans la hiérarchie de DMOZ). Un seuillage fréquentiel est utilisé pour éliminer les index de fréquence trop basse. L'espace de description est constitué par l'ensemble des index dont la fréquence est supérieure au seuil. Chacun des sites est finalement transformé en un vecteur binaire défini sur cet espace. Le tableau 2 résume les résultats du prétraitement des sites des deux corpus.

	Corpus AI-RO	Corpus OR-SP
Dimension de l'espace de description global (catégorie de référence + catégorie complémentaire)	432	204
Dimension de l'espace de description des sites associés à la catégorie de référence	288	101
Dimension de l'espace de description des sites associés à la catégorie complémentaire	259	145
Nombre de descripteurs communs entre la catégorie de référence et la catégorie complémentaire	115	42
Nombre moyen de descripteurs par site	1.972	1.238
Nombre des sites de la catégorie de référence	430	147
Nombre des sites de la catégorie complémentaire	353	224
Fréquence de seuillage	2	2

Tableau 2. *Résultats de l'indexation des sites des corpus.*

3.2. Évaluation du fonctionnement du filtre détecteur de nouveauté

Le fonctionnement du filtre détecteur de nouveauté a été évalué en utilisant les critères classiques de précision et de rappel. La précision représente le pourcentage de sites restitués qui sont effectivement pertinents par rapport à l'ensemble des sites restitués, alors que le rappel représente le pourcentage de sites pertinents qui ont été trouvés par rapport à l'ensemble des sites pertinents existants.

Dans la première partie de l'expérimentation, nous avons utilisé un seul filtre détecteur de nouveauté pour analyser les votes positifs de l'utilisateur (détecteur de choix). Pour chaque corpus, nous avons réalisé l'apprentissage en utilisant des sites appartenant à la catégorie de référence et avons ensuite évalué la capacité du filtre à détecter l'habituation, à savoir les sites appartenant à la catégorie de référence, ainsi que sa capacité à détecter la nouveauté, à savoir les sites appartenant à la catégorie complémentaire. À chaque étape, nous simulons les votes de l'utilisateur en alimentant le détecteur de choix par k sites tirés de la catégorie de référence. L'apprentissage du détecteur permet de générer deux vecteurs : un vecteur d'habituation et un vecteur de nouveauté. Nous mettons par la suite en correspondance la totalité des sites du corpus avec le vecteur d'habituation en

utilisant la similarité cosinus et nous trions les sites par ordre décroissant de pertinence relativement à ce vecteur. Pour évaluer la capacité du détecteur à détecter l'habituation, nous calculons la précision pour certaines proportions de rappel (P25%, P50%, P75%) et également le pourcentage de sites effectivement pertinents parmi les 5 (Top5) et les 10 (Top10) premiers sites restitués. Les mêmes opérations sont faites avec le vecteur de nouveauté pour évaluer la performance du détecteur relativement à la détection de nouveauté.

Dans la seconde partie de l'expérimentation, nous avons employé un autre filtre détecteur de nouveauté pour analyser les votes négatifs de l'utilisateur (détecteur de rejet) afin d'améliorer le fonctionnement du détecteur de choix ou encore de bénéficier des vecteurs créés par ce détecteur, au cas où les vecteurs engendrés par le détecteur de choix ne seraient pas significatifs : ce cas peut survenir si aucun descripteurs n'est effectivement appris ou si tous les descripteurs appris sont non discriminants. L'apprentissage de ce nouveau détecteur a été réalisé en utilisant les sites appartenant à la catégorie complémentaire dans chacun des corpus.

Pour améliorer la qualité des résultats nous avons cherché à tirer parti de la combinaison des informations fournies conjointement par les deux détecteurs. Pour cela, nous avons comparé les deux vecteurs d'habituation engendrés par le détecteur de choix et par le détecteur de rejet et nous en avons éliminé les descripteurs communs qui représentent nécessairement des descripteurs non discriminants. Nous avons également amélioré la qualité des vecteurs de nouveauté en supprimant les descripteurs communs entre ces deux vecteurs.

