OSSIR – 07 juillet 2009

Filtres statistiques de spam : Etat de l'art et utilisation collective

Jose-Marcio.Martins@mines-paristech.fr

j-chkmail

- http://www.j-chkmail
- Filtre anti-spam open source/libre – existe depuis 2002
  - Limitation de cadences, ressources, ... (première implementation depuis 2002).
  - Greylisting
  - Filtrage d'URL
  - Filtre statistique
  - ...

Sommaire

- Historique
- État de l'art
- Utilisation collective
- Conclusions

Historique

Historique

- Les acteurs :
  - Développeurs de filtres
  - AI/IT : machine learning, information retrieval, data mining, ...
  - Statisticiens
- Vers une cohabitation ???

Les développeurs

- ~ 2001 – Spam Assassin
  - Réseau de neurones (perceptron)
  - Les attributs : Nombre fixe de tests (~ 750)
- 2002 – Paul Graham – A plan for spam
  - Classificateurs „bayésiens“ : Bogofilter, SpamBayes, CRM114, ...
  - Attributs : le texte – mots, bimots, ...
- Obs : pourquoi on range ensemble ces deux types de filtre ???

De la recherche

- Domaines
  - Intelligence artificielle : machine learning
  - information retrieval, data mining
  - Statistique
- Des milliers d'articles de recherche (dont quelques centaines sont intéressants)
- Plusieurs conférences :
  - CEAS, MIT Spam Conference, Eurospam, ...

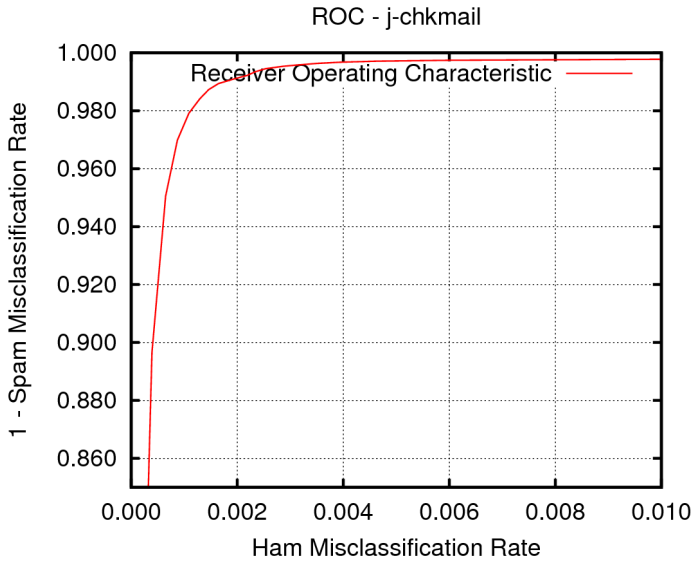
La recherche

- 1998 - M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz – A Bayesian Approach to Filtering Junk E-Mail.
- Tests comparatifs des différents types de classificateur
- 2003 – Tom Fawcett - "In vivo" spam filtering: A challenge problem for data mining
  - Définition de la problématique SPAM (presque un survey...)
    - Comportement temporel dynamique
    - « Course aux armements »
    - ...
  - -> Le filtrage de spam est un problème particulier et intéressant !

TREC – Text Retrieval Conference

- 2005/2007 – Spam Track (http://trec.nist.gov)
  - Création d'un corpus public - ENRON
  - Méthodologie d'évaluation
    - Critères – ROC, 1-ROCA, LAM, hmr, smr
    - Boîte à outils d'évaluation
  - Etude des contextes d'évaluation : en ligne vs batch, delayed feedback, active learning, ...
  - Participation différents acteurs : logiciels libre et recherche
  - 2005 winners : IJS (compression), CRM114, dbacl
  - 2007 winners : waterloo (compression+logreg), thufts (SVM)

