Vers une méthodologie de modélisation d'une signature unique des profils Web : Module de détection des auteurs des forums web

Sara El manar el bouanani[1], Ismail Kassou[2]

[1] ENSIAS – Université Mohammed V Souissi - Rabat – sara_el_manar@yahoo.fr
[2] ENSIAS – Université Mohammed V Souissi - Rabat – kassou@ensias.ma

Abstract

Our research theme is around the concept of computer profiling. The aim of our work is to propose an approach that will detect a profile web in a unique way. We have to model a print web for each profile. The general idea is to find this signature from text messages in the forums, by analyzing the vocabulary used by each user. We will focus here on the presentation of the module developed, which aims to analyze the messages in the web and combine those written by the same person in the same group.

Résumé

Notre thématique de recherche s'articule autour de la notion du profilage informatique. Notre but est de proposer une approche à même de détecter un profil Web d'une façon unique. Il s'agit en effet de modéliser une empreinte Web pour chaque profil, l'idée générale étant de retrouver cette signature à partir de messages écrits dans les forums, en analysant le vocabulaire employé par chaque internaute. Nous porterons notre attention dans ce document sur la présentation du module développé dans le cadre de notre approche , lequel a pour but d'analyser les traces écrites dans le Web et de regrouper celles écrites par une même personne au sein d'un groupe.

Mots-clés : Profilage, Datamining, Webmining, Clustering, Ontologie, Détection d'auteurs, stylométrie

1. Introduction

Le terme « profilage » ou « profiling » a trouvé son application dans la formation de «profilers», des personnes en mesure d'établir le « profil type » d'un criminel, par le moyen de l'analyse des traces laissées sur une scène de crime (Dinant *et al.*, 2008). Le terme « profil » tire aussi son origine de la sphère artistique (Dinant *et al.*, 2008) où il désigne les « contours, et traits d'un visage vu par un de ses côtés ». Par extension, ce terme a fini par désigner « l'ensemble des traits caractéristiques d'une chose, d'une situation, d'une catégorie de personnes… ». Le profilage informatique est reconnu comme le résultat d'une méthode informatisée faisant appel aux procédés de data Mining sur des entrepôts de données et permettant de situer, avec une marge réduite d'erreur, un individu dans une catégorie particulière, le but étant de formuler des décisions à son égard (Dinant *et al.*, 2008).

Le « profilage informatique » est défini dans (Bourcier., 2001) comme étant le résultat d'une technique informatique qui a pour objet de constituer des profils individuels ou des groupes à partir du traitement des données personnelles et de modèles de comportement. Pour constituer un profil, il faut recueillir un maximum de connaissances le concernant, en exploitant toutes les informations glanées à partir d'échanges sur des réseaux locaux ou sur internet.

Dans notre thématique de recherche, nous nous intéresserons particulièrement au profilage sur Internet. En effet, les internautes laissent des commentaires, expriment des opinions, des prises de position… sur le Web. Ils interviennent sur une thématique donnée et gardent généralement le même point de vue exprimé, d'une page à l'autre. Nous sommes donc en mesure, théoriquement, d'établir leur profil à partir des informations véhiculées par les traces écrites qu'ils nous ont fournies. Nous nous proposons de fonder une approche à même de définir, de manière unique, un profil spécifique, à travers des écrits sur le web. Nos outils pour brosser ces profils d'internautes seront l'analyse de leur style d'écriture et des caractéristiques de leur discours, en vue d'en dégager une empreinte unique !

Ce document présente un module, développé dans le cadre de notre approche, qui vise à détecter les messages Web écrits par le même profil. L'article est structuré comme suit : le chapitre 2 expose quelques études proposées dans le domaine du «profilage informatique». Le chapitre 3 est réservé aux caractéristiques stylométriques qui nous permettent de rapprocher les textes écrits par le même auteur. Le chapitre 4 expose notre problématique de recherche et le chapitre 5 présente le module développé et les résultats obtenus.

2. Profilage informatique

Les études menées sur le « profilage informatique » par le biais des techniques de Web Mining et d'exploration de données ne sont pas liées à un domaine particulier ; plusieurs domaines y ont en effet fait appel : criminologie, e-commerce, éducation…

(Muller., 2000) évoque la notion de « profilage criminel » comme étant le processus consistant à utiliser les informations disponibles au sujet d'un crime et de la scène du crime pour en apprendre assez sur le comportement, la personnalité et les caractéristiques physiques de l'auteur inconnu du crime. Des systèmes ont été développés par les services de police pour détecter les profils criminels. Deux exemples de ces systèmes sont CHARDON élaboré par la brigade criminelle de la préfecture de Paris inspiré du VICAP des policiers américains du FBI pour détecter les tueurs en série (Tourancheau., 2000) et COPLINK développé par des chercheurs de l'Université de l'Arizona en collaboration avec le Tucson et le Phoenix Police Département (Hauck *et al.*, 2002). (Chen *et al.*, 2004) présentent les outils de data Mining utilisés pour recueillir les informations et analyser les volumes croissants de données sur la criminalité afin de détecter des tendances criminelles et (Nath., 2006) propose une approche basée sur l'algorithme K-means pour profiler les criminels.

