SPAM ASSASSIN

Mathieu KERN - Aymeric HINDERCHIETTE

Décembre 2014

Table des matières

I Présentation de SpamAssassin **3**

1 Problématique **3**
1.1 Le SPAM 3

2 Le projet **4**
2.1 Informations 4
2.2 Développement 4
2.3 Qu'est ce que SpamAssassin 4

3 Ses fonctionnalités **7**
3.1 Comment il filtre 7

4 Les test de SpamAssassin **7**
4.1 Le score 8

5 Articulation du programme **9**
5.1 Configuration 9
5.1.1 Options de configuration 9

Première partie
Présentation de SpamAssassin

1 Problématique

Le mail (ou courriel) est aujourd'hui le moyen privilégié de communication à travers le monde. Massivement utilisé, d'une certaine fiabilité et éprouvé par des décennies d'utilisation il reste le moyen le plus répandu pour les communications entres les personnes. Malheureusement, mail est également aujourd'hui synonyme de spam, ces messages indésirables qui s'entassent dans nos boites mails. C'est ici qu'entre en jeu SpamAssassin.

1.1 Le SPAM

Avant de poursuivre sur SpamAssassin, rappelons concrètement ce qu'est le SPAM et ce qu'il implique.

Comment reconnaître un SPAM :

- De par sa nature, un SPAM n'est pas désiré par l'utilisateur qui le reçoit.
- La réception d'un SPAM résulte d'un envoi massif : une machine (souvent un bot) envoie le même message à plusieurs destinataires sans aucun discernement. Cela s'oppose aux messages ciblés par exemple de commerçant, qui n'envoie que à leurs prospects (à but publicitaire).
- Son contenu n'est pas spécifiquement destiné à l'utilisateur (chaque personne reçoit le même contenu).
- Une importante liste de destinataires.
- Entête des messages souvent corrompues ou ne respectant pas les normes.

Statut légal La loi pour la confiance dans l'économie numérique du 21 juin 2004 contient une transposition de la directive européenne du 12 juillet 2002 [1] relative à la protection de la vie privée dans le secteur des communications électroniques :

> Est interdite la prospection directe au moyen d'un automate d'appel, d'un télécopieur ou d'un courrier électronique utilisant, sous quelque forme que ce soit, les coordonnées d'une personne physique qui n'a pas exprimé son consentement préalable à recevoir des prospections directes par ce moyen.

1. Le principe introduit figura également à l'article L.34.5 du code des postes et des communications électroniques

Les SPAM sont donc connus du droit français et encadrés par des textes spécifiques.

2 Le projet

2.1 Informations

Développeur	Apache Software Foundation	
Langage	Perl Dernière version	3.4.0 (11 février 2014) [+/-]
Environnements	Multiplate-forme	
Type	Anti-spam	
Licence	Licence Apache 2.0	

SpamAsassin est donc aujourd'hui sous le giron de la Apache Software Foundation, organisation à but non lucratif qui s'occupe également du serveur Apache, Logiciel de distribution de contenu WEB le plus utilisés au monde. Elle gère également 150 autres projets. EN outre tout ses projets sont distribués sous sa propre Licence, la licence Apache(actuellement en 2.0) , qui est compatible GPL v3. Cette licence met l'accent sur le copyright tout en restant bien sur une licence libre. Les objectifs principaux de la Fondation sont de protéger juridiquement le travail des contributeurs et d'empêcher que la marque Apache soit utilisée illégalement.

Le projet SpamAssassin est actif depuis plus d'une décennie et est constamment en développement pour s'adapter aux développements des méthodes qu'utilisent les spammeurs. C'est en outre le programme anti-spam le plus utilisé à cause de son efficacité.

2.2 Développement

SpamAssasin contient environ 300 000 lignes de codes ce qui en fait un très gros projet(Graphique 1). Le projet est à maturité et il ne grossit plus depuis plusieurs années, les développeurs se concentrant sur l'optimisation du code existant. Il y actuellement 23 développeurs, avec une répartition des lignes codes assez inégales, notamment deux développeurs qui ont fait la majorité du code(Tableau 1)

2.3 Qu'est ce que SpamAssassin

SpamAssassin est un programme écrit en PERL dont le but est de filtrer activement les Emails en se basant sur des mécanismes internes. SpamAssassin n'effectue aucune action envers les mails, il ajoute seulement des informations personnalisés qui peuvent être utilisée par d'autres programmes