3.3. Résultats et discussion

3.3.1. Détection de nouveauté

La figure 5 montre que le principe de détection de nouveauté est globalement efficace, même avec un faible nombre de sites appris. Dans le cas du corpus AI-RO, les valeurs de précision en tête de liste de pertinence (Top5 et Top10) sont très rapidement élevées et stables ; ce qui prouve que le détecteur de choix fournit immédiatement une bonne représentation des sites de la catégorie complémentaire RO. L'évolution de la précision pour les fortes valeurs de rappel (P75%) est significative. On constate cependant des accidents (baisse subite, puis remontée de la précision) dans l'évolution de la précision avec l'apprentissage pour les faibles valeurs de rappel (P25%), ainsi que pour les valeurs moyennes de rappel (P50%). Dans le cas du corpus OR-SP, la précision en tête de liste de pertinence n'évolue que lentement. La précision pour certaines proportions de rappel augmente par contre plus significativement que dans le cas du corpus AI-RO. Il est également possible de remarquer que cette précision évolue de manière homogène, contrairement à celles du corpus AI-RO. Nous pouvons donc conclure que les résultats sont plus précis et que la détection de la nouveauté s'avère plus rapide dans le cas où les catégories

sont fortement corrélées (corpus AI-RO), alors que ce même apprentissage s'avère plus homogène, tout en restant cependant moins précis, dans le cas où les catégories sont faiblement corrélées (corpus OR-SP).

Deux hypothèses sont possibles pour expliquer les différences de résultats entre les deux corpus. Le nombre moyen de descripteurs par site est environ deux fois plus faible dans le cas du corpus OR-SP (voir tableau 2) ce qui rend l'apprentissage moins précis. Dans ce cas, il est en effet plus difficile au mécanisme d'apprentissage d'assimiler rapidement l'ensemble des caractéristiques à rattacher à la catégorie de référence, ce qui, en parallèle, ne lui permet pas d'isoler rapidement les caractéristiques de la catégorie complémentaire. Ce phénomène est amplifié par le fait que les deux catégories du corpus OR-SP sont peu corrélées étant donné qu'elles ne possèdent que peu de descripteurs communs. En effet, dans le cas où les catégories sont corrélées, comme dans celui du corpus AI-RO, les chances sont plus importantes d'apprendre rapidement des descripteurs non discriminants. Étant donné que ces descripteurs sont éliminés parallèlement du sous-espace de nouveauté, ce phénomène augmentera d'autant la qualité du vecteur de nouveauté. Enfin, les accidents apparaissant dans les valeurs de précision dans le cas du corpus AI-RO peuvent s'expliquer par la présence de sites qui bénéficient d'une indexation ambiguë et/ou d'une classification ambiguë dans la hiérarchie originale de DMOZ.

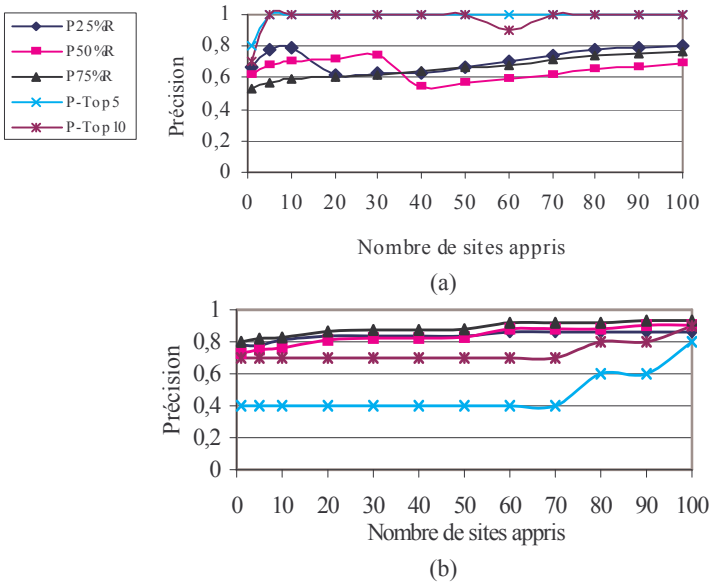

Figure 5. *Évaluation de la détection de nouveauté (a) en utilisant le corpus AI-RO (b) en utilisant le corpus OR-SP*

Comme précisé précédemment, nous avons utilisé un autre détecteur de nouveauté pour traiter les votes négatifs de l'utilisateur afin d'améliorer ces résultats. La figure 6 montre les nouveaux résultats. La comparaison entre les figures 5-a et 6-a permet de remarquer l'augmentation importante de la précision en fonction de rappel pour le corpus AI-RO. L'examen des figures 5-b et 6-b fait également apparaître que les résultats sont plus précis après l'exploitation des rejets pour le corpus OR-SP. Les valeurs de la précision en tête de liste de pertinence sont rapidement élevées et stables.

