e-Recrutement : recherche de mots-clés pertinents dans le titre des annonces d'emploi

Julie Séguéla [1, 2], Gilbert Saporta [1], Stéphane Le Viet [2]

[1] Laboratoire Cédric – CNAM – 292 rue Saint Martin – 75141 Paris cedex 03 – France

[2] Multiposting.fr – 33 rue Réaumur – 75003 Paris – France

Résumé

La multiplication du nombre de moyens de recrutement sur Internet a rendu de plus en plus importante l'analyse de la performance des annonces d'emploi et son optimisation. Notamment, celle-ci peut être évaluée à travers le volume de candidatures reçues suite à la publication d'une annonce sur un job board (rendement). Nous proposons une méthode pour repérer des mots-clés au sein du titre des annonces d'emploi permettant d'apporter des éléments d'explication du rendement obtenu. Après un pré-traitement classique d'un corpus d'offres d'emploi diffusées sur Internet et segmentées par rapport à la fonction recherchée, nous calculons les spécificités lexicales associées et relevons les formes les plus fréquentes. Les mots ainsi retenus sont codés à l'aide d'indicatrices puis introduits dans un arbre de régression en tant que prédicteurs. Les résultats obtenus sont encourageants et permettent de valider la pertinence des mots-clés trouvés pour l'explication du rendement, en complément de l'information déjà apportée par la fonction recherchée.

Abstract

The increase in the number of different ways to recruit online has made the issue of job posting performance and its optimization more and more important. In particular, performance can be assessed by the amount of applications received. In this paper, the goal is to define a strategy for detecting key-words within a job posting's title that provide information to explain the posting's performance. We will work on a corpus of job postings characterized by the underlying function (e.g. marketing, finance, …). After a pre-processing of filtering on the corpus, we will compute characteristic forms and identify the most frequent forms. In order to confirm the relevance of the key-words obtained, the latter are coded into dichotomic variables and introduced as predictors in a regression tree, in addition to function predictors. The results are encouraging and confirm key-words contribution into explaining performance.

Keywords: job offers, keywords, lexical specificities, CART

1. Contexte

Apparus vers la fin des années '90, les job boards [1] sont à l'origine de la part croissante des appariements offre-demande sur le marché du travail dus au média Internet (Fondeur and Tuchszirer, 2005). La démultiplication du nombre de job boards existant et les coûts engendrés par leur utilisation (monétaires et temporels) ont rendu essentiel l'analyse et l'optimisation de la performance des offres d'emploi diffusées par les recruteurs. Nous nous intéressons ici à la problématique liée au volume des candidatures reçues (nombre de CV) lors d'une campagne de recrutement en ligne, qui sera donc notre mesure de la performance d'une offre d'emploi.

[1] Sites internet d'offres d'emploi.

Multiposting.fr [2], outil de multidiffusion d'offres d'emploi sur Internet, met à disposition pour cette étude une base de données d'offres d'emploi ayant été postées sur une centaine de sites différents entre décembre 2008 et janvier 2010. Pour chaque offre postée sur un site, on connaît le nombre de retours (candidatures) associé. Par la suite, la notion de rendement pourra être utilisée pour désigner le volume de candidatures engendrées par la publication d'une offre d'emploi sur un job board.

Les annonces d'emploi sont des textes relativement courts (entre 1.000 et 3.000 caractères), écrits dans un format libre, mais qui s'articulent en quatre parties différenciées dans la base de données : le titre, le descriptif de la société qui recrute, la description du poste et le profil recherché. Nous concentrons ici nos recherches sur l'étude des mots présents dans le titre du poste à pourvoir, composant essentiel d'une annonce d'emploi « synthétisant » en quelques mots les qualifications recherchées pour le poste en question. Dans leur étude ciblée sur le secteur des technologies de l'information, Aureli et Iezzi (2006) mettent en évidence des formes au sein des offres d'emploi traduisant des compétences spécifiques au domaine ou des formes communes que nous appellerons transversales. Nous étudions ici l'effet de la présence dans le titre de mots traduisant des qualifications transversales ou spécifiques à une fonction [3] sur le volume de candidatures reçues suite à la diffusion de l'annonce. L'objectif poursuivi consiste à montrer que les mots-clés extraits du titre (par une méthodologie exposée en détail dans la section 2.2) permettent d'améliorer l'explication du rendement en comparaison à l'utilisation de la catégorie fonction seule. Ces mots-clés, plus précis que la fonction associée au poste, vont apporter des informations sur les compétences recherchées au sein d'une catégorie de fonction, et faire ressortir des différences de rendements selon le profil recherché.

