Profilage de candidatures assisté par *Relevance Feedback*

Rémy Kessler[1] Nicolas Béchet[2] Juan Manuel Torres-Moreno[1] Mathieu Roche[2] Marc El-Bèze[1]

(1) LIA / Université d'Avignon, 339 chemin des Meinajariès, 84911 Avignon

(2) LIRMM - UMR 5506, CNRS - Université Montpellier 2 - France

{remy.kessler, juan-manuel.torres, marc.elbeze}@univ-avignon.fr

{nicolas.bechet,mathieu.roche}@lirmm.fr

Résumé. Le marché d'offres d'emploi et des candidatures sur Internet a eu une croissance exponentielle. Ceci implique des volumes d'information (majoritairement sous la forme de texte libre) intraitables manuellement. Une analyse et catégorisation assistés nous semble pertinente pour répondre à cette problématique. Nous proposons E-Gen, système qui a pour but l'analyse et catégorisation assistés d'offres d'emploi et des réponses des candidats. Dans cet article nous présentons plusieurs stratégies, reposant sur les modèles vectoriel et probabiliste, afin de résoudre la problématique du profilage des candidatures en fonction d'une offre précise. Nous avons évalué une palette de mesures de similarité afin d'effectuer un classement pertinent des candidatures au moyen des courbes ROC. L'utilisation de *relevance feedback* a permis de surpasser nos résultats sur ce problème difficile, divers et sujet à une grande subjectivité.

Abstract. The market of online job search sites has grown exponentially. This implies volumes of information (mostly in the form of free text) manually impossible to process. An analysis and assisted categorization seems relevant to address this issue. We present E-Gen, a system which aims to perform assisted analysis and categorization of job offers and the responses of candidates. This paper presents several strategies based on vectorial and probabilistic models to solve the problem of profiling applications according to a specific job offer. We have evaluated a range of measures of similarity to rank candidatures by using ROC curves. Relevance feedback approach allows surpass our previous results on this task, difficult, diverse and higly subjective.

Mots-clés : Classification, recherche d'information, Ressources humaines, modèle probabiliste, mesures de similarité, *Relevance Feedback*.

Keywords: Classification,Information Retrieval, Human Ressources, Probabilistic Model, Similarity measure, *Relevance Feedback* .

1 Introduction

La croissance exponentielle d'Internet a permis le développement d'un grand nombre de sites d'emplois (Bizer & Rainer, 2005; Rafter *et al.*, 2000a; Rafter & Smyth, 2000) et d'un marché du recrutement en ligne en expansion significative (août 2003 : 177 000 offres, mai 2008 : 500 000 offres)[1]. Internet est devenu essentiel dans ce processus, car il permet une meilleure diffusion

[1] Site d'emploi www.keljob.com

de l'information, que ce soit par les sites de recherche d'emplois ou par les réponses à ceux-ci par courriels. Cependant ce phénomène pose divers problèmes dans leur traitement dû à la grande quantité d'information difficile à gérer rapidement et efficacement pour les entreprises (Bourse *et al.*, 2004; Morin *et al.*, 2004; Rafter *et al.*, 2000b). En outre, si le navigateur Web est devenu un outil universel, facile à employer pour les utilisateurs, la nécessité fréquente pour les internautes d'entrer des données dans les formulaires Web à partir de sources papier, de "copier et coller" de données entre différentes applications, est symptomatique des problèmes d'intégration de données communes. En conséquence, il est nécessaire de traiter cette masse de documents d'une manière automatique ou assistée. Nous proposons le système E-Gen pour résoudre ce problème. Il est composé de trois modules principaux :

– Un module d'extraction de l'information à partir de corpus de courriels provenant d'offres d'emplois extraites de la base de données d'Aktor[2].
– Un module pour analyser les e-mails de réponse de candidats (afin de distinguer lettre de motivation, abrégé LM par la suite et curriculum vitae, abrégé CV par la suite).
– Un module pour analyser et calculer un classement de pertinence du profil du candidat (LM et CV).