Author Id	Changes	Lines of Code	Lines per Change
Totals	26092 (100.0%)	1403447 (100.0%)	53.7
jm	8136 (31.2%)	721593 (51.4%)	88.6
spamassassin role	7997 (30.6%)	463448 (33.0%)	57.9
axb	741 (2.8%)	54525 (3.9%)	73.5
mmartinec	1779 (6.8%)	32348 (2.3%)	18.1
felicity	1625 (6.2%)	29134 (2.1%)	17.9
kmcgrail	605 (2.3%)	21294 (1.5%)	35.1
quinlan	1100 (4.2%)	19583 (1.4%)	17.8
parker	309 (1.2%)	10407 (0.7%)	33.6
khopesh	1369 (5.2%)	10296 (0.7%)	7.5
dos	427 (1.6%)	8427 (0.6%)	19.7
jhardin	1021 (3.9)	6845 (0.5%)	6.7
wtogami	137 (0.5%)	6470 (0.5%)	47.2
hstern	63 (0.2%)	6029 (0.4%)	95.6
sidney	275 (1.1%)	3552 (0.3%)	12.9
jquinn	28 (0.1%)	2181 (0.2%)	77.8
mss	133 (0.5%)	2072 (0.1%)	15.5
hege	95 (0.4%)	1931 (0.1%)	20.3
duncf	109 (0.4%)	1892 (0.1%)	17.3
jgmyers	66 (0.3%)	508 (0.0%)	7.6
smf	35 (0.1%)	499 (0.0%)	14.2
maddoc	11 (0.0%)	319 (0.0%)	29.0
fanf	12 (0.0%)	49 (0.0%)	4.0
kb	19 (0.1%)	45 (0.0%)	2.3

TABLE 1 – Statistique des développeurs du projets
Source [2]

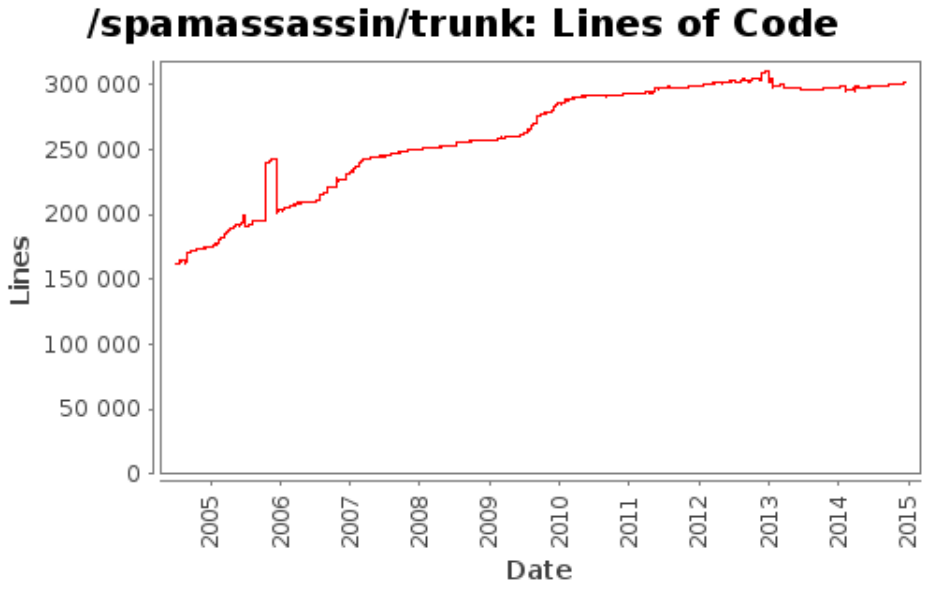

FIGURE 1 – Évolution du nombre de ligne de codes

pour effectuer des actions sur les mails (les ranger des dossiers distincts, les supprimer, les bloquer, ...)

Il peut être utilisé de plusieurs manières :

— En mode client, lancé à chaque fois que l'on fait appel à lui
— En mode demon grâce à *spamd* , les appels au demon étant fait avec l'utilitaire *spamc*.
— Comme une interface de programation : des programmes qui nécessitent des fonctionnalités de filtrage de SPAM peuvent s'interfacer avec SpamAssin pour construire des solutions utilisant ses fonctionnalités

3 Ses fonctionnalités

3.1 Comment il filtre

SpamAssassin reçoit des mails que lui redirigent d'autres programmes, y effectue des tests pour déterminer si ce sont des SPAMS, puis renvoie les mails testés au programme qui les as envoyés.

¡¡¡¡¡¡¡ HEAD
=======

4 Les test de SpamAssassin

¿¿¿¿¿¿¿ parent of 4da5823... correction + section4

Champs d'entête En se basant sur la forme des entêtes et en les comparant avec des schéma connus par SpamAssasin. En effet on peut se base sur la façon dont certains systèmes de SPAMs construisent leurs messages pour les filtrer.

Corps du message Bien sur SpamAssasin permet de filtrer les mails suivant les mots et expressions qu'ils contiennent. "ceci n'est pas un SPAM", "Bonjour jes suis une princesse d'un royaume africain", "Venez chechez votre lot", ...sont des expressions typiques pour des SPAMs.

Filtre bayésien Filtrer les entêtes et le corps d'un message résultera toujours en de multiples faux positifs. C'est ici que le filtrage bayésien se révèle interessant car il va prendre en considération ce que l'on considère comme SPam et non SPAM soit des "bon mails" ("HAM" en anglais). Il va ensuite utiliser les repertoires de SPAMs connu et de "HAM" connus, pour y identifier les mots et phrases (Définits comme "Tokens" en anglais) qui n'apparaissent que dans les SPAMs et que dans les "HAMs". Un token SPAM trouvé resultant d'une hausse du score (voir 4.1) SPAM, un token résultant en une baisse de ce niveau. Ce filtrage permet d'être plus précis et d'éviter les faux positifs, en ne se basant pas juste sur un mot ou une phrase mais des ensembles.

