UNIVERSITÉ DE
NEUCHÂTEL

Institut d'informatique

Rue Emile-Argand 11
CH-2000 Neuchâtel

Neuchâtel, le 10 août 2015

Faculté de Médecine
Secr. Ecole doctorale STIM
1, rue Gaston Veil
F-44035 Nantes Cedex 1

Thèse de Monsieur Adrien Bougouin (informatique)

Monsieur le Directeur,

Jacques SAVOY
Professeur
Jacques.Savoy@unine.ch
Tel. +41 (032) 718 13 75

La thèse de doctorat de Monsieur Adrien Bougouin porte le titre « Indexation automatique par termes-clés en domaines de spécialité ». Cette thèse se situe au carrefour de la linguistique computationnelle et de la recherche d'information, domaines dans lesquels l'équipe du prof. B. Daille est reconnue internationalement.

Dans une introduction, l'auteur expose les motivations sous-jacentes à son sujet et indique clairement la problématique de son domaine de recherche. Les trois hypothèses principales de ce travail original sont clairement présentées et justifiées. Enfin, l'auteur présente un bref survol de sa thèse.

Dans un deuxième chapitre, l'auteur expose le contexte général de ses travaux en indexation automatique en prenant le temps de définir clairement les notions et concepts essentiels de ses recherches. L'auteur synthétise clairement les stratégies les plus importantes dans la sélection et l'extraction des mots-clés. Plusieurs exemples permettent de mieux illustrer les propos de l'auteur, enrichissant l'exposé et donnant un caractère pédagogique indéniable à cette étude.

Ainsi, l'auteur présente la sélection des termes-clés via des n-grammes, des *chunks* nominaux ou par des approches basées sur des patrons syntaxiques combinés à des fréquences d'occurrence. Les méthodes

FACULTÉ DES SCIENCES

Institut d'informatique

Secrétariat
Rue Emile-Argand 11
CH-2000 Neuchâtel
Tél : +41 (0)32 718 27 00
Fax : +41 (0)32 718 27 01
secretariat.iiun@unine.ch

d'extraction sont clairement subdivisées en approche non supervisée ou supervisée. Dans le premier paradigme, l'auteur reprend des méthodes s'appuyant sur des statistiques simples comme la pondération *tf idf* avec ses adaptations dans le contexte de l'extraction automatique, la pondération Okapi, quelques variantes *ad hoc* basées sur ces considérations théoriques (taille des mots-clés, couverture, cohérence, position des termes, etc.). Le modèle LDA (*latent Dirichlet allocation*) constitue également une approche décrite par l'auteur. En deuxième lieu, l'auteur présente les méthodes par regroupement basées sur des similarités lexicales ou sémantiques. Enfin, les méthodes d'extraction fondées sur les graphes sont décrites, souvent en s'appuyant sur l'algorithme PageRank ou LDA.

Parmi les méthodes supervisées, l'auteur présente les classifieurs probabilistes (*naïve Bayes*) avec quelques adaptations tenant compte de la tâche à accomplir. Dans cette catégorie, on retrouve également les *conditional random fields.* D'autres méthodes supervisées ont été proposées comme les arbres de décision, les séparateurs à vastes marges, les réseaux de neurones multicouches ou des algorithmes génétiques.

Avec la présence d'un vocabulaire contrôlé, l'assignement automatique de termes-clés constitue une tâche plus complexe. L'auteur termine ce deuxième chapitre par la présentation des métriques les plus usitées dans l'évaluation de ces différentes tâches (précision, rappel et la mesure F1).

A ce stade, il convient de souligner la qualité du travail d'analyse présenté par l'auteur. Le texte est clair, l'exposé rigoureux et les termes essentiels sont clairement définis. On ne se limite pas à une simple description des diverses approches. Chaque solution est placée en regard à ses hypothèses sous-jacentes. L'auteur expose précisément les limites des stratégies proposées.