Afin d'éliminer les variations du rendement liées aux caractéristiques intrinsèques du job board utilisé (audience, type de visiteurs, moteur de recherche, etc.), nous nous concentrerons sur l'étude d'un seul job board de la base de données spécialisé dans les emplois cadres. Deux raisons principales motivent ce choix : l'importance du volume d'offres d'emploi postées sur ce site, et la présence d'un champ segmentant les offres selon la fonction recherchée (cf. section 2). Fig. 1 présente un extrait d'affichage d'offres d'emploi suite à une requête [4] effectuée sur le site en question.

DATE ▲▽	FONCTION PROPOSÉE ▲▼	SECTEUR D'ACTIVITÉ ▲▼	LOCALISATION ▲▼
04/01/10 LOGO	**Directeur commercial h/f (CDI)** Entreprise : xxxxxxxxxxxxxxx Contractants Régionaux, 1er réseau ... Poste : Directeur Commercial h/f - Ile de France Responsable ...	Société immobilière/promoteur	Ile de france
04/01/10 LOGO	**Conseiller clientele haut de gamme h/f (CDI)** Entreprise : xxxxxxxxx, cabinet entièrement dédié à l'intérim ... Poste : Votre mission : Au sein d'une agence à taille humaine, ...	Banque	Paris

Figure 1 : Exemple de résultat d'une recherche effectuée sur le job board étudié

[2] http://www.multiposting.fr/.

[3] Sur de nombreux sites d'emploi, le recruteur doit associer une fonction à l'offre d'emploi au moment de la publication (*Marketing*, *Administration*, *Communication*, *Production*, etc.).

[4] Critères de la recherche : type de contrat = CDI ; fonction = Commercial-Vente ; localisation = Ile-de-France.

Pour avoir accès au texte complet de l'offre, les candidats potentiels doivent cliquer sur le lien défini par le titre du poste. Suite à cela, ils pourront s'ils le souhaitent déposer une candidature. Cette méthode de présentation des offres montre bien l'importance de la rédaction du titre de l'annonce pour susciter l'intérêt des candidats et les mener à la poursuite du processus de candidature.

Dans la section 2, nous décrivons le corpus des offres d'emploi étudiées et proposons une méthode de détection des mots-clés du titre ainsi qu'un modèle pour tester l'effet de leur présence sur le rendement. Cette méthode, illustrée par une analyse exploratoire préliminaire, repose sur le codage en indicatrices des mots-clés repérés grâce à l'étude des fréquences d'apparition et des spécificités lexicales associées aux différentes catégories de fonctions des offres d'emploi (section 3). Les résultats obtenus sont présentés dans la section 4.

2. Corpus et méthodes

Le corpus étudié est un sous-ensemble de données extraites de la base de données d'annonces d'emploi historisées par Multiposting.fr. Comme nous l'avons évoqué précédemment, nous nous concentrons sur l'analyse des annonces d'emploi postées sur un job board, ce dernier étant choisi à la fois pour son volume d'annonces et leur segmentation par rapport à la fonction sous-jacente au poste (17 catégories initialement). Cependant, nous disposons également de ce champ fonction pour des offres postées sur d'autres sites d'emploi. Pour l'étape de détection des mots-clés, nous décidons d'exploiter toutes les annonces présentant le champ fonction, puis leur nombre sera réduit au périmètre du site choisi dès lors que la notion de rendement sera prise en compte (le rendement est défini par le nombre de CV reçus). Le corpus présenté dans la section 2.1 correspond à l'ensemble des annonces de la base détenant la donnée catégorielle de fonction du poste.

