Filtrage des sites Web à caractère violent par analyse du contenu textuel et structurel

Radhouane Guermazi*, Mohamed Hammami** et Abdelmajid Ben Hamadou*

*MIRACL-ISIMS, Route Mharza Km 1 BP 1030 Sfax Tunisie
rguermazi@laposte.net
abdelmajid.benhamadou@isimsf.rnu.tn
http://www.isimsf.rnu.tn/
**MIRACL-FSS, Route Sokra Km 3 BP 802, 3018 Sfax Tunisie
mohamed.hammami@ec-lyon.fr

Résumé. Dans cet article, nous proposons une solution pour la classification et le filtrage des sites Web à caractère violent. A la différence de la majorité de systèmes commerciaux basés essentiellement sur la détection de mots indicatifs ou l'utilisation d'une liste noire manuellement collectée, notre solution baptisée, « WebAngels Filter », s'appuie sur un apprentissage automatique par des techniques de data mining et une analyse conjointe du contenu textuel et structurel de la page Web. Les résultats expérimentaux obtenus lors de l'évaluation de notre approche sur une base de test sont assez bons. Comparé avec des logiciels, parmi les plus populaires, « WebAngels Filter » montre sa performance en terme de classification.

1 Introduction

L'Internet représente un extraordinaire outil d'accès à un ensemble quasi infini de ressources et un puissant outil de communication. Elle prend une place grandissante dans la vie quotidienne et dans le monde professionnel. Le public qui y a accès est de plus en plus large, mais aussi de plus en plus jeune. Les enfants trouvent chaque jour un accès plus facile à la toile. Cet accès de plus en plus large ne va pas sans inconvénients, les sites à caractère adulte, violent, raciste exposent les enfants à des contenus qui peuvent heurter leur sensibilité, voire les choquer. En effet, ces sites sont souvent en accès libre, ce qui pose un problème évident vis à vis des enfants. Ces utilisations litigieuses de l'Internet, par des individus mal intentionnés, n'ont pas occulté les énormes possibilités de progrès personnel et social, d'enrichissement culturel et éducatif offertes par ce réseau. Ainsi, un ensemble de produits commerciaux sur le marché proposent des solutions de filtrage de sites Web. La majorité de ces produits traitent principalement le caractère adulte, alors que les autres caractères, comme le caractère néonazie, raciste et violent, ont été marginalisé. C'est ce dernier caractère qui sera traité dans cet article. La section suivante présente une revue de littérature sur les travaux qui ont porté sur le filtrage de sites web. Nous décrivons dans la section 3 notre approche de classification des sites Web à caractère violent par une analyse du contenu textuel et structurel des pages Web. Les résultats de l'expérimentation de l'approche proposée seront détaillés dans la section 4. La section 5

décrit l'architecture et le principe de fonctionnement de notre solution « WebAngels Filter » ainsi que la comparaison des résultats de ce dernier avec les logiciels les plus connus sur le marché. Enfin une conclusion et quelques perspectives feront l'objet de la dernière section.

2 Filtrage Web

Plusieurs techniques de filtrage Web ont été proposées pour bloquer les pages Web à caractère litigieux. Parmi ces techniques, on peut citer :

