Filtrage d'Arnaques dans un Corpus de Spams : Une application de Filtrar-S à la sécurité du citoyen

Nicolas CAMPION[1], Thomas FONTVIELLE[2], Lieutenant-colonel Eric FREYSSINET[3]

[1]ERAKLE, 42, rue d'Artois, 75008, Paris

[2]SIGNAL-SPAM, 8, rue du Faubourg Poissonnière, 75010, Paris

[3]PJGN/STRJD/DLCC, Division de Lutte Contre la Cybercriminalité, 1 bd Théophile Sueur, 93111, Rosny-Sous-Bois cedex

nicolas_campion@yahoo.fr, thomas.fontvielle@signalspam.net, eric.freyssinet@gendarmerie.interieur.gouv.fr

Résumé – Cet article présente les résultats d'un essai des logiciels conçus et développés dans le cadre du projet Filtrar-S, financé par l'ANR, dans le cadre du programme CSOSG 2008 (http://www.filtrar-s.fr). Il s'agissait d'utiliser Filtrar-S et son module de filtrage sémantique pour filtrer des spams d'arnaques (ou Scams) dans un corpus de spams. Cette application répond à un besoin de la Division de Lutte Contre la Cybercriminalité de la gendarmerie nationale et elle a été menée en collaboration avec l'association Signal-Spam (http://www.signal-spam.fr). Les performances sont bonnes et démontrent la pertinence de Filtrar-S pour résoudre des problèmes liés à la sécurité du citoyen..

Abstract – This paper presents the testing of the softwares designed during the Filtrar-S project that is supported by ANR and belongs to the CSOSG 2008 financial session(http://www.filtrar-s.fr). The semantic filtring module of Filtrar-S has been used to find in a spam corpora the spams that are a kinds of scams and are therefore called Scams. This application responds to a need for the Division de Lutte Contre la Cybercriminalité de la gendarmerie nationale and it was conducted in collaboration with the association Signal Spam (http://www.signal-spam.fr). Actual performance is good and demonstrate the relevance of Filtrar-S to solve problems related to security of the citizen.

1. Filtrar-S et les Scams

La lutte contre les spams est un enjeu pour la sécurité des citoyens, notamment en raison de la présence d'une catégorie de spams d'arnaque appelés Scams. Il s'agit de spams comportant généralement un récit singulier, destiné à émouvoir ou simplement convaincre le destinataire, en lui expliquant pourquoi une importante somme d'argent, évidemment imaginaire, se trouve actuellement bloquée à l'étranger. Le but est de faire croire au destinataire qu'il peut récupérer tout ou partie de la somme et ensuite de lui soutirer l'argent d'une commission, soit-disant destinée au financement des intermédiaires de la transaction imaginaire. Souvent les escrocs utilisent également ces Scams pour faire de l'hameçonnage (phishing) en demandant aux destinataires leurs coordonnées postales, voire bancaires. Le procédé est apparemment grossier mais efficace, puisqu'on estime que de nombreuses victimes y perdent plusieurs millions de dollars par an. Un exemple de Scam est présenté sur la figure 1.

Les Scams sont une des infractions étroitement surveillées par le service de cybercriminalité de la gendarmerie nationale. Ils constituent à la fois une menace d'escroquerie financière et une menace d'usurpation d'identité, renvoyant par là au problème de société qu'est aujourd'hui la protection de l'identité à l'ère numérique [1].

L'objectif de cet article est de présenter un essai de filtrage sémantique de Scams. C'est une des applications possibles de Filtrar-S à la sécurité du citoyen et elle met directement à l'épreuve l'approche statistique et non supervisée du filtrage sémantique par le contenu qu'opère le module de filtrage de Filtrar-S. [2]

Dear Sir,

I am DR. PETER C. chairman- audit committee, with the Nigerian National Petroleum Corporation (NNPC),

(...) we discovered that the sum of USD 30 million (THIRTY MILLION DOLLARS) was floating unclaimed in the corporation account without beneficiary. This money emanated from an over-invoicing in the contract awarded to a foreign firm for the supply, fixing, erecting and computerized optimization of oil pipeline to one of our oil refineries during the past Military administration.