2.1. Pré-traitement des annonces et description du corpus

L'extraction a été effectuée en procédant aux filtrages suivants :

- exclusion des annonces de test (par filtrage sur la référence ou le titre de l'annonce) ;
- exclusion des annonces rédigées en anglais (notre étude porte sur les offres rédigées en français qui constituent la quasi-totalité du corpus étudié) ;

Le nombre élevé de catégories de fonctions définies qui rendra difficile les interprétations ultérieures, ainsi que le manque d'observations pour un certain nombre d'entre elles nous conduisent à limiter l'analyse à quatre de ces fonctions. Ce qui nous amène à effectuer un filtrage supplémentaire :

- sélection des offres faisant référence à des postes à pourvoir dans les fonctions *Commercial-Vente*, *Gestion-Comptabilité-Finance*, *Marketing* et *Systèmes d'Information-Télécom*.

Lors de l'étape d'évaluation des mots-clés par le test de leur influence sur le rendement, un filtrage supplémentaire des données sera effectué :

- sélection des offres postées sur le job board choisi (cf. section 1). Si une même offre, c'est-à-dire faisant référence au même poste à pourvoir, est diffusée plus d'une fois sur le job board alors seule la première diffusion est prise en compte.

Pour la suite, nous conservons uniquement le titre des annonces (partie étudiée) au sein du corpus. Ces textes ont la particularité d'être très courts et très peu pollués (pas de verbes, pas d'expressions courantes). On rencontre toutefois quelques mots-outils (*de*, *d'*, *et*, etc.) que nous choisissons de filtrer car ne présentant pas d'intérêt pour la problématique étudiée. Nous filtrons également les formes faisant référence à la déclinaison du poste au féminin (*e*, *trice*, sigle *h/f*,

etc.) pour le motif évoqué précédemment. Les titres peuvent être écrits en majuscules et/ou en minuscules. Cet aspect n'étant pas pris en considération ici, nous convertissons tous les textes en minuscules et éliminons les accents afin que deux mots ayant le même sens ne soient pas associés à deux formes différentes [5].

Tab. 1 résume les statistiques descriptives du corpus pour chacun des types de fonction étudiés après les filtrages évoqués ci-dessus.

Fonction	*Nombre d'annonces*	*Nombre d'occurrences*	*Nombre de formes*
Commercial-Vente	231	732	241
Gestion-Comptabilité-Finance	136	453	145
Marketing	144	577	176
Systèmes d'Information-Télécom	198	780	255
Total	709	2542	616

Tableau 1 : Statistiques descriptives du corpus

On constate que le corpus est assez pauvre, ce qui est du à sa particularité : les titres des annonces sont des textes très courts (3.6 mots en moyenne hors mots filtrés) dont les formes sont peu variées car généralement conventionnelles pour les recruteurs.

2.2. Méthode pour l'extraction et l'évaluation des mots-clés

Nous cherchons à mettre en évidence les mots du titre faisant référence à des qualifications spécifiques à chacune des fonctions étudiées, ainsi qu'à des qualifications transversales, c'est-à-dire pouvant être recherchées dans toutes les fonctions. Nous visualiserons dans un premier temps les profils lexicaux des différentes catégories d'offres d'emploi à l'aide d'une analyse des correspondances, puis nous compléterons cette analyse par le calcul des formes spécifiques à l'aide d'un modèle probabiliste. Par définition, les spécificités positives sont les formes qui sont significativement « sur-employées » dans la partie du corpus considérée (Lafon, 1980). Nous les utilisons ici pour détecter des compétences représentatives des fonctions étudiées mais aussi plus précises. Pour chaque fonction étudiée, nous calculerons les spécificités positives et les formes ainsi obtenues détermineront un premier ensemble de mots-clés. Ensuite, nous relèverons les formes de plus hautes fréquences parmi celles n'étant pas des spécificités lexicales afin d'obtenir des mots-clés transversaux à toutes les fonctions.

Le procédé de codage des mots-clés est ensuite très simple. Pour chaque offre du corpus, la présence ou l'absence dans le titre de l'ensemble des mots-clés retenus est codée à l'aide de variables indicatrices. L'impact des mots-clés ainsi codés sera testé à l'aide d'un arbre, construit avec la méthode CART. Dans le cas présent, la nature quantitative de la variable à expliquer (nombre de candidatures reçues) nous suggère d'utiliser un arbre de régression. Les indicatrices codées précédemment ainsi que des indicatrices codant les catégories de fonctions seront introduites en tant que variables prédicteurs. L'algorithme CART (Breiman et al., 1984), faisant référence à un partitionnement récursif binaire, semble particulièrement adapté à la nature des prédicteurs (deux modalités par construction). La représentation arborée nous permettra également de faire apparaître des interactions entre les différents mots-clés, et l'ordre des partitionnements permettra de détecter les mots les plus significatifs.