1. La technologie de l'étiquetage PICS[1] (Platform for Internet Content Selection) : c'est un standard de programmation permettant de véhiculer des informations concernant le genre de contenus qui sont représentés sur les sites. En se basant sur ce codage, le navigateur prendra la décision d'afficher ou non une page. Il est à noter que l'efficacité du PICS est relative, en effet, elle dépend fortement de l'engagement des concepteurs des sites à étiqueter leurs pages, en absence d'organisme qui les oblige à le faire.
2. La liste noire : représente un ensemble de sites, motifs génériques, ou domaines à exclure de la navigation. On garde donc la possibilité de naviguer librement d'un site à un autre, ce qui permet de conserver la spécificité de l'Internet, tout en restreignant les risques d'accéder à un site inapproprié. Cependant il est difficile de regrouper tous les sites inappropriés puisque de nouveaux sites apparaissent chaque jour. De ce fait, une liste noire ne peut jamais être exhaustive.
3. La liste blanche : contient l'ensemble de sites sur lesquels la navigation peut avoir lieu. C'est donc un ensemble de sites autorisés. Toute tentative d'accès à n'importe quel site ne figurant pas sur cette liste blanche sera automatiquement refusée. Les éditeurs de logiciels constituent rarement de telles listes blanches, dont l'élaboration est le plus souvent laissée aux parents. Cette solution, qui restreint strictement la navigation à un « jardin d'enfants », peut servir à sécuriser la navigation de très jeunes enfants. Ces listes demandent une vérification régulière par un administrateur, en effet, il arrive souvent que certains sites à contenu tout a fait licite disparaissent en laissant leurs adresses récupérées par des sites inappropriés.
4. Filtrage par le contenu
   (a) Filtrage par mots clés : il s'agit d'effectuer un contrôle des contenus par mots clés ou phrases clés. À l'aide d'un outil d'analyse de texte, le programme vérifie tous les mots de la page avant que celle-ci ne s'affiche. Si un mot « interdit » est décelé, que ce soit dans une page Web, dans le titre d'un groupe de discussion ou dans celui d'un forum de dialogue en direct, le logiciel de filtrage bloquera l'affichage de ces données. Selon Hochheiser (1997), l'un des problèmes avec ce genre de recherche est que seuls des mots bruts et décontextualisés sont recherchés, une telle recherche n'est pas capable par exemple de faire la différence entre des sites qui luttent contre la torture et des sites qui en parlent.
   (b) Filtrage par apprentissage « intelligent » : La classification des sites inappropriés par une analyse intelligente du contenu Web s'intègre dans une problématique plus générale, celle des systèmes automatiques de classification et de catégorisation de

[1] http ://www.w3.org/PICS.

sites Web. La réalisation de tels systèmes doit s'appuyer sur un processus d'apprentissage automatique et plus précisément sur un apprentissage supervisé. On peut distinguer au moins trois catégories de filtrage par apprentissage. La première concerne les travaux qui se basent sur une analyse du contenu textuel, par exemple, Du et al. (2003) représentent chaque document par un vecteur de mots et ils les classent en calculant la similarité entre eux. Caulkins et al. (2006) proposent une méthode de filtrage basée sur une analyse statistique des différents descripteurs textuels des pages Web à caractère adulte. Une évaluation des techniques de classification textuelle pour le filtrage du contenu raciste a été présentée dans (Vinot et al., 2003). La deuxième catégorie comprend les travaux basés sur une analyse du contenu structurel. Par exemple Lee et al. (2002) ont utilisé les réseaux de neurones associés à la connaissance de la structure d'une page Web pour élaborer leurs solutions de filtrage. Ho et Watters (2004) proposent une analyse statistique du contenu structurel d'une page Web en utilisant un classifieur bayésien. La troisième catégorie de filtrage concerne les travaux qui se basent sur une analyse du contenu visuel des pages Web. Par exemple Arentz et Olstad (2004) proposent une méthode de détection des images pornographiques pour discriminer les pages à caractère adulte. Quant à Denoyer et al. (2003), Grilhers et al. (2004) et Hammami et al. (2006), ils proposent une solution de filtrage des sites litigieux basée sur une analyse du contenu textuel, structurel et visuel d'une page Web.

Les travaux proposés pour la classification et le filtrage des pages Web à caractère litigieux sont assez nombreux et variés, Cependant la majorité de ces travaux ne traitent que le caractère adulte. Nous proposons dans la section suivante notre approche pour la classification des sites à caractère violent.

3 Approche proposée

On se place dans le cadre des systèmes automatiques de classification et de catégorisation de sites Web pour proposer une approche de classification des sites Web à caractère violent. En effet, pour classifier les sites à caractère violent et les sites normaux, nous nous sommes basés sur le processus d'extraction des connaissances à partir des données. Le principe général de l'approche de classification est le suivant : Soit S une population de sites concernés par le problème d'apprentissage. A cette population est associé un attribut particulier appelé « attribut classe » noté C. Cet attribut peut avoir deux valeurs, la valeur 0 si le site est violent et 1 si le site est normal. A chaque site s peut être associée sa classe C(s)

$$\begin{array}{c} C : S \text{ à } \Gamma=\{\text{Violent, nonViolent}\} \\ s \longrightarrow C(s) \end{array}$$

Dans notre étude nous cherchons un moyen pour prédire la classe C. La détermination de ce modèle de prédiction est liée à un vecteur de caractéristiques $\vec{X} = (X_i)_{1 \leq i \leq p}$ que nous avons établi a priori. Ce modèle de prédiction permet, pour un site s issu de S, pour lequel nous ne connaissons pas la classe C(s) mais nous connaissons son vecteur de caractéristiques, de prédire sa classe. La figure 1 illustre le schéma général de l'approche de classification proposée. Deux parties peuvent être distinguées : la première est celle de l'apprentissage consacrée à la