(...) we cannot stand to claim this money ourselves, hence, I decided to contact you.

(...) ***we now require from you your BANKING PARTICULARS, FAX AND TELEPHONE NUMBER for smooth transfer of this money into your bank account.*** *We intend to share the money in the following manner: 30% for you, 65% for us and 5% for any miscellaneous expenses. This business is 100% risk-free. We look forward to a fruitful business relationship with you.*

Best Regards DR. PETER C. - Account Department (NNPC)

Fig. 1. Un exemple de Scam (spam d'arnaque) avec hameçonnage (en gras)

2. L'association Signal-Spam

L'essai de filtrage sémantique que nous présentons a été réalisé dans le cadre d'une collaboration de Filtrar-S avec l'association Signal-Spam (www.signal-spam.fr). Cette association regroupe des organisations françaises concernées par la lutte contre le spam, qu'il s'agisse des pouvoirs publics ou des professionnels de l'Internet. Elle offre aux internautes une plate-forme nationale de signalement qui leur permet d'envoyer tous les spams qu'ils reçoivent à l'association. Recevant près de 10.000 signalements par jour, la plate-forme Signal-Spam est un outil précieux pour les acteurs de la sécurité sur internet et pour les chercheurs qui veulent disposer de corpus représentatifs des spams circulant à tout moment sur les messageries électroniques. C'est le cas du projet Filtrar-S dont les partenaires utilisent des corpus de spams à des fins expérimentales.

3. Les méthodes de filtrage des Spams

Nous laissons ici de côté les méthodes indépendantes du contenu (liste noire d'adresse IP, repérage des signatures numériques) bien qu'elles aient une efficacité indéniable et complémentaire des méthodes qui nous intéressent, celles qui analysent le contenu du texte des spams. Les analyses de contenu couramment effectuées par les logiciels anti-spams sont basées sur un apprentissage supervisé des caractéristiques de chaque spam ou chaque catégorie de spams.

La première difficulté est alors de disposer d'un corpus d'apprentissage reflétant la diversité des spams en circulation. La seconde difficulté tient à la sélection des caractéristiques du contenu des spams qui vont faire l'objet d'un apprentissage automatique. Divers algorithmes d'apprentissage supervisé ont été utilisés: SVM (Support Vector Machine); probabilité bayésienne naïve; algorithme Nearest Neighbors. Mais tous ces algorithmes se heurtent aux mêmes inconvénients, inhérents à l'approche supervisée : le coût de la mise à jour constante d'une base de spams; le coût de la sélection contrôlée de caractéristiques pertinentes pour l'apprentissage; le temps de réaction aux constantes modifications que les spammeurs apportent aux spams pour déjouer les filtres existants [3], [4].

A la différence des méthodes précédentes, les méthodes statistiques sont potentiellement non supervisées. Il s'agit d'utiliser des algorithmes statistiques qui exploitent la co-occurrence des mots dans les textes pour construire des représentations sémantiques de leur contenu. Les plus connus sont LSA [5] et LDA [6]. Ils ont été jusqu'ici assez peu appliqués au filtrage des spams et uniquement dans le cadre d'une approche globale, partiellement supervisée. Le paradigme consiste à prédéfinir deux corpus e-mails, l'un composé de spams et l'autre d'e-mails normaux, d'après le jugement de lecteurs. On applique ensuite l'un ou l'autre de ces algorithmes statistiques aux deux corpus d'apprentissage pour extraire de chacun une représentation sémantique globale (moyennant les représentations de chaque texte). Puis, on calcule la similarité de tout nouvel e-mail avec ces deux représentations pour le classer comme spam ou non-spam. La plupart des publications décrivant l'application de méthodes statistiques au filtrage de spam concernent LSA, appelé LSI par certains auteurs, [7]; [8]; [9]. Les résultats sont dans l'ensemble probants avec des taux de précision et de rappel avoisinant respectivement 89 et 95 %. L'étude de Tee, Gong et Kil [10] a le mérite de comparer LSA et LDA, ainsi que deux variantes de chacune de ces méthodes, et elle montre qu'avec ce paradigme global supervisé, les performances de filtrage des spams sont très proches pour toutes ces méthodes.