Ex : Receiver Operating Characteristic

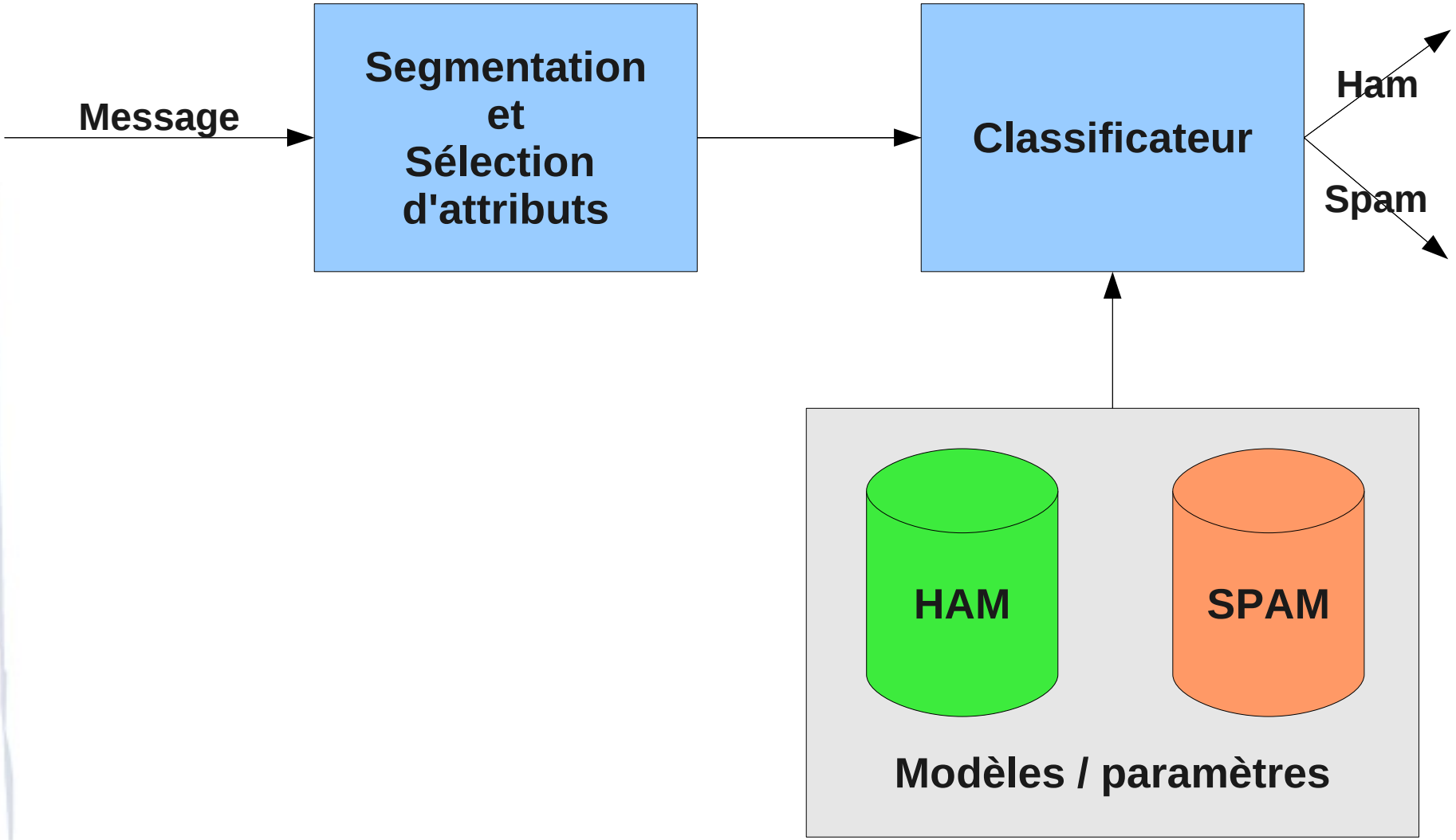


Les filtres statistiques

Deux familles de filtres

- Règles fixes (handcrafted) : SpamAssassin
  - ~ 750 tests : un vecteur de dimension fixe
  - Apprentissage : réseau de neurones -> poids de chaque test
  - Classement : somme des poids des tests positifs
- Classificateurs textuels
  - Règles d'extraction des attributs – ce sont les tests
  - Nombre de tests : taille du dictionnaire
  - Apprentissage : constitution du dictionnaire
  - Classement : dépend du classificateur

