EXTRACTION D'INFORMATIONS A PARTIR D'OFFRES D'EMPLOI ET CREATION DE DICTIONNAIRES ELECTRONIQUES

Type de contenu : Images animées

Titre(s) : EXTRACTION D'INFORMATIONS A PARTIR D'OFFRES D'EMPLOI ET CREATION DE DICTIONNAIRES ELECTRONIQUES ; BARTEYE, Olivier ; GUENTHNER, Franz ; SLT CHANAL, Justine|SLT GREGOIRE, Thibaud

Autre(s) responsabilité(s) : BARTEYE, Olivier (Directeur de thèse)
GUENTHNER, Franz (Directeur de thèse)
SLT CHANAL, Justine|SLT GREGOIRE, Thibaud (Secrétaire)

Editeur, producteur : Ecoles Militaires de Saint-Cyr Coëtquidan

Note de thèses et écrits académiques : Filière Scientifique - Option Informatique Promotion Chef d'Escadron Francoville Date de soutenance : 01/01/2011

Résumé ou extrait : > Etude : Introduction : De nos jours, Internet est devenu le moyen privilégié de la recherche d'emploi. Cependant, sa notoriété constitue également sa faiblesse. En effet, le service étant rarement gratuit, certaines entreprises ne disposent pas des fonds nécessaires ou refusent tout simplement de payer ce type de prestations. Il en est de même pour les chercheurs d'emploi. D'autre part, trouver ce que l'on cherche est loin d'être simple, car si l'information est disponible elle peut quand même être introuvable. Deux principales raisons à cela : la plus évidente vient du candidat car très souvent, la requête est mal formulée. Mais parfois, le problème vient de l'offre elle-même qui est soit mal conçue soit dissimulée sur des sites à accès restreints pour filtrer les candidatures. C'est ainsi qu'est née la start-up Jobanova. Elle possède un moteur de recherche, développé au CIS, qui récupère les offres d'emploi directement à leurs sources, et en extrait les informations lui-même afin de toutes les mettre sous une forme commune, facilitant ainsi recherches et comparaisons. Jobanova fonctionne dans plusieurs langues, dont l'allemand, l'anglais et le français, même si la première offre les meilleurs résultats. Notre tâche à donc consisté à participer à l'amélioration de la version française, notamment en créant des dictionnaires électroniques pour le logiciel Unitex qui est très utilisé au CIS. Contraintes : Il faut savoir que nous avons du travailler avec des données réelles, c'est à dire les mêmes que celles disponibles sur des sites de recherche d'emploi, ce qui est à la fois une chance mais aussi source de difficultés. La taille de ces fichiers était souvent conséquente, au point de « planter » nos ordinateurs personnels, pourtant récents. De plus nous avons pu constater par nous même de la disparité des sources, tant par leurs contenus que leurs formes. En outre, nous n'étions pas intégrés à une équipe mais travaillions uniquement en binôme, ce qui d'un côté permet une relative autonomie mais d'un autre impose auto-discipline et méfiance afin d'éviter de faire fausse route. Pour finir, notre directeur de stage avait un emploi du temps très chargé, et il nous fallait en tenir compte pour les diverses réunions de travail que nous avons faites. Démarche : Avant tout, nous avons commencé notre stage comme le souhaitait notre directeur de stage, par la lecture de la thèse de Sandra BSIRI du 27 avril 2007 intitulée : Extraction d'information : génération automatique d'une base de données d'offres d'emploi. Nous avons ensuite créé nos premiers dictionnaires de JD (Job Description ou nom de métier comme « boulanger »), étape préliminaire à l'extraction d'information à partir d'offres

d'emploi. Ces premiers dictionnaires se sont appuyés sur des bases existantes crées au CIS par Sandra BSIRI et d'autres SLT en stage international. Mais nous avons du les normaliser car ils n'étaient pas tous sous la même forme. Puis par la suite, nous avons chercher à appliquer ces dictionnaires, via le logiciel Unitex, sur de nouvelles listes de JD ou des corpus d'offres d'emploi, provenant de sites Internet, afin d'en augmenter la taille. C'est cette tâche qui nous a pris le plus de temps, d'autant plus qu'il fallait fléchir les JD, ce qui signifie générer les formes masculin pluriel ainsi que féminin singulier et pluriel à partir du masculin singulier. Mais le professeur GUENTHNER nous avait bien fait comprendre l'importance d'avoir des dictionnaires performants, c'est-à-dire contenant un maximum de JD différents. Afin de maximiser nos sources de nouveaux JD, nous avons du nettoyer des corpus d'offres d'emploi de tout ce qui n'était pas nécessaire comme les offres en langue étrangère ou la pollution graphique résultant d'erreurs d'encodage, de balises, de suites de caractère vides de sens. Grâce à des grammaires locales, qui sont des graphiques, le logiciel Unitex peut récupérer des JD complexes à partir des simples. Ceci fait, nous avons pu faire d

Sujet(s) : armée
chômeur
communication de l'information
dictionnaire électronique
gestion de l'information
information électronique
informatique
offre d'emploi
traitement de l'information