(Chen *et al.*, 2008 ; Chau. et Wu., 2007) ont mis sur pied, pour les mêmes buts, deux approches basées sur l'exploration des réseaux sociaux , des blogs, du Web… en vue d'étudier le comportement de terroristes potentiels ou de groupes prônant la ségrégation raciale… l'étude permettra de les repérer, d'extraire des informations à leur sujet et de comprendre leurs ramifications…

Une autre application des techniques du profilage informatique est liée à la croissance du nombre d'applications en ligne et du nombre de leurs utilisateurs, d'où la nécessité d'avoir un site Web comprenant les intérêts de ses utilisateurs (Yeh *et al.*, 2009 ; Chou *et al.*, 2010 ; Yang., 2010). A ce propos, (Zhang *et al.*, 2007) décrivent un ensemble d'outils exploitant les réseaux de neurones et l'algorithme SOM (Self-Organizing Map) pour identifier les modèles de navigation des internautes. (Yeh *et al.*, 2009) identifient les clients potentiels d'une librairie en ligne (nsc.gov.tw) grâce à l'exploitation du contenu des pages Web sans utiliser les fichiers logs et l'historique de navigation et (Chou *et al.*, 2010) illustrent une approche Web facilitant la détection des consommateurs potentiels d'un site de vente de produits cosmétiques.

Le profilage des clients se révèle également d'une grande utilité dans le domaine bancaire surtout pour le service des risques. Il aide les décideurs à choisir les groupes de clients potentiels pour l'octroi de crédits ou de services bancaires. (Jay et Raghavendra., 2010) utilisent SOM pour segmenter la clientèle et comprendre son comportement. (Arayaa *et al.*, 2004) proposent une méthodologie combinant l'utilisation de l'algorithme de K-means flou et des réseaux de neurones dans l'étude du comportement des clients de la banque chilienne BCI.

Un dernier volet de l'application du profilage informatique est le E-learning. La possibilité de comprendre et de détecter le comportement des utilisateurs dans ces environnements permet aux architectes de sites d'apprentissage et aux concepteurs d'outils pédagogiques et didactiques de créer et d'organiser les contenus d'apprentissage selon le profil des apprenants (Carbo *et al.*, 2005). Des exemples d'études relatives au E-learning sont relatés dans (Xu *et al.*, 2002 ; Ai et Laffey., 2007 ; Audran et Simonian., 2003 ; Ziani. et Ouinten., 2007).

3. Caractéristiques stylométriques

Il est admis aujourd'hui, l'usage aidant, que les empreintes digitales identifient sans équivoque les êtres humains (Iqbal *et al.*, 2010). Avec le développement de l'informatique, le terme « empreinte » a été adopté pour désigner des caractéristiques d'ordre comportemental de distinction entre les individus (Padmanabhan. et Yang., 2006). A ce sujet, (Miller., 1994) a montré que les individus adoptent un mode de frappe individualisé lorsqu'ils utilisent des claviers d'ordinateur et (Everitt et McOwan., 2003) ont combiné le mode d'utilisation du clavier et les mouvements de la souris pour en identifier l'utilisateur.

Les études de stylométrie ont aussi démontré que les individus peuvent avoir une empreinte liée à leur style d'écriture qu'on nommerait « writeprint » (Iqbal *et al.*, 2010). Le style d'écriture d'un individu est défini en termes d'usage des mots, de sélection de caractères spéciaux, de composition des phrases et des paragraphes, de l'organisation des phrases en paragraphes et des alinéas dans les documents (Iqbal *et al.*, 2010). Les études faites dans le domaine de la stylométrie (Li *et al.*, 2006 ; De Vel O., 2000 ; Zheng *et al.*, 2006 ; Iqbal *et al.*, 2010) ont ainsi défini quatre familles de caractéristiques :

1 – Les Caractéristiques lexicales qui sont utilisées pour en apprendre davantage sur l'utilisation préférentielle de caractères spéciaux et de mots d'un individu. Il s'agit notamment de la fréquence des alphabets différents, du nombre total de lettres majuscules, du nombre moyen de caractères par mot, et du nombre moyen de caractères par phrase.

2 – Les Caractéristiques syntaxiques qui sont appelées marqueurs de style et composées des mots outils tels que « bien », « où », « tant », « votre » et de ponctuation tels que « ! », « : », « ? »

3 – Les Caractéristiques structurelles qui sont utilisées pour connaître la façon dont un individu organise la présentation et la structure de ses textes (paragraphe, alinéa, phrases…)

4 – Les Caractéristiques liées au domaine du texte qui sont utilisées pour regrouper le champ lexical d'un thème donné.

Les différentes caractéristiques sont synthétisées dans le tableau suivant :

Caractéristiques lexicales	**Caractéristiques syntaxiques**
Nombre de caractères	Occurrence de la ponctuation (, . ? ! : ; ' ')
Nombre de caractères numériques	Occurrence des fonctions tels que (bien, où, tant, votre, notre…)
Nombre de caractères alphanumériques	
Nombre de lettres majuscules	**Caractéristiques structurelles**
Nombre des espaces	Nombre de lignes
Nombre des tabulations	Nombre de phrases
Occurrence de chaque lettre de l'alphabet (26)	Nombre de paragraphes
Occurrence des caractères spéciaux (<,>,%,|,{,},[,],\,@,#,+,-,*,$,^,&,~,÷,/)	Moyenne de la longueur des paragraphes en terme de phrases
Nombre total des mots	Moyenne de la longueur des paragraphes en termes de caractères
Moyenne de la longueur des phrases en terme de caractères	Moyenne de la longueur des paragraphes en terme de mots
Moyenne de la longueur des phrases en terme de mots	Max de la longueur de phrases en terme de mots
Moyenne de la longueur des mots en terme de caractères	Min de la longueur de phrases en terme de mots
Moyenne de la longueur des mots en terme de syllabes	Présence/absence d'une formule de politesse
Nombre de mots courts (1 à 3 caractères)	Existence de lignes vides entre les paragraphes
Nombre de mots différents / nombre total des mots (densité lexicale)	
Mesure de Yule	**Caractéristiques liées au domaine du texte**
HAPAX Legomena (mots répétés une fois)	Tous les mots du champ lexical d'une thématique donnée au sein d'un texte
HAPAX dislegomena (mots répétés deux fois)	
HAPAX trislegomena (mots répétés 3 fois)	

