Les techniques de classification de courriels

Guillaume CALAS
guillaume.calas@gmail.com

Spécialisation *Sciences Cognitives et Informatique Avancée*

ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

14-16 rue Voltaire,
94270 Le Kremlin-Bicêtre,
France

Mots clés: classification, e-mail, courriel, pourriel, SPAM.

Résumé

L'informatisation des communications a permis d'accroître la vélocité des échanges et les a considérablement enrichis en contenu. Les *e-mails*, francisés en *courriels* (contraction de *courriers électroniques*), sont de plus en plus utiisés par les particuliers et encore plus par les entreprises (70 milliards de courriels par jour). Mais comme pour le courrier traditionnel, les utilisateurs ont du très rapidement faire face à des courriers non désirés, ou *pourriel* (*SPAM* en anglais), et pour la plupart tout à fait indésirables. Pour contrer ce flot de détritus (plus de 50% de tous les courriels) et ne pas perdre les courriels qui nous sont réellement destinés, une seule solution viable : automatiser la détection et la destruction de ce type de pollution numérique avec le risque qu'un document soit mal classé.

Le présent document a pour objet de présenter l'avancement des techniques de détection de spams.

Table des matières

1 Introduction

Les courriels non désirés, ou *pourriels* (contraction de *poubelle* et *courriel*), représentent une très importante part du traffic mondial des courriels. Selon diverses analyses, sur les dizaines de milliard de courriels qui transitent sur le réseau quotidiennement, près de 50%[1] sont des pourriels et cette proportion ne cesse d'augmenter. De plus, beaucoup de ces pourriels sont les vecteurs de propagation de certains virus et autres vers numériques qui mettent la sécurité des données stockées sur nos ordinateurs à rude épreuve.

Bien que difficilement chiffrable, le coût de cette pollution numérique représenterait plus de 200 milliards de dollars[2] par an dans le Monde pour les utilisateurs (perte de productivité, coût de connexion, logiciels de détection, etc.).

On comprend alors l'importance du développement d'applications permettant de lutter efficacement contre cette forme de pollution.

2 Distinction entre courriel désirable et non désirable

> *Une forte croyance ne prouve que sa force et nullement la vérité de ce qu'elle croit.*
>
> – Friederich NIETZSCHE

Par définition, un courriel est indésirable lorsque d'une part il n'a pas été sollicité, et/ou que sont contenu n'est ni *pertinent*, *recherché* ou jugé *digne d'intérêt* par l'utilisateur et qui constitue donc une pollution des messages sains et qui fini donc directement dans la corbeille. Mais comment juger de la *pertinence* du contenu d'un courriel ? Et pire encore, comment déterminer qu'un message est *digne d'intérêt* pour l'utilisateur ? On touche là au cœur du problème de classification des courriels. En faisant appel à des considérations inhérentes à l'utilisateur et au contexte d'utilisation de son service de mail (travail, personnel, etc.), on perd immédiatement la faculté de trier les courriels sur des critères stricts reposant sur des caractéristiques simples et exemptes de toute ambiguïté.

À défaut de comprendre le contenu d'un courriel, et surtout, d'être capable d'émettre un jugement en lieu et place de l'utilisateur, on peut, dans la plupart des cas, se contenter d'analyser la structure, l'origine, la destination et les effets de bord d'un courriel afin de détecter d'éventuelles déviations par rapport à une norme déduite du comportement et du contexte de l'utilisateur.

Pour détecter les déviations et les caractéristiques suspectes, les techniques actuelles vont s'appuyer pour l'essentiel sur l'analyse du comportement de l'utilisateur et donc sur des données statistiques pour réaliser une classification probabilistique. On ne pourra donc jamais avoir une absolue certitude sur la classification d'un courriel et on se gardera donc toujours de supprimer automatiquement du courrier juger indésirable.

[1] Source : Radicati Group, 2003. Estimations pour 2007.

[2] Source : Radicati Group, 2003. Estimations pour 2007.

3 Solutions contre le SPAM

On distingue deux catégories de logiciels de classification de courriels. L'une se trouve directement chez l'utilisateur et s'appuie généralement sur son comportement pour classer les courriels, et l'autre se trouve directement au point d'entrée des courriels, chez les fournisseurs de service de courriels, et qui vont généralement (mais pas seulement) s'appuyer sur la signature des courriels et les associer au volume de courriels similaires transitant par le service pour détecter les envois massif et anormaux.