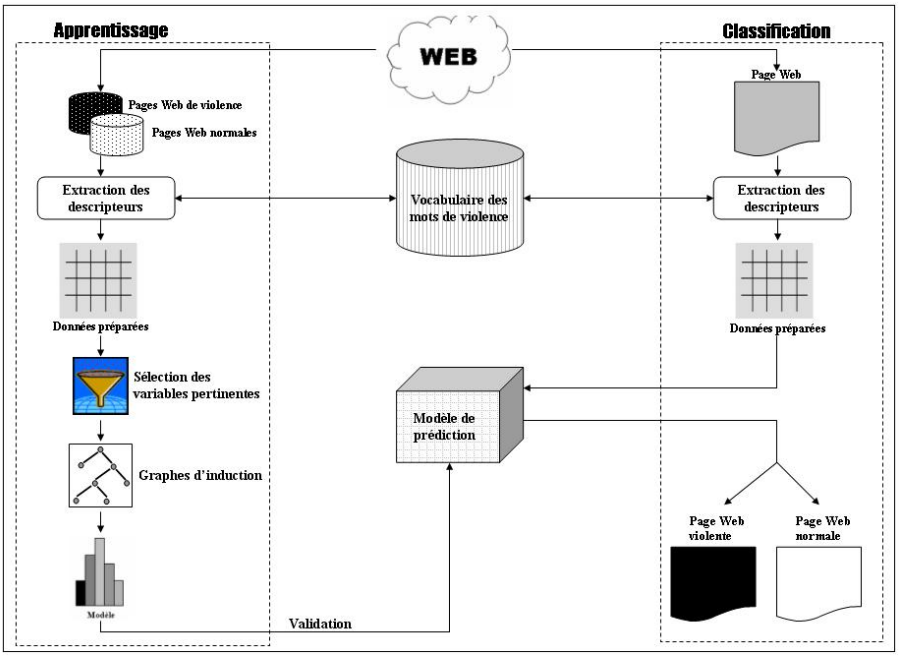

FIG. 1 – *Schéma général de l'approche proposée*

préparation du modèle de prédiction, la deuxième est celle de la classification des sites Web. Nous signalons que les modèles construits sont sensibles à la qualité des données qui leur sont fournies et nous avons été obligé de faire plusieurs itérations qui ont conduit à affiner la recherche et à élaborer de nouvelles variables ce qui nous a permis d'améliorer les résultats obtenus au fur et à mesure des différentes étapes.

Dans ce qui suit, nous détaillerons les différentes étapes de l'élaboration du modèle de prédiction des sites Web

3.1 Préparation des données

La préparation des données pour la phase d'apprentissage consiste à identifier les informations exploitables et vérifier leur qualité et leur efficacité afin de construire une table bi-dimensionnelle, à partir de notre corpus d'apprentissage. La recherche des attributs les plus informatifs est le point central de cette phase puisque c'est elle qui va conditionner la qualité des modèles établis lors de l'apprentissage, par conséquent cette étape a une influence directe sur la performance du classifieur.

3.1.1 Construction de la base d'apprentissage

Pour que l'apprentissage soit efficace, il faut que notre base d'apprentissage soit représentative de la population et que le nombre des éléments de la base sur lequel il est fait soit

important. Cette phase de collecte et de sélection des sites Web constitue une charge de travail considérable vu la diversité et le nombre énorme de sites Web sur Internet. Dans la collecte des sites violents, nous avons essayé d'avoir une base diversifiée en terme de :

- Contenu des sites : la recherche de ces sites s'est focalisée sur la guerre, l'attentat, la torture de prisonniers, l'assassinat, la violence explicite, la création de bombe, les films d'horreur, et les articles qui parlent des effets de la violence. Elle a été effectuée avec le moteur de recherche « Google » en se basant sur un ensemble de mots clés violents. Dans un premier temps on s'est intéressé sur le contenu textuel et ensuite sur les images qui représentent un contenu violent.
- Langues traitées : nous avons traité deux langues à savoir la langue anglaise et la langue française.
- Structure : certains sites collectés ne contiennent que du texte, d'autres ne contiennent que des images, et la majorité des sites contient les deux.

Pour la sélection des sites non violents, nous avons inclus ceux qui peuvent prêter à confusion, en particulier des sites qui luttent contre la violence et des sites de loi, etc. Le reste des sites a été choisi au hasard, on trouve alors des sites de téléchargement de jeu, des sites de codes sources, des sites de bandes dessinées, des sites éducatifs, des sites d'enfance, etc.

