Les défis et les opportunités techniques du fonctionnement d'un service antispam mutualisé

Laurent Aublet-Cuvelier
GIP RENATER

José-Marcio Martins da Cruz
École des Mines de Paris

Résumé

L'objet de cet article est de décrire les spécificités d'un service antispam mutualisé dans le contexte de l'enseignement supérieur et la recherche. Il ne s'agit pas de décrire le service antispam « RENATER », mais d'identifier les défis à surmonter en vue d'assurer le meilleur fonctionnement du service.

Dans une première partie, on décrit les caractéristiques d'un service antispam mutualisé, notamment en comparaison avec un service opéré au niveau d'une entité. On rappelle les choix effectués pour le service proposé par RENATER.

Dans la deuxième partie, on s'intéresse à la vie du service en exploitation : comment assurer le bon fonctionnement au quotidien. Dans un premier temps, on s'intéresse à l'évaluation de la qualité du filtrage par l'utilisateur : le ressenti de la qualité du filtrage étant a priori une donnée subjective, nous décrirons les moyens pouvant être mis en œuvre pour une évaluation objective menant à une amélioration du filtrage. On traite notamment de l'importance de la boucle de rétroaction avec l'utilisateur final et de la problématique de l'évaluation de la performance du filtrage. Dans un deuxième temps, on traite du problème de la supervision et détection d'anomalies : quelque soit le produit de filtrage, il peut toujours présenter des défauts de fonctionnement. Vue la taille de la population concernée, il est important de pouvoir détecter tout défaut de fonctionnement dans les plus brefs délais.

Enfin, on présente deux projets en cours : le premier concerne la mise en place d'outils d'aide à la mutualisation de listes blanches et noires ; le second consiste à tester un autre moteur d'analyse de contenu, en parallèle du moteur actuellement en production..

Mots clefs

Mail, antispam, mutualisation, exploitation, listes blanches, supervision

1 Introduction

Le service antispam de RENATER a été mis en production en octobre 2009. Tous les sites raccordés à RENATER peuvent utiliser ce service pour leurs domaines de messagerie : il ne s'agit pas d'un filtrage imposé, mais d'un service mutualisé offert à la communauté. En octobre 2011, 47 sites sont utilisateurs, ce qui représente 247 domaines de messagerie raccordés et plus de 550 000 boîtes aux lettres[1]. Environ 2 millions de messages sont traités par jour (voir répartition hebdomadaire, Figure 1) dont au moins 70 % sont rejetés.

L'objet de cet article est de décrire les spécificités d'un service antispam mutualisé dans le contexte de l'enseignement supérieur et la recherche. Il ne s'agit pas ici de décrire le service antispam fourni par RENATER, mais d'identifier les défis à surmonter, en vue d'assurer le meilleur fonctionnement du service.

[1]On ne comptabilise pas les adresses de courriels, qui peuvent être multiples pour un même destinataire, mais les boîtes aux lettres effectives.

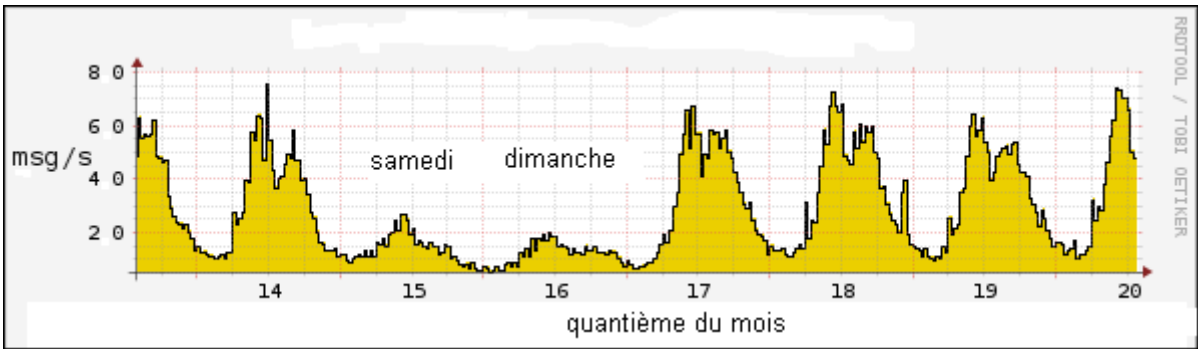

Figure 1 - Nombre de messages traités (sur une semaine)