3.1 Techniques utilisées

3.1.1 Avant-propos

Afin de juger des performances d'un classifieur, on utilise les notions suivantes :

Vrai Positif : ratio d'éléments de classe A ayant été étiquettés A par la classifieur.

Faux Positif :ratio d'éléments de classe A ayant été étiquettés B par le classifieur.

Vrai Négatif : ratio d'éléments de classe B ayant été étiquettés B par la classifieur.

Faux Négatif : ratio d'éléments de classe B ayant été étiquettés A par la classifieur.

En dehors du ratio global de bonne classification (*Vrai Positif* + *Vrai Négatif*), c'est le ratio de *Faux Positif* qui va nous intéresser en plus haut lieu car perdre un courriel important à cause d'une erreur de classification peut avoir des conséquences désastreuses. L'utilisateur, en fonction de l'importance des risques de mauvaise classification, doit pouvoir être en mesure de maîtriser ces risques et pouvoir adapter la politique de classification.

Voici une liste des techniques communément utilisées par les solutions anti-spam :

- Confrontation de signature par rapport à une base de données de pourriel
- Réputation (réseaux sociaux)
- Classifieurs probabilistiques: *Naïve Bayes*, *Quadratic Discriminant*
- *Support Vector Machine*
- Réseau de neurones
- *Data-mining*

- Listes blanches/noires
- *Clustering*
- Algorithmes génétiques

3.1.2 Classifieurs probabilistes

Pour effectuer une classification probabiliste il faut au préalable identifier les caractéristiques des courriels qui permettent de les différencier. En fonction des caractéristiques choisies, et pour chaque courriel, on va créer un vecteur de caractéristiques qui va permettre d'établir une classification.

Voici une liste non exhaustive des caractéristiques des courriels potentiellement discriminantes et sur lesquelles vont s'entrainer les classifieurs:

- longueur du message et du sujet
- nombre et type de pièces jointes
- présence d'HTML, de scripts et de liens hypertextes
- présence d'image(s)
- expéditeur: est-il connu par l'utilisateur ?
- le domaine du service d'envoie est-il sur liste noire ? est-il autorisé à effectuer ce type d'envoi ?
- destinataire(s): unique, liste de diffusion, copie caché, copie cachée.

À partir des fréquences obtenues sur chaque caractéristique, le classifieur retenu va produire une probabilité d'appartenance à chacune des classes possible et la possibilité la plus probable sera retrenu.

Avantage(s) : bonnes performances.
Inconvéniant(s) : performances dépendant de la qualité de l'entraînement ; nécessite un entraînement continu pour faire fasse aux nouvelles formes de spam.
Référence(s) associée(s) : [10], [13]

3.1.3 Data-mining

L'autre grande technique de classification est celle des algorithmes de *data-mining* supervisé à base d'arbres de décision tels que C4.5, SLIQ ou CART. Les courriels préalablement classés par l'utilisateur sont découpés selon leurs caractéristiques pertinentes et insérés dans la base de données. À partir de ces enregistrements, l'algorithme d'inférence va générer un arbre de décision qui pourra au besoin être converti en ensemble de règles et directement intégré aux clients de messagerie les supportant.

La qualité de la classification dépendra de la qualité de la préparation des données et des enregistrements retenus pour la base d'entraînement.

Avantage(s) : efficace pour peu que la base d'entraînement soit bonne.
Inconvéniant(s) : difficile à mettre en place ; nécessite la construction d'un nouvel arbre de décision pour faire face aux nouveaux types de pourriels.
Référence(s) associée(s) : [16]

3.1.4 Système immunitaire artificiel

Afin de lutter efficacement contre le spam, des chercheurs ont fait le parallèle entre les pourriels et les agents pathogènes qui sont combattus par le système immunitaire humain. À partir d'une bibliothèque de gènes, le système va générer aléatoirement des anti-corps et créer le lymphocyte correspondant afin d'être à même de détecter tout corps étranger entré dans le système. Les lymphocytes sont ensuite entraînés sur une base de courriels et de pourriels, ceux qui sont inefficace sont supprimés, et on attribue une date d'expiration à chaque lymphocyte en fonction de ses performances. À chaque message filtré par un lymphocyte, sa date d'expiration est augmentée. Arrivé à expiration les lymphocytes meurent et sont remplacés par d'autres qui redémarrent un cycle.

