Utilisation de la Machine Cellulaire pour la Détection des Courriels Indésirables

F. Barigou*, B. Atmani**, B. Beldjilali***
Equipe SIF, Laboratoire d'Informatique d'Oran
Université d'Oran BP 1524, El M'Naouer, 31 000 Oran, Algérie
* fatbarigou@gmail.com, **atmani.baghdad@univ-oran.dz, ***bouzianebeldjilali@yahoo.fr

1 Introduction

Dans ce papier, nous proposons pour la première fois une approche de filtrage de spam qui se base sur l'induction symbolique par automate cellulaire (Atmani et Beldjilali, 2007). Ce choix de cette technique a été motivé par ses propriétés intéressantes comme la réduction de l'espace de stockage et le temps de classification. Le principe de cette approche est très simple, il s'agit de construire un modèle booléen à partir d'un ensemble de courriels d'apprentissage. Ce modèle, qui sera utilisé pendant la phase de classification, va permettre de déterminer la nature d'un nouveau courriel (spam ou légitime).

2 Approche Cellulaire de Classification

L'automate cellulaire CASI (Cellular Automata for System Induction) issue des travaux de (Atmani et Beldjilali, 2007) est une méthode cellulaire de génération, de représentation et d'optimisation des graphes d'induction (Zighed,2000) générés à partir d'un ensemble d'exemples d'apprentissage. Ce système cellulo-symbolique est organisé en cellules où chacune d'elles, est reliée seulement avec son voisinage. Toutes les cellules obéissent en parallèle à la même règle appelée fonction de transition locale, qui a comme conséquence une transformation globale du système. Deux composants coopèrent entre eux pour la construction du modèle booléen : le COG (Cellular Optimization and Generation) qui s'occupe de la génération du graphe d'induction cellulaire et de son optimisation et le CIE (Cellular Inference Engine), un moteur d'inférence cellulaire, qui génère un ensemble de règles cellulaires sous formes conjonctives utilisées pendant la phase de filtrage. Pour se faire, ils utilisent une base de connaissances sous forme de deux couches finies d'automates finis. La première couche, CelFact[1], pour la base des faits et, la deuxième couche, CelRule[2], pour la base de règles. Le voisinage des cellules est défini par deux matrices d'incidence d'entrée R_E et de sortie R_S. La dynamique de l'automate cellulaire, utilise deux fonctions de transitions $\delta fact$ qui simule les phases de sélection et de filtrage dans un système expert et $\delta rule$ qui correspond à la phase d'exécution :

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{fact}} (EF, IF, \boldsymbol{EF}, ER + (R_E^T \times \boldsymbol{EF}), IR, SR)$$

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{rule}} (\boldsymbol{EF} + (R_S \times \boldsymbol{ER}), IF, SF, ER, IR, \overline{\boldsymbol{ER}})$$

[1] Toute cellule de CelFact est considérée comme fait établi si sa valeur est 1, sinon, elle est considérée comme fait à établir. Elle se présente sous trois états : état d'entrée (EF), état interne (IF) et état de sortie (SF)

[2] Toute cellule de CelRule est considérée comme une règle candidate si sa valeur est 1, sinon, elle est considérée comme une règle qui ne doit pas participer à l'inférence. Elle se présente sous trois états : état d'entrée (ER), état interne (IR) et état de sortie (SR)

2.1 Prétraitement linguistique et sélection des termes

On applique un prétraitement linguistique à chaque courriel du corpus d'apprentissage. L'objectif étant d'extraire un sous ensemble de mots représentatifs du contenu de cet ensemble de courriels. La représentation vectorielle est construite après avoir sélectionné les mots discriminants par une des méthodes de sélection (IG, IM, $\chi 2$).

2.2 Apprentissage

Nous résumons les principales étapes comme suit :

1. Transformation de la représentation vectorielle des courriels vers le format « arff »
2. Production du graphe d'induction avec la méthode Sipina et son optimisation.
3. Représentation cellulaire du graphe d'induction : l'ensemble des règles est représenté par CelRule. L'ensemble des prémisses et conclusions, est représenté par Celfact. L'interaction entre CelFact et CelRule est représenté par R_E et R_S
4. Inférence en chainage avant : le modèle cellulaire passe de la configuration G(t) vers la configuration G(t+1) en utilisant les deux fonctions de transition $\delta fact$, $\delta rule$.
5. La dynamique de la machine cellulaire (4) est répétée jusqu'à stabilisation ($G(t+1) = G(t)$)
6. Sauvegarde du modèle booléen généré.

2.3 Classification

Cette étape utilise comme entrée le modèle élaboré depuis la phase d'apprentissage :

1. Charger le modèle booléen : CelFact, CelRule, R_E, et R_S
2. Prétraiter le nouveau courriel et calculer sa représentation vectorielle : soit V.
3. Initialiser la base de faits CelFact :
   Pour chaque terme j dans CelFact faire
   Si terme $_j$ présent dans V **Alors** EF(terme$_j$ =1) ← 1
   Sinon EF(terme$_j$ =0)← 1 **Fin Si**
   Fin Pour
4. Appliquer la fonction de transition globale $\nabla = \delta fact \circ \delta rule$
5. **Si** (EF(class=spam) ==1) **Alors** le courriel est classifié spam
   Sinon (EF(class= legitimate) =1) le courriel est classifié légitime **FinSi.**

Références

Atmani B. et Beldjilali B. (2007). Knowledge Discovery in Database: Induction Graph and Cellular Automaton, Computing and Informatics Journal, 26, 171-197.

Zighed. (2000). Graphe d'induction: Apprentissage et data mining. HERMES, 2000.

Summary

Spam, also known as junk mail quickly became a major problem on the Internet. To address this growing burden of this type of spam, we propose the use of a supervised classification based on Boolean cellular automata to automatically classify incoming emails as spam or legitimate. To evaluate the performance of this new approach, we conduct a series of experiments on the corpus LingSpam.