2 Service antispam mutualisé

2.1 Spécificité mutualisation du service

Le problème de mutualisation d'un service antispam peut être compris avec deux points de vue, tous les deux ayant rapport avec la facilité de mutualisation.

Le premier point de vue concerne la diversité des utilisateurs : un fournisseur de boîtes aux lettres (Gmail, Hotmail, Yahoo...) où aucune inférence ne peut être faite sur les utilisateurs qui n'ont aucun point commun ; une entreprise où les utilisateurs peuvent être parfaitement classés et une stratégie de filtrage peut être imposée ; ou encore un organisme d'enseignement et recherche, où les utilisateurs ont clairement des points communs mais la typologie des boîtes aux lettres est moins claire.

Un autre point de vue concerne la distance entre le service mutualisé et l'utilisateur final : des utilisateurs dans le même service, du même établissement ou encore de l'ensemble des établissements. Dans les trois cas, plus on est loin de l'utilisateur final, moins on peut faire des hypothèses sur le contenu typique des boîtes aux lettres.

Du fait de la « distance » entre le système de filtrage et l'utilisateur final, on obtiendra des meilleurs résultats avec des méthodes fondées sur l'origine et la forme des messages (par exemple des listes noires et l'analyse de certains en-têtes) plutôt qu'avec des critères basés sur l'analyse du contenu textuel des messages.

2.2 Description du système antispam de RENATER

2.2.1 Traitement du flux de messages

Le diagramme simplifié de traitement du flux de message [6] donne un aperçu général des dispositifs participant au filtrage et de leur enchaînement.

2.2.2 Quelques méthodes utilisées

Même si tous les dispositifs contribuent à la qualité finale du filtrage, nous n'évoquons ici que ceux directement utiles au propos de cet article.

- Listes noires dynamiques

Une liste noire dynamique est une liste d'adresses IP identifiées comme sources de spams. Ces listes sont généralement proposées (vendues) par des fournisseurs et sont actualisées en continu. L'utilisation d'une liste noire dynamique permet d'éliminer au moins 70 à 90 % du trafic dès la connexion SMTP. C'est un phénomène très oscillatoire, le taux filtrage à un instant T pouvant varier de moins de 20 % à plus de 95 %, comme nous l'illustrons dans la partie 3.2. Cette variation est notamment fonction de l'activité humaine : la baisse du nombre de messages légitimes la nuit et le week-end fait monter le nombre de messages refusés, alors qu'il

n'y a pas de changement du nombre de spams. De même, cette méthode de filtrage réagit particulièrement à l'activation ou à la désactivation de réseaux de botnets par exemple.

Le filtrage initial par liste noire dynamique consiste en un rejet de message. Il convient donc d'être très attentif au choix de la liste utilisée. L'utilisation de ces listes est souvent une source de polémique du fait des stratégies plus ou moins sérieuses de gestion des listes. Dans certaines de ces listes, il est trop facile d'y entrer mais trop difficile d'en sortir. Et aucun serveur légitime n'est à l'abri d'un incident ponctuel. Une bonne liste doit être gérée avec des critères objectifs et avec une bonne réactivité.

Même avec une liste noire dynamique de très bonne qualité, il est néanmoins indispensable de proposer, en parallèle, un mécanisme de contournement. On utilise pour cela des listes blanches.