[5] Le problème contraire, à savoir donner le même sens à deux formes à l'origine distinctes, n'est pas rencontré ici. Ceci est du à la particularité du corpus étudié qui présente une diversité de formes assez limitée.

3. Analyses lexicales

Les analyses lexicales sont menées sur le corpus non lemmatisé, après les pré-traitements décrits dans la section 2.1. Nous fixons un seuil arbitraire de 12 occurrences minimum pour la prise en compte des formes dans les analyses. Elles sont donc réalisées [6] sur le tableau lexical croisant les 40 formes conservées (1.289 occurrences) avec les quatre catégories de fonctions.

3.1. Exploration

Afin de visualiser de manière aisée les associations entre les formes et les proximités entre les différentes catégories vis-à-vis du vocabulaire employé, nous procédons à une méthode classique d'analyse des correspondances (Lebart and Salem, 1994) sur le tableau lexical évoqué ci-dessus. Nous représentons les formes et les catégories sur le premier plan factoriel (Fig. 2). Afin de valider la représentation obtenue, nous traçons également sur le graphe les ellipses de confiance associées aux points désignant les catégories de fonctions par une méthode de bootstrap total de type 2 (Lebart, 2007).

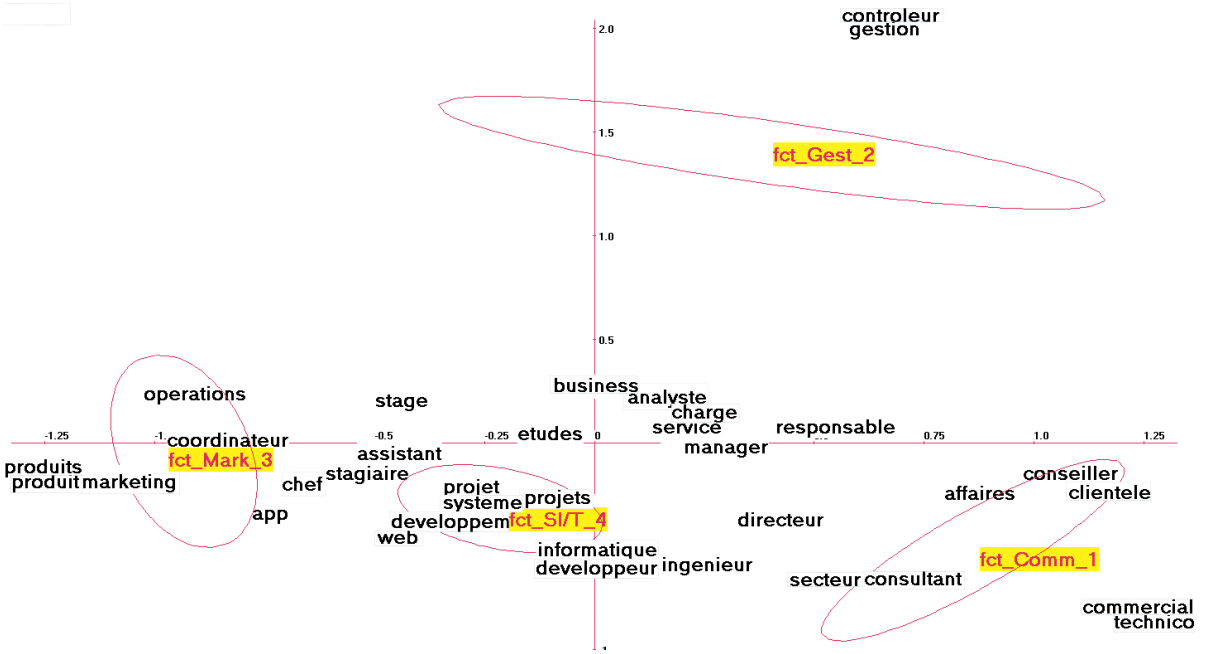

Figure 2 : Premier plan factoriel de l'AC du tableau lexical croisant formes et catégories de fonctions

