



Rapport sur le mémoire présenté par Adrien Bougouin, intitulé :

«Indexation automatique par termes-clés en domaines de spécialité»

Le travail de Adrien Bougouin porte sur la définition, l'extraction et la sélection de termes clés afin de décrire le contenu d'un texte et pouvoir le retrouver dans le cadre d'un système de recherche documentaire. Les termes-clés peuvent être extraits des documents, et le problème consiste alors à ordonner des candidats selon leur pertinence. Ils peuvent aussi appartenir à une ressource décrivant le domaine, et le problème est alors de sélectionner les plus pertinents. Le choix effectué dans cette thèse porte sur l'étude d'une méthode non supervisée reposant sur une modélisation sous forme de graphe qui permette de répondre aux deux types de problème, extraction et sélection. La thèse se situe dans le cadre d'un projet destiné à fournir des outils à des documentalistes dans leur tâche d'indexation. Ce contexte a permis à A. Bougouin de disposer de collections et d'utilisateurs pour développer et évaluer les méthodes qu'il propose. Le travail proposé a été évalué sur différents types de corpus (textes dans différents domaines de spécialité et portant sur tout domaine) et différentes langues (français et anglais).

Après une brève introduction mettant en évidence les problèmes spécifiques étudiés, un état de l'art sur l'indexation automatique est proposé. Celui-ci passe en revue les différentes approches pour la définition et l'extraction de termes-clés pertinents, en insistant sur leur hypothèses, leurs apports et leurs limites. La partie concernant l'attribution de mots clés choisis parmi un ensemble (la tâche dénommée assignation) est en revanche moins développée. Y manquent des références aux travaux posant ce problème comme un problème de classification multilabel (cf. les travaux évalués par exemple dans le cadre de BioAsq, en domaine médical). En ce qui concerne l'évaluation de termes d'indexation, l'auteur reprend le paradigme d'évaluation de Cranfield, avec des mesures de précision, rappel et F-mesure. Ce chapitre permet à l'auteur de bien positionner son travail et d'introduire et justifier les points qu'il a plus spécifiquement étudiés.

Afin d'évaluer ses propositions dans un cadre le plus large possible, l'auteur a travaillé sur des corpus diversifiés, qui sont présentés dans le chapitre suivant.

Ensuite viennent les chapitres dédiés aux différentes contributions de la thèse. L'une des hypothèses de A. Bougouin est que l'on peut améliorer l'extraction de termes clés pertinents si on restreint l'ensemble des candidats extraits des documents selon des propriétés linguistiques fines. Afin de vérifier cette hypothèse, une étude linguistique des termes-clés a été menée, à partir d'une annotation des mots par leurs catégories morpho-syntaxiques, afin de mettre en évidence des caractéristiques plus fines que celles qui sont généralement appliquées. Cette étude

a porté notamment sur les types d'adjectifs employés (relationnels, composés, autres) pour lesquels l'auteur a réalisé une annotation en exploitant différentes ressources. Il en résulte une méthode de sélection de candidats dont l'impact est mesuré ensuite via différentes méthodes de sélection de termes clés. Globalement, la méthode se montre compétitive : elle permet de considérer moins de termes sans diminuer la performance des méthodes d'extraction, voire d'améliorer leurs résultats.

Ensuite, une nouvelle méthode d'extraction de termes clés pour décrire le contenu d'un texte est proposée, consistant à appliquer un algorithme d'ordonnancement sur un graphe de sujets, chacun étant ensuite étiqueté par un terme choisi. Un sujet correspond à un regroupement de termes, ici le critère appliqué est le fait de partager des mots communs, et est destiné à éviter la redondance découlant de la considération de tous les termes candidats et à donner une représentation plus synthétique du texte. Cette méthode est évaluée et comparée à différentes méthodes de l'état de l'art et obtient des résultats intéressants. Une étude approfondie en est menée en faisant varier différents paramètres et différents choix, afin de mesurer leur contribution.

Les résultats obtenus sont différents selon les corpus et les résultats sur les corpus de spécialité du projet sont en deçà de la méthode fondée sur un calcul de poids TF.IDF. Une explication donnée concerne la taille des documents, mais on aurait aimé une discussion plus approfondie de ces résultats et des sources d'erreur afin d'essayer de caractériser les conditions dans lesquelles ce type de modèle est le plus adapté. La mesure de clustering reposant uniquement sur des mots (ou leur racine) communs, il aurait aussi été intéressant d'étudier la nécessité de tenir compte de variations lexicales, de type hyperonymie notamment, ou de voisinages sémantiques pour regrouper des termes en sujet.

Vient ensuite la proposition d'un modèle d'indexation mixte, permettant de choisir des mots-clés parmi les candidats extraits du texte et ceux décrivant le domaine. A. Bougouin propose une modélisation originale faisant coopérer deux graphes de termes, l'un pour les termes extraits et l'autre pour les termes du domaine, avec des liens entre les deux selon les nœuds (les termes) qu'ils partagent. A nouveau, cette méthode a donné lieu à des évaluations approfondies. Il semble que les résultats soient assez sensibles à la manière dont est constitué le graphe de domaine, plus ou moins recentré sur des sujets liés au texte indexé.

Il aurait été intéressant d'envisager la sélection des candidats termes du domaine pour constituer le graphe en s'appuyant sur les termes attribués à des documents similaires, ainsi que cela se fait souvent dans les méthodes d'attribution de mots-clés à des textes ou des descriptions d'image.

Enfin, une évaluation utilisateur a également été mise en place, permettant d'évaluer différemment la méthode proposée.

En conclusion sur le manuscrit, celui-ci est bien rédigé, agréable à lire et particulièrement concis tout en donnant les informations pertinentes. Les choix effectués sont toujours justifiés.

En ce qui concerne le travail réalisé dans le cadre de cette thèse, j'ai particulièrement apprécié la démarche adoptée et la rigueur avec laquelle ce travail a été mené : étude linguistique des caractéristiques des termes clés en corpus, puis proposition et évaluation approfondie des méthodes proposées, reposant sur une comparaison

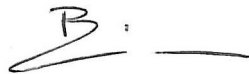
avec des méthodes de l'état de l'art et une étude de leur comportement sous différents protocoles. Les méthodes proposées sont innovantes et permettent globalement d'améliorer l'état de l'art, et les différentes remarques émises n'enlèvent rien à la qualité du travail effectué.

Je tiens à souligner tout particulièrement l'originalité de la méthode d'indexation mixte proposant de co-sélectionner termes extraits et termes du domaine modélisés en deux graphes de termes influant l'un sur l'autre, qui ouvre à mon sens de multiples perspectives.

L'ensemble des propositions a donné lieu à des publications de très bonne qualité dans des journaux et conférences nationales et internationales.

Le travail présenté porte donc sur un domaine d'importance, l'indexation de documents en domaine de spécialité, et apporte des solutions performantes qui intègrent des propositions originales. En conséquence de quoi, j'émet un avis très favorable à la présentation de ce travail pour l'obtention du grade de docteur en Informatique de l'université de Nantes Angers Le Mans.

A Orsay, le 5 octobre 2015,



Brigitte Grau,
Professeur à l'ENSIIE
LIMSI - CNRS