Ce qu'il y a dedans



Classificateur Bayésien

- Avantages
  - Facile à mettre en œuvre
  - Efficace
  - Bonne résistance au bruit
- Principe

$$P(Classe|Message) = \frac{P(Message|Classe) * P(Classe)}{P(Message)}$$

$$classement = \underset{C=ham,spam}{argmax} \; P(Message|C) * P(C)$$

- Classement : P(Message|C) ...

Détour sur la Segmentation

- Texte non structuré -> représentation vectorielle
  - VSM - vector space model
- Hypothèse d'indépendance -> modèle « bag of words »
  - Mots
  - Bi-mots
  - n mots sur N (filtres dit « markoviens »)
  - N-grams
- Sélection d'attributs – réduction de la dimension, réduction du bruit
- ***En fait : comparaison de vocabulaires***

Classificateur Bayésien

- Classification – « Naive Bayes »

$$P(Message|C) \sim P(T_1, T_2, \ldots, T_n|C) = \prod P(T_i|C)$$

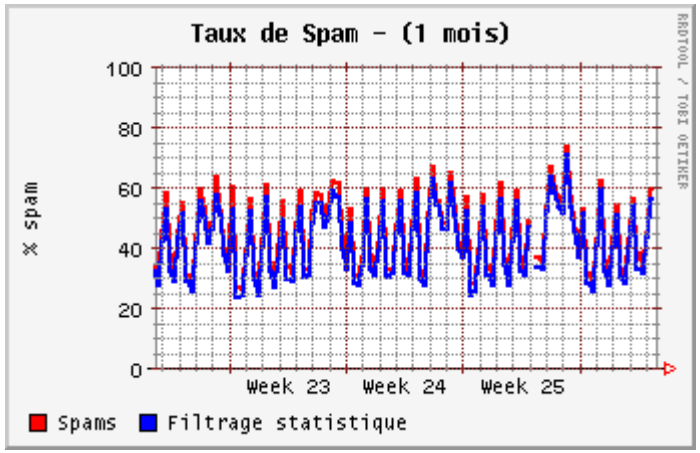
- Apprentissage
  - Comptabiliser le nombre d'apparitions de chaque terme.

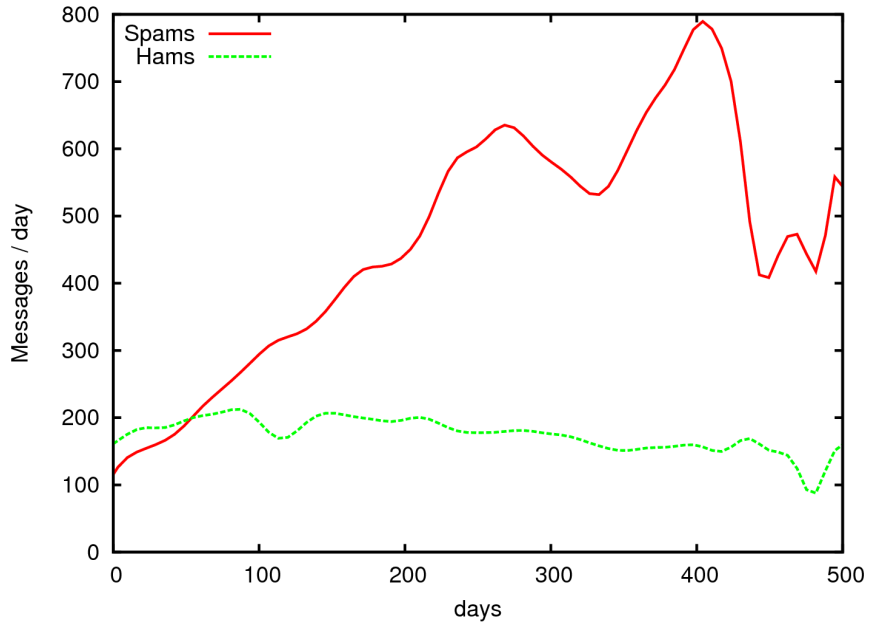
$$P(T_i|ham) \quad et \quad P(T_i|spam)$$