Nous obtenons une représentation valide dans le premier plan factoriel puisque les pourcentages d'inertie expliquée par les deux premiers axes (il y en a trois au total) sont respectivement égaux à 39% et 34%. Sur le graphique, les formes proches sont souvent utilisées dans la même catégorie fonction (par exemple *conseiller*/*clientèle*, *controleur*/*gestion*). Sur le premier axe s'opposent les offres associées aux fonctions *Marketing* et *Commercial* (contributions respectivement égales à 50% et 42%), tandis que le deuxième axe oppose la fonction *Gestion* (contribution égale à 77%) aux trois autres. La fonction *SI-Télécom*, mal discriminée sur les deux premiers axes, s'oppose aux autres fonctions sur le troisième axe (contribution égale à 66%). La position des formes sur le graphique ainsi que les zones de confiance autour des « points-fonctions » permettent de mettre en évidence des profils lexicaux assez différents pour les quatre types d'offres étudiés (deux

[6] Les analyses lexicales sont réalisées à l'aide du logiciel DTM développé par L. Lebart.

ellipses sont très étendues mais elles demeurent éloignées les unes des autres). Les recruteurs adaptent donc le vocabulaire utilisé selon la fonction recherchée. Comme on pouvait s'y attendre, il y a des compétences spécifiques à chacune de ces catégories, que nous allons identifier.

3.2. Mots-clés

Le tableau 2 présente les formes caractéristiques obtenues pour les quatre types de fonctions ainsi que les formes de plus hautes fréquences, communes à toutes les fonctions. Nous retenons les formes caractéristiques ayant un seuil de significativité d'au moins 0.5% (p-value inférieure ou égale à 0.005). Lorsqu'une forme est caractéristique de deux fonctions ou plus, alors elle sera considérée comme transversale à toutes les fonctions. Etant donné le nombre de formes étudiées, les formes transversales sont extraites parmi les formes ayant un nombre d'occurrences supérieur ou égal à 15. Tab. 2 présente également des exemples de titres caractéristiques de chaque catégorie.

Fonction Commercial commercial – rayon – conseiller – technico clientele – affaires *ex : conseiller commercial*	**Fonction Gestion-Compta-Finance** gestion – comptable – controleur – controle *ex : controleur de gestion*
Fonction Marketing marketing – produit – operations – coordinateur – produits – assistant *ex : marketing manager produits*	**Fonction SI-Télécom** ingenieur – projet – informatique – systeme – developpeur – projets – etudes – developpement *ex : ingenieur d'etudes et developpement informatique*
Toutes fonctions chef – responsable – charge – stage – stagiaire – manager – directeur – analyste	

Tableau 2 : Mots-clés extraits du titre des offres

Les mots-clés obtenus sont dans l'ensemble cohérents avec la représentation délivrée par Fig. 2. Par exemple, on retrouve les formes *charge* et *manager* près de l'origine car transversales à plusieurs fonctions. La présence de ces mots est ensuite codée à l'aide d'indicatrices [7] introduites en tant que prédicteurs dans l'algorithme CART en complément des indicatrices associées aux fonctions.

4. Résultats

Dans une première étape, nous construisons trois arbres distingués par l'ensemble des variables explicatives qui y sont introduites. Le premier arbre est construit uniquement à partir des quatre indicatrices associées aux fonctions. Dans le deuxième arbre, nous introduisons en plus des fonctions les indicatrices associées aux mots-clés identifiés précédemment. Un troisième arbre prenant en compte seulement les mots-clés est construit pour comparaison. Afin de confirmer l'intérêt de l'introduction des mots-clés dans l'arbre pour améliorer le pouvoir explicatif, la deuxième étape consiste à calculer les erreurs de prédiction par validation croisée [8] associées aux

[7] Les formes produit/produits sont considérées comme une même forme. De même pour les formes projet/projets, controleur/controle et stage/stagiaire.

