Fouille de données textuelles et systèmes de recommandation appliqués aux offres d'emploi diffusées sur le web

Julie Séguéla

CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS | le cnam

École Doctorale Informatique, Télécommunication et Électronique

Laboratoire CEDRIC du CNAM - Équipe MSDMA

THÈSE DE DOCTORAT

présentée par : **Julie SEGUELA**

soutenue le : **3 mai 2012**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline / Spécialité : **Informatique / Data mining & apprentissage**

Fouille de données textuelles et systèmes de recommandation appliqués aux offres d'emploi diffusées sur le web

THÈSE DIRIGÉE PAR

M. SAPORTA Gilbert — *Professeur, Conservatoire National des Arts et Métiers*

RAPPORTEURS

M. LEBART Ludovic — *Directeur de recherches, CNRS, TELECOM-ParisTech*
M. VIENNET Emmanuel — *Professeur, Université Paris 13*

EXAMINATEURS

M. ARTIERES Thierry — *Professeur, Université Pierre et Marie Curie*
M. CRUCIANU Michel — *Professeur, Conservatoire National des Arts et Métiers*
M. FONDEUR Yannick — *Chercheur, Centre d'Études de l'Emploi*
M. LECHEVALLIER Yves — *Directeur de recherches, INRIA*
M. LE VIET Stéphane — *Directeur associé, Multiposting*

Remerciements

Je souhaite tout d'abord remercier mon directeur de thèse, le professeur Gilbert Saporta, pour sa disponibilité tout au long de ces trois années, ses conseils avisés, son ouverture d'esprit, et pour avoir su me donner confiance aux moments cruciaux de ce projet.

En avril 2009, Stéphane Le Viet et Gautier Machelon, co-fondateurs de la société Multiposting, m'ont fait confiance pour mener à bien ce projet. Je souhaite les remercier de m'avoir permis de vivre cette expérience, de laquelle je sors grandie tant sur le plan professionnel que sur le plan humain. En arrivant il y a trois ans, je n'imaginais pas apprendre autant sur la vie d'entreprise et les relations humaines.

Je remercie chaleureusement mes deux rapporteurs, Emmanuel Viennet et Ludovic Lebart, pour le temps qu'ils ont consacré à l'étude de mes travaux et leurs précieuses critiques et remarques. Je remercie Michel Crucianu et Yannick Fondeur, pour leurs conseils et le temps qu'ils m'ont accordé tout au long de ces trois années, depuis les comités de thèse jusqu'à la soutenance finale. Je remercie également Thierry Artières et Yves Lechevallier, pour avoir accepté de faire partie du jury, le temps qu'ils ont consacré à mon manuscrit et le regard neuf qu'ils y ont apporté.

J'exprime toute ma gratitude à Alizée, Elise, Elsa, Matthieu, l'autre Matthieu, Julien, Willy, Oliver, Virginie et Marie qui n'ont cessé de m'encourager durant ces trois années. Vous avez su me redonner le sourire dans les moments difficiles. Merci Ndeye pour tous ces bons moments. Une pensée pour Anne, tiens bon car le jeu en vaut la chandelle.

Emilie, Tila, merci d'avoir accompagné mes journées de votre bonne humeur et de votre générosité. NC, je garde en mémoire ces innombrables soirées passées devant nos écrans, tes innombrables coups de main spontanés, merci pour ta générosité. Gui, merci pour ta disponiblité, ta gentillesse, ton écoute et ta patience ! Tous les deux, merci pour vos relectures et merci d'avoir été là tout simplement.

Simon, depuis qu'on se connaît tu as toujours été à mon écoute, dans les bons moments comme dans les moments difficiles. Merci pour ta présence, ta patience, ton soutien et ta compréhension durant ces derniers mois de thèse.

Enfin, un grand merci à ma petite famille, qui m'a soutenue depuis le début et qui a toujours fait en sorte de me rendre les choses plus faciles. Sachez que pendant tout ce temps, je n'ai cessé de penser à vous malgré la distance...

Résumé

L'expansion du média Internet pour le recrutement a entraîné ces dernières années la multiplication des canaux dédiés à la diffusion des offres d'emploi. Dans un contexte économique où le contrôle des coûts est primordial, évaluer et comparer les performances des différents canaux de recrutement est devenu un besoin pour les entreprises. Cette thèse a pour objectif le développement d'un outil d'aide à la décision destiné à accompagner les recruteurs durant le processus de diffusion d'une annonce. Il fournit au recruteur la performance attendue sur les sites d'emploi pour un poste à pourvoir donné. Après avoir identifié les facteurs explicatifs potentiels de la performance d'une campagne de recrutement, nous appliquons aux annonces des techniques de fouille de textes afin de les structurer et d'en extraire de l'information pertinente pour enrichir leur description au sein d'un modèle explicatif. Nous proposons dans un second temps un algorithme prédictif de la performance des offres d'emploi, basé sur un système hybride de recommandation, adapté à la problématique de démarrage à froid. Ce système, basé sur une mesure de similarité supervisée, montre des résultats supérieurs à ceux obtenus avec des approches classiques de modélisation multivariée. Nos expérimentations sont menées sur un jeu de données réelles, issues d'une base de données d'annonces publiées sur des sites d'emploi.

Mots clés : Fouille de textes, extraction des connaissances, systèmes de recommandation, offres d'emploi, recrutement sur Internet

Abstract

Last years, e-recruitment expansion has led to the multiplication of web channels dedicated to job postings. In an economic context where cost control is fundamental, assessment and comparison of recruitment channel performances have become necessary. The purpose of this work is to develop a decision-making tool intended to guide recruiters while they are posting a job on the Internet. This tool provides to recruiters the expected performance on job boards for a given job offer. First, we identify the potential predictors of a recruiting campaign performance. Then, we apply text mining techniques to the job offer texts in order to structure postings and to extract information relevant to improve their description in a predictive model. The job offer performance predictive algorithm is based on a hybrid recommender system, suitable to the cold-start problem. The hybrid system, based on a supervised similarity measure, outperforms standard multivariate models. Our experiments are led on a real dataset, coming from a job posting database.

Keywords : Text mining, knowledge discovery, recommender systems, job postings, e-recruitment

Table des matières

Introduction **21**

1 Le marché du recrutement sur Internet **29**

1.1 Internet et le marché du recrutement 29

1.1.1 Présentation des acteurs du marché 29

1.1.2 Des besoins d'amélioration pour la procédure de recherche des candidats 30

1.2 Problématiques associées à la recherche de candidats sur Internet 32

1.3 Présentation de Multiposting.fr et positionnement 35

1.3.1 Processus de diffusion d'une offre d'emploi via Multiposting.fr 35

1.3.2 Intérêts de l'utilisation de la plate-forme de diffusion Multiposting.fr 36

1.3.3 Les axes d'amélioration identifiés pour l'outil 36

1.4 Comparatif des solutions concurrentes 37

2 Indicateurs de performance d'une campagne de recrutement **45**

2.1 Évaluation de la performance d'une campagne de recrutement : état de l'art 45

2.1.1 Indicateurs post-embauche 46

2.1.2 Indicateurs pré-embauche 46

2.1.3 Apparition d'Internet et évolution des indicateurs de performance . . 47

2.2 Proposition d'indicateurs de performance 50

2.2.1 Présentation des données enregistrées par l'outil 51
2.2.2 Indicateurs de performance proposés 53
2.2.3 Indicateur de la performance relative 55
2.2.4 Discussion 56
2.3 Synthèse 57

3 Les facteurs explicatifs potentiels de la performance d'une campagne de recrutement 61

3.1 Les facteurs explicatifs de la performance d'une campagne dans la littérature 61
3.1.1 Le message transmis 62
3.1.2 Le type de poste proposé 63
3.1.3 Le recruteur 63
3.1.4 Le job board 64
3.1.5 Le calendrier 64
3.2 Propositions de facteurs explicatifs 65
3.2.1 Processus de candidature et facteurs explicatifs 65
3.2.2 Les facteurs inflexibles 67
3.2.3 Les facteurs flexibles 73
3.3 Synthèse 84

4 Fouille de textes appliquée aux offres d'emploi et extraction des connaissances 85

4.1 Présentation d'une offre d'emploi sur Internet 85
4.2 État de l'art 88
4.2.1 État de l'art des techniques de fouille de textes 88
4.2.2 Aperçu des travaux effectués sur les offres d'emploi 98
4.3 Classification des offres d'emploi en fonction du poste proposé 99

4.3.1 Objectifs poursuivis 100
4.3.2 Les tables de correspondance entre nomenclatures 100
4.3.3 Algorithmes de classification 103
4.3.4 Évaluation d'un système de classification : critères de performance . 104
4.3.5 Approche proposée 105
4.3.6 Expérimentations 110
4.3.7 Conclusions sur le système de catégorisation 121
4.4 Extraction de mots-clés pertinents 122
4.4.1 Extraction de prédicteurs candidats pour contribuer à l'explication de la performance des offres 122
4.4.2 Sélection des mots-clés à introduire dans le modèle de prédiction . . 123
4.5 Synthèse 126

5 Modélisation de la performance d'une offre d'emploi 127

5.1 Introduction 127
5.1.1 Contexte 127
5.1.2 Complexité des données et problématiques rencontrées 128
5.2 Systèmes de recommandation 129
5.2.1 Aperçu de l'état de l'art 129
5.2.2 Application innovante et cas particulier de système de recommandation 133
5.3 Modélisation de la performance d'une annonce diffusée sur un site d'emploi 134
5.3.1 Approches standards 135
5.3.2 Système hybride de recommandation 137
5.4 Expérimentations 141
5.4.1 Description des données 141
5.4.2 Comparaison des résultats 142

5.4.3 Enrichissement de la description des annonces 146
5.4.4 *Relevance feedback* 147
5.5 Illustration des résultats et discussion 147
5.6 Synthèse 150

6 Applications pour Multiposting.fr **153**
6.1 Les données 153
6.1.1 Présentation de la base de données 153
6.1.2 Enrichissement de la base de données existante 157
6.2 Impact des facteurs explicatifs et interprétation 162
6.2.1 Contribution des facteurs à la prédiction de la performance 162
6.2.2 Interprétation de l'impact des facteurs 164
6.3 Processus de diffusion d'une offre d'emploi : accompagnement et aide à la décision 166
6.3.1 Processus initial de diffusion d'une offre d'emploi 166
6.3.2 Nouvelles fonctionnalités 169
6.3.3 Mise en application des résultats 170

Conclusion et perspectives **177**

Bibliographie **179**

Annexes **193**

A Nomenclature "fonction" des offres d'emploi **195**

Glossaire **199**

Liste des tableaux

1.1 Comparaison des solutions de recrutement concurrentes 43

3.1 Indicateurs de conjoncture fournis par l'INSEE 68

3.2 Indicateurs de conjoncture fournis par le Pôle Emploi / la DARES 69

3.3 Indicateur de conjoncture fourni par *Keljob.com* 69

3.4 Indicateur de conjoncture fourni par *Apec.fr* 69

3.5 Indicateur de conjoncture fourni par *Monster.fr* 69

3.6 Indices d'image et d'attractivité des entreprises 72

4.1 Analyse sémantique latente et analyse des correspondances : comparaison . 97

4.2 Exemple de confrontation de deux nomenclatures "fonction" issues de deux sites d'emploi généralistes 102

4.3 Répartition des annonces entre les catégories de fonctions 110

4.4 Résultats obtenus au sein de chaque fonction pour la prédiction des métiers 119

4.5 Distribution des fonctions observées sur les fonctions prédites (% en colonne, les valeurs inférieures à 1% n'apparaissent pas) 119

4.6 Principales confusions entre métiers au sein de différentes fonctions : métier observé → métier prédit (fonction si différente) 120

4.7 Comparaison des termes retenus avec chaque méthode de sélection (50% des termes sont conservés à partir du score obtenu) 124

4.8 Liste des 20 termes présents dans le titre ayant les contributions les plus fortes selon le VIP-75% et valeurs associées 125

4.9 Liste des 20 termes présents dans le descriptif ayant les contributions les plus fortes selon le VIP-75% et valeurs associées 125

5.1 Résultats obtenus avec la régression PLS ($S1$) 143

5.2 Valeur(s) retenue(s) ($\times\sigma_d$) pour les fonctions gaussienne et exponentielle dans les approches $S2$ et $S3$ 143

5.3 Forces et faiblesses des approches proposées 148

5.4 MAE et $\overline{MAE}$ obtenus avec les différentes approches pour les sites d'emploi étudiés (* : estimation par le recommandeur moyen) 149

6.1 Contributions (%) des groupes de variables à la prédiction de la performance, et score de contribution moyenne indiqué entre parenthèses (* : nombre moyen de variables retenues) 163

6.2 Facteurs explicatifs ayant les plus forts impacts, valeurs du VIP et signes des coefficients associés pour deux sites d'emploi 165

A.1 Liste des fonctions et sous-fonctions de la nomenclature finale des offres d'emploi : "Architecture, Création", "Services administratifs" et "BTP" . . . 195

A.2 Liste des fonctions et sous-fonctions de la nomenclature finale des offres d'emploi : "Commercial / Vente", "Stratégie & Management", "Édition & Écriture", "Ingénierie & Recherche", "Comptabilité & Finance" et "Gestion de projet" 196

A.3 Liste des fonctions et sous-fonctions de la nomenclature finale des offres d'emploi : "Hôtellerie, Restauration", "Juridique", "Logistique & Transport", "Marketing", "Installation & Maintenance", "Production & Opérations", "Qualité / Inspection", "Formation / Éducation", "Ressources Humaines" et "Santé" 197

A.4 Liste des fonctions et sous-fonctions de la nomenclature finale des offres d'emploi : “Informatique & Technologies”, “Sécurité”, “Services clientèle” et “Autres” 198

Table des figures

1.1 Acteurs du marché du recrutement et portails de diffusion 30

1.2 Rôle des portails de diffusion dans le processus de recrutement 31

1.3 Part des postes à pourvoir ayant donné lieu à la publication d'une offre sur Internet en 2006 et 2009 (sur 100 recrutements cadre, source : étude APEC) 32

1.4 Problématiques du e-recrutement et interactions 34

1.5 Étapes de la diffusion d'une offre d'emploi avec Multiposting.fr 35

2.1 Processus de recherche d'emploi et de candidature sur Internet 48

2.2 Illustration des différents taux de conversion envisageables 50

2.3 Actions de candidature enregistrées dans l'outil Multiposting.fr 52

2.4 Chronologie du processus de recrutement et indicateurs de performance . . 58

2.5 Structure des indicateurs de performance issus des données enregistrées . . . 59

3.1 Représentation schématique du processus de candidature et intervention des facteurs potentiels 66

3.2 Répartitions des annonces postées (candidature par e-mail) et CV reçus en fonction du créneau horaire 78

3.3 Répartitions des annonces postées (candidature par URL) et clics de redirection en fonction du créneau horaire 78

3.4 Effectifs des annonces postées (candidature par e-mail ou URL) en fonction du créneau horaire 79

3.5 Retours journaliers moyens par annonce en fonction du nombre de jours de diffusion et du créneau horaire de diffusion 80

3.6 Répartitions des annonces postées et candidatures en fonction du jour de la semaine 81

3.7 Effectifs des annonces postées (candidature par e-mail ou URL) en fonction du jour de la semaine 82

3.8 Retours journaliers moyens par annonce en fonction du nombre de jours de diffusion et du jour de diffusion 83

4.1 Exemple d'offre sur le site d'emploi *Monster.fr* 87

4.2 Vue d'ensemble du processus de préparation des textes 88

4.3 Vue d'ensemble du système de catégorisation 109

4.4 Processus d'évaluation de l'erreur dans le système de catégorisation 109

4.5 Matrice des corrélations sur les 100 premiers axes issus de l'AC et de la LSA (la couleur du pixel indique le dégré de corrélation entre les axes correspondants : de bleu foncé pour une forte corrélation à blanc pour une corrélation nulle) 112

4.6 Qualité de la classification en fonction de la méthode de représentation du texte et de la mesure de dissimilarité entre documents 113

4.7 Qualité de la classification en fonction du nombre de termes conservés et de la méthode de sélection (représentation TF) 115

4.8 Qualité de la classification en fonction de la méthode de représentation et du nombre de dimensions conservées 116

4.9 Représentation des 23 catégories de fonctions dans le plan rappel × précision (taille des bulles proportionnelle à l'effectif de la catégorie) 117

5.1 Nombre cumulé moyen de CV reçus au cours de la vie d'une annonce sur un site d'emploi 129

5.2 Nombre journalier moyen de CV en fonction du nombre de jours de diffusion sur un site d'emploi 130

5.3 Les sites d'emploi représentés sur les plans (nombre d'annonces, écart-type du nombre de CV reçus) et (nombre d'annonces, nombre moyen de CV reçus) 134

5.4 Vue d'ensemble du système hybride de recommandation 137

5.5 $\overline{MAE}$ obtenu avec les systèmes $S2$ et $S3$, en fonction du paramètre de variance (écart-type "e.t.") dans les fonctions de similarité gaussienne et exponentielle (représentation TF) 144

5.6 $\overline{MAE}$ obtenu avec les systèmes $S2$ et $S3$, en fonction de la méthode de représentation du texte et de la mesure de similarité 145

5.7 Comparaison des meilleurs algorithmes de chaque approche (la fonction de similarité retenue est indiquée entre parenthèses) 146

5.8 $\overline{MAE}$ obtenu avec les descripteurs du texte seuls et avec l'ajout de variables descriptives (la fonction de similarité retenue est indiquée entre parenthèses) 147

5.9 $\overline{MAE}$ obtenu avec ou sans *relevance feedback* 148

5.10 *Généraliste 4* : représentation des résultats obtenus pour l'échantillon de test sur le plan engendré par les deux premières composantes PLS. Figure de gauche : rendement journalier réel (taille du cercle proportionnelle à la valeur). Figure de droite : lissage à partir des valeurs prédites par $S3$, courbes de niveau associées et rendement journalier réel. 151

6.1 Répartition des annonces Multiposting entre les différentes offres proposées 154

6.2 Évolution du nombre d'annonces multidiffusées (offre classique) et du nombre moyen de supports utilisés (hors écoles et associations d'anciens) 154

6.3 Répartition des annonces Multiposting selon le nombre de diffusions 155

6.4 Proportion d'annonces avec rediffusion, ajout de site(s), et recours aux écoles 155

6.5 Nombre moyen de supports utilisés pour une annonce multidiffusée (avec ou sans recours aux écoles) 156

6.6 Répartition des annonces selon le type de recruteur et le secteur d'activité des entreprises 157

6.7 Répartition des annonces selon le type de contrat, le niveau d'études requis, le niveau d'expérience requis et la région administrative 158

6.8 Diffusion d'une offre d'emploi via l'interface Multiposting.fr : étape 1 167

6.9 Diffusion d'une offre d'emploi via l'interface Multiposting.fr : étape 2 168

6.10 Nouveau processus de diffusion d'une offre d'emploi : étape 1 171

6.11 Nouveau processus de diffusion d'une offre d'emploi : étape 2 172

6.12 Nouveau processus de diffusion d'une offre d'emploi : étape 3 173

Introduction

Contexte et objectifs des travaux

Depuis les deux dernières décennies, l'utilisation d'Internet pour le recrutement s'est considérablement développée. La démocratisation d'Internet a entraîné un accroissement simultané du nombre de canaux de recrutement et du volume de personnes pouvant être atteintes par ce média-là. Nos travaux s'inscrivent dans le cadre du recrutement via les canaux Internet, et en particulier les sites web de recherche d'emploi (job board) en France. Les recruteurs en recherche de main d'oeuvre ont à disposition un grand nombre de supports web pour diffuser leurs offres d'emploi [1] : sites généralistes (*Monster.fr*, *Apec.fr*, *Pôle Emploi*, etc.), sites spécialisés (*eFinancial* spécialisé dans les métiers de la Finance, *Les Jeudis* spécialisé dans la fonction Informatique, *L'Étudiant* spécialisé dans les stages et emplois étudiants, etc.), blogs (*Developpez.com*, etc.), réseaux sociaux (*Viadeo*, *LinkedIn*, etc.), sites web d'écoles et associations d'anciens élèves. Parmi les sites généralistes les plus populaires auprès des recruteurs, seuls les sites emploi de l'APEC [2] et de Pôle Emploi [3] permettent une diffusion gratuite des offres. La diffusion des offres est également payante sur une grande partie des sites spécialisés. Pour une entreprise, le coût annuel des recrutements peut donc être très élevé. En conséquent, il est devenu indispensable pour les recruteurs d'évaluer et d'analyser les performances des différents supports utilisés, afin de pouvoir choisir objectivement les supports à utiliser lors de la diffusion d'une offre d'emploi. La performance d'une offre d'emploi est généralement mesurée par le nombre de candidatures reçues en réponse à cette offre.

1. Dans le document, les termes "offre d'emploi" et "annonce d'emploi" seront utilisés indifféremment pour faire référence au descriptif d'un poste à pourvoir publié sur un site d'emploi.
2. Association Pour l'Emploi des Cadres (`www.apec.fr`)
3. `www.pole-emploi.fr`

Aujourd'hui, un certain nombre d'outils propriétaires sont mis à disposition des entreprises afin de faciliter le processus de recrutement, depuis la mise en ligne de l'annonce jusqu'à la gestion des candidatures reçues. Dans cette thèse, nous nous intéressons à l'étape de diffusion de l'annonce, et à l'accompagnement du recruteur au cours de cette phase. Les principaux acteurs du marché des solutions de diffusion d'annonces fournissent des outils pour l'analyse de la performance de ces dernières. Cependant, la plupart de ces outils sont limités en ce qui concerne l'aide à la décision, car ils se concentrent sur l'analyse de la performance obtenue à l'issue de la diffusion de l'annonce (post-campagne de recrutement). Ces limites se justifient par l'existence de barrières au traitement et à l'analyse automatique des offres d'emploi. En effet, la multitude des sites d'emploi entraîne une multitude de structures spécifiques à ces derniers. Il n'existe pas aujourd'hui de structure uniforme admise par l'ensemble des acteurs du domaine des ressources humaines pour l'information contenue dans les offres d'emploi [4].

Dans ce contexte, nos travaux ont pour double objectif :

- l'analyse, par la structuration de l'information, des performances des offres d'emploi sur les supports Internet ;
- la mise au point d'un algorithme prédictif de cette performance.

Ces travaux donneront lieu au développement d'un outil d'aide à la décision destiné aux recruteurs, qui s'intégrera au cadre d'un outil propriétaire de multidiffusion d'annonces : *Multiposting.fr* [5]. Grâce à l'algorithme développé, nous pourrons fournir au recruteur une estimation de la performance attendue sur les différents supports lors de la diffusion d'une nouvelle offre d'emploi. Ses choix seront ainsi facilités, mais il sera également averti sur le nombre approximatif de candidatures qu'il peut attendre. Pour mener à bien cet objectif, notre approche nécessitera l'automatisation des procédés utilisés d'un point de vue global.

À notre connaissance, il n'existe pas dans la littérature de corpus d'offres d'emploi pouvant être exploité librement. Par ailleurs, ces travaux étant menés dans le cadre d'une

4. Le consortium HR-XML (`http://www.hr-xml.org/`) vise à promouvoir l'échange de données relatives à la gestion des ressources humaines au niveau mondial, notamment par la promotion d'un vocabulaire standard XML. Cependant, il n'est pas adopté par l'ensemble des acteurs du domaine des ressources humaines.

5. `www.multiposting.fr`

convention CIFRE [6] pour répondre aux besoins de la société Multiposting, nous mènerons nos expérimentations sur une extraction de la base de données détenue par la société.

Problématiques rencontrées

Pour répondre aux objectifs évoqués précédemment, nous sommes confrontés à des problématiques liées à la spécificité des données que nous traitons. Nous avons à disposition un historique d'offres d'emploi publiées sur des sites, stocké sous forme de base de données. Les données enregistrées sont les informations sur les offres et le nombre de candidatures obtenues sur les différents supports utilisés. Pour répondre au principal objectif de la thèse, il nous faut identifier l'ensemble des facteurs explicatifs potentiels de la performance d'une annonce d'emploi. Toutes les informations souhaitées n'étant pas disponibles en base de données, nous devons réaliser des traitements et des transformations sur les données initiales, et avoir recours à des données provenant de sources externes.

De plus, parmi les informations disponibles sur les offres, certaines sont structurées et d'autres non structurées. Les informations structurées concernent les caractéristiques générales de l'offre comme le type de contrat, le niveau d'études requis ou encore l'expérience souhaitée. Les informations non structurées font référence au descriptif de l'annonce d'emploi, rédigé sous forme d'un texte libre. Au sein de l'algorithme prédictif, nous devrons exploiter simultanément ces données structurées et non structurées. Par ailleurs, le traitement de données textuelles implique de travailler sur des données de très grande dimension (plusieurs milliers de descripteurs).

Enfin, nous devrons également être attentifs aux problématiques liées à la dimension temporelle des annonces. En effet, la performance finale est déterminée par un flux de candidatures reçues durant la période de présence en ligne de l'annonce.

6. Les conventions CIFRE (conventions industrielles de formation par la recherche) sont financées par le ministère de l'Enseignement supérieur et de la Recherche qui en a confié la mise en œuvre à l'ANRT (Association nationale de la recherche et de la technologie).

Contributions

Les problématiques énoncées plus haut nous ont conduits à proposer une méthodologie adaptée à la complexité des données à traiter. L'approche que nous proposons permet en effet de gérer simultanément des données structurées et non structurées au sein d'un algorithme à but prédictif. Elle permet également de gérer un très grand nombre de variables explicatives, parfois largement supérieur au nombre d'observations.

Dans cette thèse, nous introduisons un algorithme prédictif qui peut être interprété comme un cas particulier de système de recommandation, à savoir un système où les recommandations doivent être faites dans un contexte de "démarrage à froid" (les items sont nouveaux et n'ont encore jamais été notés par un utilisateur). La problématique de démarrage à froid est encore aujourd'hui un thème de recherche actif dans la littérature. L'approche que nous proposons est un système hybride, permettant de répondre à cette problématique grâce à l'usage des données de contenu.

Une partie de nos contributions concernent l'analyse des offres d'emploi. Cette thèse fournit une revue de la littérature des facteurs pouvant avoir une influence sur la performance des offres d'emploi. L'analyse du processus de candidature sur Internet nous permet de proposer de nouveaux facteurs explicatifs. Notre application permet finalement de mettre en évidence l'impact d'une partie de ces facteurs.

Des techniques de fouille de textes sont comparées à travers différentes applications aux annonces d'emploi (catégorisation, prédiction). Nous mettons en évidence les techniques qui permettent d'obtenir les meilleurs résultats selon les objectifs poursuivis.

Nous proposons une méthode pour la structuration des offres d'emploi du point de vue du métier associé au poste grâce à une nomenclature établie au préalable. Nos expérimentations montrent l'existence de vocabulaires spécifiques aux différents métiers permettant une réduction considérable de la dimension du problème par la sélection des termes.

Enfin, ces travaux ont donné lieu au développement d'un module d'aide à la décision venant compléter la solution classique de multidiffusion proposée par la société Multiposting.

Organisation du document

Le chapitre 1 est une introduction au marché du recrutement sur Internet, à travers la présentation des différents acteurs qui le composent et des mécanismes qui le régissent. La solution de multidiffusion d'annonces Multiposting.fr y est également présentée, et un comparatif avec les principales solutions concurrentes est établi.

Dans le chapitre 2, nous présentons un aperçu des indicateurs de performance d'une campagne de recrutement cités dans la littérature. Nous introduisons un ensemble d'indicateurs apparus avec le développement d'Internet pour le recrutement, et mettons en évidence les interactions entre ces derniers. La présentation des statistiques enregistrées par la société Multiposting permet finalement de statuer sur l'indicateur de performance que nous étudierons.

Le chapitre 3 débute avec un état de l'art des facteurs explicatifs de la performance d'une campagne de recrutement, par la revue de la littérature du domaine du "management des ressources humaines". Nous proposons ensuite un ensemble de facteurs explicatifs en complément ou en alternative à ceux cités en début de chapitre, en nous focalisant davantage sur l'accessibilité des données dans la pratique. Ce chapitre présente également une étude sur l'impact du jour et de l'heure de diffusion de l'annonce.

Le chapitre 4 introduit les méthodes usuelles de la fouille de textes, ainsi qu'un aperçu des études menées spécifiquement sur les offres d'emploi. Nous avons ensuite recours à ces méthodes pour obtenir une structuration uniforme des offres d'emploi du point de vue du métier (ou de la fonction) proposé. Des expérimentations menées dans des cadres supervisés et non supervisés sont présentées. Enfin, nous proposons une méthode pour extraire l'information pertinente du texte (à travers un ensemble de mots-clés) afin d'enrichir l'ensemble des facteurs explicatifs utilisés en entrée de l'algorithme prédictif.

Le chapitre 5 est consacré à la modélisation de la performance d'une offre d'emploi. Après avoir exposé les problématiques liées à la complexité des données que nous traitons, nous introduisons le lecteur aux systèmes de recommandation et présentons notre problème en tant que cas particulier de système de recommandation. Nous proposons deux variantes d'un système hybride permettant de prédire la performance d'une offre sur un site d'emploi

donné. Ses résultats sont comparés dans le cadre d'expérimentations à ceux obtenus avec des approches standards de modèle multivarié. La flexibilité de notre approche nous permet d'améliorer la qualité des résultats à l'aide d'un système de "retour de pertinence".

Enfin, le jeu de données étudié est décrit dans le chapitre 6 à travers des statistiques descriptives. Nous y présentons également les données créées et obtenues à partir de sources externes pour enrichir la description des annonces. Les résultats obtenus sont illustrés à travers la contribution des facteurs explicatifs d'un point de vue global et sur des sites utilisés comme exemples. Le chapitre s'achève par la présentation du nouveau processus de multidiffusion d'une annonce avec l'outil Multiposting.fr.

Liste des publications

Conférences internationales

- J. Séguéla et G. Saporta. A semi-supervised hybrid system to enhance the recommendation of channels in terms of campaign ROI. In *CIKM'2011 : 20th ACM Conference on Information and Knowledge Management*, pages 2265-2268, octobre 2011, Glasgow, Royaume-Uni (communication poster).

- J. Séguéla et G. Saporta. A comparison between latent semantic analysis and correspondence analysis. *CARME'2011 : International conference on Correspondence Analysis and Related Methods*, février 2011, Rennes, France.

- J. Séguéla et G. Saporta. Automatic categorization of job postings. *COMPSTAT'2010, 19th International Conference on Computational Statistics*, août 2010, Paris, France (communication poster).

- J. Séguéla, G. Saporta et S. Le Viet. e-Recrutement : recherche de mots-clés pertinents dans le titre des annonces d'emploi. In *JADT'2010 : 10es Journées internationales d'Analyse statistique des Données Textuelles*, pages 975-982, juin 2010, Rome, Italie (communication poster).

Conférences nationales

- J. Séguéla. Système pour la catégorisation automatique des offres d'emploi en une typologie de fonctions. In *EGC'2011 : 11^{e} Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances*, RNTI-E-20, pages 515-526, janvier 2011, Brest, France, *Prix du meilleur article "jeune chercheur"*.
- J. Séguéla et G. Saporta. Modèles de comptage appliqués aux décisions de candidature aux offres d'emploi sur le web. *JDS'2010 : 42es Journées de Statistique*, mai 2010, Marseille, France.

Workshop

- J. Séguéla et G. Saporta. A hybrid recommender system to predict online job offer performance. *SDA'2011 : Theory and Application of High-dimensional Complex and Symbolic Data Analysis in Economics and Management Science*, octobre 2011, Pékin, Chine.

Article soumis dans une revue avec comité de lecture

- J. Séguéla et G. Saporta. *A hybrid recommender system to predict online job offer performance*. Revue des Nouvelles Technologies de l'Information, numéro spécial.

Chapitre 1

Le marché du recrutement sur Internet

1.1 Internet et le marché du recrutement

1.1.1 Présentation des acteurs du marché

Le marché du recrutement est composé de trois principaux acteurs :

- l'*entreprise* (ou recruteur), qui souhaite trouver le candidat correspondant le mieux au profil recherché ;
- le *candidat*, qui recherche un emploi adapté à son profil et à ses goûts ;
- les *intermédiaires*, qui interviennent sur la mise en relation des deux premiers acteurs.

Les intermédiaires du marché du travail peuvent intervenir de deux manières différentes [Fondeur et Tuchszirer 2005] : soit en tant que support d'information totalement neutre, soit en orientant l'offre et la demande (au moment de la définition du besoin ou au moment de la mise en relation). Les intermédiaires sur le marché du travail sont les cabinets de recrutement, les agences d'intérim, les agences de communication RH, la presse, les intermédiaires institutionnels (ANPE, Apec), etc. Depuis les deux dernières décennies, un autre type d'intermédiaire est apparu : les *job boards* (ou sites web de recherche d'emploi). D'une manière plus générale, de nombreux canaux permettent la publication d'offres d'emploi sur Internet : nous les appelons *portails de diffusion* (job boards, sites web d'écoles, réseaux sociaux, sites web carrière d'entreprises, etc.). Il existe donc deux types d'intermédiaires : les intermédiaires dits traditionnels et les portails de diffusion.

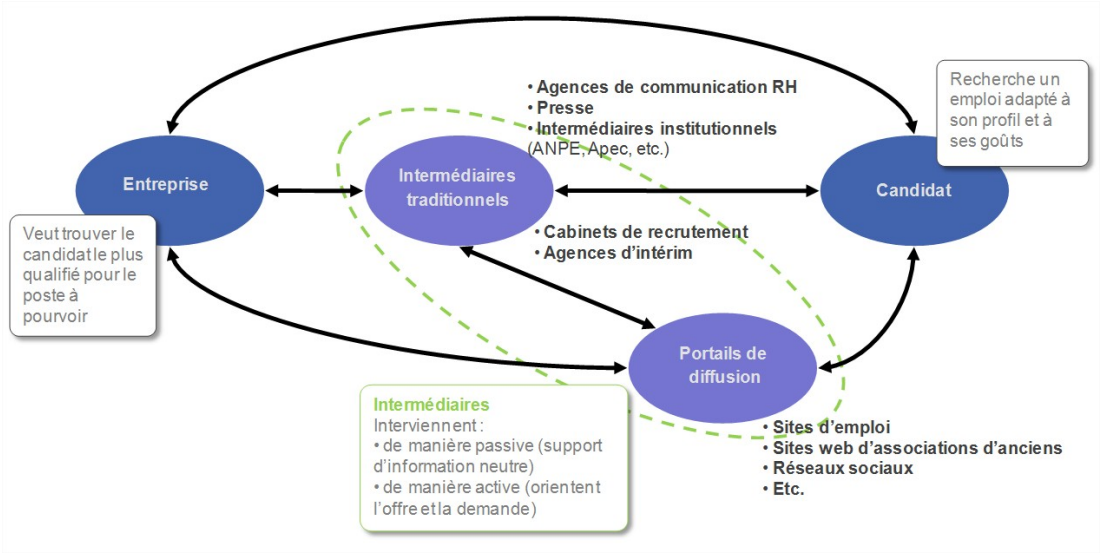

FIGURE 1.1 – Acteurs du marché du recrutement et portails de diffusion

Comme le montre la figure 1.1, certains intermédiaires traditionnels du marché de l'emploi peuvent interagir avec les portails de diffusion pour le compte d'une entreprise. En effet, l'entreprise peut confier son recrutement à un cabinet ou à une agence, qui pourra entre autres utiliser les portails de diffusion comme un moyen pour trouver des candidats correspondant au profil recherché. L'entreprise peut également faire le choix d'entrer directement en contact avec les portails de diffusion pour obtenir des candidatures.

La figure 1.2 présente plus en détail la nature des interactions entre : ceux que nous appellerons les recruteurs (entreprises, cabinets de recrutement ou agences d'intérim), les candidats et les portails de diffusion.

1.1.2 Des besoins d'amélioration pour la procédure de recherche des candidats

Une entreprise peut avoir recours à de nombreux moyens pour trouver des candidats à un poste : ANPE, Apec, presse écrite, Internet, forums et salons, cabinets de recrutement, cooptation, ou encore candidatures spontanées. Malgré la diversité des canaux utilisés (trois canaux ou plus dans 55% des recrutements), les entreprises éprouvent des difficultés pour recruter. En effet, parmi les procédures de recrutement ayant abouti, l'employeur estime tout de même que le recrutement a été difficile pour 30% des embauches sous contrat à

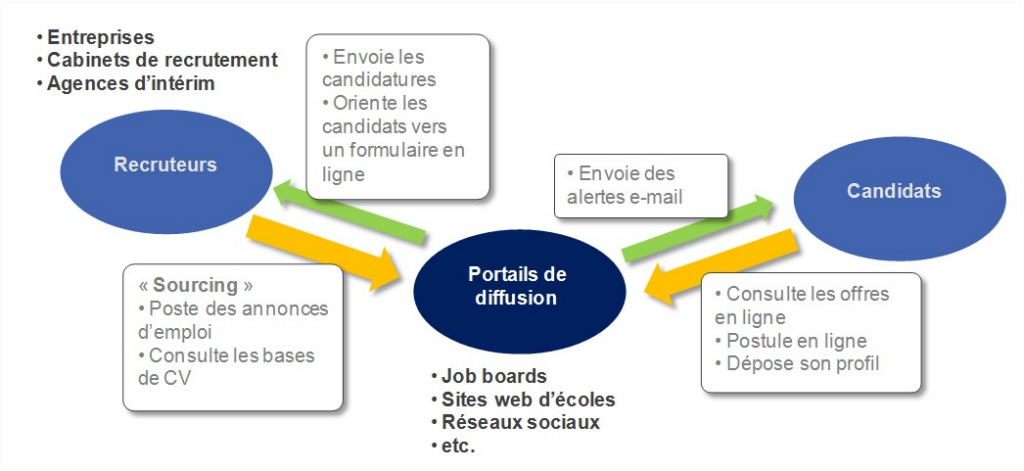

FIGURE 1.2 – Rôle des portails de diffusion dans le processus de recrutement

durée indéterminée (pour 20% des embauches sous contrat à durée déterminée). Et près d'une fois sur deux, l'employeur attribue la difficulté de recrutement à une pénurie de main-d'œuvre dans sa région (qui se traduit par un manque de candidats sur le poste à pourvoir) [1].

Pour les recrutements de cadres sous contrat CDI, c'est la diffusion d'une annonce sur Internet qui a permis d'approcher le candidat retenu dans 30% des cas où ce canal est utilisé. Ce chiffre peut paraître satisfaisant par rapport à celui des autres canaux : un taux d'efficacité [2] à 14% pour les candidatures spontanées, 39% pour l'APEC, ou encore 17% pour l'ANPE [3]. Toutefois, il reste suffisamment bas dans l'absolu pour soulever la question de la bonne utilisation des supports à disposition sur le canal Internet.

Depuis l'arrivée des job boards, l'utilisation du média Internet pour le recrutement ne cesse de se développer. Entre 2006 et 2009, la part des postes cadre à pourvoir ayant donné lieu à la publication d'une annonce sur Internet a augmenté de 16 points (cf. figure 1.3). En 2009, Internet s'avère être un média incontournable pour le recrutement avec 82% des offres d'emploi cadre qui y sont publiées. L'expansion du média Internet pour le recrutement a entrainé une multiplication des canaux permettant de trouver des candidats :

1. Les données citées dans ce paragraphe sont issues des résultats de l'enquête OFER, 2005 [voir Garner et Lutinier 2006a,b].
2. Nous entendons par taux d'efficacité la proportion de recrutements ayant effectivement abouti grâce au canal lorsque celui-ci est utilisé.
3. Voir note 1.

sites généralistes (ex. : Monster.fr, Apec.fr), sites spécialisés (ex. : FinancialCareers.fr, Lesjeudis.com), réseaux sociaux et blogs (ex. : Viadeo, Facebook), sites web d'écoles et d'associations d'anciens, CVthèques, etc. Aussi, il est de plus en plus difficile pour les recruteurs de faire un choix entre ces différents canaux, d'où la nécessité de pouvoir évaluer et comparer leurs efficacités respectives dans le cadre d'une campagne de recrutement.

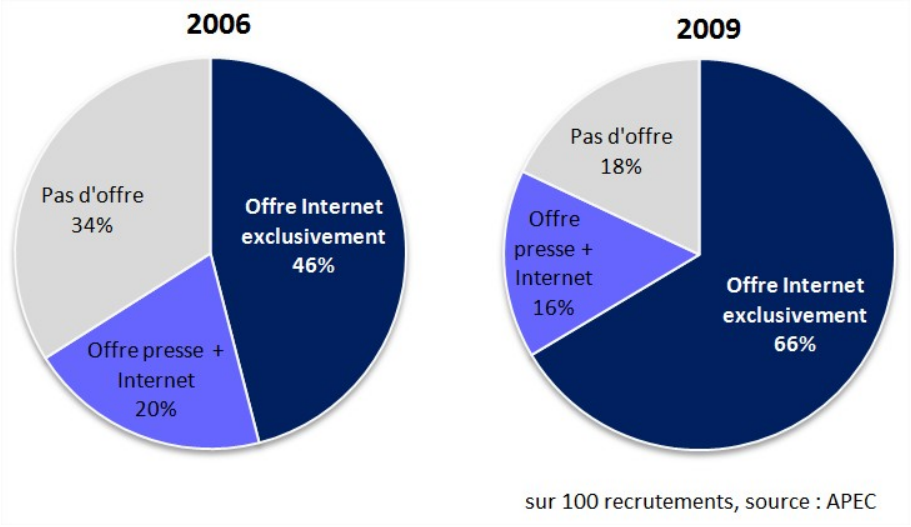

FIGURE 1.3 – Part des postes à pourvoir ayant donné lieu à la publication d'une offre sur Internet en 2006 et 2009 (sur 100 recrutements cadre, source : étude APEC)

1.2 Problématiques associées à la recherche de candidats sur Internet

Aujourd'hui, la principale préoccupation des recruteurs est de dénicher les meilleurs candidats pour un poste donné. Les problématiques sous-jacentes sont présentées ci-dessous.

Obtenir un volume important de retours. Pour avoir une chance de trouver de bons candidats, il faut recevoir un minimum de réponses à une annonce passée. Les recruteurs sont donc particulièrement attentifs au rendement des job boards qu'ils utilisent, c'est-à-dire au volume de candidatures qu'ils reçoivent en réponse à une annonce diffusée. Cependant, un gros volume de retours implique forcément un gros volume de "bruit" (candidatures non qualifiées pour le poste à pourvoir), et un travail de tri plus important. Toutefois, c'est en diffusant largement son annonce que le recruteur a des chances de ren-

contrer des candidats atypiques, avec un profil intéressant. En restreignant la visibilité de son annonce, le recruteur peut passer à côté d'un candidat qui correspondrait au poste.

Recevoir des candidatures qualifiées. Il est important de recevoir un grand nombre de CV, mais il est nécessaire qu'une bonne proportion de CV qualifiés y soit associée. Le bruit peut apparaître au niveau des annonces d'emploi (les candidats sont face à une abondance d'offres qu'ils ne peuvent pas toujours discriminer de manière pertinente), ou au niveau des candidatures. En effet, avec le développement du e-recrutement, il est de plus en plus facile de postuler en ligne (parfois un simple clic suffit pour envoyer un CV, donc non spécifique à l'offre choisie). De plus, dans une conjoncture difficile pour les chômeurs, ces derniers ont tendance à candidater à un maximum de postes, pas toujours adaptés à leur profil, car "on ne sait jamais".

Diminuer les temps de traitement des candidatures. Plus le volume de candidatures reçues est grand, plus les temps de traitement associés seront importants, et par suite les coûts qui y sont liés. Si la part de CV qualifiés est faible (et réciproquement la part de bruit élevée), les coûts "inutiles" seront d'autant plus forts.

Être visible auprès des candidats passifs et des profils rares. Les candidats en poste sont une des cibles privilégiées des recruteurs. Une des problématiques du recruteur consiste donc à être visible auprès de ces candidats au comportement passif, n'ayant pas de démarche active de recherche d'emploi. Pour cela, le choix des supports de diffusion de l'annonce est primordial. Il l'est d'autant plus lors d'une mauvaise conjoncture sur le marché de l'emploi car les individus en poste sont moins mobiles, de peur de se trouver au chômage au cas où l'expérience tournerait mal dans la nouvelle entreprise. Les annonces d'emploi doivent également être suffisamment visibles pour atteindre des profils rares ou atypiques, d'où l'importance du choix des supports de diffusion.

Optimiser le budget alloué au processus de recherche des candidats. Une problématique actuelle majeure des recruteurs est d'optimiser le rendement des annonces publiées, ainsi que les coûts liés au processus de recherche des candidats (qu'il s'agisse des

coûts liés à la publication, au traitement des retours, ou aux logiciels de recrutement).

D'une manière générale, on constate une volonté d'automatisation des processus liés au recrutement afin d'obtenir une diminution des coûts à différentes étapes. La figure 1.4 présente sous forme synthétique les problématiques évoquées ci-dessus et les interactions qui les lient.

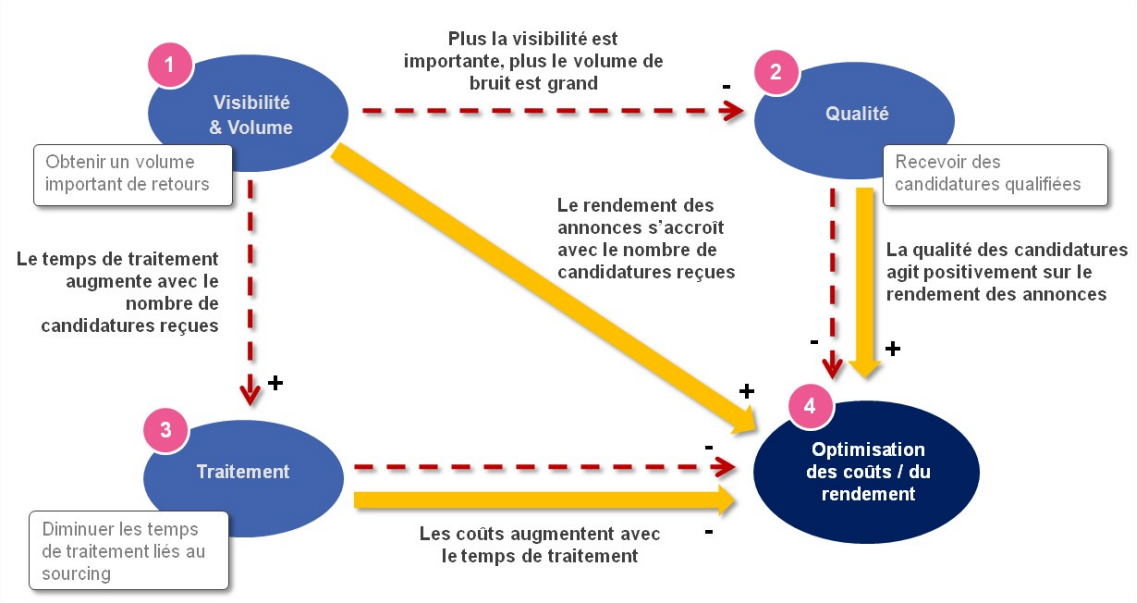

FIGURE 1.4 – Problématiques du e-recrutement et interactions

Augmenter le volume et la qualité des candidatures reçues permet d'agir positivement sur l'objectif d'optimisation des coûts. En revanche, l'augmentation des temps de traitement a une influence négative sur l'optimisation des coûts. De plus, l'augmentation de la visibilité et du volume agit indirectement et de manière négative sur le rendement des annonces. En effet, deux phénomènes en sont la cause :

- l'augmentation de la visibilité de l'annonce et du volume de candidatures reçues entraine une dégradation de la qualité globale des candidatures reçues et par suite une diminution du rendement ;
- l'augmentation du volume de candidatures reçues entraine une augmentation des temps de traitement et par suite une augmentation des coûts liés au processus de recrutement.

1.3 Présentation de Multiposting.fr et positionnement

Multiposting.fr est une plate-forme française de multidiffusion d'offres d'emploi sur Internet en activité depuis septembre 2008. En France [4], c'est la première technologie qui permet aux recruteurs de diffuser une annonce d'emploi sur un grand nombre de supports (sites d'emploi, sites d'écoles, blogs, réseaux sociaux, etc.) avec une seule saisie du contenu de l'annonce. Elle offre actuellement ses services à plus de 400 clients, principalement en France [Multiposting.fr 2011].

1.3.1 Processus de diffusion d'une offre d'emploi via Multiposting.fr

La figure 1.5 donne une représentation simplifiée des différentes étapes suivies par le recruteur lorsqu'il diffuse une annonce d'emploi via l'outil Multiposting.

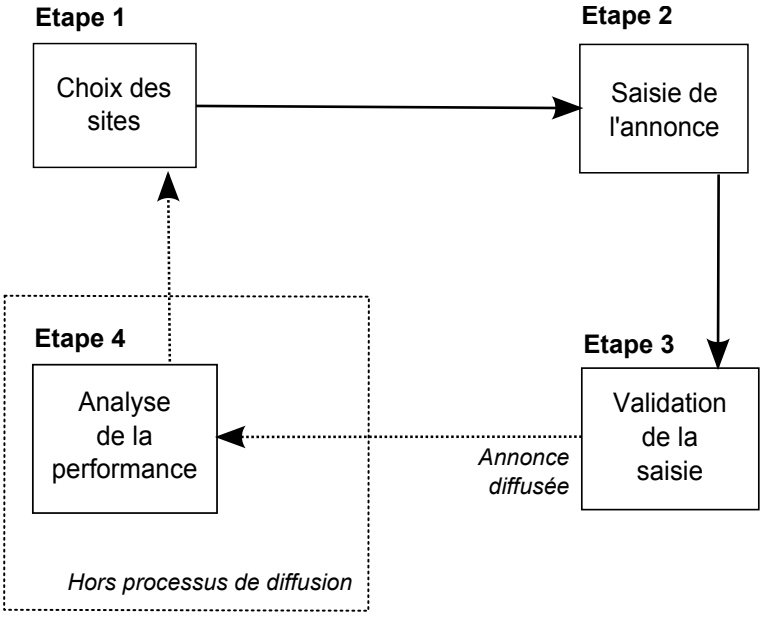

FIGURE 1.5 – Étapes de la diffusion d'une offre d'emploi avec Multiposting.fr

La première étape est consacrée au choix des sites : le recruteur coche sur l'interface les sites sur lesquels il souhaite diffuser son annonce. La deuxième étape est dédiée à la saisie de l'annonce (champs communs à remplir systématiquement et champs spécifiques dont

4. A l'international, deux concurrents sont identifiés au moment de la création de l'entreprise : *eQuest* (États-Unis) et *Broadbean* (Royaume-Uni).

la liste est déduite des sites sélectionnés à l'étape précédente). La troisième étape permet de vérifier les informations saisies et de valider la diffusion de l'annonce. A l'issue de cette étape, l'annonce est envoyée sur les différents sites choisis. Une quatrième étape (hors du processus de diffusion) peut être considérée : l'analyse de la performance des campagnes de recrutement menées. En effet, à la fin d'une campagne, le recruteur peut observer les rendements obtenus sur les différents sites et mettre à profit ses conclusions lors des futures campagnes.