Avantage(s) : efficace ; léger (seuls 200 anti-corps fonctionnent en même temps) ; utilisation possible de vaccin d'urgence.
Inconvéniant(s) : -
Référence(s) associée(s) : [12]

3.1.5 Réseau de neurones

Les réseaux de neurones permettent, après apprentissage, de reproduire une forme de raisonnement humain. Les caractéristiques des courriels ainsi que leur contenu permettent d'ajuster les coefficients synaptiques du réseau de neurones durant la phase d'apprentissage. L'apprentissage se fait à partir d'une collection de courriels préalablement triés par l'utilisateur et peut éventuellement être incrémental afin d'être le mieux adapté possible aux nouvelles formes de spam qui peuvent d'apparaître. Une fois l'apprentissage effectué, le réseau de neurone fonctionne comme un système anti-spam classique très efficace selon les cas de figure.

Comme tout les classifieurs, le risque de mauvaise classification (*Faux Positif*) est réel mais peut être contrôlé en jouant sur le seuil de sensibilité du réseau de neurone (au détriment du *Faux Négatif*).

Avantage(s) : permet de régler le taux de mauvaise classification (*Faux Positif*) en ajustant le seuil de sensibilité ; rapide.
Inconvéniant(s) : nécessite un entraînement long ; doit être régulièrement entraîné pour faire face aux nouvelles formes de spam.
Référence(s) associée(s) : [11]

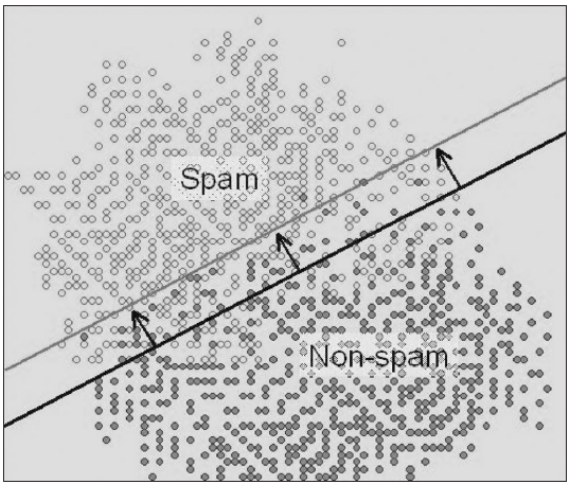

FIG. 1 – Réglage du seuil de sensibilité du réseau de neurone.

3.1.6 Signature des messages

Pour lutter contre les envois massifs de courriels, le système anti-pourriel doit se positionner au niveau du service d'envoi de courriels afin d'avoir une vision globale des courriels qui transitent sur le réseau et ainsi pouvoir détecter les envois massif caractéristiques des spammeurs.

Cependant, pour rendre les envois massif plus difficile à détecter, les spammeurs insèrent ou suppriment des séquences aléatoires dans les courriels afin que chaque message de la campagne soit unique. Ne pouvant pas se baser uniquement sur un *checksum* pour identifier les courriels identiques, les techniques de détection doivent se baser sur une autre forme de signature moins sensible à l'insertion/suppression de termes. C'est le cas de l'algorithme *I-Match*.

L'algorithme *I-Match* s'appuie sur l'ensemble des termes uniques du courriel et sur un lexicon préalablement établit pour produire la signature du message. Cette signature est alors associé à un unique *cluster* ce qui permet d'en déduire la classe du message.

Avantage(s) : détection des envois massif ; peu sensible à des modifications aléatoires du corps des courriels.
Inconvéniant(s) : doit être mis en place par le fournisseur de service de courriel ; doit être utilisé avec des techniques complémentaires ; risque réel de mauvaise classification (*Faux Positif*).
Référence(s) associée(s) : [8]

3.1.7 Réputation de l'émetteur (réseaux sociaux)

La technique la plus radicale pour trier ses courriels est encore celle des *listes blanches* (*white-list*) qui consiste à établir manuellement ou semi-automatiquement une liste de contact en qui on a confiance et dont on sait que les courriels sont « valides ». En pratique, les courriels envoyés par les personnes de confiance sont classés dans la boîte de réception et les autres sont envoyés dans une sous-répertoire pour les courriels indésirables.