- ***<u>En fait : classement = comparaison de vocabulaire</u>***

Classificateur Bayésien

- Quid de la probabilité à priori P(Classe) ???





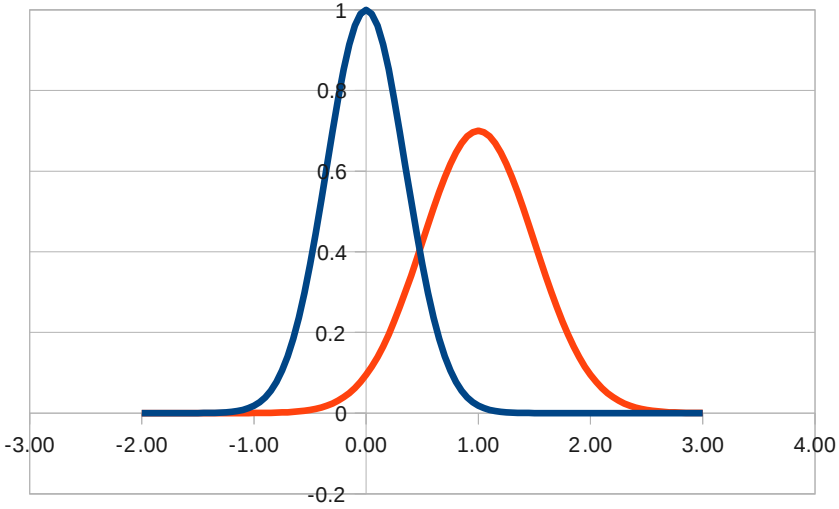
- On ne tient pas compte...
- Classificateur bayésien vs vraisemblance maximale ???

$$classement = \underset{C=ham,spam}{argmax} \; P(Message|C)$$

- Mais pourquoi ça marche ???

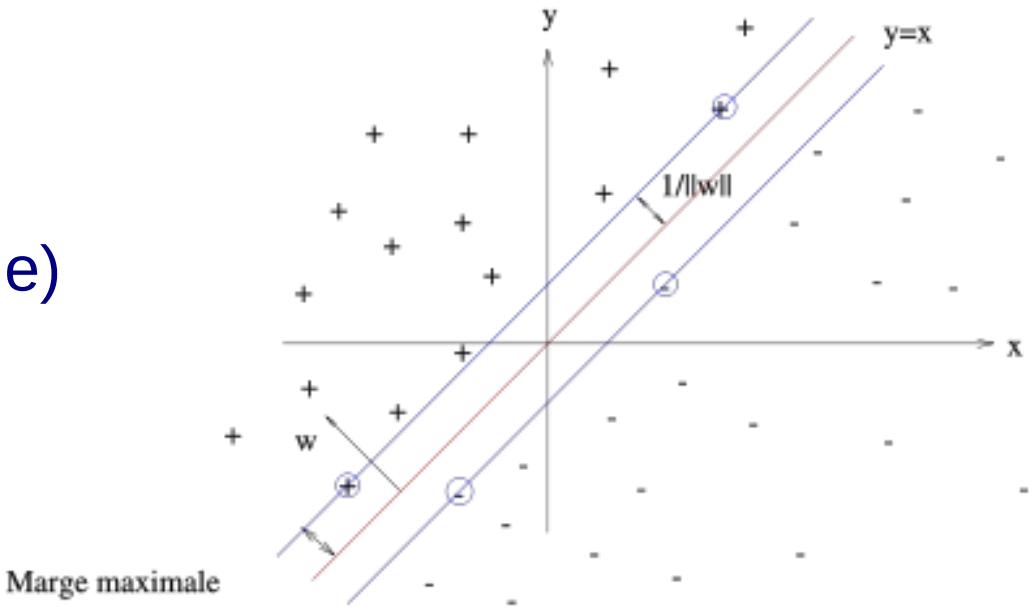
$$classement = \underset{C=ham,spam}{argmax} \; P(Message|C) * P(C)$$