List noire/blanche automatique SpamAssassin garde automatiquement une liste blanches des expéditeurs des mails. Pour chaque nouveau mail le programme compare le mail précédent provenant de la même adfresse mail et adresse IP. Comme précédemment si une adresse email a envoyé un SPAM, ce nouveau mail verra son score abaissé. A l'inverse si c'était un bon mail son score se verra baisser.

Liste noire/blanche manuelle Il est tout à fait possible de définir ses propres listes, en autorisant ou interdisant des mails de certaines adresses

Signalements En utilisant des signatures établis à partir de mails signalés par les utilisateurs. Il y a notamment les projets DCC, Pyzor, et Razor2 qui possèdent des bases de données de mails signalés comme SPAM. SpamAssassin va ainsi demander à ces bases si les mails qu'il reçoit sont présents dans leurs donnée.

DNS blocklists Ce sont des des bases de données contenant des adresses IP signalées comme expédiant du SPAM ou mal configurée(Par exemple en étant un relais ouvert). Également sont pris en compte les IP de particuliers (considérant qu'il y a peu de chance qu'un particulier envois directement des mails sans passer par son FAI. Ces signalements vont être pris en compte par SpamAsassin pour le score des mails. Il intègre nativement quelque une de ces listes.

Caractères et langues on peut spécifier des caractères et langues comme SPAM.

C'est ces ensembles de règles qui fonctionnant conjointement permettent à SpamAssassin de garantir un haut niveau de fiabilité de détection des SPAM, un test pouvant ne pas fonctionner mais sera contrebalancé par les autres.

Spam Assassin effectue sur chaque mail qui lui est donné à traiter une série de test, qui vont ensuite donner lieu à un score, qui sera indiqué dans un entête si il est considéré comme SPAM. Ce résultat sera ensuite utilisé par d'autres programmes pour déterminer des actions à entreprendre.

4.1 Le score

C'est la base du signalement des SPAM de SpamAsassin. Un mail après avoir subi des test différents se voit attribuer une note. Cette note permet ensuite de définir des actions à effectuer. Quand un entête de mail est réécrit, SpamAssassin ajoute ses propres champs avec notamment le score, mais également d'autres données (Exemple 4.1). L'utilisateur peut paramétrer ce score pour définir une marge de définitions des SPAMs(valeurs "require")

5 Articulation du programme

5.1 Configuration

La configuration de Spamassassin se fait principalement à travers le fichier *local.cf*, qui se trouve dans le repertoire *'/etc/spamassassin*. Les utilisateurs UNIX peuvent également avoir leurs propres configurations grâce au fichier *user_prefs* qui se trouve dans le répertoire *.spamassassin* de leurs home respectifs. Par défaut, un certain nombre d'options sont prédéfinies. Voici les principales

- 5 définit le score au delà duquel les mails sont considérés comme du spam ;
- fr en indique les langues que vous acceptez de recevoir (les autres auront un score plus élevé). Cette ligne n'est pas forcément prédéfinie ;
- fr pour avoir les rapports en Français ;
- *@domaine.org Cette ligne (à éditer) permet de ne pas considérer les mails du domaine comme du spam.

5.1.1 Options de configuration

Spam assassin possèdent plusieurs options dont certaine non activés par défaut. Il suffit de les ajouter aux fichiers de configuration.

Préférences utilisateurs

Options de score required_score n.nn (default :5) Définit le score par défaut considérant un message comme SPAM. la valeur peut être un réel ou un entier

score SYMBOLIC_TEST_NAME n.nn [n.nn n.nn n.nn] Assigne les scores(voir 4.1) à un test donné.

```
X-Spam-Level: *********
X-Spam-Status: Yes, score=9.0 required=5.0 tests=BAYES_99,FROM_EXCESS_BASE64,
        FR_HOWTOUNSUBSCRIBE,FR_SPAMISLEGAL,FR_SPAMISLEGAL_2,HK_RANDOM_ENVFROM,
        HTML_IMAGE_RATIO_04,HTML_MESSAGE,UNPARSEABLE_RELAY autolearn=no version=3.3.1
X-Spam-Report:
        *  3.5 BAYES_99 BODY: Bayes spam probability is 99 to 100%
        *      [score: 1.0000]
        *  0.0 HK_RANDOM_ENVFROM Envelope sender username looks random
        *  1.0 FR_SPAMISLEGAL_2 BODY: French: droit d acces de modification de
        *      rectification
        *  2.0 FR_HOWTOUNSUBSCRIBE BODY: French: how to unsubscribe
        *  1.0 FR_SPAMISLEGAL BODY: French: Conformement ou En vertu....la loi
        *  0.6 HTML_IMAGE_RATIO_04 BODY: HTML has a low ratio of text to image area
        *  0.0 HTML_MESSAGE BODY: HTML included in message
        *  1.0 FROM_EXCESS_BASE64 From: base64 encoded unnecessarily
        *  0.0 UNPARSEABLE_RELAY Informational: message has unparseable relay lines
```