Mais cette technique souffre de deux problèmes majeurs. D'une part, l'utilisateur doit lui même maintenir la *liste blanche* ce qui peut, dans certains cas, représenter une quantité de travail non négligeable. D'autre part, et c'est le plus gros défaut de cette technique, c'est quelle se base uniquement sur des contacts connus alors que bien des courriels peuvent venir de personnes ou d'organismes (connus ou non) non encore répertoriés dans la *liste blanche* et donc se retrouver mélangés dans le sous-répertoire réservé aux courriels indésirables. Retrouver ce genre de courriel malclassé dans un flot de pourriels peut vite se révéler assez périeux.

Pour pallier à ce problème, l'une des techniques consiste à utiliser un réseau de réputation[4] qui va noter (entre 0 et 10) chaque émetteur en fonction de son réseau de connaissances ce qui va permettre de lui attribuer (pour chaque utilisateur) un indice de confiance et donc de mieux classer les courriels. D'autre part, la confiance attribuée aux courriels envoyés par des émetteurs inconnus pourra être inférée pour peu que les émetteurs soient connus par une ou plusieurs personnes du réseau social de l'utilisateur.

Bien qu'améliorant sensiblement la qualité de la classification, cette technique ne résoud pas totalement le probléme des émetteurs inconnus car chaque réseau social ne comporte qu'un nombre fini d'individus. De plus, il faut toujours maintenir une *liste blanche* et noter chacun de ses contacts pour que le système puisse fonctionner de façon optimale. Il s'agit donc d'une technique intéressante mais insuffisante en elle même à utiliser en complément d'autres techniques de classification.

Avantage(s) : permet de lutter indirectement contre le taux de mauvaise classification (*Faux Positif*).
Inconvéniant(s) : difficile à maintenir ; nécessite un important réseau social ; ne résoud le problème des émetteurs inconnus.
Référence(s) associée(s) : [4]

3.1.8 Authentification de l'émetteur (test de Turing)

L'une des techniques les plus efficace est l'authentification de l'émetteur qui se base sur une fait simple : les pourriels sont envoyés automatiquement par des ordinateurs. En partant de ce fait inhérant à l'envoi massif de courriels, il suffit d'identifier automatiquement l'émetteur comme étant effectivement une personne physique en posant une question à l'émetteur à laquelle seul un humain peut répondre.

Par exemple, le système peut envoyer un *captcha* (une image contenant des caractères suffisament déformés et bruités pour compliquer sérieusement la tâche aux OCR) à l'émetteur du courriel (uniquement la première fois) et lui demander d'apporter la réponse à la question (recopier le

texte écrit dans l'image ou faire une opération mathématique).

Bien que radicale et efficace, cette solution souffre de plusieurs problèmes rédibitoires pour certaines applications. D'abord, l'utilisateur va devoir mettre en place une liste blanche pour les organismes qui envoient des messages de façon automatique (site administratif, commerce en ligne, etc.) ce qui peut vite devenir laborieux. Ensuite, c'est une méthode contrainiante pour l'émetteur du message. Dans un contexte de particulier à particulier où la volumétrie n'est pas très importante ce n'est pas particulèrement génant mais dès que lors que le volume quotidien de courriel dépasse un certain seuil – ce qui devient rapidement le cas dans le milieu professionnel – il devient très vite honnéreux pour les utilisateurs de répondre aux questions d'authentification. Enfin, cette technique nécessite l'utilisation d'une plateforme externe de contrôle, de stockage des utilisateurs authentifiés et de routage des courriels.

Avantage(s) : efficace ; rapide à mettre en place.
Inconvéniant(s) : difficile à maintenir (listes blanches) ; contrainiant pour l'émetteur.