1.3.2 Intérêts de l'utilisation de la plate-forme de diffusion Multiposting.fr

Grâce à ses fonctionnalités, Multiposting.fr constitue une aide pour les recruteurs visant à améliorer leur procédure de recrutement sur Internet à différents niveaux :

- Relativement à une procédure habituelle de recrutement, l'outil permet d'augmenter le nombre de candidatures reçues en fournissant l'accès à de nouveaux supports Internet gratuits non utilisés jusqu'à présent par le recruteur.
- L'outil permet d'augmenter la visibilité auprès des candidats qualifiés grâce à la possibilité de diffuser sur des sites spécialisés, des sites d'écoles ou d'associations d'anciens, ou encore des blogs choisis en adéquation avec le profil recherché.
- Poster une annonce sur un site d'emploi est une procédure chronophage, qui doit de plus être répétée autant de fois qu'il y a de sites utilisés. L'outil permet de réduire de manière importante le temps passé à poster l'annonce grâce à un processus simple et automatique de multidiffusion : une seule saisie suffit pour éventuellement plusieurs dizaines de sites utilisés.
- Enfin, Multiposting.fr propose le suivi de la performance des annonces et des sites utilisés (via le décompte des candidatures reçues). L'étude de ces résultats permet l'optimisation du budget alloué à la campagne de recrutement grâce à l'identification des sites ayant les meilleurs/moins bons rendements.

1.3.3 Les axes d'amélioration identifiés pour l'outil

La section précédente évoque la plus-value apportée par Multiposting.fr relativement à une utilisation classique des sites d'emploi. Cependant, ces fonctionnalités présentent certaines

limites qui permettent d'identifier des axes d'amélioration. En effet, que l'objectif soit de gagner en volume ou en qualité des candidats, le choix des sites à utiliser est laissé au recruteur, accompagné de l'expertise des commerciaux de Multiposting.fr. Aujourd'hui, ce choix n'est donc pas guidé par des critères objectifs, mais lié à des avis plus ou moins subjectifs, aux préférences des recruteurs. Bien que le système de multidiffusion propose le suivi de la performance des annonces diffusées et des sites utilisés, à nouveau, l'analyse et l'interprétation des résultats est laissée à la charge du recruteur qui ne dispose pas d'outil pour le guider dans ses futurs choix.

Les constats précédents mettent en évidence la nécessité de construire un outil d'aide à la décision, s'intégrant harmonieusement avec le système existant, et guidant le recruteur à travers les différentes étapes de la diffusion d'une annonce.

1.4 Comparatif des solutions concurrentes

Dans cette section, nous faisons un état de la concurrence des outils de recrutement intelligents disponibles sur le marché. Les informations reproduites ici sont celles communiquées par les sociétés concernées via leur site Internet, des communiqués de presse, etc., et sont donc limitées en fonction de la stratégie de communication de ces différentes sociétés. Chaque sous-section est dédiée à une société et à la présentation de l'outil (ou des outils) qu'elle propose. Nous nous concentrons sur l'étude de l'outil à travers ses fonctionnalités, ses avantages et inconvénients, et éventuellement ses divergences en termes de finalité par rapport à l'outil Multiposting.fr.

eQuest

La société eQuest se présente comme le leader mondial de la diffusion d'annonces d'emploi sur Internet [eQuest 2011a]. Cette société offre ses services à plus de 20 000 entreprises dans le monde et assure la distribution de plus de 250 millions d'annonces chaque année [eQuest 2011b]. eQuest propose des outils approfondis pour le suivi de la performance des campagnes de recrutement dans le but d'améliorer celle des prochaines campagnes. Deux outils sont proposés : *Chameleon* [eQuest 2011c] et *TRAQ24* [eQuest 2011d].

Chameleon est l'outil assurant la distribution des annonces d'emploi auprès des job boards, réseaux sociaux, associations d'élèves, etc. Quatre fonctionnalités sont disponibles :

- *JobTracker*. Cet outil permet de collecter et afficher en un seul rapport les statistiques de performance des annonces sur chaque job board utilisé (nombre d'affichages de l'offre, nombre de candidats ayant cliqué pour candidater, taux de conversion des affichages en clics pour candidater, coût par clic) et ainsi aider à allouer le budget aux job boards les plus performants. *JobTracker* permet d'évaluer et comparer l'efficacité des job boards sur la base d'une liste de critères (entité, catégorie, titre unique, poste, pays, ville).
- *Spendometer*. Lors de la diffusion d'une annonce, l'outil recherche dans la base de données les annonces diffusées pour des postes similaires dans la même localisation au cours des 30 derniers jours et identifie le job board pré-sélectionné par le recruteur ayant reçu les plus mauvaises (ou le moins de) candidatures. *Spendometer* recommande alors de désélectionner ce site afin de diminuer les coûts de recrutement.
- *Post Scheduler*. Il permet de planifier la diffusion de l'annonce à une date future et ainsi la faire apparaître en tête de liste des sites d'emploi au moment où les candidats potentiels sont les plus nombreux sur Internet.
- *Post Confirmation*. Informe de la mise en ligne et de la visibilité des annonces sur les sites d'emploi.

TRAQ24 est un outil avancé permettant le suivi en temps réel des métriques sur les différents job boards, depuis la visualisation des offres au recrutement. Sept fonctionnalités sont proposées :

- *Ticker*. Permet le suivi des statistiques de performance des annonces en temps réel (affichages et clics par heure depuis les dernières 24 heures).
- *Dashboard*. Une page résumant l'ensemble des statistiques sur les annonces postées à travers tableaux et graphiques. Pour une période donnée, l'ensemble des métriques suivies (nombre de postings, nombre d'affichages, nombre de candidatures, nombre de recrutements) sont visibles au sein d'un graphique pour le niveau d'agrégation choisi (année, trimestre, mois, semaine, jour). Il présente les répartitions des candidats selon le job board source, la fonction du poste, la localisation, le secteur et l'utilisateur.

On y trouve les fonctionnalités *My Leaderboard* et *eQuest Leaderboard*.

- *My Leaderboard*. Fournit au recruteur les taux de succès des différents job boards par rapport aux données de son entreprise. Permet d'analyser les résultats par localisation, fonction du poste, secteur et utilisateur. Indique également la provenance des candidats recrutés.
- *eQuest Leaderboard*. Cette fonctionnalité permet au recruteur de comparer les taux de succès de son entreprise à ceux des autres utilisateurs de *TRAQ24*, et constitue une aide pour la mise en place des prochaines campagnes de recrutement.
- *Track by unique job*. Permet d'analyser la performance des postes de manière individuelle sur chacun des job boards utilisés. Comme sur le *Dashboard*, les métriques sont visibles à l'aide d'un graphique à différents niveaux d'agrégation. Le recruteur peut comparer sa performance à celle des autres entreprises pour le même type de poste et la même localisation [5].
- *Google API mapping*. Permet de décompter les affichages et clics de candidature en fonction de la provenance (pays, région), et de les représenter sur une carte.
- *Search on specific job skills*. Cette fonctionnalité permet d'analyser la performance des job boards relativement à un ensemble de compétences ou postes grâce à un outil de recherche personnalisé.

TRAQ24 a été lancé en septembre 2011 et se présente comme un complément de *Chameleon* au niveau des analyses statistiques proposées. Le principe proposé consiste à se baser sur l'analyse de la performance de ses campagnes passées (comparaison des métriques entre les job boards et à celles des autres recruteurs) pour en déduire des indications sur les actions à mener pour améliorer les prochaines campagnes de recrutement. Cependant, seul *Spendometer* suggère des actions à mener en direct lors de la diffusion d'une nouvelle annonce, et cela concerne la suppression du job board payant le moins performant. Il n'y a pas d'indication en direct sur les job boards qui devraient effectivement être utilisés. *eQuest Leaderboard* permet l'estimation de la performance d'une annonce à partir de l'historique des utilisateurs de *TRAQ24* en se basant sur le type de poste et la localisation. Le nombre de critères pris en compte est donc très limité et nous ne disposons pas d'informations sur

5. Nous pouvons supposer que cette fonctionnalité a recours à la même méthodologie que celle employée par *Spendometer*.

la méthodologie employée garantissant l'absence de biais dans les statistiques délivrées. De plus, la pertinence de l'estimation vis-à-vis des besoins du recruteur n'est pas évidente. En effet, ce type d'estimation est préconisé à la fois pour évaluer la performance attendue sur des job boards donnés (*Spendometer*) et pour comparer la performance de l'offre du recruteur à celle des autres recruteurs (*Track by unique job*), ce qui paraît contradictoire.

Broadbean

La société Broadbean offre ses services à 33 000 utilisateurs à travers 55 pays [Broadbean 2011a] et propose plusieurs outils dédiés aux ressources humaines : multidiffusion d'offres d'emploi, gestion des candidatures, recherche dans des bases de CV, recrutement sur les réseaux sociaux Facebook et Twitter. Des outils sont disponibles pour aider à contrôler le budget et analyser la performance des campagnes de recrutement dans une optique d'optimisation [Broadbean 2011b].

Une première fonctionnalité offerte par Broadbean est l'enregistrement d'une sélection automatique des portails d'emploi et réseaux sociaux en fonction du type de poste diffusé [Broadbean 2011c]. Broadbean propose également des rapports d'activité permettant de suivre le nombre d'offres publiées et le nombre de candidatures reçues sur chaque job board. Les rapports permettent également de comparer la qualité globale des candidatures reçues sur les différents portails. Enfin, les statistiques fournies dans les rapports de suivi donnent des indications sur les job boards à utiliser en fonction du secteur et du type de poste à pourvoir [Broadbean 2011d], mais nous ne disposons d'aucun détail concernant la méthodologie employée.

Bien que les rapports d'activité fournissent une aide à la décision, le nombre de critères pris en compte pour la mise en évidence des job boards à utiliser est très restreint (secteur et type de poste). Omettre les autres facteurs pouvant influencer la performance des offres peut biaiser l'estimation donnée au recruteur. De plus, il n'est pas proposé d'outil de recommandation permettant de guider les choix du recruteur en direct lorsqu'il diffuse son annonce.

RFlex

RFlex est une solution de gestion du recrutement, de la mobilité et des compétences qui offre ses services à plus de 200 entreprises réparties dans 70 pays [RFlex 2011]. Lors d'un communiqué de presse datant du 23 avril 2010, RFlex fait part du lancement d'un outil d'optimisation de "sourcing" [Exclusive RH 2010]. En se basant sur une suite de métriques (nombre de candidatures reçues, nombre d'entretiens obtenus, nombre de personnes recrutées), l'outil *Profils.net* permet d'évaluer le retour sur investissement de chaque site d'emploi. *Profils.net* agrège l'ensemble des offres d'emploi des clients et des candidatures reçues métier par métier, afin de présenter au moment de la diffusion de l'annonce "les statistiques des quatre sites emploi les mieux positionnés pour le poste à pourvoir". L'outil se base sur des associations de métiers [6], et les données sont mises à jour sur un trimestre glissant.

Nous pouvons mettre en évidence plusieurs inconvénients à cette méthode. D'abord, seul le métier proposé est pris en compte dans l'estimation du ROI des différents sites d'emploi ce qui ne permet pas de fournir une évaluation précise de la performance que le recruteur peut attendre (de nombreux autres facteurs peuvent influencer le nombre de retours provenant d'un site). De plus, le système qui consiste à associer manuellement des nomenclatures provenant de différentes sources n'est pas pleinement satisfaisant car certains sites fournissent des catégories de métiers très larges qui ne permettent pas d'apprécier finement le type de poste à pourvoir.

Aktor Interactive

Aktor Interactive est une agence de Communication de recrutement et de Marketing RH présente sur les principaux marchés européens. Début 2010, le rapprochement des sociétés Kioskemploi et Aktor Interactive donne naissance à Kioskemploi-Aktor HR Software, logiciel de recrutement et de gestion des Ressources Humaines [Groupe Aktor 2011]. C'est cette solution qui nous intéresse et que nous décrivons dans cette section. Kioskemploi-

6. "un an de travail de mapping des métiers et des profils réalisé", [Exclusive RH 2010]. Notre connaissance du domaine nous permet de supposer qu'une typologie de métiers a été définie par RFlex, puis que les nomenclatures des métiers des différents sites d'emploi et/ou les titres de postes ont été associés manuellement à la typologie de métiers établie.

Aktor HR Software propose des logiciels de gestion des candidatures et deux logiciels de diffusion d'annonces et d'analyse de la performance : *Robopost* et *Jobstats*. Ces derniers permettent de publier facilement des annonces et des publicités de recrutement sur plusieurs job boards simultanément et d'analyser les performances des campagnes de recrutement [Kioskemploi-Aktor HR Software 2011]. *Robopost* est l'outil qui permet de multidiffuser les annonces sur Internet. Il fournit des statistiques par annonce, par job board, par période, etc., et permet au recruteur de définir des rapports personnalisés pour mesurer l'efficacité de ses campagnes et améliorer la performance des campagnes futures. L'outil *Jobstats* permet de visualiser les statistiques de performance (affichages, clics, ratio clics/affichages) au cours du temps par job board et par poste. Ici encore, le recruteur doit analyser les rapports fournis pour en déduire des actions à mener pour les futures campagnes, mais il n'est pas guidé au moment de la diffusion de son annonce. De plus, le recruteur n'a pas accès à des statistiques agrégées sur l'ensemble des clients pour évaluer la performance attendue sur les différents sites, ou pour se comparer aux autres recruteurs.

Remarque 1 *Des travaux de thèse ont été financés par Aktor Interactive [Kessler 2009] mais concernent des recherches sur le traitement automatique des offres (détection des différentes parties d'une offre d'emploi, distinction entre CV et lettre de motivation) et la détection de candidatures correspondant à une offre d'emploi. Nous détaillerons l'aspect pertinent de ces recherches par rapport à nos travaux dans la section 4.2.2.*

Autres solutions

D'autres solutions offrent un service de multidiffusion d'annonces ainsi que des outils permettant l'analyse des performances mais communiquent très peu sur ces derniers. Nous pouvons retenir :

- Ubiposting [Ubiposting 2011],
- Kimladi [Kimladi 2011],
- SmartRecruiters [SmartRecruiters 2011].

Synthèse

Le tableau 1.1 synthétise les informations obtenues suite à nos recherches pour les quatre principaux concurrents du marché du point de vue de l'offre d'outils d'aide à la décision pour les recruteurs. Les colonnes du tableau indiquent l'outil de recrutement concerné tandis que les lignes présentent les fonctionnalités identifiées. L'outil que nous développons est appelé "outil prédictif MP".

	eQuest	Brodbean	RFlex	Kioskemploi – Aktor HR	Objectif *outil prédictif MP*
Met à disposition des outils de rapport statistique pour l'analyse de la performance des annonces postées	×	×	×	×	×
Permet la comparaison de ses propres performances à celles des autres recruteurs	×				×
Fournit une estimation de la performance attendue pour un type de poste sur les différents job boards sur la base de statistiques agrégées	×		×		×
Prend en compte un grand nombre de critères (pertinents) pour estimer la performance d'une campagne					×
Fournit des recommandations au recruteur au moment de la diffusion d'une nouvelle annonce	×		×		×

TABLE 1.1 – Comparaison des solutions de recrutement concurrentes

Notre objectif est de développer un outil d'aide à la décision mettant à disposition du recruteur toutes ces fonctionnalités dans le but de fournir une solution complète répondant aux attentes et modes de fonctionnement variés des différents recruteurs.

Chapitre 2

Indicateurs de performance d'une campagne de recrutement

2.1 Évaluation de la performance d'une campagne de recrutement : état de l'art

Pour pouvoir introduire la notion de performance d'une campagne de recrutement, il est nécessaire de définir au préalable les objectifs de recrutement. En effet, l'évaluation d'une campagne se fait en comparant les résultats obtenus aux objectifs initiaux de l'organisation, comme le suggèrent Breaugh and Starke [2000] via leur représentation de l'organisation du processus de recrutement (la première phase y est la définition des objectifs, et la dernière phase l'évaluation des résultats).

Pendant longtemps, le principal objectif des recruteurs était d'attirer le plus grand nombre de candidatures. Rynes [1991] suggère de considérer un plus large éventail d'indicateurs, et propose notamment un ensemble d'indicateurs évalués "post-embauche". Ainsi, nous identifions deux types d'indicateurs permettant d'évaluer les résultats d'une campagne : les indicateurs "pré-embauche" (calculés avant le recrutement) et les indicateurs "post-embauche". Nous commençons par présenter les indicateurs post-embauche pour finir par les indicateurs pré-embauche, de manière à nous rapprocher progressivement des indicateurs plus intimement liés à notre problématique.

2.1.1 Indicateurs post-embauche

Des indicateurs évaluant le(s) recrutement(s) effectué(s) peuvent être étudiés. Nous avons : la performance des nouveaux employés, leur satisfaction vis-à-vis du poste, et le taux de rétention un an après les nouvelles embauches [Rynes 1991]. Mais d'autres indicateurs pouvant être mesurés les premiers jours suivant l'embauche sont également intéressants : le coût du recrutement, la durée pour pourvoir le(s) poste(s), le nombre d'individus embauchés, et la diversité des nouveaux employés [Breaugh 1992].

L'enquête "Offre d'emploi et recrutement"[1], réalisée en France par le Ministère du Travail en 2005, suggère de prendre en compte un certain nombre d'indicateurs pour étudier l'efficacité des procédures de recrutement des entreprises françaises. L'étude du dictionnaire des données nous suggère en effet de prendre en compte des indicateurs mesurés post-embauche. Certains peuvent être mesurés immédiatement après le recrutement : la durée totale du processus de recrutement (entre la diffusion du besoin et le choix d'un candidat), le coût externe total du recrutement, le coût interne en termes de durée (cumul du temps passé par des personnes de l'établissement), le niveau estimé de difficulté du recrutement (jugé sur trois niveaux). D'autres sont mesurés six mois après la prise de fonction du salarié : le candidat recruté est-il toujours présent dans l'établissement (si départ, le candidat est-il parti plus tôt que prévu), le recruteur choisirait-il le même candidat ?

2.1.2 Indicateurs pré-embauche

Comme évoqué ci-dessus, un indicateur pré-embauche majeur est le nombre d'individus qui candidatent au poste [Wanous 1992; Williams et al. 1993]. Mais d'autres indicateurs comme la qualité des candidatures reçues ou leur diversité sont également d'intérêt pour le recruteur [Williams et al. 1993]. La diversité peut concerner l'âge, l'origine ethnique ou encore l'origine géographique des candidats.

1. L'enquête "Offre d'emploi et recrutement" (OFER) s'adresse aux établissements du secteur privé d'au moins un salarié ayant recruté ou essayé de recruter au moins un salarié (hors intérim) au cours des douze mois précédant la collecte de l'enquête. L'enquête porte sur le dernier recrutement effectué. Ses objectifs sont d'améliorer la connaissance de l'organisation des procédures de recrutement du côté des entreprises, d'en apprécier l'efficacité, et d'améliorer la compréhension des notions de difficultés et d'échec du recrutement (`http://www.travail-solidarite.gouv.fr`).

L'enquête OFER citée précédemment suggère également de prendre en compte des indicateurs pré-embauche : le nombre de candidatures examinées par le recruteur (si examen des candidatures), la satisfaction du recruteur vis-à-vis de l'ensemble des candidatures reçues, le nombre de candidatures retenues à l'issue du premier tri (si premier tri avant entretien), le nombre de candidats ayant passé des entretiens individuels, et le nombre de candidats jugés intéressants qui se sont désistés.

2.1.3 Apparition d'Internet et évolution des indicateurs de performance

Les indicateurs évoqués jusqu'à présent peuvent être mesurés quel que soit le moyen utilisé pour trouver des candidats. Cependant, les mesures des indicateurs de performance ont évolué de manière importante avec l'expansion du média Internet pour recruter, et ce à deux niveaux différents.

2.1.3.1 Évolution de la performance due au média Internet

Tout d'abord, les valeurs des indicateurs de performance qui étaient mesurées avant l'apparition du e-recrutement ont été profondément modifiées. Ainsi, les principaux changements sont :

- L'augmentation de la visibilité des annonces en touchant une audience plus large [Bartram 2005; Laabs 1998; Pin et al. 2001; Zusman and Landis 2002; Veger 2006] et par suite l'augmentation du volume des candidatures reçues, parfois même de manière trop importante [Brooke 1998; Galanaki 2002], rendant impossible le traitement manuel. En particulier, l'e-recrutement permet d'atteindre une audience plus large de candidats passifs [Politt 2004; Veger 2006].
- La diminution des coûts de recrutement [Bartram 2005; Pin et al. 2001; Veger 2006].
- La réduction de la durée du processus de recrutement [Bartram 2005; Pin et al. 2001; Veger 2006]. Les gains de temps apparaissent à trois niveaux : au moment de la diffusion de l'offre (publication quasi-immédiate et automatique en ligne), au niveau de la réception des candidatures (les individus peuvent postuler immédiatement suite à la publication de l'offre et ce 24h/24), et au niveau du traitement des candidatures (elles peuvent être traitées électroniquement grâce à des logiciels de gestion

des ressources humaines).

– L'augmentation du volume de candidatures non pertinentes [Fondeur et Tuchszirer 2005; Kaydo and Cohen 1999]. En éliminant un certain nombre de barrières à l'entrée du marché du travail, Internet a apporté une plus grande transparence et entrainé un fort accroissement du taux d'informations non pertinentes (ou "bruit", Fondeur et Tuchszirer 2005; Fondeur 2006). Ce bruit apparaît au niveau de l'accès aux offres d'emploi (lié à la qualité du moteur de recherche du site d'emploi), et au niveau des choix des candidats qui peuvent postuler à des offres non adaptées à leur profil étant donné le faible coût de l'acte de candidature.

2.1.3.2 Naissance de nouveaux indicateurs de performance

De plus, le développement du recrutement sur Internet ainsi que les mécanismes de diffusion et de candidatures aux annonces d'emploi sur ce média ont donné naissance à de nouvelles mesures d'intérêt, intimement liées au processus de navigation sur les sites d'emploi. Dans un premier temps, nous présentons le processus complet de recherche d'emploi et de candidature sur Internet (cf. figure 2.1).

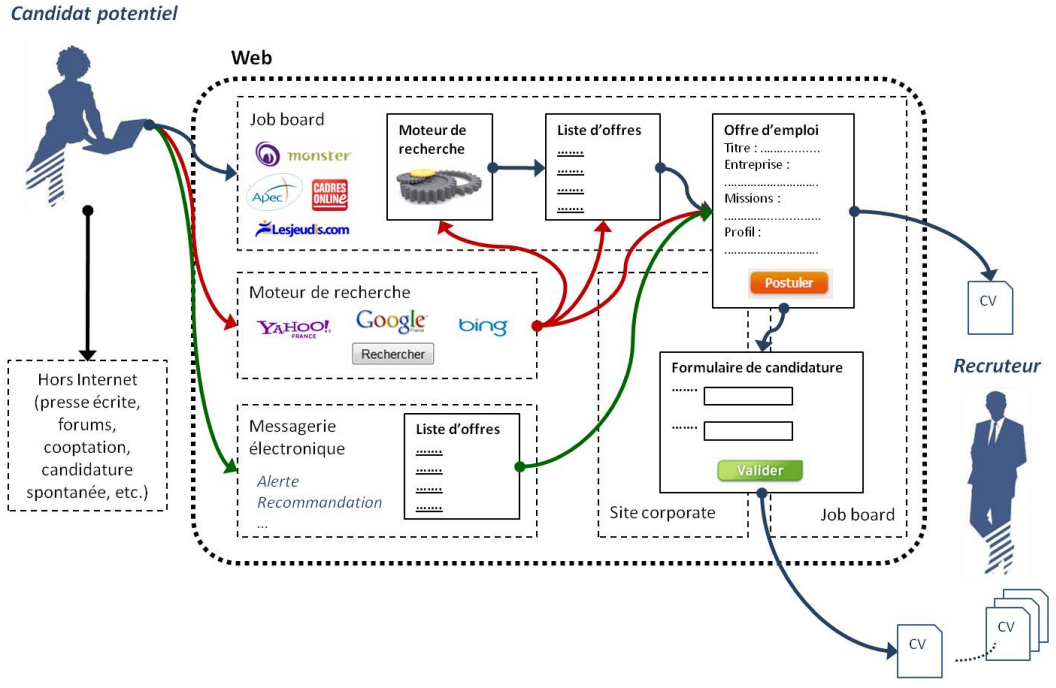

FIGURE 2.1 – Processus de recherche d'emploi et de candidature sur Internet

Comme représenté sur la figure 2.1, le chercheur d'emploi peut accéder à une offre d'emploi sur Internet principalement de trois manières différentes :

- En se rendant directement sur le site d'emploi (ou job board), puis en effectuant une requête via le moteur de recherche du site. Une liste d'offres correspondant à ses critères lui est alors proposée, les offres sont présentées sous forme de bandeaux (chaque bandeau comprend le titre et un bref extrait de l'annonce). Ensuite, le candidat peut parcourir la liste des offres et cliquer sur le bandeau de l'une d'entre elles. Le clic le renvoie sur une page où le contenu complet de l'annonce est affiché.
- Par l'intermédiaire d'un moteur de recherche du type *Google*, *Yahoo*, *Bing*, etc. qui lui suggère une liste de liens pouvant éventuellement renvoyer vers la page d'accueil d'un site d'emploi, une liste d'offres sur un site d'emploi, ou directement sur le texte complet d'une annonce d'emploi.
- Depuis sa messagerie électronique où des offres d'emploi lui sont suggérées par mail. Celles-ci peuvent provenir d'une alerte à laquelle le candidat se serait inscrit sur un site d'emploi, d'une liste de diffusion à laquelle il est inscrit, d'une recommandation, etc. Dans le cas qui nous intéresse, le mail contient des liens renvoyant au contenu des offres d'emploi.

Une fois le contenu de l'offre visualisé, le candidat potentiel peut décider de postuler. Plusieurs cas sont alors envisageables :

- Une adresse mail de candidature est indiquée dans le contenu de l'offre et l'individu est alors libre d'y envoyer directement son dossier de candidature (CV, lettre de motivation, etc.).
- Le candidat potentiel doit cliquer sur un bouton pour postuler qui le conduit à un formulaire de candidature sur le site d'emploi. Il remplit alors les informations demandées, joint son dossier de candidature et valide l'envoi par mail de sa candidature au recruteur.
- Le candidat potentiel clique sur le bouton "postuler" et est renvoyé vers le site carrière du recruteur. Il a alors accès à l'offre d'emploi, doit remplir un formulaire, et valider l'envoi de sa candidature.

Ce processus de candidature suggère la prise en compte de nouveaux indicateurs de perfor-

mance mesurés pré-embauche. En effet, nous pouvons désormais nous intéresser au nombre de fois qu'une offre donnée est visualisée. Nous pouvons également nous intéresser au nombre de candidats ayant souhaité postuler à l'offre d'emploi, c'est-à-dire au nombre de personnes ayant cliqué sur le bouton "postuler" situé sur la page de l'offre. Bien entendu, nous nous intéressons également au nombre de candidatures effectivement envoyées au recruteur (via le site d'emploi ou via le site du recruteur) comme déjà évoqué dans la section 2.1.2. Par ailleurs, il est également pertinent d'évaluer le coût d'un affichage, le coût d'un clic pour postuler ou le coût d'une candidature réelle (coût par CV) pour le recruteur. Enfin, des indicateurs de performance à prendre en compte sont les proportions d'individus se trouvant à une étape donnée de la candidature sur le site et validant l'étape suivante (taux de conversion), c'est-à-dire (cf. figure 2.2) :

- le nombre de personnes ayant visualisé l'offre par rapport au nombre de personnes ayant cliqué pour postuler *(1)* ;
- le nombre de personnes ayant cliqué pour postuler par rapport au nombre de personnes ayant réellement validé leur candidature *(2)* ;
- le nombre de personnes ayant visualisé l'offre par rapport au nombre de personnes ayant réellement validé leur candidature *(3)*.

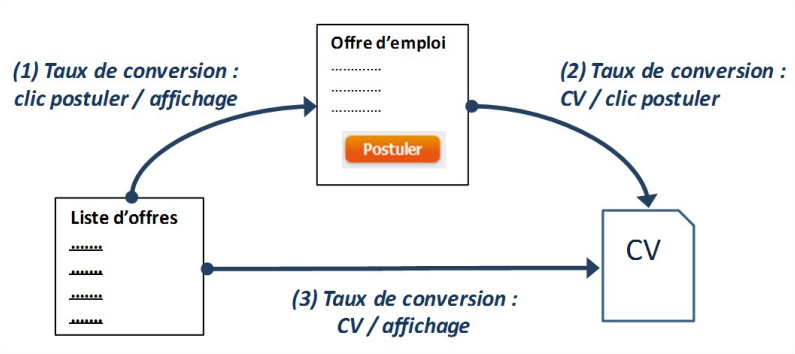

FIGURE 2.2 – Illustration des différents taux de conversion envisageables

2.2 Proposition d'indicateurs de performance

La notion de performance d'une campagne de recrutement diffère selon les entreprises car chacune a ses propres objectifs [Veger 2006]. Nous allons devoir prendre en compte

l'hétérogénéité des objectifs et attentes des recruteurs au moment du démarrage de la campagne afin de proposer un outil d'aide à la décision se révélant pertinent pour tous.

2.2.1 Présentation des données enregistrées par l'outil

Le "tracking" est l'ensemble des moyens mis en œuvre pour enregistrer, tracer les actions des internautes. Multiposting.fr enregistre en base de données les actions des internautes constituant un des actes du processus de candidature sur un site d'emploi.

Les études que nous allons mener et les résultats que nous pourrons fournir sont conditionnés par les informations enregistrées par l'outil. En effet, pour mettre au point des indicateurs de performance et tenter d'en expliquer la variabilité, nous devons disposer de données concernant les actions effectuées par les candidats en rapport avec l'offre d'emploi étudiée. En particulier, nous devons disposer de l'information lorsqu'une candidature est reçue en réponse à une offre d'emploi. Nous allons voir qu'une candidature au sens du "tracking Multiposting.fr" peut prendre différentes formes. Le schéma qui suit (figure 2.3) présente les données enregistrées lors des différentes étapes du processus de candidature. Ces données vont dépendre du site d'emploi, ainsi que de choix faits par le recruteur.

La visualisation des offres est décomptée pour certains sites d'emploi seulement (ceux qui le permettent), elle est retranscrite à travers le nombre d'affichages ($nb\_affichages$).

Le recruteur a le choix du mode de candidature, qui va déterminer la statistique décomptée pour évaluer le nombre de candidatures. Deux modes de candidature sont possibles :

- le recruteur choisit de fournir une adresse URL vers laquelle les candidats seront redirigés pour postuler (généralement la section "Carrières" du site entreprise), l'indicateur fourni au recruteur sera alors le nombre de clics de redirection (nb_clics) ;
- le recruteur choisit de fournir une adresse e-mail à laquelle les dossiers de candidature (CV et éventuellement lettre de motivation) seront envoyés. Selon le site d'emploi, le candidat enverra lui-même un e-mail de candidature ou un formulaire de candidature lui sera proposé. Dans ces deux cas, la statistique délivrée au recruteur est alors le nombre de CV (nb_cv). Ces derniers sont détectés à réception dans la boîte e-mail du recruteur et sont alors consultables dans l'interface Multiposting.fr.

Enfin, si le mode de candidature choisi est le mail, le recruteur peut avoir une option

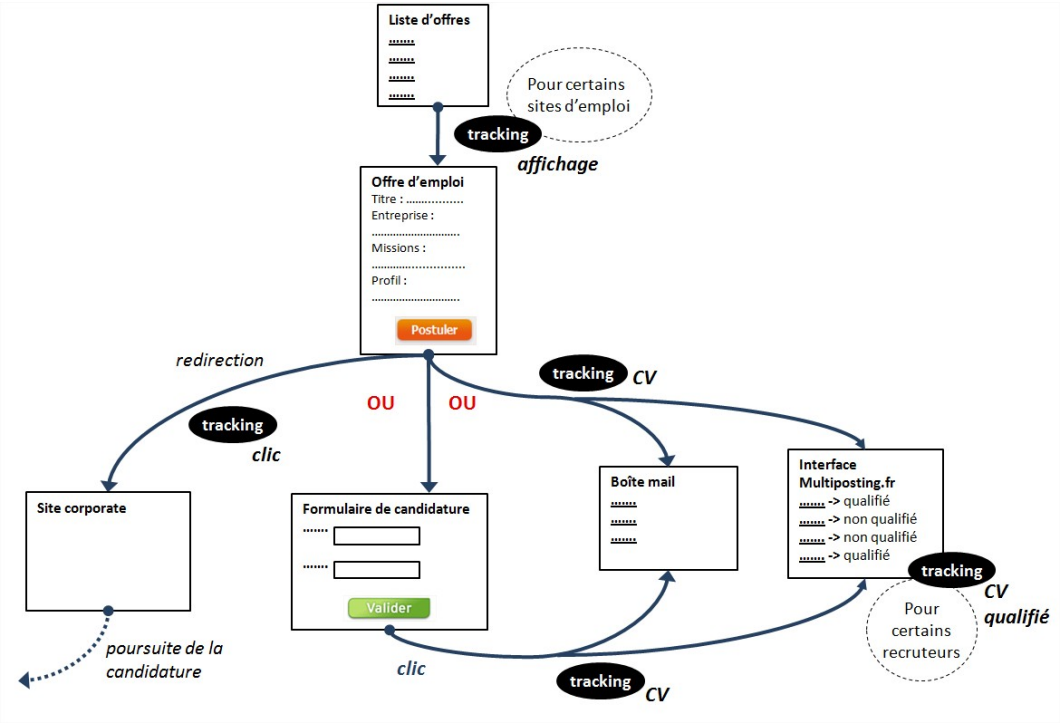

FIGURE 2.3 – Actions de candidature enregistrées dans l'outil Multiposting.fr

pour étiqueter les CV ainsi reçus sur l'interface de l'outil. Trois étiquettes sont alors possibles : "qualifié", "non qualifié", "non traité", et nous pouvons évaluer le nombre de CV qualifiés reçus (*nb_ qualifies*). Le nombre de clients ayant choisi cette option et l'utilisant régulièrement est très faible, ainsi que le nombre de candidatures étiquetées pour un client donné [2].

Remarque 2 *Une perte d'information est possible dans deux cas de figure. Si l'adresse URL ou l'e-mail de candidature fourni par le recruteur est erroné, le candidat ne sera alors pas redirigé correctement ou son dossier de candidature n'arrivera pas au recruteur. Si le candidat visualise l'annonce, quitte le site d'emploi, puis se rend ultérieurement directement sur le site du recruteur pour candidater, le lien entre visualisation et candidature ne sera pas établi.*

2. Parmi les 29 clients ayant qualifié au moins une candidature, 22 ont utilisé l'option marginalement (moins de 30 CV étiquetés), 6 ont étiqueté entre 100 et 300 CV (mais cela représente une très faible proportion du volume total de CV reçus par ces clients), et un seul client a étiqueté un nombre important de candidatures (plus de 2400 CV).

2.2.2 Indicateurs de performance proposés

Nous devons proposer aux recruteurs un ou plusieurs indicateurs très simples, facilement compréhensibles et exploitables par tous. En effet, ces indicateurs seront délivrés sur l'interface utilisateur et devront permettre à tous les recruteurs d'améliorer la diffusion de leurs annonces. Sur la base des informations enregistrées par Multiposting.fr, nous pouvons proposer un ensemble d'indicateurs de performance : certains disponibles dans tous les cas, d'autres disponibles uniquement dans des cas particuliers.

2.2.2.1 Indicateurs de performance dans le cas général

Dans le cas général, une seule statistique est disponible pour évaluer le nombre de candidats, et elle dépend du mode de candidature choisi par le recruteur (par mail ou par URL). Seul le mode de candidature par e-mail permet de connaître le nombre réel de candidatures par le décompte du nombre de CV reçus dans la boîte e-mail du recruteur. En effet, lorsque les candidats sont redirigés vers une adresse URL, Multiposting.fr ne peut alors plus tracer la suite des actions du candidat potentiel et seul le nombre de clics de redirection est décompté. Dans tous les cas, nous ne connaissons pas le nombre de candidatures qualifiées parmi celles qui ont été reçues et ne savons pas non plus si des entretiens ont été menés et si l'offre d'emploi a abouti à un recrutement.

Soit la variable $candidature\_url_i$ le mode de candidature choisi par le recruteur pour diffuser l'offre i, celle-ci pouvant prendre les valeurs "0" (candidature par e-mail) et "1" (candidature par URL). Afin que les recruteurs puissent évaluer quantativement leurs retours (ou encore le rendement de leur offre), le premier indicateur de la performance d'une offre i sur un site d'emploi j donné est :

$$P_{ij}^1 = \begin{cases} nb\_cv_{ij} & \text{si } candidature\_url_i = 0 \\ nb\_clics_{ij} & \text{si } candidature\_url_i = 1 \end{cases}$$

Soit c_j le coût de diffusion d'une annonce sur le job board j (ce coût est paramétrable pour un recruteur donné sans perte de généralité). Pour permettre aux recruteurs d'évaluer le retour sur investissement de leur offre, un second indicateur de performance est proposé :

$$P_{ij}^2 = \begin{cases} \frac{c_j}{nb\_cv_{ij}} & \text{si } candidature\_url_i = 0 \\ \frac{c_j}{nb\_clics_{ij}} & \text{si } candidature\_url_i = 1 \end{cases}$$

Soit J' l'ensemble des sites sélectionnés pour la diffusion de l'offre i, les indicateurs de la performance globale de l'offre i deviennent :

$$O_i^1 = \begin{cases} \sum_{j \in J'} nb\_cv_{ij} & \text{si } candidature\_url_i = 0 \\ \sum_{j \in J'} nb\_clics_{ij} & \text{si } candidature\_url_i = 1 \end{cases}$$

et

$$O_i^2 = \begin{cases} \frac{\sum_{j \in J'} c_j}{\sum_{j \in J'} nb\_cv_{ij}} & \text{si } candidature\_url_i = 0 \\ \frac{\sum_{j \in J'} c_j}{\sum_{j \in J'} nb\_clics_{ij}} & \text{si } candidature\_url_i = 1 \end{cases}$$

2.2.2.2 Indicateurs de performance dans les cas spécifiques

La section 2.2.1 a mis en évidence deux cas spécifiques où nous disposons d'informations supplémentaires relativement au cas général. Le premier cas apparaît lorsque le décompte des affichages de l'offre est possible sur le site étudié. Nous pouvons alors envisager pour ce site l'introduction de nouveaux indicateurs de performance :

- le nombre d'affichages de l'offre $\tilde{P}_{ij}^1 = nb\_affichages_{ij}$ si $candidature\_url_i \in \{0, 1\}$
- le coût par affichage $\tilde{P}_{ij}^2 = \frac{c_j}{nb\_affichages_{ij}}$ si $candidature\_url_i \in \{0, 1\}$
- le taux de conversion des affichages en candidatures

$$\tilde{P}_{ij}^3 = \begin{cases} \frac{nb\_cv_{ij}}{nb\_affichages_{ij}} & \text{si } candidature\_url_i = 0 \\ \frac{nb\_clics_{ij}}{nb\_affichages_{ij}} & \text{si } candidature\_url_i = 1 \end{cases}$$

Si J' est l'ensemble des sites sélectionnés pour la diffusion de l'offre i, alors les indicateurs de la performance globale de l'offre i deviennent respectivement :

- $\tilde{O}_i^1 = \sum_{j \in J'} nb\_affichages_{ij}$ si $candidature\_url_i \in \{0, 1\}$
- $\tilde{O}_i^2 = \frac{\sum_{j \in J'} c_j}{\sum_{j \in J'} nb\_affichages_{ij}}$ si $candidature\_url_i \in \{0, 1\}$
- $\tilde{O}_i^3 = \begin{cases} \frac{\sum_{j \in J'} nb\_cv_{ij}}{\sum_{j \in J'} nb\_affichages_{ij}} & \text{si } candidature\_url_i = 0 \\ \frac{\sum_{j \in J'} nb\_clics_{ij}}{\sum_{j \in J'} nb\_affichages_{ij}} & \text{si } candidature\_url_i = 1 \end{cases}$.

Le deuxième cas particulier fait référence aux recruteurs ayant choisi le mode de candidature par mail et pour lesquels nous connaissons le nombre de candidatures qualifiées

(le recruteur annote manuellement les CV) sur l'ensemble ou un sous-ensemble des offres qu'ils ont postées. L'information n'est donc disponible que pour des sites qui ont été utilisés par le recruteur. Soit $\tilde{J}$ l'ensemble des sites utilisés par le recruteur, nous pouvons alors introduire les indicateurs de performance suivants :

- le nombre de CV qualifiés $\tilde{P}^4_{ij} = nb\_qualifies_{ij}$ où $candidature\_url_i = 0$ et $j \in \tilde{J}$
- le coût par CV qualifié $\tilde{P}^5_{ij} = \frac{c_j}{nb\_qualifies_{ij}}$ où $candidature\_url_i = 0$ et $j \in \tilde{J}$
- le taux de qualification des CV reçus

  $\tilde{P}^6_{ij} = \frac{nb\_qualifies_{ij}}{nb\_cv_{ij}}$ où $candidature\_url_i = 0$ et $j \in \tilde{J}$

De même que précédemment, les indicateurs de la performance globale d'une annonce i postée sur un ensemble J' de sites ($J' \subset \tilde{J}$) sont respectivement définis par :

- $\tilde{O}^4_i = \sum_{j \in J'} nb\_qualifies_{ij}$ où $candidature\_url_i = 0$
- $\tilde{O}^5_i = \frac{\sum_{j \in J'} c_j}{\sum_{j \in J'} nb\_qualifies_{ij}}$ où $candidature\_url_i = 0$
- $\tilde{O}^6_i = \frac{\sum_{j \in J'} nb\_qualifies_{ij}}{\sum_{j \in J'} nb\_cv_{ij}}$ où $candidature\_url_i = 0$

2.2.3 Indicateur de la performance relative

Dans le chapitre 3, nous mettons en évidence deux types de facteurs explicatifs : les facteurs flexibles et inflexibles. Les facteurs inflexibles sont déterminés par la définition des besoins de l'entreprise, ou les caractéristiques de l'entreprise elle-même et sont donc considérés comme des données. L'ensemble des valeurs prises par les facteurs inflexibles déterminent le type de l'offre étudiée. L'intérêt de cette approche est qu'un recruteur puisse évaluer sa performance "dans l'absolu" et non relativement à d'autres recruteurs, car il ne peut pas agir sur certains des aspects de la campagne. Notamment, la notoriété de l'entreprise qui recrute fait partie des facteurs inflexibles.

Soit $\{X_1, \ldots, X_p\}$ l'ensemble des critères permettant de caractériser une offre d'emploi[3] (facteurs flexibles et inflexibles). Soient $\{X_{k_1}, \ldots, X_{k_l}\}$ les facteurs inflexibles et $(x_{k_1}, \ldots, x_{k_l})$ le vecteur des valeurs prises par ces facteurs. Nous souhaitons comparer

3. Ces critères seront précisés dans le chapitre 3 d'un point de vue théorique (par le listing de tous les critères que l'on souhaiterait idéalement prendre en compte pour expliquer la performance d'une offre d'emploi), et d'un point de vue pratique dans le chapitre 6 (par la définition exhaustive des variables explicatives utilisées au sein de notre algorithme prédictif).

la performance obtenue étant donné le choix du recruteur sur les facteurs flexibles à celle obtenue avec le choix "moyen" (ou choix le plus commun), pour des valeurs identiques sur les facteurs inflexibles. Par définition, cette comparaison n'est faite que pour un job board donné. Nous admettons que les facteurs flexibles $\{X_{k_{l+1}}, \ldots, X_{k_p}\}$ sont représentés par des variables catégorielles. Soit $(x^m_{k_{l+1}}, \ldots, x^m_{k_p})$ le vecteur des valeurs les plus fréquemment utilisées pour chacune des variables. Ces valeurs sont déterminées de sorte à ne pas être caractéristiques d'un faible nombre de recruteurs ayant d'importants volumes d'annonces. La méthode est la suivante :

- sélectionner le périmètre des annonces des recruteurs ayant posté au moins 30 annonces via Multiposting.fr ;
- pour chaque recruteur, retenir pour chaque facteur flexible la modalité la plus utilisée ;
- pour chaque facteur flexible, retenir la modalité associée au plus grand nombre de recruteurs.

Soit $P_j(x_{k_1}, \ldots, x_{k_l}, x^m_{k_{l+1}}, \ldots, x^m_{k_p})$ l'estimation de l'indicateur de performance pour une offre fictive dont les caractéristiques sont définies par le vecteur $(x_{k_1}, \ldots, x_{k_l}, x^m_{k_{l+1}}, \ldots, x^m_{k_p})$. Soit $P_{ij}(x_{k_1}, \ldots, x_{k_l}, x_{k_{l+1}}, \ldots, x_{k_p})$ l'indicateur de performance estimé pour l'offre i du recruteur étudié. La performance de l'offre i relativement au comparatif établi est alors estimée par : $\frac{P_{ij}(x_{k_1},\ldots,x_{k_l},x_{k_{l+1}},\ldots,x_{k_p})}{P_j(x_{k_1},\ldots,x_{k_l},x^m_{k_{l+1}},\ldots,x^m_{k_p})} \times 100$.

2.2.4 Discussion

Nous proposons de traduire la performance d'une annonce d'emploi à travers des indicateurs issus de données tracées et un indice de performance relative. Les statistiques dont nous disposons étant assez limitées, nous avons proposé des indicateurs simples. Les indicateurs de performance proposés dans le cas général ont l'avantage d'être compréhensibles et rapidement pris en main par tous types de recruteurs (expérimentés ou non dans l'analyse de la performance des campagnes). En effet, l'outil d'aide à la décision proposé en complément de l'outil classique de multiposting devra en général pouvoir être maîtrisé par un recruteur sans aide extérieure. De plus, ils sont une aide pour répondre à certaines des problématiques de la recherche de candidats sur Internet (voir section 1.2) :

- les recruteurs souhaitent obtenir un volume de retours suffisamment important pour

pouvoir y trouver des profils adaptés au poste ;
- ils souhaitent également optimiser le budget consacré au recrutement via l'optimisation du retour sur investissement des annonces (ROI ou coût par candidature).

Si les recruteurs souhaitent obtenir des recommandations sur la base des candidatures qualifiées, ils en ont la possibilité en annotant les candidatures qu'ils reçoivent sur l'interface Multiposting.fr. Nous pouvons dans ce cas analyser et prédire les rendements en termes de candidatures qualifiées. Cependant, dans ce cas, nous ne pouvons recommander à un recruteur que les sites qu'il utilise habituellement. Le module de recommandation constitue alors une aide au choix entre les sites payants utilisés par le recruteur (sites sur lesquels il a acheté des crédits pour poster des annonces). En effet, la qualification des CV est une opération subjective qui dépend du recruteur. Une même candidature pour une offre peut être jugée de deux manières différentes par deux recruteurs différents. Les résultats obtenus pour un client au niveau de l'analyse de la qualification des candidatures ne sont donc pas généralisables à l'ensemble des recruteurs. Par suite, les recommandations ne pourront pas s'appliquer à des sites pour lesquels des candidatures n'ont pas été qualifiées.

Enfin, seule une minorité de sites d'emploi permettent le décompte des affichages des offres. Les sites ne seront donc pas comparables du point de vue du nombre d'affichages ou du taux de conversion, même si ces indicateurs peuvent être estimés dans un but descriptif.

2.3 Synthèse

Ce chapitre débute par une revue des indicateurs de performance d'une campagne de recrutement suggérés par la littérature, ainsi que par notre analyse du processus de candidature sur Internet. La figure 2.4 reprend ces indicateurs en les positionnant par rapport à la chronologie du processus de recrutement. Les premiers indicateurs sont spécifiques à Internet car c'est en ligne que débute le processus de recrutement, puis les indicateurs deviennent génériques dès la réception des candidatures par le recruteur.

La spécificité de l'enregistrement des données avec Multiposting.fr nous a amenés à proposer un ensemble d'indicateurs adaptés au cas général et à des cas particuliers en fonction des informations disponibles. La figure 2.5 présente la structure des indicateurs issus des

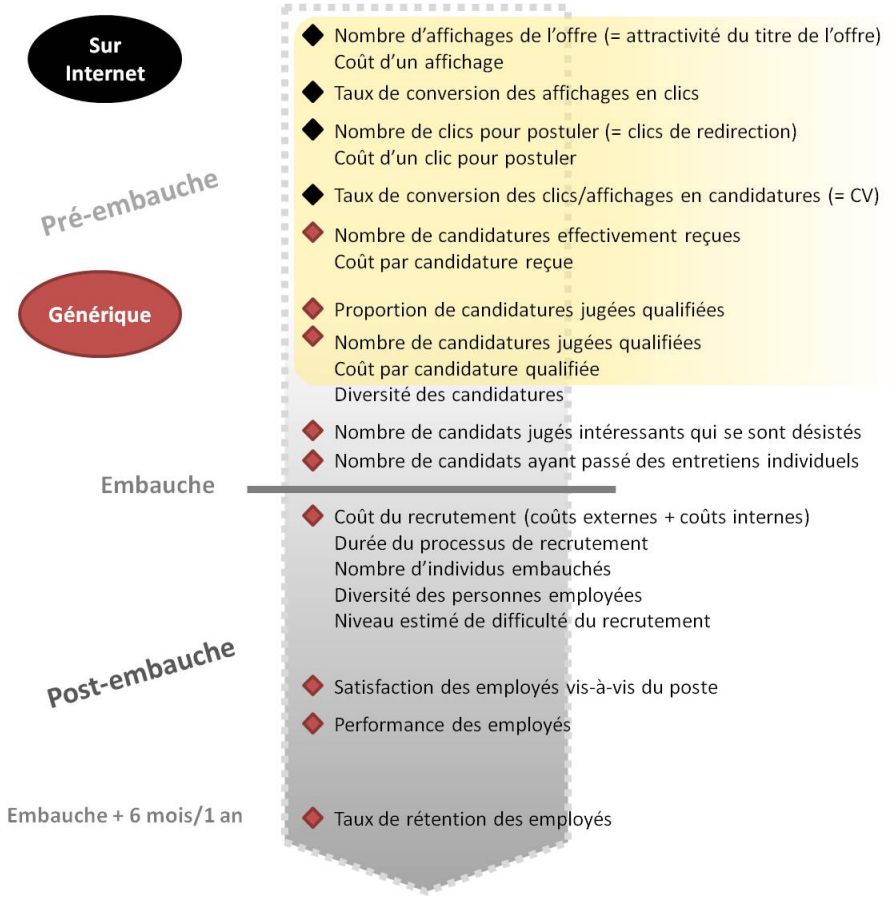

FIGURE 2.4 – Chronologie du processus de recrutement et indicateurs de performance

données enregistrées et disponibles ponctuellement ou pour l'ensemble des offres (surface jaune dans la figure 2.4).