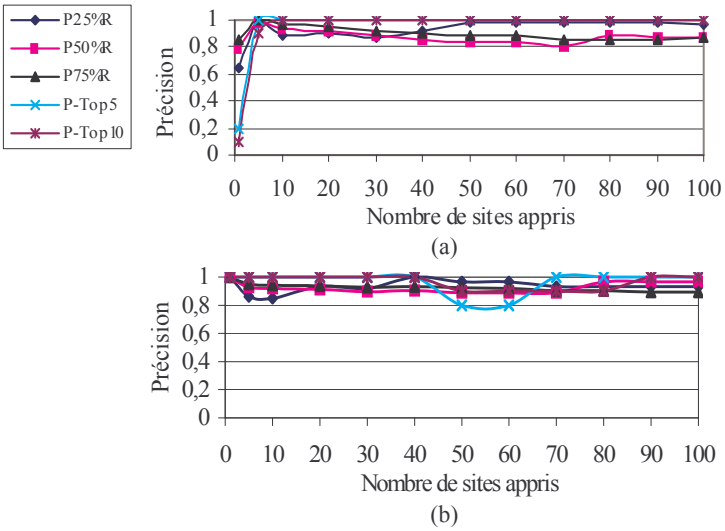

Figure 6. *Évaluation de la détection de nouveauté après l'exploitation des rejets (a) en utilisant le corpus AI-RO (b) en utilisant le corpus OR-SP*

3.3.2. Détection d'habituation

L'examen de la figure 7 montre encore une fois que l'apprentissage est globalement efficace, même avec un faible nombre de sites appris. Contrairement au cas de la détection de nouveauté, l'apprentissage s'avère meilleur sur le corpus OR-SP que sur le corpus AI-RO. La précision en tête de liste de pertinence est parfaite dans le cas du corpus OR-SP. Cette même précision reste très bonne, bien que plus instable, dans le cas du corpus AI-RO. Pour le corpus OR-SP, les valeurs de précision sont stables quelque soit le rappel. Pour le corpus AI-RO les valeurs de précision sont très bonnes pour les faibles valeurs de rappel (P25%), mais plus on cherche à augmenter le rappel, plus la précision chute de manière importante.

Selon l'hypothèse de l'apprentissage de descripteurs non discriminants (voir section 3.3.1), il est logique que la tendance s'inverse concernant la nouveauté et l'habituation entre les deux corpus, autrement dit que la qualité de l'habituation soit meilleure dans le cas du corpus OR-SP, alors que la qualité de la nouveauté est

meilleure dans le cas du corpus AI-RO. En effet, comme le corpus OR-SP contient peu de descripteurs non discriminants, seuls les descripteurs discriminants ont tendance à être appris ; ce qui améliore sensiblement la qualité de l'habituation. À l'inverse, comme nous l'avons déjà fait remarquer précédemment, il s'avère plus facile de détecter la nouveauté dans le cas du corpus AI-RO, étant donné que l'habituation, en apprenant plus facilement des descripteurs non discriminants, les élimine parallèlement de la nouveauté.

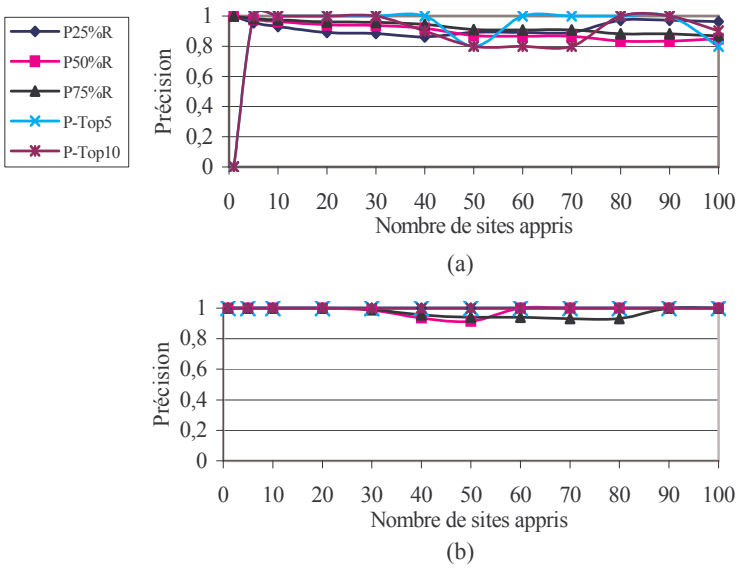

Figure 7. *Évaluation de la détection d'habituation (a) en utilisant le corpus AI-RO (b) en utilisant le corpus OR-SP*