4. La nouveauté de l'approche Filtrar-S

Le module de filtrage de Filtrar-S explore une voie nouvelle pour l'analyse sémantique automatique des textes d'un corpus, qu'il s'agisse ou non de spams. Le module applique de façon entièrement non supervisée l'algorithme génératif bayésien LDA (Latent Dirichlet Allocation) avec échantillonnage de Gibbs. [11]. LDA analyse les co-occurrences des mots dans les textes et produit des Topics qui sont des groupes de mots sémantiquement associés

entre eux et formant la face lexicale de nos connaissances sur les situations que décrivent les textes du corpus.

Ainsi d'un corpus d'articles du journal Le Monde Filtrar-S a extrait le Topic:

«police violence jeune cité délinquance bande...».

En outre, l'algorithme produit des liens pondérés entre les textes du corpus analysé et les Topics, conférant ainsi aux Topics le rôle de classifieurs sémantiques.

Une première conséquence importante de l'approche est qu'il n'est pas nécessaire de paramétrer le logiciel à partir d'exemplaires des spams qu'il doit filtrer. Le principal inconvénient des méthodes supervisées, qui en limite notablement le champ d'action, est donc supprimé. De plus, la classification des textes est mécaniquement adaptée au corpus traité puisqu'elle repose sur les Topics et que ceux-ci sont produits uniquement par l'analyse statistique des mots des textes.

Un autre résultat notable est que les classifications par Topics sur lesquels reposent le filtrage des textes ne sont pas tributaires de la présence de mots clés particuliers. Les textes sont sémantiquement représentés par des liens pondérés avec des Topics, c'est-à-dire non pas avec des mots particuliers, mais avec des séries pondérées de mots associés. Ainsi, seuls quelques mots des Topics sont généralement présents dans les textes associés et la composition lexicale des textes associés aux mêmes Topics peut être très variable.

Enfin, une conséquence majeure de l'approche est que, sous certaines conditions, liées notamment à la composition et à l'extension des corpus, les regroupements de mots opérés par les Topics sont la composante verbale de représentations sémantiques associatives proches de celles qui, d'après les sciences cognitives, composent la mémoire sémantique humaine. Une manifestation de cette validité psychologique s'observe lorsque les Topics discriminent correctement les mots en fonction de leurs différentes significations dans différents contextes sémantiques. Les Topics offrent ainsi l'opportunité de résoudre les problèmes que crée la polysémie de la plupart des mots des langues naturelles.

5. L'essai de filtrage d'arnaques

L'essai du module de filtrage de Filtrar-S a été réalisé en réponse à un besoin particulier de la Division de Lutte Contre la Cybercriminalité de la gendarmerie nationale (DLCC) : la lutte contre les arnaques sur internet, et en particulier les Scams.

En outre, c'est une expérience nouvelle, à notre connaissance inédite, puisqu'il s'agit, conformément à l'approche Filtrar-S, de filtrer une catégorie de spams (les Scams) parmi toute sorte de spams et, cela sans connaissances a priori sur les contenus filtrés, contrairement aux approches classiques.