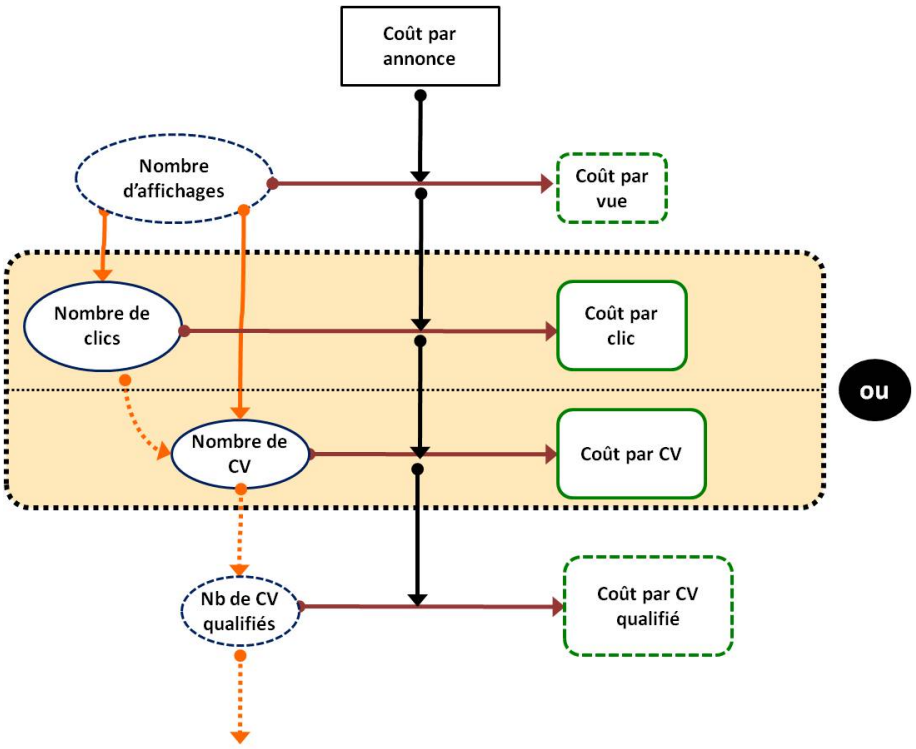

FIGURE 2.5 – Structure des indicateurs de performance issus des données enregistrées

Au final, les indicateurs qui seront estimés dans le cas général sont le nombre de candidatures reçues (nombre de clics ou nombre de CV selon le mode de candidature choisi) et le coût par candidature reçue. Ceux-ci sont complétés par un indice de performance relative permettant leur analyse dans la pratique. Ces indicateurs s'avèrent pertinents par rapport à certaines des problématiques liées à la recherche de candidats (identifiées dans la section 1.2), et leur simplicité assure une facilité d'usage en pratique par tous types de recruteurs. Dans la quasi-totalité des cas, il ne nous est pas possible de fournir des recommandations basées sur le niveau de qualification des candidatures en alternative à des recommandations basées sur le volume de candidatures reçues, car nous ne disposons pas de l'information.

Chapitre 3

Les facteurs explicatifs potentiels de la performance d'une campagne de recrutement

3.1 Les facteurs explicatifs de la performance d'une campagne dans la littérature

La littérature issue du "Management des Ressources Humaines" relève un ensemble de facteurs intervenant dans le processus de recrutement et influençant la décision de candidature. Nous ne faisons pas ici une revue exhaustive de la littérature de ce domaine. L'objectif est de rapporter les éléments qui y sont mentionnés et pouvant s'avérer utiles (directement ou indirectement) pour faire émerger des propositions de facteurs pertinents pour expliquer la performance d'une campagne de recrutement. De plus, la plupart des études examinant les effets des annonces d'emploi qui ont été menées ont fait appel à des échantillons d'étudiants jouant le rôle de candidats potentiels. Elles peuvent donc présenter des biais dans la mesure où d'autres profils pourraient avoir des comportements différents vis-à-vis des annonces d'emploi. En effet, Breaugh [2008] note que les limites méthodologiques de ces études restreignent la possibilité de tirer de fermes conclusions (e.g. il faudrait avoir recours à de plus larges et plus divers échantillons d'individus). Aussi, comme cela a été évoqué précédemment, nous les utiliserons donc comme des pistes de facteurs explicatifs.

3.1.1 Le message transmis

3.1.1.1 Attractivité du message

Les études menées ont montré que le titre a un impact important sur l'attractivité du message et la décision de candidature. En effet, des titres attirants, mis en avant avec des polices ou couleurs particulières vont influencer l'efficacité de la campagne [e.g. Dessler et al. 1999; Koch 1976]. Thompson et al. [2008] testent l'impact du formatage d'une annonce sur internet et montrent qu'une annonce ayant recours à des listes à puces est plus attractive et incite davantage les individus à poursuivre le processus de candidature qu'une annonce présentée sous forme d'un paragraphe.

A travers une étude menée sur des étudiants en commerce de dernière année, Blackman [2006] montre que l'usage du mot "diplômé" dans le titre de l'annonce la rend plus attractive aux yeux de ces derniers.

Les messages incluant des images sont également perçus comme plus attractifs [e.g. Tybout and Artz 1994; Blackman 2006].

3.1.1.2 Information transmise dans le message

Rafaeli et al. [2005] rapportent dans leur étude que les annonces de recrutement parues dans la presse génèrent un plus grand nombre de candidatures, d'embauches, et un coût par embauche plus faible si elles sont ciblées géographiquement (relativement à des annonces non ciblées).

Les recherches menées ont également mis en évidence que les annonces contenant le plus d'informations sont perçues comme les plus attractives [e.g. Allen et al. 2007] et les plus crédibles [e.g. Allen et al. 2004]. Barber and Roehling [1993] suggèrent que le fait qu'une entreprise fournisse peu d'informations sur elle et/ou sur le poste est perçu comme un manque de professionnalisme et un désintéressement envers le candidat. Gatewood et al. [1993] notent que les candidats potentiels réagissent positivement aux longues descriptions d'entreprises qui mettent l'accent sur les points forts de celles-ci. De plus, des chercheurs [e.g. Roberson et al. 2005] ont aussi montré que les annonces contenant des informations plus spécifiques à propos d'un poste augmentaient l'intérêt du candidat envers ce poste.

3.1.2 Le type de poste proposé

Les caractéristiques du poste proposé peuvent influencer la décision de candidature. Dans la littérature, les caractéristiques suivantes ont été identifiées :

- la localisation du poste [Cable and Graham 2000; Lievens and Highhouse 2003; Turban et al. 1998] ;
- le salaire [Cable and Graham 2000; Lievens and Highhouse 2003; Turban et al. 1998] ;
- le profil recherché [Mathews and Redman 1998; Petrick and Furr 1995].

3.1.3 Le recruteur

3.1.3.1 Attractivité du recruteur

Comme évoqué par Rynes and Cable [2003], les candidats potentiels vont être attirés par une entreprise selon qu'elle présente ou non certaines caractéristiques. En particulier, Chapman et al. [2005] rapportent que les candidats sont davantage attirés par les grandes entreprises. D'autres montrent que la notoriété de l'entreprise a une influence sur l'efficacité des annonces de recrutement [Aaker 1997; Belt and Paolillo 1982; Mathews and Redman 1998].

Des études ont également montré que l'image de l'entreprise [Gatewood et al. 1993; Lemmink et al. 2003] et que l'image en tant qu'employeur [Lemmink et al. 2003] influencent positivement les décisions de candidature.

3.1.3.2 Site web du recruteur

Bien que le site web du recruteur soit un support de diffusion sortant du champ de notre étude, nous pouvons nous inspirer de ses caractéristiques affectant la décision de candidature pour identifier de manière générale les aspects d'un site web potentiellement pertinents pour expliquer la performance d'une campagne.

Des sondages menés auprès de professionnels des ressources humaines ont montré que les sites web d'entreprises sont perçus comme une méthode de recrutement très efficace [e.g. Stone et al. 2005]. En particulier, ils permettent selon les praticiens de générer un

grand nombre de candidatures à un coût relativement faible. L'efficacité du site web est de plus fortement corrélée avec la visibilité et la réputation du recruteur [Rynes and Cable 2003].

Des chercheurs ont analysé l'effet des sites d'entreprises en menant des expérimentations et ont montré que des attributs comme l'esthétique (ex. : l'inclusion d'images, l'utilisation de polices distinctives), le contenu (ex. : aborder les caractéristiques importantes des emplois), et la fonctionnalité (ex. : facilité de navigation) sont importants pour expliquer les réactions des candidats potentiels [e.g. Braddy et al. 2006; Cober et al. 2003; Williamson et al. 2003]. Par exemple, Thompson et al. [2008] évaluent la facilité de navigation sur le site web par le nombre de pages et le temps nécessaires pour atteindre l'offre d'emploi étudiée.

Cependant, bien qu'un certain nombre d'études aient été menées récemment dans la littérature sur l'utilisation des sites web d'entreprises comme moyen de recrutement, de nombreux points restent encore à éclaircir. Par exemple, nous ne savons pas quel type d'information est recherché en premier sur le site web, ou ce qui cause le départ d'un visiteur sans avoir soumis de candidature [Breaugh 2008].

3.1.4 Le job board

Bien qu'ils permettent de générer de gros volumes de candidatures, les job boards (sites d'emploi) ont encore reçu peu d'attention de la part des chercheurs. **?** ont étudié les différences entre les candidatures générées par les sites généralistes (e.g. *Monster.com*) et celles générées par les sites spécialisés (secteur ou fonction). Les résultats obtenus montrent que les candidats issus des sites spécialisés ont un meilleur niveau d'études et de compétences, mais moins d'expérience professionnelle que ceux issus des sites généralistes.

3.1.5 Le calendrier

Selon le type de profil ciblé, le moment où débute le processus de recrutement est important. Turban and Cable [2003] ont montré que les employeurs cherchant à recruter sur les campus d'université avaient plus de candidats et des candidats de meilleure qualité si le processus commence tôt dans l'année.

3.2 Propositions de facteurs explicatifs

Nous proposons ici un ensemble de facteurs pouvant contribuer à expliquer la performance d'une campagne de recrutement. Certains sont inspirés et découlent de l'état de l'art présenté en section 3.1, tandis que d'autres sont de nouvelles propositions provenant de nos échanges avec des experts du recrutement, ainsi que de nos connaissances acquises sur le domaine. En reproduisant le processus de candidature standard sur Internet, nous mettons en évidence à chacune des étapes de manière "macro" les facteurs pouvant intervenir sur la performance d'une offre d'emploi postée sur un site d'emploi donné. L'ensemble de ces facteurs est détaillé dans un second temps. Nous distinguons deux types de facteurs :

- les facteurs *inflexibles*, qui sont déterminés par la définition des besoins de l'entreprise, l'entreprise elle-même, ou encore les données conjoncturelles, et qui de ce fait doivent être considérés comme des données ;
- les facteurs *flexibles*, découlant de choix faits par le recruteur.

Le chapitre 6 présente le détail de l'obtention de ces facteurs dans le cadre de l'application à notre étude (les facteurs sont calculés à partir de la base de données Multiposting.fr, ainsi qu'à partir de données obtenues par des sources externes). Le calcul des facteurs proposés pourra alors être adapté afin de répondre à d'éventuelles contraintes méthodologiques.

3.2.1 Processus de candidature et facteurs explicatifs

La figure 3.1 reproduit le parcours standard d'un candidat potentiel qui recherche des offres sur un site d'emploi donné. Les ellipses en trait plein représentent les étapes pour lesquelles nous pouvons mesurer [1], selon le site et le mode de candidature choisi, les effectifs concernés : le nombre de candidatures effectives ou le nombre de clics pour postuler, et le nombre de visualisations (sur certains sites). Lors du passage d'une étape à une autre, les catégories de facteurs explicatifs candidats intervenant (sur le volume de candidats potentiels ou sur la décision du candidat de passer à l'étape suivante du processus) sont mentionnées. Une même catégorie peut intervenir à plusieurs niveaux car différents facteurs sont impliqués. Les catégories (et sous-catégories) de facteurs sont détaillées dans les sections suivantes.

1. Grâce aux données enregistrées par Multiposting.fr.

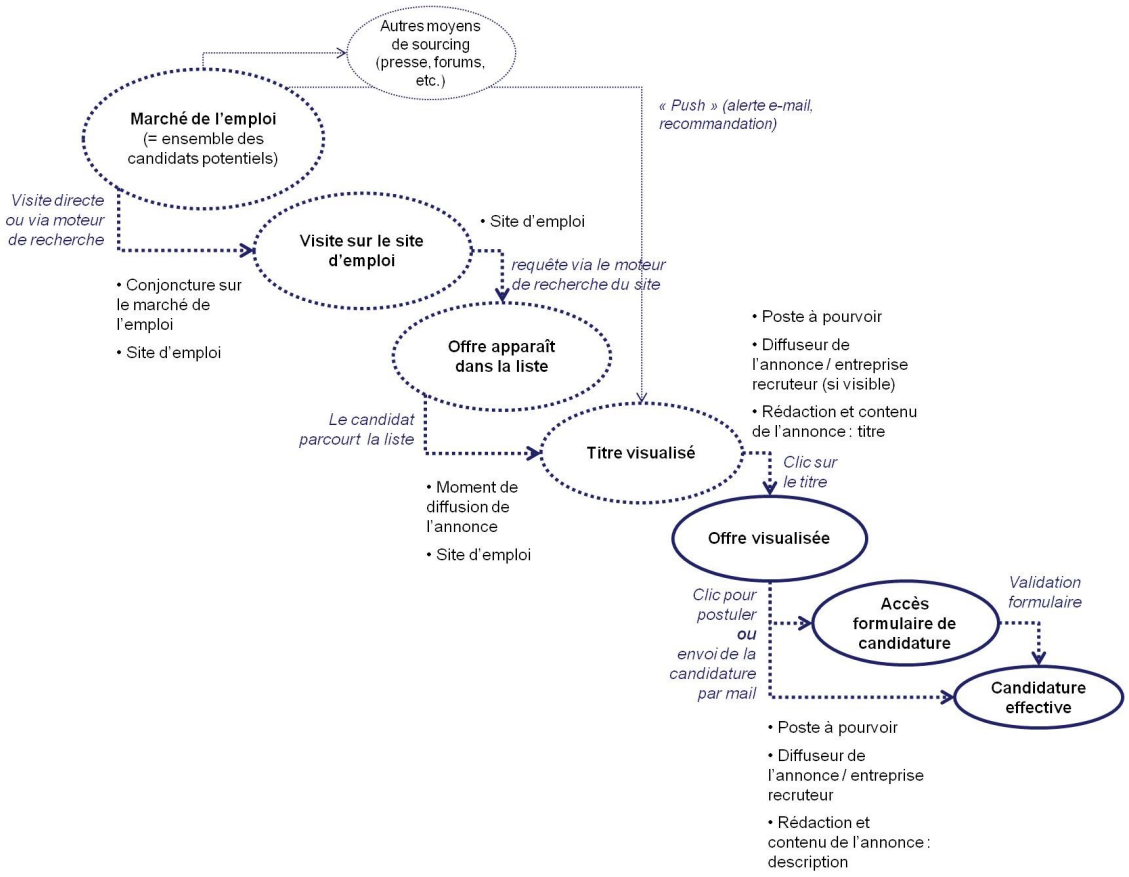

FIGURE 3.1 – Représentation schématique du processus de candidature et intervention des facteurs potentiels

3.2.2 Les facteurs inflexibles

3.2.2.1 Conjoncture sur le marché de l'emploi

Dans le but d'écarter toute confusion, nous commençons par définir les termes "offre" et "demande" sur le marché de l'emploi. Afin de nous adapter au vocabulaire employé dans les études menées sur le domaine du recrutement, nous appellerons "offre" les offres d'emploi proposées par les entreprises cherchant à recruter, et "demande" les candidats potentiels aux offres d'emploi, en référence aux "demandeurs d'emploi" [2].

Selon les périodes, les niveaux d'offre et de demande sur les différents secteurs et fonctions du marché de l'emploi peuvent varier. Aussi, les rapports de tension entre offre et demande varient également au cours du temps. Plus généralement, tous les critères permettant de décrire un poste sont sujets à des mesures de tensions entre offre et demande. Cependant, les données conjoncturelles ne sont disponibles publiquement que pour certains de ces critères et pour certaines périodes. La conjoncture peut également avoir une dimension géographique. Les tensions entre offre et demande dépendent du temps et de la localisation. Ces informations conjoncturelles sont des facteurs candidats pertinents pour expliquer la performance d'une campagne de recrutement car :

- Pour des critères de recherche donnés, plus le volume d'offres est important, plus les candidats potentiels ont le choix parmi les offres pouvant leur correspondre. Pour un nombre de candidats potentiels donné (et un nombre de candidatures supposé fixe pour chaque candidat potentiel), le nombre de candidatures reçues par offre diminue avec le nombre d'offres.
- Symétriquement, pour des critères de recherche donnés et un nombre d'offres fixe, le nombre de candidatures par offre augmente avec le nombre de candidats potentiels (ou la demande).

Les niveaux d'offre et de demande peuvent être connus ou traduits indirectement par des indicateurs de conjoncture comme par exemple la taille de la population active ou le nombre de chômeurs.

2. Il est intéressant de remarquer que la terminologie employée en Économie est contraire à celle que nous utilisons. En effet, l'offre désigne alors les chercheurs d'emploi tandis que la demande émane des entreprises.

Nous recensons ici des données de conjoncture disponibles publiquement. Les éléments utilisés en pratique seront explicités dans le chapitre 6. Les sources de données de conjoncture peuvent être classées en deux catégories : les institutions et les sites d'emploi. Parmi les institutions, l'INSEE [3] (cf. tableau 3.1), Pôle Emploi [4] et la DARES [5] (cf. tableau 3.2) fournissent des données sur la population active, les chômeurs, les volumes d'offres et de demandes selon la période et la localisation. Du côté des sites d'emploi, ce sont *Keljob.com* (cf. tableau 3.3), *Apec.fr* [6] (cf. tableau 3.4) et *Monster.fr* (cf. tableau 3.5) qui fournissent des indicateurs traduisant les tendances du marché du recrutement en ligne. Le nombre d'indicateurs fournis pouvant être très grand, nous proposons ici une synthèse de ceux qui semblent pertinents au vu de la problématique.

Indicateur	**Unité**	**Date / période**	**Échelle géographique**
Population légale	effectif	2008	commune, département, ZE, région
Nombre de personnes actives de 15 à 64 ans	effectif	2007	commune, département, ZE, région
Nombre de personnes actives occupées de 15 à 64 ans	effectif	2007	commune, département, ZE, région
Nombre de chômeurs de 15 à 64 ans au sens du recensement	effectif	2007	commune, département, ZE, région
Demandeurs d'emploi de catégories A, B, C au sens du BIT [7]	effectif	au 31 décembre, 2009 et 2010	commune, département, ZE, région
Demandeurs d'emploi de catégorie A au sens du BIT	effectif	au 31 décembre, 2001 à 2010	commune, département, ZE, région

Définition INSEE d'une zone d'emploi (ZE) : espace géographique à l'intérieur duquel la plupart des actifs résident et travaillent, et dans lequel les établissements peuvent trouver l'essentiel de la main d'œuvre nécessaire pour occuper les emplois offerts.

TABLE 3.1 – Indicateurs de conjoncture fournis par l'INSEE

Les valeurs de l'indice de diffusion Apec sont disponibles en ligne au niveau global. Cet indice est spécifique aux offres d'emploi "cadre" diffusées sur Internet. Les valeurs de l'indice Monster sont disponibles en ligne au niveau global et par secteur d'activité. Bien que le baromètre Keljob soit détaillé par secteur, les valeurs exactes ne sont pas disponibles

3. Institut National de la Statistique et des Études Économiques (`www.insee.fr`).

4. `www.pole-emploi.fr`

5. Direction de l'Animation de la Recherche, des Études et des Statistiques, qui dépend du Ministère du Travail, de l'Emploi et de la Santé (`www.travail-emploi-sante.gouv.fr`)

6. Association Pour l'Emploi des Cadres (`www.apec.fr`)

Indicateur	**Unité**	**Date / période**	**Échelle géographique**
Demandeurs d'emploi inscrits en fin de mois à Pôle Emploi en catégories A, B, C	effectif	1997 à 2011, par mois	département, région
Demandeurs d'emploi inscrits en fin de mois à Pôle Emploi en catégorie A	effectif	1997 à 2011, par mois	département, région
Offres d'emploi collectées à Pôle Emploi	effectif	1997 à 2011, par mois	département, région

TABLE 3.2 – Indicateurs de conjoncture fournis par le Pôle Emploi / la DARES

Indicateur	**Unité**	**Date / période**	**Échelle géographique**
Baromètre Keljob du marché de l'emploi sur Internet	indice (base 100)	2008 à 2011, par mois	France

TABLE 3.3 – Indicateur de conjoncture fourni par *Keljob.com*

Indicateur	**Unité**	**Date / période**	**Échelle géographique**
Indice de diffusion des offres d'emploi cadre sur Internet	indice (base 100)	août 2005 à 2011, par mois	France

TABLE 3.4 – Indicateur de conjoncture fourni par *Apec.fr*

Indicateur	**Unité**	**Date / période**	**Échelle géographique**
Monster index de l'emploi en ligne	indice (base 100)	2006 à 2011, par mois	France

TABLE 3.5 – Indicateur de conjoncture fourni par *Monster.fr*

en ligne de manière exhaustive. Nous ne pourrons donc pas l'exploiter dans la pratique.

3.2.2.2 Poste à pourvoir

L'état de la littérature a permis de mettre en évidence que les caractéristiques du poste à pourvoir peuvent influencer les individus dans leur décision de candidature. Dans le processus de candidature (cf. figure 3.1), les caractéristiques du poste vont intervenir dès lors que l'individu en prend connaissance : les premières informations apparaissent dans le titre de l'annonce [8], et le détail dans la description si l'individu décide de visualiser l'offre. En reprenant les facteurs cités précédemment et en complétant la liste, nous proposons de prendre en compte comme facteurs explicatifs :

- le type de contrat (ex : contrat à durée indéterminée, contrat à durée déterminée, stage, contrat d'apprentissage, etc.) ;
- le niveau d'études requis (ex : BEP/CAP, Bac, Bac+2, Bac+5, etc.) ;
- le nombre d'années d'expérience requises ;
- le type de formation recherchée ;
- les compétences recherchées ;
- la localisation (ville / zone d'emploi / département / région administrative / région basée sur les indicatifs téléphoniques) ;
- la tranche de salaire proposée ;
- la fonction sous-jacente aux missions proposées (générique).

3.2.2.3 Entreprise recruteur

La revue de la littérature a mis en évidence que la notoriété de l'entreprise influence la décision de candidature. Les grandes entreprises sont perçues comme plus attractives aux yeux des candidats potentiels. Lorsque le candidat potentiel prend connaissance de l'entreprise qui recrute ou du diffuseur de l'annonce si celle-ci n'est pas connue, cela peut influencer sa décision de poursuivre le processus de candidature à cette offre.

Des recherches nous ont permis d'identifier des études menées par des instituts délivrant

8. Il faut noter que l'individu a déjà connaissance de certaines informations sur le poste car il a effectué une requête sur certains critères via le moteur de recherche du site.

des “indices” d'image et d'attractivité en tant qu'employeur. Le tableau 3.6 synthétise les différents indices identifiés ainsi que leurs caractéristiques majeures.

Comme cela est mis en évidence, l'information sur les indices d'image et attractivité n'est disponible que pour une sélection de grandes entreprises, ou pour les premières du classement. Le périmètre des clients de Multiposting.fr n'étant pas limité aux grandes entreprises, ces données sont manquantes pour une partie importante des clients diffuseurs d'annonces. Nous proposons donc dans le paragraphe suivant une solution alternative.

Comme alternative à ces indices, nous proposons de prendre en compte dans le modèle des facteurs explicatifs permettant de traduire indirectement le degré de notoriété et d'attractivité de l'entreprise. En d'autres termes, nous supposons ces facteurs corrélés avec le degré de notoriété et d'attractivité de l'entreprise et proposons de les employer en substitution. Le modèle statistique retenu permettra par la suite de confirmer ou infirmer l'impact de ces variables sur la performance des campagnes de recrutement. Les facteurs candidats sont :

- le nombre de salariés ;
- le secteur d'activité de l'entreprise ;
- le chiffre d'affaires annuel.

Parfois, l'entreprise qui recrute n'est pas connue des candidats car celle-ci a fait appel à des services extérieurs pour trouver des candidats (par exemple un cabinet de recrutement). Il faut alors s'intéresser aux caractéristiques du “diffuseur de l'annonce”. Ne connaissant pas l'entreprise qui souhaite recruter, les candidats potentiels vont alors s'intéresser à la notoriété et santé financière de l'entreprise qui a diffusé l'annonce.

3.2.2.4 Diffuseur de l'annonce

Les facteurs cités précédemment pour l'entreprise recruteur sont d'intérêt pour décrire le diffuseur de l'annonce :

- le nombre de salariés ;
- le chiffre d'affaire annuel.

Étant donné que le diffuseur de l'annonce peut ou non être l'entreprise employeur, nous proposons d'introduire une variable catégorielle indiquant le type de diffuseur : entreprise

Source	Type d'indice	Périmètre	Méthodologie	Disponibilité
IPSOS	Baromètre de suivi d'image (différence entre le % d'interviewés déclarant avoir une bonne image et le % déclarant avoir une mauvaise image de l'entreprise)	30 grandes entreprises françaises	-Enquête réalisée par téléphone au domicile des interviewés, sur 2 jours -Échantillon de plus de 900 personnes de 18 ans et plus représentatif de la population française (méthode des quotas)	Information publique
TNS Sofres	-Palmarès spontané des entreprises les plus attractives (en % d'interviewés) -Score d'attractivité des entreprises (% de notes comprises entre 5 et 10)	Plus de 100 grandes entreprises françaises	-Enquête réalisée en face-à-face dans plus de 50 écoles (de commerce et d'ingénieurs), sur une vingtaine de jours -Échantillon de plus de 500 élèves	Information publique pour le top 25 des entreprises (écoles de commerce vs écoles d'ingénieurs)
Universum	Indice d'attractivité en tant qu'employeur (rang évalué à partir du nombre de fois où l'entreprise est choisie comme "employeur idéal")	Grandes entreprises dans le monde	Enquête réalisée auprès de 130 000 étudiants issus d'écoles/universités prestigieuses	Information publique : top 50 des entreprises (commerce vs ingénierie)
Randstat (Awards)	Score d'attractivité en tant qu'employeur (% des interviewés qui voudraient travailler pour l'entreprise)	200 plus grandes entreprises françaises	Enquête réalisée en ligne auprès d'un échantillon représentatif de 10 000 salariés potentiels (étudiants et population active de 18 à 65 ans)	Information publique : entreprises nominées et lauréates de chaque secteur
Great Place to work	Palmarès des entreprises les plus attractives en tant qu'employeur (rang fourni à partir du score obtenu)	Entreprises en Europe	Enquête réalisée par courrier ou par internet auprès des salariés des entreprises	Information publique : top 30 des entreprises françaises, top 25 des grandes entreprises en Europe, top 50 des PME en Europe

TABLE 3.6 – Indices d'image et d'attractivité des entreprises

(hors SSII), SSII, cabinet de recrutement, agence d'intérim, PME, collectivité. Ces différents types de diffuseurs ont différentes stratégies de recrutement et peuvent donc susciter des réactions différentes auprès des candidats potentiels.

3.2.3 Les facteurs flexibles

3.2.3.1 Site d'emploi

Rigoureusement, le site d'emploi doit être considéré comme un facteur semi-flexible, car bien que le choix du site d'emploi puisse être sujet à des recommandations, choisir un site implique l'acceptation de l'ensemble de ses caractéristiques comme une donnée. Chaque caractéristique du site ne peut pas être choisie indépendamment. Pour choisir le(s) site(s) adéquat(s) pour une campagne de recrutement donnée, nous proposons de nous baser sur les informations de signalétique :

- le type de portail (généraliste, spécialisé, blog, réseau social, école / association d'anciens, etc.) ;
- la spécialisation si elle existe (fonction, secteur, type de contrat, discrimination positive, etc.) ;
- la gratuité (site payant ou site gratuit pour la diffusion d'annonces).

3.2.3.2 Rédaction et contenu de l'annonce

Titre. Sur la plupart des sites d'emploi, les candidats potentiels visualisent une liste d'offres (générée par la requête qu'ils ont effectuée sur le moteur de recherche) avant de visualiser le contenu d'une offre d'emploi. La liste générée présente le titre des offres ainsi que d'autres éléments sur le poste dépendant du site en question. Le candidat potentiel est donc d'abord confronté à une première décision : celle de cliquer ou non sur le lien défini par le titre de l'offre afin d'en consulter le détail. Par conséquent, nous pouvons supposer que la manière de rédiger le titre de l'annonce va avoir une influence sur la décision de visualiser le texte complet de l'offre (et par suite la décision de candidature).

Pour étudier l'influence du titre de l'annonce, nous proposons d'analyser les facteurs suivants :

- longueur du titre (nombre de caractères) ;

- nombre de mots dans le titre ;
- présence/absence de mots-clés dans le titre (une liste de mots-clés candidats pertinents doit être établie, voir chapitre 4 pour la méthodologie d'obtention).

Description. Au moment d'accéder au contenu de l'annonce, le candidat potentiel a déjà connaissance de certaines caractéristiques du poste : éventuellement l'entreprise qui recrute, le titre, ainsi que les critères sur lesquels il a effectué sa recherche. Une fois que le candidat potentiel accède au contenu de l'annonce, il prend alors connaissance des détails sur le poste à pourvoir : description de la société qui recrute (sauf si le diffuseur n'est pas l'entreprise qui recrute et qu'il ne souhaite pas donner de détails sur son client), description des missions et description du profil recherché. En fonction de tous ces critères, il va alors décider si le poste lui correspond ou non et envoyer ou non une candidature. Nous proposons d'intégrer au sein du modèle statistique le texte de l'annonce grâce à des méthodes de fouille de données textuelles. En effet, une partie importante des informations dont nous disposons sur l'offre d'emploi est située dans le texte de l'annonce, et ce, de manière très peu ou non structurée (c'est un texte libre, parfois découpé en trois parties comme évoqué ci-dessus). Nous allons donc utiliser la fouille de textes afin d'analyser ce contenu (identification des mots, de leurs rôles grammaticaux, etc.) et d'en extraire des connaissances exploitables au sein d'un modèle statistique. Les mots-clés présents dans le texte ainsi que les composantes extraites par analyse sémantique serviront de facteurs explicatifs au sein de notre modèle (cette partie sera développée dans le chapitre 4).

Même si le candidat potentiel s'intéresse au poste dont il est en train de consulter le descriptif, d'autres critères peuvent également influencer son choix. En effet, la construction de l'annonce, le style rédactionnel (complexité du texte, ponctuation), ainsi que la richesse de l'expression peuvent faire "pencher la balance" dans un sens ou dans l'autre.

Pour ce qui est de la construction de l'annonce, nous suggérons de prendre en compte la répartition en volume des différentes parties de l'annonce :

- proportion (en longueur) de la partie "descriptif société" ;
- proportion (en longueur) de la partie "missions" ;
- proportion (en longueur) de la partie "profil recherché".

Pour obtenir des suggestions d'indicateurs traduisant le style rédactionnel d'un texte ou la richesse de l'expression, nous avons parcouru la littérature de l'analyse de données textuelles [e.g. Monière et Labbé 2002; Habert et al. 2000]. Les indicateurs doivent être envisagés au vu de leur pertinence vis-à-vis du type de textes que nous étudions : des textes courts (90% des annonces d'emploi de notre base ont une longueur comprise entre 1000 et 3000 caractères), très stéréotypés.

Dans un premier temps, nous nous intéressons à la longueur du texte, ainsi qu'à la longueur et structure des phrases. Une phrase est définie par une suite de mots délimitée par le point, les trois points de suspension, le point d'interrogation ou le point d'exclamation. Par exemple, des phrases longues dans une annonce d'emploi peuvent rendre la compréhension du texte plus difficile relativement à des annonces plus fragmentées avec de courtes phrases. De même, l'usage intensif de la virgule indique une phrase compliquée. Nous suggérons les indicateurs suivants :

- longueur du texte (égale au nombre total de caractères) ;
- nombre de phrases ;
- nombre de mots ;
- nombre moyen de mots par phrase ;
- nombre moyen de virgules par mot ;
- nombre moyen de parenthèses par phrase ;
- nombre de points d'exclamation (effectif absolu car très faible en pratique) ;
- nombre moyen de caractères par mot.

Nous nous intéressons ensuite à l'usage des différentes catégories grammaticales dans le texte. Celles-ci sont reconnues grâce à un lemmatiseur qui identifie et étiquette les mots du texte. Nous proposons de calculer les proportions de différentes catégories de mots relativement au nombre total de mots :

- proportion de verbes ;
- proportion de noms communs.

Un excédent du groupe verbal signifie que l'accent est mis sur l'action tandis qu'un excédent de noms implique la stabilité.

Le nombre de mots dans un texte peut également être appelé le nombre d'occurrences. Dans

l'ensemble de ces occurrences, un certain nombre de formes distinctes peuvent être relevées (le nombre de mots différents). Le nombre d'hapax désigne le nombre de formes qui n'ont qu'une seule occurrence dans le texte. Enfin, les nombres d'hapax et de formes distinctes sont utilisés pour traduire la richesse et la diversité du vocabulaire employé. Étant donné que les textes ont des longueurs variables, nous proposons les indicateurs suivants :

- nombre de formes divisé par le nombre d'occurrences ;
- nombre d'hapax divisé par le nombre de formes.

Un inconvénient bien connu de ces indicateurs est que de par leur nature, ils décroissent avec la longueur du texte. En effet, une caractéristique générale des textes est que le nombre de formes distinctes et d'hapax augmente moins que proportionnellement avec le nombre d'occurrences.

Le calcul des indicateurs cités ci-dessus peut être décliné pour chaque partie de l'annonce (si nous disposons de cette information).

3.2.3.3 Moment de diffusion

Cette section est consacrée à une étude sur l'impact à court terme et à long terme du moment de diffusion de l'annonce. Cependant, le moment de diffusion ne pourra pas être pris en compte en tant que facteur explicatif car sur la plupart des sites d'emploi, le décalage entre diffusion via l'outil et mise en ligne sur le site est très important (il peut aller de quelques heures à plusieurs jours).

Heure de diffusion. Les annonces peuvent être diffusées via Multiposting.fr à toute heure de la journée. L'heure exacte de mise en ligne sur le(s) site(s) d'emploi choisi(s) dépend de l'heure à laquelle l'annonce est postée avec l'outil. Nous nous intéressons à l'impact que peut avoir l'heure de mise en ligne sur la performance finale de l'annonce. En effet, dans les moteurs de recherche d'annonces des sites d'emploi, la date de publication a une grande importance sur le classement des offres dans la liste qui est présentée au candidat potentiel. Plus la date est récente, plus l'annonce aura de chances d'être présentée parmi les premiers résultats de la recherche.

Pour obtenir de premiers éléments de réponse, nous avons recours à une extraction de la

base de données de Multiposting.fr. Ne connaissant pas le moment exact de mise en ligne sur les différents sites, nous devons nous focaliser sur un site pour lequel la diffusion est très rapide : Monster.fr. Notre échantillon d'étude se compose de 6 930 annonces postées sur Monster.fr entre le 1er janvier 2010 et le 5 août 2011, et ayant été diffusées sur le site pendant 25 jours ou plus[9]. L'étude du moment de réception des candidatures sur un sous-échantillon de 45 annonces ayant obtenu 25 clics ou plus au cours des premières 24h de diffusion indique un délai compris entre 60 et 160 minutes pour obtenir les premières candidatures. Le délai avant que l'annonce soit visible auprès des candidats potentiels est donc approximativement de 1h à 2h40. Bien que conscients des limites que cela implique, nous considérons dans ce qui suit que le délai pour la visibilité sur Monster.fr est constant d'une annonce à l'autre. Ce délai est estimé par la médiane évaluée sur le sous-échantillon de 45 annonces. Le délai médian étant de 1h30, l'erreur maximale peut alors être estimée à 1h10. Le moment de diffusion via l'outil Multiposting.fr est donc translaté de 1h30 pour approcher le moment de mise en ligne réel. Cependant, afin de donner de la robustesse aux statistiques, nous étudions des créneaux agrégés de deux heures.

La figure 3.2 représente les répartitions des annonces postées avec mode de candidature par e-mail et des CV reçus en fonction du créneau horaire. Symétriquement, la figure 3.3 représente les répartitions des annonces postées avec mode de candidature par URL et des clics en fonction du créneau horaire.

Les répartitions des CV reçus et clics de redirection en fonction du créneau horaire étant très proches, l'heure de candidature des visiteurs du site d'emploi semble indépendante du moment de diffusion des annonces (nous ne détectons pas de schéma), mais plutôt liée aux habitudes de connexion de ces internautes. Ces figures montrent que les 3% d'annonces avec mode de candidature par e-mail (ou les 7% d'annonces avec mode de candidature par URL) arrivant en ligne de 20h à 22h se retrouvent en tête de liste pour les 30% de candidats potentiels en ligne de 20h à 10h le lendemain matin. De plus, la concurrence avec des offres postées de manière proche dans le temps est faible, contrairement aux annonces arrivant en ligne de 10h à 20h pour lesquelles la concurrence avec les autres offres est forte étant

9. Cette durée est choisie arbitrairement de sorte que nous puissions étudier les retours journaliers sur presque un mois (les 25 premiers jours de diffusion), la durée "standard" de diffusion sur une grande partie des sites d'emploi.

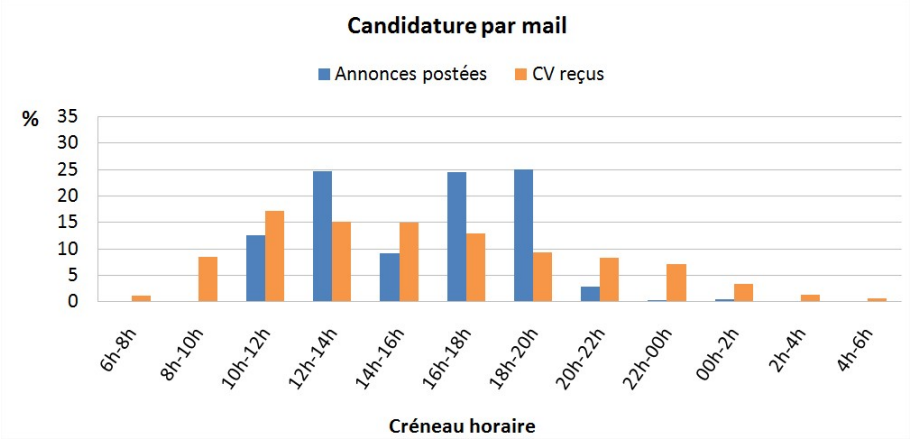

FIGURE 3.2 – Répartitions des annonces postées (candidature par e-mail) et CV reçus en fonction du créneau horaire

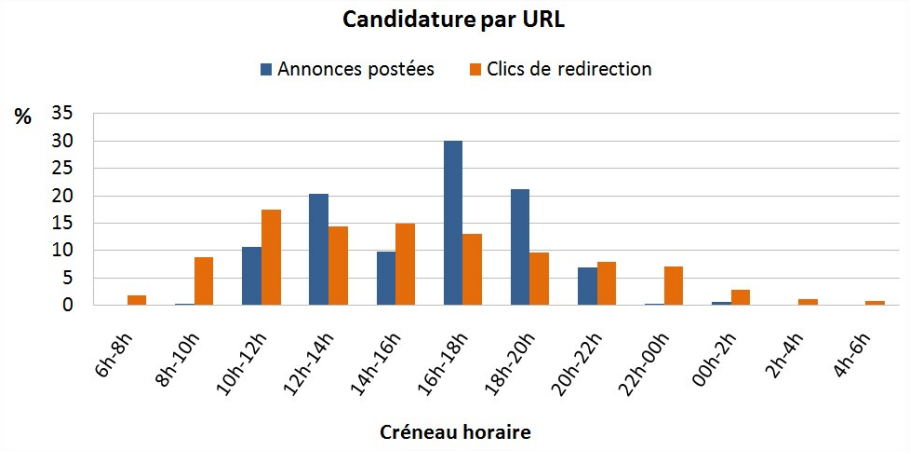

FIGURE 3.3 – Répartitions des annonces postées (candidature par URL) et clics de redirection en fonction du créneau horaire

donné les flux importants d'annonces tout au long de la journée.

Dans la pratique nous n'avons pas de contrôle direct sur l'heure de mise en ligne de l'annonce. Pour la suite des analyses, nous nous focalisons donc sur l'étude des créneaux horaires de diffusion via Multiposting.fr calculés à partir de l'heure enregistrée en base de données. La proportion d'offres postées avant 8h et après 20h étant égale à 1%, nous éliminons ces dernières de l'échantillon étudié. L'échantillon des annonces est divisé en 12 sous-ensembles : 6 créneaux horaires (créneaux de 2h de 8h à 20h) et 2 modes de candidature (cf. figure 3.4).

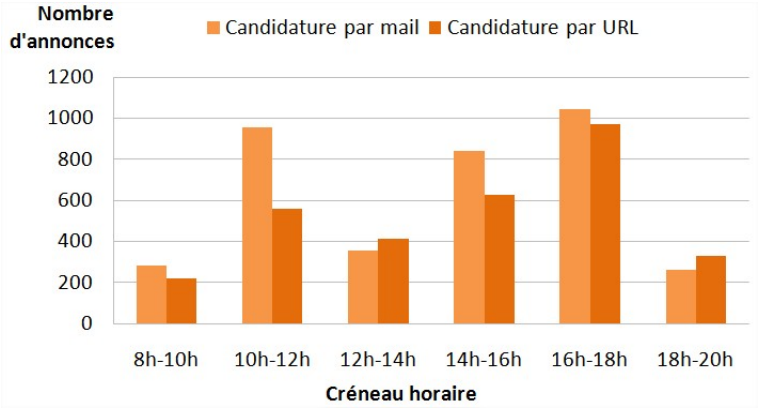

FIGURE 3.4 – Effectifs des annonces postées (candidature par e-mail ou URL) en fonction du créneau horaire

Nous étudions le volume de retours obtenus à l'issue des premières 24h de diffusion en fonction de l'heure où l'annonce est postée via l'outil. La figure 3.5 représente le nombre de retours journaliers (CV ou clics selon le mode de candidature) moyen en fonction du nombre de jours de diffusion. Chaque créneau horaire est représenté par une courbe. Nous avertissons le lecteur que ces deux graphiques sont présentés dans un but illustratif et ne permettent pas de tirer une conclusion générale sur l'effet de l'heure de diffusion (possible présence de biais non identifiés, étude sur un seul site, etc.).

Les courbes moyennes de réception de candidatures sont globalement assez proches pour les différents créneaux horaires. Toutefois, le créneau 18h-20h semble apporter de plus nombreux retours en moyenne à l'issue des premières 24h (également le créneau 14h-16h pour les candidatures par URL). Afin de savoir si ces écarts observés pour le premier jour de diffusion peuvent être considérés comme significatifs, nous utilisons un modèle ANOVA

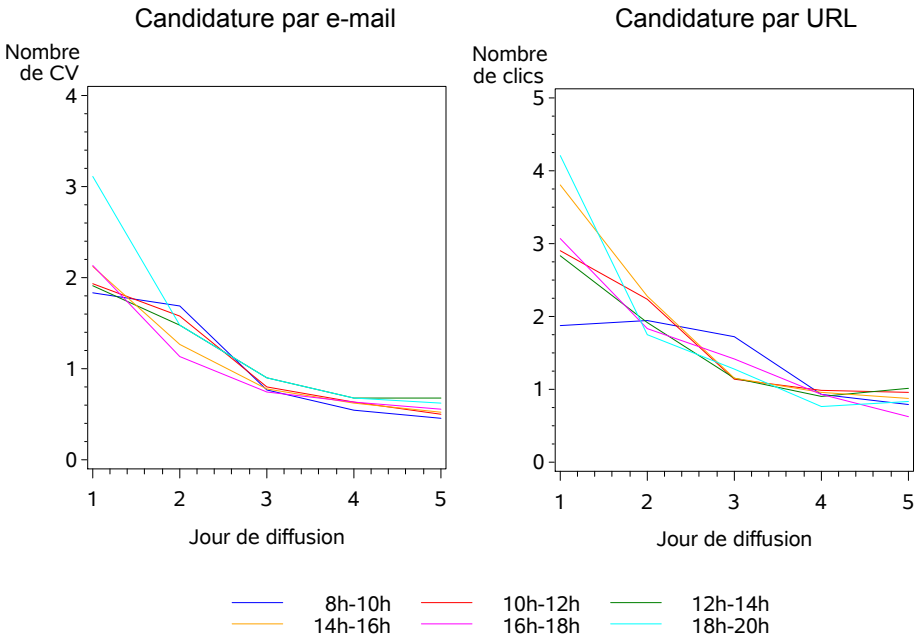

FIGURE 3.5 – Retours journaliers moyens par annonce en fonction du nombre de jours de diffusion et du créneau horaire de diffusion

à un facteur. Le test d'homogénéité des variances [Levene 1960] effectué dans un premier temps indique un non rejet de l'hypothèse d'égalité des variances. Le test d'égalité des moyennes permet de confirmer qu'il existe des différences significatives entre les retours moyens associés aux différents horaires. De plus, une annonce postée entre 18h et 20h obtient significativement plus de retours qu'une annonce postée pendant les autres tranches horaires (pas de différence significative avec le créneau 14h-16h pour les candidatures par URL).

Cette différence observée le premier jour de diffusion n'est plus observée au bout de 25 jours de diffusion. En effet, les tests statistiques effectués ne montrent alors pas de différence significative entre les moyennes associées aux différents créneaux horaires.

L'heure de diffusion dans la journée peut donc constituer un facteur candidat pour expliquer des variations dans la performance de l'annonce durant les premiers jours de diffusion. Toutefois, l'étude de son impact est très complexe car celui-ci dépend du fonctionnement du moteur de recherche (influence de la date de diffusion plus ou moins importante sur le classement des offres relativement à celle de la pertinence de l'offre au regard de la recherche

effectuée).

Remarque 3 *Nous connaissons les fonctions associées aux postes à pourvoir par les annonces de notre échantillon. Nous avons étudié la répartition des candidatures en fonction de l'heure de la journée pour les différentes fonctions ("Commercial / Vente", "Informatique et Technologies", "Ressources Humaines", etc.), mais n'avons pas pu faire ressortir de résultats. En effet, nous avons observé des résultats non consistants (peu d'écart avec la répartition globale et répartitions différentes entre les deux modes de candidature) ne permettant pas de faire ressortir des différences au niveau des heures de candidatures entre les différentes fonctions et par rapport à la moyenne.*

Jour de diffusion. Nous nous intéressons maintenant à l'impact que peut avoir le jour de diffusion sur la performance de l'annonce. Le jour est ici étudié en tant que jour de la semaine, les modalités du facteur étant le lundi, mardi, ..., dimanche. Nous reprenons l'échantillon étudié dans le paragraphe précédent. La figure 3.6 représente la répartition des annonces postées et des candidatures reçues via Monster.fr en fonction du jour de la semaine, pour chaque mode de candidature. Les annonces étant postées au plus tard à 20h, et étant donné le délai de diffusion sur le site Monster.fr, nous pouvons supposer que le jour de diffusion est correctement identifié pour toutes les annonces.

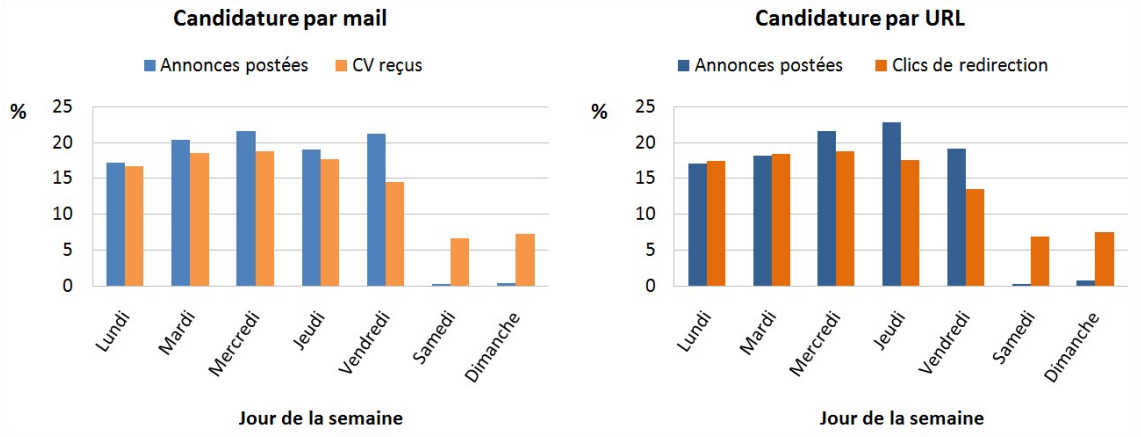

FIGURE 3.6 – Répartitions des annonces postées et candidatures en fonction du jour de la semaine

86% des candidatures sont envoyées du lundi au vendredi, avec une légère baisse le vendredi

(environ 14% des candidatures). Le week-end, le volume de candidatures est beaucoup plus faible, avec environ 7% de candidatures chaque jour. Nous souhaitons étudier l'impact du jour de diffusion sur le volume des retours observés durant les premiers jours de diffusion de l'annonce.

L'effectif des annonces postées le samedi ou le dimanche étant très faible (cf. figure 3.7), nous les retirons de l'échantillon étudié. L'échantillon des annonces est donc divisé en 10 sous-ensembles : 5 jours de la semaine (de lundi à vendredi) et 2 modes de candidature.

