

# Introduction générale

---

## i Contexte et motivations

Une décision judiciaire peut être définie soit comme le résultat rendu par les juges à l'issue d'un procès, soit comme un document décrivant une affaire judiciaire. Un tel document rapporte, notamment, les faits, les procédures judiciaires antérieures, le verdict des juges, et les explications associées. Dans cette thèse, nous désignons par « décision » le document, et par « résultat » une conclusion ou réponse des juges. Une jurisprudence est un ensemble de décisions rendues par les tribunaux. Elle représente la manière dont ces derniers interprètent les lois pour résoudre un problème juridique donné (type de contentieux). Les juristes doivent alors collecter des décisions traitant de situations similaires, les sélectionner, et les analyser afin de mener, par exemple, des recherches empiriques en droit [Ancel, 2003; Jeandidier & Ray, 2006]. Les avocats exploitent aussi les décisions passées pour anticiper les résultats des juges. Ils peuvent ainsi mieux conseiller leurs clients sur le risque judiciaire que ces derniers encourent, et sur la stratégie à adopter pour faire accepter leurs demandes et faire rejeter celles de leurs adversaires. Cette activité de collecte et d'analyse, centrale pour de nombreux métiers du droit, est généralement effectuée manuellement. Elle est par conséquent sujette à plusieurs difficultés liées à l'accès et à l'exhaustivité des documents traités même lors de l'étude d'une question spécifique. Il faut notamment souligner ici que les documents sont dispersés dans les nombreux tribunaux, et que les procédures administratives ne facilitent pas toujours leur accès du fait de la nécessité de préserver la confidentialité des parties. En effet, les décisions n'étant pas « anonymisées » la plupart du temps, elles restent alors inaccessibles aux juristes qui en font la demande. Un certain nombre de documents sont néanmoins accessibles sur internet grâce à des sites de publication de don-

nées ouvertes gouvernementales<sup>1</sup>. Ces sites publient régulièrement des décisions récemment prononcées.

(a) Formulaire de Légifrance.

(b) Formulaire de Dalloz.

Figure 1 – Exemples de critères des moteurs de recherche juridique.

Il existe aussi des moteurs de recherche juridiques qui permettent de retrouver des décisions intéressantes. Cependant, qu'ils soient payants (Lexis-

1. Données ouvertes gouvernementales : data.gouv.fr en France, judiciary.uk en Grande-Bretagne, scotusblog.com aux Etats-Unis, et scc-csc.ca au Canada.

Nexis<sup>2</sup>, Dalloz<sup>3</sup>, Lamyline<sup>4</sup>,...) ou gratuits (CanLII<sup>5</sup>, Légifrance<sup>6</sup>, ...), les critères de recherche offerts par leurs moteurs de recherche limitent grandement la pertinence des résultats pouvant être obtenus. En effet, il ne s'agit en général que de combinaisons de mots-clés et autres méta-données (date, type de juridiction, ...), ou d'expressions régulières, comme l'illustre la Figure 1 Page 2. La manipulation de tels critères est difficile pour constituer des échantillons pertinents suivant une sémantique souhaitée tels que l'ensemble des décisions traitant d'une catégorie de demande ou d'une circonstance factuelle donnée.

Justice	2013	2014	2015	2016	2017
civile	2 761 554	2 618 374	2 674 878	2 630 085	2 609 394
pénale	1 303 469	1 203 339	1 206 477	1 200 575	1 180 949
administrative	221 882	230 477	228 876	231 909	242 882

Source : <http://www.justice.gouv.fr/statistiques-10054/chiffres-cles-de-la-justice-10303/>

Tableau 1 – Nombre de décisions prononcées en France par an de 2013 à 2017.

Plus de 4 millions de décisions sont prononcées en France chaque année d'après les chiffres du ministère français de la justice (Tableau 1 Page 3). Dans ce contexte, l'analyse manuelle ne peut être limitée qu'à une infime proportion de documents disponibles. En effet, au regard de la croissance rapide du nombre de décisions, même une étude sur une question très précise nécessite la constitution d'un large corpus de décisions pertinentes. Par ailleurs, il peut s'avérer très pénible de les lire pour en identifier les données d'intérêt. Les documents sont très souvent longs et complexes dans leur style de rédaction. Par exemple, Certaines phrases comprennent plusieurs clauses discutant d'aspects différents (Figure 2 Page 4). On y retrouve aussi des références à des jugements antérieurs (Figure 3 Page 4).

Il est évident qu'une automatisation du traitement des corpus de décisions s'impose pour répondre aux diverses difficultés d'accès, de volumétrie, et de complexité liées à la compréhension des décisions. Une telle

2. <https://www.lexisnexis.fr/>

3. <http://www.dalloz.fr>

4. <http://lamyline.lamy.fr>

5. <https://www.canlii.org>

6. <https://www.legifrance.gouv.fr>

69 Exposant subir un trouble anormal de voisinage pour être privée d'une vue sur la  
 70 mer dont elle disposait auparavant, ce en raison de l'absence de taille de haies  
 71 implantées à proximité de son jardin privatif, elle a attiré devant le juge des  
 72 référés du tribunal de grande instance de Marseille, le syndicat des  
 73 copropriétaires LES CATALANS (ci-après désigné : le syndicat des copropriétaires)  
 74 , et son syndic recherché personnellement, le Cabinet L., à l'effet, au visa de  
 75 l'article 809 du code de procédure civile d'obtenir leur condamnation sous  
 76 astreinte de 200 euros par jour de retard à tailler les haies qui bouchent sa  
 77 vue et la condamnation personnelle du Cabinet L. à lui régler une provision de  
 78 2.000 euros à valoir sur dommages et intérêts, outre 1.500 euros sur le  
 79 fondement des dispositions de l'article 700 du code de procédure civile.

*Source : extrait de la décision R.G. 15/10226 de la Cour d'Appel d'Aix-en-Provence du 2 Juin 2016*

Figure 2 – Exemple de phrases composée de plusieurs clauses dont une demande de condamnation sous astreinte, une autre de dommages et intérêts pour trouble anormal de voisinage, et une dernière de dommages et intérêts sur l'article 700 du code de procédure civile.

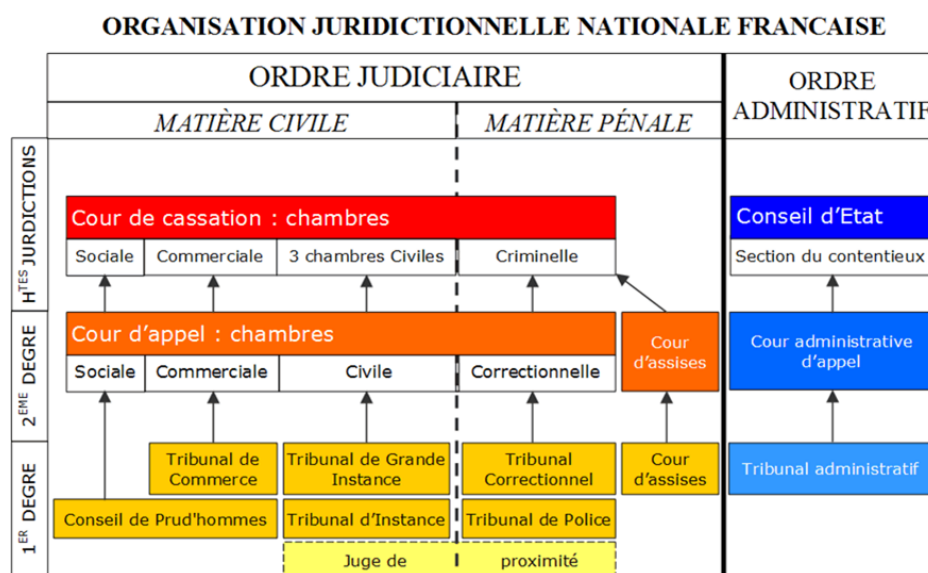
73 Vu le jugement du tribunal de grande instance de Versailles du 5 décembre 2013  
 74 qui a :  
 75 - rejeté la demande de démolition de la construction litigieuse,  
 ...  
 119 SUR CE LA COUR,  
 ...  
 278 PAR CES MOTIFS,  
 ...  
 281 Confirme le jugement en toutes ses dispositions à l'exception de celle relative  
 282 au montant des dommages et intérêts ...