Les caractéristiques stylométriques sont utilisées pour déterminer les auteurs potentiels de textes. (Zheng *et al.*, 2006) présentent un Framework pour identifier les auteurs des messages Web en utilisant les quatre familles de caractéristiques. L'expérience a montré que ce Framework détecte les auteurs anonymes de messages Web avec une probabilité de 70 à 90% et que le classificateur SVM donne de meilleurs résultats que les réseaux de neurones ou les arbres de décision. (Farkhund *et al.*, 2010) exposent une approche d'analyse des emails, d'extraction

de leur style d'écriture et de repérage des messages écrits par des criminels ciblés et (De Vel O., 2000) présente une approche de détection des auteurs des mails en utilisant l'algorithme SVM. (Mohtasseb et Ahmed., 2009) recommandent une méthodologie de détection des auteurs de blogs s'appuyant sur les caractéristiques des textes et sur l'exploitation du logiciel LIWC qui introduit la notion de sentiments éprouvés à partir du vocabulaire des auteurs. (Orebaugh et Allnutt., 2009) identifient les auteurs potentiels des messages web instantanés grâce à l'utilisation de 3 classificateurs implémentés dans WEKA (J48, IBk, Naive Bayes).

4. Problématique

La problématique à laquelle nous avons été confrontés peut se résumer de cette manière : comment, à partir de textes écrits dans les forums Web et du langage spécifique utilisé par une personne , pouvons-nous identifier ce profil de façon unique ? Comment pouvons-nous modéliser une signature unique, à partir de ces traces, pour déterminer un profil ?

L'idée générale de notre approche est de recueillir le maximum d'informations sur les utilisateurs et d'employer les techniques de Web Mining, de Text Mining et de détection d'auteurs… pour extraire les informations pertinentes et repérer notre profil cible. Notre contribution tentera de combiner ces méthodes en vue de modéliser la « signature du profil ».

La figure 1 décrit notre approche de modélisation de signature, articulée autour de 4 modules : extraction d'information et ontologie du discours, détection d'auteurs, modèle utilisateur et signature du profil.