Par ailleurs cela représente un intérêt pour l'évaluation des capacités de filtrage sémantique de Filtrar-S, en raison des difficultés formelles et sémantiques que posent le type de texte traité. La qualité du texte contenu dans les spams présente de multiples difficultés : suites de caractères sans signification (codage des images, signe html …), erreurs typographiques, fautes d'orthographes, néologismes, multilinguisme …. Un point important ici est que pour peu qu'elles soient la source de redondances, ces difficultés ne seront pas ignorées mais exploitées par les algorithmes de Filtrar-S.

Du point de vue sémantique, les Scams sont un cas intéressant. Ils sont en apparence d'une grande richesse sémantique et d'une grande diversité en raison des récits singuliers qu'ils développent. La tâche du module de filtrage de Filtrar-S est donc de dégager des structures sémantiques invariantes, liées au transfert de fond que l'on fait toujours miroiter aux destinataires du Scam. Ces transferts sont décrits avec des mots variables, liés à la complexité des scénarios inventés par les auteurs. Mais Filtrar-S permet justement un filtrage indépendant de la présence de mots clés particuliers dans un texte. C'est au niveau du contexte sémantique, créé par la régularité des significations transmises, que se situe Filtrar-S pour extraire des réseaux de mots associés, les Topics, et pour opérer un filtrage non subordonné à la présence de mots clés.

5.1 Le corpus

5.1.1 Le corpus de Signal Spam

L'association Signal Spam a donc fourni à Filtrar-S un corpus de spams. Rappelons qu'un spam est un courrier électronique envoyé en grand nombre, de façon anonyme ou sous une fausse identité, à des destinataires qui ne l'ont pas sollicité et que l'on maintient dans l'incapacité de s'opposer à cette diffusion. C'est pourquoi les courriers électroniques publicitaires doivent offrir aux destinataires la possibilité de refuser de futures diffusions pour ne pas tomber dans la catégorie des spams.

Les corpus livrables par Signal Spam présentent un double intérêt pour Filtrar-S : premièrement ils sont potentiellement importants en volume et représentatifs de l'existant, deuxièmement leur contenu a été contrôlé grâce au signalement, et ils sont uniquement composés de spams, sauf erreurs probablement exceptionnelles lors du signalement.

La principale difficulté a toutefois été de satisfaire aux exigences de la CNIL qui a réclamé une anonymisation des documents du corpus, afin que le nom des personnes ayant reçu les spams soit masqué. Puisque le corpus n'est utilisé que pour opérer un filtrage des spams par le contenu et évaluer les performances de Filtrar- S dans cette tâche, il a été décidé que Filtrar-S ne traiterait que les corps de texte des spams et non pas leurs en-têtes (partie du message contenant les adresses mail des expéditeurs et des destinataires, le chemin de transmission via les ordinateurs relais, et le titre du message). En outre, un accord de confidentialité a été signé entre Signal Spam et chacun des partenaires du projet Filtrar-S.

Une autre difficulté était de constituer un corpus contenant une proportion non négligeable du type de spams que nous cherchons à filtrer. Puisque les Scams sont une catégorie relativement abondante de spams, la solution

retenue a été de récupérer un corpus composé de l'ensemble des spams signalés pendant 6 mois durant par les adhérents de Signal Spam.

5.1.2 L'échantillon de spams analysé

Le corpus livré par l'association Signal Spam à Filtrar-S correspond à 6 mois de signalement par ses adhérents, de mai à octobre 2010. Une mise en forme a été nécessaire pour extraire chaque message, en supprimer l'en-tête et placer le corps du message dans un fichier texte, un fichier distinct étant créé pour chaque spam. Au total, 690 009 fichiers contenant des corps de spams ont été créés.

Pour cet essai un échantillon aléatoire de 5 000 spams a été prélevé dans le corpus total. Le principal avantage de travailler sur un échantillon restreint est de pouvoir disposer d'une estimation fiable de la quantité de spams contenus dans l'échantillon. En outre, les temps de traitement sont réduits à quelques heures. Enfin, il est intéressant de savoir si l'efficacité de l'algorithme est satisfaisante avec un nombre de spams de taille modérée : un dixième environ de la taille du corpus type que constitue l'ensemble des articles publiés dans le journal Le Monde. Est-ce suffisant pour faire émerger des catégories homogènes de spams ?