Nos précédents travaux présentaient le premier module (Kessler *et al.*, 2007), l'identification des parties d'une offre d'emploi et l'extraction d'informations pertinentes (contrat, salaire, localisation, etc.). Le deuxième module permet, à l'aide d'une solution combinant règles et méthodes d'apprentissage (Machines à Support Vectoriel SVM), de distinguer correctement les parties de la candidature (CV ou LM) avec une précision de 0,98 et un rappel de 0,96 (Kessler *et al.*, 2008b). Cependant le flux de réponses à une offre d'emploi entraîne un long travail de lecture des candidatures par les recruteurs. Afin de faciliter cette tâche, nous souhaitons mettre en place un système capable de fournir une première évaluation automatisée des candidatures selon divers critères. Nous présentons ici les travaux concernant le dernier module du système E-Gen. La section 2 présente un bref état de l'art sur le traitement automatique des documents issus du domaine des ressources humaines. La section 3 décrit l'architecture globale du système. Nous présentons en section 4 et 5 les stratégies pour le profilage des candidatures, le protocole expérimental, les statistiques sur le corpus et les résultats obtenus, avant de conclure en section 6.

2 Etat de l'art

Dans le but d'automatiser certains processus souvent longs et coûteux liés à la gestion des ressources humaines, divers travaux ont été menés. La spécificité des informations contenues dans les documents d'une candidature à une offre d'emploi a permis le développement de différentes approches sémantiques. (Desmontils *et al.*, 2002; Morin *et al.*, 2004) proposent une méthode d'indexation sémantique de CV fondée sur le système BONOM (Cazalens & Lamarre, 2001). La méthode proposée consiste à exploiter les caractéristiques dispositionnelles du document afin d'identifier chacune des parties et l'indexer en conséquence. Par ailleurs, une description d'une approche sémantique du processus de recrutements et des différents impacts économiques est proposée par (Bizer & Rainer, 2005; Tolksdorf *et al.*, 2006) au sein de gouvernement allemand. (Rafter *et al.*, 2000a) décrivent les lacunes des systèmes actuels face à la problématique de recherche d'emploi et proposent un système sur la base de filtre collaboratif (ACF) permettant d'effectuer des profilages automatiques sur le site JobFinder. Mochol (Mocho *et al.*, 2006) dé-

[2] Aktor Intéractive (www.aktor.fr)

crit l'importance d'une ontologie commune (*HR ontology*) afin de pouvoir traiter efficacement ce type de documents et (Bourse *et al.*, 2004) décrit un modèle de compétence et un processus dédié à la gestion des compétences dans le cadre de l'e-recrutement (principalement des CV ou des offres d'emploi). De la même façon, fondé sur la technologie HR-XML mise en place par (Allen & Pilot, 2001), (Dorn *et al.*, 2007; Dorn & Naz, 2007) décrit un prototype de méta-moteur spécifique à la recherche d'emploi. Celui-ci privilégie la récolte des informations importantes (catégorie de l'emploi, lieu du travail, compétences recherchées, intervalle de salaire etc.) sur un ensemble de sites web (Jobs.net, aftercollege.com, Directjobs.com etc.).
L'étude du document principal d'une candidature, le CV, a fait l'objet de différents travaux afin de l'analyser de manière automatique. (Clech & Zighed, 2003) décrit une approche de fouille de données ayant pour but la mise au point d'automates capables d'apprendre à identifier des typologies de CV, de profils de candidats et/ou de postes. Les travaux présentent une première approche limitée à la catégorisation de CV de cadres et de CV généraux. La méthode mise en œuvre s'appuie sur sur l'extraction de termes spécifiques permettant une catégorisation à l'aide de C4.5 (Quilan, 1993) et un modèle à base d'analyse discriminante. Les résultats obtenus mettent en évidence la spécificité de certains termes ou concepts (tel que le niveau d'étude, les compétences mises en avant) afin d'effectuer cette classification mais reste décevant au niveau de la validation (environ 50- 60% de CV correctement classés). (Roche & Kodratoff, 2006; Roche & Prince, 2008) décrivent une étude réalisée sur l'extraction de terminologie spécifique sur un corpus de CV[3]. Celle-ci permet d'extraire un certain nombre de collocations contenues dans les CV sur la base de motifs (tels que *Nom-Nom*, *Adjectif-Nom*, *Nom-préposition-Nom*,*etc.*) et de les classer en fonction de leur pertinence en vue de la construction d'une ontologie. Notre approche diffère des différentes méthodes proposées puisqu'elle s'appuie sur une combinaison de mesures de similarité avec une phase de *Relevance Feedback* comme nous le décrirons par la suite.