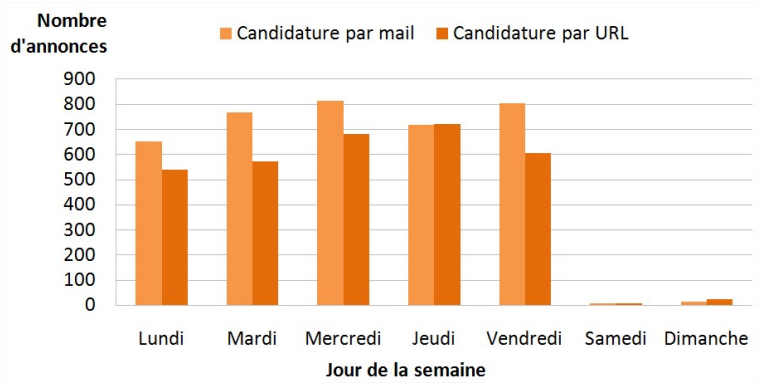

FIGURE 3.7 – Effectifs des annonces postées (candidature par e-mail ou URL) en fonction du jour de la semaine

Comme précédemment, nous étudions le volume de retours obtenus à l'issue des premières 24h de diffusion en fonction du jour où l'annonce est postée via l'outil. La figure 3.8 représente le nombre de retours journaliers (CV ou clics selon le mode de candidature) moyen en fonction du nombre de jours de diffusion. Chaque jour de la semaine est représenté par une courbe. De même que pour l'heure de diffusion, ces deux graphiques sont présentés dans un but illustratif mais ne permettent pas de tirer une conclusion générale sur l'effet du jour de diffusion.

Pour les deux modes de candidature, les retours moyens à l'issue du premier jour de diffusion sont plus élevés pour les annonces ayant été diffusées un lundi, un mardi, ou un mercredi. De plus, les remontées observées à l'issue du quatrième et troisième jour de diffusion pour les courbes associées respectivement au jeudi et vendredi correspondent à une hausse du volume de candidatures le lundi.

Le test d'homogénéité des variances effectué dans le cadre d'un modèle ANOVA permet

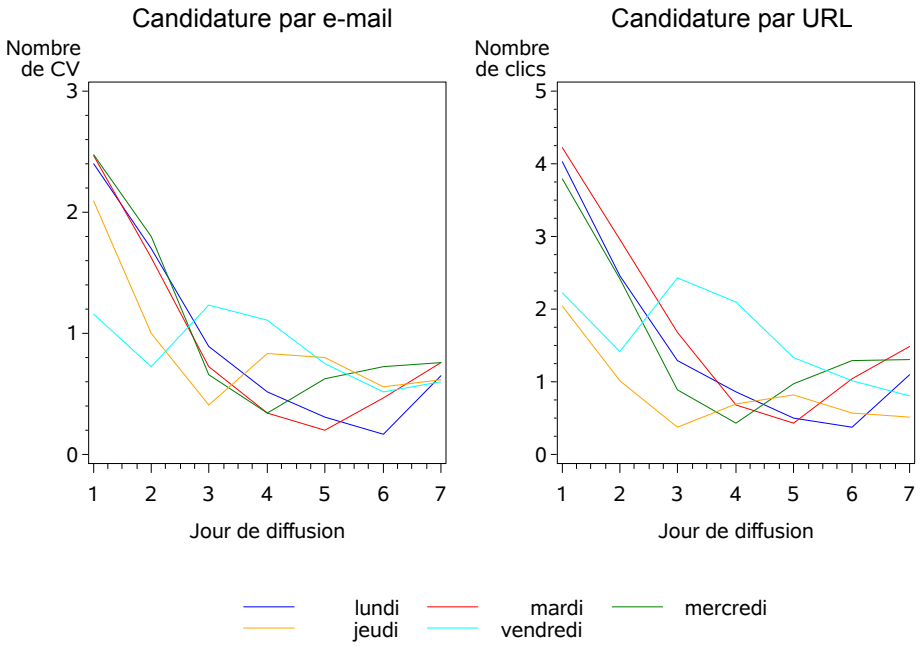

FIGURE 3.8 – Retours journaliers moyens par annonce en fonction du nombre de jours de diffusion et du jour de diffusion

d'accepter l'hypothèse d'égalité des variances. Le test d'égalité des moyennes indique des différences très significatives entre les jours de la semaine. Les retours moyens à l'issue des premières 24h de diffusion sont significativement plus élevés pour les annonces diffusées le lundi, mardi ou mercredi, comparé aux annonces diffusées le vendredi (et le jeudi pour les annonces avec candidature par URL). Toutefois, les différences ne sont plus significatives au 25^{e} jour de diffusion de l'annonce.

Remarque 4 *Selon le profil des candidats, les jours de recherche (et par suite de candidature) peuvent différer. Nous avons observé les répartitions des candidatures en fonction du jour de la semaine associées à chaque fonction. Cependant, les écarts par rapport à la répartition globale sont faibles et la répartition des annonces postées pour une fonction peut influencer celle des candidatures reçues. Nous n'avons donc pas pu faire ressortir de résultat sur les comportements de connexion des candidats pour les différentes fonctions.*

Mois de diffusion. Pour certains types de contrat, comme les stages ou les formations par apprentissage, les candidats potentiels sont disponibles pendant une période précise dé-

terminée par leur formation. De plus, les candidats effectuent généralement leurs recherches plusieurs semaines ou mois à l'avance et le recruteur doit être visible à ce moment-là. Si la campagne de recrutement est lancée trop tardivement, il y aura moins de candidats potentiels et leur profil sera de moins bonne qualité (cf. section 3.1.5), les meilleurs candidats ayant déjà trouvé un poste. Le mois de diffusion est donc à prendre en considération pour les contrats impliquant un calendrier de formation.

3.3 Synthèse

A partir de la littérature et de nos propres hypothèses (formulées grâce aux connaissances que nous avons acquises du domaine ou à partir de l'observation de statistiques descriptives), nous avons proposé un ensemble de facteurs potentiels pour expliquer la performance d'une annonce diffusée sur Internet. Nous avons divisé ces facteurs en deux sous-ensembles : les facteurs inflexibles et les facteurs flexibles. Les facteurs inflexibles devront être considérés dans le modèle mais sont déterminés par le poste, l'entreprise ou des facteurs de la conjoncture et sont donc non modifiables. En revanche, le recruteur a une influence directe ou indirecte sur les facteurs flexibles.

Certains de ces facteurs pourront être obtenus directement grâce aux informations enregistrées dans la base de données Multiposting (informations sur le poste à pourvoir comme le niveau d'expérience, le niveau d'études requis, le type de contrat, etc.). D'autres nécessiteront des pré-traitements pour être extraits et rendus exploitables (localisation, indicateurs sur le style rédactionnel, etc.). Nous verrons par la suite que certains facteurs ne seront pas exploitables en pratique (par exemple, le salaire est très peu renseigné et les parties de l'annonce sont rarement identifiées). Par ailleurs, pour obtenir certains facteurs, des études plus conséquentes seront nécessaires (cf. chapitre 4). Des techniques d'apprentissage et de fouille de données textuelles seront utilisées pour extraire la fonction associée à l'offre d'emploi, ainsi que le vecteur des descripteurs du texte de l'offre (mots-clés, composantes issues de l'analyse sémantique). Enfin, des données externes devront être utilisées pour les facteurs décrivant la conjoncture ou le recruteur/diffuseur de l'annonce.

Chapitre 4

Fouille de textes appliquée aux offres d'emploi et extraction des connaissances

4.1 Présentation d'une offre d'emploi sur Internet

La publication massive d'offres d'emploi sur les sites web dédiés à l'emploi a nécessité la mise en place de nomenclatures afin d'offrir aux internautes et candidats potentiels la possibilité d'atteindre les offres d'emploi qui les intéressent à l'aide de moteurs de recherche. Ainsi, en plus d'un outil de recherche par mots-clés parcourant l'ensemble du texte de l'offre, les sites d'emploi proposent une liste de critères à plusieurs modalités pour la recherche des offres d'emploi. Cette liste de critères est spécifique à chaque site, ainsi que la liste des modalités correspondantes. Cependant, on peut citer une liste de critères communément trouvés sur les sites d'emploi et dont les modalités peuvent être facilement rapprochées :

- le nom de la société qui diffuse l'annonce,
- le type de contrat,
- le niveau d'études requis,
- le nombre d'années d'expérience requises,
- le pays,
- le code postal.

De plus, les offres d'emploi sont communément présentées à travers trois champs texte :

- le descriptif de la société qui recrute,

- le titre,
- la description de l'annonce [1].

Une offre présente également des champs semi-structurés qui ne peuvent être exploités sans pré-traitement :

- la localisation (champ texte libre qui peut être exploité à travers le code postal),
- le salaire annuel (donné par une fourchette salaire minimum – salaire maximum). Très peu renseigné dans la pratique et non exploitable, nous n'étudierons pas les effets du salaire proposé.

La fonction associée au poste à pourvoir est la plupart du temps représentée par une variable catégorielle à un ou deux niveaux. Cependant, malgré la structuration de cette information, celle-ci demeure peu exploitable en pratique car il existe autant de nomenclatures que de sites d'emploi, et les données ne peuvent donc être agrégées qu'au niveau d'un seul site d'emploi. Pour permettre des analyses statistiques générales par fonction et la comparaison entre les sites d'emploi, il est nécessaire d'uniformiser l'ensemble de ces nomenclatures : c'est l'objet de la section 4.3. Les objectifs poursuivis à travers la classification des offres d'emploi y sont présentés plus en détail.

La figure 4.1 présente un exemple d'offre d'emploi diffusée sur le site d'emploi généraliste *Monster.fr* et les champs qui peuvent en être extraits. Comme illustré, une offre présente simultanément des champs structurés (critères définis par un ensemble de modalités) et des champs non-structurés ou semi-structurés. Nous devons donc gérer des données de types différents dans les analyses que nous menons sur les offres d'emploi. Beaucoup d'informations sont contenues dans le texte des offres et il nous faut les extraire pour enrichir la description des offres d'emploi.

Après un état de l'art des techniques de la fouille de textes et des travaux spécifiques aux offres d'emploi, nous détaillons les applications y faisant appel et permettant de répondre à nos problématiques. La section 4.3 est dédiée à la classification des offres d'emploi : nous souhaitons étiqueter chaque offre en fonction du type de poste à pourvoir. La section 4.4 présente la méthode que nous proposons pour extraire des mots-clés du texte afin d'enrichir les facteurs explicatifs et améliorer la qualité du modèle de prédiction.

1. La description de l'annonce peut parfois contenir également le descriptif de la société qui recrute.

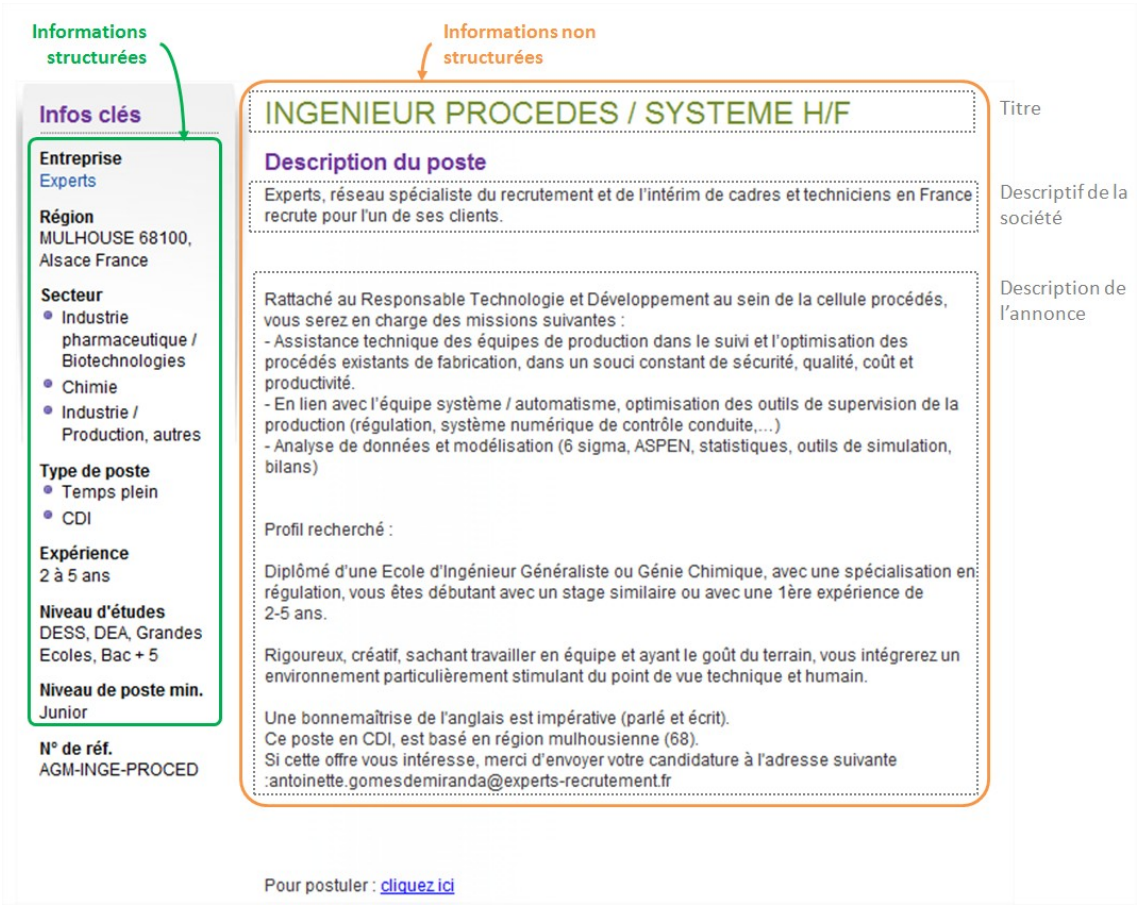

FIGURE 4.1 – Exemple d'offre sur le site d'emploi *Monster.fr*

4.2 État de l'art

4.2.1 État de l'art des techniques de fouille de textes

Cette section est destinée à présenter les principales étapes de la préparation des textes pour l'extraction des connaissances. Selon les applications, certains auteurs font le choix d'appliquer très peu de traitements, tandis que d'autres préfèrent en combiner plusieurs afin de réduire de manière importante le volume de termes avant les analyses. L'ensemble des textes (ou documents) étudiés est appelé "corpus". Dans un texte, un terme (ou une forme) a un certain nombre d'occurrences.

La figure 4.2 présente une vue d'ensemble du processus de préparation des textes pour l'extraction des connaissances.

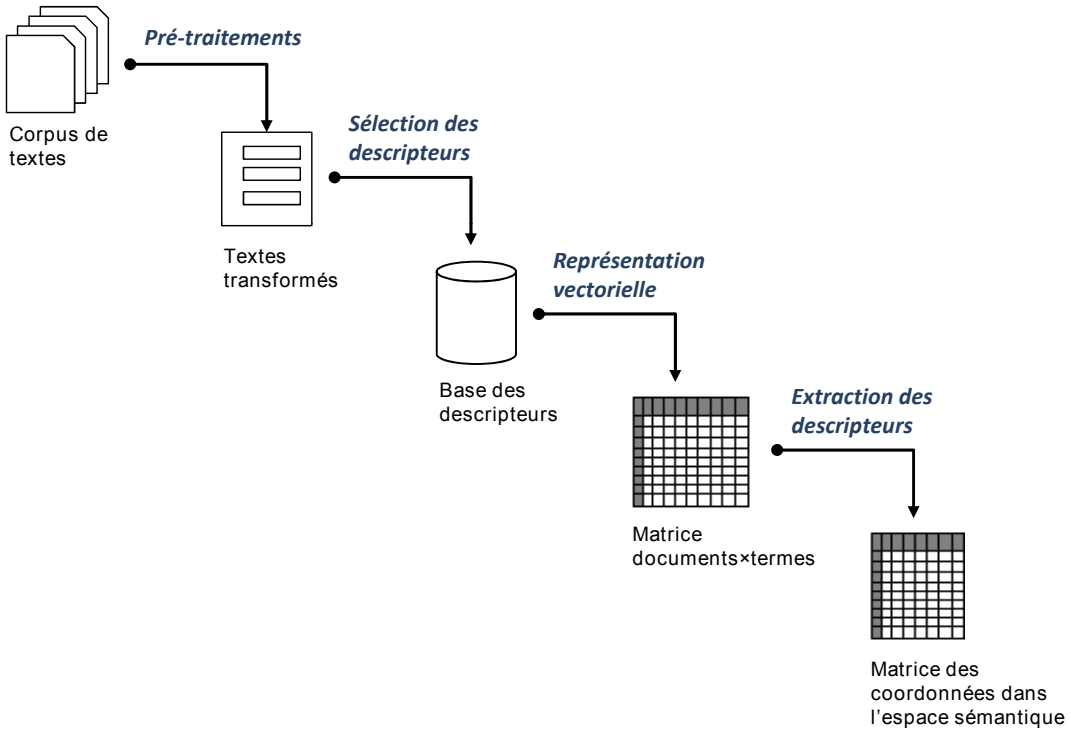

FIGURE 4.2 – Vue d'ensemble du processus de préparation des textes

4.2.1.1 Applications de la fouille de textes

Nous débutons cette section par une présentation des différents types d'applications mises en œuvre sur un corpus de documents et nécessitant un traitement des textes au préalable. Nous pouvons distinguer les applications ayant pour objectif l'automatisation d'une tâche (classification, étiquetage de documents, etc.) de celles ayant pour objectif la compréhension approfondie d'un corpus de textes à travers l'analyse de son vocabulaire et de la stylistique de ses auteurs.

Parmi les applications les plus communes ayant pour finalité l'automatisation d'une tâche, nous retenons :

- La catégorisation de documents, qui consiste à affecter automatiquement un document à une ou plusieurs catégories connues au préalable à partir de l'analyse du texte et d'un algorithme de classification supervisée. La détection de spam [Cormack 2006; Gansterer et al. 2007] et la catégorisation selon le thème [Joachims 1998; Kessler et al. 2006] sont deux applications très répandues.
- Le clustering de documents. Par opposition à la catégorisation, le clustering vise à construire des classes de documents homogènes du point de vue du contenu textuel à l'aide d'un algorithme de classification non supervisé [Steinbach et al. 2000; Zhao and Karypis 2002; Azzag et al. 2006].
- La recherche de documents ("document retrieval" en anglais) et plus généralement la recherche d'information ["information retrieval" en anglais ; Baeza-Yates and Ribeto-Neto 1999; Manning et al. 2008]. Ce domaine de l'informatique englobe les travaux dont le but est d'obtenir un classement des documents relativement à leur pertinence vis-à-vis d'une requête utilisateur.
- Le résumé automatique de texte.
- L'extraction automatique de mots-clés pour représenter un document.

Parmi les applications visant à obtenir des conclusions sur un corpus de texte spécifique, nous citons :

- La lexicométrie (étude statistique de l'usage des mots) et la stylométrie (mesure du style). Ces disciplines de l'analyse textuelle sont par exemple utilisées pour la

comparaison de discours [Monière et Labbé 2002], la caractérisation lexicale d'un corpus de conversations [Cailliau et Poudat 2008], ou encore l'attribution d'auteur [Labbé et Labbé 2010].

- Le traitement statistique des questions ouvertes dans les enquêtes [Lebart 2003].

Dans ce type d'application, les pré-traitements appliqués aux données sont généralement peu nombreux afin de ne pas modifier la sémantique du texte.

4.2.1.2 Pré-traitements usuels

Lemmatisation, racinisation et étiquetage morphosyntaxique. La taille du vocabulaire initial d'un corpus de documents peut parfois dépasser une centaine de milliers de termes distincts. Des pré-traitements permettent une première réduction de la taille du vocabulaire en éliminant des termes jugés non pertinents. En premier lieu, les auteurs ont parfois recours à l'étiquetage morphosyntaxique des termes et à la lemmatisation/racinisation. L'étiquetage morphosyntaxique consiste à identifier la catégorie grammaticale associée à chaque terme au sein de la phrase (nom, verbe, adjectif, pronom, déterminant, adverbe, etc.). Cette identification permet ensuite la lemmatisation, procédé qui prend en compte la flexion [2] des mots afin de les ramener au lemme (masculin singulier pour un adjectif, infinitif pour un verbe conjugué, etc.). Initialement, les différentes flexions d'un mot rencontrées dans le texte sont associées à des formes (ou termes) différentes. La lemmatisation permet alors de réunir toutes les flexions rencontrées sous un unique terme. Un des algorithmes les plus utilisés pour la lemmatisation et adapté à la langue française est l'algorithme de Schmid [1994]. Contrairement à la lemmatisation, les algorithmes de racinisation suivent une démarche de troncation visant à réduire les différentes formes d'un mot à une racine commune (ils ne prennent pas en compte la flexion des mots). L'algorithme le plus connu pour la racinisation ("stemming" en anglais) dans la langue anglaise est l'algorithme de Porter [1980]. Bien que développé également pour la langue française, cet algorithme est jugé peu adapté au français, langue ayant un fort taux de flexion [Namer 2000]. Par ailleurs, certains auteurs (en particulier en analyse de textes) préfèrent garder les formes initiales telles quelles sans appliquer de lemmatisation/racinisation. Ils considèrent que ces

2. Variation de la forme d'un mot en fonction de facteurs grammaticaux.

regroupements sont à l'origine d'une perte d'information importante dans la mesure ou des mots peuvent avoir des sens différents selon qu'ils sont, par exemple, au singulier ou au pluriel.

Filtrage des termes non pertinents. Le praticien peut décider de réduire de manière significative le nombre de termes distincts en :

- Filtrant les mots vides (ou mots-outils : "de", "et", "ou", "mais", "on", "pour", etc.). Ces mots apparaissant dans l'ensemble des documents, "stop words" en anglais, sont éliminés car ne contiennent pas d'information sémantique et ne permettent pas la discrimination entre les documents. Leur liste est spécifique à chaque langue et définie de manière arbitraire. Certains outils linguistiques proposent des listes de mots vides[3]. Une alternative permettant également l'élimination des mots vides consiste à filtrer les termes les plus fréquents (ces termes présents sur l'ensemble des textes n'apportent rien pour leur caractérisation).
- Filtrant certaines catégories grammaticales qu'il juge sans intérêt pour sa problématique (par exemple les prépositions, déterminants, pronoms, etc., mais cela dépend du type d'application). Dans certaines applications, les auteurs choisissent de conserver uniquement les noms, verbes et adjectifs.
- Filtrant les termes les moins fréquents ou en fixant un nombre de documents minimum dans lesquels doit apparaître le terme. Ces suppressions ne sont pas forcément justifiées d'un point de vue sémantique mais dans la pratique, elles sont contraintes par des aspects techniques car un nombre minimum d'occurrences est nécessaire pour pouvoir appliquer des méthodes statistiques à base d'apprentissage.

Les seuils supérieur et inférieur du nombre d'occurrences acceptées pour un terme donné sont définis de manière arbitraire et dépendent du type d'application.

4.2.1.3 Sélection des descripteurs

Des techniques sont utilisées afin de réduire le nombre de termes étudiés tout en conservant l'information pertinente. La réduction du nombre de termes permet de gagner en temps de

3. Par exemple, se reporter au langage *Snowball* (`http://snowball.tartarus.org/`).

calcul lorsque les algorithmes sont appliqués à des matrices de très grande dimension comme c'est le cas en fouille de données textuelles. Elle permet parfois également d'améliorer les résultats obtenus grâce à l'élimination des termes constituant un bruit. Nous distinguons différents types d'indices permettant d'attribuer des scores aux termes en fonction de leur pouvoir discriminant selon qu'il s'agit d'une application dans un cadre supervisé (catégorisation) ou non supervisé (clustering, recherche d'information). Les termes sont ensuite ordonnés selon leur score et éliminés sur la base d'un seuil minimal fixé ou du nombre de termes que l'on souhaite conserver. Ce choix est arbitraire et demeure une tâche délicate car il dépend du type d'application. Idéalement, des expérimentations doivent être menées afin de décider du seuil/du nombre de termes à conserver.

Cas supervisé. Dans le cas supervisé, les catégories de documents sont connues et le but est d'identifier la catégorie d'un nouveau document à partir de son contenu textuel. Des indicateurs permettent d'évaluer le pouvoir discriminant d'un terme vis-à-vis des différentes catégories en se basant sur les occurrences de ce terme au sein de chaque catégorie. Parmi les méthodes les plus souvent utilisées :

- La statistique du χ^2. La statistique $\chi^2(t_j, c_i)$ mesure le manque d'indépendance entre le terme t_j et la catégorie c_i. Elle évalue l'importance de l'écart entre la fréquence observée et la fréquence attendue s'il y avait indépendance. Le score mesurant la pertinence d'un terme pour la discrimination entre les m catégories peut être calculé ainsi : $\chi^2_{max}(t_j) = \max_{i=1}^{m}\{\chi^2(t_j, c_i)\}$. La moyenne sur les m catégories (pondérée par le poids de chaque catégorie) peut être une alternative pour évaluer ce score.
- L'information mutuelle ("mutual information" en anglais). Dans la théorie des probabilités et la théorie de l'information, l'information mutuelle est une quantité mesurant la dépendance statistique de deux variables aléatoires. Elle est utilisée dans le domaine de la modélisation statistique du langage [Church and Hanks 1990]. Le critère $I(t_j, c_i)$ mesure donc la dépendance entre les occurrences du terme t_j et la catégorie c_i. La pertinence globale du terme pour discriminer les catégories est évaluée ainsi : $I_{max}(t_j) = \max_{i=1}^{m}\{I(t_j, c_i)\}$, ou de manière alternative par la moyenne pondérée.
- Le gain d'information ("information gain" en anglais). Fréquemment employé dans le

domaine de l'apprentissage automatique [Quinlan 1986; Mitchell 1996], notamment pour la construction des arbres de décision, le gain d'information $G(t_j)$ mesure la contribution du terme t_j à la prédiction des catégories sachant la présence ou absence du terme dans un document.

Nous citons également deux autres indicateurs, basés sur des lois, provenant de l'analyse statistique de données textuelles :

- La spécificité lexicale. Introduites par Lafon [1980], les spécificités positives sont très utilisées en analyse textuelle pour décrire une partie d'un corpus à travers les formes qui y sont significativement sur-employées par rapport aux autres parties. La formule de Lafon basée sur le modèle hypergéométrique est utilisée pour calculer le score de spécificité $S^+(t_j, c_i)$ du terme t_j au sein de la catégorie c_i.
- Le Z-score. Le score $Z(t_j, c_i)$ est une statistique indiquant le degré d'appartenance d'un terme t_j au vocabulaire spécifique d'une catégorie c_i. Son calcul est basé sur l'hypothèse que le nombre d'occurrences d'un mot au sein d'une catégorie suit une loi binomiale.

Afin de mesurer la pertinence du terme t_j pour la discrimination entre les m catégories, les scores suivant peuvent être calculés, respectivement pour la spécificité lexicale et le Z-score : $S^+_{max}(t_j) = \max_{i=1}^{m}\{S^+(t_j, c_i)\}$ et $Z_{max}(t_j) = \max_{i=1}^{m}\{Z(t_j, c_i)\}$. Ces scores peuvent être évalués de manière alternative par la moyenne sur les m catégories (moyenne pondérée par le poids de chaque catégorie).

Cas non supervisé. Dans le cas non supervisé, il n'y a pas de connaissance a priori sur les documents et seul leur contenu textuel est utilisé pour, par exemple, obtenir une classification des documents, ou les ordonner en fonction de leur pertinence vis-à-vis d'une requête utilisateur. Les méthodes les plus souvent utilisées pour la sélection des termes sont les suivantes :

- Le nombre de documents ("document frequency" en anglais). Les termes apparaissant dans un nombre (ou une proportion) de documents dépassant un seuil fixé sont retirés car présents dans la quasi-totalité des documents, ils n'apportent pas d'information utile pour décrire ces derniers.

– La force du terme ("term strength" en anglais). Utilisé pour la réduction du vocabulaire en recherche d'information [Wilbur and Sirotkin 1992], ce critère évalue l'importance d'un terme en se basant sur sa propension à apparaitre dans des documents semblables. Dans un premier temps, des paires de documents sont formées (documents dont la similarité cosinusoïdale est au-dessus d'un certain seuil). Puis la force du terme est calculée comme la probabilité qu'il apparaisse dans le deuxième document d'une paire sachant qu'il apparaît dans le premier document.

4.2.1.4 Représentation vectorielle des documents

Pour rendre les textes exploitables, chaque document est transformé en un vecteur de mots [Salton et al. 1975]. Ce type de représentation est également appelée "sacs de mots". Les documents ainsi vectorisés forment une matrice dont :

– les lignes sont les documents,
– les colonnes sont les termes,
– les cellules sont les fréquences des termes au sein des documents.

Cette dernière est appelée matrice documents×termes. Une matrice termes×documents peut également être construite de la même manière selon la préférence du praticien. Les dimensions de la matrice sont alors déterminées par le nombre de documents du corpus et le nombre total de termes distincts. Les termes ayant le moins d'occurrences sont responsables du caractère "sparse" de la matrice documents×termes (la matrice obtenue est très creuse), qui peut être réduit grâce à la suppression des termes n'atteignant pas un seuil minimum d'occurrences ou n'apparaissant pas dans un nombre minimum de documents. Toutefois, la représentation vectorielle des textes est souvent critiquée car elle ne prend pas en compte les relations sémantiques entre les termes ou la structure des phrases. Les termes ainsi détachés de leur contexte peuvent être considérés comme un même descripteur d'un texte à l'autre alors qu'ils ont des sens différents.

Une première méthode pour pallier ce problème consiste à introduire dans le vecteur d'un document, en plus des termes "simples", les suites de deux mots consécutifs. Ceci permet de faire apparaître des groupes de mots ayant un sens différent de celui des deux mots pris indépendamment. Nous illustrons cela avec un exemple du marché du recrutement :

"responsable comptabilité générale" est décrit par les termes "générale", "comptabilité", "responsable", "responsable comptabilité" et "comptabilité générale".

La pondération des termes et l'analyse sémantique constituent une deuxième approche permettant d'améliorer la prise en compte du sens des mots et de leur importance au sein d'un texte. Cette approche est présentée dans la section qui suit (4.2.1.5).

4.2.1.5 Extraction des descripteurs

Afin d'améliorer la qualité de représentation des textes du point de vue sémantique, de nouveaux descripteurs sont extraits à partir de la représentation vectorielle des documents. Pour cela, la première étape consiste à appliquer des fonctions de pondération aux termes afin de mettre en avant les termes les plus importants et de "masquer" les termes peu pertinents. Pour améliorer encore la prise en compte de la sémantique des textes, les auteurs ont fréquemment recours à l'analyse sémantique latente (LSA pour "latent semantic analysis" en anglais) suite à une pondération des termes. Appliquée en recherche d'information, cette technique de réduction de la dimension est appelée LSI ("latent semantic indexing").

Pondération des termes. Il existe un grand nombre de fonctions de pondération des termes, et celles-ci peuvent être séparées en deux catégories :

- les fonctions de pondération locale, qui décrivent la fréquence relative d'un terme au sein d'un document ;
- les fonctions de pondération globale, qui décrivent la fréquence d'un terme relativement à l'ensemble du corpus.

Soit f_{ij} la fréquence d'apparition du terme j dans le document i, et soient respectivement $l_{ij}(f_{ij})$ et $g_j(f_{ij})$ les pondérations locale et globale associées.

Les pondérations locales les plus couramment utilisées sont :

- la fréquence du terme dans le document ("term frequency" ou TF), $l_{ij}(f_{ij}) = f_{ij}$, pondération la plus basique qui découle de la représentation vectorielle des textes ;
- la pondération binaire, $l_{ij}(f_{ij}) = \begin{cases} 1 \text{ si le terme } j \text{ apparaît dans le document } i \\ 0 \text{ sinon} \end{cases}$;
- la pondération logarithme, $l_{ij}(f_{ij}) = log(f_{ij} + 1)$, qui atténue les plus hautes fréquences.

Parmi les fonctions de pondération globale, nous retenons :

- la normalisation, $g_j(f_{ij}) = \frac{1}{\sqrt{\sum_i f_{ij}^2}}$;
- la pondération IDF ("inverse document frequency"), $g_j(f_{ij}) = 1 + log(\frac{n}{n_j})$, où n est le nombre total de documents et n_j le nombre de documents contenant le terme j ;
- l'entropie, $g_j(f_{ij}) = 1 + \sum_i \frac{\frac{f_{ij}}{\sum_i f_{ij}} log(\frac{f_{ij}}{\sum_i f_{ij}})}{log(n)}$, où n est le nombre total de documents.

Généralement, une pondération locale est combinée avec une pondération globale, ce qui donne naissance à de nouvelles fonctions de pondération. Parmi les plus courantes :

- la pondération TF-IDF ["term frequency – inverse document frequency" ; Salton 1989], $l_{ij}(f_{ij}).g_j(f_{ij}) = f_{ij}(1 + log\frac{n}{n_j})$;
- la pondération *log-entropie*, $l_{ij}(f_{ij}).g_j(f_{ij}) = log(f_{ij} + 1)\left[1 + \sum_i \frac{\frac{f_{ij}}{\sum_i f_{ij}} log(\frac{f_{ij}}{\sum_i f_{ij}})}{log(n)}\right]$.

Ces fonctions de pondération ont pour but de diminuer le poids des termes apparaissant dans un grand nombre de documents et qui sont inutiles pour la discrimination entre les documents. Il n'existe pas de pondération "optimale" car la qualité des résultats obtenus dépend fortement du type d'application et du type de textes manipulés.

Projection dans l'espace sémantique. Brevetée [4] en 1988, l'analyse sémantique latente [Deerwester et al. 1990; Martin and Berry 2007], ou LSA, est une technique de réduction de la dimension reposant sur la décomposition en valeurs singulières ("singular value decomposition" ou SVD) de la matrice documents×termes. L'hypothèse justifiant la mise en œuvre de cette méthode est qu'il existe une structure sémantique latente, sous-jacente aux données d'usage des mots, partiellement cachée ou bruitée par l'alea lié au choix des mots. Cet alea fait référence aux problèmes de synonymie et polysémie. D'une part, différentes formes peuvent avoir un même sens (synonymie) mais seront considérées comme des descripteurs différents ; d'autre part, une même forme peut avoir des sens différents selon le contexte (polysémie) mais sera représentée par un seul descripteur. La LSA, qui établit des relations sémantiques entre les termes, permet de pallier ces problèmes. Dans la pratique, entre 40 et 400 dimensions sont conservées, mais cela dépend fortement de l'application. Le nombre de descripteurs est donc réduit de quelques milliers à quelques centaines, voire moins. De plus, la réduction de la dimension élimine le caractère "sparse"

4. US Patent - 4 839 853

de la matrice documents×termes, tout en retirant la partie bruitée des données.

L'analyse des correspondances, technique largement utilisée en analyse statistique des données, est très proche de l'analyse sémantique lorsqu'elle est appliquée dans le contexte des données textuelles [Lebart et al. 1998]. Nous présentons ici une comparaison des deux méthodes à travers leurs différentes étapes (tableau 4.1).

Analyse sémantique latente	**Analyse des correspondances**
Étape 1 : Construction de la matrice documents×termes	
$T = [f_{ij}]_{i,j}$	$T = [f_{ij}]_{i,j}$
Étape 2 : Pondération	
$T_W = [l_{ij}(f_{ij}).g_j(f_{ij})]_{i,j}$	$T_W = \left[\frac{f_{ij}}{\sqrt{f_{i.}.f_{.j}}}\right]_{i,j}$
Étape 3 : Décomposition en valeurs singulières	
$T_W = U\Sigma V'$	$T_W = U\Sigma V'$
	Étape 3b : Normalisation des vecteurs propres $\tilde{U} = diag\left(\sqrt{\frac{f_{..}}{f_{i.}}}\right)U$
Étape 4 : Calcul des coordonnées des documents	
$C = U_k\Sigma_k$	$C = \tilde{U}_k\Sigma_k$

TABLE 4.1 – Analyse sémantique latente et analyse des correspondances : comparaison

La méthodologie générique de la LSA implique l'application de fonctions de pondération locale (l) et globale (g). Étant donné les étapes de l'analyse des correspondances énumérées ci-dessus, nous pouvons identifier les pondérations locale et globale associées : $l_{ij}(f_{ij}) = \frac{f_{ij}}{\sqrt{f_{i.}}}$ et $g_j(f_{ij}) = \frac{1}{\sqrt{f.j}}$. Les deux méthodes sont similaires, à l'exception d'une étape supplémentaire de normalisation dans l'analyse des correspondances.

Il existe dans la littérature d'autres méthodes non supervisées permettant de construire des modèles sémantiques, notamment l'analyse sémantique latente probabiliste ["probabilistic latent semantic analysis" ou pLSA ; Hofmann 1999] et la LDA ["latent Dirichlet allocation" ; Blei et al. 2003]. Toutefois, il n'est pas ressorti de la littérature que ces méthodes permettent d'obtenir de meilleures performances que la LSA d'une manière générale [e.g. Gehler et al. 2006].

4.2.2 Aperçu des travaux effectués sur les offres d'emploi

4.2.2.1 Travaux visant à rapprocher une offre d'emploi des candidatures les plus adéquates

Les job boards sont capables de mettre en relation de très grands volumes de recruteurs et de chercheurs d'emploi, mais ne permettent pas de garantir un bon niveau de qualification des candidatures qui en découlent. Bartram [2000] indique que repositionner le recrutement en ligne comme un processus de rapprochement entre les compétences des candidats et les exigences du poste à pourvoir permettrait de produire des listes de candidats de haute qualité, et éviterait ainsi aux recruteurs de devoir *"embrasser des grenouilles pour trouver des princes"*[5]. Depuis, de nombreux travaux ont été publiés dans la littérature sur cette thématique. Nous n'en faisons pas la liste exhaustive mais présentons au lecteur les travaux considérés comme pertinents au regard de la problématique du traitement textuel des offres d'emploi.

Dans le contexte du rapprochement offres-CV, visant à détecter de manière automatique les CV les plus qualifiés pour une offre d'emploi (et inversement), certains auteurs adoptent la représentation vectorielle des textes. C'est le cas de Kessler et al. [2009], qui comparent différentes méthodes pour représenter les documents. Après étiquetage grammatical, seuls les noms, verbes et adjectifs sont conservés, et différentes combinaisons et pondérations sont expérimentées. Parallèlement, les n-grammes de caractères [Damashek 1995] sont testés sans appliquer de pré-traitements aux textes. Les auteurs testent également deux types d'enrichissement sémantique du descriptif du poste : en utilisant le vocabulaire issu de la base ROME[6] de l'ANPE, puis par une technique de *relevance feedback*[7] (prendre en compte les choix du recruteur lors d'une première évaluation de quelques CV). Enfin, les résultats pour les pondérations TF et TF-IDF sont également comparés. L'ensemble des candidatures est ensuite ordonné par rapport aux offres d'emploi étudiées grâce à une mesure de similarité. Les meilleurs résultats sont obtenus avec les pondérations TF et TF-IDF, ceux-ci étant améliorés par le *relevance feedback*.

5. Texte original : "kissing frogs to find princes".
6. Répertoire Opérationnel des Métiers et des Emplois, (`http://www2.pole-emploi.fr/espacecandidat/romeligne/RliIndex.do`).
7. "Retour de pertinence".

D'autres auteurs choisissent d'avoir recours aux ontologies [e.g. Bizer et al. 2005; Yahiaoui et al. 2006; Trichet et al. 2006; Radevski et al. 2006]. L'inconvénient majeur de ces approches est qu'elles nécessitent une annotation manuelle importante de la part du recruteur et du candidat pour décrire respectivement les offres d'emploi et les CV (compétences, fonctions du poste, secteur), en ayant recours à des vocabulaires spécifiques.

4.2.2.2 Le projet SIRE

Mené dans le cadre du Pôle de Compétitivité Cap Digital [8], le projet SIRE [9] a pour objectif de mettre au point un dispositif permettant de recueillir et d'analyser automatiquement les offres d'emploi disponibles sur Internet afin de produire un observatoire du marché du travail accessible à tous. Le projet vise à construire un modèle de compétences basé sur des ontologies, et à le mettre en œuvre dans des applications pratiques comme "l'observatoire des métiers".

Un prototype se basant sur des ontologies RH et des méthodes issues de la fouille de textes (e.g. clustering de documents) a été élaboré afin d'annoter sémantiquement et de classifier les offres d'emploi [Loth et al. 2010]. La structuration des offres ainsi obtenue sera utilisée pour servir plusieurs objectifs : dédoublonnage, statistiques du marché du travail, rapprochement des offres d'emploi et des candidatures, etc.

4.3 Classification des offres d'emploi en fonction du poste proposé

Comme évoqué précédemment, nous souhaitons obtenir une nomenclature des offres indépendante du site d'emploi. Nous présentons dans un premier temps les motivations de cette classification, puis introduisons les algorithmes usuels de classification. Enfin, nous présentons l'approche que nous proposons pour classifier efficacement les offres d'emploi.

8. Créé en 2006, Cap Digital, outil de développement de la filière des contenus numériques en Île-de-France, a déjà soutenu plus de 150 projets de recherche collaborative.
9. Sémantique, Internet, Recrutement et Emploi (`http://www.sire-project.eu/index.php`).

4.3.1 Objectifs poursuivis

Cette nomenclature uniformisée des offres d'emploi nous permettra d'une manière générale d'améliorer la structuration de l'information relative aux annonces. Cette structuration nous permettra ensuite :

- de fournir des statistiques de comparaison entre les sites, de suivre les performances des annonces selon le type de poste proposé, d'établir des comparaisons entre les recruteurs (toutes ces informations pouvant être intégrées à des rapports statistiques destinés aux recruteurs) ;
- de prendre en compte l'information traduite par la nomenclature au sein du modèle de prédiction ;
- d'établir les règles expertes permettant la recommandation des sites gratuits adaptés aux caractéristiques du poste à pourvoir (en particulier les sites spécialisés et les écoles) ;
- d'exploiter les associations avec les autres nomenclatures de sites afin de permettre le remplissage automatique des champs spécifiques [10] à la deuxième étape de la diffusion d'une annonce via l'outil Multiposting.fr. Ce dernier point sort du cadre des problématiques abordées par nos travaux de thèse, mais est à prendre en considération car représente un apport significatif au processus de diffusion d'une annonce.

4.3.2 Les tables de correspondance entre nomenclatures

Pour obtenir cette nomenclature unique, une première méthode élémentaire consisterait à établir la nomenclature de notre choix puis à rechercher les correspondances pour chaque valeur des nomenclatures de l'ensemble des sites d'emploi. Le fichier des correspondances entre les valeurs de deux champs différents est appelé "mapping". Cette solution présente deux inconvénients majeurs :

- elle est très chronophage et extrêmement coûteuse en termes de ressources humaines (car les correspondances doivent être effectuées manuellement) ;
- les catégories choisies par les différents sites pour la répartition des annonces sont

10. Lors de la diffusion d'une annonce via Multiposting.fr, l'étape de choix des sites est succédée par une étape dédiée au remplissage des champs spécifiques à chaque site d'emploi choisi. Cette étape indispensable peut se révéler très chronophage lorsqu'un grand nombre de sites ont été choisis.

plus ou moins nombreuses, et par suite de précisions différentes. Or, s'il est possible d'associer des catégories "fines" (d'un point de vue sémantique) à une catégorie plus large, il n'est pas possible d'affecter les annonces d'une catégorie large à des catégories plus précises. De plus, nous recherchons une catégorisation fine des offres d'emploi, et certains sites ne proposent qu'une nomenclature "macro" de la fonction du poste. L'exemple qui suit illustre cette problématique.

Exemple 1 *Nous considérons les nomenclatures "fonction" de deux sites d'emploi généralistes largement utilisés par les recruteurs. Afin de simplifier l'exemple, nous considérons les fonctions d'un point de vue macro même si des sous-fonctions plus précises existent dans les nomenclatures. Nous verrons ainsi que même des fonctions larges d'un point de vue sémantique peuvent ne pas trouver de correspondance entre elles. Le tableau 4.2 présente la confrontation des deux nomenclatures en disposant les libellés de sorte à mettre en évidence les meilleures correspondances possibles. Les libellés signalés en gras indiquent l'existence de sous-fonctions à l'intérieur de la fonction. Nous pouvons déjà remarquer que certaines fonctions ne sont pas découpées sur le site 1 et les annonces correspondantes ne peuvent donc pas être réparties dans des catégories plus précises. De plus, certaines fonctions correspondent à des regroupements de fonctions sur le site 2 (par exemple, "Ressources Humaines – Personnel – Formation", site 1) et ne peuvent pas trouver de correspondance exacte.*

Une alternative consistant à rechercher automatiquement les correspondances entre nomenclatures en se basant sur l'historique des multidiffusions (diffusions simultanées sur plusieurs sites d'emploi) permet la réduction du temps alloué au travail manuel mais n'identifie pas les correspondances de manière exhaustive et présente les problèmes liés au "mapping" évoqués ci-dessus.

Cette solution n'étant pas satisfaisante, nous nous tournons vers la classification, technique faisant appel à l'apprentissage automatique permettant de répondre à la tâche de catégorisation des documents.

Nomenclature du site 1	**Nomenclature du site 2**
	Architecture, Création & Spectacle
Administration - Services généraux	**Services administratifs**
Audit	
	BTP & second-œuvre
Commercial - Vente	**Commercial / Vente**
Communication - Création	
Conseil	
Direction générale - Direction centre de profits	**Stratégie & Management**
	Edition & Écriture
Études - Recherche	**Ingénierie**
	Recherche & Analyses
Export	
Gestion-Comptabilité-Finance	**Comptabilité & Finance**
	Gestion de projet / programme
	Hôtellerie, Restauration & Tourisme
Internet - e-Commerce	
Juridique Fiscal	**Juridique**
Logistique - Achat - Stock - Transport	**Logistique, Approvisionnement & Transport**
Marketing	**Marketing**
Production - Maintenance - Qualité - Sécurité - Environnement	**Installation, Maintenance & Réparation**
	Production & Opérations
	Qualité / Inspection
Ressources Humaines - Personnel - Formation	**Formation / Éducation**
	Ressources Humaines
Santé	**Santé**
Systèmes d'informations - Télécom	**Informatique & Technologies**
	Sécurité
	Services clientèle & aux particuliers
	Autres

TABLE 4.2 – Exemple de confrontation de deux nomenclatures "fonction" issues de deux sites d'emploi généralistes

4.3.3 Algorithmes de classification

Il existe dans la littérature un grand nombre d'algorithmes de classification, pouvant être répartis en deux grandes catégories : les algorithmes supervisés et les algorithmes non supervisés. Dans le cas d'algorithmes non supervisés, il s'agit d'une problématique de clustering : l'objectif est d'obtenir des classes de documents homogènes en se basant sur leur contenu textuel mais sans connaissance préalable autre. Dans le cas d'algorithmes supervisés, il s'agit d'un problème de catégorisation (ou d'étiquetage) de documents. Ces derniers algorithmes utilisent l'existence d'une nomenclature a priori afin d'apprendre sur les documents déjà étiquetés et de catégoriser les nouveaux documents selon les valeurs de cette nomenclature. En d'autres termes, il s'agit d'une problématique de prédiction.

Dans le cas non supervisé, les algorithmes les plus répandus sont *k-means* [MacQueen 1967; Hartigan and Wong 1979], les réseaux de Kohonen [*Self-Organizing Map* ; Kohonen 2001], et la classification ascendante hiérarchique [e.g. Ward 1963; Jardine and van Rijsbergen 1971; Candillier 2006]. Cette dernière se base sur le regroupement pas à pas des deux classes les plus proches (les classes initiales sont les observations) au regard d'un critère de distance à choisir (*single linkage*, *complete linkage*, *average linkage*, méthode des centroïdes, ou encore critère de Ward). La CAH se révèle limitée lorsque le nombre de documents est très important. Une alternative consiste à effectuer un premier regroupement des observations grâce à un algorithme de type *k-means* (spécifier par exemple un nombre de classes égal à un dixième du nombre initial d'observations) et réaliser la CAH en démarrant avec les barycentres des classes obtenues.

Dans le cas supervisé, les algorithmes les plus utilisés sont les machines à vecteurs de support ("support vector machines" ou SVM), les arbres de décisions, les réseaux de neurones [perceptron multicouche ; Rumelhart and Williams 1986; Stricker 2000], et l'analyse discriminante [e.g. Benzécri 1977; Lebart et al. 1998]. Parmi les nombreux algorithmes existants permettant la construction d'arbres de décision, nous retenons les algorithmes CART [Breiman et al. 1984] et C4.5 [Quinlan 1993]. Les SVM [Vapnik 1995] reposent sur la projection des données dans un espace de grande dimension par une transformation basée sur un noyau, dont les plus répandus sont les noyaux polynomiaux et gaussiens. Il est

également possible d'utiliser un noyau linéaire, ramenant au cas d'un classifieur linéaire, sans changement d'espace.

4.3.4 Évaluation d'un système de classification : critères de performance

Pour évaluer la qualité d'un système, les données sont scindées en deux sous-ensembles : échantillon d'apprentissage et échantillon de test (cf. section 4.3.5.3). La qualité du système est alors évaluée sur l'échantillon de test contenant des documents n'ayant pas servi à l'apprentissage. Dans le cadre supervisé, les documents de l'échantillon de test sont étiquetés avec les labels des catégories connues a priori. Il est alors possible de comparer les catégories prédites aux catégories réelles. Dans le cadre non supervisé, les classes obtenues ne sont pas étiquetées, il convient alors d'utiliser des critères spécifiques pour comparer les classes prédites aux classes réelles.

4.3.4.1 Cas supervisé

La précision et le rappel sont deux critères largement utilisés pour évaluer la qualité d'un algorithme de classification. Dans le contexte de la catégorisation de documents, la précision p_i permet d'évaluer la qualité de prédiction de la classe i au regard des documents qui lui sont attribués : $p_i = \frac{\text{nombre de documents correctement attribués à la classe } i}{\text{nombre de documents attribués à la classe } i}$; tandis que le rappel r_i permet d'évaluer la capacité de l'algorithme à retrouver les documents appartenant à la classe i : $r_i = \frac{\text{nombre de documents correctement attribués à la classe } i}{\text{nombre de documents appartenant à la classe } i}$.

Dans le cadre multi-classes, la précision P et le rappel R d'un point de vue global sont calculés comme des macro-moyennes sur l'ensemble des m catégories :

$$P = \frac{\sum_{i=1}^{m} p_i}{m} \text{ et } R = \frac{\sum_{i=1}^{m} r_i}{m}$$

La mesure F_β [van Rijsbergen 1979] est un indicateur de synthèse qui permet de prendre en compte simultanément la précision et le rappel en accentuant l'importance de l'un ou de l'autre grâce au paramètre β, réel positif :

$$F_\beta = \frac{(1+\beta^2) \times P \times R}{\beta^2 \times P + R}$$

Dans le cas usuel, $\beta = 1$. Choisir $\beta < 1$ permet d'accentuer l'importance de la précision, tandis que choisir $\beta > 1$ accentue celle du rappel.

