



**THESE / UNIVERSITE DE BRETAGNE-SUD**

*sous le sceau de l'Université Bretagne Loire*

pour obtenir le titre de  
**DOCTEUR DE L'UNIVERSITE DE BRETAGNE-SUD**

*Mention : Informatique*  
**Ecole doctorale SICMA**

Présentée par

**VU Hai Hieu**

Préparée dans l'équipe **EXPRESSION**  
Laboratoire **IRISA**

# Indexation aléatoire et similarité inter-phrases appliquées au résumé automatique

**Thèse soutenue le 29 janvier 2016**

*devant le jury composé de :*

**Pierre-François MARTEAU**

Professeur, université de Bretagne Sud / directeur de thèse

**Jeanne VILLANEAU**

MCF, université de Bretagne Sud / co-directrice de thèse

**Farida SAÏD**

MCF, université de Bretagne Sud / co-directrice de thèse

**Sophie ROSSET**

Chercheuse, LIMSI – CNRS / rapporteuse

**Emmanuel MORIN**

Professeur, université de Nantes / rapporteur

**Gwénoélé LECORVÉ**

MCF, université de Rennes 1 / examinateur

UNIVERSITE DE BRETAGNE-SUD

# *Résumé*

IRISA  
EXPRESSION

Docteur en informatique

## **Indexation aléatoire et similarité inter-phrases appliquées au résumé automatique**

par VU Hai Hieu

Face à la masse grandissante des données textuelles présentes sur le Web, le résumé automatique d'une collection de documents traitant d'un sujet particulier est devenu un champ de recherche important du Traitement Automatique des Langues. Les expérimentations décrites dans cette thèse s'inscrivent dans cette perspective. L'évaluation de la similarité sémantique entre phrases est l'élément central des travaux réalisés. Notre approche repose sur la similarité distributionnelle et une vectorisation des termes qui utilise l'encyclopédie Wikipédia comme corpus de référence. Sur la base de cette représentation, nous avons proposé, évalué et comparé plusieurs mesures de similarité textuelle ; les données de tests utilisées sont celles du défi SemEval 2014 pour la langue anglaise et des ressources que nous avons construites pour la langue française. Les bonnes performances des mesures proposées nous ont amenés à les utiliser dans une tâche de résumé multi-documents, qui met en oeuvre un algorithme de type PageRank. Le système a été évalué sur les données de DUC 2007 pour l'anglais et le corpus RPM2 pour le français. Les résultats obtenus par cette approche simple, robuste et basée sur une ressource aisément disponible dans de nombreuses langues, se sont avérés très encourageants.

# *Remerciements*

Je tiens à remercier, en tout premier lieu, mon directeur et mes co-directeurs de thèse, Monsieur le Professeur Pierre-François MARTEAU, Mesdames Jeanne VILLANEAU et Farida SAÏD pour m'avoir accueilli, guidé et mis dans les meilleures conditions pour préparer ma thèse au sein de l'équipe EXPRESSION du Laboratoire IRISA, l'Université de Bretagne-Sud. Je tiens à leur exprimer ma gratitude pour leurs qualités pédagogiques et scientifiques, leur franchise, leur sympathie, leur confiance. J'ai appris beaucoup auprès d'eux. Je leur suis également reconnaissant pour leur écoute, leur partage et leur soutien dans les moments difficiles. J'ai pris un grand plaisir à travailler sous leur direction.

Je voudrais aussi remercier les rapporteurs de cette thèse : Madame Sophie ROSSET, Directrice de Recherche du Laboratoire LIMSI, CNRS et Monsieur le Professeur Emmanuel MORIN au Laboratoire d'Informatique de Nantes-Atlantique, l'Université de Nantes pour l'intérêt qu'ils ont porté à mon travail.

Mes remerciements s'adressent également à Monsieur Gwénoél LECORVÉ de l'Université de Rennes 1 pour avoir accepté d'examiner mon travail et de participer au jury.

Je souhaite remercier tous les membres du laboratoire IRISA, Lab-STICC, ENSIBS : les enseignants, techniciens, administratifs et doctorants qui m'ont aidé et accompagné dans mon travail durant ces quatre années en France.

Je n'oublie pas non plus tous les amis de France qui nous ont aidés, ma famille et moi : Brigitte ENQUEHARD, Evelyne BOUDOU, Alain BOUDOU, Lucien MOREL, Gildas TRÉGUIER, Sylvain CAILLIBOT..., les étudiants vietnamiens et les familles vietnamiennes de Lorient.

Pour terminer, je remercie du fond du cœur mes beaux-parents NONG Quoc Chinh - TRAN Thi Doan, mes parents VU The Huan - LE Thi Nhi et tous les membres de ma famille qui m'ont toujours soutenu, tout au long de ma vie, de mes études, sans lesquels je n'en serais pas là aujourd'hui. Ma reconnaissance va surtout à mon épouse NONG Thi Quynh Tram et à nos deux enfants VU Quynh Mai et VU Hai Minh qui sont toujours à mes côtés et me donnent la force de relever les défis.