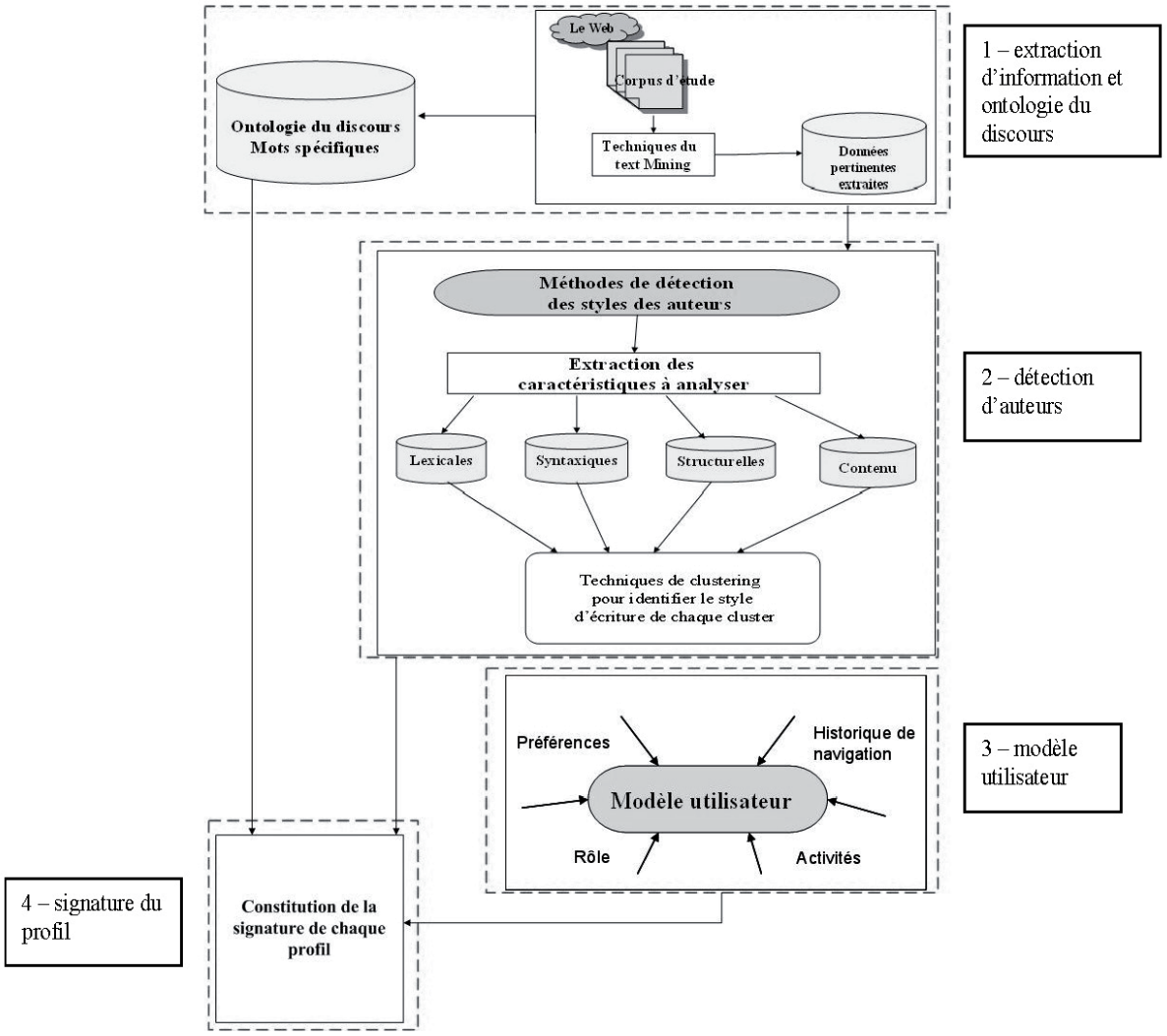

Figure 1 : schéma général de l'approche

Au Module 1, *Extraction d'informations et ontologie du discours*, nous travaillerons à construire un univers du discours relatif à un profil donné. En effet, les internautes écrivent des commentaires, expriment des opinions et des idées sur le Web. Ils interviennent sur une thématique donnée et ont généralement le même point de vue exprimé d'une page à l'autre. Les traces écrites contiennent donc des connaissances qui nous conduisent à avoir toutes les données nécessaires pour l'analyse du profil d'un individu.

Ce premier module va user de techniques de fouille de textes afin d'extraire des pages Web, les informations pertinentes. Nous procéderons dès lors à la création d'une ontologie de discours qui comprend tous les termes et mots spécifiques utilisés par un individu. Nous devons, par conséquent, à partir de ce premier module, être en mesure de regrouper tout le vocabulaire employé par chaque profil.

Le Module 2, *Détection d'auteurs*, articulé autour des travaux initiés en stylométrie, nous permettra de rassembler dans des clusters les textes écrits par la même personne. L'objectif sera

donc de retrouver, à partir de divers textes extraits des forums, ceux rédigés par le même profil. Ce module étant l'objet de notre article, la description en sera détaillée dans le chapitre suivant.

Le Module 3, *Modèle Utilisateur*, permettra de créer un modèle utilisateur regroupant toutes les informations que nous pouvons extraire ou sauvegarder sur cette même personne, lesquelles peuvent concerner son identité, bien que les internautes optent le plus souvent pour un nom virtuel ou un pseudo, les pages qu'il a visitées, les transactions qu'il a traitées via Internet, les préférences qu'il a manifestées…

Il convient de signaler, à ce propos, que l'étude que nous allons mener, sera effectuée sur une période déterminée ; nous intégrerons donc la notion de « période » à notre modèle.

Le dernier module, *Signature du profil* sera pour nous, une tentative de contribution individuelle à la technique des recherches de profils. Nous avons estimé pratique et pertinent de combiner toutes les informations fournies par chaque module pour modéliser une signature unique à chaque utilisateur

5. Module de détection des auteurs et résultats expérimentaux

Le module de détection des auteurs aura pour finalité de rapprocher les textes écrits par le même auteur et de les rassembler au sein d'un cluster. En effet, soit N textes extraits du Web de M auteurs, l'outil développé aura à dégager les M clusters regroupant avec une certaine probabilité les textes rédigés par la même personne. Nous notons que nous ne connaissons pas d'avance, le nombre d'auteurs (nombre de clusters). En nous basant sur les études faites dans le domaine de la stylométrie et de la détection des auteurs (Farkhund *et al.*, 2010 ; Li *et al.*, 2006 ; Zheng *et al.*, 2006 ; Iqbal *et al.*, 2010 ; De Vel O., 2000), nous avons développé un prototype expérimental en JAVA prenant en entrée les caractéristiques stylométriques à étudier , la collection de textes à analyser et générant un fichier .arff que nous utilisons en entrée à l'outil de Mining WEKA. Les deux approches adoptées pour le clustering sont les deux algorithmes k-means et EM.

Pour que nous puissions calculer la performance du prototype développé et évaluer les résultats, nous avons fait appel aux deux paramètres *Rappel* et *Précision* issus de la statistique tels qu'ils ont été exploités dans (Iqbal *et al.*, 2010 ; Anderson *et al.* 2001). La combinaison de ces deux paramètres calcule la moyenne harmonique (noté *F-mesure*), laquelle nous instruit sur la performance du système.

P(i,j) = nombre de textes de l'auteur i présents dans le cluster j / nombre total de textes de l'auteur i

R(i,j) = nombre de textes de l'auteur i présents dans le cluster j / nombre total de textes du cluster i

$$F(i,j) = (2 * P(i,j) * R(i,j)) / (P(i,j) + R(i,j))$$

5.1. Présentation des algorithmes utilisés : K-means et EM

5.1.1. Algorithme K-means

L'algorithme *k-means* mis au point par McQueen en 1967 est l'un des algorithmes de clustering les plus connus. *K-means* est un algorithme de minimisation alternée qui, étant donné un entier K, va chercher à séparer un ensemble de points en K clusters. Il est basé sur la méthode des centroïdes. Le principe de cette méthode est le suivant : nous nous donnons pour commencer, *k* centres arbitraires *c1, c2,..., ck* où chaque *ci* représente le centre d'une classe Ci. Chaque classe Ci est représentée par un ensemble d'individus plus proches de *ci* que de tout autre centre. Après cette initialisation, nous effectuons une deuxième partition en regroupant les individus autour des mj qui prennent alors la place des *cj* (*mj* est le centre de gravité de la classe Cj, calculé en utilisant les nouvelles classes obtenues). Le processus est ainsi réitéré jusqu'à l'atteinte d'un état de stabilité où aucune amélioration n'est possible.