- Ex : classement à une seule dimension.
- Moins ça « chevauche », moins la probabilité à priori est essentielle :
  « Distributions séparables » classificateurs discriminants !



SVM - Support Vector Machine

- Hyperplan séparateur
- Classification : trouver de quel côté de l'hyperplan se trouve le message à classer
- Apprentissage = trouver l'équation de l'hyperplan (programmation quadratique)
- Obs : le résultat a une interprétation géometrique.



Marge maximale

Régression Logistique

- Classement : calcul direct de la probabilité conditionnelle

$$P(Classe|Message) = \frac{1}{1+e^{\vec{w}\cdot\vec{M}}}$$

- Apprentissage : trouver les coefficients w – optimisation par descente de gradient

Compression

- MDL – Minimum Description Length
- Classement : Le message est « compressé » avec chaque modèle et associé à la classe dont le modèle donne le meilleur taux de compression
- Méthodes (compression sans perte) :
  - DMC – Dynamic Markov Compressor – G. Cormack - 1987
  - PPM – Prediction by Partial Matching - Cleary, Witten - 1984

Efficacité

- Efficacité typique (TREC 1-ROCA %)
  - Classificateur bayésien : ~ 0;05 % - 0.1 %
  - SVM, Regression Logistique, Compression : ~ 0.01 %

Apprentissage

- Créer un modèle (statistique, représentatif) pour chacune des classes
- La qualité du filtrage dépend fortement de la qualité de l'apprentissage
- Processus non stationnaire - méthodes canoniques
  - Fenêtre temporelle glissante
    - Stockage des messages (« instance based »)
  - Incrémental
    - Les messages sont supprimés après utilisation
    - Ne convient pas à tous les types de classificateur

Apprentissage

- Heuristiques courantes (elles ne sont pas toutes bonnes) :
  - Apprentissage active
  - Tous les messages
  - Une seule classe
  - Seulement les erreurs
  - Jusqu'à ne plus avoir d'erreurs
  - Suppression des termes peu/pas utilisées
  - Apprentissage non/semi supervisée

Apprentissage

- Quelques problèmes ouverts :
  - Heuristiques efficaces d'apprentissage
  - Retour d'information avec retard
  - Traitement du bruit : erreur de jugement, malveillance, « inclassables », ...
  - Apprentissage pour utilisation collective
  - ...

Utilisation collective de filtres statistiques

Pour quoi ?

- L'apprentissage d'un filtre n'est pas une tâche triviale pour des non spécialistes
- Mutualiser !
- ...