3 Vue d'ensemble du système

Nous avons choisi de développer un système répondant aussi rapidement et judicieusement que possible aux besoins d'Aktor, et donc aux contraintes du marché de recrutement en ligne. Dans ce but, une adresse électronique a été créée afin de recevoir les offres d'emploi. Après l'identification de la langue par n-grammes de caractères, E-Gen analyse le message afin d'extraire le texte pertinent de l'offre d'emploi du message ou du fichier attaché. Deux modules externes sont utilisés, *wvWare*[4] et *pdftotext*[5] produisent une version texte de l'annonce découpée en segments. Après l'étape de filtrage et racinisation, nous utilisons la représentation vectorielle pour chaque segment afin de lui attribuer une étiquette en fonction de son rôle dans le texte à l'aide des SVM ou des n-grammes de mots. Cette séquence d'étiquettes, qui donne une représentation de l'enchaînement des différents segments de l'annonce, est traitée par un processus correctif qui la valide ou qui propose une meilleure séquence (tâche 1) (Kessler *et al.*, 2007). Lors de la publication d'une offre d'emploi, Aktor génère une adresse électronique afin de répondre a cette offre. Chaque courriel est ainsi redirigé vers un logiciel de ressources humaines, Gestmax[6] afin d'être lu par un consultant en recrutement. Lors de la réception d'une candidature, le sys-

[3] corpus fournis par la société Vedior Bis (http ://www.vediorbis.com)

[4] `http://wvware.sourceforge.net`. La segmentation de textes MS-Word étant un vrai casse tête, on a opté par un outil existant. Dans la majorité des cas, il sectionne en paragraphes le document.

[5] http ://www.bluem.net/downloads/pdftotext_en

[6] http ://www.gestmax.fr

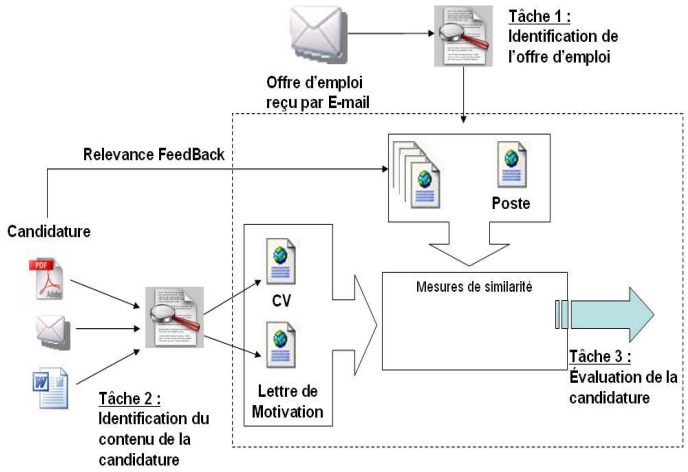

FIG. 1 – Vue d'ensemble du système.

tème extrait le corps du message, ainsi que les différentes pièces jointes. Une version texte des différents documents contenus dans la candidature est alors produite. Différents processus de filtrage et racinisation permettent au système d'identifier à l'aide de SVM et de règles le type du document (CV et/ou LM présente dans le corps du mail ou dans les pièces jointes, pour plus de détails, (Kessler *et al.*, 2008b)) (tâche 2). Une fois le CV et la LM identifiés, le système effectue un profilage automatisé de cette candidature à l'aide de mesures de similarité en s'appuyant sur un petit nombres de candidatures préalablement validé comme candidatures pertinentes par un consultant en recrutement (tâche 3). La figure 1 présente une vue d'ensemble du système.

4 Approches de classement des candidatures

4.1 Prétraitements

Dans un premier temps, nous avons effectué un pré-traitement des données textuelles (CV et LM) permettant de supprimer des informations non pertinentes tels que le nom des candidats, la suppression des adresses, courriers électroniques, noms de villes. La suppression des accents et des majuscules est également effectuée. Nous avons par la suite utilisé différents processus de filtrage et racinisation sur la représentation en mots afin de réduire le lexique. Pour réduire le bruit dans le modèle à base de mots[7], nous avons supprimé les mots fonctionnels (*être, avoir, pouvoir, falloir* ...), les expressions courantes (*par exemple, c'est-à-dire, chacun de* ...), les chiffres et nombres (numériques et/ou textuelles) et les symboles comme <$>, <#>,<*>, ainsi que l'anti-dictionnaire de Jean Veronis [8]. Enfin, nous avons ramené les verbes fléchis à leur racine et les mots pluriels et/ou féminins au masculin singulier.