L'échantillon de 5 000 spams contient 137 062 types de chaîne de caractères (mots ou autres) séparées de chaque côté par un espace. Ces chaînes types, par leurs répétitions, fournissent à l'analyse 1 360 269. On peut donc calculer la taille moyenne d'un spam de l'échantillon et l'on trouve 272,05 mots par spams.

5.1.3 Les Scams recherchés

Une double lecture des spams par l'un des auteurs (Nicolas Campion) a permis d'identifier 112 Scams. Ce nombre est probablement exact, car on a trouvé une quantité à peu près égale de Scams dans chaque tranche de 1000 spams, les spams étant lus dans l'ordre initialement déterminé par l'échantillon aléatoire. Il y avait en moyenne 22,4 Scams par tranche, avec un écart-type de 4,4.

5.2. Le traitement

5.2.1 Mise en forme du texte

Aucune mise en forme n'a été faite. C'est un point important. Nous testons ici la capacité de l'algorithme à travailler à partir du texte brut, sans lemmatisation, sans élimination de mots vides du type articles pronoms, auxiliaires, sans élimination des mots ou séquences de caractères, inconnus et absents des dictionnaires, sans filtrage des occurrences lexicales fréquentes ou rares.

5.2.2 Paramétrage du logiciel de filtrage

Le logiciel de filtrage CalcLDA a été développé dans le cadre du projet Filtrar-S. Il est paramétrable par l'utilisateur et effectue les traitements nécessaires au filtrage sémantique non supervisé, par extraction de thèmes sémantiques représentatifs du contenu des documents.

Pour l'essai, le nombre de Topics produits a été fixé à 100. Les autres paramètres de CalcLDA ont été les suivants : $\alpha = 0,5$; $\beta = 0,1$; Itérations = 1000. Le temps de traitement nécessaire à la production des Topics a été de 8 heures.

Une autre fonctionnalité paramétrable de CalcLDA est l'optimisation qui réduit la taille des matrices résultats. En effet, dans ces matrices, tous les mots et tous les documents du corpus sont associés à tous les Topics par des valeurs de probabilité souvent infinitésimales. L'optimisation solutionne ce problème en ramenant les probabilités trop faibles à 0. Ainsi, on obtient une série de Topics auxquels est associé un nombre restreint de mots, et des documents auxquels est associé un nombre restreint de Topics.

5.3 Résultats

Deux matrices ont été produites par CalcLDA. Une matrice Mot X Topics et une matrice Topics X documents, les documents étant en l'occurrence le corps de texte des spams de l'échantillon analysé. Chacune de ces matrices relie par des valeurs de probabilité d'une part chaque mot de l'échantillon et chacun des 100 Topics, et d'autre part, chacun des documents de l'échantillon et chacun des 100 Topics. Puis, l'optimisation sélectionne dans ces matrices les liaisons mots-Topics et Topics-documents les plus fortes, celles pour lesquelles les valeurs de probabilité sont les plus élevées.

5.3.1 L'espace sémantique multilingue des Topics

Par la matrice Mot X Topics, chacun 100 Topics est associé à une série de mots qui appartiennent généralement à une même langue. On observe ainsi la présence de 42 Topics en anglais, 22 Topics en français, 7 Topics mixtes anglais français (en raison de la présence de spams bilingues dans l'échantillon), 1 Topic en espagnol, 1 Topic en allemand et 1 Topic en italien. Les autres Topics sont indécidables : les mots qui leur sont associés sont des caractères sans signification ou bien ces Topics sont vides car, après optimisation, aucun mot n'est associé avec une probabilité suffisante au Topic. Il faut noter que l'étendue de cette dernière catégorie est relative au seuil d'optimisation choisi.