*Source : extrait de la décision R.G. 14/01640 de la Cour d'Appel de Versailles du 7 Avril 2016*

Figure 3 – Exemple de référence à un jugement antérieur dans une décision d'appel.

automatisation ferait gagner du temps aux juristes lors de tâches d'analyse métier préalables à leur raisonnement d'experts, tout en leur fournissant une vue exhaustive de la jurisprudence. D'autre part, Cretin [2014] fait remarquer que la justice est complexe dans son organisation (Figure 4 Page 5) et son fonctionnement, et que son langage est peu compréhensible. Il est donc presque impossible pour les profanes en droit d'estimer leurs droits et le risque judiciaire qu'ils encourent dans leur quotidien sans consul-

ter un initié du droit. L'exigence pour le profane étant l'exacte pertinence des ressources, leur accessibilité, et l'intuitivité du processus de leur exploitation [Nazarenko & Wyner, 2017], l'automatisation de l'analyse de la jurisprudence pourrait ainsi améliorer l'accessibilité du droit dans d'innombrables situations. Par exemple, en comparant le montant qu'on peut espérer d'une juridiction et le coût d'un procès, on peut plus aisément se décider entre un arrangement à l'amiable et la poursuite du litige en justice [Langlais & Chappe, 2009]. Le traitement automatique de la jurisprudence constituerait alors une aide précieuse non seulement pour les professionnels du droit, mais aussi pour les particuliers et entreprises tous soucieux de voir l'issue de leur affaire leur être favorable.



Source : [https://fr.wikipedia.org/wiki/Organisation\\_juridictionnelle\\_\(France\)](https://fr.wikipedia.org/wiki/Organisation_juridictionnelle_(France))

Figure 4 – Organisation de la justice en France.

## ii Objectifs

Ce mémoire propose des approches pour automatiser l'extraction de connaissances judiciaires à partir des décisions françaises. Le but est de faciliter la structuration et l'analyse descriptive et prédictive de corpus de décisions de justice en adressant les difficultés de l'approche traditionnelle

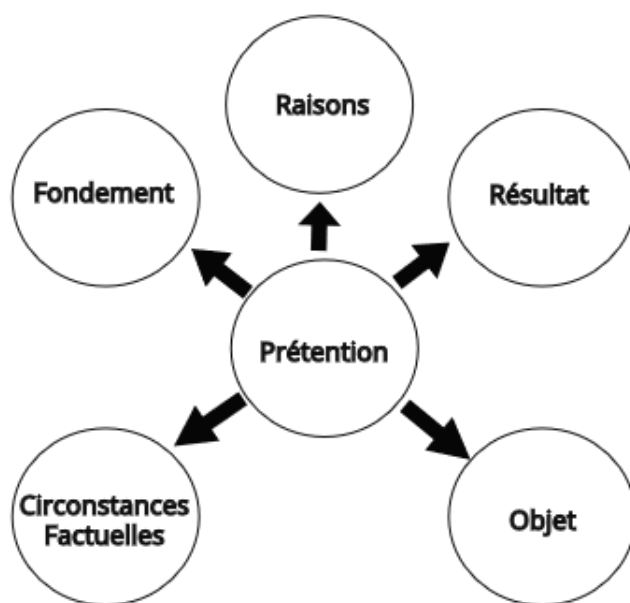
d'analyse de contentieux. L'étude de la jurisprudence pour un contentieux donnée consiste à [Ancel, 2003] :

1. **Choisir un échantillon représentatif** : Des décisions sont collectionnées suivant des contraintes définies : période précise, couverture géographique, types d'affaires, etc.
2. **Sélectionner les décisions** : élimination des décisions qui ne correspondent pas au type de demande d'intérêt.
3. **Élaborer la grille d'analyse** : Un modèle de grille est créé et permet d'enregistrer les informations potentiellement importantes. Chaque ligne de la grille correspond à une demande, et les colonnes font référence aux différents types d'informations qu'il est possible d'extraire sur une demande. Ces variables vont de la procédure suivie, aux solutions proposées, en passant par la nature de l'affaire. Les champs à remplir ne sont pas connus à l'avance ; ce n'est généralement qu'au cours de la lecture des décisions que l'on distingue les informations pertinentes pour l'étude.
4. **L'analyse des décisions et l'interprétation des informations** : Les informations retrouvées dans les décisions sont saisies dans la grille, et des calculs statistiques sont effectués par la suite.

Ancel [2003] évoque principalement le problème de la différence entre l'état capté de la jurisprudence et son état présent. En effet, les longs délais de travail sont caractéristiques de ces études. L'étude de son équipe portait sur les décisions d'expulsion d'occupants sans droit ni titre. La saisie des informations à elle seule a duré 9 mois. De plus, il est difficile d'observer l'évolution des pratiques judiciaires dans le temps et leur différence entre les villes du fait de la faible taille de l'échantillon choisi. Par exemple, Jeandidier & Ray [2006] n'ont analysé que 399 dossiers d'affaires de pension alimentaire correspondant aux audiences s'étalant de fin 1999 à fin 2000 d'un seul tribunal de grande instance. L'équipe de Ancel [2003] n'a quant à elle analysé que 3865 décisions sélectionnées parmi 5656 décisions rendues du 1<sup>er</sup> juillet au 31 décembre 2001.

La problématique de notre étude est « **comment donner accès à l'analyse automatique de la sémantique d'un corpus jurisprudentiel pour comprendre la prise de décision des juges ?** ». La complexité de cette analyse s'explique notamment par l'interprétation subjective des règles juridiques, l'application non déterministe de la loi, et la technicité du langage judiciaire. Cette problématique intéresse des entreprises telles que

LexisNexis<sup>7</sup> et Lexbase SA<sup>8</sup>, et plusieurs startups telles que Predictice<sup>9</sup> et CASE LAW ANALYTICS<sup>10</sup>. Afin d'y répondre, nous nous intéressons aux concepts d'informations mentionnées dans les décisions, au centre desquels se trouvent les demandes des parties (prétentions) sur lesquelles portent les conclusions rendues. Ainsi, l'analyse sémantique d'un corpus jurisprudentiel vise l'identification de connaissances sur les demandes (Figure 5 Page 7).



*Sens des flèches : l'extrémité est l'information sur l'origine.*

Figure 5 – La demande au centre de la compréhension des décisions.

Une demande peut être caractérisée par :

- l'objet qui a été demandé (par ex. dommages et intérêts) quantifié par un quantum ;
- le résultat associé qui est décrit par une polarité (« accepte » ou « rejette »), souvent lié à un quantum accordé, par exemple 5000 euros de dommages et intérêts ou 2 mois d'emprisonnement ;

7. <https://lexmachina.com>

8. <https://www.legalmetrics.fr>

9. <http://predictice.com>

10. <http://caselawanalytics.com>

- le fondement ou la norme juridique qui est la règle ou l'ensemble de règles juridiques qui légitiment la prétention ou le résultat ;
- les circonstances factuelles définissent des types d'affaires et qui caractérisent les différentes situations dans lesquelles sont formulées les demandes d'une catégorie donnée ; Une catégorie de demande étant définie de manière unique par un objet et un fondement ;
- les divers arguments apportés par les parties pour justifier leurs requêtes (raisons des demandes) ;
- les motivations des solutions des juges (raisons des résultats).

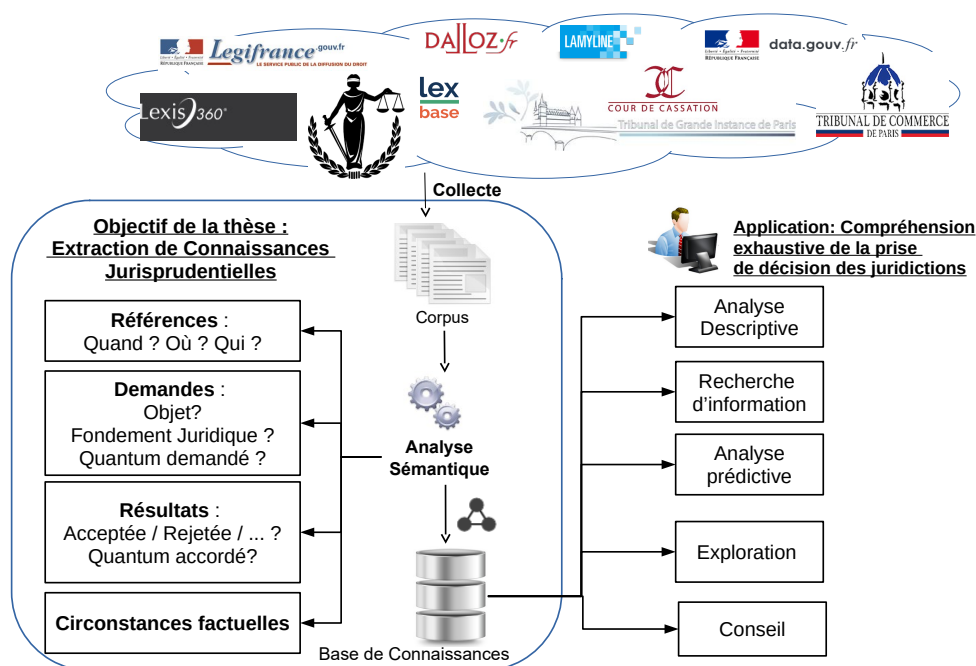


Figure 6 – Objectifs et exemples d'application de la thèse.