Les points durs

- Difficulté de construction du modèle
  - Disponibilité des exemples – confidentialité / vie privée
  - Représentativité du modèle
  - Évolution temporelle (processus non stationnaire)
- Évaluation de la qualité du filtrage pas triviale
- Retours d'information incertain et avec retard
- Idée :
  - Constitution d'un corpus synthétique d'apprentissage

Synthèse d'un corpus

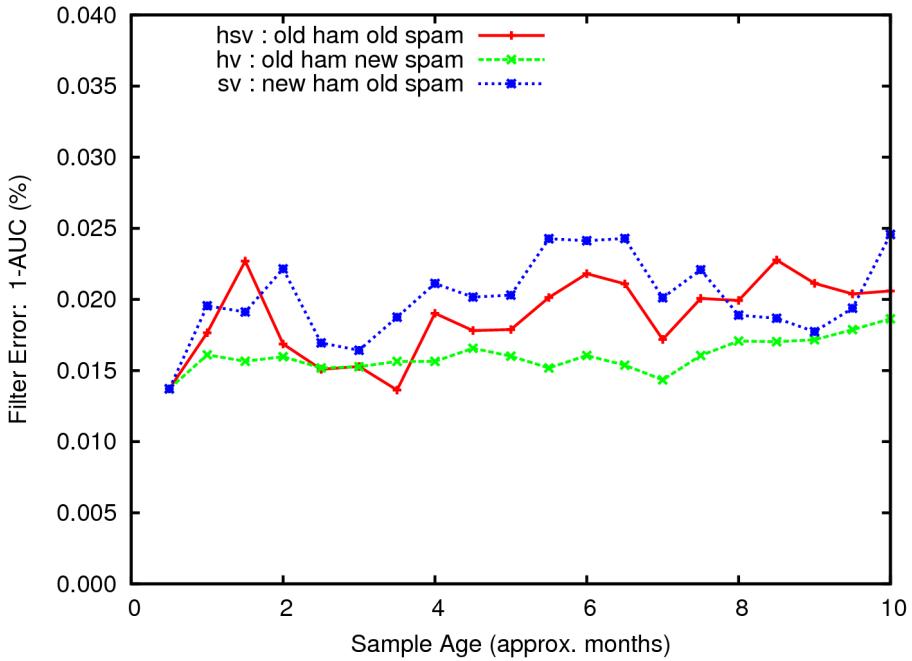
- Une boîte aux lettres individuelle est l'union de « sub-boîtes » : les sujets.
- Une boîte aux lettres collective est une « combinaison linéaire » de boîtes aux lettres individuelles, donc de sujets
  - Modèle hiérarchique
  - Des sujets en commun
  - Modèle statistique de mélange de distributions
- Les caractéristiques statistiques du corpus, vues par le classificateur doivent être équivalentes à celles des messages réels

Dans le temps...

- Constats :
  - Certaines caractéristiques des messages ne sont pas aussi changeantes qu'on le pense
  - La corrélation entre les deux classes est faible
- Conclusions :
  - On peut traiter les deux classes indépendamment
  - On peut utiliser des messages non récents (avec quelques précautions) pour entrainer un filtre
-
  - J.M. Martins and G. V. Cormack (http://www.j-chkmail.org/ceas)
    - On the Relative Age of Spam and Ham Training Samples for Email Filtering
    - Using old ham and spam samples to train email filters

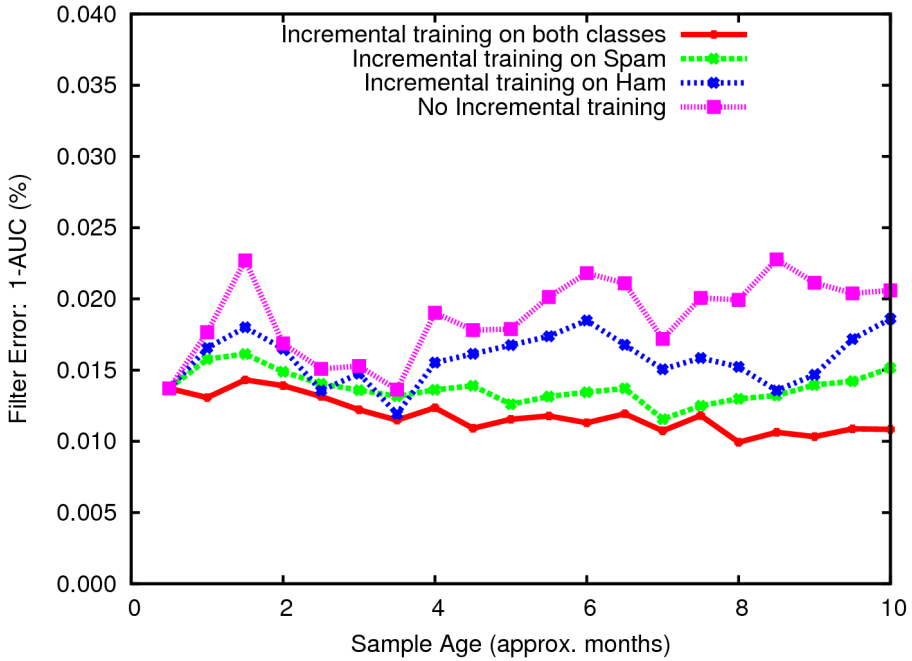
Dans le temps...