4.2 Comparaison Candidature/Offre d'emploi par mesure de similarité

Nous avons transformé chaque document en une représentation vectorielle (Salton, 1991) avec des poids représentant la fréquence des termes (*Tf*) et le *Tf-idf* (Salton & Mcgill, 1986). Nous

[7] ces pré traitements ne sont pas appliqués lors de la représentation en n-grammes décrites en 4.3.2
[8] http ://www.up.univ-mrs.fr/ veronis/data/antidico.txt

avons mis en place une approche par mesure de similarité, afin de pouvoir ordonnancer automatiquement l'ensemble des candidatures par rapport aux offres d'emploi proposées. Nous avons combiné pour cela un certain nombre de mesures de similarité entre les candidatures (CV et LM) et l'offre d'emploi associée. Les mesures de similarité que nous avons utilisées dans nos travaux sont décrites dans (Bernstein *et al.*, 2005) : cosinus (formule 1), qui permet de calculer l'angle entre l'offre d'emploi et la réponse de chaque candidat, les distances de Minkowski (formule 2) ($p = 1$ pour Manhattan, $p = 2$ pour la distance euclidienne). Une autre mesure testée est Okabis (formule 3) (Bellot & El-Bèze, 2001), fondée sur la formule Okapi (Robertson *et al.*, 1994), souvent utilisée en Recherche d'Information :

$$sim_{cosine}(j,d) = \frac{j_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} |j_i|^2 \cdot \sum_{i=1}^{n} |d_i|^2}} \quad (1)$$

$$sim_{Minkowski}(j,d) = \frac{1}{1 + (\sum_{i=1}^{n} |j_i - d_i|^p)^{\frac{1}{p}}} \quad (2)$$

Avec j l'offre d'emploi, d la candidature, i un terme.

$$Okabis(j,d) = \sum_{i \in d \cap j} \frac{TF_{i,d}}{TF_{i,d} + \frac{\sqrt{|d|}}{M_S}} \quad (3)$$

Avec j une offre d'emploi, d la candidature, i un terme, $TF_{i,d}$ le nombre d'occurrence de i, N le nombre total de réponses à l'offre et M_S leur taille moyenne. D'autres mesures (Overlap, Enertex, Needleman-Wunsch, Jaro-Winkler) ont été expérimentées, mais n'ont pas été prises en comptes suite aux résultats obtenus. Afin de combiner ces mesures, nous utilisons un algorithme de décision (AD) (Boudin & Torres Moreno, 2007) qui pondère les valeurs λ obtenues par chaque mesure de similarité. Deux moyennes différentes sont calculées : la décision est calculée à partir de ces moyennes, la tendance positive (lorsque la mesure obtient un résultat $\lambda > 0.5$) et la tendance négative ($\lambda < 0.5$).

4.3 Extraction des descripteurs

Dans les sections suivantes, nous décrivons un certain nombre de descripteurs qui seront utilisés pour représenter les documents. Ces descripteurs s'appuient sur des informations grammaticales (section 4.3.1), des informations "lexicales" fondées sur les n-grammes de caractères (section 4.3.2) et des informations sémantiques (section 4.3.3).

4.3.1 Filtrage et pondération des mots selon leur étiquette grammaticale

Afin d'améliorer les résultats obtenus par les mesures de similarité (section 4.2), nous avons effectué une extraction d'informations grammaticales du corpus à l'aide de Treetagger[9] (Schmid, 1994). (Roche & Kodratoff, 2006) et nos observations du corpus montrent que les CV sont des documents courts (le plus souvent ne dépassant pas une page) et syntaxiquement pauvres : absence de sujet et de verbe dans les phrases, phrases sous forme de résumé, nombreuses énumérations de noms et d'adjectifs, etc. Les mots respectant des étiquettes grammaticales spécifiques

[9]http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ : Treetagger est un système d'étiquetage automatique des catégories grammaticales des mots

ont été extraits : **N** (Nom) **A**(adjectif) **V**(Verbe). Ces seuls mots sélectionnés seront la base de la représentation vectorielle des documents. Par ailleurs, différentes combinaisons (NV, NA,VA et NVA) et pondérations[10] (2N4A, 3N2A, 4N2A, 2N3A, 4N3A, 3N3A) ont été expérimentées.