Comme illustré par la Figure 6 Page 8, l'analyse sémantique identifie ou découvre différentes informations descriptives d'un corpus constitué par des décisions collectées à partir de divers moteurs de recherche juridique et des juridictions. Cette thèse s'inscrit dans un projet qui vise, entre autres, à automatiser la constitution d'une base de connaissances sur la jurisprudence française. Une telle base permettrait notamment de mener une grande variété de recherches et d'études expertes. Elle aurait aussi



naturellement une importance certaine pour la définition de modèles prédictifs par exemple pour la prédiction des types de demandes à formuler et la prédiction de la solution des juges.

## **ii.a Collecte, gestion et pré-traitement des documents**

Il est nécessaire de trouver des moyens pour collecter le maximum de documents bruts non-structurés, les pré-traiter, et organiser leur gestion afin de les indexer pour faciliter leur traitement. Les décisions de cours d'appel de justice civile sont les plus accessibles à partir des moteurs de recherche juridique (Lexis360, Dalloz, LamyLine, Lexbase, Legifrance, etc.) et de la grande base de données JuriCa alimentée d'environ 180k décisions civiles par an [Lamanda, 2010]. Cependant, l'accès à ces décisions est généralement payant, et le nombre de documents simultanément téléchargeables est très faible sur les sites payants (généralement 10 à 20 décisions au maximum à la fois). La base JuriCa est la plus grosse base de décisions de cours d'appel en France. Elle est gérée par la Cour de cassation. L'accès à cette base est offert par le Service de Documentation, des Etudes et du Rapport<sup>11</sup> (SDER). L'accès est payant pour les professionnels et gratuit pour les universités et centres de recherche en partenariat avec le SDER. Lexbase dispose depuis une dizaine d'année d'une licence pour vendre les décisions de JuriCa [Emmanuel, 20 janvier 2010]. Legifrance, le moteur de recherche du ministère de la justice, fournit quant à lui un accès public et gratuit à un nombre considérable de documents. Les décisions y sont identifiées à l'aide de numéros consécutifs et accessibles à partir d'un service Web. Ce dernier a l'avantage de proposer des décisions de tous les ordres et de tous les degrés. Cependant, les décisions des juridictions du premier degré (appelées jugements) restent plus rares sur internet et principalement disponibles auprès des tribunaux. Il faut préciser que nos expérimentations se sont concentrées sur les décisions d'appel en justice civile, et ce choix a été motivé par le fait qu'elles sont les plus nombreuses sur internet.

Les décisions sont ainsi collectées à partir de diverses sources pouvant contenir des documents identiques. Il est donc important de les identifier de manière unique pour éviter des redondances de traitement. Par ailleurs,

---

11. [https://www.courdecassation.fr/institution\\_1/composition\\_56/etudes\\_rapport\\_28.html](https://www.courdecassation.fr/institution_1/composition_56/etudes_rapport_28.html)

l'uniformisation préalable des documents permet d'analyser les décisions indépendamment de leur format d'origine.

## ii.b Extraction de connaissances

La difficulté d'extraire des connaissances des corpus jurisprudentielles découle de l'état non-structuré des documents et de la complexité du langage employé. L'extraction des connaissances nécessite de mettre en œuvre des techniques de fouille de textes adaptées à la nature des éléments à identifier. Ce mémoire aborde l'annotation des références de l'affaire (juridiction, ville, participants, juges, date, numéro R.G., normes citées, ...), l'extraction des demandes et résultats correspondants, et l'identification des circonstances factuelles.

Les méta-données de référence sont des segments de texte qu'on peut directement localiser dans le document. Elles sont donc semblables aux entités nommées dont la reconnaissance est une problématique intensivement étudiée en traitement automatique du langage naturel [Yadav & Bethard, 2018] dans plusieurs travaux et compétitions, aussi bien pour des entités communes [Tjong Kim Sang & De Meulder, 2003; Grishman & Sundheim, 1996], que pour des entités spécifiques à un domaine [Kim *et al.*, 2004; Persson, 2012; Hanisch *et al.*, 2005], et dans diverses langues [Li *et al.*, 2018; Alfred *et al.*, 2014; Amarappa & Sathyanarayana, 2015].

Le problème d'identification des demandes consiste à reconnaître, dans la décision analysée, l'objet, le fondement, le quantum demandé, le sens du résultat correspondant, et le quantum accordé de chaque prétention. La demande s'apparente donc aux entités structurées telles que les événements ACE [2005] qui sont décrits par un type, un terme-clé, des participants, un temps, une polarité.

Le problème d'identification des circonstances factuelles consiste à constituer des regroupements au sein du corpus des décisions traitant une certaine catégorie de demande (objet+fondement). Le but est, comme indiqué précédemment, de repérer les différentes situations dans lesquelles cette catégorie de demande est formulée. Chacun des groupes représente donc une situation particulière partagée par les membres du groupe mais bien distinctes de celles reflétées par les autres groupes. Ce problème évoque des problématiques de similarité entre textes, de catégorisation non-supervisée (*clustering*), et de « modélisation thématique » (*topic modeling*). A l'issue du processus d'extraction, les connaissances extraites sont destinées à enrichir

progressivement une base de connaissances. Cette dernière est amenée à faciliter les diverses analyses automatiques applicables aux décisions et demandes judiciaires.

### **ii.c Application : analyse descriptive**

L'analyse descriptive exploite l'ensemble des connaissances extraites et organisées pour répondre aux diverses questions que l'on pourrait se poser sur l'application de la loi. Il est intéressant par exemple de comparer les fréquences de résultats positifs et négatifs pour une catégorie de prétention donnée dans une situation précise. Les quantités extraites servent à visualiser les différences entre les montants accordés et réclamés. D'autres analyses plus complexes permettraient d'étudier l'évolution dans le temps et les différences d'opinion entre les juges suivant leur localisation géographique (ville, département, région, etc.).

## **iii Méthodologie**

Les tâches sont définies par le métier. Comme illustrées précédemment (§ ii.b), les problématiques propres aux textes juridiques trouvent généralement des analogies avec les problèmes étudiés en analyse de données textuelles. Ainsi, les méthodes issues de ce domaine sont applicables aux textes juridiques. Cependant, quelques adaptations sont généralement nécessaires pour obtenir des résultats de bonne qualité hors des domaines pour lesquels ces approches ont été développées [Waltl *et al.*, 2016]. De plus, la recherche en fouille de textes est souvent réalisée sur des échantillons qui ne reflètent pas toujours la complexité des données réelles. Effectuant l'une des premières études d'analyse sémantique des décisions françaises, nous avons axé notre travail sur le rapprochement des problèmes de l'analyse des décisions jurisprudentielles de ceux généralement traités en analyse de textes. Il s'agit ensuite d'établir des protocoles d'évaluation et d'annotation manuelle de données. Selon les problématiques identifiées et les protocoles d'évaluations définis, des méthodes adaptées ont été proposées et expérimentées sur les données réelles annotées manuellement par un expert juriste.

## iv Résultats

Les principaux résultats de cette thèse sont un ensemble de proposition d'approches d'extractions de connaissances et les discussions d'un grand nombre de résultats empiriques. Premièrement, le sectionnement a pour but d'organiser l'extraction des informations qui sont réparties dans des sections selon leur nature. Par exemple, les méta-données de références sont dans l'entête du documents, et les fondements utilisés sont cités dans le reste du document. L'application de deux modèles probabilistes, les champs aléatoires conditionnels ou CRF (*conditional random fields*) et les modèles cachés de Markov ou HMM (*hidden Markov Model*), est étudiée pour le sectionnement et l'annotation d'entités juridiques en considérant plusieurs aspects de la conception des systèmes d'extraction d'information par étiquetage de séquence.

