Profilage de candidatures assisté par relevance Feedback

Rémy Kessler, Nicolas Béchet, Juan-Manuel Torres-Moreno, Mathieu Roche, Marc El Bèze

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 connaît une croissance exponentielle. Ceci implique des volumes d'information (majoritairement sous la forme de texte libre) qu'il n'est plus possible de traiter manuellement. Une analyse et catégorisation assistées nous semble pertinente en réponse 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 d'une forme de *relevance feedback* a permis de surpasser nos résultats sur ce problème difficile 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 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) et d'un marché du recrutement en ligne en expansion significative (août 2003 : 177 000 offres, mai 2008 : 500 000 offres)[1]. Internet est

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

devenu essentiel dans ce processus, car il permet une meilleure diffusion de l'information, que ce soit par les sites de recherche d'emplois ou par les échanges de courriels. Cependant divers problèmes se posent dans son traitement notamment en raison de la grande quantité d'information difficile à gérer rapidement et efficacement pour les entreprises (Bourse *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" des données entre différentes applications, est une problématique fréquente lors de l'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, chacun d'eux étant chargés de :

– Extraire de l'information à partir de corpus de courriels provenant d'offres d'emplois extraites de la base de données d'Aktor[2].
– Analyser les e-mails de réponses des candidats pour distinguer lettre de motivation (LM) et curriculum vitæ(CV).
– Analyser et calculer un classement de pertinence des candidatures (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 Vecteurs Support MVS), 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 sections 4 et 5 les stratégies pour le profilage des candidatures, le protocole expérimental, les statistiques sur le corpus et les résultats obtenus.

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. (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) en partenariat avec le 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) décrit l'importance d'une ontologie commune (*HR ontology*) afin de pouvoir traiter efficacement ce type de

[2] Aktor Intéractive (www.aktor.fr)

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, s'appuyant sur la technologie HR-XML mise en place par (Allen & Pilot, 2001), (Dorn *et al.*, 2007; Dorn & Naz, 2007) décrivent 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) décrivent 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 non cadres. 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. Cette méthode permet de mettre 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 décevante au niveau des résultats obtenus (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]. Leur approche permet d'extraire un certain nombre de collocations contenues dans les CV sur la base de patrons (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 spécialisée. Notre approche diffère des méthodes proposées par le fait qu'elle s'appuie sur une combinaison de mesures de similarité afin d'effectuer un classement des candidatures avec une phase de *Relevance Feedback*. Ceci nous permet de prendre en compte l'opinion de l'utilisateur.

3 Vue d'ensemble du système

Nous avons choisi de développer un système répondant 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 à l'aide de deux modules externes : *wvWare*[4] et *pdftotext*[5]. 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 l'annonce à l'aide des MVS et 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 système extrait le corps du message ainsi que les pièces jointes. Une version texte des documents contenus dans la candidature est

[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

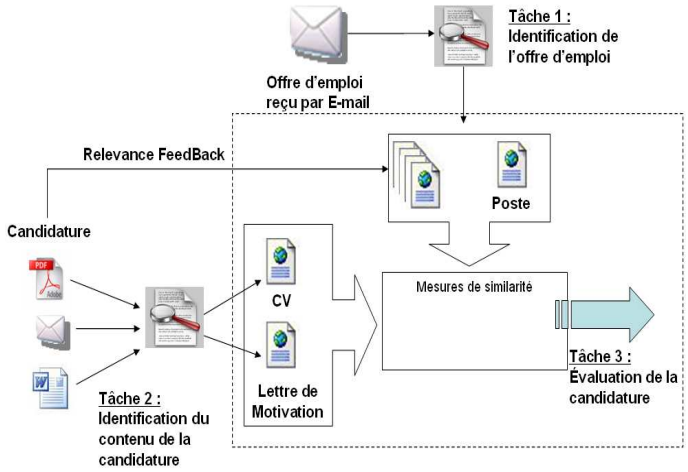

FIG. 1 – Vue d'ensemble du système.

alors produite. Différents processus de filtrage et racinisation permettent au système d'identifier à l'aide de MVS et de règles le type du document (CV et/ou LM). Ce processus (tâche 2) est détaillé dans (Kessler *et al.*, 2008b). Une fois la LM et le CV identifiés, le système effectue un profilage automatisé de cette candidature à l'aide de mesures de similarité en s'appuyant sur un nombre de candidatures préalablement validées comme candidatures pertinentes par un consultant en recrutement (tâche 3). Le processus mis en œuvre est détaillé dans la section suivante. La figure 1 résume le processus global 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 telles que les noms des candidats, les adresses, les courriers électroniques, les 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 etc.*), les expressions courantes (*par exemple, c'est-à-dire, chacun de, etc.*), les chiffres et nombres (numériques et/ou textuelles), les symboles comme "$", "#","*", ainsi que les termes contenu dans un anti-dictionnaire. Enfin, nous avons ramené les verbes fléchis à leur racine et les adjectifs pluriels et/ou féminins à une forme canonique à l'aide d'un dictionnaire[8].