Notre base d'apprentissage se compose de 700 sites dont 350 sont violents, et 350 sites non violents.

3.1.2 Analyse du contenu textuel et structurel

L'analyse du contenu textuel et structurel d'une page vise à extraire des variables structurelles et textuelles permettant de mieux discriminer les pages Web violentes de celles inoffensives. Dans cette phase nous nous sommes basé sur les connaissances acquises suite à notre étude des travaux de recherche existants et sur la sélection manuelle des pages Web lors de la construction de notre base d'apprentissage.

La fréquence des mots interdits dans une page Web nous semble la variable la plus discriminante. C'est pourquoi nous proposons d'utiliser deux variables textuelles qui sont n_v_mots, et pourcentage v_mots, qui présentent respectivement le nombre de mots violents qui figurent dans la page et leur pourcentage.

La structure d'une page Web est fondée sur un système de balises (chaînes de caractères délimitées par les symboles < et >) qui décrit leur type (liens hypertexte, images, mots clés, etc.). Glover et al. (2002) ont prouvé que l'analyse de cette structure combinée à une analyse textuelle ne peut qu'améliorer la classification et la description de la page Web. L'analyse de différentes balises nous a permis également d'extraire et de calculer d'autres variables dites structurelles comme n_v_url qui représente le nombre de mots violents dans l'URL, et le n_v_meta qui décrit le nombre de mots violents dans les balises meta. Pour récapituler ce qui précède, le vecteur de caractéristiques que nous avons utilisé pour classifier les pages Web est représenté par le tableau 1.

L'extraction des différentes caractéristiques précédentes nécessite l'analyse du code HTML d'une page Web. Il nous a donc fallu nous doter de : (1) un client HTTP, qui prend en paramètre une URL et renvoie une page de code HTML ; (2) un analyseur syntaxique (parser HTML), qui lit le code de la page, calcule les valeurs associées aux différents critères et stocke ces valeurs dans un fichier, qui sera utilisé par la suite dans une phase d'apprentissage.

Nom variable	Description
n_v_mots_page	nombre de mots violents dans la page web.
%v_mots_page	fréquence des mots violents de la page.
n_v_mots_url	nombre de mots violents dans l'url de la page web.
%v_mots_url	fréquence des mots violents de l'url de la page.
n_v_mots_title	nombre de mots violents qui figurent dans la balise title.
%v_mots_title	fréquence des mots violents dans la balise title.
n_v_mots_body	nombre de mots violents qui figurent dans la balise body.
%v_mots_body	fréquence des mots violents dans la balise body.
n_v_mots_meta	nombre de mots violents qui figurent dans la balise meta.
%v_mots_meta	fréquence des mots violents dans la balise meta.
n_lien	nombre de liens d'une page web.
n_lien_v	nombre de liens contenant au moins un mot violent.
%liens_v	fréquence des liens contenant un mot violent.
n_img	nombre d'images dans la page web.
n_img_v	nombre d'image dont le nom contient au moins un mot violent.
n_v_img_src	nombre de mots violents dans l'atribut src des balises img.
n_v_img_alt	nombre de mots violents dans l'atribut alt des balises img.
%img_v	fréquence des images contenant un mot violent.

TAB. 1 – *Vecteur de caractéristiques d'une page Web.*

Il est à noter que le calcul de la majorité des variables a été effectué en se basant sur un vocabulaire de mots violents rassemblés dans un dictionnaire. Ce dernier a été construit manuellement et il comprend des mots clés français et anglais.

La phase suivante de notre démarche est celle de la sélection de variable. Dans cette phase, nous avons déterminé les variables qui ont une influence sur notre problème. La sélection des variables contribue à réduire la taille du problème en isolant les variables exogènes les plus pertinentes. L'élimination des variables inutiles et redondantes permet d'accélérer le processus d'apprentissage et d'augmenter la fiabilité du classifieur obtenu. Afin de sélectionner les variables les plus pertinentes nous avons utilisé une approche de type filtre et plus précisément l'algorithme Relief (Kira et l. Rendel, 1992) vue qu'il est capable de travailler avec des variables bruitées et corrélées et de traiter des données nominales et continues.

