Avis de Soutenance

Monsieur Wenjun SUN

Spécialité : Informatique et Applications

Soutiendra publiquement ses travaux de thèse intitulés

« Calcul de fonctionnalités de blocs de texte multimodaux et détection de communauté pour la séparation des articles de presse historiques »

dirigés par Monsieur Antoine DOUCET et LABAHN ROGER

Soutenance prévue le **mardi 30 septembre 2025 à 14h00**

Lieu : la Rochelle Université
45 Rue François de Vaux de Foletier,
Bâtiment Droit
Salle : Amphi RIVERO -

Composition du jury proposé

M. Antoine DOUCET	Université de La Rochelle	Directeur de thèse
Mme Aurelie NEVEOL	CNRS	Examinatrice
Mme Alicia FORNES	CVC	Examinatrice
M. Roger LABAHN	University of Rostock	Co-directeur de thèse
M. Apostolos ANTONACOPOULOS	University of Salford	Rapporteur
M. Gaël LEJEUNE	Sorbonne Université	Rapporteur
M. Carlos-Emiliano GONZALEZ-GALLARDO	Université de Tours	Invité
M. Coustaty MICKAEL	La Rochelle Université	Invité

Résumé :

Les journaux constituent des sources d'information exhaustives sur les événements culturels, politiques et sociaux, surpassant les autres archives publiques par l'étendue de leur couverture. Depuis leur apparition au XVIIe siècle, les journaux ont quotidiennement documenté d'innombrables événements, récits et personnalités dans presque toutes les langues et dans divers pays. Entre le milieu du XIXe siècle et le milieu du XXe siècle, ils sont devenus le premier média de masse, exerçant une influence considérable sur l'opinion publique. Tout au long de l'histoire, les journaux ont toujours joué un rôle central dans la diffusion des points de vue publics et politiques, des créations littéraires, des essais et des expressions artistiques. Cette richesse thématique les place au premier plan pour toute personne intéressée par le patrimoine culturel européen. Cependant, l'extraction automatique d'informations à partir de ces journaux reste un défi à relever. Cette difficulté provient principalement de l'analyse de documents détériorés par le temps et de la mise en page particulière des journaux historiques, qui diffère de celle des journaux contemporains, lesquels ont été au centre des recherches existantes. Dans le contexte des journaux historiques, les informations sont généralement organisées sous forme d'articles. Par conséquent, la tâche principale consiste à séparer l'ensemble du document en articles distincts correspondant à des informations individuelles. Sur la base de ce qui précède, nous proposons la séparation des articles, qui consiste à segmenter une page de journal en articles individuels. La séparation des articles est un problème intrinsèquement complexe, multimodal et interdisciplinaire. Une solution idéale nécessite le traitement simultané d'informations multidimensionnelles, notamment le contenu textuel, la structure de la mise en page et les caractéristiques visuelles. À la lumière de l'état actuel de la recherche et des défis liés à la séparation des articles, cet article propose des méthodes plus universelles et applicables pour la séparation des articles de journaux historiques. Le processus de lecture d'un texte ou d'un journal peut être considéré comme la construction d'un flux d'informations textuelles, où le texte véhicule des informations et suit l'ordre de lecture. Les frontières entre les différents articles représentent alors des points de rupture dans ce flux d'informations. Ainsi, cet article définit la tâche de séparation des articles comme suit : reconstruire l'ordre de lecture pour obtenir le flux d'informations, puis identifier les points de transition sémantique au sein de ce flux afin de le segmenter. Les paragraphes connectés restant dans le flux segmenté sont considérés comme un seul article, ce qui permet d'achever la tâche de séparation des articles. En conceptualisant la séparation des articles comme la reconstruction et la segmentation du flux d'informations, ce travail décompose la tâche en deux sous-tâches : (1) reconstruire l'ordre de lecture et (2) détecter les points de transition sémantique. Cet article aborde ces deux aspects : premièrement, comment identifier les points de segmentation sémantique à partir d'un ordre de lecture connu, et deuxièmement, comment reconstruire l'ordre de lecture à partir des caractéristiques multimodales du texte. De plus, des travaux connexes ont montré que les modèles linguistiques modernes sont peu performants dans l'extraction sémantique à partir de textes historiques. Afin d'améliorer la compréhension sémantique des textes historiques par le modèle linguistique, cette étude explore également des techniques d'amélioration sémantique au niveau des tokens et des paragraphes.