5.3.2 La récupération des Scams

La plupart des Scams trouvés (91 soit 81,2 %) sont en anglais et associés avec une forte probabilité ($p > 0,49$) dans un même Topic que nous appellerons le Topic-arnaque, qui est aussi le Topic principal de ces documents,

c'est-à-dire le plus probablement associé à leur contenu par la matrice Topic X Documents.

TAB. 1 : premiers mots du Topic-arnaque, classés de haut en bas par probabilités décroissantes

to
the
you
... articles pronoms (30) ...
Please
US
bank
any
more
contact
Now
account
money
name

Pour la plupart, les Scams du Topic-arnaque sont du même type que l'exemple présenté sur la figure 1. On trouve aussi 7 exemplaires d'une variante consistant à annoncer au destinataire qu'il a gagné le gros lot d'une prétendue loterie. Par exemple voici un extrait de ce type de Scam:

"….your email address emerged as one of the online Winning emails in the 1st category and therefore attracted a cash award of EUR1,500,000.00 (One Million Five Hundred Thousand Euros)…"

On a trouvé aussi : 4 exemplaires de Scams usurpant l'identité d'un service connu de livraison et faisant croire au destinataire que s'il donne son adresse, et une somme d'argent, on lui adressera un colis de valeur ou une somme d'argent ; 6 exemplaires de Scams proposant au destinataire de recevoir temporairement l'argent de dons humanitaires égarés (en relation avec le terrible tsunami de 2006). Ces 6 Scams, ainsi que 8 autres sont écrits par des auteurs qui se disent asiatiques, tandis que deux Scams se déclarent en provenance du Koweït, 1 Scam en provenance d'Espagne, 2 de Hollande, 2 d'Angleterre, 4 des USA. Les autres , c'est-à-dire 75% des Scams se disent provenir d'Afrique : Burkina-Faso, Nigeria, Ghana, Côte d'ivoire, Sénégal, Afrique du Sud.

On note enfin la présence d'hameçonnage dans 46% des Scams du Topic-arnaque dans lesquels on demande au destinataire de fournir ses coordonnées : nom, e-mail, adresse postale, téléphone, fax. Dans 9 Scams, on demande une copie de la carte d'identité du destinataire. Dans 2 Scams enfin, on demande au destinataire son numéro de compte bancaire.

Les Scams filtrés sont très variés du point de vue du contenu et le nombre de Scams répétés est faible : 8 Scams sont présents en double exemplaire et 1 Scam est présent en triple exemplaire.

5.3.2 Comptage des mots dans les Scams du Topic-arnaque

Les types de chaînes de caractères trouvés dans les Scams du Topic-arnaque sont au nombre de 3 828, pour un nombre d'occurrences égal à 38 041. La taille moyenne d'un document du Topic-arnaque est donc de 333,69 mots, donc légèrement supérieure mais assez proche de 272, le nombre moyen d'occurrences trouvées dans un spam de l'échantillon.

Les fréquences d'occurrences des chaînes de caractères correspondant aux 50 premiers mots du Topic-arnaque ont une fréquence d'occurrence dans les 91 Scams du Topic-arnaque qui est environ le triple de la fréquence d'occurrence des mêmes mots dans l'échantillon de 5000 spams. Les fréquences d'occurrences sont calculées en rapportant le nombre d'occurrences des mots présents dans les documents de l'échantillon au nombre de documents de l'échantillon.

5.3.3 Les Scams hors du Topic-arnaque

La plupart des Scams dont le contenu n'est pas représenté par le Topic-arnaque sont écrits dans une autre langue que l'anglais (cf. TAB. 2).