Par la suite, nous proposons une méthode d'extraction des demandes et des résultats en fonction des catégories présentes dans la décision. L'approche consiste en effet à identifier dans un premier temps les catégories (objet+fondement) présentes par classification supervisée. Un vocabulaire d'expression des demandes et résultats est exploité pour identifier les passages. Puis à l'aide de termes propres à chacune des catégories identifiées, les trois attributs (quantum demandé, sens du résultat, quantum accordé) des paires demande-résultat sont reconnus.

Par ailleurs, nous analysons l'extraction du sens du résultat par classification binaire des documents. L'objectif est de s'affranchir de l'identification préalable de l'expression des demandes et résultats. En effet, les décisions comprenant des demandes d'une catégorie donnée semblent ne contenir, dans une forte proportion, qu'une seule demande. A partir d'une représentation adéquate du contenu de la décision, il est possible d'identifier le sens du résultat de la seule demande d'intérêt par classification binaire de documents.

L'identification des circonstances factuelles, quant à elle, est modélisée comme une tâche de regroupement non supervisé de documents. Nous proposons dans ce cas une méthode d'apprentissage d'une distance entre textes, à l'aide d'un algorithme de régression. La métrique apprise est utilisée dans l'algorithme des « K-moyennes » (*k-means*) [Forgey, 1965] et celui des « K-medoides » (*k-medoids*) [Kaufman & Rousseeuw, 1987], et comparée à d'autres distances établies en recherche d'information.

## **v Structure de la thèse**

La thèse est organisée en 6 chapitres. Le chapitre 1 positionne nos travaux par rapport à ceux qui ont été réalisés précédemment sur des problématiques d'analyse automatique de décisions de justice. Le chapitre ?? présente les architectures et modèles proposés pour la structuration des décisions et la reconnaissance des entités juridiques ; il discute notamment des différents résultats empiriques obtenus par application des modèles CRF et HMM. Ensuite, le chapitre ?? détaille le problème d'extraction des demandes, puis présente notre méthode et les résultats obtenus. Le chapitre ?? traite de l'identification du sens du résultat par classification directe des décisions, cela en comparant différents algorithmes de classification et de représentations des textes. Le chapitre ?? discute de l'usage de l'apprentissage proposé d'une distance qui est comparée à d'autres distances pour la découverte des circonstances factuelles. Enfin, le chapitre ?? présente des résultats de scénarios d'analyses descriptives pour illustrer l'exploitation potentielle de nos propositions sur un corpus de grande taille.

# Chapitre 1

---

## Analyse automatique de corpus judiciaires

---

L'étude bibliographique de ce chapitre est focalisée sur l'application de techniques d'analyse de données textuelles aux décisions judiciaires. Une synthèse bibliographique plus technique sur les algorithmes de fouille de texte est détaillée dans les chapitres qui traitent, dans la suite, des méthodes que nous avons mises en œuvre. Plus précisément, suivant la structure du présent chapitre, il s'agit des chapitres ?? et ?? pour l'extraction d'information, du chapitre ?? pour la classification des documents, et du chapitre ?? pour la similarité entre documents.

### 1.1 Introduction

Les deux grands paradigmes de jugement se distinguent par l'importance qu'ils accordent aux règles juridiques [Tumonis, 2012]. D'une part, les adeptes du Formalisme Juridique, plus pertinent dans le droit civil, considèrent que toutes les considérations normatives ont été incorporées dans les lois par leurs auteurs. D'autre part, l'école du Réalisme Juridique, plus proche du « *Common Law* », permet un pouvoir discrétionnaire entre les jugements en raisonnant selon le cas. Les premières tentatives d'anticipation des comportements judiciaires s'appuyaient sur une formalisation des lois. Il en est né le « droit computationnel », qui est une sous-discipline de l'« informatique juridique<sup>1</sup> ». Il s'intéresse, en effet, au raisonnement juridique automatique axé sur une représentation sémantique riche et plus formelle de la loi, des régulations, et modalités de contrat [Love & Genereth, 2005]. Il vise à réduire la taille et la complexité de la loi pour la

---

1. Application des techniques modernes de l'informatique à l'environnement juridique, et par conséquent aux organisations liées au droit.

rendre plus accessible. Plus précisément, le « droit computationnel » propose des systèmes répondant à différentes questions, comme « Quel montant de taxe dois-je payer cette année ? » (planification juridique), « Cette réglementation contient-elle des règles en contradiction » (analyse réglementaire), « L'entreprise respecte-t-elle la loi ? » (vérification de la conformité) [Genesereth, 2015]. Les techniques pro Formalisme Juridique étaient déjà critiquées au début des années 60, parce qu'excessivement focalisées sur les règles juridiques qui ne représentent qu'une partie de l'institution juridique [Llewellyn, 1962]. Pour analyser le comportement judiciaire, plusieurs variables plus ou moins contrôlables, comme le temps, le lieu et les circonstances, doivent aussi être prises en compte [Ulmer, 1963]. Etant donné que les juristes s'appuient sur la recherche de précédents, Ulmer [1963] conseille de se concentrer sur les motifs réguliers que comprennent les données pour réaliser des analyses quantitatives. Il est possible d'exploiter la masse de décisions pour identifier de telles régularités car une collection suffisante d'une certaine forme de données révèle des motifs qui une fois observés sont projetables dans le futur [Ulmer, 1963]. Il s'agit de raisonnements à base de cas qui se distinguent de ceux à base de règles.

Les premiers outils automatiques d'anticipation des décisions étaient généralement des systèmes experts juridiques. Ces derniers raisonnent sur de nouvelles affaires en imitant la prise de décision humaine par la logique en général et souvent par analogie. Ils s'appuient sur un raisonnement à base de règles, c'est-à-dire à partir d'une représentation formelle des connaissances des experts ou du domaine. En droit, il s'agit de la connaissance qu'a l'expert des normes juridiques, et de l'ordre des questions à traiter lors du raisonnement sur un cas (appris par expérience). Le modèle explicite de domaine nécessaire ici se trouve dans une base de connaissances où les normes juridiques sont représentées sous forme de « SI ... ALORS ... », et les faits sont généralement représentés dans la logique des prédicats. Un système expert juridique doit s'appuyer sur une base de connaissances juridiques exhaustive et disposer d'un moteur d'inférence capable de trouver les règles pertinentes et le moyen efficace, par déduction, de les appliquer afin d'obtenir la solution du cas d'étude aussi rapidement que possible. Les systèmes experts ont échoué dans leur tentative de prédire les décisions de justice [Leith, 2010]. La première raison découle de ce que Berka [2011] a appelé le « goulot d'acquisition de connaissances » c'est-à-dire le problème d'obtention des connaissances spécifiques à un domaine d'expertise sous la forme de règles suffisamment générales.

L'autre raison tient à l'interprétation ouverte du droit et à la complexité de la formalisation applicable sans tenir compte des particularités de l'affaire.

Deux paradigmes s'affirment comme de bonnes alternatives aux raisonnements à base de règles. La première est le raisonnement à base de cas qui concerne une recherche de solution, une classification ou toute autre inférence pour un cas courant à partir de l'analyse d'anciens cas et de leurs solutions [Moens, 2002]. Un tel système juridique affecte à un nouveau cas en retrouvant la solution des cas les plus similaires dont sont connues les solutions [Berka, 2011]. Pour un problème de classification, l'algorithme des  $k$ -plus-proches-voisins est une méthode adéquate de raisonnement à base de cas [Poole & Mackworth, 2017]. L'algorithme du plus proche voisin (1-plus-proche-voisin) est utilisé notamment par Ashley & Brüninghaus [2009] pour identifier les types de faits (« Facteurs ») d'une affaire. Pour d'autres problèmes plus complexes, les différences entre les deux cas peuvent exiger une adaptation de la solution du cas le plus similaires. La seconde alternative est l'apprentissage automatique. Contrairement aux paradigmes de raisonnement précédents qui nécessitent de programmer explicitement des étapes ou instructions à exécuter, concerne le développement de programmes qui apprennent automatiquement à accomplir une tâche à partir des données auxquelles ils ont accès. L'apprentissage automatique est plus récemment utilisé pour la prédiction de l'issue d'affaires. Pour exemple, Katz *et al.* [2014] entraînent des forêts aléatoires [Breiman, 2001] sur les cas de 1946-1953 pour prédire si la Cour Suprême des États-Unis infirmera ou confirmera une décision de juridiction inférieure. Leur approche parvient à prédire correctement 69,7% des décisions finales pour 7700 cas des années 1953-2013. Ils ont amélioré ce résultat par la suite en augmentant le nombre d'arbres et la quantité de données [Katz *et al.*, 2017]. Toujours pour la prédiction des décisions de la Cour Suprême des États-Unis, Walzl *et al.* [2017b] utilisent des techniques de traitement automatique du langage naturel (TALN) pour extraire moins d'attributs caractéristiques de décisions que [Katz *et al.*, 2014] à partir des décisions d'appel de la Cour Fiscale allemande (11 contre 244). Ils obtiennent des valeurs de  $F_1$ -mesures entre 0,53 et 0,58 (validation croisée à 10 itérations) pour la prédiction de la confirmation ou l'infirmerie d'un jugement en appel avec un classifieur bayésien naïf.