5.1.2. Algorithme EM (Expectation Maximisation)

L'algorithme EM ou « Espérance Maximisation » proposé par Dempster en 1977, est une classe d'algorithmes qui permettent de trouver le maximum de vraisemblance des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables. EM est souvent utilisé pour la classification de données, en apprentissage machine, ou en vision artificielle. EM alterne des étapes d'évaluation de l'espérance (E), où est calculée l'espérance de la vraisemblance en tenant compte des dernières variables observées, et une étape de maximisation (M), où est estimé le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E. Nous utilisons ensuite les paramètres trouvés en M comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi.

5.2. Expérimentation du prototype sur une collection d'articles de journaux

Nous avons utilisé toutes les caractéristiques définies au chapitre 3 et avons fait le test sur une collection d'articles extraits de journaux. Nous notons que les caractéristiques liées au domaine n'ont pas été utilisées. Il a été démontré dans (Iqbal *et al.*, 2010) que l'ajout de ces caractéristiques n'augmente pas vraiment la performance du système. Il a été aussi démontré dans (Iqbal *et al.*, 2010) que la combinaison des 3 types de caractéristiques syntaxiques, lexicales et structurelles donne un résultat meilleur.

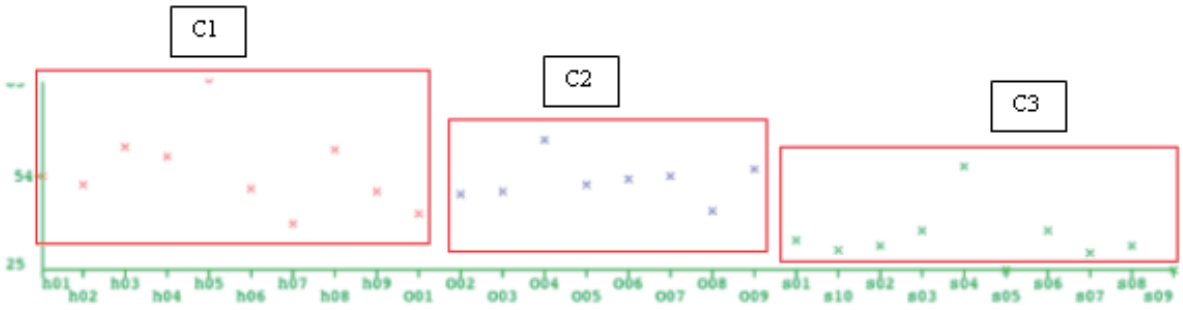

Figure 2 : Clustering d'une collection d'articles de journaux avec l'algorithme EM (Cas de trois auteurs avec un nombre de neuf textes par auteur)

Nous pouvons relever aisément les 3 clusters renvoyés par l'algorithme EM dans la figure 2. Nous remarquons également que seul le texte « 001 » a été mal classé ; il a été regroupé avec les textes de l'auteur « h ». Le calcul de la F-mesure pour cette première expérience donne une valeur de 0,997.

Une fois que nous avons appliqué l'algorithme de k-means à la même collection de textes, nous avons obtenu les clusters suivants :

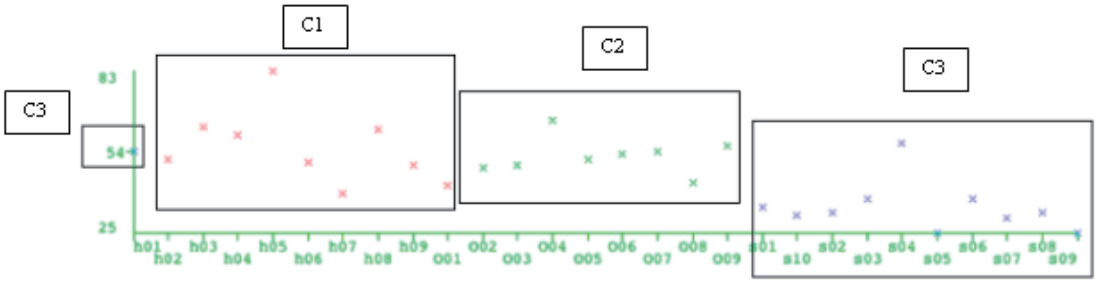

Figure 3 : Clustering d'une collection d'articles de journaux avec l'algorithme K-means (Cas de trois auteurs avec un nombre de neuf textes par auteur)

Nous signalons que pour l'algorithme de k-means, nous avons introduit le nombre de clusters que nous souhaitions obtenir. Nous remarquons, d'après la figure 3, que nous avons deux textes mal classés : le texte « h01 » a été regroupé avec les textes de l'auteur « s » et le texte « 001 » a été regroupé avec les textes de l'auteur « h ». La valeur de la F-mesure pour ce deuxième algorithme est de 0,88.

Nous pouvons conclure, à partir de ces premiers tests, que les deux mesures obtenues sont satisfaisantes ; néanmoins, pour l'algorithme K-means, nous avons introduit le nombre de clusters à obtenir contrairement à l'algorithme EM, lequel a calculé le nombre de clustering optimal sans notre intervention. Nous en déduisons que l'algorithme EM est plus performant que k-means.

D'autre part , et à dessein de mesurer la performance de notre prototype et de choisir l'algorithme de clustering optimal de l'approche de détection des auteurs, nous avons effectué deux autres expériences ; nous avons choisi pour la première, 20 textes par auteur, à raison 3 auteurs . Pour la deuxième, nous avons 20 textes pour l'auteur 1, 30 textes pour l'auteur 2 et 30 textes pour l'auteur 3.