Dans le troisième chapitre, l'auteur décrit les diverses ressources utilisées dans ses travaux de recherche. Ainsi, il aura recours aux collections tests de Termith, Deft, WikiNews, SemEval et DUC. Ces collections couvrent deux langues (français, anglais) et des domaines différents (articles scientifiques, notices bibliographiques, journaux). La longueur moyenne des documents et le nombre de mots-clés associés varient aussi d'une collection à l'autre permettant à l'auteur d'évaluer ses propositions sur des fondements plus solides. Ce chapitre se termine par la description des divers outils informatiques utilisés lors du pré-traitement en langue anglaise ou française.

Dans un quatrième chapitre, l'auteur propose une stratégie originale pour la sélection des termes-clés candidats (LR-NP). Une étude préparatoire indique clairement que les syntagmes nominaux sont les plus fréquents en indexation automatique, en particulier les noms et adjectifs. Pour ces derniers, l'adjectif relationnel s'avère nettement le plus fréquent en langue française, tandis que la langue anglaise utilise fréquemment les adjectifs relationnels et qualificatifs. A côté des ces critères grammaticaux, l'auteur considère la position de l'adjectif (e.g., antéposé pour l'anglais) et la fréquence d'occurrence. Afin d'évaluer cette proposition, l'auteur la compare à trois autres approches (n-grammes, NP-*chunks*, et plus longue séquence du patron « N|A »). Sur sept collections tests, la proposition de l'auteur (LR-NP) apporte la meilleure performance en évaluation intrinsèque, et très souvent la meilleure valeur F1 (évaluation extrinsèque).

Dans un second volet, l'auteur décrit une approche novatrice dans l'extraction (non supervisée) de termes-clés nommé TopicRank. Cette solution s'appuie sur un graphe dont les sommets représentent des sujets (termes-clés candidats) et les arcs les distances entre les termes-clés candidats dans le document à indexer. Les termes-clés sont ensuite classés selon un algorithme de marche aléatoire. Finalement, la sélection des meilleurs termes-clés s'opère en fonction également de la position dans le document, la fréquence ou son aspect central. L'évaluation de TopicRank comparée à trois méthodes efficientes indique une performance souvent significativement meilleure obtenue par la proposition de l'auteur. De plus, les paramètres sous-jacents de TopicRank semblent robustes (peu de variabilité dans la performance). L'auteur a procédé à une analyse statistique de ses résultats et indique les raisons de la difficulté d'une sélection optimale.

Un cinquième chapitre aborde la question de l'indexation dans un domaine de spécialité. Dans ce cas, l'indexation s'appuie souvent sur un vocabulaire contrôlé dont les mots-clés réflètant la sémantique du texte n'apparaissent pas forcément dans le document. De manière similaire à TopicRank, la solution proposée (nommée TopicCoRank) s'appuie sur deux graphes. Le premier est construit à l'aide du document à indexer et le second en fonction de documents indexés appartenant au domaine visé (approche supervisée). Ce dernier comprend donc la terminologie du domaine. Les liens s'établissent en fonction des contextes (document ou du domaine). L'auteur présente ensuite une méthode d'ordonnancement et de sélection des termes-clés. L'évaluation compare trois méthodes efficientes reflétant l'état de l'art et deux variantes de TopicCoRank. Sur sept collections, la proposition de l'auteurp permet d'obtenir très souvent des performances statistiquement supérieures aux autres approches.

L'auteur effectue également une évaluation de TopicRank et TopicCoRank hors des domaines de spécialité. Les solutions proposées démontrent à nouveau leur efficacité mais sans pouvoir certifier clairement quelle méthode (TopicRank ou TopicCoRank) s'avère la meilleure dans tous les contextes possibles. Une des difficultés de cette tâche est d'assigner un terme-clé n'apparaissant jamais dans le document ou d'attribuer un sujet très fréquent dans toutes les indexations (e.g., *Europe*). Parmi les erreurs d'assignation, l'auteur constate également la difficulté de choisir le bon synonyme (e.g., entre *rite funéraire* et *pratique funéraire*).