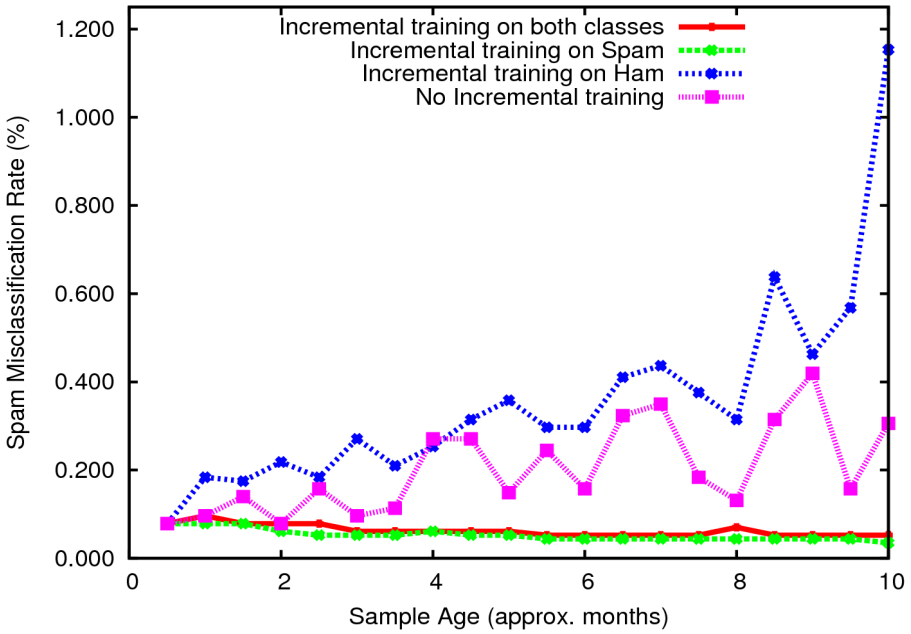
- Des vieux messages



Dans le temps...

- Apprentissage incrémental





J-chkmail

- Classificateur « bayésien » (ML, en fait)
- Base d'apprentissage téléchargeable – hash MD5 –
  - quelques universités en France
- Apprentissage (fenêtre temporelle glissante) :
  - Hams : messages d'origine diverses, quelques listes de diffusion, ... Mise à jour éventuelle.
  - Spams : les N derniers mois de mes spams + 1 mois de messages de pots de miel. Mise à jour quotidienne.

J-chkmail

- Efficacité estimée : FPR ~ 0.1 %, FNR ~ 1- 5 %
  - Pas pire que certains produits commerciaux
- Largement perfectible !!!
- Travail actuel :
  - coefficients de mélange
  - Intégration d'un classificateur régression logistique

Conclusions

- Des nouvelles méthodes de filtrage sont en train de voir le jour, mais le filtrage bayésien reste compétitif
- L'utilisation collective d'un filtre statistique donne des résultats moins bons que l'utilisation individuelle, mais n'est pas à exclure.