4.3.2 N-grammes de caractères

Utilisée principalement en reconnaissance de la parole, la notion de n-grammes de caractères prit davantage d'importance avec les travaux de (Damashek, 1995) sur le traitement de l'écrit. Ils montrent que ce découpage, bien que différent d'un découpage en mots, ne faisait pas perdre d'information. De nombreux travaux depuis ont montré l'efficacité des n-grammes comme méthode de représentation des textes (Mayfield & Mcnamee, 1998; Juola, 1998; Teytaud & Jalam, 2001; Hurault-Plantet *et al.*, 2005). Un n-gramme est une séquence de n caractères consécutifs. Pour un document quelconque, l'ensemble des n-grammes que l'on peut générer est le résultat que l'on obtient en déplaçant une fenêtre de n cases sur le corps de texte. Ce déplacement se déroule par étapes, une étape correspondant à un caractère. Ensuite les fréquences des n-grammes trouvés sont calculées. Par exemple, la phrase "Développeur php mysql" se représente avec des 3-grammes par [dev, eve, vel, elo, lop, opp, ppe, eur, ur_, r_p, _ph, php, hp_, p_m, _my, mys, ysq, sql]. Nous représentons les n-grammes en utilisant le caractère "_" pour caractériser les espaces. L'intérêt de cette représentation est qu'elle nous permet de capturer automatiquement les racines des mots les plus fréquents (Greffenstette, 1995). Un tel processus ne nécessite pas d'étape de recherche des racines lexicales. Le second intérêt de cette représentation est sa tolérance aux fautes d'orthographe et aux erreurs typographiques souvent présentes dans les CV et LM[11]. Nous avons testé différents n-grammes (*3/4/5/6-grammes*). Les résultats présentés en section 5.2 sont obtenus avec des *5-grammes* qui donnent les meilleurs résultats.

4.3.3 Enrichissement sémantique de la Mission et Relevance Feedback

L'observation des mots ayant le plus d'influence lors du calcul de la mesure de similarité, nous a conduit à envisager un enrichissement du contenu de la mission à l'aide d'une ontologie obtenue à partir de la base ROME[12] de l'ANPE[13]. Ainsi, pour chaque mission, nous effectuons un enrichissement de celle-ci à l'aide des compétences et niveaux d'études nécessaires afin de remplir cette fonction[14]. Les résultats de ces tests sont présentés en section 5.2 sous l'appellation *Mission enrichie*. Ceux-ci n'apportant pas toujours d'amélioration,, nous avons modifié le système afin d'intégrer un processus de retour de pertinence. Le retour de pertinence (*Relevance Feedback*) (Spärck Jones, 1970), a pour idée d'exploiter les documents retournés en réponse à une première requête pour améliorer le résultat de la recherche (Salton & Buckley, 1990). Nous simulerons ainsi la première requête à l'aide d'un tirage aléatoire de cinq candidatures étiquetées comme **positives** par un consultant en recrutement. Les résultats obtenus sont présentés en section 5.2 sous l'appellation *Relevance Feedback*.

[10] NA signifie que l'on combine les noms et les adjectifs, 2N3A signifie que l'on double le poids des noms combiné avec un poids triple pour les adjectifs, etc.

[11] Par exemple, un système fondé sur les mots aura des difficultés à reconnaître le mot "Dévelopeur" mal orthographié (avec un seul p).

[12] Répertoire Opérationnel des Métiers et des Emplois

[13] http ://www.anpe.fr/espacecandidat/romeligne/RliIndex.do

[14] Exemple : 32321/développeur/Bac+2 à Bac+4 en informatique CFPA, BTS, DUT ;Participe au développement et à la maintenance des applications informatiques,l'analyse fonctionnelle, la conception technique, le codage, la mise au point et la documentation des programmes etc..