Notre objectif est d'alimenter les analyses quantitatives de corpus jurisprudentiels en proposant des méthodes d'extraction de connaissances pertinentes telles que les méta-données d'affaires, les règles juridiques as-



sociées, les demandes des parties, les réponses des tribunaux, et les liens entre ces données. L'un des postulats évalués empiriquement dans cette thèse est que l'identification de ces diverses connaissances est possible par l'analyse des textes judiciaires basée sur des méthodes du TALN, de la fouille de texte et de la recherche d'information. Cependant, l'application de ces méthodes exigent certaines adaptations pour surmonter les divers défis décrits par Nazarenko & Wyner [2017] : textes très longs et en grande quantité, corpus régulièrement mis à jour, influence subjective de facteurs sociaux et d'opinions politiques, couverture de problématiques économiques, sociales, politiques très variées, langage complexe, etc. Dans la suite de ce chapitre, nous passons en revue des travaux qui ont été menés dans ce sens pour traiter des problématiques proches des nôtres, en particulier celles décrites dans l'introduction générale (§ ii.b).

## 1.2 Annotation et extraction d'information

L'annotation consiste à enrichir les documents pour les préparer à des analyses, faciliter la recherche d'affaires pertinentes, et faire la lumière sur des connaissances linguistiques sous-jacentes au raisonnement juridique. Les éléments annotés peuvent être de courts segments de texte mentionnant des entités juridiques [Waltl *et al.*, 2016; Wyner, 2010] comme la date, le lieu (juridiction), les noms de juges, des citations de loi. L'annotation de passages plus longs consiste à identifier des instances de concepts juridiques plus complexes comme les faits [Wyner, 2010; Wyner & Peters, 2010; Shulayeva *et al.*, 2017], les définitions [Waltl *et al.*, 2016, 2017a], des citations de principes juridiques [Shulayeva *et al.*, 2017], ou des arguments [Wyner *et al.*, 2010].

Différentes méthodes ont été expérimentées pour la reconnaissance d'information dans les documents judiciaires. La plupart reposent sur l'entraînement d'algorithmes d'apprentissage automatique supervisé sur un ensemble d'exemples annotés manuellement (résultats attendus). Parmi ces algorithmes, on retrouve par exemple les modèles probabilistes HMM (Modèles Cachés de Markov, cf. ??) et CRF (Champs Aléatoires Conditionnels, cf. ??) que dont l'application est étudiée au chapitre ??. Ces modèles peuvent être combinés à d'autres approches dans un système global. En effet, après avoir segmenté les documents à l'aide d'un modèle CRF, Dozier *et al.* [2010] ont par exemple combiné plusieurs approches pour recon-

naître des entités dans les décisions de la Cour Suprême des États-Unis. Ils ont défini manuellement des détecteurs distincts à base de règles pour identifier séparément la juridiction (zone géographique), le type de document, et les noms des juges, en plus de l'introduction d'une recherche lexicale pour détecter la cour, ainsi qu'un classifieur entraîné pour reconnaître le titre. Ces différents détecteurs ont atteint des performances prometteuses, mais avec des rappels limités entre 72% et 87%. Suivant la complexité des éléments à extraire, un système peut exploiter un lexique pour les motifs simples et non-systématiques (indicateurs de mentions de résultats ou de parties) et des règles pour des motifs plus complexes et systématiques (e.g., noms de juges, énoncés de décisions) [Waltl *et al.*, 2016, 2017a; Wyner, 2010]. Cardellino *et al.* [2017] ont par ailleurs utilisé un modèle CRF et des réseaux de neurones pour la reconnaissance d'entités nommées juridiques dans des jugements de la Cour Européenne des Droits de l'Homme. Ils définissent une hiérarchie des entités nommées distinguant au niveau 1, les entités nommées et des non-entités, spécialisées par 6 classes au niveau 2 (par exemple, Personne, Document), spécialisées par 69 classes au niveau 3 (par exemple, Rôle Juridique, Règlement), spécialisées par 358 classes au niveau 4 (par exemple Juge, Code Juridique). Les basses performances qu'ils rapportent sur le corpus juridique illustrent bien la difficulté de la détection d'entités juridiques dans les décisions judiciaires ( $F_1$ -mesures de 0.25, 0.08, 0.03 en moyenne respectivement pour les niveaux 2, 3, 4). Plus récemment encore, Andrew & Tannier [2018] proposent une approche pour l'extraction d'entités nommées d'une transaction d'investissement<sup>2</sup> et des relations qu'elles partagent dans des décisions du Luxembourg rédigées en français. Ils combinent un modèle CRF pour les entités à une grammaire GATE JAPE [Thakker *et al.*, 2009] pour les relations, et obtiennent un faible taux d'erreur pour le CRF de 3.12%.

Pour la détection des arguments, par contre, Moens *et al.* [2007] proposent une classification binaire des phrases : *argumentative* / *non argumentative*. Ils comparent notamment le classifieur bayésien multinomial et le classifieur d'entropie maximum tout en explorant plusieurs caractéristiques textuelles. Mochales & Moens [2008] proposent, pour la même tâche, une méthode d'extraction basée sur une formalisation de la structure des arguments dans les jugements par une grammaire sans contexte.

---

2. Entités : Personne, Nom, Adresse, Société Principale, Société Secondaire, Rôle, Fonction, Type Société.

## 1.3 Classification des jugements

La classification de texte permet d'organiser un corpus en rangeant les documents dans des catégories généralement prédéfinies par des experts. Pour la classification des décisions, le principe des propositions de la littérature est d'entraîner un modèle statistique traditionnel sur une représentation des documents généralement définie à partir des connaissances du domaine. Par exemple, par classification binaire avec une Machine à Vecteurs de Support (SVM) [Vapnik, 1995] à noyau linéaire (cf. ??), Aletras *et al.* [2016] identifient s'il y a eu une violation d'un article donné de la convention des droits de l'homme sur les jugements de la Cour Européenne des Droits de l'Hommes (CEDH)<sup>3</sup>. Les vecteurs représentant les documents sont construits sur la base des 2000 n-grammes les plus fréquents. Certaines composantes sont les fréquences normalisées des n-grammes sélectionnés (modèle sac-de-mots [Salton *et al.*, 1975; Salton & McGill, 1983]), calculées distinctement pour différentes parties du document (Procédure, Circonstances, Faits, Loi applicable, la Loi et le document entier); ce qui résulte en une matrice document-terme  $C$ . D'autres composantes sont définies par la fréquence des thématiques extraites par une catégorisation non supervisée (*clustering*) avec la similarité cosinus des n-grammes les plus fréquents représentés par leurs vecteurs dans  $C$ , i.e. le vecteur de leurs scores d'occurrence dans les différentes parties précédemment citées du document. Aletras *et al.* [2016] obtiennent une précision moyenne de 79% sur les 3 articles qu'ils ont manipulés. Notons tout de même la sélection des régions particulières (circonstances, faits, lois, etc.) du document à partir desquelles sont extraits les n-grammes. Cette sélection est un ajustement de la représentation des textes qui paraît nécessaire pour obtenir de bons résultats. La structuration préalable des documents est ainsi utile pour réduire le bruit qui occupe généralement plus d'espace que les passages ou éléments d'intérêt. Medvedeva *et al.* [2018] étendent ces travaux à neuf articles de loi, tout en montrant empiriquement, entre autres, la possibilité de prédire la violation des articles sur des périodes futures à celles couvertes par les données utilisées lors des phases d'entraînement. Şulea *et al.* [2017a] traitent, d'autre part, l'identification des résultats dans des arrêts<sup>4</sup> de la Cour Française de Cassation.