Les F-mesure calculées pour cette deuxième expérience sont les suivantes :

	EM	**K-means**
3 auteurs / 20 textes par auteur	0,8	0,9
3 auteurs / 20 textes pour l'auteur 1 ; 30 pour l'auteur 2 et 30 pour l'auteur 3	0,99	0,96

Nous remarquons pour le premier cas que les résultats fournis par K-means sont meilleurs que ceux fournis par EM. Pour le deuxième cas, EM dépasse K-means. Toutefois, comme déjà mentionné, nous considérons que les résultats fournis par EM sont plus performants car nous n'avons pas besoin de lui spécifier le nombre de clusters.

5.3. Expérimentation du prototype sur une collection de messages des forums Web

Nous avons mené les tests qui suivent sur un ensemble d'écrits extraits du forum *« forum.sports.fr »* dans le but de détecter les messages écrits par le même profil. Notre prototype va extraire les caractéristiques stylométriques de chaque texte pour repérer les messages correspondant à un même profil.

Dans un premier temps, nous avons utilisé les mêmes caractéristiques que celles employées pour le cas des articles de journaux sur une collection de 126 textes écrits par 5 auteurs (le nombre de textes n'est pas le même pour tous les auteurs).

Nous avons obtenu les résultats ci-après :

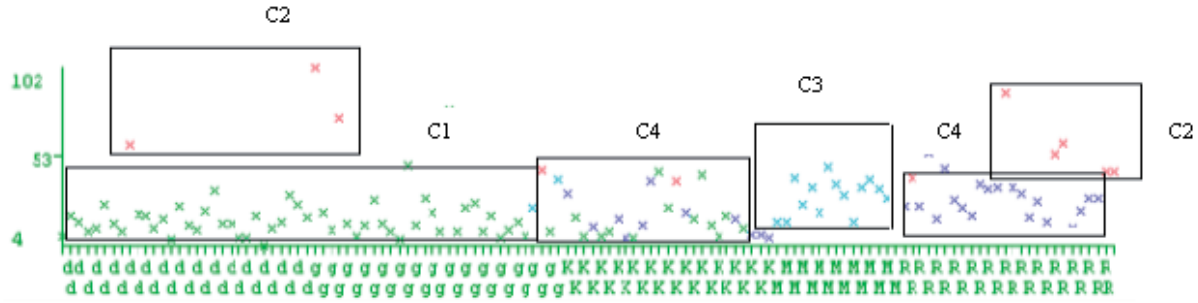

Figure 4 : Clustering d'une collection de 126 textes de forums avec l'algorithme EM (5 auteurs)

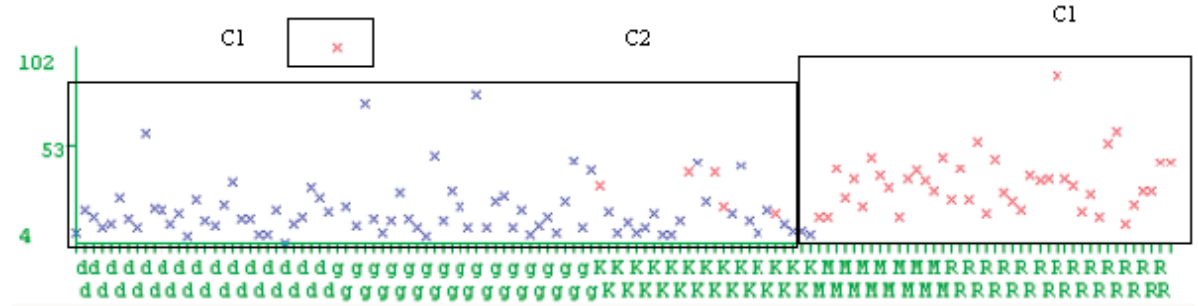

Figure 5 : Clustering d'une collection de 126 textes de forums avec K-means (5 auteurs)

Nous remarquons d'après les figures 4 et 5 que les clusters n'ont pas été correctement détectés pour les deux algorithmes. A titre d'exemple, l'algorithme EM n'a généré que 4 clusters, k-means n'en a généré que 2 alors que nous avons 5 auteurs, les textes des auteurs « d », « g » et « k » font partie du même cluster. Les F-mesure obtenues à l'issue de cette expérience sont 0,4 pour EM et 0,35 pour k-means. Nous avons entrepris d'autres tests sur d'autres textes issus d'autres forums et n'avons obtenu qu'une F-mesure variant entre 0,22 et 0,58.

Il s'avère donc que les caractéristiques que nous avons adoptées ne s'appliquent pas aux textes Web. Nous avons estimé que nous devons introduire d'autres caractéristiques qui prennent en considération la particularité des messages Web.

Les textes saisis dans les forums ne sont pas des textes structurés et n'ont pas de taille définie contrairement aux articles de journaux. Ils ont aussi la particularité d'être écrits dans un langage informel et simplifié, de contenir des phrases courtes, des mots écrits incorrectement, d'utiliser un grand nombre d'abréviations et de termes spécifiques aux internautes. Certains messages peuvent également contenir des images ou des pièces jointes.

(Anderson *et al.* 2001) proposent d'utiliser les ratios au lieu d'utiliser les nombres dans certaines caractéristiques. Des exemples sont donnés ci-dessous :

Ratio des lignes vides Vs nombre de lignes vides
Ratio des mots courts Vs nombre de mots courts
Ratio des HAPAX Legomena Vs HAPAX Legomena
Ratio des caractères numériques Vs nombre de caractères numériques
Ratio des espaces Vs nombre d'espaces

D'autre part, nous avons pu détecter certains mots fréquents utilisés par les internautes que nous avons rajoutés aux caractéristiques.