5 Expérimentations

Nous avons sélectionné un sous corpus de la base de données d'Aktor. Le corpus *Corpus Mission*, d'environ 10Mo, est un ensemble d'offres d'emploi avec des thématiques différentes (emplois en comptabilité, entreprise, informatique, etc) associées aux réponses des candidats. Chaque candidature est identifiée comme **positive** ou **negative**. Une valeur **positive** correspond à un candidat potentiellement intéressant pour un emploi donné et une valeur **négative** a été attribuée à une candidature non pertinente, selon l'avis d'un consultant en recrutement. Le tableau 1 présente différentes statistiques de ce corpus (Kessler *et al.*, 2008b; Kessler *et al.*, 2008a).

Numéro	Mission	Nombre de Candidatures	Candidatures	
			positives	**négatives**
34990	collaborateur comptable	32	7	25
34861	ingénieur commercial(e)	40	14	26
31702	comptable, département fournisseurs	55	23	32
32461	développeur web php/mysql	60	7	53
33633	ingénieur commercial	65	18	47
34865	assistant comptable	67	10	57
34783	assistant comptable	108	9	99
33746	3 chefs de cuisine	116	60	56
33553	un délégué commercial	117	17	100
33725	conseiller commercial urbain	118	43	75
31022	assistant(e) en recrutement	221	28	193
31274	assistant comptable junior	224	26	198
34119	assistant commercial	257	10	247
31767	assistant comptable junior	437	51	386
Total		1917	323	1594

TAB. 1 – Statistiques du *Corpus Mission*.

5.1 Protocole expérimental

Nous souhaitons mesurer la similarité entre une offre d'emploi et ses candidatures. Le *Corpus Mission* contient 14 offres d'emploi associés à au moins cinq candidatures identifiées **positives**. Nous représentons les documents dans un espace vectoriel adéquat (Salton, 1991). Les mesures présentées en section 4.2 permettent d'effectuer un classement des candidats relativement aux offres proposées L'algorithme de décision (section 4.2) combine différentes mesures de similarité dans le but d'attribuer un score à chaque candidat. Pour évaluer la qualité de l'ordonnancement obtenu, nous calculons des courbes ROC (Ferri *et al.*, 2002), utilisées à l'origine dans le traitement du signal. Cette méthode est fréquemment employée en médecine afin d'évaluer automatiquement la validité d'un diagnostique de tests. On trouve en abscisse des axes représentant une courbe ROC le taux de faux positifs. La surface sous la courbe ROC ainsi créée est appelée AUC (*Area Under the Curve*). Dans l'ordonnancement de candidatures, une courbe ROC parfaite (AUC = 1) reviendrait à obtenir les candidatures pertinentes en tête du classement et celles non pertinentes, à la fin. La ligne diagonale correspond à la performance d'un système aléatoire : $AUC = 0.5$. Pour qu'un système soit considéré efficace, les AUC doivent avoir les valeurs les plus élevées possibles. Ceci revient à minimiser la somme des rangs des exemples positifs. Le principal avantage des courbes ROC est sa résistance au déséquilibre entre les exemples **positifs** et **négatifs** (Roche & Kodratoff, 2006). Pour chaque offre

d'emploi de notre corpus, nous évaluons la qualité du classement obtenu avec cette méthode. Les candidatures étudiées sont celles composées d'un CV et d'une LM.

5.2 Résultats

Le tableau 2 présente l'ensemble des résultats obtenus en fonction de chaque partie de la candidature (CV/LM) ou de la candidature globale (All). La colonne *Word TF* présente les résultats obtenus avec une représentation vectorielle de mots avec comme unité la fréquence du terme. *Word TF*IDF* utilise le produit de la fréquence des termes et de la fréquence inverse de documents. La colonne n-grammes présente les résultats obtenus avec des 5-grammes. Les colonnes *Mission enrichie* et *Relevance Feedback* présentent les résultats obtenus avec l'enrichissement sémantique et les meilleurs résultats obtenus avec la méthode Relevance Feedback présentée en section 4.3.3. Les deux dernières colonnes, *NVA* et *2N4A*, résument les résultats obtenus par filtrage et pondération des informations grammaticales présentés en section 4.3.1.