Nous considérons également l'exactitude E, indicateur rapportant la qualité globale de l'algorithme. À la différence de la précision et du rappel, l'exactitude est influencée par les effectifs des différentes classes (plus une classe est grande, plus la qualité de prédiction de cette classe a de poids sur l'indicateur) :

$$E = \frac{\text{nombre de documents correctement attribués}}{\text{nombre total de documents}}$$

4.3.4.2 Cas non supervisé

Nous souhaitons ici comparer deux partitions P_1 et P_2 composées chacune de n objets (nombre total de documents) répartis en m classes distinctes. Soit $N = [n_{ij}]_{\substack{i=1,..,m \\ j=1,..,m}}$ la table de contingence associée. Une des deux partitions n'étant pas étiquetée, nous devons donc utiliser un critère indépendant des labels des classes pour la comparaison.

L'indice brut de Rand est le pourcentage global de paires en accord entre les deux partitions. Soient a le nombre de paires dans une même classe de P_1 et dans une même classe de P_2, et b le nombre de paires séparées dans P_1 et séparées dans P_2. L'indice de Rand est alors défini par : $R = \frac{a+b}{C_n^2}$. Pour le calcul, nous utilisons la variante de Marcotorchino et El Ayoubi [1991] où toutes les paires (y compris celles identiques) sont considérées :

$$R = \frac{2\sum_i \sum_j n_{ij}^2 - \sum_i n_{i.}^2 - \sum_j n_{.j}^2 + n^2}{n^2}, \qquad 0 \leq R \leq 1$$

La mesure F_β et l'exactitude sont initialement des critères d'évaluation dépendants des labels des classes. Ils peuvent toutefois être adaptés au cas de partitions non étiquetées. Une solution peut alors consister à affecter à chaque classe obtenue l'étiquette de la nomenclature initiale la plus fréquente au sein de la classe.

4.3.5 Approche proposée

Des premiers résultats peu satisfaisants avec une méthode de classification non supervisée (voir section 4.3.6.1) nous ont conduit à opter pour une méthode supervisée, permettant d'obtenir une catégorisation des offres d'emploi de bonne qualité grâce à l'exploitation des connaissances expertes disponibles. En effet, nous utilisons l'étiquetage manuel qui a été effectué par les recruteurs sur un ensemble d'annonces ayant été multidiffusées via Multiposting.fr.

4.3.5.1 Méthodologie de mise au point de la nomenclature métier

La première étape consiste à préparer l'échantillon des offres étiquetées selon la nomenclature que nous souhaitons utiliser par la suite. Afin d'obtenir des offres déjà étiquetées par les recruteurs, nous utilisons l'historique des annonces postées sur un site généraliste choisi pour la qualité de sa nomenclature "fonction". Il s'agit d'une nomenclature à deux niveaux, fonction et sous-fonction, contenant respectivement 24 et 278 catégories. Le premier niveau permet une répartition des postes d'un point de vue macro ("Marketing", "Commercial/Vente", "Services administratifs", "Comptabilité et Finance", etc.). Le deuxième niveau permet, au sein de chaque fonction, une dénomination plus précise du type de poste à pourvoir d'un point de vue métier (par exemple : "Assistanat de direction", "Accueil/Réception", "Services Généraux", etc., pour la fonction "Services administratifs" ; ou "Analyse financière", "Actuariat", "Fiscalité", etc., pour la fonction "Comptabilité et Finance"). Une étude de cette nomenclature par des experts métier a conclu à un bon niveau de précision et une bonne couverture des postes sur le marché du travail. Les offres d'emploi étiquetées avec la nomenclature de ce site peuvent se voir affecter un à trois labels de sous-fonctions différentes appartenant à une même fonction. Le grand nombre de sous-fonctions induisent des catégories de sous-fonctions avec de très faibles effectifs. De plus, notre analyse de la nomenclature a mis en évidence des subdivisions non nécessaires ainsi que des recouvrements possibles entre les sous-fonctions. Nous souhaitons donc en diminuer le nombre grâce à des regroupements pertinents.

Pour détecter les regroupements pertinents, nous allons exploiter les co-occurrences de labels sur une même offre d'emploi, et nous nous ramenons à un problème de classification de variables qualitatives. Nous avons recours au coefficient de Tchuprow pour le calcul des distances entre variables qualitatives (ici variables binaires), puis effectuons les regroupements à l'aide de la classification ascendante hierarchique. L'observation du dendrogramme et la validation des regroupements par un expert métier nous conduisent à conserver 107 sous-fonctions, pour un total de 23 fonctions. Les fonctions "Ingénierie" et "Recherche & Analyses" ont été regroupées, et six fonctions ne sont pas découpées ("Édition & Écriture", "Gestion de projet / programme", "Formation / Éducation", "Santé", "Sécurité" et "Autres"). La nomenclature détaillée est présentée en annexe dans le tableau 4.4.

Après étiquetage des offres à partir de la nouvelle nomenclature, les offres ayant plusieurs labels deviennent rares (environ 10% du volume des offres étiquetées), nous décidons donc de les écarter. De plus, pour des raisons de praticité dans les développements techniques, notre objectif est d'affecter un unique couple (fonction, sous-fonction) à chaque offre.

4.3.5.2 Algorithme pour la classification des annonces

Nous disposons maintenant d'un échantillon d'offres étiquetées par un couple de libellés (fonction, sous-fonction). Chaque offre est décrite par un texte (titre, descriptif des missions). Des pré-traitements usuels ainsi qu'une sélection des descripteurs sont appliqués. Cela nous permet de réduire les temps d'apprentissage tout en conservant l'information pertinente présente dans les documents. Nous appliquons ensuite une méthode d'analyse sémantique au tableau de fréquences croisant documents et termes afin d'en réduire la dimension et d'éliminer la partie bruitée des données (se référer à la section 4.3.6 pour les expérimentations). Les coordonnées dans l'espace sémantique ainsi obtenu sont utilisées pour représenter les documents. Afin d'obtenir l'étiquetage des offres par rapport à la fonction et la sous-fonction, nous établissons dans un premier temps un modèle de prédiction de la fonction à l'aide d'un algorithme supervisé. Dans un deuxième temps, un modèle est construit au sein de chaque fonction, toujours à l'aide du même algorithme, afin de prédire la sous-fonction associée à l'offre.

L'algorithme SVM s'est montré efficace à de nombreuses reprises pour des tâches de catégorisation sur données textuelles. Joachims [1998] présente les principales caractéristiques des données textuelles et montre en quoi les SVM se révèlent particulièrement adaptés à ce type de données, notamment :

- ils ont la capacité de gérer un très grand nombre de dimensions avec un faible risque de sur-apprentissage ;
- ils sont adaptés à des matrices de données très creuses comme peut l'être la matrice documents×termes.

Par ailleurs, les expérimentations menées par Joachims [1998] sur deux jeux de données de la littérature montrent la supériorité des SVM relativement à d'autres méthodes de

classification (un classifieur bayésien naïf, l'algorithme de Rocchio, l'algorithme C4.5 et un classifieur de type k plus proches voisins). Dans le cadre du traitement textuel des offres d'emploi, les SVM ont été utilisés avec succès par Kessler et al. [2007] pour l'étiquetage des différentes parties d'une annonce d'emploi. C'est donc l'algorithme que nous choisissons pour réaliser notre classification des offres d'emploi. Des premières expérimentations sur le corpus des annonces ont montré que les noyaux les plus complexes n'apportaient pas un gain de performance dans la classification par rapport au noyau linéaire. Étant donné la grande dimension du problème et afin de minimiser les risques de sur-apprentissage, les SVM linéaires sont utilisés pour la classification des annonces. L'optimisation du paramètre de coût est effectuée par validation croisée (*10-fold*[11]). Enfin, l'observation de la matrice des confusions nous permet de mettre en évidence des métiers dont les missions sont proches sémantiquement, à l'origine des erreurs de catégorisation. La procédure dans son ensemble est illustrée par la figure 4.3.

4.3.5.3 Évaluation de la performance du système

Échantillon de test. Pour mesurer la qualité de l'algorithme de catégorisation, le corpus est divisé en deux parties : 65% du corpus sont dédiés à l'apprentissage, tandis que les 35% restants constituent l'échantillon de test. Les données de cet échantillon sont complètement extérieures à la définition de l'ensemble des descripteurs et à l'apprentissage du modèle. Les échantillons de test et d'apprentissage sont construits de sorte à préserver la répartition des sous-catégories (sous-fonctions) observée sur l'ensemble du corpus, ceci dans le but de se prémunir d'éventuels biais et d'assurer un apprentissage sur l'ensemble des sous-catégories. Au sein de chaque sous-catégorie, la répartition entre apprentissage et test est faite par un tirage aléatoire. Le processus d'évaluation de l'erreur sur l'échantillon de test est illustré par la figure 4.4.

Critère de performance du système de catégorisation. Dans notre application, les effectifs des classes sont fortement déséquilibrés car la clientèle de la société Multiposting

11. La méthode de validation croisée *10-fold* est la suivante : découper l'échantillon d'apprentissage en 10 parties d'effectifs égaux ; estimer le modèle sur 9 des parties et évaluer la prédiction sur la partie restante ; reproduire l'opération jusqu'à obtenir une prédiction pour l'ensemble de l'échantillon. L'erreur du système est ensuite évaluée sur les prédictions ainsi obtenues.

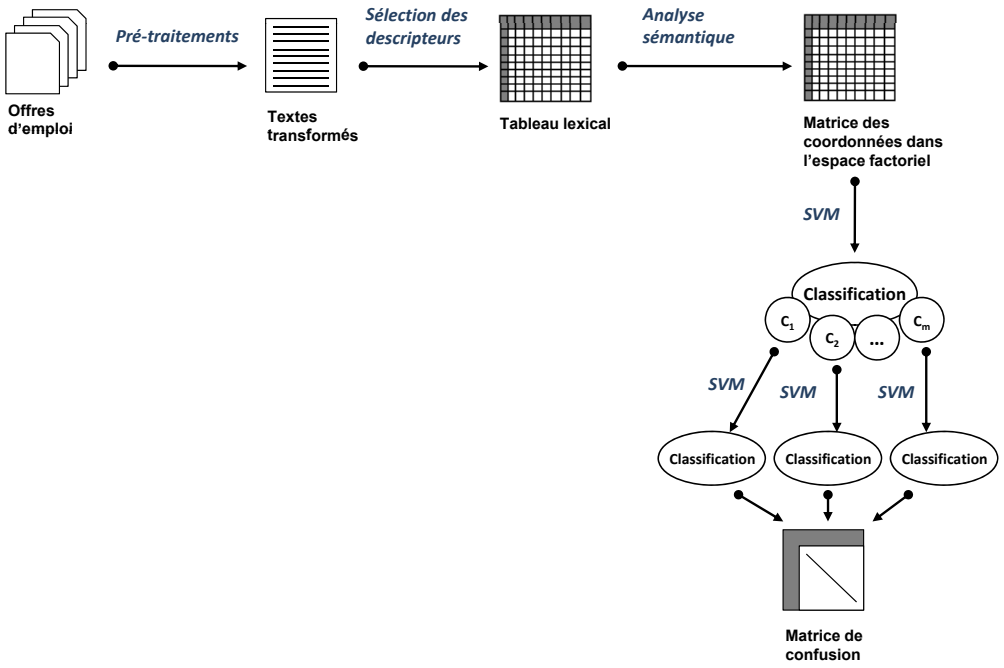

FIGURE 4.3 – Vue d'ensemble du système de catégorisation

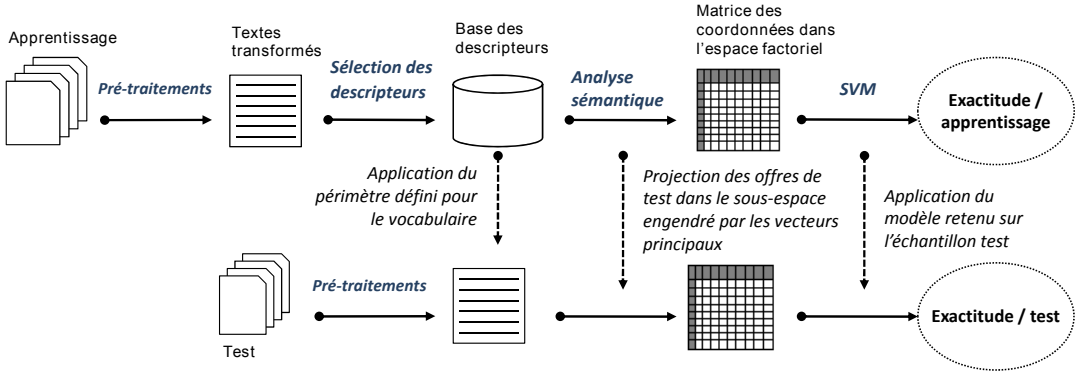

FIGURE 4.4 – Processus d'évaluation de l'erreur dans le système de catégorisation

recrute en grande partie dans une minorité des domaines représentés. En effet, dans notre échantillon d'apprentissage, 43% des postes concernent les fonctions "Comptabilité & Finance" (17%), "Informatique & Technologies" (14%) et "Commercial / Vente" (12%) pour un total de 24 fonctions initiales. Malgré ce déséquilibre dans les effectifs des classes, nous choisirons le meilleur système en nous basant sur l'exactitude car nous souhaitons privilégier les catégories majoritairement représentées, afin d'obtenir un système performant pour les fonctions les plus recherchées par les utilisateurs de Multiposting.fr. De plus, certaines classes ayant des effectifs très faibles, la mesure F_β en découlant se révèle être peu stable. Toutefois, les résultats obtenus seront également illustrés d'un point de vue global avec la mesure F_β, et au niveau des classes à l'aide de la précision et du rappel.

4.3.6 Expérimentations

Nous étudions une base de 9 305 offres d'emploi, se répartissant entre échantillon d'apprentissage et échantillon de test à hauteur de 65% et 35% respectivement. Chaque offre de la base est étiquetée selon une nomenclature à deux niveaux obtenue à partir de la méthode décrite dans la section 4.3.5.1. La répartition entre les 23 catégories de fonctions (premier niveau de la nomenclature) est visible dans le tableau 4.3.

Fonction	**Effectif**	**%**	**Fonction**	**Effectif**	**%**
Architecture, Création	64	0.7	Juridique	143	1.5
BTP & second-œuvre	317	3.4	Logistique & Transport	496	5.3
Commercial / Vente	1154	12.4	Stratégie & Management	180	1.9
Comptabilité & Finance	1565	16.8	Marketing	302	3.3
Édition & Écriture	55	0.6	Production & Opérations	403	4.3
Formation / Éducation	106	1.1	Ingénierie & Recherche	713	7.7
Hôtellerie, Restauration	112	1.2	Ressources Humaines	547	5.9
Informatique & Technologies	1286	13.8	Santé	120	1.3
Gestion de projet	173	1.9	Services administratifs	506	5.4
Qualité / Inspection	190	2.0	Services clientèle	318	3.4
Installation & Maintenance	383	4.1	Sécurité	30	0.3
Autres	142	1.5	**Total**	**9305**	**100**

TABLE 4.3 – Répartition des annonces entre les catégories de fonctions

Les pré-traitements suivants sont effectués afin de rendre exploitables les textes des offres et de procéder à une première réduction du nombre de termes :

- lemmatisation et étiquetage grammatical à l'aide de l'algorithme proposé par Schmid [1994] ;

- filtrage de certaines catégories grammaticales de sorte à ne conserver que les noms, verbes et adjectifs, à l'instar de Kessler [2009] (les offres d'emploi sont des textes courts composés en grande partie d'énumérations de tâches) ;
- filtrage des termes apparaissant dans moins de cinq documents.

Dans un premier temps, nous allons nous concentrer sur l'identification de la fonction (premier niveau de la nomenclature). Les étiquettes des offres de l'échantillon de test sont considérées comme inconnues, notre objectif étant de les identifier grâce à l'apprentissage effectué.

4.3.6.1 Expérimentations menées dans un cadre non supervisé

Nous tentons en premier lieu de réaliser l'étiquetage des documents grâce à une approche non supervisée. Pour cela, nous réalisons une classification ascendante hierarchique sur les offres de l'échantillon d'apprentissage, et en déduisons les coordonnées des barycentres associés aux 23 classes [12] issues de la découpe du dendrogramme. Une réaffectation des offres est ensuite réalisée à l'aide d'une seule itération de l'algorithme *k-means*, initialisé à l'aide des barycentres calculés précédemment.

Trois méthodes de représentation du texte sont comparées (TF, AC, et LSA avec pondération TF-IDF), ainsi que deux mesures de dissimilarité entre offres (dissimilarité cosinus et distance euclidienne) pour les regroupements effectués via la CAH. Les résultats obtenus sont comparés à l'aide de l'indice de Rand et de l'exactitude. L'indice de Rand ne nécessite pas de connaître les étiquettes des classes obtenues, contrairement à l'exactitude. Le nombre de classes étant assez élevé, nous adoptons la même approche que Slonim et al. [2002] et attribuons à chaque cluster l'étiquette de la fonction d'origine la plus représentée dans le cluster [13].

En comparant les différentes méthodes de représentation du texte, nous souhaitons voir si une des deux techniques de réduction de la dimension (AC et LSA) se montre supérieure à l'autre dans le cadre d'une tâche de classification de documents. Cette comparaison

12. Nous utilisons la connaissance du nombre réel de classes.
13. Cette approche a pour inconvénient que plusieurs clusters peuvent être associés à une même fonction. Toutefois, cela a pour effet d'avantager le critère de performance obtenu dans l'approche non supervisée, ce qui vient confirmer la supériorité de l'approche supervisée.

nous paraît justifiée dans la mesure où les coordonnées sémantiques fournies par les deux techniques se révèlent assez différentes. En effet, la figure 4.5 montre de faibles corrélations sur les 100 premiers axes issus des deux méthodes.

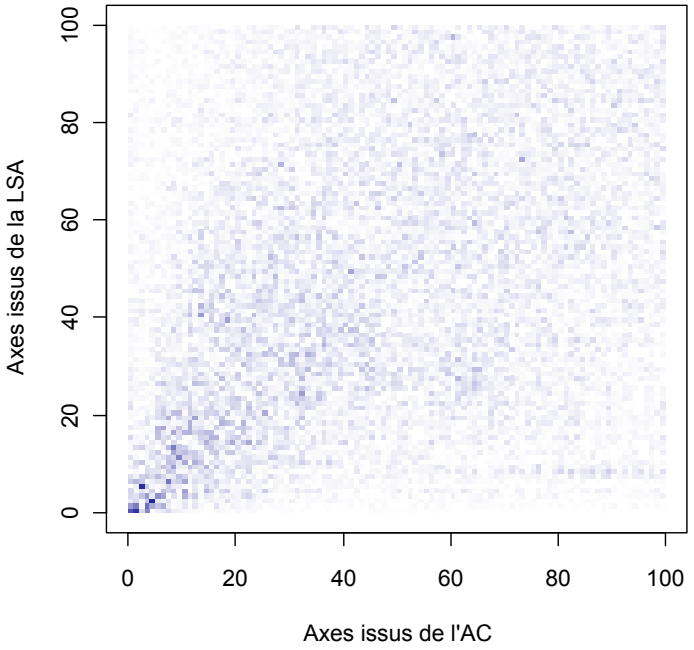

FIGURE 4.5 – Matrice des corrélations sur les 100 premiers axes issus de l'AC et de la LSA (la couleur du pixel indique le dégré de corrélation entre les axes correspondants : de bleu foncé pour une forte corrélation à blanc pour une corrélation nulle)

Les résultats des expérimentations sont présentés par la figure 4.6. Avec une mesure de dissimilarité par la distance euclidienne et sur un grand nombre de dimensions, la représentation LSA permet d'obtenir de bien meilleurs résultats que la représentation par AC pour les deux indicateurs de qualité étudiés. Sur un petit nombre de dimensions (une dizaine d'axes), l'AC permet d'obtenir des résultats de qualité comparable. Par ailleurs, la LSA montre plus de stabilité avec le nombre de dimensions retenues.

Avec la dissimilarité cosinus, les deux méthodes de représentation du texte présentent une plus grande stabilité avec le nombre de dimensions. La qualité de classification obtenue avec la représentation LSA est comparable pour les deux mesures de dissimilarité entre vecteurs de coordonnées. En revanche, la mesure de dissimilarité cosinus permet d'apporter une forte stabilité aux résultats obtenus lorsque la représentation AC est utilisée, alors

que la qualité décroissait rapidement avec le nombre de dimensions quand la distance euclidienne était employée pour identifier les regroupements de documents.

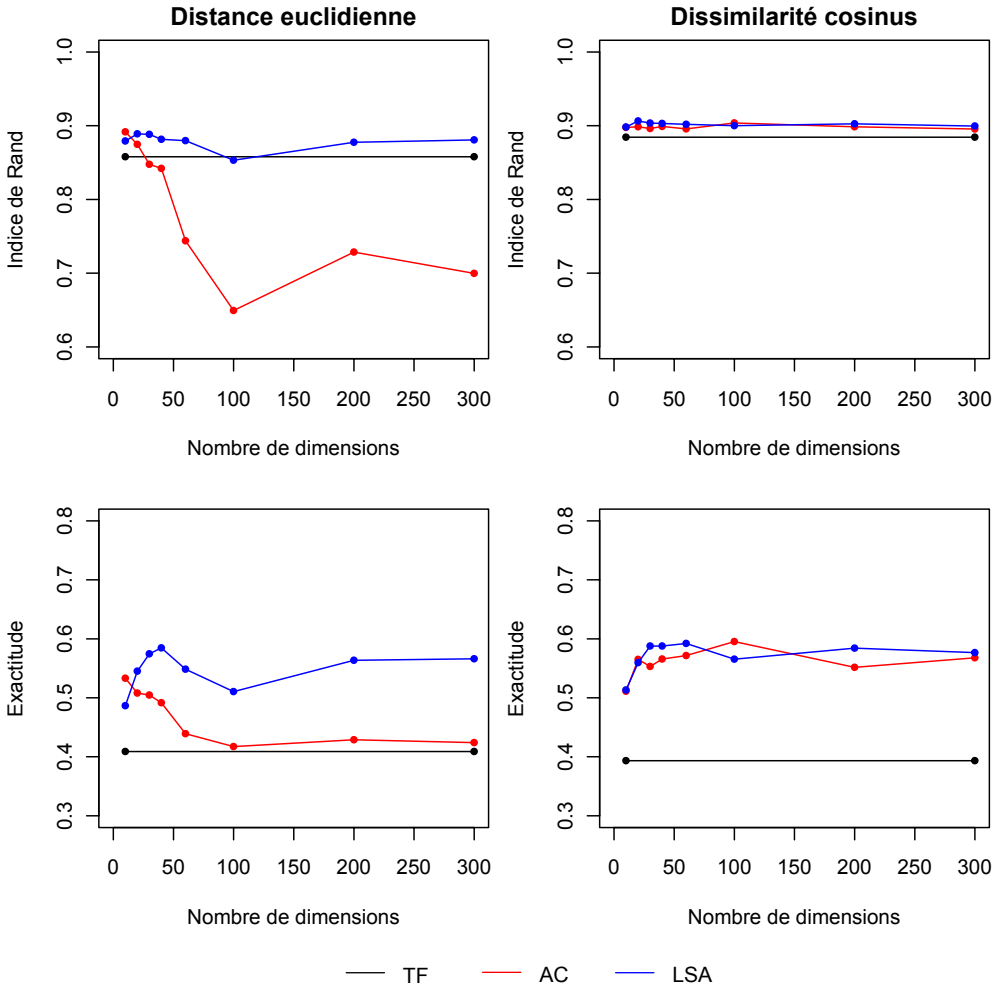

FIGURE 4.6 – Qualité de la classification en fonction de la méthode de représentation du texte et de la mesure de dissimilarité entre documents

4.3.6.2 Expérimentations menées dans un cadre supervisé

Nous nous plaçons maintenant dans un cadre supersivé pour réaliser la tâche de catégorisation des offres d'emploi. Pour cela, nous avons recours aux SVM linéaires avec optimisation du paramètre de coût par validation croisée, notre objectif étant la maximisation de l'exactitude (voir section 4.3.4.1). Comme précédemment, nous comparons plusieurs méthodes de représentation du texte : TF, AC et LSA (avec pondération TF-IDF). Nous étudions

également l'impact de la sélection de descripteurs sur la qualité de la classification obtenue. Dans un premier temps, l'objectif est l'étiquetage des offres selon la fonction du poste à pourvoir. Puis dans un second temps, nous étiquetons les offres de chaque fonction relativement au métier (sous-fonction) associé au poste.

Identification de la fonction du poste. Nous souhaitons limiter le vocabulaire utilisé pour décrire les offres à l'ensemble des termes pertinents pour la discrimination des différentes catégories, afin de réduire la dimension du problème de manière significative (et ainsi réduire les temps d'apprentissage) tout en maintenant le pouvoir de généralisation de l'algorithme. Nous étudions la qualité de la classification en fonction du nombre de termes retenus pour la représentation vectorielle des offres et de la méthode choisie pour la sélection. Nous faisons varier le nombre de termes conservés depuis la totalité du vocabulaire (un peu plus de 4000 termes suite aux pré-traitements) jusqu'à 500 termes (par décrément de 500 termes), pour chacune des statistiques de score calculées.

Nous avons vu dans la section 4.2.1.3 un ensemble de méthodes employées pour la sélection des descripteurs dans le cadre supervisé. Yang and Pederson [1997] mettent en évidence dans une étude comparative les très bons résultats de la statistique du χ^2 pour la tâche de catégorisation de textes, relativement à ceux obtenus avec l'information mutuelle ou le gain d'information. Nous souhaitons ici comparer la statistique du χ^2 à des statistiques issues de l'analyse textuelle permettant d'évaluer le sur-emploi d'un mot au sein d'une catégorie : la spécificité lexicale positive et le Z-score. Les expérimentations menées sur l'impact de la sélection des termes sont réalisées avec une représentation TF des textes. Les résultats obtenus sont présentés dans la figure 4.7.

Les trois méthodes permettent de réduire le nombre de termes à 1500 tout en conservant la qualité de classification obtenue avec la totalité des termes (74% d'exactitude). La spécificité positive et le Z-score permettent la réduction à 1000 termes alors que l'exactitude décroît fortement avec le χ^2 en-dessous de 1500 termes. La F_1-mesure, présentée à titre illustratif, est beaucoup moins stable car fortement influencée par la qualité obtenue sur les petites classes. Toutefois, la sélection des termes a permis d'augmenter la valeur de la F_1-mesure (par exemple, jusqu'à 64.4% avec le Z-score contre 62% avec l'ensemble des

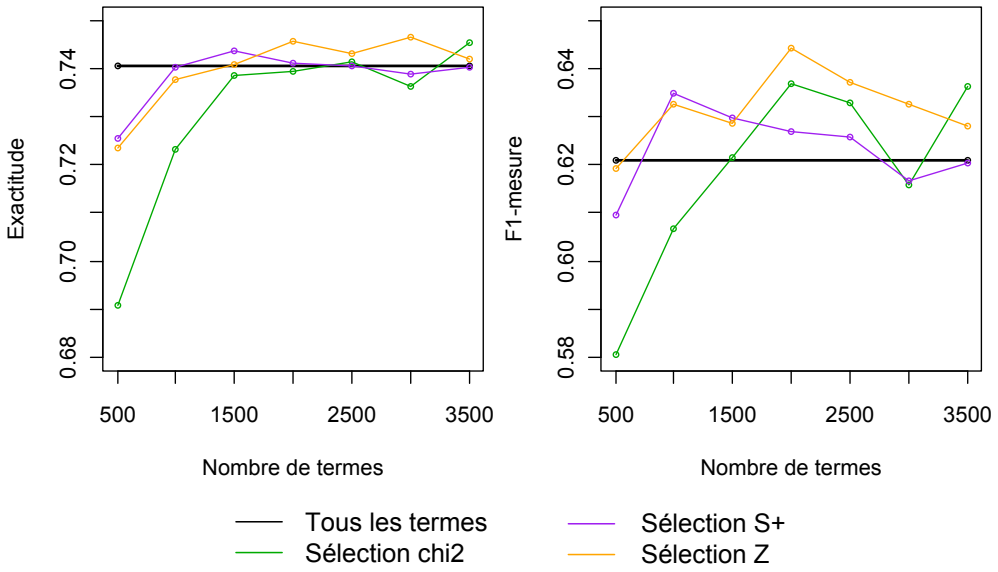

FIGURE 4.7 – Qualité de la classification en fonction du nombre de termes conservés et de la méthode de sélection (représentation TF)

termes).

Nous nous intéressons maintenant à l'évolution du critère de performance en fonction de la méthode choisie pour la représentation des textes (TF, AC ou LSA) et du nombre de dimensions choisi dans le cas des représentations AC et LSA. Nous faisons varier le nombre d'axes retenus de 50 à 400 (par incrément de 50). Nous choisissons de conserver 2000 termes descripteurs (soit presque 50% de l'ensemble de départ), sélectionnés à partir du Z-score, et de comparer les résultats obtenus avec ou sans réduction du nombre de termes. Les résultats sont présentés dans la figure 4.8.

Relativement à une représentation TF avec l'ensemble des termes, l'AC et la LSA permettent une légère amélioration des résultats, tout en réduisant de manière importante la dimension du problème. Lorsqu'un nombre réduit de termes est utilisé (ici 2000 termes sélectionnés à partir du Z-score), l'AC et la LSA permettent d'obtenir une qualité de classification comparable à celle obtenue avec une simple représentation TF. L'intérêt de ces méthodes dans ce contexte applicatif repose donc principalement sur la réduction de la dimension. Pour la suite, nous choisissons une représentation LSA basée sur la sélection de termes, et conservons 250 dimensions. Pour cette représentation, l'exactitude sur l'échan-

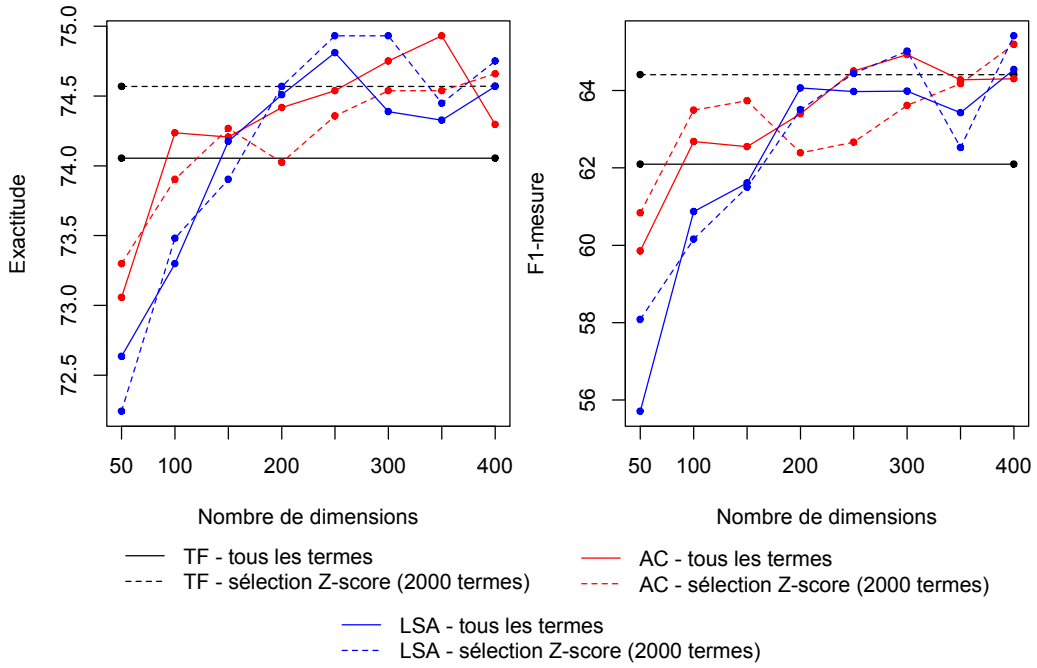

FIGURE 4.8 – Qualité de la classification en fonction de la méthode de représentation et du nombre de dimensions conservées

tillon de test est évaluée à 74.9%, pour une F_1-mesure à 64.4%. Les résultats obtenus pour chaque fonction sont présentés à travers la précision et le rappel dans la figure 4.9.

Nous pouvons constater des disparités assez marquées en termes de performance pour la prédiction des différentes catégories. Pour la plupart des catégories, un niveau de précision d'au moins 60% est assuré, et 17 classes ont une précision supérieure à 70%. Bien qu'un faible niveau de rappel ait été obtenu pour certaines catégories, il demeure supérieur à 70% pour la moitié des classes. Des résultats très satisfaisants sont obtenus pour certaines catégories (*Juridique*, *Hôtellerie / Restauration*, *Ressources Humaines* ou encore *Comptabilité / Finance*), ce qui laisse supposer qu'un vocabulaire très spécifique de ce type de fonction est employé lors de la rédaction des annonces. A contrario, certaines catégories (*Stratégie / Management*, *Services clientèle* et *Gestion de projet*) semblent plus difficiles à prédire. Plusieurs raisons possibles à cela, notamment :

– un faible effectif de la catégorie au départ, induisant un vocabulaire moins représentatif de l'ensemble de la fonction, et par suite une généralisation à de nouvelles annonces plus difficile ;

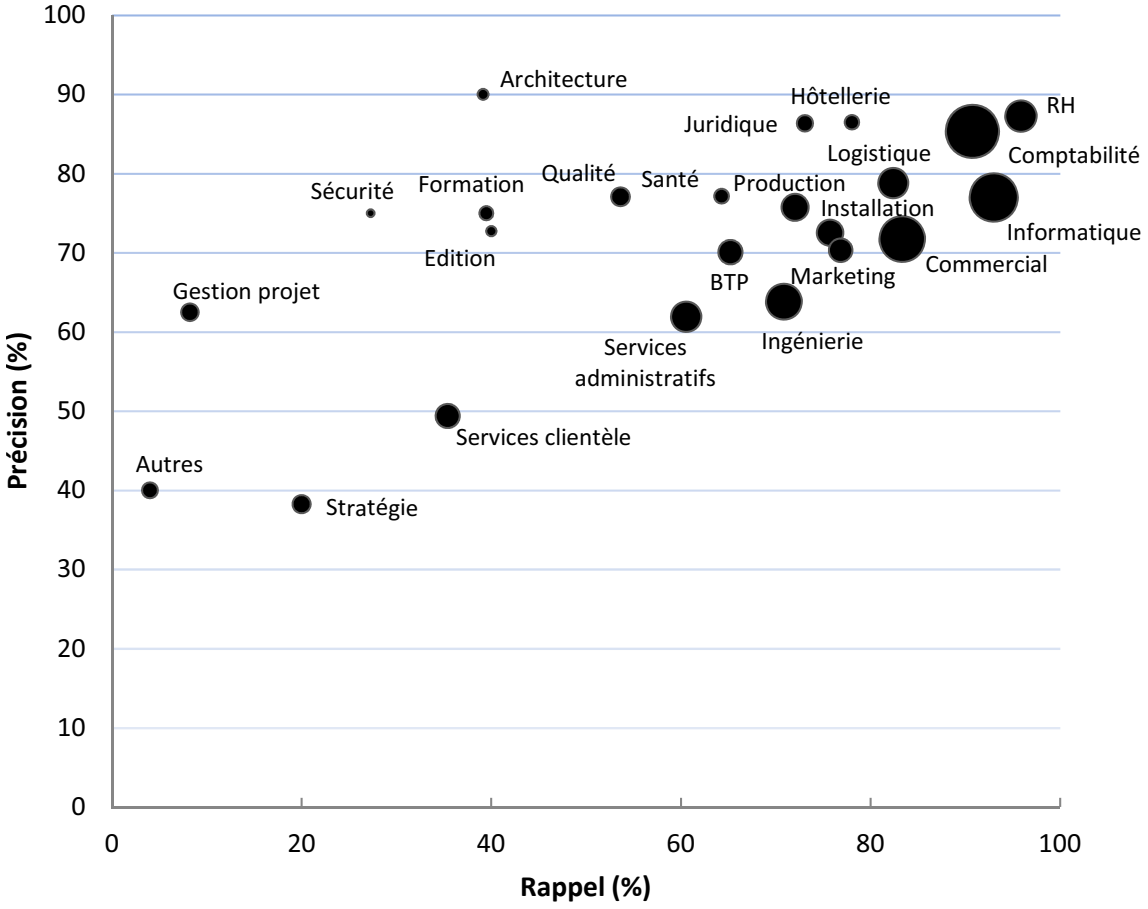

FIGURE 4.9 – Représentation des 23 catégories de fonctions dans le plan rappel × précision (taille des bulles proportionnelle à l'effectif de la catégorie)

– un vocabulaire peu spécifique, commun à plusieurs fonctions, entrainant une discrimination des catégories plus complexe.

La catégorie *Autres* est par construction peu spécifique. Le vocabulaire associé est varié et non représentatif d'une fonction donnée, ce qui entraine une prédiction particulièrement difficile.

Remarque 5 *En plus des représentations TF, AC et LSA, nous avons essayé une représentation TF enrichie des bigrammes de mots. Les offres sont alors représentées par les fréquences de termes, ainsi que les fréquences des bigrammes (suites de deux mots consécutifs extraites à l'issue des pré-traitements). Cependant, les résultats n'ont pas été améliorés par rapport à une représentation TF.*

Identification du métier. Nous souhaitons maintenant identifier les différents métiers existant au sein de chaque fonction. Pour cela, la méthode de représentation du texte retenue pour la catégorisation en fonctions (sélection des termes et LSA) est utilisée et l'algorithme SVM appliqué sur les sous-ensembles issus des fonctions contenant un découpage en métiers. Le nombre de dimensions à retenir avec la LSA est évalué pour chaque fonction à partir de l'exactitude obtenue. Pour la tâche de catégorisation selon le métier, nous obtenons une exactitude globale de 71.7%.

Remarque 6 *La représentation des textes utilisée (coordonnées LSA) est issue de l'espace sémantique obtenu avec un apprentissage sur l'ensemble des fonctions. Des expérimentations ont été menées sur des espaces sémantiques obtenus avec un apprentissage au sein de chaque fonction étudiée. Les résultats obtenus se sont montrés inférieurs.*

Le processus dans son ensemble (prédiction de la fonction puis du métier) donne une exactitude à 55.8% pour l'identification des métiers sur l'échantillon de test (74.9% pour l'identification des fonctions). L'exactitude obtenue lors de la prédiction des métiers est présentée dans le tableau 4.4 pour l'ensemble des fonctions.

Afin d'obtenir des éléments d'explication, nous observons la distribution des fonctions observées sur les fonctions prédites (tableau 4.5) et listons les plus fortes confusions entre métiers afin de visualiser plus en détail la nature des erreurs (tableau 4.6).

Fonction	Effectif test	Nombre de métiers	Exactitude (%)
Architecture, Création	23	2	100.0
Services administratifs	180	6	73.9
BTP & second-œuvre	115	8	64.3
Services clientèle	113	4	77.0
Comptabilité & Finance	552	10	76.8
Hôtellerie, Restauration	41	3	82.9
Informatique & Technologies	457	12	62.4
Installation & Maintenance	136	5	77.9
Ingénierie & Recherche	254	12	57.9
Juridique	52	4	61.5
Logistique & Transport	176	6	74.4
Marketing	108	4	71.3
Production & Opérations	143	4	64.3
Qualité / Inspection	69	4	69.6
Ressources Humaines	193	4	87.6
Stratégie & Management	65	4	70.8
Commercial / Vente	408	9	74.5
Total	**3085**	**101**	**71.7**

TABLE 4.4 – Résultats obtenus au sein de chaque fonction pour la prédiction des métiers

	Fonctions observées																						
	Architecture	Services admin.	Autres	BTP	Services client.	Édition	Comptabilité	Formation	Gest. projet	Hôtellerie	Informatique	Installation	Ingénierie	Juridique	Logistique	Marketing	Production	Qualité	RH	Santé	Sécurité	Stratégie	Commercial
Architecture	39																						
Services admin.		61	12		6	5	2	16	5	7				10	1			1			18		4
Autres			4																	2			
BTP			2	65								2	7				1	1			9	2	
Services client.		2	10		35					2								3					5
Édition			4			40																	
Comptabilité		10	18		18		91		7					12	1	3		1	2			9	2
Formation			2					39												2			
Gest. projet				2					8														
Hôtellerie										78												5	
Informatique	26	5	4	2	4	5		3	51		93	4	13			6		6	1	10		15	
Installation				13	3				2			76	3				5	1			9	2	
Ingénierie	9	1	4	11		5		11	15		2	10	71		2	2	13	22		5	18	3	
Juridique		2				5								73									
Logistique		5	12					3		5					82		3	1		2	9	5	2
Marketing	26		2			25			5		1		2			77	1					5	
Production		1	12			5						5	2		3		72	6		2		3	
Qualité													2		1		1	54		5	9		
RH			4				2	16						4				1	96				
Santé														2						64		5	
Sécurité																		1			27		
Stratégie			2			5		5								6			1	7		20	
Commercial		12	8	3	32	5	3	8	8	7					6	6	1					28	83
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Effectif	23	180	50	115	113	20	552	38	61	41	457	136	254	52	176	108	143	69	193	42	11	65	408

TABLE 4.5 – Distribution des fonctions observées sur les fonctions prédites (% en colonne, les valeurs inférieures à 1% n'apparaissent pas)

Commercial / Vente
Vente d'espace → Commercial
Télévendeur/Commercial sédentaire → Service clientèle/Centre d'appel (Services client.)
Télévendeur/Commercial sédentaire → Commercial
Installation, Maintenance
Matériels informatiques/électriques/télécom → Equipement industriel
Autres métiers- Encadrement et Généralistes → Equipement industriel
Chauffagiste/Climatisation → Equipement industriel
Services administratifs
Assistant Import-Export → Assistanat commercial (Commercial)
Accueil/Réception → Secrétariat/Assistanat de direction
BTP
Electricité → Equipement industriel (Installation)
Maçonnerie/Béton/Sols & Murs → Autres métiers- Encadrement et Généralistes
CAO-DAO / Dessinateur-projeteur → CAO-DAO / Dessinateur-projeteur (Ingénierie)
Ingénierie & Recherche
Hardware/Systèmes embarqués → Développement/Programmation (Informatique)
Télécom - Sans fil & RF → Télécommunication (Informatique)
Génie civil/structures → Autres métiers- Encadrement et Généralistes (BTP)
Qualité / Inspection
Certification/Inspection-batîments → Génie électrique et Génie industriel/Process et méthodes (Ingénierie)
Architecture, Création
Autres métiers- Encadrement et Généralistes → Evènementiel/Communication/Relations Presse (Marketing)
Informatique & Technologies
Développement Jeux → Développement/Programmation
Gestion de projet
Gestion de projet → Gestion de projet IT (Informatique)

TABLE 4.6 – Principales confusions entre métiers au sein de différentes fonctions : métier observé → métier prédit (fonction si différente)

Dans le tableau 4.6, les exemples listés concernent des erreurs représentant plus de 24% des annonces du métier observé. L'analyse du tableau 4.5 peut s'accompagner de celle de la figure 4.9, les valeurs en diagonale n'étant autres que le rappel associé à chaque catégorie. Nous pouvons alors expliquer certains mauvais résultats. En effet, les offres de *Stratégie / Management* semblent principalement confondues avec celles de la fonction *Commercial*, de même que les offres de *Services clientèle*. La fonction *BTP* est prédite également en *Ingénierie* et *Installation*, tandis que la fonction *Services administratifs* est confondue avec les fonctions *Commercial* et *Comptabilité*. La fonction *Gestion de projet* est confondue en grande partie avec la fonction *Informatique*, en particulier à cause du métier *Gestion de projet IT* existant au sein de cette dernière.

Une certaine logique semble ressortir d'une partie des erreurs de prédiction, dans la mesure où les postes des catégories confondues sont proches sémantiquement ainsi qu'en termes de tâches à accomplir. De plus, il ne faut pas négliger le fait que l'étiquetage humain induit une interprétation plus ou moins subjective du poste à proposer. Ainsi, si le système est en désaccord avec l'étiquetage humain, cela désigne potentiellement un recruteur en désaccord avec les autres ou une erreur d'étiquetage.

4.3.7 Conclusions sur le système de catégorisation

L'étude de l'impact du nombre de descripteurs et de leur nature sur la performance du système de catégorisation nous a amenés à plusieurs conclusions. Grâce à la sélection des descripteurs selon leur pouvoir discriminant, leur nombre a pu être réduit de manière importante tout en maintenant l'efficacité de l'algorithme de classification. Trois scores basés sur des statistiques de test ont été comparés. Les trois méthodes de sélection ont montré des efficacités comparables, bien que le Z-score et la statistique de spécificité permettent une réduction plus importante du nombre de termes relativement au χ^2. De plus, l'AC et la LSA appliquées au tableau lexical ont permis de réduire encore le nombre de descripteurs tout en préservant la qualité de la classification. L'évaluation du système sur des échantillons test d'annonces a permis de valider la capacité du système à fournir des catégories homogènes (bonne précision). Par la suite, nous utiliserons ce système dans le cadre de l'analyse et de la compréhension des performances des offres d'emploi. Ce système présente

l'avantage de pouvoir associer une fonction à une offre de manière automatique, à partir du seul descriptif du poste.

4.4 Extraction de mots-clés pertinents

Les annonces d'emploi sont décrites par des données structurées et non structurées (cf. section 4.1). Or, les données structurées ne sont pas suffisantes pour assurer une bonne qualité de prédiction, et une partie importante de l'information est contenue dans le texte de l'annonce. Nous allons donc extraire des prédicteurs supplémentaires à partir du texte (mots-clés) afin d'augmenter le pouvoir explicatif du modèle de prédiction de la performance des annonces. Toutefois, nous sommes confrontés aux problèmes habituellement rencontrés avec les données textuelles (grande dimension, matrices creuses) et devons donc sélectionner les variables à introduire dans le modèle selon leur pertinence. La méthode pour l'identification des mots-clés candidats au modèle explicatif est décrite dans la section 4.4.1, tandis que la section 4.4.2 présente la méthode de sélection parmi les mots-clés candidats.

4.4.1 Extraction de prédicteurs candidats pour contribuer à l'explication de la performance des offres

La méthode la plus simple consisterait à coder l'intégralité des termes apparaissant dans le texte sous forme de variables de fréquences, et de s'intéresser ensuite à leur pouvoir explicatif. Cependant, le nombre initial de termes est très grand (plus de 10 000 termes distincts) et :

- une grande partie de ces termes n'a pas d'utilité pour la prédiction de la performance ;
- introduire un si grand nombre de facteurs dans un modèle explicatif rend l'évaluation de la contribution de chacun d'entre eux difficile et peu fiable.

Nous souhaitons donc effectuer une première sélection des termes afin d'éliminer les mots représentant un bruit (ils ne portent pas d'information pertinente au regard de la problématique de compréhension de la performance des annonces). Nous voulons que cette sélection contienne des mots-clés permettant de décrire plus précisément les missions associées au poste. Pour cela, nous avons recours aux pré-traitements usuels (section 4.2.1.2) et à des techniques de sélection des termes dans le cadre supervisé (section 4.2.1.3). Après un

premier filtrage grâce aux pré-traitements usuels, chaque terme (ou lemme) se voit affecter trois scores issus de la statistique de spécificité lexicale, de la statistique du χ^2, et du Z-score, dont les calculs sont basés sur les catégories de sous-fonctions. Pour chaque méthode, les termes ayant obtenu les 50% de scores les plus élevés sont conservés. La liste des termes candidats est finalement la réunion des ensembles obtenus par les trois méthodes citées précédemment. Chaque terme est ensuite codé par sa fréquence d'apparition à l'intérieur du descriptif des missions. Nous créons également un ensemble de variables binaires indiquant la présence ou absence de ces mots-clés à l'intérieur du titre de l'annonce.

4.4.2 Sélection des mots-clés à introduire dans le modèle de prédiction

A l'issue de la pré-sélection décrite dans le paragraphe précédent, un grand nombre de termes sont encore candidats (2645 termes distincts, pouvant apparaître dans le titre et/ou dans le descriptif). Nous souhaitons sélectionner uniquement ceux qui sont pertinents pour expliquer la performance d'une offre d'emploi. Étant donné que nous sommes en présence d'un très grand nombre de variables explicatives, pouvant être très corrélées entre elles, nous devons avoir recours à une méthode de sélection de variables en présence de multicolinéarité. La régression PLS [Abdi 2010], combinée avec l'usage du critère VIP (*Variable Importance in the Projection*), s'est montrée supérieure aux méthodes de sélection de variables de type Lasso [*Least absolute shrinkage and selection operator* ; Tibshirani 1996] ou régression *stepwise*[14] dans la littérature [Chong and Jun 2005]. Aussi, nous choisissons la méthode PLS-VIP pour déterminer l'ensemble des variables qui seront prises en compte dans le modèle explicatif. La régression PLS et le calcul du VIP sont présentés dans la section 5.3.1.2. Soit p le nombre de variables explicatives, dans la mesure où $\frac{\sum_{j=1}^{p} VIP_j^2}{p} = 1$, la valeur 1 est souvent utilisée comme seuil minimal pour la sélection d'une variable. Cette règle pouvant se révéler très stricte dans la pratique, le seuil 0.8 est parfois choisi comme alternative. Nous choisissons d'utiliser la règle "$VIP_j \geq 1$" pour la sélection des variables à introduire dans le modèle de prédiction, afin d'en diminuer le nombre de manière importante et de ne conserver que les prédicteurs très pertinents.

14. La régression *stepwise* est une méthode de sélection de variables classique consistant à introduire séquentiellement et un par un les prédicteurs dans le modèle. Un critère est choisi pour déterminer : le prédicteur à introduire à chaque étape, si un prédicteur doit être retiré, et l'arrêt de la procédure.

4.4.2.1 Illustration

La méthodologie décrite précédemment est appliquée afin d'obtenir les trois listes de mots-clés candidats, associées aux trois statistiques de score. La majorité des mots-clés sont suggérés par les trois méthodes simultanément, mais de légères différences sont visibles (cf. tableau 4.7). La réunion des trois ensembles fournit une liste de 2645 mots-clés candidats.