- Liste blanches et noires

Il convient donc de proposer des listes noires et blanches statiques afin de corriger les défauts de filtrage. Par exemple, ajouter une adresse IP en liste noire statique peut bloquer immédiatement un flux illégitime, voire dangereux (phishing, nouveau malware), avant que les mécanismes « normaux » n'aient appris à le bloquer à leur tour. Au contraire, il peut être nécessaire d'autoriser tout le flux venant d'une source précise pour éviter qu'il ne soit filtré par un mécanisme trop sensible, comme le moteur d'analyse de contenu, qui n'est pas exempt d'erreurs de classement. C'est une action corrective (après détection d'une erreur de filtrage) ou préventive (pour éviter qu'une source légitime ne soit bloquée).

Dans le cas de l'antispam RENATER mutualisé, il existe des listes blanches et noires qui sont soit globales (valables quel que soit le domaine de destination), soit spécifiques à chaque domaine. Elles peuvent comporter des adresses IP émettrices (IP source), des adresses de courriel d'émetteurs (MAIL FROM du protocole SMTP [1]), des adresses de courriel de destinataires (RCPT TO du protocole SMTP) et même le contenu du champ « objet » d'un courriel (en-tête « Subject: »).

- Analyse de contenu

Plusieurs mécanismes intervenant dès les premières phases du protocole SMTP (connexion, Mail From, Rcpt To) permettent de filtrer une majorité de messages indésirables (70 à 90 % des messages rejetés en général). En complément, on utilise sur la partie restante (10 à 20 %) un moteur d'analyse de contenu. Comme son nom l'indique, il analyse essentiellement le contenu transmis lors de la phase "DATA" du protocole SMTP. Il s'agit d'un classificateur, qui rend généralement 3 niveaux de résultat : spam, ham[2] et suspect (on trouve également d'autres indications de classement, telles que « virus », « message à caractère commercial », « message de type "large diffusion" », etc.).

Un classificateur idéal devrait être binaire (spam ou ham), sans zone d'incertitude, et surtout ne commettre aucune erreur. Mais les techniques de classement ne sont pas sûres à 100 % et, de surcroît, la notion même de message indésirable peut varier d'un individu à l'autre, comme nous le verrons dans la partie 3.1.1.

L'indication de la zone d'incertitude est généralement exprimée par un score ou une sorte de probabilité (0 = ham, 1= spam, et par exemple, suspect entre 0,90 et 0,99). Dans le cas du moteur choisi pour l'antispam mutualisé de RENATER, le score est un entier relatif : par défaut, un score supérieur à 300 indique un spam certain, un score inférieur à 100 indique un ham, et un score entre 100 et 300 est un message suspect. Le comportement par défaut est donc de rejeter les spams certains, de livrer les hams, et de marquer les suspects (le marquage est paramétrable par domaine de destination : ajout d'en-tête, modification du sujet, etc.) .

Certains outils de filtrage de spam gèrent une quarantaine par utilisateur, mais cela nécessite de connaître et d'identifier chaque utilisateur, ce qui dépasse le but visé par cette mutualisation : la quarantaine, qui n'est finalement qu'une boîte aux lettres particulière de l'utilisateur, doit être gérée par l'hébergeur des boîtes aux lettres.

[2]spam/ham : l'utilisation du mot « spam » pour désigner les messages indésirables a été inspiré par un sketch des Monty Python sur un produit de Hormel Foods à base de viande épicée (Spiced Ham), par opposition, on a coutume d'appeler les « bons » messages « ham » (jambon).

De plus, nous avons fait le choix de fournir le maximum d'information au site de destination, qui peut donc prendre des décisions à son tour. Par exemple, le score obtenu par le moteur d'analyse de contenu est propagé par l'ajout d'un en-tête spécifique, qui peut alors être utilisé par le MDA pour aiguiller les messages d'une certaine classe vers une quarantaine.

Enfin, les seuils sont également paramétrables par domaine de destination. Ainsi, un site peut décider qu'aucun rejet ne doit être effectué par le moteur d'analyse de contenu, afin de traiter sur son infrastructure le comportement final en fonction du score obtenu.