3.1.9 Autres techniques

RBL (*Realtime Blackhole List*) : il s'agit d'immenses bases de données communes contenant la liste de tous les serveurs connu pour être utilisés pour du pollupostage. Si le serveur d'envoi est listé dans un RBL, alors c'est qu'il s'agit d'un pourriel.

SPF (*Sender Policy Framework*) : on vérifie que, dans la zone DNS du domaine, le serveur est autorisé à effectuer des envois.

Liste grise : cette technique s'appuie sur les normes décrites dans la RFC 2821[3] qui stipulent qu'un serveur de réception de courriel, en cas d'indisponibilité, doit retourner le code d'erreur 421 au serveur d'émission qui devra attendre un certain temps avant de ré-émettre le courriel. Les spammeurs, pour gagner du temps, renvoient les courriels beaucoup plus tôt que le temps minimum défini. Il suffit alors de ne laisser passer que les courriels qui sont renvoyés après le temps minimum d'attente. Cette technique, très simple à mettre en place par l'administrateur du serveur de réception est très efficace mais rajoute un temps de *lag* pour le destinataire.

3.2 Solutions pour les serveurs de courriels

SpamAssassin : système *Open Sources*, gratuit, sous licence GPL, réputé et efficace mais difficile à configurer et plus lent que d'autres systèmes commerciaux comme *M-Switch Anti-Spam*.

M-Switch Anti-Spam : système payant pour serveur de courriel. L'un des plus efficace selon plusieurs études (dont celle d'Isode), surtout en ce qui concerne les *Faux Positif*.

SpamGuru : système développé par IBM[15] sur la base de l'algorithme TEIRESIAS[14] (détection de séquences de gènes) auquel sont rajoutés plusieurs autres filtres et heuristiques. Le meilleur anti-spam du marché (98% de détection de spam et seulement 0,1% de *Faux Positif*).

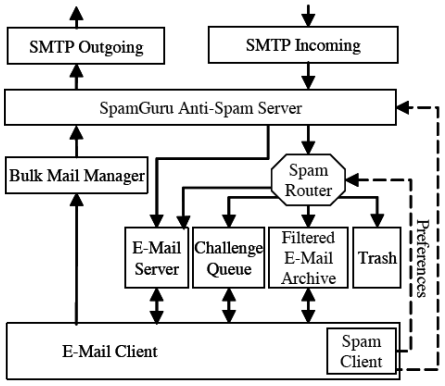

FIG. 2 – Architecture de SpamGuru.

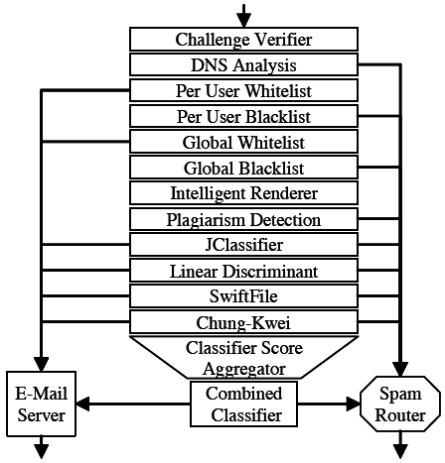

FIG. 2 – SpamGuru pipeline.

3.3 Solutions pour l'utilisateur final

Parmi les solutions présentes sur le marché, on trouve :

- **BogoFilter:** licence GPL, gratuit, multi-plateforme, filtre bayesien.

[3] http ://www.ietf.org/rfc/rfc2821.txt

- **Vade Retro:** solution française très performante et gratuite (sans mise à jour).
- **SpamFighter:** fonctionne uniquement avec Outlook®, existe en version gratuite ou payante.
- **SpamPal:** fonctionne uniquement sous Windows®, gratuit, RBL.
- **Spamihilator:** fonctionne uniquement sous Windows®, gratuit, filtre bayesien.

4 Perspectives d'évolution

Les techniques de détection de spam à base de gènes, d'anti-corps ou d'algorithmes génétiques sont actuellement très efficace et très en vogue mais les nouvelles avancées en matière de détection de chemins temporels[6] risquent de redonner de l'attrait aux techniques probabilistes ainsi qu'aux algorithmes de *data mining*.

D'autre part, les avancées dans le domaine du *text mining* pourrait bien un jour ouvrir une nouvelle voie dans la grande famille des techniques de classification des courriels.