Méthodes comparées	**Termes en commun (%)**
Statistique du χ^2 et spécificité lexicale	77%
Statistique du χ^2 et Z-score	85%
Spécificité lexicale et Z-score	83%
Statistique du χ^2, spécificité lexicale et Z-score	73%

TABLE 4.7 – Comparaison des termes retenus avec chaque méthode de sélection (50% des termes sont conservés à partir du score obtenu)

Ces 2645 mots-clés sont associés à 2967 variables (553 variables indiquant la présence ou l'absence d'un mot dans le titre et 2414 variables de fréquences des mots à l'intérieur du descriptif). La sélection des variables est réalisée de manière indépendante pour chaque site d'emploi étudié. Les valeurs observées des VIP associés sont donc différentes pour chaque site, et par suite les ensembles de variables retenues également. Afin d'illustrer les résultats obtenus, nous associons à chaque variable le troisième quartile du VIP (nous l'appelons VIP-75%) sur l'ensemble des sites étudiés. Cet indicateur est robuste car garantit que la valeur du VIP est supérieure à ce seuil pour 25% des sites d'emploi. Les tableaux 4.8 et 4.9 présentent respectivement la liste des 20 mots-clés présents dans le titre et la liste des 20 mots-clés présents dans le descriptif ayant les contributions les plus fortes selon le VIP-75%. Ces tableaux présentent également le signe du coefficient associé lorsque la valeur du VIP-75% est obtenue.

La méthode PLS-VIP nous conduit à conserver en moyenne 515 variables par site d'emploi (en moyenne 51 indicatrices de mots-clés dans le titre et 464 variables de fréquences dans le descriptif) au sein du modèle explicatif.

Terme	VIP	signe	Terme	VIP	signe
assistant	4.9	+	ingenieur	2.6	-
produit	2.5	+	administratif	2.3	+
recrutement	2.1	+	marketing	2.1	+
developpement	2.1	-	charger	2.0	+
direction	1.9	+	comptable	1.7	-
commercial	1.7	+	stagiaire	1.7	-
responsable	1.7	+	gestion	1.6	-
technicien	1.6	-	manager	1.6	+
stage	1.6	-	systeme	1.5	-
rh	1.5	+	gestionnaire	1.4	-

TABLE 4.8 – Liste des 20 termes présents dans le titre ayant les contributions les plus fortes selon le VIP-75% et valeurs associées

Terme	VIP	signe	Terme	VIP	signe
assistant	3.5	+	gestion	3.4	+
courrier	3.4	+	ingenieur	3.3	+
direction	3.3	+	marketing	3.0	+
suivre	2.8	+	pack	2.8	+
experience	2.8	-	technique	2.7	-
administratif	2.7	+	assistanat	2.7	+
commerce	2.7	-	commercial	2.7	-
accueil	2.7	+	java	2.7	-
dossier	2.6	-	projet	2.6	+
organisation	2.6	+	charger	2.6	-

TABLE 4.9 – Liste des 20 termes présents dans le descriptif ayant les contributions les plus fortes selon le VIP-75% et valeurs associées

4.5 Synthèse

Le chapitre 4 est consacré à la présentation des méthodes et applications permettant d'extraire des connaissances à partir du texte des offres d'emploi. Dans un premier temps, nous avons fourni un état de l'art des méthodes usuelles de la fouille de textes, et donné un aperçu des applications aux offres d'emploi rencontrées dans la littérature.

Dans un deuxième temps, nous présentons une méthodologie permettant de classer l'ensemble des offres d'emploi du point de vue des missions proposées en une nomenclature uniforme. Nous avons mené des expérimentations dans des cadres supervisés et non supervisés, et comparé différentes méthodes de représentation du texte. Dans les deux types d'approche, sous des conditions particulières, l'analyse sémantique latente et l'analyse des correspondances permettent d'obtenir des résultats de qualités comparables. Les résultats obtenus étant peu satisfaisants dans le cadre non supervisé, nous adoptons un algorithme supervisé (SVM) pour l'étiquetage des offres d'emploi (apprentissage basé sur une nomenclature établie au préalable). Dans ce contexte, une réduction importante de la dimension est possible grâce à la sélection des termes pertinents et aux techniques d'analyse sémantique (LSA et AC appliquée aux données textuelles). L'étude des confusions entre les catégories prédites et les catégories réelles nous a permis de mettre en évidence des métiers proches du point de vue de la sémantique, et de relativiser l'importance de ces erreurs.

Enfin, des méthodes issues de l'analyse textuelle nous ont permis d'extraire un ensemble de mots-clés candidats qui, recodés en variables, seront utilisés pour enrichir la description des annonces et augmenter le pouvoir prédictif de notre modèle explicatif.

Chapitre 5

Modélisation de la performance d'une offre d'emploi

5.1 Introduction

5.1.1 Contexte

Dans le chapitre 2, nous avons identifié l'indicateur de performance qui pourra être prédit pour toutes les nouvelles campagnes (offres diffusées) : le nombre de CV reçus ou le nombre de clics de redirection selon le mode de candidature choisi par le recruteur. Le nombre de CV ou le nombre de clics pourront être désignés de manière générique par le "nombre de candidatures". Nous avons parcouru les facteurs explicatifs potentiels de la performance d'une campagne dans le chapitre 3. Ce chapitre est dédié à la présentation et à la comparaison de différentes approches pour modéliser la performance d'une offre d'emploi. Après avoir introduit les problématiques liées à la complexité des données que nous traitons, nous présentons la littérature des systèmes de recommandation et positionnons notre système en tant que cas particulier et application innovante. Nous présentons des approches classiques possibles et proposons une nouvelle approche à travers un système hybride de recommandation. Les différentes approches sont comparées à travers des expérimentations. Nous nous focaliserons sur le mode de candidature par e-mail et la prédiction du nombre de CV reçus, la même méthodologie pouvant être appliquée pour la prédiction du nombre de clics de redirection.

5.1.2 Complexité des données et problématiques rencontrées

5.1.2.1 Données en grande dimension

Les annonces d'emploi sont caractérisées par des données structurées et des données non structurées (données textuelles) que nous devons gérer simultanément. Les données structurées (type de contrat, niveau d'études requis, niveau d'expérience requis, secteur, caractéristiques de la localisation, etc.) peuvent être de types différents : variables qualitatives ou quantitatives. Les modalités des variables qualitatives sont recodées en variables indicatrices.

Nous avons montré dans le chapitre 4 comment extraire des variables structurées à partir des données textuelles. Des milliers de facteurs sont extraits de la description du poste grâce à des techniques issues de la recherche d'information et de la fouille de données textuelles.

Nous devons donc gérer des données de très grande dimension et avons affaire à un déséquilibre entre le nombre de variables extraites des données structurées et le nombre de variables extraites des données textuelles, largement supérieur. Il est donc essentiel de filtrer et pondérer les différentes variables au sein de l'algorithme prédictif selon leur pouvoir explicatif.

5.1.2.2 Dimension temporelle

Lorsqu'une annonce est publiée sur un site d'emploi, sa durée de diffusion est déterminée au préalable, et peut être différente d'un site d'emploi à l'autre. La diffusion d'une annonce peut également être interrompue avant son terme par le recruteur. Les candidatures se cumulant au cours du temps (illustration avec le nombre de CV, figure 5.1), l'indicateur de performance doit être considéré simultanément à la durée de diffusion qui lui est associée. De plus, le nombre de candidatures à une offre doit pouvoir être estimé pour toute durée de diffusion. Pour répondre à cette problématique, l'indicateur estimé sera le nombre de candidatures obtenues par jour.

Le flux de candidatures n'étant pas régulier au cours du temps, le nombre journalier de candidatures a une relation décroissante avec la durée de diffusion (voir figure 5.2). En

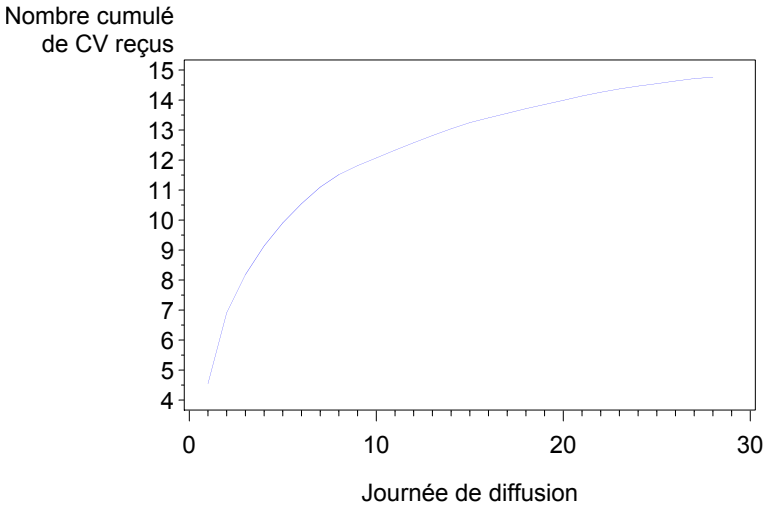

FIGURE 5.1 – Nombre cumulé moyen de CV reçus au cours de la vie d'une annonce sur un site d'emploi

effet, pour une annonce donnée, le nombre journalier de candidatures sera d'autant plus faible que la durée de diffusion est longue, car le nombre de candidatures reçues chaque jour diminue au cours du temps (l'annonce initialement en tête de liste se retrouve plus difficile d'accès sur le site et une grande partie des candidats intéressés postule dès les premiers jours de diffusion).

D'après la figure 5.2, la relation entre le nombre journalier de candidatures et le nombre de jours de diffusion est non linéaire. Nous introduirons donc le logarithme de la durée de diffusion effective de l'annonce en tant que variable au sein de l'algorithme prédictif pour expliquer le nombre journalier de candidatures.

5.2 Systèmes de recommandation

5.2.1 Aperçu de l'état de l'art

5.2.1.1 Contexte

Les entreprises disposent aujourd'hui de bases de données très volumineuses stockant les préférences, notations, achats de l'ensemble de leurs clients ou utilisateurs. Leur objectif est de tirer profit de ces informations passées afin de fournir des suggestions personnalisées aux clients pour leurs prochaines utilisations ou consommations. Sans la suggestion, les

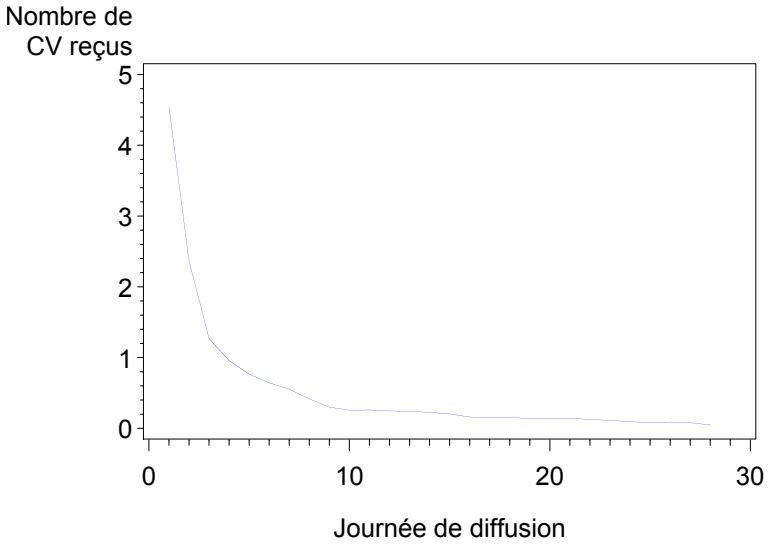

FIGURE 5.2 – Nombre journalier moyen de CV en fonction du nombre de jours de diffusion sur un site d'emploi

clients n'auraient peut-être pas eu connaissance de l'existence du produit. Les systèmes de recommandation s'avèrent donc bénéfiques pour l'entreprise comme pour le client. Un exemple classique est le système de recommandation de films : une matrice de données permet de stocker pour chaque utilisateur (en ligne) la note attribuée à l'ensemble des films (en colonne) qu'il a visionnés, et le système suggère à l'utilisateur les films qu'il n'a pas encore vus et auxquels il est susceptible d'attribuer une note élevée. Les objets (les films dans l'exemple précédent) notés par les utilisateurs sont appelés "items" (i). L'objectif est d'obtenir un classement des items préférés (parmi les items non utilisés) pour un utilisateur (u) quelconque.

Des problématiques d'ordre général apparaissent. D'abord, la plupart du temps, le nombre total d'items étant très élevé et chaque utilisateur n'ayant noté qu'un faible nombre d'items, on est en présence d'une matrice de données très creuse. De plus, certains items sont susceptibles d'avoir très peu voire aucune note attribuée, et de manière symétrique, un nouvel utilisateur n'a encore attribué aucune note. Ce dernier problème est connu sous le nom anglophone de *cold-start problem* (pour un nouvel utilisateur ou un nouvel item). Enfin, les systèmes de recommandation nécessitent des temps de calcul qui croissent de manière non linéaire avec le nombre d'utilisateurs et le nombre d'items. Pour qu'ils soient

utilisables en temps réel, il peut s'avérer nécessaire de mettre en place une architecture particulière pour la structuration des données.

5.2.1.2 Les différents types de système de recommandation

Les systèmes de recommandation suggèrent des items aux utilisateurs sur la base de leurs préférences *explicites* (les notes) ou *implicites* (les achats, les pages visitées, etc.), des préférences des autres utilisateurs, et des attributs des items et utilisateurs (informations sur le contenu, les caractéristiques socio-démographiques, etc.). Trois types de systèmes de recommandation sont distingués dans la littérature.

Systèmes basés sur le filtrage collaboratif. Ces méthodes basent leurs recommandations sur l'historique des préférences des utilisateurs vis-à-vis des items, ignorant les attributs caractérisant les utilisateurs et items [e.g. Shardanand and Maes 1995; Konstan et al. 1997]. Pour proposer un classement des items à l'utilisateur étudié, on exploite les préférences des utilisateurs "similaires" (ceux qui ont attribué des notes proches de celles attribuées par l'utilisateur étudié) pour en déduire une estimation des notes qu'il attribuerait.

Plusieurs algorithmes sont employés pour le calcul des similarités entre utilisateurs/items : corrélation de Pearson [Pearson 1900], similarité cosinus, corrélation de Spearman, etc. Dans le cadre du filtrage collaboratif, il a été montré que le coefficient de corrélation de Pearson s'avère plus efficace que la similarité cosinus [Breese et al. 1998] et le coefficient de corrélation de Spearman [Herlocker et al. 1999]. Un algorithme de prédiction peut être basé sur la similarité entre les vecteurs de préférences des utilisateurs (*user-based model*) ou sur la similarité entre les vecteurs de notes attribuées à différents items (*item-based model*).

Un problème apparaît : si deux utilisateurs similaires (resp. deux items similaires) n'ont jamais noté le même item (resp. jamais été notés par le même utilisateur), alors il ne sera pas possible d'évaluer une corrélation entre eux deux (*non transitive item association*).

Systèmes basés sur le contenu. Ces méthodes sont utilisées la plupart du temps pour recommander des items décrits par un contenu textuel, et recommandent des items similaires à ceux que l'utilisateur a aimé dans le passé [e.g. Pazzani and Billsus 1997; Mooney and Roy 1999]. Les systèmes basés sur le contenu ne tiennent pas compte des préférences explicites ou implicites des autres utilisateurs. Ils permettent d'apporter une solution au problème de "démarrage à froid" grâce à un calcul de proximité entre items (resp. utilisateurs) basé sur les attributs.

Systèmes hybrides. Les systèmes hybrides combinent les deux types précédents pour tirer profit des avantages de chaque méthode [e.g. Balabanovic and Shoham 1997; Schein et al. 2002]. L'hybridation peut être obtenue par différents moyens :

- développer indépendamment un système collaboratif et un système basé sur le contenu, puis combiner les prédictions obtenues ;
- incorporer des caractéristiques basées sur le contenu au sein d'une approche collaborative ;
- incorporer des caractéristiques obtenues de manière collaborative au sein d'une approche basée sur le contenu ;
- construire un modèle général incorporant à la fois des caractéristiques basées sur le contenu et obtenues de manière collaborative.

Remarque 7 *Il existe de nombreuses variantes de systèmes de recommandation dans la littérature, certains basés sur des méthodes heuristiques, d'autres basés sur des fondements théoriques. Les deux types de systèmes sont capables de fournir des résultats de qualités comparables.*

5.2.1.3 Évaluation de la qualité des systèmes

Pour évaluer la capacité d'un système à fournir de bonnes estimations des préférences des utilisateurs, les notes estimées sont habituellement comparées aux notes réelles à l'aide de l'erreur absolue moyenne (*mean absolute error* ou MAE), ou de la racine carrée de l'erreur carrée moyenne (*root mean squared error* ou RMSE). Les indicateurs de précision et rappel (cf. section 4.3.4.1) peuvent également être utilisés lorsqu'il est d'intérêt d'évaluer

la capacité du système à fournir une liste ordonnée des items préférés de l'utilisateur.

5.2.2 Application innovante et cas particulier de système de recommandation

5.2.2.1 Problématiques communes

Dans ce chapitre, nous introduisons un algorithme prédictif de la performance (ou rendement) d'une annonce diffusée sur un site d'emploi. Pour cela, nous avons recours à une application innovante de système de recommandation dans laquelle :

- les items sont les annonces d'emploi,
- les utilisateurs sont les sites d'emploi,
- les notes sont les rendements (performances) des annonces observés sur les sites d'emploi.

Lorsqu'une nouvelle annonce doit être diffusée, l'objectif est alors d'identifier les sites d'emploi les plus susceptibles "d'apprécier" cette annonce, ce qui se traduit par un rendement élevé sur le site. Le périmètre des sites d'emploi est limité, contrairement à celui des annonces d'emploi. En effet, chaque nouvelle campagne de recrutement peut être différente de toutes les précédentes campagnes. Cependant, si une nouvelle annonce est identifiée comme étant une rediffusion d'une ancienne annonce, les rendements obtenus pour cette annonce antérieure sont utilisés pour améliorer la prédiction fournie pour la rediffusion. Les observations de notre jeu de données sont des annonces postées simultanément sur un ou plusieurs sites d'emploi, et pour lesquelles le rendement associé à chaque canal utilisé est observé. Pour chaque annonce, peu de canaux sont utilisés, et donc peu de rendements sont observés. Par suite, la matrice qui en résulte est une matrice très creuse.

5.2.2.2 Problématiques particulières

Étant donné le contexte de notre application, le système que nous proposons peut être considéré comme un cas particulier de système de recommandation car nous sommes confrontés à des problématiques supplémentaires relativement à celles rencontrées habituellement dans la littérature [Adomavicius and Tuzhilin 2005]. En effet, la plupart des applications concernent de très grands jeux de données (des milliers d'items et d'utilisateurs) où les

utilisateurs ont déjà noté plusieurs items en fonction de leurs préférences. Notre application concerne seulement quelques dizaines d'utilisateurs (les sites d'emploi), et le cas le plus fréquent (nouvelle annonce) nous confronte au problème de démarrage à froid : aucune préférence observée pour l'item d'intérêt. Dans ce cas, seules les variables descriptives des items sont utilisées. Par ailleurs, il n'est pas possible d'obtenir des "notations" pour ces items car chaque nouvel item est recommandé une unique fois à un ou plusieurs utilisateurs à la fois.

Enfin, les notations sont très variables au sein d'un utilisateur et les notations moyennes très variables entre les utilisateurs (voir figure 5.3), tandis que dans les systèmes de recommandation usuels, les notations sont des entiers compris entre 0 et 5, ou entre 0 et 10.

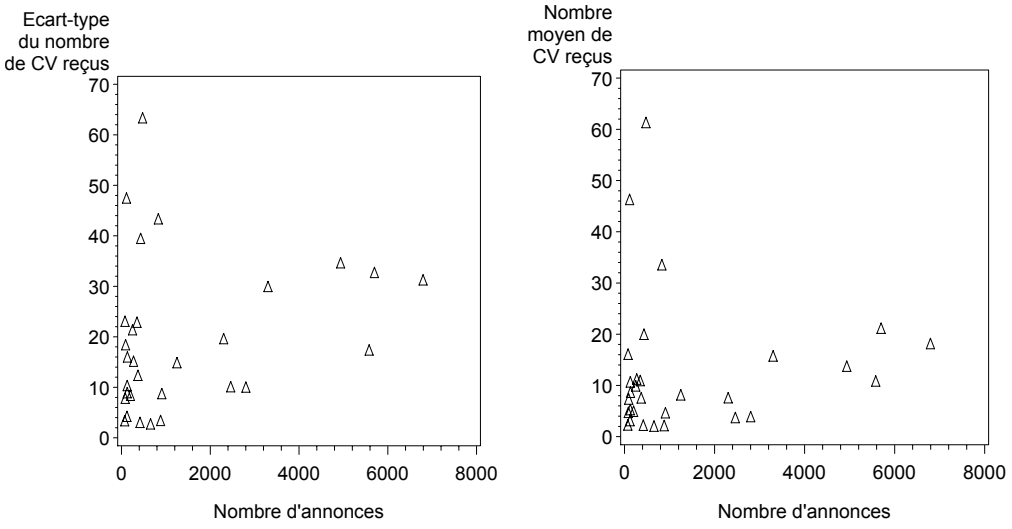

FIGURE 5.3 – Les sites d'emploi représentés sur les plans (nombre d'annonces, écart-type du nombre de CV reçus) et (nombre d'annonces, nombre moyen de CV reçus)

5.3 Modélisation de la performance d'une annonce diffusée sur un site d'emploi

Nous présentons différentes approches possibles pour obtenir un algorithme prédictif de la performance d'une annonce sur un site d'emploi. Dans une première section, nous rappelons brièvement les fondements de quelques approches classiques envisageables étant donné notre contexte applicatif. Dans une deuxième section, nous introduisons deux variantes

d'un système hybride de recommandation incorporant des éléments basés sur le contenu au sein d'une approche collaborative.

5.3.1 Approches standards

5.3.1.1 Le modèle linéaire

Le modèle linéaire peut constituer une première approche pour la modélisation du rendement journalier des offres d'emploi. Toutefois, il présente rapidement des lacunes lorsque les variables explicatives sont nombreuses (le nombre de variables doit être inférieur au nombre d'observations), et très corrélées. Nous présentons succintement la régression linéaire multiple, méthode envisageable dans le cadre de notre application. Elle sera appliquée en utilisant une sélection [1] des variables structurées comme variables explicatives. Soit $X = (x_1, \ldots, x_p)$ l'ensemble des variables explicatives et r la variable dépendante. La relation entre la variable r et les variables explicatives est modélisée par :

$$r = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + u$$

où u est le terme d'erreur. Sous l'hypothèse de normalité du terme d'erreur, les paramètres du modèle peuvent être estimés indifféremment par la méthode des moindres carrés ordinaires ou par maximum de vraisemblance.

5.3.1.2 La régression PLS (PLS-R)

Introduite par Wold et al. [1983a,b] et d'abord présentée comme un algorithme de calcul de vecteurs propres, la régression PLS (*Partial Least Squares regression* ou PLS-R) est rapidement interprétée dans un cadre statistique [e.g. Helland 1990]. La régression PLS est une technique qui généralise et combine des caractéristiques de l'analyse en composantes principales et la régression multiple [voir Abdi 2010, pour plus de détails]. Cette méthode est utilisée lorsque le nombre de prédicteurs est très grand et/ou grand comparé au nombre d'observations. Plus le nombre de prédicteurs est grand, plus le risque de corrélations entre eux est important. En cas de fortes corrélations entre les prédicteurs, les méthodes

1. L'ensemble des variables retenues dans le modèle est déterminé à l'aide d'une méthode de sélection de type *backward*.

de régression classiques ne peuvent pas fournir d'estimations fiables des coefficients. En alternative, la régression PLS fournit des composantes robustes (combinaisons linéaires des prédicteurs), indépendantes et hautement corrélées à la variable à prédire. Dans ces travaux, nous sommes en présence d'un très grand nombre de prédicteurs (des milliers de variables extraites du texte de l'annonce et des variables structurées associées au poste) et de fortes corrélations, cette technique se révèle donc très adaptée.

Calcul des composantes. La première étape de la régression PLS consiste à calculer des composantes indépendantes, et hautement corrélées à la variable à prédire. Le calcul des composantes est basé sur l'algorithme NIPALS, initialement introduit par Wold [1966] pour l'analyse en composantes principales. Soit $(x_1, ..., x_p)$ l'ensemble des variables explicatives, potentiellement corrélées entre elles, et une variable à prédire r. La première composante est calculée ainsi : $t_1 = w_{11}x_1 + \cdots + w_{1p}x_p$, où $w_{1j} = \frac{cov(x_j,r)}{\sqrt{\sum_{j=1}^{p} cov^2(x_j,r)}}$. Le résidu r_1 est obtenu en régressant r sur t_1.

La seconde composante t_2 est une combinaison linéaire des résidus x_{1j} provenant des régressions des x_j sur t_1 : $t_2 = w_{21}x_{11} + \cdots + w_{2p}x_{1p}$, où $w_{2j} = \frac{cov(x_{1j},r_1)}{\sqrt{\sum_{j=1}^{p} cov^2(x_{1j},r_1)}}$. La régression de r sur t_1 et t_2 donne le résidu r_2.

Ce processus itératif continue jusqu'à obtenir H composantes, où H est déterminé par validation croisée [Hoskuldsson 1988]. Finalement, r est régressé sur les composantes PLS : $r = c_1t_1 + c_2t_2 + \ldots + c_Ht_H + r_H$.

Calcul du VIP. Pour mesurer l'importance de la variable x_j pour expliquer r à travers les composantes t_h, nous utilisons le critère VIP (*Variable Importance in the Projection*) :

$$VIP_{hj} = \sqrt{\frac{p}{\sum_{l=1}^{h} cor^2(r,t_l)} \sum_{l=1}^{h} cor^2(r,t_l) w_{lj}^2}$$

Dans la mesure où $\sum_{j=1}^{p} VIP_{hj}^2 = p$, le seuil minimal pour conserver une variable dans le modèle est souvent fixé à 1.

5.3.2 Système hybride de recommandation

Dans notre contexte, les items (annonces d'emploi) sont décrits par un ensemble de variables extraites de leur description et un ensemble de variables extraites de données structurées (type de contrat, localisation, niveau d'études requis, etc.). La description est un texte auquel nous avons appliqué des techniques issues de la recherche d'information et de la fouille de données textuelles (présentées dans la section 4.2.1.5). L'approche est présentée dans son ensemble par la figure 5.4.

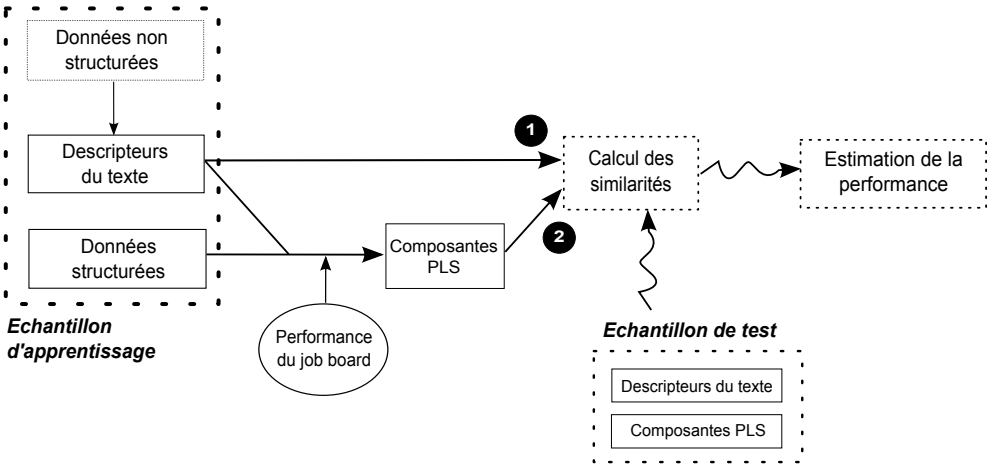

FIGURE 5.4 – Vue d'ensemble du système hybride de recommandation

Les notations suivantes seront utilisées dans cette section :

- $X = (x_1, \ldots, x_p)$ l'ensemble des variables décrivant les items ;
- $X_i = (x_{1i}, \ldots, x_{pi})$ le vecteur des caractéristiques de l'item i ;
- $r_{u,i}$ la note réelle de l'item i pour l'utilisateur u ;
- $p_{u,i}$ la note prédite de l'item i pour l'utilisateur u.

5.3.2.1 Calcul de similarité entre deux offres

Les systèmes hybrides que nous proposons sont basés sur le calcul de similarités entre items, supposant que des items similaires ont des notes similaires pour un utilisateur donné. Nous proposons deux méthodes pour le calcul des similarités entre items :

- une similarité "naïve" entre deux vecteurs de caractéristiques,

– et une similarité "supervisée" permettant d'identifier les plus proches voisins relativement aux variables pertinentes (celles qui contribuent à l'explication des rendements sur les sites d'emploi).

Lorsque la note d'un nouvel item doit être estimée pour un utilisateur donné, sa similarité par rapport à l'ensemble des items déjà notés par cet utilisateur est calculée.

Fonctions de similarité. Pour comparer deux vecteurs de descripteurs extraits de textes, nous avons recours dans un premier temps à la similarité cosinus [Baeza-Yates and Riberto-Neto 1999] entre les items i_1 et i_2 :

$$sim(i_1, i_2) = cos(X_{i_1}, X_{i_2}) = \frac{\sum_{j=1}^{p} x_{ji_1} x_{ji_2}}{\sqrt{\sum_{j=1}^{p} x_{ji_1}^2} \sqrt{\sum_{j=1}^{p} x_{ji_2}^2}}$$

D'autres fonctions de similarité basées sur la distance euclidienne entre les items (d_{i_1,i_2}) sont utilisées (la similarité étant une fonction décroissante de la distance euclidienne). D'abord, nous testons la mesure de similarité constante en tant que cas particulier. Nous utilisons ensuite la même méthode que Shardanand [1994] pour obtenir une troisième mesure de similarité :

$$sim(i_1, i_2) = \max_{i_k \in K}(d_{i_1,i_k}) - d_{i_1,i_2}$$

où K est l'ensemble des $|K|$ plus proches voisins de i_1 selon la distance euclidienne.

Enfin, nous testons deux fonctions de similarité additionnelles : des fonctions gaussienne et exponentielle décroissantes avec la distance euclidienne. Ces fonctions dépendent d'un paramètre d'écart-type, pour lequel nous testerons différentes valeurs. Soit σ_d l'écart-type de la distance entre deux items sur le périmètre de l'utilisateur étudié, les fonctions de similarité gaussienne et exponentielle sont respectivement définies par :

$$sim(i_1, i_2) = exp\left\{-\frac{1}{2}\left(\frac{d_{i_1,i_2}}{\sigma_d}\right)^2\right\} \text{ et } sim(i_1, i_2) = exp\left\{-\frac{d_{i_1,i_2}}{\sigma_d}\right\}$$

Similarité "naïve" et similarité "supervisée". Deux approches différentes sont proposées. Dans la première approche, les notes sont prédites à l'aide d'une méthode heuristique. Cette approche suppose que des offres d'emploi (items) similaires du point de vue de leurs

caractéristiques devraient avoir des rendements (notes) proches pour un site d'emploi (utilisateur) donné. Cette similarité est dite "naïve" dans la mesure où elle repose sur l'hypothèse que toutes les variables décrivant les offres d'emploi sont utiles, et dans une même mesure, pour expliquer les rendements sur les sites d'emploi.

Afin de ne prendre en compte que les variables pertinentes pour la recherche des voisins les plus proches, nous proposons comme seconde approche une mesure de similarité supervisée, orientée sur la prédiction des rendements des offres. Au lieu d'évaluer la distance euclidienne sur les vecteurs de caractéristiques des offres, elle est évaluée sur les vecteurs de composantes PLS (voir section 5.3.1.2), combinaisons linéaires des prédicteurs initiaux extraites de sorte à expliquer le mieux possible les rendements des offres.

5.3.2.2 Estimation de la performance

Les rendements sont estimés à l'aide d'une fonction d'aggrégation [voir par exemple Adomavicius and Tuzhilin 2005], évaluée sur le voisinage de l'offre. Le rendement attendu sur le job board u pour l'offre i_1 est donné par :

$$p_{u,i_1} = \frac{\sum_{i_{k'} \in K} sim(i_1, i_{k'}) \times r_{u,i_{k'}}}{\sum_{i_{k'} \in K} sim(i_1, i_{k'})}$$

où K est l'ensemble des $|K|$ plus proches voisins de i_1 relativement à la mesure de similarité correspondante. Pour la similarité constante, qui signifie estimer le rendement par la moyenne sur le voisinage de i_1, les plus proches voisins sont déterminés par la distance euclidienne.

5.3.2.3 Amélioration de la prédiction par *relevance feedback*

Parfois, une nouvelle offre est en fait identique à une offre précédemment postée (mêmes critères, même descriptif). En effet, certains postes à pourvoir sont des postes récurrents et les recruteurs sont susceptibles de rediffuser les anciennes annonces correspondantes. Une offre qui a déjà été postée par le passé est appelée une *rediffusion*. Quand une rediffusion est identifiée, nous détenons une information additionnelle : les rendements observés sur les job boards choisis par le recruteur lors de la (des) diffusion(s) précédente(s). Le fait

d'exploiter cette information dans le but d'améliorer la prédiction peut être assimilé à un système de "retour de pertinence", connu sous le nom anglophone de *relevance feedback* [Rocchio 1971]. En effet, nous proposons un item (une offre d'emploi) à l'utilisateur (le job board) et celui-ci nous retourne un jugement de pertinence sur l'item qui lui a été proposé (la performance observée de l'offre). Nous exploitons ce retour pour améliorer les propositions qui lui seront faites les fois suivantes.

Nous rappelons que $X_i = (x_{1i}, \ldots, x_{pi})$ est le vecteur des caractéristiques de l'offre i. Les variables de rendements associés aux job boards $u_1, \ldots, u_N$ sont discrétisées en l catégories codées à travers des variables indicatrices. Nous créons un vecteur enrichi :

$$\tilde{X}_i = (x_{1i}, \ldots, x_{pi}, r^1_{u_1 i}, \ldots, r^l_{u_1 i}, \ldots, r^1_{u_N i}, \ldots, r^l_{u_N i})$$

qui sera utilisé pour modéliser et estimer les rendements sur les job boards en cas de rediffusion. Dans le vecteur $\tilde{X}_i$, r^l_{ui} est égal à 1 si le rendement observé de l'offre i sur le job board u appartient à la catégorie l.

5.3.2.4 Évaluation du système

Critère de qualité. La qualité de nos systèmes de recommandation sera comparée avec la qualité d'un système basique fournissant des recommandations basées sur la note moyenne de l'utilisateur (évaluée sur l'ensemble des notes passées) et ne prenant pas en compte les caractéristiques des items. Ce système sera appelé "recommandeur moyen" (*average recommender* en anglais ou AR) par la suite. $p^{AR}_{u,i}$, la note prédite par le recommandeur moyen pour l'utilisateur u et l'ensemble des items, est définie par :

$$p^{AR}_u = \frac{\sum_{i \in D_u} r_{u,i}}{|D_u|}$$

où D_u est l'ensemble des items précédemment notés par l'utilisateur u.

Pour évaluer la capacité d'un système à fournir de bonnes estimations des préférences, les valeurs estimées sont habituellement comparées aux valeurs réelles à travers les indicateurs MAE et RMSE. Dans notre contexte, les utilisateurs ont noté des volumes d'items très différents. Cela peut biaiser notre critère de qualité car il y a une grande variabilité entre les notes moyennes des différents utilisateurs (voir figure 5.3).

Nous utiliserons le critère $\overline{MAE}$, la moyenne des critères MAE associés aux différents utilisateurs, pour comparer la qualité des approches sur la tâche de prédiction :

$$\overline{MAE} = \frac{1}{N} \sum_{u \in U} \frac{\sum_{i \in D_u} |p_{u,i} - r_{u,i}|}{|D_u|}$$

où U est l'ensemble des utilisateurs, $N = |U|$ le nombre d'utilisateurs, et $p_{u,i}$ la note estimée pour l'utilisateur u et l'item i.

Méthode d'évaluation du critère de qualité. Comme l'illustre la figure 5.4, l'extraction des descripteurs textuels (définition du périmètre du vocabulaire, pondération des termes, analyse sémantique latente et analyse des correspondances) est réalisée sur un échantillon d'apprentissage. De même, les composantes PLS sont déterminées à partir de l'échantillon d'apprentissage. L'estimation de la performance est entièrement réalisée sur un échantillon de test, permettant ainsi de reproduire le contexte réel de l'application. Les textes et variables associés aux annonces de l'échantillon de test sont projetés dans les espaces définis en apprentissage. Puis l'estimation est faite sur le voisinage de l'annonce de test sur la base des rendements obtenus par les annonces d'apprentissage les plus similaires.

5.4 Expérimentations

5.4.1 Description des données

Le jeu de données étudié est constitué d'annonces postées sur des sites généralistes et spécialisés disposant d'un historique d'au moins 70 annonces. Nous étudions les annonces postées avec un mode de candidature par e-mail et prédisons le nombre de CV attendus pour une annonce postée sur un site d'emploi donné. Seuls les sites ayant un rendement moyen supérieur ou égal à 2 CV sont étudiés afin d'évaluer la performance du système pour les sites d'emploi où la problématique de prédiction est réellement d'intérêt.

Ce type d'application est assez inhabituel dans la littérature des systèmes de recommandation car nous étudions un petit jeu de données : quelques utilisateurs (30 job boards), et environ 42 000 notations (rendements observés). Nous étudions environ 18 000 items

(annonces), chacun ayant en moyenne trois notes attribuées (une annonce est postée en moyenne sur trois sites d'emploi du jeu de données).

Les rediffusions sont identifiées et seront traitées séparément (environ 1400 annonces). Le jeu de données restant est affecté aux échantillons d'apprentissage et de test, avec une répartition de 60% et 40% respectivement. Dans les expérimentations, l'échantillon des rediffusions sera traité de la même manière que l'échantillon de test.

Les variables explicatives listées ci-dessous seront utilisées au sein du système hybride pour enrichir la description des offres d'emploi :

- nom du recruteur (si plus de 50 annonces multidiffusées) ;
- type de recruteur (cabinet de recrutement, entreprise, SSII, etc.) ;
- secteur d'activité ;
- variables démographiques au niveau du code postal, de la zone d'emploi et du département de la localisation (population totale, population active, nombre de chômeurs, etc.) ;
- type de contrat ;
- niveau d'études requis ;
- niveau d'expérience requis ;
- zone géographique (région basée sur l'indicatif téléphonique) ;
- fonction ;
- mois et année de publication de l'annonce ;
- indicateurs du style rédactionnel (longueur du descriptif, nombre moyen de mots par phrases, nombre moyen de caractères par mot, etc.).

5.4.2 Comparaison des résultats

Nous souhaitons dans un premier temps comparer différentes méthodes de représentation du texte afin d'identifier celle qui permet d'obtenir le plus faible $\overline{MAE}$. Les représentations TF, LSA avec pondération TF-IDF, et analyse des correspondances (AC) seront comparées. Pour ces expérimentations, seuls les prédicteurs extraits du texte seront utilisés dans un premier temps. Ces méthodes de représentation du texte seront comparées à travers trois algorithmes prédictifs : la régression PLS ($S1$), un système de recommandation hy-

bride basée sur une similarité naive ($S2$) et un système de recommandation hybride basé sur une similarité supervisée ($S3$). Dans les approches $S2$ et $S3$, plusieurs fonctions de similarité sont comparées (voir section 5.3.2.1) et les résultats présentés en fonction de la taille du voisinage utilisé pour l'estimation. La similarité cosinus n'est pas pertinente pour l'approche $S3$ car le nombre de composantes conservées est très petit (moins de 10 composantes) et parfois égal à 1. Dans la mesure où $S2$ se base sur une approche non supervisée, le nombre de dimensions conservées suite à la LSA et l'AC peut influencer l'efficacité du système de recommandation : des expérimentations préliminaires nous conduisent à conserver 50 dimensions pour les deux méthodes.

Les résultats obtenus avec la régression PLS sont présentés dans le tableau 5.1.

Approche	AR	PLS-R		
Représentation du texte	×	TF	LSA (TF-IDF)	AC
$\overline{MAE}$	12.5	10.8	11.1	11.2

TABLE 5.1 – Résultats obtenus avec la régression PLS ($S1$)

Pour les approches $S2$ et $S3$, nous discutons d'abord le choix du paramètre de variance dans les fonctions de similarité gaussienne et exponentielle. Les valeurs suivantes seront testées respectivement pour les fonctions gaussienne et exponentielle : $\sigma_g \in \{\sigma_d, \frac{1}{2}\sigma_d, \frac{1}{3}\sigma_d, \frac{1}{4}\sigma_d, \frac{1}{6}\sigma_d\}$ et $\sigma_e \in \{\sigma_d, \frac{1}{3}\sigma_d, \frac{1}{6}\sigma_d, \frac{1}{8}\sigma_d, \frac{1}{10}\sigma_d\}$. Les résultats obtenus sont illustrés par la figure 5.5, pour la représentation TF (les courbes obtenues avec les représentations LSA et AC présentent la même allure). Les valeurs retenues sont reportées dans le tableau 5.2.

Approche	$S2$			$S3$		
Représentation du texte	TF	LSA	AC	TF	LSA	AC
Fonction gaussienne	1/3	1/3	1/4	1, 1/2	1, 1/2	1, 1/2
Fonction exponentielle	1/6, 1/8	1/6, 1/8	1/6, 1/8	1, 1/3	1	1/3

TABLE 5.2 – Valeur(s) retenue(s) ($\times\sigma_d$) pour les fonctions gaussienne et exponentielle dans les approches $S2$ et $S3$

Dans l'approche $S2$, le plus faible $\overline{MAE}$ est atteint avec $\sigma_g^{opt,S2} \in \{\frac{1}{3}\sigma_d, \frac{1}{4}\sigma_d\}$ et $\sigma_e^{opt,S2} \in \{\frac{1}{6}\sigma_d, \frac{1}{8}\sigma_d\}$ respectivement pour les fonctions gaussienne et exponentielle. Dans l'approche $S3$, réduire la valeur initiale du paramètre de variance ne permet pas d'améliorer la qualité de prédiction.

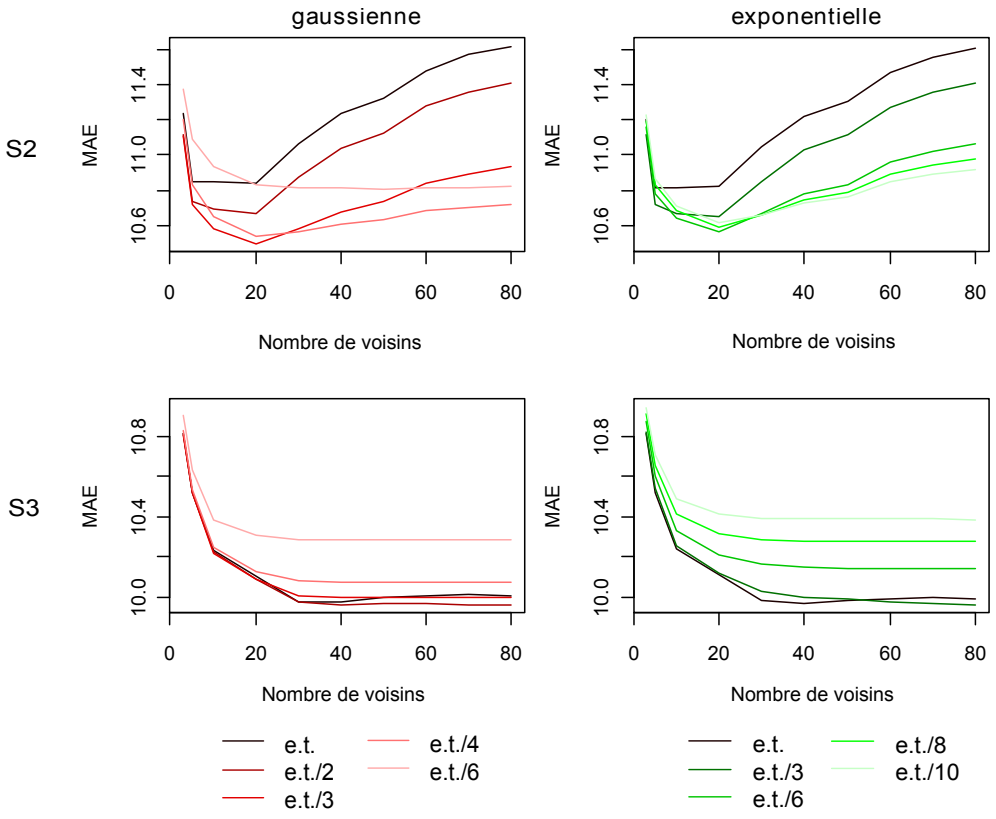

FIGURE 5.5 – $\overline{MAE}$ obtenu avec les systèmes $S2$ et $S3$, en fonction du paramètre de variance (écart-type "e.t.") dans les fonctions de similarité gaussienne et exponentielle (représentation TF)

Nous retenons les meilleurs paramètres pour les fonctions gaussienne et exponentielle, et comparons les méthodes de représentation du texte pour les approches $S2$ et $S3$ dans la figure 5.6.

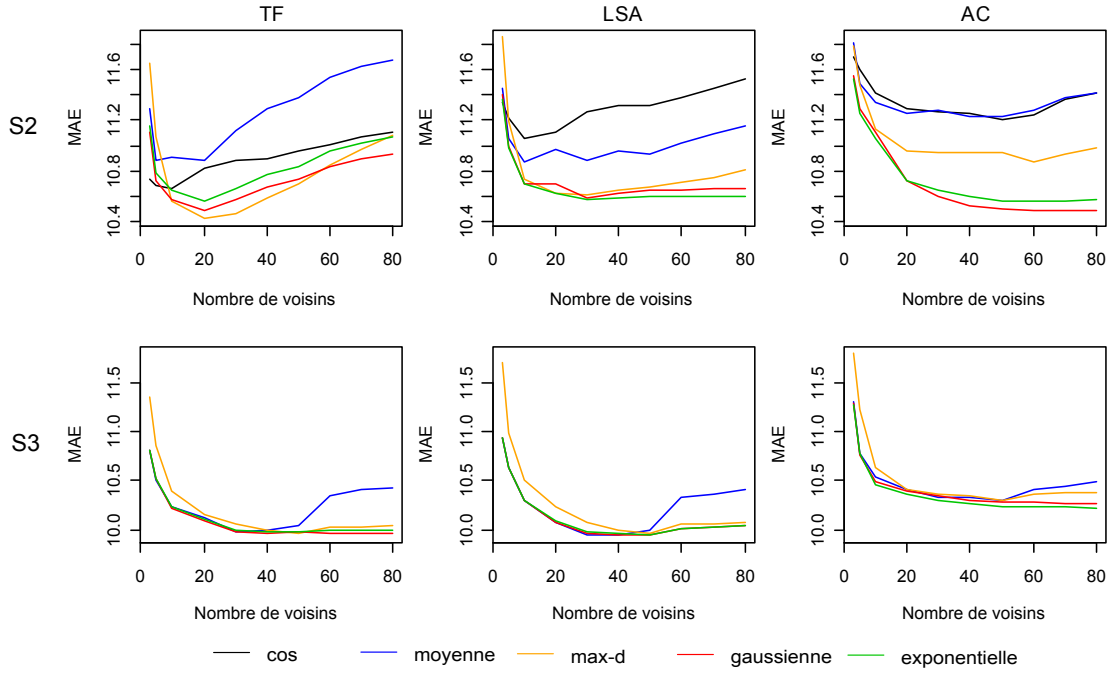

FIGURE 5.6 – $\overline{MAE}$ obtenu avec les systèmes $S2$ et $S3$, en fonction de la méthode de représentation du texte et de la mesure de similarité

Dans le système $S3$, les représentations TF et LSA sont équivalentes tandis que l'analyse des correspondances fournit des résultats de qualité légèrement inférieure. L'indicateur $\overline{MAE}$ est stable à partir de 30 voisins pris en compte, sauf pour la méthode d'estimation par la moyenne sur le voisinage. Dans le système $S2$, les trois méthodes de représentation permettent d'atteindre une qualité de prédiction comparable. Toutefois, la LSA et l'AC montrent plus de stabilité avec le nombre de voisins retenus. Dans cette approche, le choix de la mesure de similarité utilisée est important car les performances associées diffèrent.

Dans ce qui suit, nous choisissons la représentation TF car elle fournit des résultats légèrement meilleurs pour les approches $S1$ et $S2$. Cependant, il est important de noter que la LSA a permis de réduire la dimension des données de manière importante tout en préservant la qualité de la prédiction. Les trois approches sont finalement comparées à

travers la figure 5.7. La mesure de similarité $max(d) - d$ est retenue pour le système hybride $S2$, tandis que la similarité gaussienne est préférée pour le système hybride $S3$. Les trois approches surpassent largement le "recommandeur moyen" (AR), mais c'est le système $S3$ qui permet d'atteindre le plus faible $\overline{MAE}$ et qui apporte le plus de stabilité.

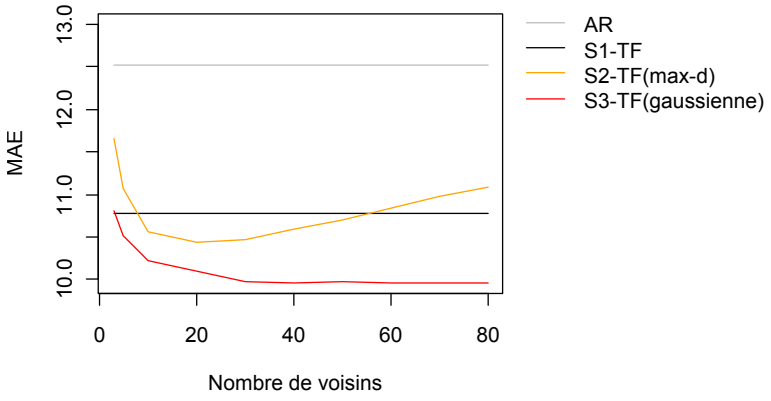

FIGURE 5.7 – Comparaison des meilleurs algorithmes de chaque approche (la fonction de similarité retenue est indiquée entre parenthèses)

5.4.3 Enrichissement de la description des annonces

Nous ajoutons maintenant les variables structurées à l'ensemble des descripteurs d'une annonce d'emploi. Dans la mesure où l'approche $S2$ est basée sur une similarité textuelle, elle ne permet pas de gérer l'ajout de variables qualitatives et quantitaves au vecteur des descripteurs.