Plus que le réglage du seuil de rejet, l'enjeu majeur de mesure de la qualité d'un moteur d'analyse porte plutôt sur les taux de faux positifs (messages légitimes classés à tort comme spam) et faux négatifs (messages indésirables non détectés). Ces taux d'erreurs, leur évaluation, et leur ordre d'importance sont discutés dans la partie suivante.

3 Maintien en condition opérationnelle et évolution

Le flot de messages n'est pas stationnaire : ses caractéristiques évoluent en permanence. Il n'est pas aberrant de considérer que les messages légitimes évoluent peu : les expéditeurs changent rarement leurs habitudes d'écriture et, sauf dans certains cas spécifiques, la topologie de leurs réseaux de correspondants reste relativement stable. Néanmoins, dans un antispam mutualisé, les caractéristiques des expéditeurs et destinataires et des messages échangés couvrent un large spectre (différence de population, de discipline scientifiques, etc. et différence d'utilisation du courrier électronique).

Le spam, par contre, évolue en permanence, surtout pour déjouer les filtres. Ceci explique le besoin constant de mise à jour des mécanismes de filtrage, et d'autre part, de mettre en œuvre des mécanismes de contrôle de la qualité du filtrage.

3.1 Évaluation de la qualité du filtrage

3.1.1 Subjectivité implicite de l'utilisateur

Il est toujours difficile de demander aux utilisateurs d'évaluer la qualité d'un service de filtrage, les réponses obtenues sont souvent plus affaire de ressenti que de données objectives, et cela pour plusieurs raisons. Tout d'abord, le jugement de l'utilisateur ne porte que sur les messages qui lui sont effectivement remis. Il n'a souvent pas conscience de tout ce qui a été éliminé en amont (qui peut représenter plus de 90 % des messages qui lui sont destinés) alors que l'administrateur du service, lui, s'en félicite !

Ensuite, un taux de faux positifs trop élevé peut s'avérer très préjudiciable : un seul courrier important vous manque et tout est dépeuplé ! On préfèrera donc, du point de vue technique de réglage du filtre, avoir plutôt des faux négatifs que des faux positifs, ce qui nuit à la perception de la qualité du filtrage.

Enfin, la notion même de spam est subjective. Dans le monde réel, une personne peut être ravie de recevoir dans sa boîte aux lettres les avis de promotion de son supermarché préféré alors que son voisin déteste être obligé de trier attentivement son courrier pour retrouver les vraies lettres au milieu de tous ces prospectus qui encombre sa boîte.

Il faut donc utiliser des moyens plus objectifs. Pour cela, on cherche plutôt à inciter l'utilisateur à signaler les erreurs de classement dont il est victime.

3.1.2 Boucle de rétroaction utilisateur

Les moteurs d'analyse de contenu nécessitent un apprentissage constant, qu'ils soient basés sur des méthodes statistiques, des ensembles de règles, ou d'autres méthodes. Il doivent donc être alimentés par le retour des erreurs de classement. Seul l'utilisateur final est à même de détecter efficacement ces erreurs de classement. Il faut donc impérativement mettre en œuvre une boucle de rétroaction qui permet de remonter à l'éditeur du filtre de contenu le maximum d'erreurs de classement, en sollicitant le concours des utilisateurs.

Au départ, le seul mécanisme de signalement d'erreurs de classement proposé par l'antispam mutualisé est la mise à disposition des postmasters de chaque site d'une boîte aux lettres « Spam » et d'une boîte aux lettres « NoSpam », dans lesquelles ils peuvent

déposer respectivement les faux négatifs et les faux positifs. Dans ce mécanisme, fonctionnant en deux étapes, l'absence de formalisme dans la récupération des messages par les postmasters rend le processus fastidieux et inefficace : les solutions variant entre une redirection manuelle des messages mal-classés et une collecte automatique des boîtes personnelles « Spam ».

Le développement récent d'un plugin « report-spam » (uniquement pour le client de courriel Thunderbird pour l'instant) [5], devrait permettre d'améliorer la boucle de retour, à la fois sur le nombre et la qualité des messages.

