



Karim Sayadi

Formations

- 2016–2017 **Attaché temporaire d'enseignement et de recherche**, *EPHE*, Ecole Pratique des Hautes Etudes, Paris, France.
- 2013–2016 **Thèse de doctorat en informatique**, *UPMC*, Université Pierre et Marie Curie, Paris, France.
- 2011–2013 **Master Cognitions naturelle et artificielle. Étude des systèmes complexes**, *Ecole Pratique des Hautes Études*, Paris, France.
Spécialité informatique. Mention très bien
- 2008–2011 **Licence fondamentale en informatique.**, *Institut d'informatique et de mathématiques*, Monastir, Tunisie.
Mention bien.

Sujet de thèse

Titre *Construction d'algorithmes d'apprentissage automatique pour l'analyse des documents.*

Description Différentes disciplines des sciences humaines telles la philologie ou la paléographie font face à des tâches complexes, fastidieuses et consommatrices en temps pour l'examen des sources de données. La proposition d'approches computationnelles en humanités permet d'adresser les problématiques rencontrées telles que la lecture, l'analyse et l'archivage de façon systématique. Les modèles conceptuels élaborés reposent sur des algorithmes et ces derniers donnent lieu à des implémentations informatiques qui automatisent ces tâches fastidieuses. La première partie de la thèse vise, d'une part, à établir la structuration thématique d'un corpus, en construisant des espaces sémantiques de grande dimension. D'autre part, elle vise au suivi dynamique des thématiques qui constitue un réel défi scientifique, notamment en raison du passage à l'échelle. Une contribution de cette thèse concerne la distribution de ces calculs sur une grille d'ordinateurs, avec des outils tels que Spark, Hadoop et MapReduce. La seconde partie de la thèse traite de manière holistique la page d'un document numérisé sans aucune intervention préalable. Le but est d'apprendre automatiquement des représentations du trait de l'écriture ou du tracé d'un certain script (langue écrite) par rapport au tracé d'un autre script. Il faut dans ce cadre tenir compte de l'environnement où se trouve le tracé : image, artefact, bruits dus à la détérioration de la qualité du papier, etc. Notre approche propose un empilement de réseaux de neurones auto-encodeurs afin de fournir une représentation alternative des données reçues en entrée. Les résultats obtenus avec les représentations apprises par les réseaux de neurones sont comparables à ceux fabriqués par des experts. De plus, nous proposons un retour d'expériences autour de l'initialisation des paramètres d'apprentissage en vue de pistes d'améliorations futures. Les contributions de ce travail concernent la conception d'algorithmes pour la classification des textes numériques et numérisés. La classification des textes numériques est opérée à l'aide d'un modèle basé sur des espaces sémantiques quant à la classification des textes numérisée, elle s'appuie sur le développement d'architectures de réseaux de neurones.

Mots clés *Topic Modeling, Probabilistic Graphical Model, Machine Learning, Deep Learning, Digital Humanities, Natural Language Processing*

Directeur Professeur Marc Bui, France

Autres formations

- Sept 2015 Ecole d'été du Helmholtz-Zentrum Berlin sur l'analyse des manuscrits anciens, Frauenchiemsee, Germany (1 semaine).
- Juil 2014 Ecole d'été du labex Dynamite sur les systèmes complexes et la modélisation des territoires, Florence, Italie (1 semaine).

Jan-Juil 2013 Semestre Michel Serres, heSam, École des Arts et Metier, École Pratique des Hautes Études. Sujet de travail : proposition d'un design pattern d'une ontologie interdisciplinaire. Paris, France. (6 mois)

Activités d'enseignement

- 2015-2016 TD sur l'utilisation de \LaTeX . École Pratique des Hautes Études (8h). Formation humanités numériques EPHE.
- 2014-2015 TD sur l'utilisation de processing. École Pratique des Hautes Études (12h). Formation humanités numériques EPHE.
- 2014-2015 TD sur la modélisation avec UML. École Pratique des Hautes Études (8h). Master 1 CNA-PC.
- 2014-2015 TD sur le traitement des informations spatial avec python. Université Paris 8 (16h). Master 1 géomatique.
- 2013-2014 Cours sur les méthodes d'apprentissage automatique pour l'analyse des corpus de textes. École Pratique des Hautes Études (8h). Master 2 CNA-PC.