3.2 Apprentissage supervisé

Il s'agit de trouver une fonction de classement efficace pour prédire les valeurs d'une variable catégorielle, dite à prédire, en fonction des valeurs d'une série de variables continues et/ou catégorielles, dites prédictives. Dans la littérature, il existe plusieurs techniques d'apprentissage supervisé comme les réseaux de neurones (Herault et Jutten, 1994), les graphes d'induction (Zighed et Rakotomalala, 2000), les réseaux bayesiens (Naïm et al., 2004), les machines à vecteurs supports (Schölkopf et al., 1998). Ces classifieurs se différencient selon leur mode de construction et selon leurs caractéristiques. Dans notre approche, nous avons utilisé les graphes d'induction vue qu'ils produisent des règles de type si…alors, ce qui facilite le

travail de validation et de communication du modèle. De plus les graphes d'induction sont applicables sur tout type de données et ils sont robustes au bruit et aux valeurs manquantes, donc ils sont convenables pour notre approche.
Dans cette étape de notre démarche, notre objectif est de trouver le graphe d'induction le mieux adapté à notre problème. Nous avons étudié quatre algorithmes de data-mining à savoir ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), IMPROVED C4.5 (Rakotomalala et Lallich, 1998), SIPINA (avec λ=1 et λ=0.2) (Zighed, 1996).

3.3 Validation

Après la phase d'apprentissage, nous avons évalué la qualité et la stabilité des modèles obtenus à partir des quatre algorithmes de data mining par le biais de la méthode des taux d'erreur. En effet, il est délicat de formuler des indicateurs généraux pour valider les modèles, et dans la plupart des cas, les chercheurs travaillent sur le taux d'erreur parce qu'il est l'un des meilleurs indicateurs qui soit véritablement comparable d'un algorithme à un autre.

Les mesures d'évaluation que nous avons effectué sont, le taux d'erreur global, le taux d'erreur a priori et le taux d'erreur a posteriori. Rappelons que le taux d'erreur global est le complément du taux de classification tandis que le taux d'erreur a priori (respectivement le taux d'erreur a posteriori) est le complément du taux de rappel classique (respectivement le taux de précision). Ainsi, plus le taux d'erreur a priori obtenu est faible, meilleur est le taux de rappel. La même règle s'applique au rapport entre le taux d'erreur global et le taux global de classification ainsi qu'entre le taux d'erreur a posteriori et le taux de précision.

4 Expérimentations

Cette section présente les différentes expérimentations réalisées afin de trouver le modèle de prédiction le plus pertinent pour notre application. L'étape de recherche du modèle consiste à extraire la connaissance utile de l'ensemble de données que nous avons collecté dans les phases décrites auparavant. Avant de présenter les séries d'expérimentations et afin de clarifier leurs conditions, nous allons décrire brièvement les conditions d'expérimentation.

4.1 Conditions d'expérimentations et techniques de validation

Dans nos expérimentations, nous présentons deux séries de tests :
La première est le résultat de notre système de classification sur les 700 sites qui constituent notre base d'apprentissage. Après une première phase d'apprentissage, nous avons évalué la qualité et la stabilité des modèles obtenus à partir des quatre algorithmes de data mining par le biais de la méthode des taux d'erreur.
La deuxième série de test est le résultat de notre système de classification sur une base de test composée de 300 sites : 150 à caractère violent et 150 à caractère non violent. Cette série d'expérimentation a été réalisée afin d'éviter le phénomène de « surapprentissage » (overfitting). En effet, il est fréquent que certains classifieurs « apprennent » les données plutôt que le modèle.

4.2 Résultats

La figure 2 illustre les différents taux d'erreur correspondants à l'utilisation des quatre algorithmes d'apprentissage. Le meilleur algorithme étant SIPINA que ce soit avec λ=1 ou λ=0.2. Ces résultats peuvent s'expliquer par le fait que cet algorithme tente de réduire les inconvénients des méthodes arborescentes d'une part par l'introduction de l'opération de fusion et d'autre part par l'utilisation d'une mesure sensible aux effectifs. Nous pouvons signaler que, pour la majorité des algorithmes, le taux d'erreur a priori et a posteriori violent sont faibles par rapport à ceux non violent. Cela signifie que ces algorithmes fournissent une décision plus fiable en ce qui concerne la classification des sites violents.