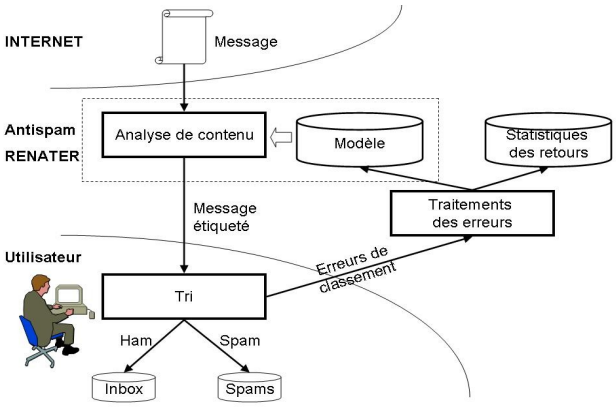

Figure 2 - Boucle de rétroaction

Cette boucle de rétroaction sert donc bien évidemment à alimenter l'apprentissage du moteur d'analyse de contenu, mais c'est aussi un élément précieux pour l'évaluation plus objective de la qualité du filtrage.

Ainsi, contrairement aux signalements effectués via le postmaster du domaine, pour lesquels il n'y pas de format imposé (tantôt message brut, message « attaché », message « inclus », etc.), le plugin « report-spam » permet un retour des messages mal classés dans un format unique facilitant le traitement automatique de la collecte. Le robot auquel sont envoyés les messages ainsi signalés effectue 2 actions (cf. Figure 2) : d'une part la remontée vers l'éditeur, pour correction du filtre, et d'autre part, une collecte de données statistiques (par exemple, le nombre de faux positifs/faux négatifs par domaine raccordé) permettant d'évaluer l'évolution de la qualité du filtrage dans le temps.

3.2 Supervision et détection d'anomalies

Lorsqu'on opère un service, quel qu'il soit, il est impératif d'en assurer la supervision. On dispose pour cela d'outils classiques comme la supervision des serveurs et de leurs processus (disponibilité, consommation des ressources, etc.). Dans notre cas, ces outils ne suffisent pas : il faut également être capable de détecter toute anomalie de fonctionnement au plus tôt. En effet, alors qu'on traite plusieurs millions de messages par jour, un dysfonctionnement peut immédiatement avoir des conséquences importantes.

Par exemple, l'arrêt du filtrage par la liste noire dynamique signifierait une hausse brutale du nombre de messages indésirables livrés (donc une gêne importante pour l'utilisateur) de surcroît potentiellement dangereux (phishing, etc.), mais aurait aussi certainement des conséquences sur l'infrastructure, puisque les serveurs des sites ne sont plus dimensionnés pour une telle capacité de traitement. On aurait alors vraisemblablement une perte de service plus globale. Un dysfonctionnement du moteur d'analyse de contenu pourrait, quant à lui, entraîner une hausse de faux positifs préjudiciable pour les utilisateurs.

Le véritable enjeu n'est donc pas de détecter la panne franche, mais bien de détecter un fonctionnement anormal, dans un délai court, pour en minimiser l'impact.