Activités de recherche

- Fév-Juin 2016 Séjour de recherche au sein du groupe de recherche sur l'analyse de documents, images et voix. Université de Fribourg, Suisse. (Pr. Rolf Ingold, Pr. Marcus Liwicki) (5 mois)
- Nov-Dec 2015 Séjour de recherche au sein du groupe d'analyse des données et calcul haute performance. Institut John Von Neumann. Université Nationale de Ho Chi Minh Ville, Vietnam. (Pr. Vu Duong) (1 mois)
- Sept 2015 Workshop sur l'analyse des documents historiques. Centre de recherche allemand pour l'intelligence artificielle. Kaiserslautern, Allemagne. (Pr. Marcus Liwicki) (1 semaine)
- Dec 2014 Workshop sur la modélisation du socio écosystème de l'île de Moorea. Université de Berkeley, Californie, États-Unis d'Amérique. (Dr. Joachim Claudet) (1 semaine)
- Nov-Dec 2014 Séjour de recherche au sein du groupe de recherche sur les technologies d'information et de communication. Institut polytechnique de Hanoi, Vietnam. (Associate Pr. Ha Quoc Trung) (1 mois)

Travaux de recherche

Mes travaux de recherche s'inscrivent dans le cadre du Knowledge Data Discovery (KDD) où il s'agit de fournir à l'utilisateur l'information pertinente en rendant efficace la phase de recherche d'informations.

Travaux effectués

Dans le cadre de ma thèse, j'ai développé des applications pour la détection et la visualisation sous la forme de réseau complexe, des similarités thématiques entre les différents documents dans un corpus de textes. Les documents sont représentés par des sommets et les liens thématiques entre les différents documents sont représentés par des arcs. Les méthodes employées permettent l'analyse de grands corpus de documents grâce à une distribution du calcul sur une grappe d'ordinateurs. Elles sont en mesure de traiter des tâches de catégorisations thématiques, multilingues, multi domaines, mais l'originalité de l'approche proposée consiste à aborder l'aspect dynamique des contenus à analyser en vue de produire une expertise automatisée de qualité. Il s'agit par exemple de détecter les centres d'intérêt exprimés sur la Blogosphère mondiale ou les corpus Twitter pour un suivre les tendances.

Travaux en cours de réalisation

Dans le cadre d'une collaboration scientifique avec l'université de Fribourg en Suisse, je travaille actuellement sur des solutions pour automatiser complètement ou partiellement la tâche de la reconnaissance des styles d'écritures dans un document ancien. J'approche ce problème avec une structure d'apprentissage profonde implémenté avec les réseaux de neurones.

Publications

- Avr 2017 Quang Vu Bui, Karim Sayadi and Marc Bui,. *Combining Latent Dirichlet Allocation and K-means for Documents Clustering : Effect of Probabilistic Based Distance Measures*, To appear in ACIIDS proceedings Avril 2017 intl conference. Kanazawa, Japon, 3-6 Avril 2017.
- Avr 2016 Karim Sayadi, Marcus Liwicki, Rolf Ingold, Marc Bui,. *Tunisian Dialect and Modern Standard Arabic Dataset for Sentiment Analysis : Tunisian Election Context*, IEEE-CICLing (Computational Linguistics and Intelligent Text Processing) Intl. conference, Konya, Turkey, 7-8 Avril 2016.