[8] Nous procédons à une validation croisée par *leave-one-out*. Il est important de noter que les arbres construits lors du processus de validation croisée ne sont pas identiques car estimés à partir d'échantillons différents mais partagent le même ensemble de prédicteurs et le même paramètre de complexité (celui de l'arbre retenu à la première étape).

deux arbres et à comparer la distribution des erreurs absolues. Le coefficient de détermination R^2 est donné à titre indicatif car évalué sur les données d'apprentissage, il n'atteste pas de la capacité du modèle à généraliser. Les statistiques obtenues sont présentées dans Tab. 3.

Modèle	R^2	*Erreurs absolues*				
		Quantile 25%	*Médiane*	*Quantile 75%*	*Maximum*	*Somme*
Fonctions seules	23.3%	5.9	12.2	20.9	143.9	5.486
Fonctions + mots-clés	38.6%	3.4	10.4	20.3	135.6	5.328
Mots-clés seuls	30.8%	5.3	10.4	21.5	137.8	5.532

Tableau 3 : Statistiques de comparaison des différents modèles

Les résultats montrent une légère amélioration de la prédiction suite à l'introduction des mots-clés. En effet, la somme des écarts de la prédiction à la valeur réelle ainsi que les principaux quantiles sont inférieurs pour le deuxième modèle. Par ailleurs, le pouvoir prédictif du modèle avec mots-clés seuls semble comparable à celui du modèle avec fonctions seules. Cela peut s'avérer utile dans les cas où l'on ne dispose pas de l'information sur la fonction recherchée (c'est le cas sur certains job boards), les mots-clés pourront alors éventuellement se substituer au champ fonction dans un modèle.

La figure 3 présente l'arbre obtenu avec la totalité des prédicteurs (fonctions et mots-clés). Pour des raisons de lisibilité, une seule règle de partition est affichée au-dessus de chaque nœud : elle s'applique à la branche gauche. Le reste des annonces est affecté à la branche droite.

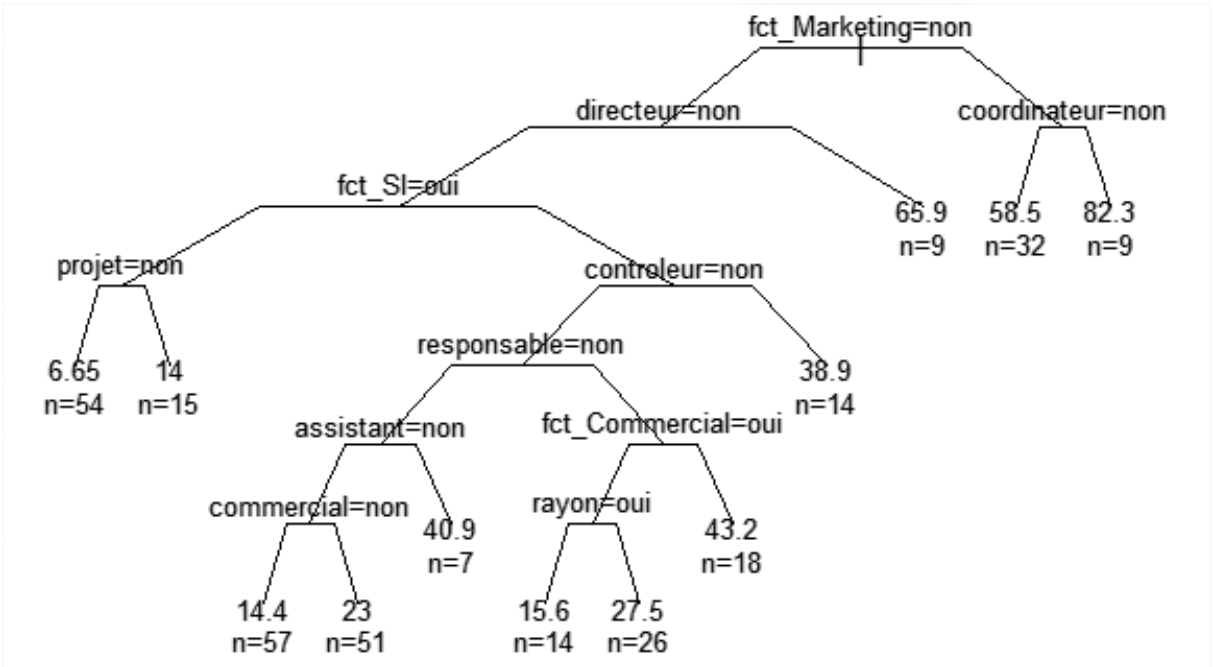

Figure 3 : Arbre obtenu pour la modélisation du rendement avec tous les prédicteurs