---

3. HUDOC ECHR Database : <http://hudoc.echr.coe.int>.

4. Documents de <https://www.legifrance.gouv.fr>.

Après un essai [Şulea *et al.*, 2017b] avec un SVM entraîné sur une représentation des documents par le modèle TF-IDF [Salton & Buckley, 1988], ils améliorent les résultats à l'aide d'un classifieur ensembliste de SVM à probabilité moyenne, parvenant à des  $F_1$ -mesures de plus de 95% Şulea *et al.* [2017a]. Un classifieur SVM à probabilité moyenne combine plusieurs modèles SVM dits « faibles » (ou de base) entraînés chacun sur un sous-ensemble de la base d'apprentissage. Lors de la prédiction, chacun des SVM estime une probabilité d'appartenance du document classifié à chaque classe. La classe du document est celle dont la probabilité moyenne (robustement estimé par la médiane [Kittler *et al.*, 1998]) est maximale.

Par ailleurs, Ashley & Brüninghaus [2009] identifient les « Facteurs » (*Factors* [Ashley, 1990]) qui s'appliquent à une affaire par classification de la description des faits de cette affaire. Les Facteurs sont en effet des aspects juridiques spécifiques à un domaine qui sont importants dans un contentieux [Bench-Capon, 1997]. Ils font abstraction des faits dans les raisonnements à base de cas où ils sont définis sous forme de prédicats favorables soit au plaignant soit au défendeur. Sur l'appropriation illícite de secrets commerciaux (*trade secret misappropriation*), l'environnement d'enseignement CATO<sup>5</sup> [Aleven & Ashley, 1997; Aleven, 2003] comprend 26 Facteurs. On y retrouve par exemple les Facteurs *Unique-Product* (le produit est unique), *Agreed-Not-To-Disclose* (il existait un accord de non-divulgaration entre le défendeur et le plaignant), *Info-Reverse-Engineerable* (les informations du produit peuvent être apprises par ingénierie inverse), et *Disclosure-In-Negotiations* (le demandeur a divulgué les informations concernant son produit lors de négociations avec le défendeur). Les deux premiers Facteurs favorisent le plaignant, propriétaire du produit, et les deux derniers Facteurs favorise le défendeur accusé. Un Facteur s'applique à une affaire si la description de cette dernière contient un fait correspondant. Ashley & Brüninghaus [2009] définissent un classifieur (le-plus-proche-voisin) par Facteur pour identifier ceux qui s'appliquent à la décision. En effet, les phrases de faits des cas résolus sont labellisées par le Facteur auxquels il correspondent. Ensuite, la classification d'un nouveau cas consiste à comparer les différentes phrases annotées à chacune des phrases de faits du nouveau cas, et à affecter à ces derniers le Facteur de la phrases

---

5. CATO est un environnement intelligent d'enseignement, aux étudiants de droit, de compétences de construction d'arguments à partir de cas à travers la pratique de tâches de test de théorie et d'argumentation à base de cas [Aleven & Ashley, 1997].

annotées la plus similaires. Au cours de leurs expérimentations, les auteurs démontrent que des adaptations de la représentation des cas par sacs-de-mots sont nécessaires pour améliorer les résultats de classifications. Ils proposent deux méthodes améliorées de représentation [Brüninghaus & Ashley, 2001] : l' « abstraction des noms par les rôles » (*roles-replaced representation*) et les « schémas propositionnels » (*propositional patterns*). La représentation par abstraction des noms par les rôles consiste à remplacer les noms des parties et les informations sur le produit par leur rôle respectifs : *plaintiff* (demandeur), *defendant* (defendeur), *information* (produit). Quant à la représentation par schémas propositionnels, elle consiste à définir, à l'aide de techniques de TALN, des attributs sous forme de propositions logiques du texte qui captent la signification du Facteur. Pour le Facteur *Disclosure-In-Negotiations*, pour capter le fait que le demandeur a divulgué quelque chose, les auteurs définissent, par exemple, la proposition ( $\pi$  disclose) chaque fois qu'un synonyme du verbe *disclose* (divulger) est identifié. Réalisées sur 146 affaires, les expérimentations de validation croisée *leave-one-out*<sup>6</sup> montre l'impact des améliorations avec une  $F_1$ -mesure moyenne de 0.211 pour les sacs-de-mots, 0.26 pour les schémas propositionnels, et 0.28 pour l'abstraction des noms par les rôles.

D'autres catégorisations sont tout aussi utiles pour faciliter la recherche d'information. Par exemple, Şulea *et al.* [2017b,a] expérimentent la classification pour identifier la formation judiciaire (chambre civile, chambre commerciale, chambre sociale, etc.) et la période (Intervalle d'années dans laquelle la décision a été prononcée) des décisions. La classification peut aussi servir à évaluer d'autres problématiques comme la similarité [Ma *et al.*, 2018].

## 1.4 Similarité entre décisions judiciaires

La similarité entre textes est indispensable pour des applications qui nécessitent de regrouper des textes traitant de sujets similaires, et séparer ceux dont les sujets sont différents. La mesure de similarité doit être définie de sorte à rapprocher ou éloigner les documents suivant l'aspect sémantique que l'on souhaite révéler. Nair & Wagh [2018] exploitent les

---

6. Pour  $N$  cas annotés, une validation croisée *leave-one-out* réalise  $N$  expérimentation qui utilise à tour de rôle 1 cas pour le test et le reste ( $N - 1$ ) comme cas d'apprentissage.

citations de lois et précédents<sup>7</sup> pour retrouver les textes juridiques qui ont une similarité. Ils analysent le réseau de 597 citations<sup>8</sup> sous l'Acte 2000 des Technologies de l'Information (*Information Technology Act, 2000*<sup>9</sup>) dans des jugements indiens. Leur proposition est d'utiliser des règles d'association générées par l'algorithme Apriori [Agrawal *et al.*, 1994] pour regrouper les jugements susceptibles d'être cités ensemble. Cet algorithme recherche les ensembles singletons de citations suffisamment fréquentes (seuil nécessaire), puis fusionne de manière itérative les ensembles tant que la co-occurrence des citations de la fusion est suffisamment fréquente dans le réseau. Une règle d'association  $\{c_1, \dots, c_n\} \rightarrow \{c'\}$  indique qu'une citation  $c'$  est observable si l'on observe une co-occurrence d'un ensemble donné de citations  $\{c_1, \dots, c_n\}$ . A chaque règle est associé un score de confiance calculé à partir d'une métrique appelé score de support  $sc()$  qui indique, pour un ensemble  $\{c_1, \dots, c_n\}$ , la fréquence de co-occurrence des citations de cet ensemble. Le support d'un singleton est sa fréquence d'occurrence. La similarité est confirmée si le score de confiance de la règle  $conf(\{c_1, \dots, c_n\} \rightarrow \{c'\}) = sc(\{c_1, \dots, c_n, c'\}) / sc(\{c_1, \dots, c_n\})$  est suffisamment élevé. Nair & Wagh [2018] démontrent au travers de scénarios (aucune évaluation quantifiée de l'efficacité de l'approche n'est proposée) que les documents qui sont fréquemment cités ensemble sont similaires car traitant de thématiques proches. Cette relation permet par transitivité de retrouver les documents pertinents dans une base de données.

Les métriques traditionnelles de similarité ne sont pas toujours très efficaces sur les décisions judiciaires. La raison peut être une représentation inadéquate des textes qui ne permet pas de traduire une fois comprise la notion de similarité telle qu'entendue dans la plupart des travaux. Thenmozhi *et al.* [2017] comparent par exemple, l'utilisation de la similarité cosinus sur trois représentations différentes des jugements dans le cadre de la campagne de recherche d'affaires antérieures pertinentes IR-LeD@FIRE2017 [Mandal *et al.*, 2017] : (1) TF-IDF des concepts (noms), (2) TF-IDF des concepts et relations (verbes), (3) et la moyenne des plongements lexicaux *Word2Vec* [Le & Mikolov, 2014] des concepts et relations. Au vu des résultats (0.1795, 0.178, 0.0755 de précision@10<sup>10</sup> et 0.681, 0.661,

7. Les jugements du « Common Law » citent des décisions antérieures similaires.

8. Disponibles sur <https://indiankanoon.org>.

9. <https://www.meity.gov.in/content/information-technology-act-2000>.

10. Précision@N, rappel@N : précision et rappel calculées sur les N premiers résultats retournés par un système de recherche d'information.