A la fin de ce cinquième chapitre, l'auteur procède à une évaluation manuelle sur un sous-ensemble de la collection Termith. Cette expérience démontre les limites d'une évaluation automatique. De plus, elle met clairement en lumière la capacité de l'approche TopicRank à éviter la redondance (une des défauts majeurs de la stratégie *tf idf*).

Dans sa conclusion, l'auteur résume ses principales contributions (sélection des termes-clés, ordonnancement des termes-clés, indexation dans un domaine spécifique) et dresse les perspectives ouvertes pour des recherches futures.

UNIVERSITÉ DE
NEUCHÂTEL

L'ensemble du travail est agréable à lire, l'exposé est bien structuré et la présentation soignée. L'auteur a su rédiger de manière appropriée l'état de l'art, les diverses questions importantes, ainsi que ses contributions. Dans ce travail, l'auteur démontre qu'il connaît les principales publications dans les divers domaines de son travail de doctorat et les différentes méthodologies d'évaluation. L'exposé est clair et des exemples complètent bien les propos de l'auteur. Ce travail démontre une maîtrise des concepts liés à l'indexation automatique, et plus généralement, au traitement automatique de la langue naturelle. La méthodologie utilisée correspond à celle admise dans le domaine. L'auteur a démontré qu'il est capable de mener des recherches scientifiques, de sélectionner les collections tests appropriées, de recourir à une méthodologie reconnue et d'appliquer les tests statistiques appropriés.

Sur la base de ces considérations, je propose d'accepter le travail de thèse de Monsieur Adrien Bougouin et de l'admettre à soutenir sa thèse devant le jury désigné par la Faculté.

Tout en restant à votre disposition pour de plus amples renseignements, je vous prie de croire, Monsieur le Directeur, en mes sentiments les meilleurs.

Prof. Jacques Savoy
Institut d'informatique

Commentaires

page de titre et page de couverture. 2 fois. Mettre une majuscule à université après le nom de M. Gelgon

Page 11 , Ligne 7 à y accèdes -> à y accéder

Vous utilisez des petites capitales pour les acronymes, sauf pour la première lettre (e.g., DVD). Toute la thèse est ainsi… j'aurais opté pour les mettre toutes en capitales ou petites capitales. Pas grave

Page 12 depuis des bases de données -> depuis des entrepôts de données (afin de ne pas impliquer l'idée que c'est seulement des SGBD.)

Page 13 1er ligne. les méthodes d'indexation par termes-clés -> les méthodes d'indexation automatique par termes-clés.

Page 16. J'aurais écrit chunks en italique (mot étranger).

Page 21. sous la formule 2.4. Dire que le DL sont des longueur en nombre de mots (et non en nombre de caractères, bytes, etc.).

Page 22. Sous l'équation 2.8, indiquer que ML(candidat) et ML'(candidat) sont des probabilités (et non pas un score..)

Page 27. Bilan des méthodes non supervisés. On peut y ajouter le besoin de paramètres et de fixer des valeurs par défaut (par toujours évident.)…

Page 31 : SVM. sur un plan selon la -> sur un plan (ou hyperplan) selon la
Page 31 : SVM. puis construisent l'hyperplan -> puis construisent la droite (l'hyperplan)

Page 33. fin section 2.3. On peut ajouter que l'on doit disposer d'un ensemble pour l'entrainement et que celui ci doit être proche de celui de production..

Page 62. Ne pas terminer une page avec un mot coupé en deux. Idem en page 82.

Page 68. Légende table 4.10. est –ce bien 0,001 (un pour mille) et non 0,01 (un pourcent) ? Idem pour la légende de la table 5.2

Page 75. de l'indexations : -> de l'indexation :