	Word TF	*Word TF*IDF*	*n-grammes*	*Mission enrichie*	*Relevance Feedback*	*NVA*	*2N4A*
Mission/All	0,64	0,64	0,60	0,62	**0,67**	0,61	0,64
LM	0,57	**0,62**	0,56	0,60	0,58	0,58	0,59
CV	0,62	0,60	0,60	0,61	**0,70**	0,59	0,60

TAB. 2 – Comparaison de résultats obtenus.

Nous observons que la représentation *Word TF*IDF* présente des résultats globalement similaires (Mission/All) au *Word TF*. Nous attribuons cette performance à la faible taille du corpus. Les combinaisons et pondération grammaticales présentent des résultats sensiblement équivalents au *Word TF*IDF*. Notons que les traitements que nous avons apportés (filtrage grammatical, pondération grammaticale, enrichissement sémantique) permettent d'améliorer les résultats fondés sur la fréquence pour certains types de documents (LM). Les n-grammes donnent des résultats d'AUC plus faibles que les autres représentations. Cependant, nous envisageons d'intégrer divers post-processus dans le but d'éliminer les séquences de caractères trop fréquentes ou à l'inverse non significatives. L'enrichissement sémantique ne semble pas améliorer la performance générale du système. Nous observons cependant une amélioration significative des résultats obtenus avec le retour de pertinence.

6 Conclusion et perspectives

Le traitement des offres d'emploi est une tâche extrêmement difficile car l'information est en format libre malgré une structure conventionnelle. Ces travaux ont mis en avant le module de traitement des réponses à des offres d'emplois, troisième module du projet E-Gen, système pour le traitement des offres d'emploi sur Internet. Celui-ci a pour but la mise en place d'un système d'aide à la décision pour le recruteur, ce dernier effectue une évaluation des premières candidatures afin de guider le système par la suite. Après différentes étapes de filtrage et de racinisation afin de produire une représentation vectorielle, nous effectuons un classement des candidatures à l'aide de différentes mesures de similarité et différentes représentations des documents. Les

premiers résultats à l'aide du retour de pertinence montrent une amélioration significative des AUC obtenues. Nous envisageons divers tests complémentaires (recherche de critères discriminants sur les candidatures identifiées comme négatives, pondération en fonction de l'importance de chacune des parties de la mission, etc.) pouvant apporter des améliorations. Nous souhaitons par ailleurs inclure divers paramètres tels que la richesse du vocabulaire, l'orthographe afin d'évaluer la lettre de motivation, ceux-ci étant à l'heure actuelle faiblement exploités lors de la prise de décision par les recruteurs. Nous envisageons par ailleurs la mise en place d'un système d'évaluation de CV sur le portail *jobmanager*[15] afin d'effectuer l'opération inverse (le candidat dépose son CV et le système lui propose les missions les plus adaptées à son profil).

Références

ALLEN C. & PILOT L. (2001). Hr-xml : Enabling pervasive hr- e-business. In *XML Europe 2001, Int. Congress Centrum (ICC), Berlin, Germany*.

BELLOT P. & EL-BÈZE M. (2001). Classification et segmentation de textes par arbres de décision. In *Technique et Science Informatiques (TSI)*, volume 20. Hermès.

BERNSTEIN A., KAUFMANN E., KIEFER C. & BÜRKI C. (2005). *SimPack : A Generic Java Library for Similarity Measures in Ontologies*. Rapport interne.

BIZER R. H. & RAINER E. (2005). Impact of Semantic web on the job recruitment Process. *International Conference Wirtschaftsinformatik.*

BOUDIN F. & TORRES MORENO J. M. (2007). Neo-cortex : A performant user-oriented multi-document summarization system. In *CICLing*, p. 551–562.

BOURSE M., LECLÈRE M., MORIN E. & TRICHET F. (2004). Human resource management and semantic web technologies. In *ICTTA.*

CAZALENS S. & LAMARRE P. (2001). An organization of internet agents based on a hierarchy of information domains. In *In Y. DEMAZEAU & F. J. GARIJO, Eds. Proceedings MAAMAW.*

CLECH J. & ZIGHED D. A. (2003). Data mining et analyse des cv : une expérience et des perspectives. In *Extraction et la Gestion des Connaissances, EGC'03*, p. 189–200.

DAMASHEK M. (1995). Gauging similarity with n-grams : Language-independent categorization of text. *Science 1995 ; 267*, p. 843–848.