0.435 de `rappel@1010` respectivement pour les méthodes 1, 2, et 3), la première représentation semble mieux capter la similarité contrairement à l'utilisation des verbes et de la représentation distribuée. Ma *et al.* [2018] utilisent une forme de connaissances a priori définie dans des modèles de type ontologie pour estimer la similarité entre décisions. Ils proposent notamment d'aligner le document sur une ontologie des concepts et relations d'un corpus judiciaire. L'idée est de calculer la similarité sur un résumé du texte qui regroupe des aspects pertinents. Cette méthode permet ainsi de mieux capter la sémantique des jugements, d'avoir une meilleure précision, et de réduire la complexité temporelle inhérente à l'exploitation de longs documents notamment lors de l'utilisation de la « distance du déménageur de mots » ou WMD (*Word Mover's Distance*) de Kusner *et al.* [2015] (cf. ??). L'amélioration est observée sur une tâche de classification des jugements Chinois relatifs aux crimes de la circulation routière dans quatre catégories correspondant à des sentences d'emprisonnement<sup>11</sup> (précision de 90.3% et 92.3% pour le résumé contre 84.8% et 82.4% pour le document original respectivement sur les deux ensembles de données utilisés).

Toujours dans l'objectif d'une représentation pertinente des textes, Kumar *et al.* [2011] proposent quatre méthodes pour l'estimation de la similarité entre deux jugements  $x$  et  $y$  de la Cour Suprême indienne :

1. *all-term cosine similarity* : le cosinus de similarité entre les représentations TF-IDF de  $x$  et  $y$  dont tous les termes présents dans les jugements sont les dimensions.
2. *legal-term cosine similarity* : le cosinus de similarité sur les réductions des dimensions précédentes uniquement aux termes apparaissant dans un dictionnaire juridique.
3. *bibliographic coupling similarity* : la similarité de couplages bibliographiques égal au nombre de citations de jugements communes à  $x$  et  $y$ .
4. *co-citation similarity* : la similarité de co-citation qui est le nombre de citations de  $x$  et  $y$  dans un même jugement.

Les résultats ont été interprétés sur de très faibles proportions des données utilisées (paires de jugements pris parmi 2430 cas et annotées par des experts avec des scores de similarité à valeurs réelles entre 0 pour faible et

---

11. Détention, emprisonnement à durée déterminée de moins de 3 ans, emprisonnement de durée déterminée de 3 à 7 ans et emprisonnement de plus de 7 ans.

1 pour fort). En effet, la comparaison des méthodes de calcul de la similarité est analysée sur les 18 paires de jugements ayant un score de similarité de couplages bibliographiques supérieur ou égal à 3. Deux jugements sont considérés similaires pour un score d'expert  $\geq 0.50$  (parmi les 18 paires, une seule a un score  $\leq 0.50$ ). Il en ressort que le cosinus de similarité avec les termes juridiques (8/18 pour un seuil minimum à 0.5) et le couplage bibliographique (17/18 pour un seuil minimum à 3) correspondent plus à la notion de similarité des experts, que la similarité basée sur tous les termes (3/18 pour un seuil minimum à 1) du corpus ou sur la co-citation (6/18 pour un seuil minimum à 1).

En synthèse, la similarité entre documents est utilisée pour répondre à plusieurs tâches, comme par exemple, la recherche de décisions similaires [Thenmozhi *et al.*, 2017], le regroupement non-supervisé de jugements [Raghuveer, 2012] et la classification supervisée de ces derniers [Ma *et al.*, 2018]. Ces diverses applications définissent aussi la sémantique juridique liée à la notion de similarité. Parmi les questions liées à la conception d'une mesure de la similarité entre documents, on distingue : la sémantique experte qui fonde cette similarité, sa métrique de mesure, la représentation des documents, le contexte d'exploitation et les critères d'évaluation.

## 1.5 Conclusion

En résumé, les travaux portant sur l'analyse automatique des décisions ont donné des résultats encourageants grâce aux éléments spécifiques aux affaires. Ces éléments peuvent être extraits des décisions grâce aux techniques de TALN et de fouille de texte. L'analyse des données textuelles juridiques a pour but la structuration des documents, l'extraction d'information, et l'organisation sémantique de corpus. Le domaine est très actif depuis déjà plusieurs décennies, au point que des librairies de développement, spécifiques au domaine, commencent à voir le jour [Bommarito *et al.*, 2018]. Dans la littérature, nous remarquons que le concepteur investit un minimum d'ingénierie d'adaptation que ce soit pour la définition des caractéristiques pertinentes pour les modèles à apprentissage automatique, ou pour définir les règles pour les méthodes à base de règles ou à base de grammaire. Notons aussi l'effort d'évaluation quantitative avec la participation d'experts pour l'annotation d'exemples de référence même pour des tâches qui peuvent paraître subjectives comme la mesure de similarité.



---

## Bibliographie

---

- ACE. 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*. 5.4.3 edn. Linguistic Data Consortium.
- Agrawal, Rakesh, Srikant, Ramakrishnan, *et al.* . 1994. Fast algorithms for mining association rules. *Pages 487–499 of : Proc. 20th int. conf. very large data bases, vldb*, vol. 1215.
- Aletras, Nikolaos, Tsarapatsanis, Dimitrios, Preoȕiuc-Pietro, Daniel, & Lampos, Vasileios. 2016. Predicting judicial decisions of the European Court of Human Rights : A Natural Language Processing perspective. *Peerj computer science*, **2**, e93.
- Aleven, Vincent. 2003. Using background knowledge in case-based legal reasoning : a computational model and an intelligent learning environment. *Artificial intelligence*, **150**(1-2), 183–237.
- Aleven, Vincent, & Ashley, Kevin D. 1997. Evaluating a learning environment for case-based argumentation skills. *Pages 170–179 of : Proceedings of the 6th international conference on artificial intelligence and law*. ACM.
- Alfred, Rayner, Leong, Leow Chin, On, Chin Kim, & Anthony, Patricia. 2014. Malay named entity recognition based on rule-based approach. *International Journal of Machine Learning and Computing*, **4**(3), 300.
- Amarappa, S., & Sathyanarayana, S. V. 2015. Kannada named entity recognition and classification (NERC) based on multinomial naïve bayes (MNB) classifier. *International Journal on Natural Language Computing (IJNLC)*, **4**(4).
- Ancel, Pascal. 2003. *Les décisions d’expulsion d’occupants sans droit ni titre - connaissance empirique d’un contentieux hétérogène*. Tech. rept. [Rapport de recherche] Ministère de la Justice.

- Andrew, Judith Jeyafreeda, & Tannier, Xavier. 2018. Automatic extraction of entities and relation from legal documents. *Pages 1–8 of : Proceedings of the seventh named entities workshop.*
- Ashley, Kevin D. 1990. *Modeling legal arguments : Reasoning with cases and hypotheticals*. MIT press.
- Ashley, Kevin D., & Brüninghaus, Stefanie. 2009. Automatically classifying case texts and predicting outcomes. *Artificial intelligence and law*, 17(2), 125–165.
- Bench-Capon, Trevor J.M. 1997. Arguing with cases. *Pages 85–100 of : Jurix.*
- Berka, Petr. 2011. Nest : A compositional approach to rule-based and case-based reasoning. *Advances in artificial intelligence*, 2011, 15.
- Bommarito, II, Michael, J, Katz, Daniel Martin, & Detterman, Eric M. 2018. Lexnlp : Natural language processing and information extraction for legal and regulatory texts. *arxiv preprint arxiv :1806.03688*.
- Breiman, Leo. 2001. Random forests. *Machine learning*, 45(1), 5–32.
- Brüninghaus, Stefanie, & Ashley, Kevin D. 2001. Improving the representation of legal case texts with information extraction methods. *Pages 42–51 of : Proceedings of the 8th international conference on artificial intelligence and law*. ACM.
- Cardellino, Cristian, Teruel, Milagro, *et al.* . 2017. A low-cost, high-coverage legal named entity recognizer, classifier and linker. *Pages 9–18 of : Proceedings of the 16th edition of the international conference on artificial intelligence and law*. ACM.
- Cretin, Laurette. 2014. L’opinion des français sur la justice. *Infostat justice*, 125(Janvier).
- Dozier, Christopher, Kondadadi, Ravikumar, Light, Marc, Vachher, Arun, Veeramachaneni, Sriharsha, & Wudali, Ramdev. 2010. Named entity recognition and resolution in legal text. *Pages 27–43 of : Semantic processing of legal texts*. Springer.