Dans le cas des deux corpus, les résultats obtenus pour la détection d'habituation sont presque parfaits en exploitant seulement les choix. Par conséquent, l'amélioration de ces résultats par l'exploitation des rejets (figure 8) est moindre que dans le cas de la détection de nouveauté.

4. Conclusion

Dans cet article nous avons présenté le modèle de filtre détecteur de nouveauté pour la modélisation du profil de l'utilisateur dans un système de filtrage basé sur le contenu. Ce modèle permet d'apprendre le besoin de l'utilisateur à travers des bouclages de pertinence positif et négatif. Notre expérimentation sur le fonctionnement de ce modèle a permis de montrer sa pertinence pour la construction

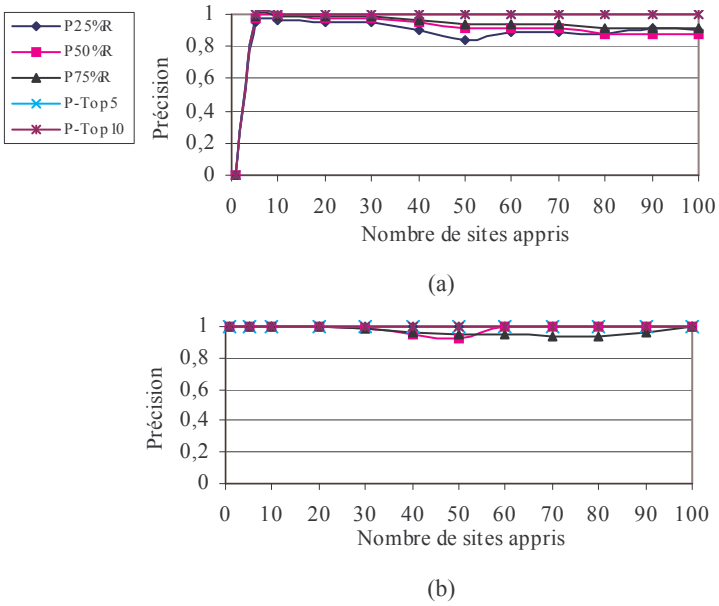

Figure 8. *Évaluation de la détection d'habituation après l'exploitation des rejets (a) en utilisant le corpus AI-RO (b) en utilisant le corpus OR-SP*

des profils des utilisateurs dès lors que la qualité de l'indexation est suffisamment riche, comme c'est ce cas de celle obtenue à partir des descripteurs synthétiques de sites associés aux catégories de DMOZ. De nombreux points méritent cependant d'être approfondis, en particulier ce qui concerne l'utilisation d'une indexation directement basée sur le contenu. Dans nos futures expérimentations, nous étudierons donc plus particulièrement l'effet du processus d'indexation sur le fonctionnement du modèle. Nous mettons également en place un processus de comparaison approfondi avec les modèles existants.

5. Bibliographie

Annika W., User involvement in automatic filtering: an experimental study. *In Information Processing and Management*. 14(2-3):201-237, 2004.

Dunlop M. D., Multimedia Information Retrieval. PhD thesis. Computing Science Dep. University of Glasgow. 1991.

Ide E., New experiments in relevance feedback. In *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall Inc., Englewood Cliffs, USA. 1971.

Kassab R., Méthodes d'analyse intelligente des profils utilisateurs et de l'évolution du contenu des fonds pour le filtrage d'information. Master Report. Université Henri Poincaré, Nancy1, France. 2004.

Kohonen T., *Self organization and associative memory*. Springer Verlag, USA. 1984.

Lamirel J-C., Application d'une approche symbolico-connexionniste pour la conception d'un système documentaire hautement interactif. PhD thesis. Université Henri Poincaré, Nancy1, France. 1995.

Rocchio J.J., Relevance feedback in information retrieval. In The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall Inc., Englewood Cliffs, USA. 1971.

Salton G., Yang C. S., Wong A., A vector space model for automatic indexing. Communication of the ACM. 18(11):229-237, 1975.