DESMONTILS E., JACQUIN C. & MORIN E. (2002). Indexation sémantique de documents sur le web : application aux ressources humaines. In *AS CNRS Web Sémantique 2002*.

DORN J. & NAZ T. (2007). Meta-search in human resource management. In *in Proceedings of 4th International Conference on Knowledge Systems ICKS'07 Bangkok,Thailand*.

DORN J., NAZ T. & PICHLMAIR M. (2007). Ontology development for human resource management. In *International Conference on Knowledge Management Vienna*.

FERRI C., FLACH P. & HERNANDEZ-ORALLO J. (2002). Learning decision trees using the area under the ROC curve. In *Proceedings of ICML'02*, p. 139–146.

GREFFENSTETTE G. (1995). Comparing two language identification schemes. *Communications JADT 1995*, p. 85–96.

HURAULT-PLANTET M., JARDINO M. & ILLLOUZ G. (2005). Modèles de langage n-grammes et segmentation thématique. *Actes TALN & RECITAL, vol 2*, p. 135–144.

[15] http ://www.jobmanager.fr

JUOLA P. (1998). Cross-entropy and linguistic typology. *Proceedings of New Methods in Language Processing, 3, Sydney, Australie.*, p. 843–848.

KESSLER R., BÉCHET N., ROCHE M., EL-BÈZE M. & TORRES-MORENO J. M. (2008a). Automatic profiling system for ranking candidates answers in human resources. In *OTM '08 in Monterrey, Mexico*, p. 625–634.

KESSLER R., TORRES-MORENO J. M. & EL-BÈZE M. (2007). E-Gen : Automatic Job Offer Processing system for Human Ressources. *MICAI*.

KESSLER R., TORRES-MORENO J. M. & EL-BÈZE M. (2008b). E-Gen : Profilage automatique de candidatures. *TALN 2008, Avignon, France*.

MAYFIELD J. & MCNAMEE P. (1998). Indexing using both n-grams and words. *NIST Special Publication*, p. 500–242.

MOCHO M., PASLARU E. & SIMPERL B. (2006). Practical Guidelines for Building Semantic eRecruitment Applications. *I-Know'06 Special track on Advanced Semantic Technologies*.

MORIN E., LECLÈRE M. & TRICHET F. (2004). The semantic web in e-recruitment (2004). In *The First European Symposium of Semantic Web (ESWS'2004)*.

QUILAN J. (1993). C4.5 : Programs for machine learning. In *Kaufmann, San Mateo, CA*.

RAFTER R., BRADLEY K. & SMYT B. (2000a). Automated Collaborative Filtering Applications for Online Recruitment Services. p. 363–368.

RAFTER R. & SMYTH B. (2000). Passive Profiling from Server Logs in an Online Recruitment Environment.

RAFTER R., SMYTH B. & BRADLEY K. (2000b). Inferring Relevance Feedback from Server Logs : A Case Study in Online Recruitment.

ROBERTSON S., WALKER S., JONES S., HANCOCK-BEAULIEU M. M. & GATFORD M. (1994). Okapi at trec-3. *NIST Special Publication 500-225 : TREC-3*, p. 109–126.

ROCHE M. & KODRATOFF Y. (2006). Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. In *OTM'06, Montpellier, France*, p. 1107–1116.

ROCHE M. & PRINCE V. (2008). Evaluation et détermination de la pertinence pour des syntagmes candidats à la collocation . *JADT2008*, p. 1009–1020.

SALTON G. (1991). Developments in automatic text retrieval. *Science 253*, p. 974–979.

SALTON G. & BUCKLEY C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, p. 288–297.

SALTON G. & MCGILL M. J. (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA : McGraw-Hill, Inc.

SCHMID G. (1994). *TreeTagger - a language independent part-of-speech tagger.*

SPÄRCK JONES K. (1970). Some thoughts on classification for retrieval. *Journal of Documentation*, **26**, 89–101.

TEYTAUD O. & JALAM R. (2001). Kernel-based text categorization. *IJCNN'01, Washington,DC, USA.*

TOLKSDORF R., MOCHO M., HEESE R., OLDAKOWSKI R. & CHRISTIAN B. (2006). Semantic-Web-Technologien im Arbeitsvermittlungsprozess . *International Conference Wirtschaftsinformatik*.