Sujet de thèse ICCARE-LAB

# Extraction et structuration automatique frugale d'événements culturels

Laboratoire Informatique d'Avignon / Laboratoire des Sciences du numérique de Nantes

**Direction de thèse :** Richard Dufour, richard.dufour@univ-nantes.fr

L'information décrivant les événements culturels (concerts, pièces de théâtre, expositions, etc.) est disponible en abondance, cependant, elle souffre de deux problèmes majeurs. D'une part, elle est éparpillée sur les dizaines de milliers de sites Web des acteurs culturels (artistes, troupes, salles de spectacle, musées, festivals, etc.) ; d'autre part sa forme et sa structure ne sont absolument pas standardisées. En conséquence, pour faire connaître leurs événements dans des médias (journaux, guides et applications culturels), les acteurs culturels sont aujourd'hui forcés de recopier manuellement les informations de leurs événements dans autant de formulaires différents que de médias. C'est un processus long, fastidieux, et improductif, limitant fortement la diffusion de l'information culturelle, et par conséquent l'accès des publics à l'offre culturelle.

L'objectif de la thèse est de développer un système automatique capable (i) de détecter la présence d'une liste d'événements culturels sur le site web d'un acteur culturel, (ii) d'y identifier chacun des événements individuels, (iii) d'extraire de cette liste (et généralement d'une page de détail) et de hiérarchiser par degré de pertinence les éléments descriptifs de l'événement (titre, image, date, lieu, description, mots-clefs, prix, lien de réservation, etc.) et (iv) de comparer ces informations entre elles pour détecter les doublons et décoder comment consolider ces informations.

Les secteurs des médias et de la culture ayant des contraintes budgétaires fortes, il faut trouver une solution frugale, avec des solutions ad hoc qui ne recourent pas à des services tiers (par exemple des LLM). En outre, la méthode proposée doit minimiser le risque d'erreur et être exempte d'hallucination, car il ne saurait être question de diffuser des informations portant sur des événements fictifs.

Du point de vue informatique, les tâches à effectuer pour remplir ces objectifs relèvent du traitement automatique des langues naturelles (TALN) et de la recherche d'information (reconnaissance d'entités nommées, désambiguïsation d'entités, extraction d'information), appliqués à des contenus Web non structurés, en tirant parti non seulement du contenu textuel des pages Web, mais aussi de leur structure. La thèse mettra en œuvre deux approches, afin de les comparer à la solution technique existante (manuelle), et entre elles. La première approche consiste à construire un modèle capable d'extraire la structure de l'information de chaque site, et d'ensuite exploiter cette structure pour lire les événements de ce site. La seconde approche consiste à entraîner le modèle à lire directement les événements d'une grande variété de sites (sans passer par la recherche de la structure de l'information sur un site donné). Ces deux approches seront testées sur des ensembles de sites Web d'acteurs culturels présélectionnés pour leur niveau de difficulté croissante. Le travail portera d'abord sur l'extraction d'évènements en soi (ii et iii ci-dessus), puis, si possible sur les sujets i et iv.

La thèse sera menée en partenariat avec ideactiv, plateforme française qui indexe aujourd'hui plusieurs milliers de sites web d'acteurs culturels. D'une part, ideactiv fournira des données (sites Web, pages d'événements, données extraites) en abondance, et des outils de bas niveau (comme la détection de dates, villes, adresses, etc.), et d'autre part, elle exploitera les résultats de la thèse pour offrir ce service d'indexation culturelle aux acteurs culturels français.



RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030 PROGRAMME DE RECHERCHE INDUSTRIES CULTURELLES ET CRÉATIVES

cnrs

anr