THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
ÉCOLE NATIONALE SUPÉRIEURE DES MINES D'ALÈS (IMT MINES ALÈS)

En Informatique

École doctorale Risques et société

Centre de recherche LGI2P de l'IMT Mines Ales
Equipe d'accueil CHROME de l'Université de Nîmes

Méthodes D'Analyse Sémantique De Corpus De Décisions Jurisprudentielles

Présentée par Gildas TAGNY NGOMPÉ

Le 24 Janvier 2020

Sous la direction de Stéphane MUSSARD
Et Jacky MONTMAIN

Devant le jury composé de

Sandra BRINGAY, Professeur, Université Paul Valéry Montpellier	Rapporteur
Mohand BOUGHANEM, Professeur, Université Toulouse III Paul Sabatier	Rapporteur
Françoise SEYTE, Maître de Conférences (HDR), Université de Montpellier	Examinateur
Fabrice MUHLENBACH, Maître de Conférences, Université Jean Monnet de Saint-Étienne	Examinateur
Stéphane MUSSARD, Professeur, Université de Nîmes	Directeur de thèse
Jacky MONTMAIN, Professeur, IMT Mines Alès	Co-directeur de thèse
Guillaume ZAMBRANO, Maître de Conférences, Université de Nîmes	Encadrant de proximité
Sébastien HARISPE, Maître Assistant, IMT Mines Alès	Encadrant de proximité

Collège Doctoral Languedoc-Roussillon

IMT Mines Alès
École Mines-Télécom

Résumé

Titre : MÉTHODES D'ANALYSE SÉMANTIQUE DE CORPUS DE DÉCISIONS JURISPRUDENTIELLES

Une jurisprudence est un corpus de décisions judiciaires représentant la manière dont sont interprétées les lois pour résoudre un contentieux. Elle est indispensable pour les juristes qui l'analysent pour comprendre et anticiper la prise de décision des juges. Son analyse exhaustive est difficile manuellement du fait de son immense volume et de la nature non-structurée des documents. L'estimation du risque judiciaire par des particuliers est ainsi impossible car ils sont en outre confrontés à la complexité du système et du langage judiciaire. L'automatisation de l'analyse des décisions permet de retrouver exhaustivement des connaissances pertinentes pour structurer la jurisprudence à des fins d'analyses descriptives et prédictives. Afin de rendre la compréhension d'une jurisprudence exhaustive et plus accessible, cette thèse aborde l'automatisation de tâches importantes pour l'analyse métier des décisions judiciaires. En premier, est étudiée l'application de modèles probabilistes d'étiquetage de séquences pour la détection des sections qui structurent les décisions de justice, d'entités juridiques, et de citations de lois. Ensuite, l'identification des demandes des parties est étudiée. L'approche proposée pour la reconnaissance des quanta demandés et accordés exploite la proximité entre les sommes d'argent et des termes-clés appris automatiquement. Nous montrons par ailleurs que le sens du résultat des juges est identifiable soit à partir de termes-clés prédéfinis soit par une classification des décisions. Enfin, pour une catégorie donnée de demandes, les situations ou circonstances factuelles où sont formulées ces demandes sont découvertes par regroupement non supervisé des décisions. A cet effet, une méthode d'apprentissage d'une distance de similarité est proposée et comparée à des distances établies. Cette thèse discute des résultats expérimentaux obtenus sur des données réelles annotées manuellement. Le mémoire propose pour finir une démonstration d'applications à l'analyse descriptive d'un grand corpus de décisions judiciaires françaises.

Mots-clés : analyse de données textuelles, décisions jurisprudentielles, extraction d'information, classification de textes, regroupement non-supervisé.

Abstract

Title : METHODS OF SEMANTIC ANALYSIS OF CORPORA OF CASE LAW DECISIONS

A case law is a corpus of judicial decisions representing the way in which laws are interpreted to resolve a dispute. It is essential for lawyers who analyze it to understand and anticipate the decision-making of judges. Its exhaustive analysis is difficult manually because of its immense volume and the unstructured nature of the documents. The estimation of the judicial risk by individuals is thus impossible because they are also confronted with the complexity of the judicial system and language. The automation of decision analysis enable an exhaustive extraction of relevant knowledge for structuring case law for descriptive and predictive analyses. In order to make the comprehension of a case law exhaustive and more accessible, this thesis deals with the automation of some important tasks for the expert analysis of court decisions. First, we study the application of probabilistic sequence labeling models for the detection of the sections that structure court decisions, legal entities, and legal rules citations. Then, the identification of the demands of the parties is studied. The proposed approach for the recognition of the requested and granted quanta exploits the proximity between sums of money and automatically learned key-phrases. We also show that the meaning of the judges' result is identifiable either from predefined keywords or by a classification of decisions. Finally, for a given category of demands, the situations or factual circumstances in which those demands are made, are discovered by clustering the decisions. For this purpose, a method of learning a similarity distance is proposed and compared with established distances. This thesis discusses the experimental results obtained on manually annotated real data. Finally, the thesis proposes a demonstration of applications to the descriptive analysis of a large corpus of French court decisions.

Keywords : textual data analysis, case law decisions, information extraction, text classification, document clustering.