5 Conclusion

Quelle que soit la technique nous avons vu qu'il n'est pas possible d'obtenir une classification automatique correcte à 100%. Cependant, quelle que soit la technique retenue, on voit que les solutions atteignent un taux de bonne classification et surtout le taux de *Faux Positif* tout à fait correct voir même très correct pour certaines.

Mais quoi qu'il en soit, il ne faut pas perdre de vu que le choix de la technique à utiliser ainsi que les taux de performance attendus sont fortement liés au contexte d'utilisation et aux impératifs qui s'y rattache. Pour une entreprise, il n'est pas acceptable de perdre (par mauvaise classification) un courriel valide, alors que pour un particulier ce peut être beaucoup moins grave. En fonction de ses attentes, on choisira donc de privilégier un taux de *Faux Positif* faible ou un taux global de bonne classification élevé. De même, le choix de la solution doit prendre en compte la volumétrie de courriels à traiter et les sacrifices que l'on est prêt à concéder.

Enfin, il ne faut pas oublier qu'il est souvent possible – et préférable – de combiner les différentes techniques afin d'obtenir les performances escomptées.

Références

[1] David A. Bader and Ashfaq A. Khokhar, editors. Proceedings of the ISCA 17th International Conference on Parallel and Distributed Computing Systems, September 15-17, 2004, The Canterbury Hotel, San Francisco, California, USA. ISCA, 2004.

[2] Richard Clayton. Stopping spam by extrusion detection. In CEAS, 2004.

[3] Ernesto Damiani, Sabrina De Capitani di Vimercati, Stefano Paraboschi, and Pierangela Samarati. An open digest-based technique for spam detection. In Bader and Khokhar [1], pages 559–564.

[4] Jennifer Golbeck and James A. Hendler. Reputation network analysis for email filtering. In CEAS, 2004.

[5] Christian Jacob, Marcin L. Pilat, Peter J. Bentley, and Jonathan Timmis, editors. Artificial Immune Systems : 4th International Conference, ICARIS 2005, Banff, Alberta, Canada, August 14-17, 2005, Proceedings, volume 3627 of Lecture Notes in Computer Science. Springer, 2005.

[6] Svetlana Kiritchenko, Stan Matwin, and Suhayya Abu-Hakima. Email classification with temporal features. In Klopotek et al. [7], pages 523–533.

[7] Mieczyslaw A. Klopotek, Slawomir T. Wierzchon, and Krzysztof Trojanowski, editors. Intelligent Information Processing and Web Mining, Proceedings of the International IIS : IIPWM'04 Conference held in Zakopane, Poland, May 17-20, 2004, Advances in Soft Computing. Springer, 2004.

[8] Aleksander Kolcz, Abdur Chowdhury, and Joshua Alspector. The impact of feature selection on signature-driven spam detection. In CEAS, 2004.

[9] Vijay Krishnan and Rashmi Raj. Web spam detection with anti-trust rank. In AIRWeb, pages 37–40, 2006.

[10] Steve Martin, Blaine Nelson, Anil Sewani, Karl Chen, and Anthony D. Joseph. Analyzing behavioral features for email classification. In CEAS, 2005.

[11] Chris Miller. Neural network-based antispam heuristics. In Symantec, white paper, 2003.

[12] Terri Oda and Tony White. Immunity from spam : An analysis of an artificial immune system for junk email detection. In Jacob et al. [5], pages 276–289.

[13] Jefferson Provost. Naive-bayes vs. rule-learning in classification of email. Technical Report AI-TR-99-284, The University of Texas at Austin, Department of Computer Sciences, 1999.

[14] Isidore Rigoutsos and Aris Floratos. Combinatorial pattern discovery in biological sequences : The teiresias algorithm [published erratum appears in bioinformatics 1998 ;14(2) : 229]. Bioinformatics, 14(1) :55–67, 1998.

[15] Richard Segal, Jason Crawford, Jeffrey O. Kephart, and Barry Leiba. Spamguru : An enterprise anti-spam filtering system. In CEAS, 2004.

[16] Ellen Spertus. Smokey : Automatic recognition of hostile messages. In AAAI/IAAI, pages 1058–1065, 1997.