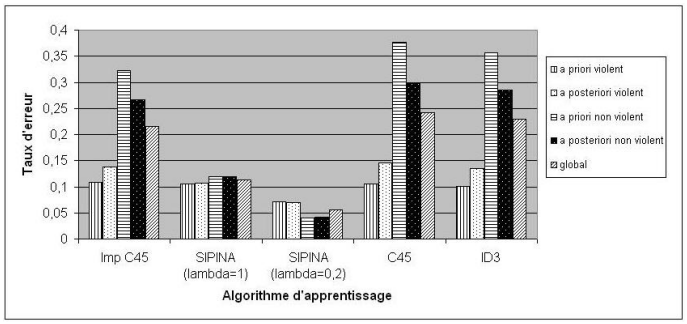

FIG. 2 – *Classification des sites Web à caractère violent*

Encouragés par les résultats précédents, nous avons alors testé les différents modèles de prédiction, obtenus lors de la phase d'apprentissage, sur notre base de test.
Les résultats des différentes expérimentations sont décrits par la figure 3.

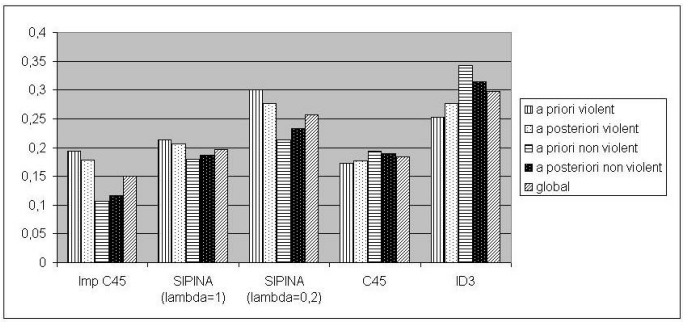

FIG. 3 – *Résultats des évaluations des algorithmes d'apprentissage sur la base de test*

Contrairement aux résultats trouvés dans la phase d'apprentissage, l'algorithme SIPINA avec λ=0.2 donne un taux d'erreur élevé par rapport aux autres algorithmes, ceci est du au phénomène de sur apprentissage. Concernant les deux algorithmes C4.5 et Sipina(λ=1), ils présentent les taux d'erreur a priori violent et non violent les moins élevés en tenant compte

des deux phases d'apprentissage et celle du test avec un taux d'erreur a priori violent de 0.17 pour C4. 5 et de 0.21 pour SIPINA. En revanche Sipina a montré sa performence au niveau de classification des sites non violents avec un taux d'erreur de 0.18 Alors que C4.5 enregistre un taux d'erreur plus élevé de 0.01.

Un filtre efficace détermine les sites à filtrer et les sites à ne pas filtrer. En d'autres termes, le logiciel identifie tous les sites à caractère violent, disponibles sur le web. Ceci constitue le rappel. Ce qui différencie les bons filtres des moins bons filtres est leur capacité à correctement distinguer les sites trouvés. Les sites contenant le mot « violent » ne doivent pas tous être filtrés. L'accès aux sites violents doit être bloqué, mais les sites qui luttent contre la violence doivent rester accessibles. Cette capacité à distinguer les différents sites constitue la précision. Le rappel et la précision sont inversement proportionnels. Ainsi, si un filtre est capable d'identifier tous les sites à contenus inappropriés, l'accès à certains sites inoffensifs sera également bloqué. Mais, si le logiciel est très spécialisé et capable de trouver uniquement des contenus préjudiciables sur un sujet spécifique, de nombreux contenus inadéquats pourront toujours être consultés.
La stratégie que nous avons choisie consiste à utiliser le modèle qui assure un meilleur compromis entre le rappel et la précision. Compte tenu des résultats obtenus dans les phases d'apprentissage et de test, nous avons opté pour l'utilisation du modèle de prédiction produit par l'algorithme SPINA (λ=1).

5 L'outil « WebAngels Filter »

5.1 Principe et Architecture

Les règles de prédictions obtenues par l'algorithme SIPINA (λ=1) ont été exploitées pour proposer un outil de détection et de filtrage du contenu violent sur Internet nommé « WebAngels Filter ».

Afin d'accélérer la navigation, nous avons choisi de mettre en oeuvre une liste noire qui sera crée et mise à jour d'une façon automatique.

D'un point de vue architecture système, « WebAngels Filter » s'exécute automatiquement lors de l'ouverture d'un navigateur et tourne en tâche de fond. Il doit agir à chaque demande d'une URL (c'est à dire chaque fois qu'une requête HTTP est lancée) et effectuer les actions suivantes :

- récupérer le code source HTML de la page demandée ;
- vérifier si l'URL appartient à une liste noire, et sinon analyser le code ;
- déclarer cette page autorisée ou interdite ;
- mettre à jour la liste noire ;
- mettre à jour l'historique de navigation ;
- afficher ou non la page.