Pr, fdp, ke, lol, vs, qd, mag, ds, coucou, pti, bin, hello, bonjour, ☺, ☺, ☹, :-P, :P, :-D…

Nous avons procédé à l'utilisation des ratios au lieu des nombres et avons introduit les mots fréquents utilisés sur le Web aux caractéristiques étudiées. En rajoutant ces nouveaux critères, nous avons pu obtenir les résultats suivants, pour un nombre d'auteurs égal à 3.

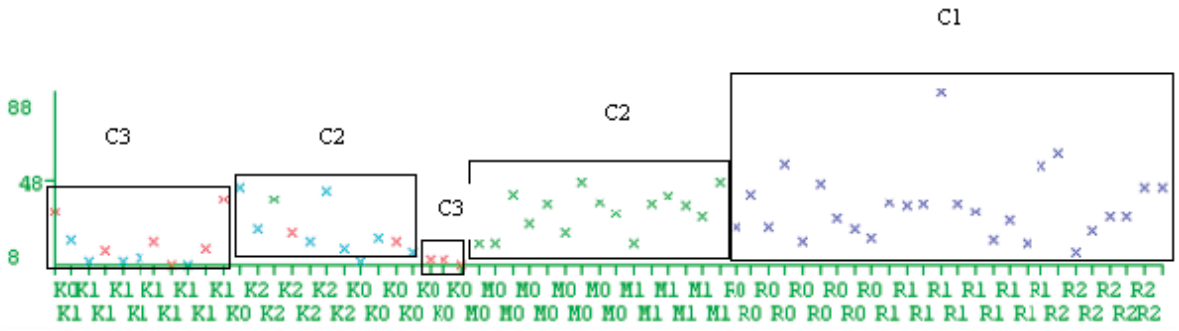

Figure 6 : Clustering d'une collection de 66 textes de forums avec l'algorithme EM (3 auteurs)

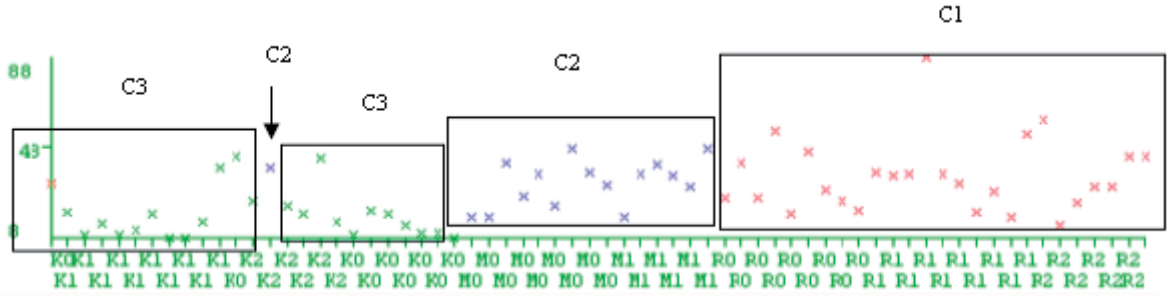

Figure 7 : Clustering d'une collection de 66 textes de forums avec l'algorithme K-means (3 auteurs)

Nous pouvons distinguer clairement les clusters détectés dans les figures 6 et 7. La F-mesure obtenue dans ce cas est de 0.827 pour EM et 0.978 pour K-means. La performance du système a donc augmenté considérablement après l'introduction des caractéristiques spécifiques aux messages web. Nous avons effectué d'autres expériences sur d'autres échantillons des textes des forums et avons obtenu les résultats décrits dans la figure ci-dessous :

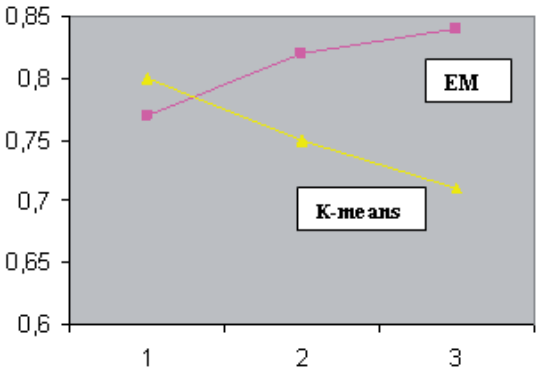

Figure 8 : F-mesures obtenues après clustering des textes de forums
(Cas de 3 auteurs avec une collection de 10 puis de 15 puis de 20 textes par auteur)

Les résultats démontrent que la F-mesure varie de 0,78 à 0,84 pour EM et de 0,71 à 0,8 pour K-means. Nous remarquons également que plus le nombre de textes augmente, plus la F-mesure augmente pour EM et diminue pour K-means. Toutefois, nous considérons, comme précédemment, que c'est l'algorithme EM qui donne de meilleur résultat et considérons donc que pour avoir un bon clustering des textes, il faudra rassembler le maximum de textes pour chaque profil.

→ Les caractéristiques analysées dans ce cas de figure ont donné de bons résultats et seront retenues pour implémenter notre prototype de détection d'auteurs au sein de notre approche générale.

6. Conclusion

Dans ce travail, nous avons proposé une démarche de détection d'une signature unique de profil web. En premier lieu, nous avons développé un prototype expérimental regroupant les

textes d'un auteur au sein du même cluster. Le module ainsi développé est l'un des composants majeurs de notre approche. Il nous fournira une première classification des profils, classification à laquelle nous allons devoir rajouter d'autres paramètres liés au discours pour modéliser la signature unique de ce profil.

Dans la suite de nos travaux de recherche, nous allons utiliser les caractéristiques qui ont donné les meilleurs résultats et opter pour l'algorithme EM comme algorithme de clustering. Nous développerons également les autres modules de détection de signature de profils Web, objet de notre recherche.