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 caractérisant 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 section 4.3.2

[8] Ainsi les mots *développe, développent, développé, développeront, développement* et éventuellement *développeur* seront ramenés à la même forme.

avons mis en place une approche par mesure de similarité, afin de pouvoir ordonner 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 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 utilisée est Okabis (formule 3). Cette mesure est une version simplifié de la formule Okapi, souvent utilisée en Recherche d'Information, qui permet d'obtenir de meilleures performances (Bellot & El-Bèze, 2001) :

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

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

$$\text{Okabis}(d,j) = \sum_{i \in d \cap j} \frac{\sum_{i=1}^{n} j_i \cdot d_i}{\sum_{i=1}^{n} j_i \cdot d_i + \frac{\sqrt{|d|}}{M_d}} \tag{3}$$

Avec j une offre d'emploi, d la candidature, i un terme, j_i et d_i le nombre d'occurrences de i respectivement dans j et d et M_d leur taille moyenne. D'autres mesures (Overlap, Enertex, Needleman-Wunsch, Jaro-Winkler) ont été expérimentées, mais n'ont pas été prises en compte suite aux résultats obtenus. Les mesures utilisées ainsi que leur combinaison sont détaillées dans (Kessler *et al.*, 2008a).

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). Le processus de retour de pertinence est également décrit dans la 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 du TreeTagger[9] (Schmid, 1994). (Roche & Kodratoff, 2006) et nos observations du corpus à partir des données textuelles montrent que les CV sont des documents courts (le plus souvent ne dépassant pas une page) et syntaxiquement pauvres : peu de sujets et de verbes dans les phrases, phrases sous forme de résumé, nombreuses énumérations de noms et d'adjectifs, etc. Les mots respectant des étiquettes grammaticales spécifiques peuvent donc être intéressants. Nous proposons donc d'extraire les mots suivants : **N** (Nom) **A**(adjectif) **V**(Verbe). Ces seuls mots sélectionnés seront la base de la représentation vectorielle des documents. Par ailleurs, différentes combinaisons et pondérations ont été expérimentées.

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

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 ne fait pas perdre d'information. De nombreux travaux depuis ont montré l'efficacité des n-grammes comme méthode de représentation des textes (Mayfield & Mcnamee, 1998; Hurault-Plantet *et al.*, 2005). Un n-gramme est une séquence de n symboles consécutifs. Pour un document quelconque, l'ensemble des n-grammes que l'on peut générer est le résultat que nous obtenons 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. Notons qu'aucun des prétraitements présenté en 4.1 ne sont utilisé dans cette représentation. Par exemple, la phrase "Développeur php mysql" est représenté 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 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[10]. Nous avons testé différents n-grammes (*3/4/5/6-grammes*).

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 de connaissances sémantiques obtenues à partir de la base ROME[11] de l'ANPE[12]. 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[13]. 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 (*Relevance Feedback*) (Spärck Jones, 1970). (Rafter *et al.*, 2000a) propose l'utilisation du *Relevance Feedback* afin de guider l'internaute dans sa recherche d'emploi à partir d'informations récoltées sur le site JobFinder[14]. Dans notre système, la méthode *Relevance Feedback* permet de prendre en compte les choix du recruteur lors d'une première évaluation de quelques CV. Cette approche repose sur l'exploitation des documents retournés en réponse à une première requête pour améliorer le résultat de la recherche (Salton & Buckley, 1990). Dans notre contexte, nous effectuons un tirage aléatoire de quelques candidatures (de une à six dans nos expérimentations) parmi l'ensemble des candidatures étiquetées comme **pertinentes**. Celles-ci sont ajoutées à la Mission. Nous enrichissons ainsi l'espace vectoriel par les termes appartenant à des candidatures jugées pertinentes par un consultant en recrutement. Ceci nous permet d'effectuer un nouveau calcul de similarité entre la candidature que nous évaluons et la mission enrichie de l'opinion du recru-

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

[11] Répertoire Opérationnel des Métiers et des Emplois

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

[13] 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.

[14] JobFinder (jobfinder.com)

teur. Les résultats obtenus sont présentés en section 5.2 sous l'appellation *Relevance Feedback*.

5 Expérimentations

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

Numéro	Mission	Nombre de Candidatures	Candidatures	
			pertinentes	**non pertinentes**
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 associées à au moins sept candidatures **pertinentes**. Pour évaluer la qualité des classements, nous construisons des courbes ROC (Ferri *et al.*, 2002) utilisées à l'origine dans le traitement du signal. On trouve en abscisses d'une courbe ROC le taux de faux positifs et les ordonnées sont relatives aux taux de vrais positifs. La surface sous la courbe ROC ainsi créée est appelée AUC (*Area Under the Curve*). Le principal avantage des courbes ROC est leur résistance au déséquilibre entre les exemples positifs et négatifs (Roche & Kodratoff, 2006). Pour chaque offre d'emploi du corpus, nous évaluons la qualité du classement avec cette méthode. Les candidatures étudiées sont celles composées d'un CV et d'une LM.