# Table des matières

Résumé	ii
Remerciements	iii
Table des matières	iv
Liste des figures	ix
Liste des tableaux	xi
<b>1 Introduction</b>	<b>1</b>
<b>2 Représentation sémantique d'un terme</b>	<b>5</b>
2.1 Quelques approches de la sémantique lexicale . . . . .	5
2.1.1 Modèles graphiques . . . . .	6
2.1.2 Modèles d'espaces vectoriels et modèles neuronaux . . . . .	7
2.1.3 Modèles géométriques . . . . .	9
2.1.4 Modèles logico-algébriques . . . . .	10
2.2 Les espaces vectoriels sémantiques . . . . .	11
2.2.1 Différentes représentations sémantiques . . . . .	12
2.2.1.1 Matrice terme-document et similarité entre docu- ments . . . . .	12
2.2.1.2 Matrice mot-contexte et similarité entre mots . . . . .	13
2.2.1.3 Matrice paire-patron et similarité relationnelle . . . . .	14
2.2.1.4 Autres représentations . . . . .	15
2.2.2 VSM et types de similarité . . . . .	16
2.3 Traitements mathématiques des VSM . . . . .	17
2.3.1 Construction de la matrice des fréquences brutes . . . . .	18
2.3.2 Pondération des fréquences brutes . . . . .	18
2.3.3 Lissage de la matrice . . . . .	23
2.3.4 Comparaison des vecteurs . . . . .	26
2.3.5 Algorithmes aléatoires . . . . .	28
2.4 Notre approche pour la représentation des mots . . . . .	29

---

2.4.1	Wikipédia comme ressource linguistique . . . . .	30
2.4.2	Random Indexing pondéré . . . . .	32
<b>3</b>	<b>Espace sémantique et sélection automatique des articles Wikipédia</b>	<b>35</b>
3.1	Les principes . . . . .	35
3.2	Construction du Web crawler . . . . .	36
3.3	Calcul de la relation entre concepts Wikipédia . . . . .	38
<b>4</b>	<b>Calculs de similarité entre phrases</b>	<b>43</b>
4.1	Introduction . . . . .	43
4.2	Similarité par définition d'un vecteur sémantique de phrase . . . . .	44
4.2.1	Expérimentations concernant les groupes de deux termes et modification des pondérations . . . . .	45
4.2.1.1	Introduction du paramètre $\alpha$ . . . . .	46
4.2.1.2	Introduction de deux paramètres : $\alpha$ et $\beta$ . . . . .	48
4.3	Similarité par optimisation des similarités entre termes . . . . .	51
<b>5</b>	<b>WikiRI et similarité entre phrases : évaluations</b>	<b>55</b>
5.1	Évaluations du calcul de similarités entre phrases : langue anglaise . . . . .	55
5.1.1	Les corpus SemEval . . . . .	56
5.1.2	Étude des paramètres $\alpha$ et $\beta$ (WikiRI <sub>1</sub> ) . . . . .	57
5.1.2.1	Introduction du paramètre $\beta$ . . . . .	58
5.1.3	Résultats obtenus par les différentes versions de WikiRI sur les corpus de SemEval 2014 . . . . .	58
5.2	Évaluations du calcul de similarités entre phrases : langue française . . . . .	61
5.2.1	Les corpus d'évaluation . . . . .	61
5.2.2	Résultats obtenus par les différentes versions de WikiRI sur les corpus de langue française . . . . .	64
5.2.2.1	WikiRI sur sélection d'articles . . . . .	64
5.2.2.2	Comparaison entre WikiRI <sub>1</sub> et WikiRI <sub>2</sub> . . . . .	66
5.3	Conclusion . . . . .	66
<b>6</b>	<b>Application de WikiRI à une tâche de résumé multi-documents</b>	<b>69</b>
6.1	Principes généraux . . . . .	69
6.2	Description de l'algorithme DivRank . . . . .	71
6.3	Expérimentations en langue française . . . . .	72
6.3.1	Le corpus de tests . . . . .	73
6.3.2	Les résultats . . . . .	74
6.4	Expérimentations en langue anglaise . . . . .	75
6.4.1	Les données de test . . . . .	76
6.4.2	Les résultats de WikiRI <sub>1</sub> . . . . .	76
6.5	Conclusion . . . . .	78
<b>7</b>	<b>Bilan et perspectives</b>	<b>79</b>
7.1	Objectifs initiaux et déroulement des travaux . . . . .	79

7.2	Bilan . . . . .	80
7.3	Pistes d'amélioration et perspectives . . . . .	81
<b>A Liste des publications</b>		<b>85</b>
<b>Bibliographie</b>		<b>87</b>



# Table des figures

2.1	Pondérations $TF$ . . . . .	20
2.2	Pondération $BM25$ . . . . .	20
2.3	Pondération $IDF$ . . . . .	21
2.4	Normalisation pivot de la longueur des documents . . . . .	21
2.5	Structure en noeud-papillon de Wikipédia . . . . .	30
3.1	SourceWikipedia . . . . .	38
3.2	Wikipedia Graph . . . . .	40
4.1	Valeur de $\log\left(\frac{N+1}{n_i+1}\right)^\alpha$ en fonction du taux de documents qui contiennent le terme pour différentes valeurs de $\alpha$ . . . . .	47
4.2	Logarithme décimal du nombre de termes en fonction de leur taux d'apparition dans les articles du Wikipédia français . . . . .	49
4.3	Logarithme décimal du nombre de termes en fonction de leur taux d'apparition dans les articles du Wikipédia anglais . . . . .	50
4.4	Valeurs de $l'icf_{\alpha,\beta}$ en fonction du taux de documents qui contiennent le terme pour différentes valeurs de $\beta$ avec $\alpha = 3$ . . . . .	51
5.1	Tool . . . . .	64