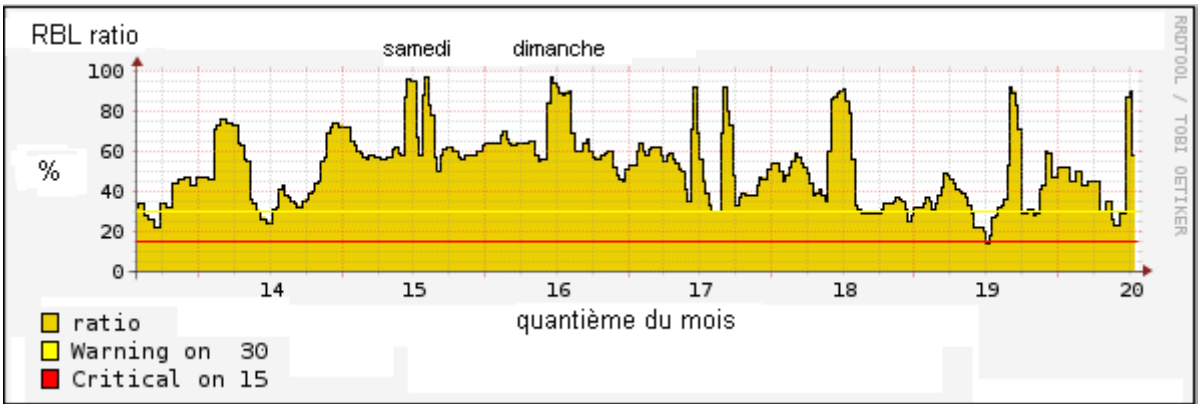

Figure 3 - Taux de filtrage par la liste noire dynamique (sur une semaine)

On peut croire qu'il suffit de mettre en place une simple surveillance de compteurs permettant, par exemple, d'observer le taux de filtrage et déclencher une alerte si ce taux est en dessous d'un seuil fixé. Malheureusement, comme le montre la Figure 3, l'intensité du flot de messages n'est pas constante. On observe dans une journée, plusieurs ruptures franches du taux de messages rejetés passant, par exemple, de 95 à 20 % sur une courte durée. Cela peut notamment s'expliquer par des cycles d'attaques contrées (activation/désactivation d'un réseau de botnets, entrée en liste noire dynamique des adresses IP correspondantes...) ou alors par les cycles d'activité professionnelle des correspondants légitimes.

Sur une période courte, on ne peut donc pas distinguer dégradation du service et comportement normal. Ainsi, on ne sait souvent détecter automatiquement une anomalie que sur une période aussi longue que celle nécessaire à l'utilisateur pour la constater et en subir les effets. Une éventuelle dégradation du service ou comportement anormal sera donc assez souvent ressentie par un certain nombre d'utilisateurs avant d'être détectée par des outils automatiques de surveillance.

4 Perspectives

Comme on l'a vu, la gestion d'un service antispam nécessite des processus constants d'amélioration: évaluation de la qualité, boucle de rétroaction, alimentation des processus d'apprentissage des moteurs d'analyses, détection d'anomalie.

4.1 Mutualisation de listes blanches et noires statiques

Chaque administrateur de site peut mettre en œuvre des listes blanches et noires pour les domaines qu'il gère. Il apparaît parfois que l'ajout d'une entrée dans une liste intéresse une grande partie, voire l'ensemble de la communauté. Ce sont souvent des mises en liste blanche préventive. Par exemple, on voudra s'assurer que les messages du CERT RENATER ne soient pas filtrés. Plus récemment, lors d'une campagne de vote électronique initiée par le ministère de tutelle de nombreux établissements raccordés, on a souhaité mettre en liste blanche le serveur SMTP du fournisseur du service de vote.

Au départ, le seul moyen de gérer ces listes passe par un opérateur du GIP RENATER qui alimente, en fonction des demandes, les listes effectivement en production, avec des délais inévitables de mise en œuvre des modifications. Nous avons donc développé une interface Web accessible à chaque administrateur de site. Du fait de l'implémentation interne des listes noires et blanches sur les serveurs en production (solution du fournisseur, qu'on ne maîtrise pas), les tests de bon fonctionnement de la chaîne de modification d'une liste à partir de cette interface Web sont particulièrement sensibles. En effet, il serait assez facilement possible d'écraser totalement les listes en place de tous les domaines, en cas de bug ou manque de contrôle.

Néanmoins, l'automatisation de ces procédures de modification des listes blanches et noires de chaque domaine raccordé offre des perspectives intéressantes dans le cas d'un service mutualisé. Nous travaillons ainsi à ajouter des mécanismes permettant de détecter qu'une entrée de liste est, par exemple, ajoutée par plusieurs sites : ce n'est pas trivial, la forme utilisée pouvant varier, pour un même but. Ces mécanismes visent à détecter, par exemple, des anomalies (action corrective) ou des possibilités de mutualisation (action préventive).