- Dec 2015 Quang Vu Bui, Karim Sayadi, Marc Bui. *A multi-criteria document clustering method based on topic modeling and pseudoclosure function*, ACM-SOICT (Symposium on Information and Communication Technology) Intl. conference, Hue, Vietnam, 3-4 December 2015. (Extended version submitted to informatica journal)
- Juil 2015 Karim Sayadi, Quang Vu Bui, Marc Bui. *Multilayer classification of web pages using Random Forest and semi-supervised version of the Latent Dirichlet Allocation*, IEEE-I4CS (International Conference on Innovations for Community Services) Intl. conference, Nuremberg, Germany. 8-10 July 2015.
- Juil 2013 Karim Sayadi, Marc Bui, Michel Lamure. *Predictive topic modeling : Complex Networks approach using dynamics of author's communities*, EURO INFORMS (Operational Research), Rome, Italy, 1-4 July 2013.
- Mai 2012 Karim Sayadi, Marc Bui, Vigile Hoareau, Sofian Ben Amor, *Une approche prétopologique pour la catégorisation des données de microblogging*, Conférence nationale VSST'12 (Veille Scientifique et Technologique), Ajaccio, Corse, France, 24-25 Mai 2012.
- Rapport de recherche**
- Oct 2013 Marc Bui, Karim Sayadi. *Modèle de recommandation de contacts basé sur l'analyse thématique des échanges*. Rapport du projet n 45 Nexboo. Convention Techno Pole de la Réunion. 20 Octobre 2013

Présentation et séminaires

- Nov 2015 Towards an automatic annotation in the computer science ontologies. University El Manar et Institute Pasteur. Tunis, Tunisie. 19-20 November 2015
- Sept 2015 Poster : Multilayer classification of the text in ancient documents. Frauenchiemsee, Germany. 10 September 2015
- Dec 2014 Data science, and modeling tools in investigating social-ecological systems. IDEA project, University of California Berkeley, California, USA. 9 December 2014
- Juin 2014 Modeling the Seine Estuary with an ontology, Global Estuaries Forum, Deauville, France. 1 July 2014
- Oct 2013 Identification des thématiques dans les corpus de textes non structurés : application à un corpus santé (QALY), Conférence intl, VSST 2013, Nancy, France, 24 Octobre 2013.

Projets de recherche

- Dec 2014 Moorea IDEA. <http://mooreaidea.org>.
- Fév 2016 HisDoc DIVA Group <https://diuf.unifr.ch/main/diva/research/research-projects/hisdoc-2-towards-computer-assisted-palaeography>

Expériences professionnelles

- Juil-Oct 2013 Système de recommandation d'amis basé sur l'analyse des interactions sur les réseaux sociaux. Nexboo. Paris, France (4 mois)
- Fév-Juil 2013 Construction d'un référentiel conceptuel pour la recherche d'informations autour de l'estuaire de la Seine. École des Arts et Métiers. Paris, France (6 mois)
- Sept-Déc 2012 Approche prétopologique pour la catégorisation de textes. École Pratique des Hautes Études. Stage effectué au sein du laboratoire Laisc. Paris, France (4 mois)
- Fév-Juil 2012 Analyse des réseaux sociaux et visualisation des interactions entre utilisateurs. Semdee. Stage effectué au sein du laboratoire Laisc. Paris, France (6 mois)
- Juil-Aout 2010 Construction de base de données virtuelles. Metro Group IT. Düsseldorf, Allemagne (2 mois)

Compétences informatiques

- Langage PYTHON, JAVA, R, HTML, C
- Utilitaires LaTeX, BibTex, Ipe
- Logiciels Matlab, Anylogic, Scilab, Scikit learn, Hadoop, Spark, Lucene
- Systèmes UNIX, Linux, Microsoft Windows
- Production de logiciels Pretopological Semantic Analyzer (PSA for health). Déposé à l'agence de protection des programmes le 18/10/2012.

Langues

Langue maternelle **Arabe, Français**
Maîtrisée **Anglais**
Basique **Allemand**

Vie associative & centres d'intérêt

2012–2014 Responsable logistique et événementiel pour les activités culturelles à la cité internationale universitaire de Paris.

2010 Fondateurs et président d'honneur du club des logiciels libres à L'ISIMM Monastir

Centres d'intérêt

- Cuisine
- Water Polo, cyclisme
- Philosophie