La première partition des données sépare les offres d'emploi à pourvoir dans la fonction *Marketing* des autres. En effet, les offres proposées dans cette fonction génèrent un nombre de candidatures largement supérieur la moyenne globale de 27.2 CV. Le mot-clé *directeur* apparaît au deuxième niveau de l'arbre, avant une seconde partition sur les catégories de fonctions. Ainsi, il y a un rendement moyen de 65.9 CV pour les offres des fonctions *SI*, *Commercial*,

et *Gestion* mentionnant le mot *directeur* dans le titre. Une tension apparaît pour les offres dans la fonction *SI* ne faisant pas référence à un poste de *directeur* avec un rendement moyen très faible (6.6 ou 14 CV en moyenne selon que le mot *projet* est mentionné ou non). L'arbre permet de mettre en évidence des interactions entre les mots-clés, à savoir ici des combinaisons de mots ayant des rendements particuliers. Par exemple, le nombre de CV moyen est de 15.6 pour un *responsable* de *rayon* (dans la fonction *Commercial*) tandis qu'il est de 43.2 pour un *responsable* dans la fonction *Gestion*.

5. Conclusion et perspectives

Nous avons présenté une méthode pour détecter des mots-clés dans le titre des annonces d'emploi et tester leur influence sur le volume des retours obtenus. Une application réalisée sur un corpus de 709 offres d'emploi nous a permis de mettre en évidence des mots-clés pertinents apportant un complément d'information par rapport à la fonction seule et permettant d'expliquer une plus grande partie de la variation du rendement des offres.

Nous n'avons cependant étudié qu'un aspect très restreint de la problématique de l'analyse de la performance des offres d'emploi. De nombreux autres facteurs contribuent à l'explication du rendement (caractéristiques du poste, du recruteur, etc.) et devraient être pris en compte dans un modèle dont le but serait la prédiction. Suite à ces résultats encourageants, nous renouvellerons cette analyse sur un corpus de taille plus importante [9] puis introduirons les mots-clés dans un modèle à but prédictif faisant appel à des facteurs variés (comme évoqué ci-dessus). Cela nous permettra à la fois de confirmer leur impact sur le rendement et d'améliorer le pouvoir prédictif de notre modèle. Nous projetons également d'établir un processus de recommandations sur les mots-clés à intégrer dans le titre afin d'accroître le rendement dans le cas de profils rares ou de le réduire dans le cas de profils trop génériques. Pour cela, il nous faudra être capable d'identifier les mots-clés adaptés à un poste donné et mesurer l'effet de leur présence ou absence sur la performance de l'annonce diffusée. Une taille de corpus importante est nécessaire, ainsi qu'une analyse du texte complet de l'offre.

Références

Aureli E. and Iezzi D.F. (2006). Recruitment via web and information technology : a model for ranking the competences in job market. In *Proceedings of JADT 2006*, pp. 79-88.

Breiman L., Friedman J., Olshen R.A. and Stone C.J. (1984). *Classification and regression trees*. London : Chapman & Hall/CRC.

Fondeur Y. and Tuchszirer C. (2005). *Internet et les intermédiaires du marché du travail*. Rapport IRES.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, vol. 1 : 127-165.

Lebart L. and Salem A. (1994). *Statistique textuelle*. Paris : Dunod.

Lebart L. (2007). Which bootstrap for principal axes methods ? In Brito, P., Cucumel, G., Bertrand, P. and de Carvalho, F., editors, *Selected Contributions in Data Analysis and Classification*. Heidelberg-Berlin : Springer, pp. 581-588.

[9] Le corpus dont nous disposons, lié à l'activité de Multiposting.fr, s'enrichit au cours du temps de nouvelles offres d'emploi diffusées sur Internet.