Références

Anderson A., Corney M. et Mohay G. (2001). Mining E-mail Content for Author Identification Forensics. *ACM SIGMOD Record,* Vol.(30), Issue 4.

Ai J. et Laffey J. (2007). Web Mining as a Tool for Understanding Online Learning. *MERLOT Journal of Online Learning and Teaching*, Vol.(3), No. 2.

Arayaa S., Silvab M. et Weberc R. (2004). A methodology for web usage mining and its application to target group identification. *Fuzzy Sets and Systems,* Vol(148) :139–152.

Audran J. et Simonian S. (2003). Profiler les apprenants à travers l'usage du forum. *ISDM N°10 – Spécial Colloque TICE.*

Bourcier D. (2001). De l'intelligence artificielle à la personne virtuelle : émergence d'une entité juidique ?. *Droit et société,* Vol.(49): 847-871.

Carbo J.M., Mor E. et Minguillon J. (2005). User navigational behavior in e-learning virtual environments. *IEEE/WIC/ACM International Conference on Web Intelligence.*

Chau M. et Wu J. (2007). Mining communities and their relationships in blogs: a study of online hate group. *Int. J. Human-Computer Studies,* pp.57-70.

Chen H., Chung W., Xu J.J., Wang G., Qin Y. et Chau M. (2004). Crime Data Mining : A General Framework and Some Examples. *Journal Computer*. Vol.(37), Issue 4.

Chen H., Chung W., Qin J., Reid E., Sageman M. et Weimann G. (2008). Uncovering the Dark Web: A Case Study of Jihad on the Web. *Journal of the American Society for Information Science and Technology,* Vol.(59), Issue 8, pp: 1347–1359.

Chou P.H., Li P.H., Chen K.K. et Wu M.J. (2010). Integrating web mining and neural network for personalized e-commerce automatic service. *Expert System with applications,* Vol.(37): 2898-2910.

De Vel O. (2000).Mining e-mail authorship. *In: Proc. of the Workshop on text mining in ACM international conference on knowledge discovery and data mining (KDD).*

Dinant J.M., Lazaro C., Poullet Y., Lefever N. et Rouvroy A. (2008). L'application de la Convention 108 au mécanisme de profilage. Eléments de réflexion destinés au travail futur du Comité consultatif (T-PD). *Comité consultatif de la Convention pour la protection des personnes à l'égard du traitement Automatisé des données à caractère personnel.* 24ème réunion. Strasbourg.

Everitt R.A.J. et McOwan P.W. (2003). Java-Based Internet Biometric Authentication System. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.(25), Issue 9.

Farkhund I., Binsalleeh H., Fung B.C.M. et Debbabi M. (2010). Mining writeprints from anonymous e-mails for forensic investigation. *digital investigation,* Vol.(7): 56-64.

Hauck R.V., Atabakhsh H., Ongvasith P., Gupta H. et Chen H. (2002). Using Coplink to Analyze Criminal-Justice Data. University of Arizona.

Iqbal F, *et al.* (2010). Mining writeprints from anonymous e-mails for forensic investigation. *Digit. Investig, doi:10.1016/j.diin.2010.03.003.*

Jay B.S. et Raghavendra B.K. (2010) A Neural Network based framework for Customer Profiling for Risk analysis. *International Journal of Advanced Computing (IJAC).* Vol.(2), Issue 4.

Li J., Zheng R. et Chen H. (2006). From fingerprint to writeprint. *Communications of the ACM - Supporting exploratory search.* Vol.(49), Issue 4, pp: 76-82.

Miller B. (1994). Vital signs of identity. *IEEE Spectru.* Vol.(31) : 22-30.

Mohtasseb H. et Ahmed A. (2009). Mining Online Diaries for Blogger Identification. *Proceedings of the World Congress on Engineering* (*WCE).* London, U.K.

Muller D.A. (2000). Criminal profiling: Real science or just wishful thinking. *Homicide Studie*s, Vol.(4): 234-264, ISNN 1552-6720 (Online) 1088-7679 (Print).

Nath S.V. (2006). Crime Pattern Detection Using Data Mining. *Web Intelligence and Intelligent Agent Technology Workshops*.

Orebaugh A. et Allnutt J. (2009). Classification of Instant Messaging Communications for Forensics Analysis. *the International Journal of Forensic Computer Science*, Vol.(1): 22-28.

Padmanabhan B. et Yang Y. (2006). Clickprints on the Web: Are there signatures in Web browsing data, published *in knowledge@wharton.*

Tourancheau P. (2000). Projet Chardon : Un fichier contre les tueurs en série. *Journal Libération.*

Xu D., Wang H. et SuK. (2002). Intelligent Student Profiling with Fuzzy Models. *Proceedings of the 35th Hawaii International Conference on System Sciences.*

Yang Y.C. (2010). Web user behavioral profiling for user identification. *Decision Support Systems*, Vol. (49): 261–271.

Yeh I.C., Lien C.h., Ting T.M. et Liu C.H. (2009). Applications of web mining for marketing of online bookstore. *Expert System with applications*, Vol.(36) :11249-11256.

Ziani B. et Ouinten Y. (2007). Etude de cas en Web Usage Mining: Catégorisation des utilisateurs de la connexion Internet de l'UATL. *RIST,* Vol.(17).

Zhang X., Edwards J. et Harding J. (2007). Personalised online sales using web usage data mining. *Computers in Industry,* Vol.(58): 772–782.

Zheng R., Li J., Chen H. et Huang Z. (2006).A framework for authorship Identification of Online Messages: writing-Style features and classification Techniques. *Journal of The American Society For Information Science And Technology,* pp: 378-393.