



MINISTÈRE DE L'ÉDUCATION  
NATIONALE, DE L'ENSEIGNEMENT  
SUPÉRIEUR ET DE LA RECHERCHE

>>>

## AVIS DE SOUTENANCE DE THESE DE DOCTORAT

**Monsieur Cédric VIDREQUIN** soutiendra une thèse  
**le 17 décembre 2008 à 10h**

**LIA**

**SPÉCIALITÉ : INFORMATIQUE ED 166**

Titre de la thèse : Constitution automatique de bases de connaissances à partir de données  
textuelles non structurées

Membres du jury :

ANTOINE Jean-Yves, PR Informatique, Université François-Rabelais  
QUAFAFOU Mohamed, PR Informatique, Université de la Méditerranée (U2)  
SCHNEIDER Jean-Jacques, Doct. Informatique SEMANTIA  
TORRES-MERENO Juan Manuel, MCF Informatique, Université d'Avignon et  
des Pays de Vaucluse,  
EL-BÈZE Marc PR Informatique Université d'Avignon et des Pays de Vaucluse.

Résumé de la thèse :

L'objectif de cette thèse est l'exploration de modèles permettant la constitution automatisée de bases de connaissances. Ces bases sont constituées en appliquant des procédures d'apprentissage sur des corpus de textes statiques ou évoluant dans le temps. Le mémoire s'articule autour de deux grandes parties, décrivant chacune une méthode d'enrichissement automatique de bases de connaissances, appliquée à un domaine spécifique. Ces deux méthodes ont en commun de travailler sur des documents textuels peu ou non structurés, et d'utiliser très peu de connaissances en amorce. Pour pouvoir être transposables d'une langue à l'autre, ou d'un domaine à l'autre, on a recours aussi peu que possible aux ressources linguistiques. On donne donc une préférence marquée pour les méthodes numériques. Mais l'objectif étant de produire des bases de connaissances, on ne rejette pas pour autant les méthodes symboliques. Au contraire, il s'agit d'étudier dans quelle mesure on peut produire certains des éléments sur lesquels elles s'appuient, au moyen de méthodes numériques. Le premier volet est orienté vers les moteurs de questions-réponses. Lors des dernières campagnes d'évaluation, les systèmes les plus performants ont employé des quantités massives d'expressions régulières. L'écriture de ces règles nécessitant un investissement énorme, il n'est donc pas raisonnable de se contenter de ce mécanisme pour gérer le tout venant des questions pouvant être posées par un utilisateur, dans un système interactif. On envisage donc de déduire automatiquement ces règles à partir de textes. Chaque question induit une relation entre deux éléments, dont l'un est connu et l'autre pas. Pour chacune des relations, on construit manuellement une courte amorce, utilisée dans un algorithme d'amorce mutuelle à double niveau, entre patrons et couples de termes. À partir des couples d'amorce, le système génère des listes de patrons qui sont enrichies de façon semi-supervisée, puis utilisées pour trouver de nouveaux couples. Ces couples sont à leur tour réutilisés pour générer, par itérations successives, de nouveaux patrons. Les différentes étapes de l'enrichissement automatique utilisent les patrons de phrase dans des recherches d'informations sur Internet.

Le second volet traite de l'extraction automatique d'informations sur des micro-textes. Le cas pratique des petites annonces est étudié, avec pour objectif d'extraire automatiquement l'ensemble des critères de petites annonces. Ces critères doivent définir au mieux les informations structurant les petites annonces. Dans la plupart des cas, une amorce est malgré tout nécessaire. Elle est donc construite de façon à être courte et facilement adaptable. On profite de cette dernière pour enrichir la liste des critères extraits, en fournissant des informations sous-jacentes aux données extraites. Nous comparons enfin deux méthodes d'extraction, l'une utilisant un découpage de l'annonce en fonction des caractères de ponctuation, l'autre utilisant des collocations et ratios de vraisemblance.

UNIVERSITÉ D'AVIGNON  
ET DES PAYS DE VAUCLUSE  
MAISON DE LA RECHERCHE  
COLLEGE DES ETUDES DOCTORALES  
Campus centre-ville  
Site Ste Marthe  
74 rue Louis Pasteur  
84029 AVIGNON CEDEX 1  
<http://www.univ-avignon.fr>  
tél : +33(0)4 90 16 25 29  
fax : +33(0)4 90 16 25 31  
joelle.derbaise@univ-avignon.fr