Nous menons les mêmes expérimentations que précédemment afin d'identifier les meilleurs paramètres pour les fonctions de similarité gaussienne et exponentielle : $\sigma_g^{opt} = \frac{1}{3}\sigma_d$ et $\sigma_e^{opt} = \frac{1}{3}\sigma_d$. Les résultats obtenus pour les approches $S1$ et $S3$, avec ou sans l'ajout des variables structurées, sont présentés dans la figure 5.8. L'allure des courbes reste la même, mais l'enrichissement de la description des annonces (caractéristiques du poste, du recruteur, etc.) a permis d'améliorer la qualité de l'algorithme prédictif.

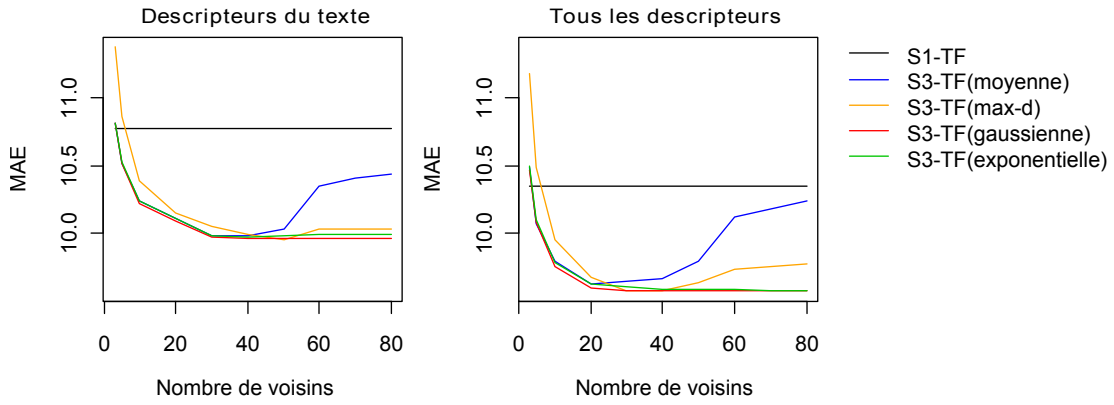

FIGURE 5.8 – $\overline{MAE}$ obtenu avec les descripteurs du texte seuls et avec l'ajout de variables descriptives (la fonction de similarité retenue est indiquée entre parenthèses)

5.4.4 *Relevance feedback*

Nous étudions maintenant l'échantillon des rediffusions. Ces offres ont déjà été diffusées par le passé sur un ou plusieurs sites d'emploi, et notre objectif est d'exploiter les rendements observés lors de cette première diffusion afin d'améliorer la prédiction pour cet échantillon. Dans l'approche avec *relevance feedback* (RF), l'apprentissage est fait sur l'ensemble des autres annonces, y compris les annonces initiales associées aux rediffusions. Les annonces sont décrites par un vecteur enrichi, comme présenté dans la section 5.3.2.3. L'approche sans *relevance feedback* nécessite seulement les descripteurs extraits du texte ainsi que les variables descriptives complémentaires. Comme précédemment, le vecteur descriptif enrichi ne peut être utilisé qu'avec les approches $S1$ et $S3$. La figure 5.9 montre les résultats obtenus pour ces deux approches, avec ou sans le recours au *relevance feedback*. Nous constatons que grâce à celui-ci, l'algorithme est plus efficace lors de la prédiction du rendement associé à une rediffusion (l'indicateur $\overline{MAE}$ est réduit de 1.5).

5.5 Illustration des résultats et discussion

Les forces et faiblesses des approches proposées sont résumées dans le tableau 5.3. La régression PLS implique une relation linéaire entre les composantes et la variable dépendante, tandis que les autres approches sont non linéaires avec une estimation au voisinage des annonces. De plus, le risque de surapprentissage est plus élevé avec la régression PLS (à

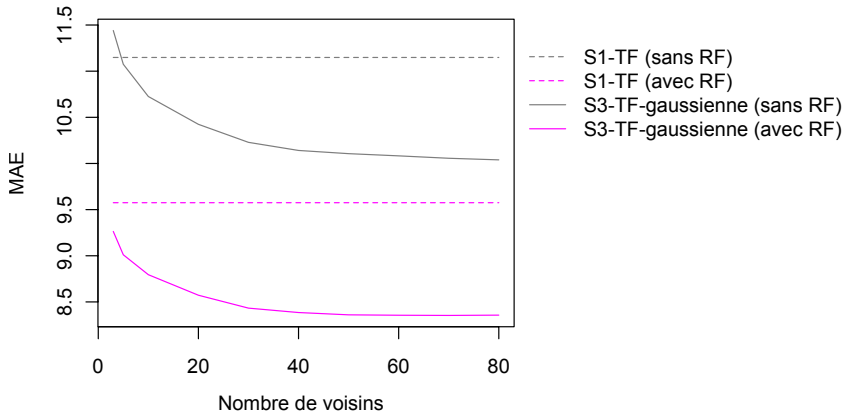

FIGURE 5.9 – $\overline{MAE}$ obtenu avec ou sans *relevance feedback*

cause du très grand nombre de prédicteurs pris en compte et de l'estimation des paramètres basée sur un échantillon d'apprentissage) qu'avec les autres approches. Nous apprécions la capacité des systèmes $S1$ et $S3$ à fournir des outils d'interprétation de l'impact des variables durant la phase de modélisation PLS. Ces deux systèmes permettent également l'ajustement des poids en fonction de la pertinence du descripteur vis à vis de l'estimation du rendement. En revanche, le système $S2$ ne permet pas d'expliquer l'impact des facteurs pris en compte, ni de contrôler les poids affectés aux descripteurs.

Type d'approche	**Contrainte de linéarité**	**Risque de surapprentissage**	**Interprétation des facteurs**	**Ajustement des poids**
PLS-R ($S1$)	oui	oui	oui	oui
Hybride $S2$	non	non	non	non
Hybride $S3$	non	faible	oui	oui

TABLE 5.3 – Forces et faiblesses des approches proposées

Le tableau 5.4 présente le détail des meilleurs résultats obtenus avec les différentes approches pour chaque site d'emploi étudié. Les résultats obtenus avec la régression linéaire multiple sont également présentés dans le tableau à titre de comparaison. Pour le système hybride $S2$, la fonction $max(d) - d$ est utilisée pour le calcul des similarités entre annonces, et les 20 plus proches voisins sont considérés. Pour l'approche $S3$, c'est la similarité gaussienne qui est choisie, et les 40 plus proches voisins sont pris en compte dans l'estimation.

Pour certains sites d'emploi, le nombre d'observations en apprentissage est trop faible

Site d'emploi	Nombre d'annonces	Ecart-type	Recommandeur moyen	Régression linéaire	Régression PLS	Système hybride $S2$	Système hybride $S3$
Généraliste 1	5698	32,8	20,2	17,0	17,1	17,8	16,1
Généraliste 2	3300	30,0	16,3	14,1	13,7	12,9	13,2
Généraliste 3	6795	31,3	18,8	16,0	15,7	14,1	15,3
Généraliste 4	4937	34,7	18,4	14,9	16,4	13,6	14,8
Généraliste 5	133	16,0	9,0	13,2	8,8	8,3	8,5
Généraliste 6	2803	10,0	5,1	4,3	4,7	3,6	4,3
Généraliste 7	5580	17,4	11,1	8,6	9,0	8,1	8,3
Généraliste 8	908	8,8	4,7	4,9	4,5	4,1	4,3
Généraliste 9	881	3,4	2,3	2,1	1,9	2,0	2,0
Généraliste 10	128	10,4	8,5	8,5*	6,6	9,2	6,6
Généraliste 11	370	12,4	7,9	6,1	7,0	7,4	6,7
Généraliste 12	272	15,2	11,1	11,1	10,1	12,3	9,7
Généraliste 13	349	22,9	13,2	14,3	13,0	11,8	12,5
Généraliste 14	84	7,8	7,0	7,0*	3,7	4,3	4,2
Généraliste 15	122	4,3	3,1	2,9	2,9	2,8	2,6
Spécialisé 1	432	39,5	25,5	20,2	16,8	17,4	16,6
Spécialisé 2	1250	14,9	9,1	6,9	7,4	6,6	6,7
Spécialisé 3	476	63,4	52,3	48,0	39,1	45,5	36,7
Spécialisé 4	830	43,4	27,4	25,9	24,3	22,8	23,2
Spécialisé 5	2303	19,7	10,5	8,3	10,2	7,6	9,0
Spécialisé 6	93	18,4	8,7	8,7*	3,7	12,3	3,3
Spécialisé 7	201	8,4	5,8	4,6	4,8	5,0	4,8
Spécialisé 8	2461	10,1	4,9	4,1	4,4	4,3	4,1
Spécialisé 9	79	23,1	14,6	14,6*	12,2	13,1	12,7
Spécialisé 10	113	47,5	28,8	28,8*	25,5	27,8	24,0
Spécialisé 11	73	3,4	2,9	2,9*	2,8	2,6	2,7
Spécialisé 12	129	8,9	5,6	5,1	5,6	4,0	4,1
Spécialisé 13	654	2,8	1,9	1,6	1,6	1,6	1,6
Spécialisé 14	249	21,4	18,9	17,0	14,7	7,5	6,3
Spécialisé 15	417	3,1	2,3	2,3	2,3	2,4	2,3
Total	**42120**	**28,1**	**12,5**	**11,5**	**10,3**	**10,4**	**9,6**

TABLE 5.4 – MAE et $\overline{MAE}$ obtenus avec les différentes approches pour les sites d'emploi étudiés (* : estimation par le recommandeur moyen)

relativement au nombre de variables explicatives pour permettre l'estimation du modèle de régression linéaire. Ayant déjà opéré une première sélection des variables descriptives candidates au modèle, nous ne souhaitons pas réduire encore leur nombre car ce cas apparaît sur une minorité de sites. Lorsque le cas se présente, le recommandeur moyen est appliqué pour le site en question.

Comme vu précédemment, le système hybride $S3$ permet d'obtenir les meilleurs résultats au global grâce à une réduction importante de l'erreur lorsque les autres approches montrent des faiblesses (exemple avec les sites *Spécialisé 3*, *Spécialisé 6* et *Généraliste 1*).

Afin d'illustrer les résultats obtenus sur l'échantillon test avec le système de recommandation basé sur une similarité supervisée, nous prenons l'exemple du site d'emploi *Généraliste 4*. Les valeurs réelles du rendement journalier (nombre moyen de CV reçus par jour) sont comparées aux valeurs prédites à l'aide d'un lissage sur l'espace engendré par les deux premières composantes PLS (voir figure 5.10). Le lissage des valeurs prédites est effectué à l'aide d'un noyau gaussien.

Remarque 8 *Dans le contexte réel comme dans notre base d'apprentissage, une annonce peut être présente simultanément sur plusieurs sites d'emploi. Dans nos travaux, l'objectif est de fournir une estimation de la performance attendue sur un site d'emploi, indépendamment des autres sites utilisés. Ce choix se justifie par le fait que la prédiction est fournie au recruteur pour chaque site d'emploi avant de connaître l'ensemble des sites sélectionnés. Nous ne connaissons donc ni le nombre ni les sites sur lesquels l'annonce sera présente lorsque nous délivrons la prédiction.*

5.6 Synthèse

Nous avons introduit dans ce chapitre différentes approches permettant de prédire le rendement des annonces sur les sites d'emploi. Nous avons proposé une application innovante des systèmes de recommandation en introduisant deux systèmes hybrides adaptés à la problématique du nouvel item. Dans le principal cas étudié, les systèmes sont uniquement basés sur les caractéristiques des annonces (pas de "notes" attribuées pour l'item étudié). Le système hybride basé sur une similarité supervisée dépasse les autres approches en com-

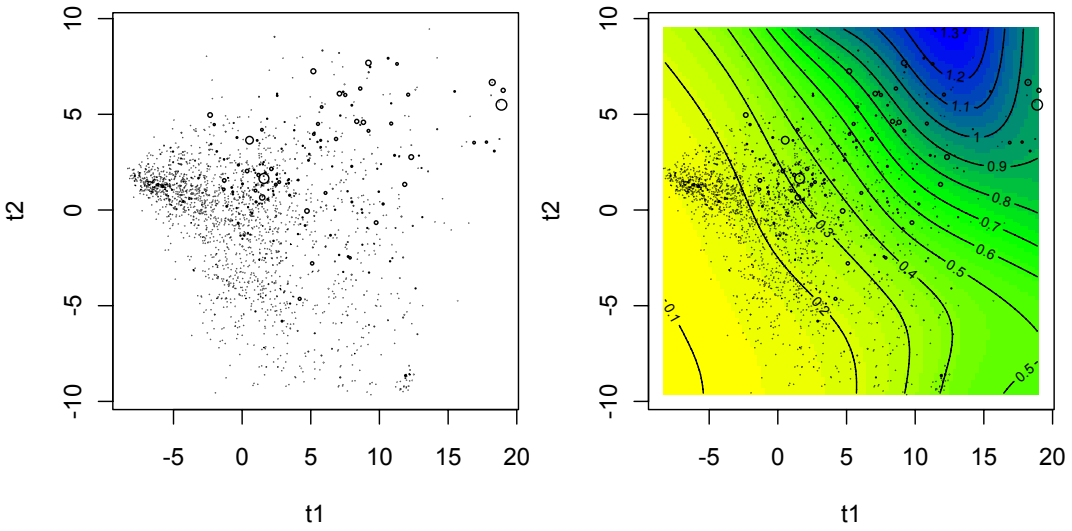

FIGURE 5.10 – *Généraliste 4* : représentation des résultats obtenus pour l'échantillon de test sur le plan engendré par les deux premières composantes PLS. Figure de gauche : rendement journalier réel (taille du cercle proportionnelle à la valeur). Figure de droite : lissage à partir des valeurs prédites par $S3$, courbes de niveau associées et rendement journalier réel.

binant la compréhension de l'impact des facteurs explicatifs sur le rendement du job board et une recommandation collaborative. Le système permet également l'ajout de variables structurées pour améliorer la qualité de prédiction et assure la stabilité de l'algorithme avec l'augmentation de la taille du voisinage. Dans cette approche, les représentations TF et LSA fournissent des résultats similaires, de même que les fonctions de similarité $max(d)-d$, gaussienne et exponentielle. En cas de rediffusion, l'apprentissage sur un vecteur de descripteurs enrichi des connaissances acquises lors de la première diffusion permet d'améliorer la performance de l'algorithme prédictif.

Chapitre 6

Applications pour Multiposting.fr

Ce chapitre est dédié à la présentation de la mise en application des résultats obtenus pour la société Multiposting.fr. Nous présentons dans une première section la base de données initiale ainsi que nos apports ayant permis l'enrichissement des connaissances stockées. Dans une deuxième section, nous illustrons l'impact des facteurs explicatifs introduits à travers leurs contributions et interprétons l'effet de certains d'entre eux pour des sites en particulier. Enfin, nous parcourons dans une troisième section le processus classique de diffusion d'une offre d'emploi et introduisons les nouvelles fonctionnalités accompagnées d'un nouvel enchaînement des étapes de la diffusion.

6.1 Les données

6.1.1 Présentation de la base de données

6.1.1.1 Description du mode d'utilisation de l'outil et des supports

La société Multiposting.fr propose différentes offres adaptées aux besoins des recruteurs. En plus de l'offre de multidiffusion classique (diffusion sur les sites généralistes et spécialisés, les écoles et associations d'anciens), Multiposting.fr propose :

- la multidiffusion sur des sites d'emploi dédiés au handicap et à la diversité ;
- la diffusion des offres d'emploi sur la page Facebook de l'entreprise.

Ces deux dernières offres sont très spécifiques et sortent du cadre de notre problématique car impliquent l'automatisation du processus de diffusion des annonces, qui ne font alors l'objet d'aucune recommandation.

Une annonce multidiffusée est composée de plusieurs postings sur différents job boards. Nous appellerons également les annonces multidiffusées des “annonces Multiposting”. Au 1er septembre 2011, les annonces (multidiffusées) de la base de données se répartissent comme présenté dans la figure 6.1. Toutes les statistiques qui suivent feront référence au même périmètre de données, et plus précisément au périmètre des annonces multidiffusées dans le cadre de l’offre classique Multiposting.

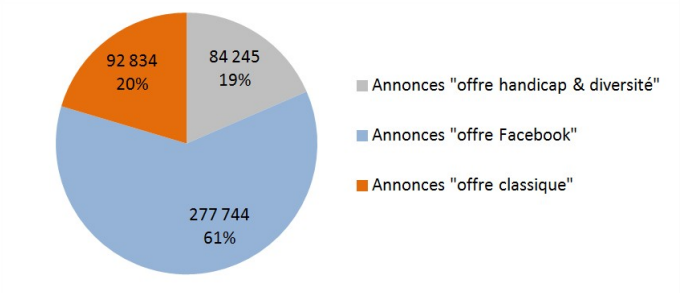

FIGURE 6.1 – Répartition des annonces Multiposting entre les différentes offres proposées

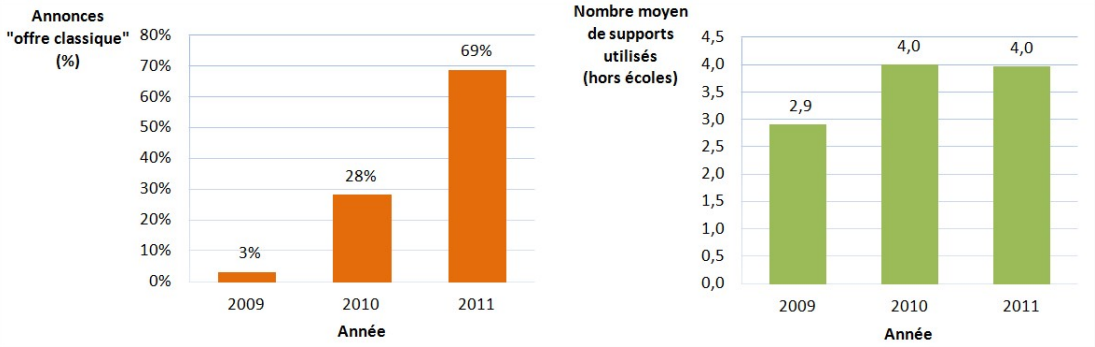

FIGURE 6.2 – Évolution du nombre d’annonces multidiffusées (offre classique) et du nombre moyen de supports utilisés (hors écoles et associations d’anciens)

Le nombre d’annonces multidiffusées via l’outil augmente fortement d’année en année. En effet, ce sont plus de 64 000 annonces qui sont multidiffusées en 2011, alors que moins de 3 000 offres étaient multidiffusées en 2009. Le nombre moyen de supports différents utilisés (hormis les écoles et associations d’anciens) pour une offre est quant à lui resté stable au cours du temps : 3 supports en moyenne en 2009, et 4 supports en 2010 et 2011 (cf. figure 6.2).

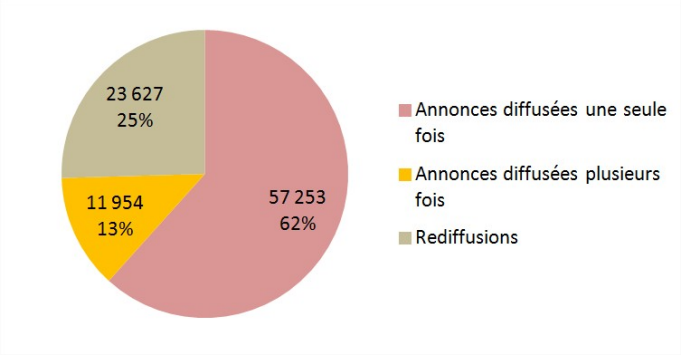

FIGURE 6.3 – Répartition des annonces Multiposting selon le nombre de diffusions

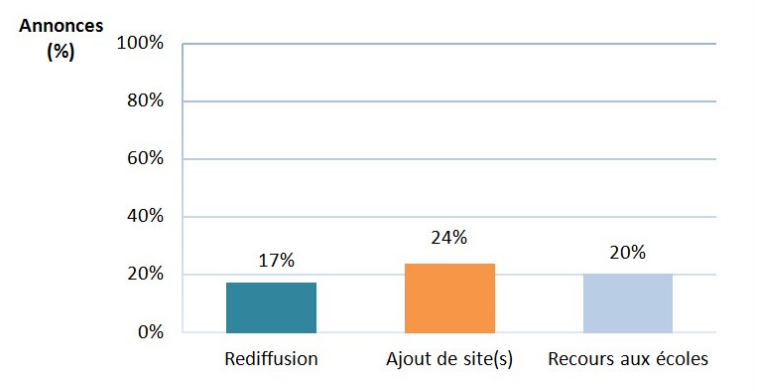

FIGURE 6.4 – Proportion d'annonces avec rediffusion, ajout de site(s), et recours aux écoles

Comme cela a été évoqué dans le chapitre 5, l'outil permet au recruteur de diffuser à nouveau une annonce déjà postée sur un site d'emploi par le passé. À ce moment-là, le contenu de l'annonce diffusée est en tout point identique à celui de l'annonce publiée la première fois. Ce type d'annonce est appelé une "rediffusion". Sur le périmètre étudié, 62% des annonces sont des annonces ayant été diffusées une unique fois, 13% sont des annonces qui ont par la suite été rediffusées et 25% sont des rediffusions (figure 6.3). Cela revient à dire que 17% des annonces multidiffusées ont été rediffusées par la suite. De plus, 24% des annonces sont sujettes à l'ajout de site(s) [1]. Une annonce en cours diffusée sur des sites supplémentaires peut alors être considérée comme une rediffusion. Enfin, les recruteurs ont recours à l'utilisation d'écoles pour 20% des annonces multidiffusées (figure 6.4).

La figure 6.5 présente le nombre moyen de supports utilisés, par type de support, selon que l'annonce fait recours aux écoles ou non. Une annonce sans recours aux écoles fait appel en moyenne à 2 sites généralistes et 2 sites spécialisés. Une annonce avec recours aux écoles fait appel à 2 sites généralistes et 3 sites spécialisés en moyenne, avec une diversification des supports s'opérant davantage sur les sites gratuits. Un panier "Écoles" moyen est composé de 37 écoles.

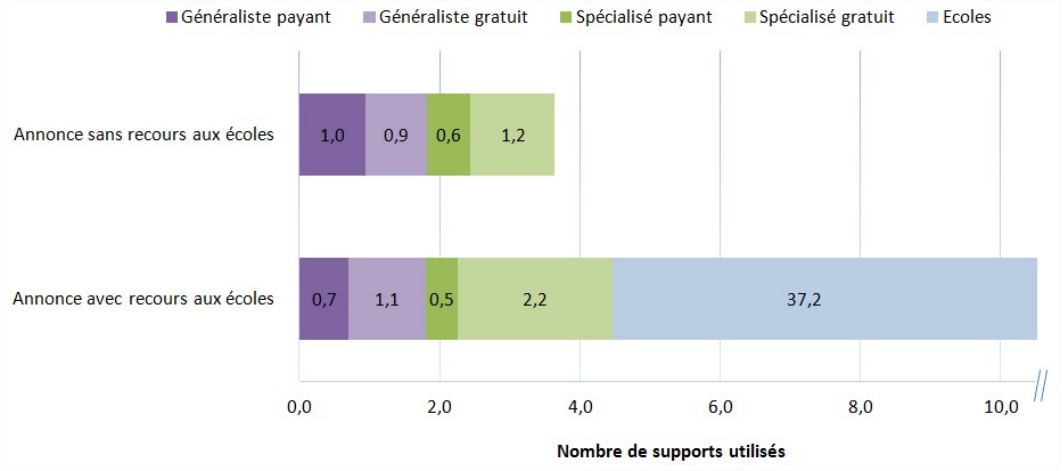

FIGURE 6.5 – Nombre moyen de supports utilisés pour une annonce multidiffusée (avec ou sans recours aux écoles)

1. L'outil Multiposting offre la possibilité d'ajouter des supports de diffusion à une annonce déjà en cours.

6.1.1.2 Description du type d'annonces multidiffusées

Nous décrivons maintenant les annonces qui sont multidiffusées via l'outil. Afin d'être en accord avec les expérimentations menées dans le chapitre 5, nous nous focalisons sur la description des annonces postées avec un mode de candidature par e-mail (parmi les annonces multidiffusées, 29% ont un mode de candidature par e-mail, et 71% par URL).

Les recruteurs peuvent être de cinq types différents. Les volumes d'annonces associés à chaque type de recruteur sont indiqués dans la figure 6.6. La figure présente également la répartition des annonces entre les différents secteurs d'activité des entreprises (hors cabinets de recrutement, agences d'intérim et SSII).

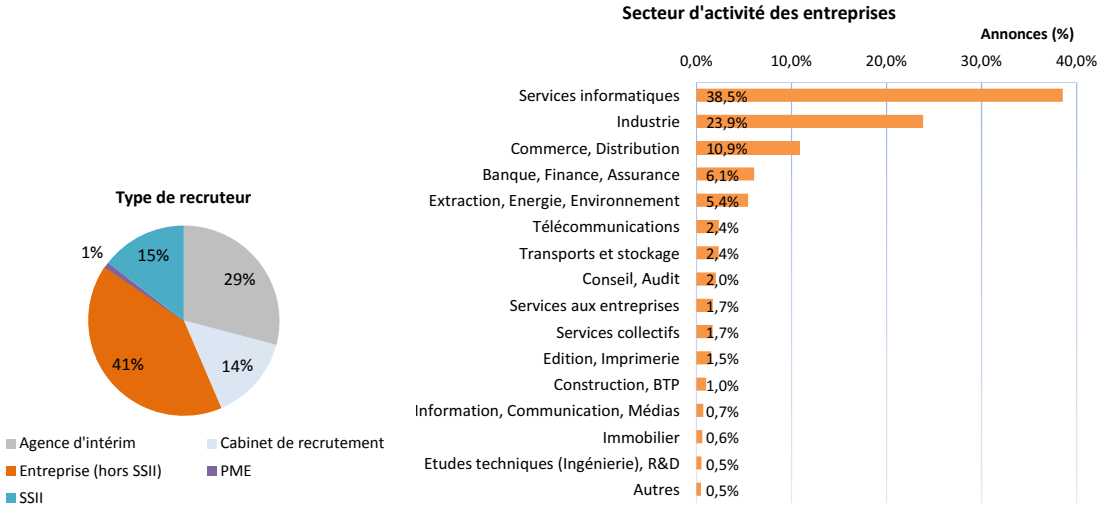

FIGURE 6.6 – Répartition des annonces selon le type de recruteur et le secteur d'activité des entreprises

Le type d'annonces multidiffusées est décrit à travers le type de contrat, le niveau d'études requis, le niveau d'expérience requis et la région où le poste est localisé. Les répartitions des annonces sont présentées dans la figure 6.7.

6.1.2 Enrichissement de la base de données existante

Les données structurées disponibles en base de données ne permettent pas de couvrir l'ensemble des facteurs proposés dans le chapitre 3 et d'assurer une bonne qualité de prédiction de la performance des offres. Pour améliorer la description des offres d'emploi, nous enri-

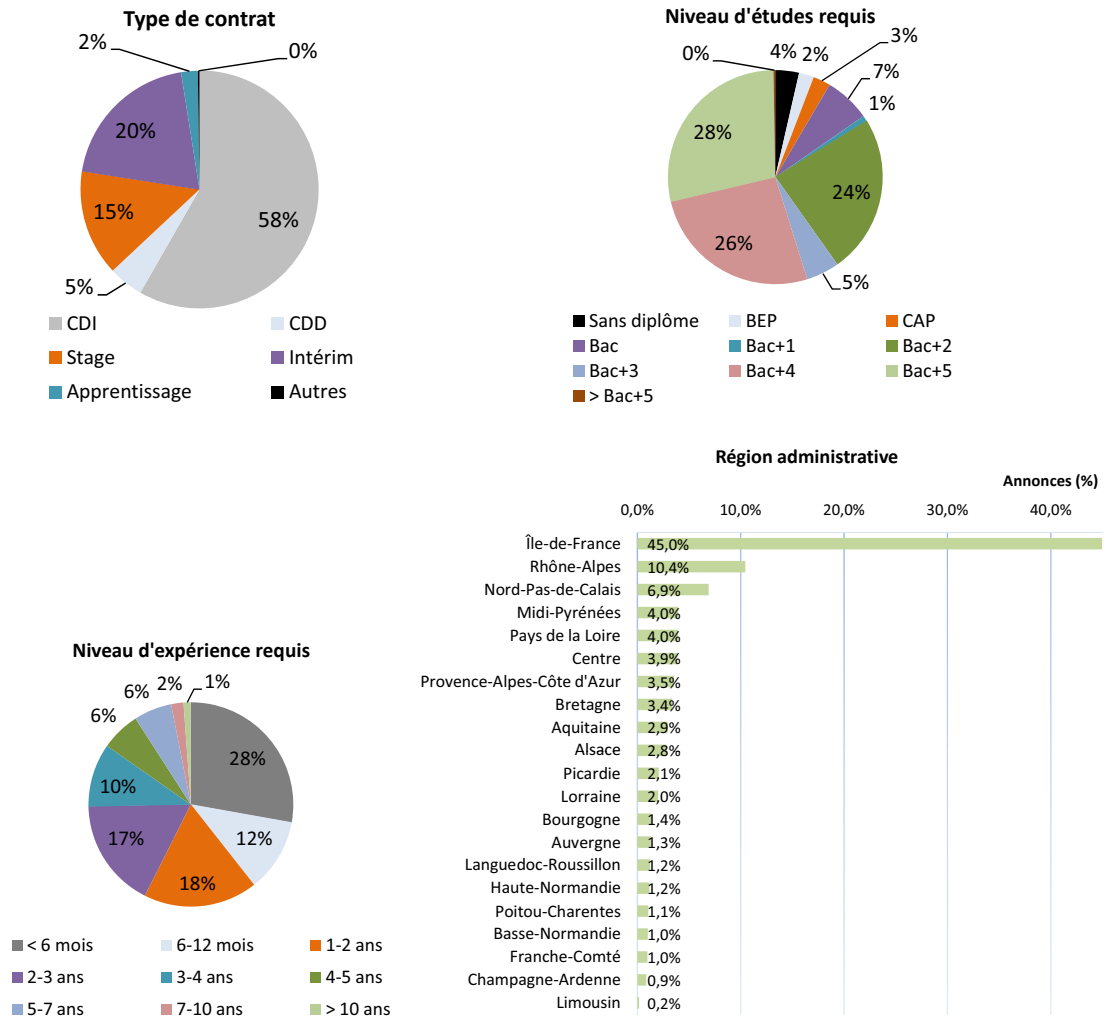

FIGURE 6.7 – Répartition des annonces selon le type de contrat, le niveau d'études requis, le niveau d'expérience requis et la région administrative

chissons la base de données en créant de nouvelles variables à partir des données existantes et en faisant appel à des données externes.

6.1.2.1 Enrichissement de la description des offres d'emploi

En addition du type de contrat, du niveau d'études requis, du niveau d'expérience requis et de la localisation du poste (région administrative identifiée à partir du code postal), nous créons un ensemble de variables descriptives des offres d'emploi à partir de la description du poste et du profil recherché. Les deux parties sont traitées de manière concaténée car ces dernières ne sont pas identifiées pour une partie importante des annonces.

À partir du texte des annonces, nous avons pu identifier la fonction et le métier associés à chaque annonce comme présenté dans la section 4.3. Nous ajoutons à la description un ensemble de mots-clés candidats et des composantes sémantiques obtenues grâce à des méthodes de fouille de données textuelles et de recherche d'information. Enfin, un ensemble d'indicateurs de stylistique sont extraits du texte des annonces.

Mots-clés et composantes sémantiques. Les expérimentations menées dans le chapitre 5 ont mis en évidence une qualité de prédiction équivalente si le texte est représenté en fréquences de termes ou avec les coordonnées issues de l'analyse sémantique latente (LSA). Nous choisissons la représentation LSA (500 premières dimensions) et la complétons avec les fréquences de termes associées aux mots-clés présents dans le descriptif des missions. Nous ajoutons à cette description l'ensemble des indicatrices codant la présence ou l'absence de mots-clés dans le titre. Les variables de fréquences et indicatrices à conserver sont sélectionnées à partir de la méthodologie décrite dans la section 4.4.

Indicateurs de la stylistique. Afin de prendre en compte le style rédactionnel des offres d'emploi, nous créons un ensemble d'indicateurs à partir du texte comme évoqué dans le chapitre 3. Pour calculer ces différents indicateurs, le pré-traitement du texte (identification et étiquetage grammatical des mots) est réalisé à l'aide de l'algorithme de Schmid [1994] [2]. Les indicateurs calculés sont les suivants :

2. Outil *TreeTagger* (`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`)

- longueur du titre (nombre de caractères) ;
- nombre de mots dans le titre ;
- proportion (en longueur) de la partie “descriptif société” ;
- proportion (en longueur) de la partie “missions + profil recherché” ;

Les indicateurs qui suivent sont déclinés pour la partie “descriptif société” et la partie “missions et profil recherché” :

- longueur du texte (égale au nombre total de caractères) ;
- nombre de phrases (une phrase est une suite de caractères délimitée en début et en fin par les signes de ponctuation suivants : “ .!? ...”) ;
- nombre total de mots ;
- nombre moyen de mots par phrase ;
- nombre moyen de virgules par mot ;
- nombre moyen de parenthèses par phrase ;
- nombre de points d’exclamation (effectif absolu car très faible en pratique) ;
- nombre moyen de caractères par mot ;
- proportion de verbes (par rapport au nombre total de mots) ;
- proportion de noms communs (par rapport au nombre total de mots) ;
- nombre de formes [3] divisé par le nombre total de mots ;
- nombre d’hapax [4] divisé par le nombre de formes.

6.1.2.2 Enrichissement de la description des entreprises / des recruteurs

Dans la base de données actuelle, nous ne disposons pas d’informations sur les recruteurs (les clients de la solution Multiposting). Seuls les identifiants de comptes peuvent être utilisés pour caractériser un recruteur. Nous souhaitons enrichir la description des recruteurs. Par ailleurs, plusieurs comptes peuvent être associés à une même entreprise (par exemple, un compte pour l’offre classique et un compte pour l’offre handicap). Comme nous voulons prendre en compte l’image de l’entreprise, nous avons créé un identifiant “recruteur” de niveau hierarchique supérieur afin de caractériser une entreprise (à un identifiant recruteur peut être associé un ou plusieurs identifiants de comptes). Nous avons également complété

3. Le nombre de formes indique le nombre de mots distincts dans le texte.
4. Le nombre d’hapax indique le nombre de formes ayant une seule occurrence dans le texte.

la description des entreprises à l'aide des variables suivantes (étiquetage manuel) :

- type de recruteur (entreprise, cabinet de recrutement, agence d'intérim, etc.) ;
- secteur d'activité de l'entreprise ;
- indicatrice de recruteur, pour les recruteurs ayant multidiffusé plus de 50 annonces avec la solution Multiposting. Sur notre échantillon d'étude, cela concerne 39 recruteurs sur un total de 265.

6.1.2.3 Intégration de données de conjoncture

Pour expliquer les fluctuations conjoncturelles de la performance, nous enrichissons la description des annonces à l'aide des variables suivantes :

- année de publication de l'annonce ;
- mois de publication de l'annonce ;
- indice *Monster* de l'emploi en ligne (indice base 100, niveau France) ;
- indice *Apec* de diffusion des offres d'emploi cadre sur Internet (indice base 100, niveau France).

Nous ajoutons à ces variables une variable qualitative croisant le mois de diffusion de l'annonce et le type d'emploi (stage/apprentissage ou autre emploi), car les volumes de candidats aux stages et contrats d'apprentissage sont associés à un calendrier particulier.

Nous prenons également en compte des données démographiques pour caractériser la conjoncture d'un point de vue géographique. Pour cela, nous associons à la localisation du poste des variables descriptives calculées au niveau code postal, département, et zone d'emploi. Les données utilisées sont fournies par l'INSEE [5] au niveau commune. Nous utilisons une table de correspondances commune-code postal (un code postal peut être associé à une ou plusieurs communes), et agrégeons les données pour calculer les valeurs au niveau code postal, département et zone d'emploi. Les variables suivantes sont utilisées :

- population totale (en 2008) ;
- nombre de personnes actives de 15 à 64 ans (en 2007) ;
- nombre de chômeurs de 15 à 64 ans au sens du recensement (en 2007) ;

5. `http://www.insee.fr/fr/themes/theme.asp?theme=3`

- demandeurs d'emploi de catégorie A au sens du BIT [6] (en 2010) ;
- demandeurs d'emploi de catégories A, B, C au sens du BIT (en 2010).

6.2 Impact des facteurs explicatifs et interprétation

Pour la prédiction des performances sur les sites d'emploi, nous avons recours à l'algorithme prédictif retenu à l'issue des expérimentations du chapitre 5. L'ensemble des variables citées dans les sections 6.1.1.2 et 6.1.2 sont utilisées au sein du modèle.

6.2.1 Contribution des facteurs à la prédiction de la performance

Nous souhaitons maintenant identifier les facteurs et groupes de facteurs explicatifs ayant les plus fortes contributions pour la prédiction de la performance des annonces. Comme un modèle de prédiction est réalisé pour chaque site d'emploi, les contributions sont différentes selon le site et ses caractéristiques. Nous avons identifié 7 groupes de variables explicatives dont nous souhaitons étudier l'importance :

- les variables descriptives du type de poste (*poste*) ;
- les coordonnées issues de l'analyse sémantique latente (*coord. LSA*) ;
- les indicatrices de mots-clés dans le titre (*mots-clés titre*) ;
- les fréquences de mots-clés dans le descriptif (*mots-clés desc.*) ;
- les indicateurs de stylistique (*style*) ;
- les variables descriptives du recruteur (*recruteur*) ;
- les variables descriptives de la conjoncture (*conjoncture*).

Nous étudions également l'impact du logarithme de la durée de diffusion sur le rendement journalier des annonces. Nous rappelons que cette variable est introduite dans le modèle afin de prendre en compte la décroissance du rendement journalier au cours de la vie d'une annonce publiée sur un site (voir section 5.1.2.2).

Soient VIP_{x_j} le VIP associé à la variable x_j, p le nombre total de variables explicatives et X_g l'ensemble des variables du groupe g. La contribution C_g du groupe de variables g à la

6. Bureau international du travail

prédiction de la performance est mesurée ainsi :

$$C_g = \sum_{x_j \in X_g} \frac{VIP_{x_j}^2}{p}$$

Les contributions sont telles que : $\sum_{g=1,\dots,7} C_g = 1$. Le tableau 6.1 présente les contributions de ces groupes de variables pour quatre sites d'emploi servant d'exemples. Il présente également l'indicateur C-75%, troisième quartile de la contribution sur l'ensemble des sites d'emploi. Cette contribution étant fortement influencée par le nombre de variables du groupe, nous nous intéressons au score de "contribution moyenne", défini par :

$$\overline{C}_g = \sum_{x_j \in X_g} \frac{VIP_{x_j}^2}{p_g}$$

où p_g est le nombre de variables du groupe g. Ce score permet d'avoir une estimation de l'impact moyen des variables du groupe sur la performance des offres. Comme précédemment, le troisième quartile du score de contribution moyenne est présenté.

Groupe	Nombre de variables	Gén. 1	Gén. 4	Spé. 5	Spé. 4	C-75%
poste	156	8.2(0.8)	10.3(1.0)	7.9(0.7)	5.7(0.5)	7.3(0.6)
coord. LSA	500	13.9(0.4)	6.2(0.2)	19.2(0.5)	36.5(1.0)	33.4(0.9)
mots-clés titre	51*	8.1(1.4)	6.3(1.3)	8.4(1.3)	8.7(1.9)	7.3(1.7)
mots-clés desc.	464*	39.4(0.9)	49.6(1.2)	49.7(1.4)	41.1(1.1)	53.2(1.5)
style	24	2.8(1.8)	5.4(3.3)	2.1(1.2)	1.6(0.9)	3.2(1.7)
recruteur	65	5.3(1.2)	5.6(1.3)	5.1(1.1)	2.6(0.5)	5.3(1.1)
conjoncture	45	21.7(7.2)	14.3(4.6)	5.6(1.7)	3.8(1.2)	7.6(2.4)
log(durée)	1	0.7(10.5)	2.2(32.3)	2.0(27.2)	0.03(0.5)	0.9(13.1)
Total	**1306**	**100(1)**	**100(1)**	**100(1)**	**100(1)**	

TABLE 6.1 – Contributions (%) des groupes de variables à la prédiction de la performance, et score de contribution moyenne indiqué entre parenthèses (* : nombre moyen de variables retenues)

Nous constatons que les contributions des facteurs peuvent être très différentes d'un site à l'autre, ce qui justifie l'intérêt d'un modèle de prédiction spécifique à chaque site d'emploi. De part leur grand nombre, les variables associées aux mots-clés représentent une forte contribution. Le score moyen de contribution permet de relativiser leur impact, qui reste toutefois important avec un score moyen supérieur à 1.5 pour 25% des sites d'emploi. Le logarithme de la durée de diffusion a une grande importance dans le modèle de prédiction, de même que les indicateurs de conjoncture introduits. D'un point de vue global, les variables

descriptives associées au poste et les coordonnées de l'analyse sémantique latente sont les facteurs qui ont l'impact le moins fort sur la performance des offres. Cependant, au sein de ces groupes, certaines variables présentent un fort impact qui justifie leur introduction dans le modèle de prédiction.

6.2.2 Interprétation de l'impact des facteurs

Nous souhaitons obtenir des éléments d'interprétation de l'impact des facteurs sur la performance des offres d'emploi. Pour cela, nous identifions les facteurs les plus importants au sein de chaque groupe. Nous nous basons sur l'indicateur VIP-75% (troisième quartile du VIP sur l'ensemble des sites étudiés) pour mesurer l'importance d'un facteur explicatif et illustrons les résultats avec les cinq premiers facteurs de chaque groupe. Le tableau 6.2 présente, pour les facteurs ainsi identifiés, les valeurs du VIP et les signes des coefficients associés pour deux sites d'emploi utilisés comme exemples. L'impact des facteurs issus des coordonnées de l'analyse sémantique latente n'est pas étudié car ceux-ci sont difficilement interprétables. Par ailleurs, l'impact des variables associées aux mots-clés a déjà été illustré dans la section 4.4.2.1.

Nous proposons des interprétations pour quelques uns des résultats présentés dans le tableau 6.2. Dans la mesure où nous ne présentons pas l'intégralité des résultats obtenus, nous ne pouvons pas tirer de conclusions générales sur l'impact des différentes modalités des variables qualitatives.

Parmi les indicateurs de conjoncture, ce sont les indices Monster et APEC (indices de diffusion d'offres d'emploi et d'offres d'emploi cadre sur Internet) qui ont le plus d'impact sur la performance des annonces. Cet impact est négatif, et retrouvé sur les deux sites illustrés. En effet, plus le volume d'offres sur Internet est important plus la concurrence est forte entre les recruteurs, et plus le nombre de candidatures par offre est faible pour un volume de candidats potentiels donné. Parmi les variables démographiques, celles associées à la zone d'emploi semblent être les plus pertinentes. Le volume de chômeurs et la taille de la population active ont une forte influence sur le rendement des annonces. En effet, plus le volume de main d'œuvre disponible dans la zone d'emploi est élevé, plus le nombre de candidats potentiels par offre est élevé pour un volume d'offres donné.

Groupe	Variable	VIP-75%	Généraliste		Spé. Informatique	
			VIP	signe	VIP	signe
conjoncture	indice Monster	2.5	2.7	-	2.0	-
	indice APEC	2.4	2.8	-	1.7	-
	année 2011	2.4	2.4	-	1.9	-
	chômeurs ABC 2010 (ZE)	2.2	4.3	+	1.4	+
	pop. active 2007 (ZE)	2.2	4.3	+	1.6	+
poste	fct. Serv. administratifs	2.2	1.4	+	0.4	-
	fct. Marketing	2.1	2.9	+	0.1	+
	fct. Informatique	2.0	2.0	-	1.0	+
	région IDF	2.0	3.8	+	2.6	+
	Bac+5	1.8	1.0	-	2.2	-
recruteur	type entreprise	2.6	3.4	+	1.2	+
	sect. Commerce	2.2	3.8	+	1.8	+
	type agence d'intérim	1.9	2.1	-	2.2	+
	type SSII	1.9	1.1	+	1.1	-
	recr. P...	1.8	4.2	+	1.8	+
style	nb mots desc. société	1.8	2.2	+	0.8	-
	nb mots desc. missions	1.8	1.4	-	0.7	-
	longueur desc. société	1.7	1.9	+	0.8	-
	nb phrases desc. société	1.6	2.8	+	1.4	+
	proportion desc. société	1.5	2.1	+	0.6	+
	log(durée)	3.6	3.2	-	0.7	+

TABLE 6.2 – Facteurs explicatifs ayant les plus forts impacts, valeurs du VIP et signes des coefficients associés pour deux sites d'emploi

En ce qui concerne le type de poste proposé, nous constatons que les fonctions "Services administratifs", "Marketing" et "Informatique" se distinguent. Sur le site généraliste étudié, le nombre de candidatures par offre est plus faible pour des postes à pourvoir en Informatique que pour les deux autres fonctions citées. Sur le site spécialisé en Informatique, seule la fonction "Informatique" a un léger impact parmi les trois fonctions, et celui-ci est positif.

Nous constatons que pour un poste et des caractéristiques donnés, le recruteur *P...*[7] obtiendra généralement des performances plus élevées qu'un recruteur quelconque (c'est le cas pour les deux sites d'emploi illustrés).

D'une manière générale, les indicateurs liés à la taille du descriptif de la société semblent avoir un impact positif sur la performance des offres sur le site généraliste. Sur le site spécialisé, parmi les cinq indicateurs observés, seul le nombre de phrases dans le descriptif de la société semble avoir un impact positif.

Comme pour les contributions, nous observons des résultats différents selon le site d'emploi. Les différences résident dans l'importance des facteurs introduits et dans le sens de l'impact observé si un impact est effectivement observé.

6.3 Processus de diffusion d'une offre d'emploi : accompagnement et aide à la décision

6.3.1 Processus initial de diffusion d'une offre d'emploi

Dans le processus actuel de diffusion d'une offre d'emploi via Multiposting.fr, les recruteurs ne disposent que de leurs propres connaissances pour choisir les sites d'emploi qu'ils vont utiliser. La première étape est l'étape de choix des supports (sites d'emploi généralistes, sites d'emploi spécialisés et écoles) pour la diffusion de l'offre (figure 6.8).

La deuxième étape est consacrée au remplissage des champs pour la description de l'offre : informations générales sur le recruteur, champs communs remplis pour toutes les annonces et champs spécifiques associés aux sites d'emploi sélectionnés (figure 6.9).

La dernière étape est consacrée à la vérification de la description de l'annonce et à la validation finale avant diffusion.

7. Achat en ligne de matériel.

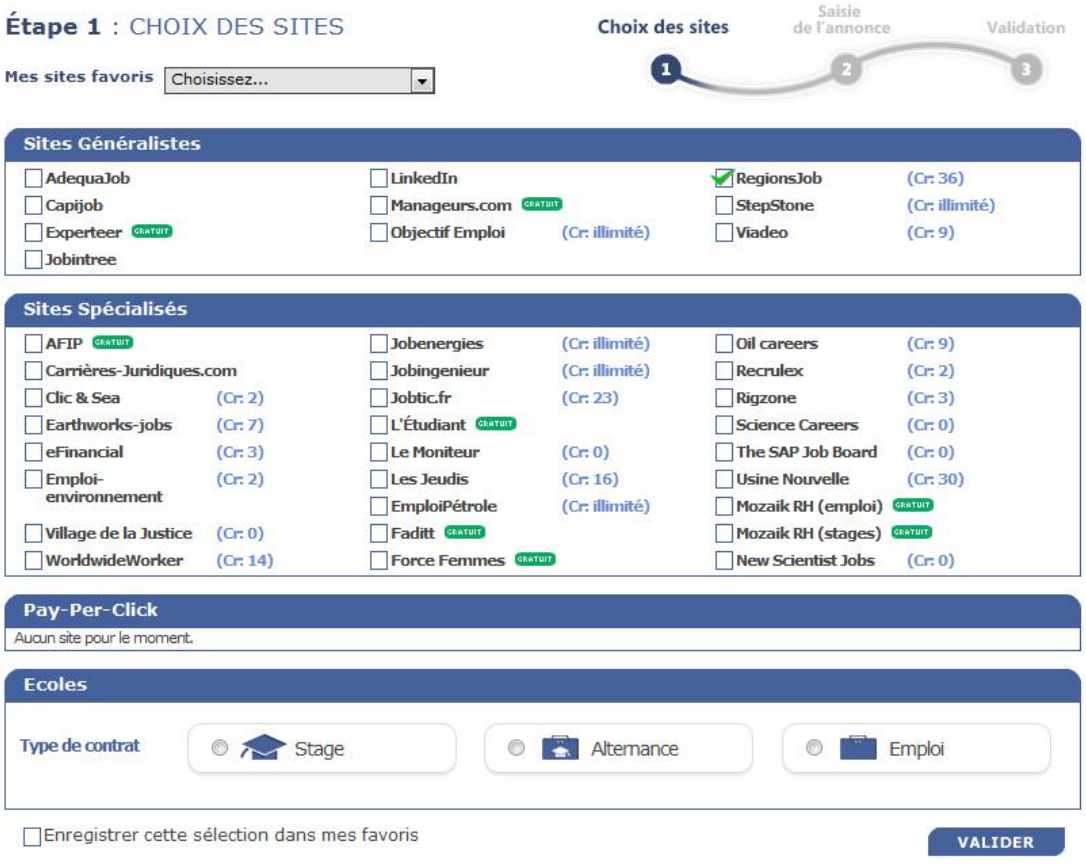

FIGURE 6.8 – Diffusion d'une offre d'emploi via l'interface Multiposting.fr : étape 1

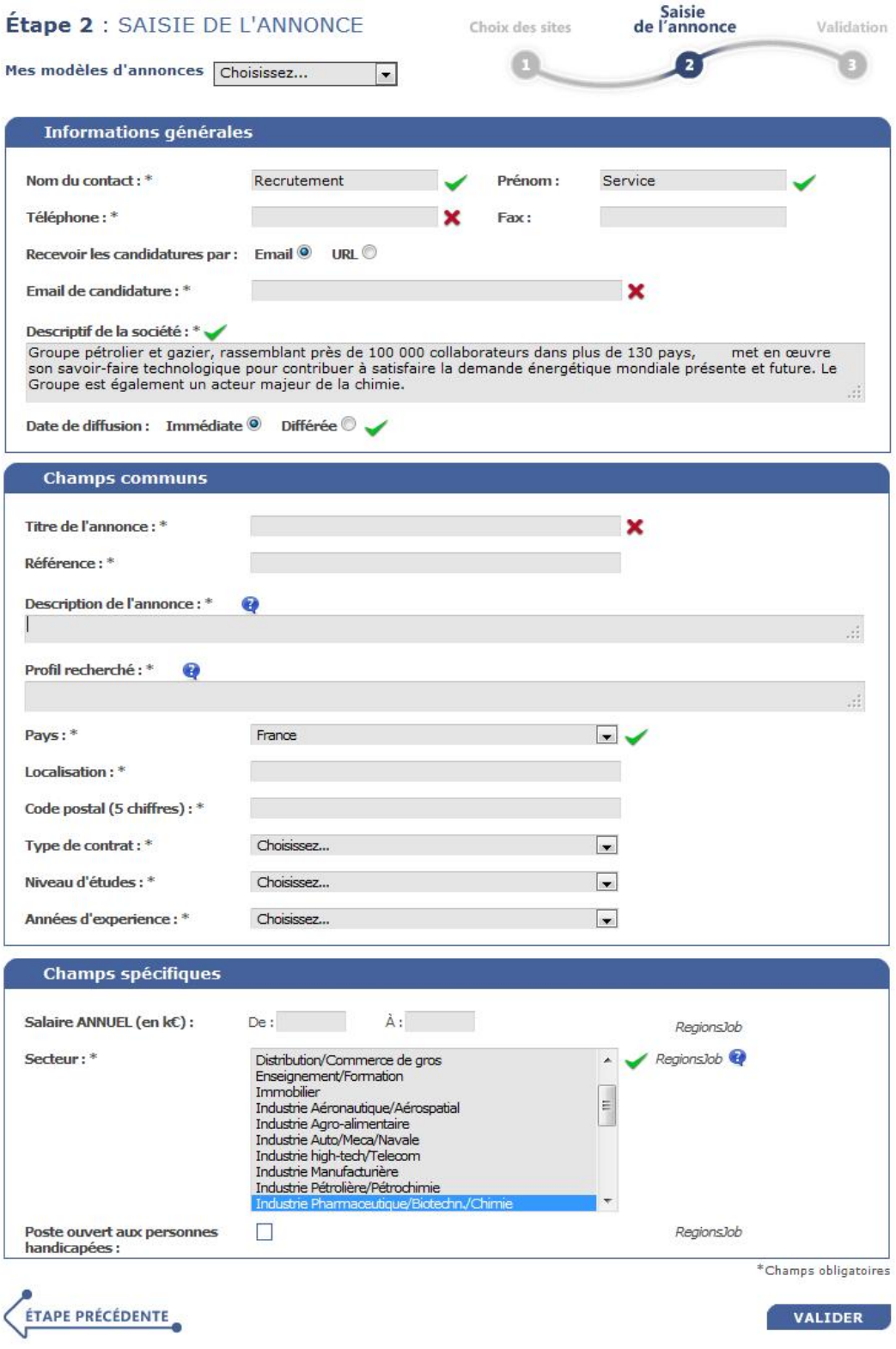

FIGURE 6.9 – Diffusion d'une offre d'emploi via l'interface Multiposting.fr : étape 2

Tel qu'est défini le processus actuel, aucune information ne peut être récoltée sur l'offre d'emploi avant la sélection des supports de diffusion. Le processus doit donc être revisité et les étapes réordonnées pour permettre de délivrer au recruteur des recommandations et statistiques sur la performance attendue.