La figure 4 résume le fonctionnement de notre logiciel.

5.2 Comparaison avec quelques produits commerciaux

Afin de mieux évaluer notre modèle de prédiction produit par l'algorithme SIPINA (λ=1), nous avons mené une étude comparative de notre solution avec quatre produits commerciaux, à

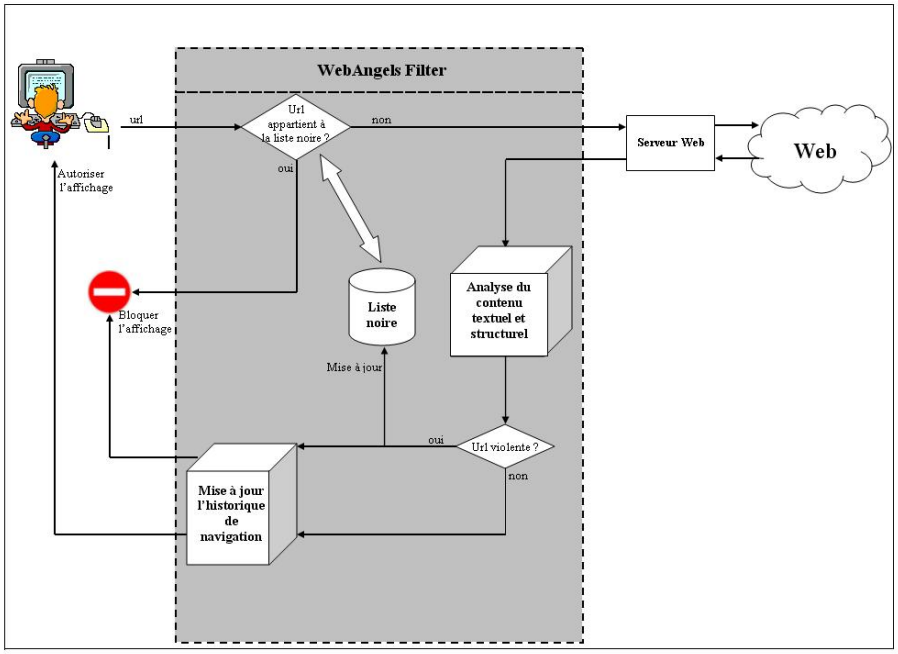

FIG. 4 – *Schéma du fonctionnement de l'outil « WebAngels Filter »*

savoir control kids[2] , content protect[3] , k9-webprotection [4] et Cyber patrol[5]. Ces logiciels sont censés être capable de filtrer des sites à caractère litigieux, et ils ont été paramétrés de façon à filtrer que le caractère violent. Notre objectif, ici, est d'évaluer si nos résultats théoriques avaient un sens en les comparant aux résultats réels du logiciel. Cette étude a été effectuée sur notre base de test, totalement indépendante de celle utilisée dans la phase d'apprentissage.
La figure 5 montre la performance de « WebAngels Filter » par rapport aux logiciels existants sur le marché avec un taux de classification égale à 81%. Ceci peut être expliqué par le fait que la majorité de ces logiciels traitent principalement le caractère pornographique des sites Web, alors que le caractère violent à été omis , et donc échappe à leurs filtres.

6 Conclusion et perspectives

Ce papier présente une solution de classification et de filtrage des sites Web à caractère violent par un apprentissage qui s'appuie sur plusieurs algorithmes de data mining avec non seulement une analyse du contenu textuel mais aussi du contenu structurel. L'étude compa-

[2]http ://www.controlkids.com/fr
[3]http ://www.contentwatch.com
[4]http ://www.k9webprotection.com
[5]http ://www.cyberpatrol.com

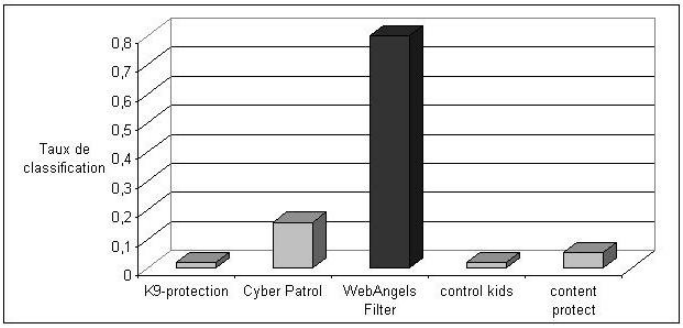

FIG. 5 – *Etude comparative de l'approche proposée avec quelques produits du marché.*

rative de notre solution avec les logiciels du marché les plus connus, sur notre base de test, montre la performance de notre système.