TAB. 2 : Scams non classés dans le Topic- arnaques

Langue	Nombre	Topic	Traits
français	9	17	8 transferts de fonds 1 loterie
allemand	3	98	Loterie
espagnol	1	3	Loterie
français	1	78	Loterie
anglais	1	56	Argent pour victimes de Scams
anglais	6	aucun	Transferts de fonds

5.3.4 Récupération automatique des Scams

Entrer dans un moteur de recherche qui fouille le contenu lexical des Topics, des mots comme « bank », « account » ou « money », permet la récupération du Topic-arnaque. Dans l'outil Filtrar-S, c'est une extension du moteur de recherche d'Exalead qui assure cette fonctionnalité.

Ensuite, la matrice Topic X Documents permet d'accéder aux Scams associés au Topic-arnaque. Dans cette matrice,

les documents sont classés par le degré de probabilité qui caractérise l'association du Scam au Topic-arnaque : le premier document étant celui auquel le Topic-arnaque est associé avec la probabilité la plus forte. L'examen de cette liste de documents permet de calculer deux indicateurs classiques de performances des systèmes de récupération de document : la précision et le rappel.

Rappelons que la précision est le nombre de documents pertinents pour la requête rapporté au nombre de documents récupérés par le système en réponse à la requête, tandis que le rappel est le nombre de documents pertinents récupérés par rapport au nombre de documents récupérables (c'est-à-dire le nombre total de documents pertinents dans l'échantillon analysé).

La récupération automatique des Scams associés au Topic-arnaque, via un moteur de recherche, a été possible car ils sont nombreux et la plupart de ces Scams sont associés au Topic-arnaque par des valeurs de probabilité élevées. Ils sont donc classés dans les premiers rangs du Topic-arnaque. A l'inverse, la récupération automatique des Scams associés à d'autres Topics que le Topic-arnaque n'a pas été possible. Les faibles niveaux de probabilité qui les associent aux autres Topics ne les classent pas dans les premiers rangs de ces Topics.

La figure 2 présente la précision de la récupération de Scams dans le Topic-arnaque, la précision étant calculée jusqu'à la récupération des 91 Scams associés au Topic-arnaque. On constate que la précision est de 100 % pour les 10 premiers rangs, qu'elle reste supérieure à 90 % pour les 35 premiers rangs. La précision au rang 91 qui est le nombre de Scams associés au Topic-arnaque, la précision est encore supérieure à 70 %.

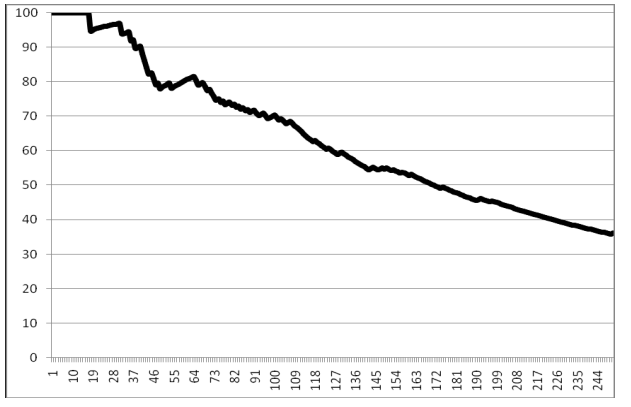

Fig. 2 : précision du rappel des Scams du Topic-arnaque

On peut considérer que le taux de rappel des Scams est le nombre de Scams associés au Topic-arnaque (c'est-à-dire 91) rapporté au nombre de Scams dans l'échantillon (c'est-à-dire 112). Le taux de rappel est donc de 81,2 %.

6. Conclusion

L'essai fait apparaître les bonnes performances de Filtrar-S pour récupérer dans un Topic produit par l'algorithme lui-même, les spams du type Scam, dont le contenu sémantique est relativement riche et présente une certaine homogénéité. Cette bonne performance se traduit par une précision de la récupération qui reste supérieure à 80 % jusqu'au rang 46 de classement des documents dans le Topic-arnaque. En outre le taux de rappel global est de 80% environ.