6.3.2 Nouvelles fonctionnalités

Ces travaux ont pour finalité l'intégration de fonctionnalités complémentaires permettant d'accompagner le recruteur dans ses décisions lors de la diffusion d'une offre d'emploi via l'interface Multiposting.fr. Ces fonctionnalités doivent être utiles pour le recruteur et intégrées en harmonie avec le processus actuel. Des entretiens avec des professionnels du recrutement (en SSII, en agence d'intérim et en entreprise) ont mis en évidence les besoins suivants :

- l'information sur le nombre de candidatures qui peut être attendu sur chaque site d'emploi étant donné le poste à pourvoir ;
- la suggestion de supports gratuits adaptés (en particulier les sites spécialisés et les écoles) que le recruteur n'a pas l'habitude d'utiliser et qu'il ne pense pas à sélectionner ;
- être informé des sites sur lesquels on peut s'attendre à très peu de retours, voire aucun (et ainsi éviter la sélection de sites payants non adaptés au poste à pourvoir) ;
- la suggestion de sites gratuits adaptés qui ne sont pas sur l'interface du recruteur [8].

Il est également ressorti de ces entretiens qu'idéalement, une bonne sensibilisation aux objectifs de l'outil serait bénéfique pour que les recruteurs puissent en tirer tous les avantages.

Suite à ces conclusions, nous avons décidé d'intégrer les fonctionnalités suivantes, répondant à tous les besoins cités plus haut :

- pour chaque site d'emploi, indiquer sur l'interface le nombre de candidatures attendues étant donné les caractéristiques du poste à pourvoir (nombre de CV ou nombre de clics de redirection selon le mode de candidature choisi) ;
- à l'étape du choix des supports de diffusion, placer un bouton permettant la sélection

8. Le nombre de sites d'emploi étant très grand, chaque utilisateur dispose d'une liste de sites déterminée selon les besoins au moment de la création du compte. Ces sites sont par la suite visibles et sélectionnables sur l'interface utilisateur.

automatique d'un panier de sites gratuits (généralistes, spécialisés et écoles) adaptés au poste à pourvoir, avec suggestion possible de sites gratuits qui ne sont pas sur l'interface du recruteur.

Grâce à l'information sur la performance prédite, le recruteur pourra ainsi éviter de sélectionner des sites payants sur lesquels peu de retours sont attendus, et par suite la probabilité de trouver un bon matching très faible.

L'intégration de ces fonctionnalités à l'interface Multiposting.fr est présentée à travers une maquette dans la section 6.3.3.1.

6.3.3 Mise en application des résultats

6.3.3.1 Nouveau processus de diffusion : maquette

Pour pouvoir mettre en place ces fonctionnalités, il est nécessaire de modifier l'ordre des étapes du processus actuel, afin de disposer d'informations sur le poste à pourvoir lors de l'étape de sélection des sites d'emploi. La première étape est donc désormais consacrée aux champs d'informations générales sur le recruteur et aux champs communs des annonces. Ces derniers sont complétés (par rapport à la version classique de l'outil) par des champs secteur et fonction génériques. La deuxième étape est l'étape de sélection des sites. La performance attendue est indiquée à côté de chaque site d'emploi, et un bouton intitulé "Sélection Multiposting" permet la sélection automatique d'un ensemble de supports gratuits. Enfin, la troisième étape est dédiée au remplissage des champs spécifiques aux sites choisis à l'étape précédente. Le nouveau processus comprend une étape supplémentaire par rapport au processus classique de diffusion, toutefois, le nombre d'informations saisies au global reste le même. De plus, les champs communs remplis à la première étape sont exploités afin d'obtenir des associations et de remplir automatiquement une partie des champs spécifiques à l'étape 3. Au final, un gain de temps est obtenu grâce à la sélection rapide des sites gratuits à l'étape 2 et au pré-remplissage des champs spécifiques à l'étape 3.

Le nouveau processus de diffusion est illustré à travers une maquette réalisée en collaboration avec l'équipe Design de Multiposting.fr (figures 6.10, 6.11 et 6.12).

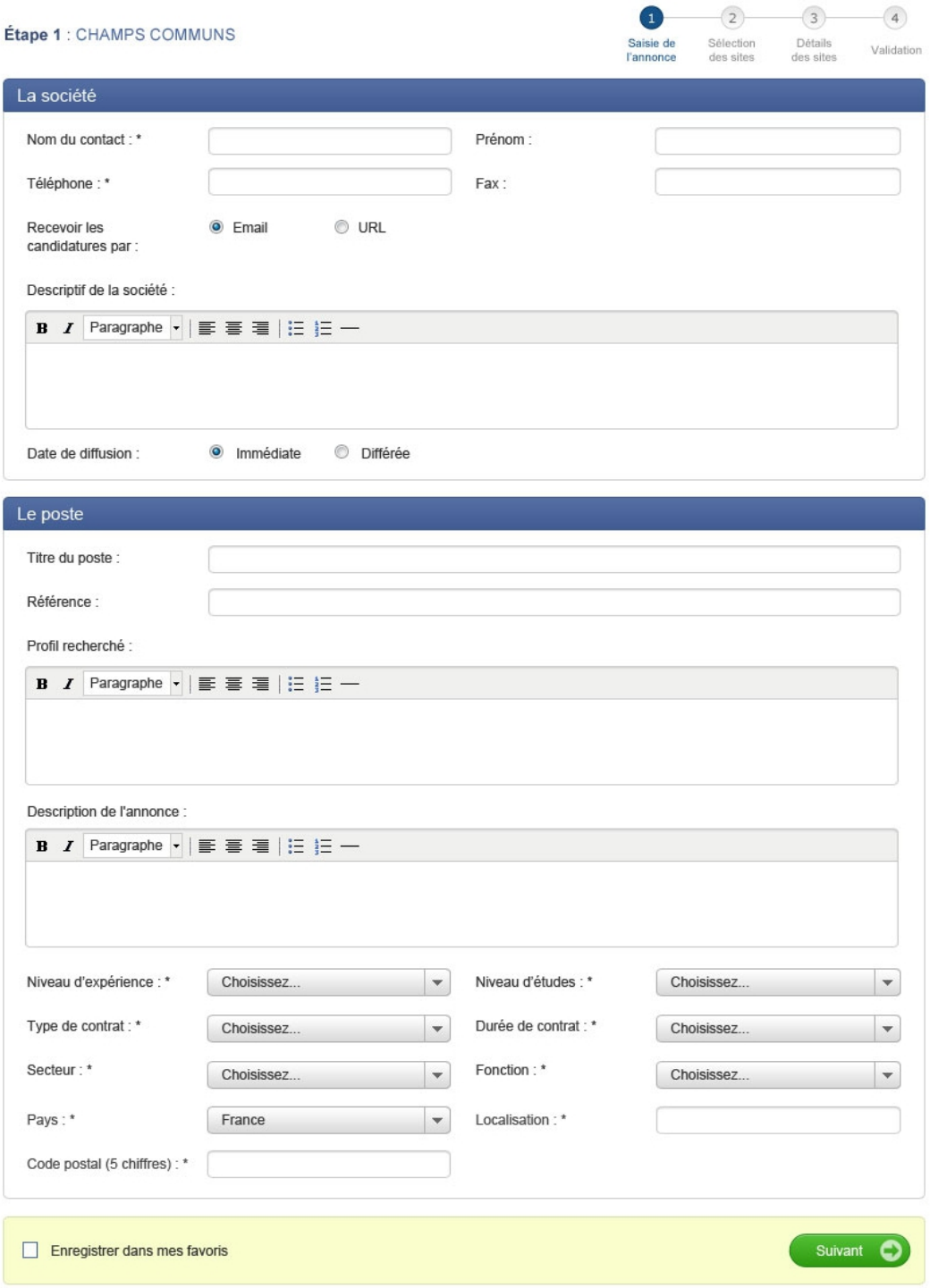

FIGURE 6.10 – Nouveau processus de diffusion d'une offre d'emploi : étape 1

FIGURE 6.11 – Nouveau processus de diffusion d'une offre d'emploi : étape 2

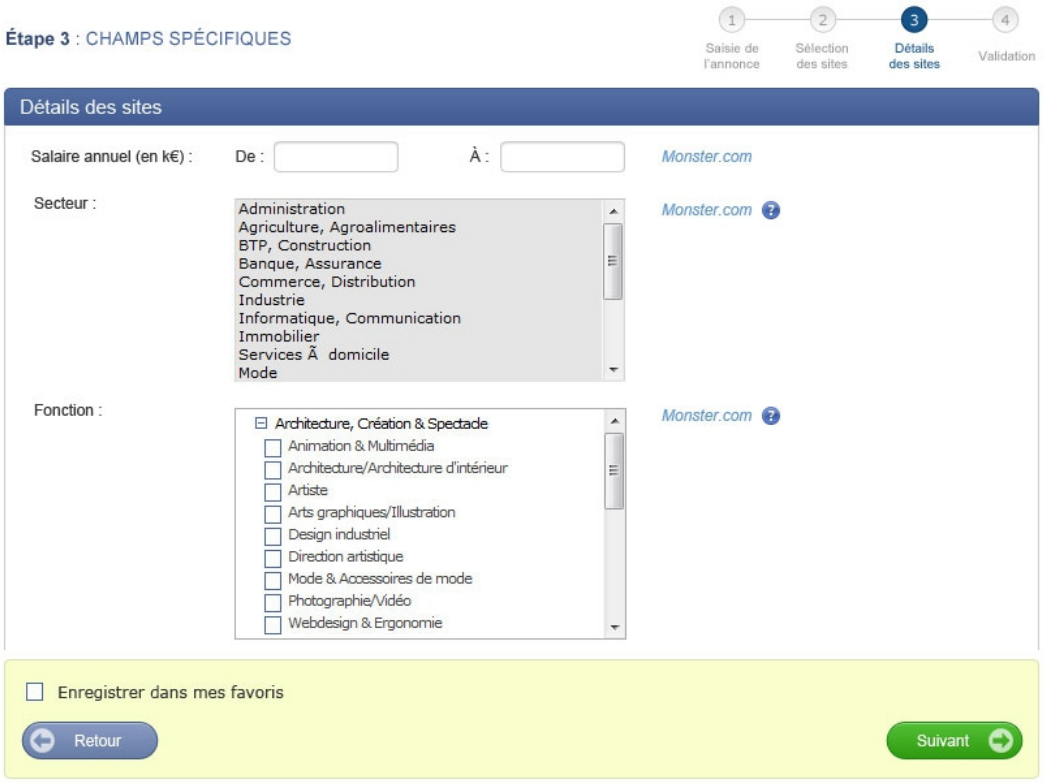

FIGURE 6.12 – Nouveau processus de diffusion d'une offre d'emploi : étape 3

6.3.3.2 Méthodologie de recommandation des sites gratuits

L'étape 2 de la diffusion d'une annonce avec l'outil Multiposting est l'étape de sélection des sites. Nous proposons à l'utilisateur un bouton ("Sélection Multiposting", voir figure 6.11) permettant la sélection automatique d'un ensemble de sites gratuits comme évoqué dans la section 6.3.2. Nous proposons deux méthodologies différentes : une pour les sites généralistes et spécialisés et une spécifique aux écoles.

À l'étape 2 de la diffusion d'une annonce, nous connaissons le nombre de candidatures attendues pour l'ensemble des sites gratuits dont l'historique dépasse 70 annonces diffusées. Dans la mesure où ces sites ne représentent pas de coût pour le recruteur, et où nous ne connaissons pas les taux de qualification des candidatures, notre objectif est alors de proposer un ensemble de sites permettant d'assurer un nombre minimum de candidatures et la diversité des supports utilisés, en accord avec le type de poste proposé.

Recommandation des écoles. Pour faciliter le choix des écoles lors de la diffusion d'une annonce, celles-ci ont été structurées par l'intermédiaire d'un ensemble de variables catégorielles. Les variables permettant de caractériser et classifier les écoles sont les suivantes :

- le type de contrat (stage, alternance ou emploi) ;
- le niveau d'études (Bac/Bac+1, Bac+2, Bac+3, Bac+4/5, $>$ Bac+5) ;
- la localisation (régions administratives) ;
- la formation, décrite par l'intermédiaire de catégories (ex : "Commerce, Gestion, Economie"), sous-catégories (ex : "Management, Gestion", "Comptabilité, Finance, Audit", etc.) et spécialisations (ex : "Administration et gestion des organisations", "Marketing", etc.).

Grâce aux informations obtenues sur l'annonce d'emploi à la première étape de la diffusion, nous pourrons sélectionner un panier d'écoles dont les critères correspondent au poste à diffuser. Les champs communs "type de contrat" et "niveau d'études" permettront une première sélection des écoles adaptées, qui sera affinée à l'aide de la fonction et du métier associés au poste à pourvoir. La fonction et le métier sont des informations spécifiées par le recruteur à la première étape, et correspondent à la nomenclature établie dans la

section 4.3.5.1. Nous établissons une table de correspondances entre la nomenclature métier et les sous-catégories d'écoles. La localisation est spécifiée par défaut à "France entière" (car la localisation de l'école ne donne pas d'indication sur le lieu de travail souhaité des personnes y étudiant, ou y ayant étudié), mais pourra être ajustée par le recruteur. Toutes les écoles éligibles seront automatiquement sélectionnées pour la diffusion.

Recommandation des sites généralistes et spécialisés. Afin d'identifier les supports adaptés au poste à pourvoir, nous annotons manuellement l'ensemble des sites d'emploi à l'aide de trois valeurs, pour l'ensemble des types de contrat, niveaux d'études, niveaux d'expérience, secteurs et métiers existant. Pour chacun de ces critères, les trois valeurs possibles sont :

- 0 (ne pas recommander) : le site d'emploi ne permet pas la diffusion de ce type de poste ;
- 1 (recommandation possible) : la diffusion de ce type de poste est possible mais le critère ne constitue pas une spécialisation du site ;
- 2 (recommandation) : le critère correspond à une spécialisation du site.

Ce travail doit être effectué manuellement car ces informations ne sont pas disponibles en base de données. En plus de permettre la recommandation, nous enrichissons les connaissances sur les sites d'emploi.

Pour chaque site, les valeurs correspondant aux critères spécifiés par le recruteur pour le poste à diffuser sont extraites. Si une des valeurs est égale à 0, le site ne sera pas recommandé. Si une des valeurs est égale à 2, le site sera recommandé. Parmi les recommandés, certains sont disponibles sur l'interface du recruteur, d'autres non. Nous proposons jusqu'à deux nouveaux sites : ceux qui ont les performances attendues les plus élevées dans la mesure où celles-ci dépassent un seuil fixé. Nous recommandons en complément le site dont toutes les valeurs sont égales à 1 pour les critères du poste (il s'agit dans la pratique d'un site généraliste) et dont la performance attendue est la plus élevée. La performance globale attendue est ensuite évaluée sur l'ensemble des sites sélectionnés par cette méthode. Si celle-ci est inférieure à la performance moyenne (sur les sites gratuits sélectionnés) évaluée sur les annonces de toute la base correspondant au même type de poste (contrat, secteur,

métier), nous recommandons en complément le site gratuit ayant la performance attendue la plus élevée, n'étant pas déjà recommandé et appartenant à l'interface du recruteur.

Conclusion et perspectives

Bilan

L'usage de la fouille de textes sur les offres d'emploi nous a permis dans un premier temps de structurer l'information qu'elles contiennent du point de vue des missions proposées grâce à l'utilisation d'un algorithme supervisé. Les offres d'emploi sont désormais catégorisées en fonction du métier qui leur est associé. Cette nomenclature rend possible l'analyse et la comparaison des performances des sites d'emploi relativement au métier proposé, critère de décision essentiel pour les recruteurs.

Dans un deuxième temps, nous avons eu recours à des méthodes issues de l'analyse textuelle pour extraire un ensemble de mots-clés associés à l'offre, et les avons utilisés pour enrichir la description de cette dernière au sein d'un algorithme prédictif. L'analyse des résultats obtenus a montré la pertinence de ces prédicteurs pour contribuer à expliquer la performance des offres d'emploi sur Internet.

Pour prédire la performance des offres, nous avons proposé comme approche un système hybride de recommandation, adapté à la situation de démarrage à froid. Lors de nos expérimentations sur un jeu de données réelles, cette nouvelle approche s'est montrée supérieure aux approches standards de modélisation multivariée. De plus, la flexibilité de notre algorithme permet de mettre en place un système pouvant être assimilé à du "retour de pertinence", et améliorant la qualité des résultats pour des offres rediffusées.

Dans un contexte où les entreprises consacrent une partie importante de leur budget aux recrutements, et où les coûts doivent être réduits autant que possible, il est nécessaire de contrôler les dépenses effectuées pour la diffusion des annonces sur Internet. Les travaux que

nous avons menés permettent de comparer avant la diffusion d'une offre les performances attendues sur les sites d'emploi pour un poste à pourvoir donné. Principalement à but informatif pour les supports gratuits, l'indication sur la performance attendue permet au recruteur de mettre en regard les rendements des supports payants. La mise en application de ces résultats au sein d'une solution de multidiffusion d'annonce permet d'accompagner le recruteur durant le processus de diffusion de son annonce, et lui fournit une aide pour le choix des sites d'emploi à utiliser.

Perspectives de recherche

Suite à ces résultats encourageants, notre premier axe de recherche concerne l'amélioration de la qualité de prédiction obtenue, grâce à l'introduction de facteurs explicatifs complémentaires. Notamment, nous souhaitons prendre en compte dans le modèle des indicateurs de la conjoncture sur le marché de l'emploi associée aux différentes fonctions. Les travaux actuels prennent en compte la conjoncture sur le marché de l'emploi d'un point de vue général grâce à l'exploitation des indices de diffusion d'offres d'emploi sur Internet. Dans de futurs travaux, nous exploiterons des indices de diffusion d'offres d'emploi spécifiques aux différentes fonctions. Ce travail sera permis grâce à la structuration des offres que nous avons mise en place. En effet, selon la fonction associée à une offre donnée, il sera possible de rattacher l'indicateur de conjoncture correspondant.

Un deuxième axe de recherche concerne la gestion des sites d'emploi nouveaux, c'est-à-dire sur lesquels très peu d'annonces ont été diffusées, voire aucune. Sans historique de diffusion, il n'est pas possible de comprendre et prédire les performances par le biais d'un modèle explicatif standard. Nous proposons donc de décrire les sites d'emploi à travers un ensemble de variables choisies pour leur pouvoir explicatif, ce qui permettra une nouvelle application des systèmes de recommandation. Les similarités entre sites pourront être évaluées et l'estimation des performances attendues sur les sites nouveaux se basera sur les performances obtenues sur les sites similaires.

Enfin, notre troisième axe de recherche est lié à l'ajout de fonctionnalités au module d'aide à la décision. Nous souhaitons, sur la base du métier identifié, suggérer au recruteur des

mots-clés adaptés à son offre et qui, insérés dans le titre ou dans le texte de l'annonce, permettront d'accroître la visibilité de son offre sur les sites d'emploi. Le texte de l'annonce étant commun à l'ensemble des sites, il nous faudra valider la pertinence de ces mots-clés de manière globale. En parallèle, il nous faudra tester et mettre en évidence l'impact de la rédaction du titre sur la performance des offres d'emploi.

Bibliographie

J. Aaker. Dimensions of brand personality. *Journal of Marketing Research*, 34 :347–356, 1997. 63

H. Abdi. Partial least square regression, projection on latent structure regression, PLS-regression. *Wiley Interdisciplinary Reviews : Computational Statistics*, 26 :97–106, 2010. 123, 135

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6) :734–749, 2005. 133, 139

D. G. Allen, J. R. Van Scotter, and R. F. Otondo. Recruitment communication media : Impact on prehire outcomes. *Journal of Applied Psychology*, 57 :143–171, 2004. 62

D. G. Allen, R. V. Mahto, and R. F. Otondo. Web-based recruitment : Effects of information, organizational brand, and attitudes toward a web site on applicant attraction. *Journal of Applied Psychology*, 92 :1696–1708, 2007. 62

H. Azzag, C. Guinot, et G. Venturini. Classification hiérarchique et visualisation de pages web. In *Actes de l'atelier Fouille du Web - 6es journées francophones EGC*, pages 5–16, 2006. 89

R. Baeza-Yates and B. Riberto-Neto. *Modern information retrieval.* New York : ACM Press, 1999. 89, 138

M. Balabanovic and Y. Shoham. Fab : Content-based, collaborative recommendation. *Communications of the ACM*, 40(3) :66–72, 1997. 132

A. E. Barber and M. V. Roehling. Job posting and the decision to interview : A verbal protocol analysis. *Journal of Applied Psychology*, 78 :845–856, 1993. 62

D. Bartram. Internet recruitment and selection : Kissing frogs to find princes. *International Journal of Selection and Assessment*, 8(4) :261–274, 2000. 98

D. Bartram. Testing on the internet : Issues, challenges and opportunities in the field of occupational assessment. In D. Bartram and R. Hambleton, editors, *Computer-based testing and the Internet : Issues and Advances*, pages 13–37. John Wiley and Sons, Chichester, 2005. 47

A. Belt and J. Paolillo. The influence of corporate image and specificity of candidate qualifications on response to recruitment advertisement. *Journal of Management*, 8 : 105–112, 1982. 63

J.-P. Benzécri. Analyse discriminante et analyse factorielle. *Les Cahiers de l'Analyse des Données*, 2(4) :369–406, 1977. 103

C. Bizer, R. Heese, M. Mochol, R. Oldakowski, R. Tolksdorf, and R. Eckstein. The impact of semantic web technologies on job recruitment process. In *WI'05 : International Conference on Wirtschaftsinformatik*, 2005. 99

A. Blackman. Graduating students' responses to recruitment advertisements. *Journal of Business Communication*, 43(4) :367–388, 2006. 62

D. M. Blei, A. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3 :993–1022, 2003. 97

P. W. Braddy, A. W. Meade, and C. M. Kroustalis. Organizational recruitment website effects on viewers' perceptions of organizational culture. *Journal of Business and Psychology*, 20 :525–543, 2006. 64

J. A. Breaugh. *Recruitment : Science and practice.* Boston : PWS-Kent, 1992. 46

J. A. Breaugh. Employee recruitment : Current knowledge and important areas for future research. *Human Resource Management Review*, 18 :103–118, 2008. 61, 64

J. A. Breaugh and M. Starke. Research on employee recruitment : So many studies, so many remaining questions. *Journal of Management*, 26(3) :405–434, 2000. 45

J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, 1998. 131

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Chapman & Hall/CRC, 1984. 103

Broadbean. Job posting distribution and resume search software – broadbean.com, novembre 2011a. URL `http://www.broadbean.com`. 40

Broadbean. Online Recruitment Software – broadbean.com, novembre 2011b. URL `http://www.broadbean.com/recruiting-software.html`. 40

Broadbean. Always be in control of your budget and yours users – broadbean.com, novembre 2011c. URL `http://www.broadbean.com/recruitment_budgeting.html`. 40

Broadbean. Track and measure online recruitment performance – broadbean.com, novembre 2011d. URL `http://www.broadbean.com/roi-reporting.html`. 40

B. Brooke. Explosion of internet recruiting. *Hispanic*, 11(12) :68, 1998. 47

D. Cable and M. Graham. The determinants of job seekers' reputation perceptions. *Journal of Organizational Behavior*, 21 :929–947, 2000. 63

F. Cailliau et C. Poudat. Caractérisation lexicale des contributions clients agents dans un corpus de conversations téléphoniques retranscrites. In *JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles*, pages 267–275, 2008. 90

L. Candillier. *Contextualisation, visualisation et évaluation en apprentissage non supervisé*. PhD thesis, Université Charles de Gaulle – Lille 3, 2006. 103

D. S. Chapman, K. L. Uggerslev, S. A. Carroll, K. A. Piasentin, and D. A. Jones. Applicant attraction to organizations and job choice : A meta-analytic review of the correlates of recruiting outcomes. *Journal of Applied Psychology*, 90 :928–944, 2005. 63

I.-G. Chong and C. H. Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78 :103–112, 2005. 123

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1) :22–29, 1990. 92

R. T. Cober, D. J. Brown, P. E. Levy, A. B. Cober, and L. M. Keeping. Organizational web sites : Web site content and style as determinants of organizational attraction. *International Journal of Selection and Assessment*, 11 :158–169, 2003. 64

G. V. Cormack. Email spam filtering : A systematic review. *Foundations and Trends in Information Retrieval*, 1(4) :335–455, 2006. 89

M. Damashek. Gauging similarity with n-grams : language independent categorization of text. *Science*, 267 :843–848, 1995. 98

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41 :391–407, 1990. 96

G. Dessler, J. Griffiths, B. Lloyd-Walker, and A. Williams. *Human resource management.* Maryborough, Australia : Prentice Hall, 1999. 62

eQuest. eQuest | The World's Leading Job Distributor, novembre 2011a. URL `http://www.equest.com`. 37

eQuest. eQuest | The World's Leading Job Distributor, novembre 2011b. URL `http://www.equest.com/about`. 37

eQuest. eQuest Chameleon | Job Posting Delivery Gateway, novembre 2011c. URL `http://chameleon.equest.com`. 37

eQuest. TRAQ24 by eQuest | Metrics. Realtime. Anytime, novembre 2011d. URL `http://www.equest.com/traq24/overview.php`. 37

Exclusive RH. R.Flex vous offre le ROI de chaque jobboard, avril 2010. URL `http://exclusiverh.com/logiciel-rh/r-flex-vous-offre-le-roi-de-chaque-jobboard.htm`. 41

Y. Fondeur. Le recrutement par internet : Le dilemme transparence/bruit. *Personnel*, 472 : 46–48, 2006. 48

Y. Fondeur et C. Tuchszirer. *Internet et les intermédiaires du marché du travail.* IRES, 2005. 29, 48

E. Galanaki. The decision to recruit online : A descriptive study. *Career Development International*, 7(4) :243–251, 2002. 47

W. N. Gansterer, A. G. K. Janecek, and R. Neumayer. Spam filtering based on latent semantic indexing. In M. W. Berry and M. Castellanos, editors, *Survey of Text Mining : Clustering, Classification, and Retrieval, Second Edition*, pages 165–183. Springer, 2007. 89

H. Garner et B. Lutinier. Les procédures de recrutement : canaux et modes de sélection. *Premières Synthèses, DARES*, 48.1, 2006a. 31

H. Garner et B. Lutinier. Des difficultés pouvant aller jusqu'à l'échec du recrutement. *Premières Synthèses, DARES*, 48.2, 2006b. 31

R. D. Gatewood, M. A. Gowan, and G. J. Lautenschlager. Corporate image, recruitment image, and initial job choice decision. *Academy of Management Journal*, 36(2) :414–427, 1993. 62, 63

P. Gehler, A. Holub, and M. Weilling. The rate adapting poisson model for information retrieval and object recognition. In *ICML'06 : 23rd International Conference on Machine Learning*, pages 337–344, 2006. 97

Groupe Aktor. Groupe Aktor Conseil en stratégie de communication, novembre 2011. URL `http://www.aktor-group.com`. 41

B. Habert, G. Illouz, P. Lafon, S. Fleury, H. Folch, S. Heiden, et S. Prévost. Profilage de textes : cadre de travail et expérience. In R. Rajman, editor, *JADT 2000 : 5e Journées internationales d'Analyse statistique de Données Textuelles*, 2000. 75

J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28 : 100–108, 1979. 103

I. S. Helland. PLS regression and statistical models. *Scandinavian Journal of Statistics*, 17 :97–114, 1990. 135

J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999. 131

T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR'99 : 22nd International Conference on Research and Development in Information Retrieval*, pages 50–57, 1999. 97

A. Hoskuldsson. PLS regression methods. *Journal of Chemometrics*, 2 :211–228, 1988. 136

N. Jardine and C. J. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7(5) :217–240, 1971. 103

T. Joachims. Text categorization with support vector machines : learning with many relevant features. In *ECML'98 : European Conference on Machine Learning*, pages 137–142, 1998. 89, 107

C. Kaydo and A. Cohen. The hits and misses of online hiring. *Sales and Marketing Management*, 153(10) :13, 1999. 48

R. Kessler. *Traitement automatique d'informations appliqué aux ressources humaines.* PhD thesis, Université d'Avignon et des Pays de Vaucluse, 2009. 42, 111

R. Kessler, J.-M. Torres-Moreno, et M. El-Bèze. Classification automatique de courriers électroniques par des méthodes mixtes d'apprentissage. *Revue des Sciences et Technologies de l'Information - Série ISI*, 11 :93–112, 2006. 89

R. Kessler, J.-M. Torres-Moreno, and M. El-Bèze. E-gen : automatic job offer processing system for human resources. In *MICAI'07 : Mexican International Conference on Artificial Intelligence*, pages 985–995, 2007. 108

R. Kessler, N. Béchet, J.-M. Torres-Moreno, M. Roche, et M. El-Bèze. Profilage de candidatures assisté par relevance feedback. In *TALN'09 : Traitement Automatique des Langues Naturelles*, 2009. 98

Kimladi. Multi-diffusion d'offres d'emploi | Kimladi, novembre 2011. URL `http://www.kimladi.com/page.php?id_page=22`. 42

Kioskemploi-Aktor HR Software. Jobposting et gestion des publicités de recrutement : Kioskemploi.com | Aktor HR Software, novembre 2011. URL `http://www.kioskemploi.com/logiciel-statistiques-recrutement.html`. 42

J. Koch. *The economics of affirmative action.* Lexington, MA : Lexington Books, 1976. 62

T. Kohonen. *Self-Organizing Maps. Third, extended edition.* Springer, 2001. 103

J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. Grouplens : Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3) :77–87, 1997. 131

J. J. Laabs. Recruiting in the global village. *Workforce*, 77(4) :30–33, 1998. 47

C. Labbé et D. Labbé. Ce que disent leurs phrases. In *JADT 2010 : 10es Journées internationales d'Analyse statistique des Données Textuelles*, pages 297–307, 2010. 90

P. Lafon. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1 :127–165, 1980. 93

L. Lebart. Traitement statistique des questions ouvertes ; quelques pistes de recherche. *Journal de la Société Française de Statistique*, pages 7–21, 2003. 90

L. Lebart, A. Salem, and L. Berry. *Exploring textual data.* Kluwer, 1998. 97, 103

J. Lemmink, A. Schuijf, and S. Streukens. The role of corporate image and company employment image in explaining application intentions. *Journal of Economic Psychology*, 24 :1–15, 2003. 63

H. Levene. Robust tests for the equality of variance. In I. Olkin, editor, *Contributions to Probability and Statistics*, page 278–292. Palo Alto, CA : Stanford University Press, 1960. 80

F. Lievens and S. Highhouse. The relation of instrumental and symbolic attributes to a company's attractiveness as an employer. *Personnel Psychology*, 56(1) :75–103, 2003. 63

R. Loth, D. Battistelli, F.-R. Chaumartin, Mazancourt H., J.-L. Minel, and A. Vinckx. Linguistic information extraction for job ads (sire project). In *RIAO'10 : Adaptivity, Personalization and Fusion of Heterogeneous Information*, pages 222–224, 2010. 99

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press, 1967. 103

C. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval.* Cambridge University Press, 2008. 89

J.F. Marcotorchino et N. El Ayoubi. Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association. *Revue de Statistique Appliquée*, 39(2) : 25–46, 1991. 105

D. I. Martin and M. W. Berry. Mathematical foundations behind latent semantic analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*, pages 35–55. Lawrence Erlbaum Associates, 2007. 96

B. Mathews and T. Redman. Managerial recruitment advertisements – just how market oriented are they ? *International Journal of Selection and Assessment*, 6(4) :240–248, 1998. 63

T. Mitchell. *Machine Learning.* McCraw Hill, 1996. 93

D. Monière et D. Labbé. Essai de stylistique quantitative. In A. Morin and P. Sébillot, editors, *JADT 2002 : 6es Journées internationales d'Analyse statistique des Données Textuelles*, pages 561–569, 2002. 75, 90

R. J. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *Proceedings of ACM SIGIR'99 Workshop Recommender Systems : Algorithms and Evaluation*, 1999. 132

Multiposting.fr. Multiposting.fr | Multi-diffusion d'annonces d'emploi et de stage, novembre 2011. URL `http://www.multiposting.fr`. 35

F. Namer. Un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues*, 41 :247–523, 2000. 90

M. Pazzani and D. Billsus. Learning and revising user profiles : The identification of interesting web sites. *Machine Learning*, 27 :313–331, 1997. 132

K. Pearson. Mathematical contribution to the theory of evolution - vii. on the correlation of characters not quantitatively measurable. In *Philosophical Transactions of the Royal Society of London. Series A*, volume 195, pages 1–47. 1900. 131

J. Petrick and D. Furr. *Total quality in managing human resources.* Delray Beach, FL : St. Lucie, 1995. 63

R. J. Pin, M. Laorden, and I. Saenz-Diez. Internet recruiting power : Opportunities and effectiveness. Research Paper 439, IESE, 2001. 47

D. Politt. E-recruitment helps xerox to pick the cream of the crop. *Human Resource Management*, 12(5) :33–35, 2004. 47

M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3) :130–137, 1980. 90

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1) :81–106, 1986. 93

J. R. Quinlan. *C4.5 : Programs for machine learning.* Morgan Kaufmann Publishers, 1993. 103

V. Radevski, Z. Dika, and F. Trichet. Common : A framework for developing knowledge-based systems dedicated to competency-based management. In *ITI'06 : 28th International Conference on Information Technology Interfaces*, pages 419–424, 2006. 99

A. Rafaeli, O. Hadomi, and T. Simons. Recruiting through advertising or employee referrals : Costs, yields, and the effect of geographic focus. *European Journal of Work and Organizational Psychology*, 14 :355–366, 2005. 62

RFlex. RFLEX – Logiciels RH | Présentation, novembre 2011. URL `http://fr.rflex.com/presentation.html`. 41

Q. M. Roberson, C. J. Collins, and S. Oreg. The effects of recruitment message specificity on applicant attraction to organizations. *Journal of Business and Psychology*, 19 :319–339, 2005. 62

J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART retrieval system : Experiments in automatic document processing*, pages 313–323, 1971. 140

Hinton G. E. Rumelhart, D. E. and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*, pages 318–362. The MIT Press, 1986. 103

S. L. Rynes. Recruitment, job choice, and post-hire consequences. In M. D. Dunette, editor, *Handbook of industrial and organizational psychology*, pages 399–444. Palo Alto, CA : Consulting Psychologists Press, 1991. 45, 46

S. L. Rynes and D. M. Cable. Recruitment research in the twenty-first century. In W. C. Borman, D. R. Ilgen, and R. J. Klimoski, editors, *Handbook of Psychology : Industrial and organizational psychology*, volume 12, pages 55–76. Hoboken, NJ : John Wiley and Son, 2003. 63, 64

G. Salton. *Automatic text processing*. Addison-Wesley, 1989. 96

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18 :613–620, 1975. 94

A. I. Schein, A. Popescul, L. H. Ungar, and D.M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, 2002. 132

H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, 1994. 90, 110, 159

U. Shardanand. Social information filtering for music recommendation. Master's thesis, Massachusetts Institute of Technology, 1994. 138

U. Shardanand and P. Maes. Social information filtering : Algorithms for automating "word of mouth". In *Proceedings of Conference on Human Factors in Computing Systems*, 1995. 131

N. Slonim, N. Friedman, and N. Tishby. Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, 2002. 111

SmartRecruiters. SmartRecruiters Free recruiting software Great Candidates | SmartRecruiters, novembre 2011. URL `http://www.smartrecruiters.com/static/product/find-great-candidates`. 42

M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, 2000. 89

D. L. Stone, K. Lukaszewski, and L. C. Isenhour. e-recruiting : Online strategies for attracting talent. In H. Gueutal and D. L. Stone, editors, *The Brave New World of EHR : Human Resources in the Digital Age*, pages 22–53. New York : John Wiley & Sons, 2005. 63

M. Stricker. *Réseaux de neurones pour le traitement automatique du langage : conception et réalisation de filtres d'informations.* PhD thesis, ESPCI ParisTech, 2000. 103

L. F. Thompson, P. W. Braddy, and K. L. Wuensch. E-recruitment and the benefits of organizational web appeal. *Computers in Human Behavior*, 24 :2384–2398, 2008. 62, 64

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1) :267–288, 1996. 123

F. Trichet, M. Bourse, M. Leclère, and E. Morin. Human resource management and semantic web technologies. In *ICTTA'04 : 1st International Conference on Information and Communication Technologies*, pages 641–642, 2006. 99

D. Turban, M. Forret, and C. Hendrickson. Applicant attraction to firms : Influences of organization reputation, job and organizational attributes, and recruiter behaviors. *Journal of Vocational Behavior*, 52 :24–44, 1998. 63

D. B. Turban and D. M. Cable. Firm reputation and applicant pool characteristics. *Journal of Organizational Behavior*, 24 :733–751, 2003. 64

A. M. Tybout and N. Artz. Consumer psychology. In M. R. Rosenzweig and L. W. Porter, editors, *Annual review of Psychology*, volume 45, pages 131–169. Palo Alto, CA : Annual Reviews Inc., 1994. 62

Ubiposting. Fonctionnalités | Ubiposting logiciel de multidiffusion d'offres d'emploi, novembre 2011. URL `http://www.ubiposting.com/Fonctionnalités.aspx`. 42

C. J. van Rijsbergen. *Information Retrieval.* London : Butterworths, 1979. 104

V. Vapnik. *The nature of statistical learning theory.* Berlin : Springer Verlag, 1995. 103

M. Veger. How does internet recruitment have effect on recruitment performance ? In *Proc. of the 4th Twente Student Conference on IT*, 2006. 47, 50

J. P. Wanous. *Organizational entry.* Reading, MA : Addison-Wesley, 1992. 46

J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301) :236–244, 1963. 103

W. J. Wilbur and K. Sirotkin. The automatic identification of stop words. *Journal of Information Science*, 18(1) :45–55, 1992. 94

C. R. Williams, C. E. Labig, and T. H. Stone. Recruitment sources and post-hire outcomes for job applicants and new hires : A test of two hypothesis. *Journal of Applied Psychology*, 78 :163–172, 1993. 46

I. O. Williamson, D. P. Lepak, and J. King. The effect of company recruitment web site orientation on individuals' perceptions of organizational attractiveness. *Journal of Vocational Behavior*, 63 :242–263, 2003. 64

H. Wold. Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiaah, editor, *Multivariate Analysis*, pages 391–420, 1966. 136

S. Wold, C. Albano, W. J. D. III, K. Esbensen, S. Hellberg, E. Johansson, and H. Sjostrom. Pattern recognition : Finding and using regularities in multivariate data. In *Proceedings of IUFOST Conference : Food Research and Data Analysis*, pages 147–188, 1983a. 135

S. Wold, H. Martens, and H. Wold. The multivariate calibration problem in chemistry solved by the PLS method. In *Proceedings of the Conference on Matrix Pencils*, pages 286–293, 1983b. 135

L. Yahiaoui, Z. Boufaïda, and Y. Prié. Semantic annotation of documents applied to e-recruitment. In *SWAP'06 : 3rd Italian Semantic Web Workshop*, 2006. 99

Y. Yang and J. P. Pederson. A comparative study on feature selection in text categorization. In *ICML'97 : International Conference on Machine Learning*, pages 412–420, 1997. 114

Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. Technical report, Department of Computer Science and Engineering - University of Minnesota, 2002. 89

R. R. Zusman and R. S. Landis. Applicant preferences for web-based versus traditional job postings. *Computers in Human Behavior*, 18(3) :285–296, 2002. 47

Annexe A

Nomenclature “fonction” des offres d’emploi

Le tableau ci-dessous présente la nomenclature exhaustive des offres d’emploi suite aux prétraitements effectués sur la nomenclature initiale (issue d’un site d’emploi généraliste). La nomenclature obtenue présente un découpage en 23 fonctions (classes de permier niveau) et 107 sous-fonctions ou métiers (classes de deuxième niveau).

Fonctions	**Sous-fonctions**
Architecture, Création	Architecture/Design industriel
	Autres métiers
Services administratifs	Accueil/Réception
	Employé administratif/Saisie informatique
	Secrétariat/Assistanat de direction
	Assistant Import-Export
	Gestion locative/de copropriété
	Autres métiers
BTP & second-œuvre	CAO-DAO / Dessinateur-projeteur
	CVC - Conception/Installation
	Chef de chantier & Conducteur de travaux
	Electricité
	Géomètre & Economiste de la construction
	Maçonnerie/Béton/Sols & Murs
	Plomberie/Tuyauterie
	Autres métiers

TABLE A.1 – Liste des fonctions et sous-fonctions de la nomenclature finale des offres d’emploi : “Architecture, Création”, “Services administratifs” et “BTP”

Fonctions	**Sous-fonctions**
Commercial / Vente	Agent Immobilier / Courtage
	Assistanat commercial
	Commercial
	Courtage - Assurances
	Distribution/Vente en gros/Revendeur
	International
	Télévendeur/Commercial sédentaire
	Vente d’espace
	Autres métiers
Stratégie & Management	Consulting
	Responsable de rayon/Direction d’agence ou magasin
	Direction de département/Direction générale
	Autres métiers
Édition & Écriture	
Ingénierie & Recherche	CAO-DAO
	Environnement/Géologie
	Génie civil/structures
	Génie électrique et Génie industriel/Process et méthodes
	Hardware/Systèmes embarqués
	Mécanique/Aéronautique
	Nouveaux produits
	Pétro-chimie/Chimie/Energie/Nucléaire
	Recherche clinique et pharmaceutique
	Recherche en Chimie/Biologie/Science de la matière
	Télécom - Sans fil & RF
	Autres métiers
Comptabilité & Finance	Analyse financière/Analyste crédits
	Audit/Contrôle de gestion
	Banque-Particulier/Conseiller financier/Gestion de patrimoine
	Comptabilité générale/Facturation
	Courtage/Rédacteur d’assurance
	Expert-Comptable/Finance/Fiscalité
	Gestion des Risques
	Placements/Investissements - Opérations sur titres
	Recouvrement
	Autres métiers
Gestion de projet	

TABLE A.2 – Liste des fonctions et sous-fonctions de la nomenclature finale des offres d’emploi : “Commercial / Vente”, “Stratégie & Management”, “Édition & Écriture”, “Ingénierie & Recherche”, “Comptabilité & Finance” et “Gestion de projet”

Fonctions	Sous-fonctions
Hôtellerie, Restauration	Accueil/Réception et Service
	Cuisinier
	Autres métiers
Juridique	Droit du travail/Droit fiscal
	Paralégal & Secrétariat juridique
	Propriété Intellectuelle/Juriste-Mandataire
	Autres métiers
Logistique & Transport	Achats/Import-Export
	Approvisionnements/Gestion des stocks
	Livraisons/Coursiers/PL/SPL
	Magasinier/Chargement/Entreposage
	Transport maritime & aérien
	Autres métiers
Marketing	Evènementiel/Communication/Relations Presse
	Marketing Direct/CRM/Etudes/Enquêtes
	Marketing Produit/Chef de Produit
	Autres métiers
Installation & Maintenance	Chauffagiste/Climatisation
	Equipement industriel
	Matériels informatiques/électriques/télécom
	Réparation, Entretien et Carrosserie auto
	Autres métiers
Production & Opérations	Assemblage/Façonnage/Opérateur machine
	Gestion de la production/des opérations
	Production agro-alimentaire
	Autres métiers
Qualité / Inspection	Certification/Inspection-batîments
	Environnement/Sécurité
	Qualité - Production
	Autres métiers
Formation / Éducation	
Ressources Humaines	Paie & Administration du personnel
	Politique de Recrutement
	Recrutement
	Autres métiers
Santé	

TABLE A.3 – Liste des fonctions et sous-fonctions de la nomenclature finale des offres d’emploi : “Hôtellerie, Restauration”, “Juridique”, “Logistique & Transport”, “Marketing”, “Installation & Maintenance”, “Production & Opérations”, “Qualité / Inspection”, “Formation / Éducation”, “Ressources Humaines” et “Santé”

Fonctions	**Sous-fonctions**
Informatique & Technologies	Architecture/Systèmes
	Bases de données/Data warehouse/Décisionnel
	Développement Jeux
	Développement/Programmation
	ERP/Progiciels/CRM
	Etudes/Analyste Système
	Gestion de projet IT
	Réseaux & serveurs/Support/Helpdesk
	Sécurité
	Télécommunications
	Webdesign & Ergonomie
	Autres métiers
Sécurité	
Services clientèle	Agent de réservation/Guichetier/Vendeur détail
	Service après vente/Support technique
	Service clientèle/Centre d’appel
	Autres métiers
Autres	

TABLE A.4 – Liste des fonctions et sous-fonctions de la nomenclature finale des offres d’emploi : “Informatique & Technologies”, “Sécurité”, “Services clientèle” et “Autres”

Glossaire

- CV : Curriculum Vitae ;
- job board : site web de recherche d'emploi, les entreprises peuvent y poster des annonces d'emploi ;
- mapping : procédé qui consiste à établir une correspondance entre deux nomenclatures différentes ;
- ROI : return on investment, retour sur investissement ;
- sourcing : recherche de candidats qualifiés pour un poste à pourvoir ;
- tracking : moyens mis en œuvre pour enregistrer les actions des internautes sur l'ensemble du processus de candidature ;
- URL : Uniform Resource Locator, adresse web.

Julie SEGUELA

Fouille de données textuelles et systèmes de recommandation appliqués aux offres d'emploi diffusées sur le web

Résumé :
L'expansion du média Internet pour le recrutement a entraîné ces dernières années la multiplication des canaux dédiés à la diffusion des offres d'emploi. Dans un contexte économique où le contrôle des coûts est primordial, évaluer et comparer les performances des différents canaux de recrutement est devenu un besoin pour les entreprises. Cette thèse a pour objectif le développement d'un outil d'aide à la décision destiné à accompagner les recruteurs durant le processus de diffusion d'une annonce. Il fournit au recruteur la performance attendue sur les sites d'emploi pour un poste à pourvoir donné. Après avoir identifié les facteurs explicatifs potentiels de la performance d'une campagne de recrutement, nous appliquons aux annonces des techniques de fouille de textes afin de les structurer et d'en extraire de l'information pertinente pour enrichir leur description au sein d'un modèle explicatif. Nous proposons dans un second temps un algorithme prédictif de la performance des offres d'emploi, basé sur un système hybride de recommandation, adapté à la problématique de démarrage à froid. Ce système, basé sur une mesure de similarité supervisée, montre des résultats supérieurs à ceux obtenus avec des approches classiques de modélisation multivariée. Nos expérimentations sont menées sur un jeu de données réelles, issues d'une base de données d'annonces publiées sur des sites d'emploi.

Mots clés :
Fouille de textes, extraction des connaissances, systèmes de recommandation, offres d'emploi, recrutement sur Internet

Abstract :
Last years, e-recruitment expansion has led to the multiplication of web channels dedicated to job postings. In an economic context where cost control is fundamental, assessment and comparison of recruitment channel performances have become necessary. The purpose of this work is to develop a decision-making tool intended to guide recruiters while they are posting a job on the internet. This tool provides to recruiters the expected performance on job boards for a given job offer. First, we identify the potential predictors of a recruiting campaign performance. Then, we apply text mining techniques to the job offer texts in order to structure postings and to extract information relevant to improve their description in a predictive model. The job offer performance predictive algorithm is based on a hybrid recommender system, suitable to the cold-start problem. The hybrid system, based on a supervised similarity measure, outperforms standard multivariate models. Our experiments are led on a real dataset, coming from a job posting database.

Keywords :
Text mining, knowledge discovery, recommender systems, job postings, e-recruitment