Ces résultats encourageants nous incitent à approfondir nos travaux de classification et filtrage de sites Web par une analyse conjointe de plusieurs modalités et à les appliquer à d'autres problèmes comme par exemple le filtrage de sites adultes, racistes, etc. D'autres pistes d'amélioration concernent l'élaboration du dictionnaire des mots clés qui a joué un rôle central dans les performances de « WebAngels Filter ». Or, l'élaboration de ce dictionnaire a été très laborieuse car manuellement améliorée étape après étape, et elle n'a vraisemblablement possible que grâce à la compréhensibilité des modèles obtenus par les techniques d'extraction de connaissance. Il serait donc intéressant d'automatiser par l'apprentissage à partir d'un corpus la construction d'un tel dictionnaire. Enfin, penser à une manière d'intégrer le traitement de l'aspect visuel dans les modèles de prédictions afin de remédier aux difficultés de classifier les sites Web violentes qui ne comprennent que des images.

Références

Arentz, W.-A. et B. Olstad (2004). Classifying offensive sites based on image content. *Computer Vision and Image Understanding 94*, 295–310.

Caulkins, J.-P., W. Ding, G. Duncan, R. Krishnan, et E. Nyberg (2006). A method for managing access to web pages: Filtering by statistical classification (fsc) applied to text. *Decision Support Systems*, (To apear).

Denoyer, L., J.-N. Vittaut, P. Gallinari, S. Brunessaux, et S. Brunessaux (2003). Structured multimedia document classification. *ACM Symposium on Document Engineering*, 153–160.

Du, R., R. Safavi-Naini, et W. Susilon (2003). Web filtering using text classification. *IEEE International Conference on Networks*, 325–330.

Glover, E.-J., K. Tsioutsiouliklis, S. Lawrence, D.-M. Pennock, et G.-W. Flake (2002). Using web structure for classifying and describing web pages. *International World Wide Web Conference*, 562–569.

Grilhers, B., S. Brunessaux, et P. Leray (2004). Combining classifiers for harmful document filtering. *RIAO'2004, Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, 1–10.

Hammami, M., Y. Chahir, et L. Chen (2006). A web filtering engine combining textual, structural, and visual content-based analysis. *IEEE Transactions on Knowledge and Data Engineering*, 272–284.

Herault, J. et C. Jutten (1994). *Les réseaux de neurones et le traitement du signal*. Hermès.

Ho, W. et P. Watters (2004). Statistical and structural approaches to filtering internet pornography. *IEEE Conference on Systems, Man and Cybernetics*, 4792–4798.

Hochheiser, H. (1997). Filtering faq. Technical report, Computer Professionals for Social Responsibility(http://www.cpsr.org/filters/faq.html).

Kira, K. et L. l. Rendel (1992). A practical approach to feature selection. *International Conference on Machine Learning*, 249–256.

Lee, P., S. Hui, et A. Fong (2002). Neural networks for web content filtering. *IEEE Intelligent Systems*, 48–57.

Naïm, P., P. Wuillemin, P. Leray, O. Pourret, et A. Becker (2004). *Réseaux bayésiens*. Eyrolles.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning 1*, 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Rakotomalala, R. et S. Lallich (1998). Handling noise with generalized entropy of type beta in induction graphs algorithm. *International Conference on Computer Science and Informatics*, 25–27.

Schölkopf, B., C. Burges, et A. Smola (1998). *Advances in Kernel Methods Support Vector Learning*. Eyrolles.

Vinot, R., N. Grabar, et M. Valette (2003). Applications d'algorithmes de classification automatique pour la détection des contenus racistes sur l'internet. *Traitement Automatique des Langues Naturelles*, 257–284.

Zighed, D. et R. Rakotomalala (2000). *Graphes d'Induction - Apprentissage et Data Mining*. Hermes.

Zighed, R. (1996). A method for non arborescent induction graphs. Technical report, Loboratory ERIC, University of Lyon 2.

Summary

In this paper, we propose a technique for automatically detect and filter violent content on the Web. While the most commercial filtering products on the marketplace are mainly based on textual content-based analysis such as indicative keywords detection or manually collected black list checking, the originality of our work resides on the addition of structural content-based analysis to the classical textual content-based analysis along with several major-data mining techniques for learning and classifying. Our results show that it can detect and filter violent content effectively.