Au-delà de la performance de récupération des spams d'arnaque, c'est la performance de récupération de texte par le contenu qui est ici démontrée. Toutefois, les résultats indiquent aussi qu'un nombre critique de documents relevant d'un même thème ou contexte sémantique est nécessaire pour que ce Topic soit représenté et qu'il est impératif que ces documents soient écrits dans une même langue.

Il est important de remarquer que les difficultés inhérentes à la forme des textes qui contiennent des chaînes de caractères imprévisibles ne sont pas un obstacle au fonctionnement de l'algorithme.

On observe également que l'algorithme regroupe les textes écrits dans une même langue en les associant à des Topics spécifiques pour chaque langue. Dans le cas présent, les textes écrits dans une autre langue que l'anglais n'étaient pas présents en nombre suffisant, aussi le traitement n'a produit que peu de Topics pour chaque langue et ces Topics sont relativement hétéroclites du point de vue du contenu.

L'essai démontre enfin que l'algorithme donne de bons résultats avec un échantillon de documents de taille restreinte, 5 000 documents dans le cas présent. Toutefois, la taille de l'échantillon doit être déterminée par une contrainte statistique : les thèmes ou contextes sémantiques que l'on veut filtrer doivent être représentées par un nombre suffisant d'occurrences via les documents de l'échantillon, et chaque occurrence doit se traduire par un développement de plusieurs mots différents.

Références

[1] G. Desgens-Pasanau, et E. Freyssinet. *L'Identité à l'ère numérique*, Dalloz, 2009.

[2] N. Campion, J. Closson, O. Ferret, R. Besançon, W. Wang, J. Shin, J.-M. Floret, B. Grau, X. Tannier, A.-D. Mezaour, et J. -M. Lazard. *FILTRAR-S : Nouveaux développements.* Actes du cinquième Workshop Interdisciplinaire sur la Sécurité Globale - WISG'11, Université de Technologie, Troyes, 2011.

[3] K. Menghour, et L. Souici-Melsati. *Sélection de Caractéristiques pour le Filtrage de Spams*. CORIA : Conférence en Recherche d'Information et Applications, Sousse, 2010.

[4] E. P. Sanz, J. M. G. Hidalgo, et J. C. C. Perez. *Email Spam Filtering*, Advances in Computers, 74, 45-109, 2008.

[5] T. K. Landauer, et S. T. Dumais. *A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge.* Psychological Review, 104, 211–240, 1997.

[6] T. L. Griffiths, M. Steyvers, et J. B. Tenenbaum, . *Topics in semantic representation*. Psychological Review, 114, 211-244, 2007.

[7] J.R. Bellegarda, D. Naik, et K.E.A. Silverman. *Automatic junk e-mail filtering based on latent content.* Automatic Speech Recognition and Understanding. ASRU '03, IEEE Workshop ; 2003.

[8] W.N. Gansterer, A. G. K. Janecek, et R. Neumayer. *Spam Filtering Based on Latent Semantic Indexing*. In Michael W. Berry and Malu Castellanos, editors, Survey of Text Mining II: Clustering, Classification, and Retrieval, 165-183. Springer, 2003.

[9] J. Sun, Q. Zhang, Z. Yuan. *Application of Refined LSA and MD5 Algorithms in Spam Filtering*. Journal of Computers, 4, 245-250, 2009.

[10] S. Lee, J. Song, et Y. Kim. *An Empirical Comparison of Four Text Mining Methods*. Journal of Computer Information Systems. 1-10, 2010.

[11] N. Campion, J. Closson, O. Ferret , J. Shin, B. Grau, J. -M., Lazard, D. Lahbib, R. Besançon, J.-M. Floret, A.-D. Mezaour et X. Tannier. *FILTRAR-S : Un outil de filtrage sémantique et de fouille de textes pour la veille.* Actes du colloque VSST : Veille stratégique scientifique et technique, Université Paul Sabatier, Toulouse, 2010.