5.2 Résultats

Le tableau 2 présentent un comparatif des meilleurs résultats obtenus par chaque méthode. Chaque test a été effectué une centaine de fois avec une distribution aléatoire des candidatures **pertinentes** ajoutés au *Relevance Feedback*. Nous effectuons une moyenne des AUC obtenues pour chaque mesure. *TF* montre les résultats obtenus à partir d'une représentation en fréquence de termes. *TF-IDF* utilise le produit de la fréquence des termes et de leur fréquence inverse en documents. Les représentations *TF* et *TF-IDF* donnent des résultats sensiblement similaires avec des scores AUC de 0,64, la taille réduite de nos corpus pouvant expliquer ces résultats.

	N-grams	*Offre d'emploi enrichie*	*TF*	*TF-IDF*	*Étiquettes grammaticales*	*Relevance Feedback*
Offre d'Emploi /CV et LM	0,60	0,62	0,64	0,64	0,64	**0,66**

TAB. 2 – Comparaison entre les scores AUC obtenues par chaque méthode.

Les combinaisons et pondérations d'étiquettes grammaticales (voir section 4.3.1) ne semblent pas apporter d'amélioration (*Étiquettes grammaticales*). Les résultats N*-grams* sont obtenus à partir des 5-grammes. Avec un score AUC de 0,6 au mieux, ceux- ci restent relativement faibles. Dans de futurs travaux, nous envisageons d'étudier les raisons précises de la faible performance obtenue avec les n-grammes. Différents post-processus dans le but d'éliminer les séquences de caractères trop fréquentes ou non significatives ont été envisagés mais sans réellement d'amélioration majeure. Avec un score AUC de 0,62, l'enrichissement sémantique (*Offre d'emploi enrichie*) ne semble pas améliorer la performance générale du système. Nous observons cependant une amélioration des résultats obtenus avec le retour de pertinence. La figure présente les résultats pour chaque taille de *Relevance Feedback* (RF1 correspond à une candidature ajoutés a la mission, RF2 deux, etc.). RF1 obtient une moyenne de 0,65 et RF6 de 0,66. Nous utilisons actuellement la méthode *residual ranking* (Billerbeck & Zobel, 2006) : les documents utilisés pour le *Relevance Feedback* sont retirés de la collection avant d'effectuer la requête reformulée.

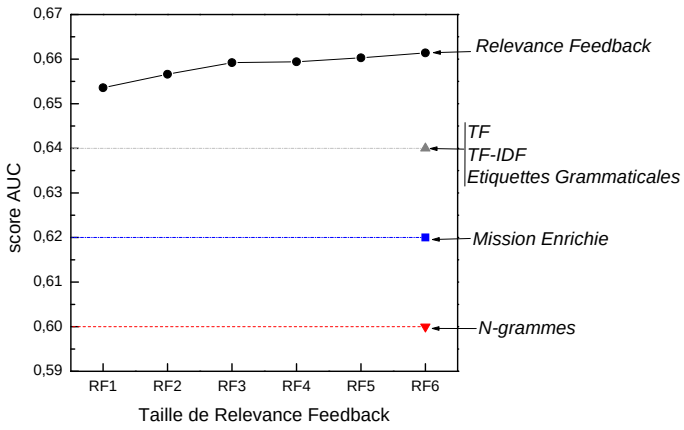

FIG. 2 – Comparaison entre les scores AUC pour chaque méthode.

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 mesures de similarité et diverses représentations des documents. Les premiers résultats à l'aide du retour de pertinence montrent une amélioration des AUC obtenues. Nous envisageons quelques tests complémentaires (recherche de critères discriminants sur les candidatures identifiées comme non pertinentes, pondération en fonction de l'importance de chacune des parties de la mission, etc.) pouvant apporter de nouvelles améliorations. Nous souhaitons par ailleurs inclure d'autres paramètres tels que la richesse du vocabulaire et 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, p. 107–134. Hermès.

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

BILLERBECK B. & ZOBEL J. (2006). Efficient query expansion with auxiliary data structures. *Inf. Syst.*, (7), 573–584.

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

BOURSE M., LECLÈRE M., MORIN E. & TRICHET F. (2004). Human resource management and semantic web technologies. In *ICTTA*, p. 641–642.

CAZALENS S. & LAMARRE P. (2001). An organization of internet agents based on a hierarchy of information domains. In *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.

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*, p. 105 – 110.

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

DORN J., NAZ T. & PICHLMAIR M. (2007). Ontology development for human resource management. In *International Conference on Knowledge Management Vienna*, p. 109–120.

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.

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, 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 2007, Aguscalientes, Mexique, pp 985-995*.

KESSLER R., TORRES-MORENO J. M. & EL-BÈZE M. (2008b). E-Gen : Profilage automatique de candidatures. *TALN 2008, Avignon, France*, p. 370–379.

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

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. *In 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. *In Proceedings of EACL-SIGDAT 1995. Dublin, Ireland.*, p. 44–49.

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

TOLKSDORF R., MOCHO M., HEESE R., OLDAKOWSKI R. & CHRISTIAN B. (2006). Semantic-Web-Technologien im Arbeitsvermittlungsprozess. *International Conference Wirtschaftsinformatik*, p. 17–26.