4.2 Test d'un moteur d'analyse de contenu alternatif

Afin d'évaluer, voire d'améliorer la qualité du filtrage, outre les systèmes d'observation décrits précédemment, la plate-forme RENATER accueille pour un test de quelques semaines un deuxième système d'analyse de contenu. Ce second moteur est interrogé en parallèle de celui en production, mais ne sert pas à filtrer : il n'est qu'indicatif. La comparaison des classements effectués par le moteur 1 (production) et le moteur 2 (test) pour chaque message montre surtout une différence au niveau de la zone grise d'incertitude (spam suspecté). Ainsi pour les hams francs et les spams francs du moteur 1, le résultat est similaire avec le moteur 2. Par contre, pour les messages classés « incertains » par le moteur 1, le moteur 2 est plus souvent plus catégorique en classant franchement spam. On pourrait croire que le moteur 2 serait donc plus performant, puisque capable d'éliminer plus de spams. Malheureusement, rien ne permet de savoir, automatiquement, si le classement du moteur 2 est pertinent. Seul l'humain, encore une fois, est capable de confirmer ou d'infirmer le résultat. Or, comme nous avons vu, un taux de faux positifs important est inacceptable.

Là encore, l'utilisation de la boucle de rétroaction va nous aider. Le principe, en cours de mise en œuvre, est d'utiliser les signalement d'erreur de classement (du moteur 1, seul résultat utilisé visiblement pour le filtrage en production), et de vérifier si le moteur 2 a vu juste ou non. Malheureusement, cela reste encore imparfait : en effet, les utilisateurs signalent volontiers les faux négatifs (spam non détectés) mais moins les faux positifs (messages légitimes classés suspects) car les messages légitimes ont souvent un caractère confidentiel pour l'utilisateur. La statistique portera donc surtout sur les faux négatifs. Par ailleurs, ce cas de figure ne permet pas d'évaluer les faux positifs du moteur 2 qui sont bien classés par le moteur 1.

5 Conclusion

La mise en service d'une plate-forme antispam mutualisée au sein de RENATER pose des défis uniques, du fait d'une qualité de service attendue très supérieure aux offres grand public, et de la diversité des établissements qui constituent le réseau RENATER.

Nous avons évoqué dans cet article les éléments caractéristiques de ces défis : l'éloignement entre l'opérateur et l'utilisateur ainsi que la diversité des contextes d'utilisations. Cette diversité, intrinsèque à la mutualisation, va à l'encontre des caractéristiques des plate-formes de filtrage antispam, plutôt conçues pour une organisation centralisée. Ces éléments caractérisques du service antispam de RENATER constituent autant de défis à surmonter pour atteindre une qualité de service optimale.

Cet objectif de qualité de service nécessite la mise en place de nouveaux outils de supervision et de communication. La communication doit avoir lieu avec les informaticiens des entités raccordées (statistiques, signalement d'anomalie, supervision) mais également avec les utilisateurs de la messagerie électronique (boucle de rétroaction). Les développements en cours constituent un effort important au service de la communauté, et contribuent à renforcer l'attractivité du service.

6 Bibliographie

[1] J. Klensin, RFC 5321 : Simple Mail Transfer Protocol, IETF, octobre 2008.

[2] P. Resnik, RFC 5322 : Internet Message Format IETF, octobre 2008.

[3] D. Crocker, RFC 5598 : Internet Mail Architecture, IETF, july 2009.

[4] J.-M. Martins da Cruz, Contribution au classement statistique mutualisé de messages électroniques (spam) – Thèse de doctorat Mines-ParisTech, Paris, Octobre 2011

[5] S. Aumont, E. Méléard, H. Lascaux, Report-spam : un plug-in Thunderbird pour escalader les anomalies de filtrage antispam, JRES 2011.

[6] RENATER, Schéma fonctionnel du processus de filtrage antispam, http://www.renater.fr/IMG/pdf/WF-simplifie-20101018.pdf