- Emmanuel, Barthe. 20 janvier 2010. *Arrêts des cours d'appel : la base juridique enfin en service chez lexbase*. <https://www.precisement.org/blog/Arrets-des-cours-d-appel-la-base.html>.
- Forgey, Edward. 1965. Cluster analysis of multivariate data : Efficiency vs. interpretability of classification. *Biometrics*, **21**(3), 768–769.
- Genesereth, Michael. 2015. Computational law : The cop in the backseat. *The standford center for legal informatics hosted the third annual futurelaw 2015 conference*.
- Grishman, Ralph, & Sundheim, Beth. 1996. Message understanding conference-6 : A brief history. In : *Coling 1996 volume 1 : The 16th international conference on computational linguistics*, vol. 1.
- Hanisch, Daniel, Fundel, Katrin, *et al.* . 2005. ProMiner : rule-based protein and gene entity recognition. *BMC bioinformatics*, **6**(1), 14.
- Jeandidier, Bruno, & Ray, Jean-Claude. 2006. Pensions alimentaires pour enfants lors du divorce - [ Les juges appliquent-ils implicitement un calcul fondé sur le coût de l'enfant ? ]. *Revue des politiques sociales et familiales*, **84**(1), 5–18.
- Katz, Daniel Martin, Bommarito, Michael James, & Blackman, Josh. 2014. Predicting the behavior of the supreme court of the united states : A general approach. *Available at ssrn 2463244*.
- Katz, Daniel Martin, Bommarito II, Michael J, & Blackman, Josh. 2017. A general approach for predicting the behavior of the supreme court of the united states. *Plos one*, **12**(4), e0174698.
- Kaufman, Leonard, & Rousseeuw, Peter J. 1987. Clustering by means of medoids. *Page 405–416 of : Dodge, Yadolah (ed), Statistical data analysis based on the l1-norm*. North Holland/Elsevier. Amsterdam.
- Kim, Jin-Dong, Ohta, Tomoko, Tsuruoka, Yoshimasa, Tateisi, Yuka, & Collier, Nigel. 2004. Introduction to the bio-entity recognition task at JNLPBA. *Pages 70–75 of : Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*. Association for Computational Linguistics.

- Kittler, Josef, Hatef, Mohamad, Duin, Robert PW, & Matas, Jiri. 1998. On combining classifiers. *Ieee transactions on pattern analysis and machine intelligence*, **20**(3), 226–239.
- Kumar, Sushanta, Reddy, P Krishna, Reddy, V Balakista, & Singh, Aditya. 2011. Similarity analysis of legal judgments. *Page 17 of : Proceedings of Compute 2011 - Fourth Annual ACM Bangalore Conference*. ACM.
- Kusner, Matt, Sun, Yu, Kolkin, Nicholas, & Weinberger, Kilian. 2015. From word embeddings to document distances. *Pages 957–966 of : International conference on machine learning*.
- Lamanda, Vincent. 2010. *Discours du premier président de la cour de cassation vincent lamanda lors de l'audience solennelle de début d'année 2010*. [https://www.courdecassation.fr/institution\\_1/occasion\\_audiences\\_59/debut\\_annee\\_60/discours\\_m.\\_lamanda\\_14858.html](https://www.courdecassation.fr/institution_1/occasion_audiences_59/debut_annee_60/discours_m._lamanda_14858.html).
- Langlais, Eric, & Chappe, Nathalie. 2009. *Analyses économiques du droit : principes, méthodes, résultats*. Editions de Boeck Université. Chap. 4. Analyse économique de la résolution des litiges.
- Le, Quoc, & Mikolov, Tomas. 2014. Distributed representations of sentences and documents. *Pages 1188–1196 of : International conference on machine learning*.
- Leith, Philip. 2010. The rise and fall of the legal expert system. *European journal of law and technology*, **1**(1), 179–201.
- Li, Jianqiang, Zhao, Shenhe, Yang, Jijiang, Huang, Zhisheng, Liu, Bo, Chen, Shi, Pan, Hui, & Wang, Qing. 2018. Wcp-rnn : a novel rnn-based approach for bio-ner in chinese emrs. *The journal of supercomputing*, 1–18.
- Llewellyn, Karl Nickerson. 1962. *Jurisprudence : Realism in theory and practice*. The University of Chicago Press.
- Love, Nathaniel, & Genesereth, Michael. 2005. Computational law. *Pages 205–209 of : Proceedings of the 10th international conference on artificial intelligence and law*. ACM.
- Ma, Yinglong, Zhang, Peng, & Ma, Jiangang. 2018. An Efficient Approach to Learning Chinese Judgment Document Similarity Based on Knowledge Summarization. *arXiv preprint*, 23. arXiv :1808.01843.

- Mandal, Arpan, Ghosh, Kripabandhu, Bhattacharya, Arnab, Pal, Arindam, & Ghosh, Saptarshi. 2017. Overview of the fire 2017 irded track : Information retrieval from legal documents. *Pages 63–68 of : Fire (working notes)*.
- Medvedeva, Masha, Vols, Michel, & Wieling, Martijn. 2018. Judicial decisions of the european court of human rights : Looking into the crystal ball. *In : Proceedings of the conference on empirical legal studies*.
- Mochales, Raquel, & Moens, Marie-Francine. 2008. Study on the structure of argumentation in case law. *Pages 11–20 of : Proceedings of the 2008 conference on legal knowledge and information systems*.
- Moens, Marie-Francine. 2002. What information retrieval can learn from case-based reasoning. *Pages 83–91 of : Legal knowledge and information systems*. Amsterdam : T.J.M. Bench-Capon, A. Daskalopulu and R.G.F. Winkels (eds.), for Jurix 2002 : The Fifteenth Annual Conference.
- Moens, Marie-Francine, Boiy, Erik, Palau, Raquel Mochales, & Reed, Chris. 2007. Automatic detection of arguments in legal texts. *Pages 225–230 of : Proceedings of the 11th international conference on artificial intelligence and law*. ACM.
- Nair, Akhil M., & Wagh, Rupali Sunil. 2018. Similarity Analysis of Court Judgements Using Association Rule Mining on Case Citation Data - A Case Study. *International Journal of Engineering Research and Technology*, **11**(3), 373–381.
- Nazarenko, Adeline, & Wyner, Adam. 2017. Legal NLP Introduction. *Traitement automatique de la langue juridique / Legal Natural Language Processing - Revue TAL*, **58**(2), 7–19.
- Persson, Caroline. 2012. *Machine learning for tagging of biomedical literature*. Closing project report, Technical University of Denmark, DTU Informatics.
- Poole, David, & Mackworth, Alan. 2017. *Artificial intelligence : Foundations of computational agents*. Cambridge University Press. Chap. 7 Supervised Machine Learning.

- Raghuveer, K. 2012. Legal documents clustering using latent dirichlet allocation. *Iaes int. j. artif. intell*, **2**(1), 34–37.
- Salton, Gerard, & Buckley, Christopher. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, **24**(5), 513–523.
- Salton, Gerard, & McGill, Michael J. 1983. *Introduction to modern information retrieval*. mcgraw-hill.
- Salton, Gerard, Wong, Anita, & Yang, Chung-Shu. 1975. A vector space model for automatic indexing. *Communications of the acm*, **18**(11), 613–620.
- Shulayeva, Olga, Siddharthan, Advait, & Wyner, Adam. 2017. Recognizing cited facts and principles in legal judgements. *Artificial intelligence and law*, **25**(1), 107–126.
- Şulea, Octavia-Maria, Zampieri, Marcos, Malmasi, Shervin, Vela, Mihaela, P. Dinu, Liviu, & van Genabith, Josef. 2017a (June). Exploring the Use of Text Classification in the Legal Domain. *Page 5 of : Proceedings of 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts. ASAIL'2017*, London, United Kingdom.
- Şulea, Octavia-Maria, Zampieri, Marcos, Vela, Mihaela, & van Genabith, Josef. 2017b. Predicting the Law Area and Decisions of French Supreme Court Cases. *Pages 716–722 of : Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2017*.
- Thakker, Dhaval, Osman, Taha, & Lakin, Phil. 2009. *Gate jape grammar tutorial*.
- Thenmozhi, D., Kannan, Kawshik, & Aravindan, Chandrabose. 2017. A Text Similarity Approach for Precedence Retrieval from Legal Documents. *Pages 90–91 of : Proceedings of Forum for Information Retrieval Evaluation - FIRE (Working Notes)*.
- Tjong Kim Sang, Erik F., & De Meulder, Fien. 2003. Introduction to the conll-2003 shared task : Language-independent named entity recognition. *Pages 142–147 of : Proceedings of the seventh conference on natural*

- language learning at hlt-naacl 2003 - volume 4*. CONLL '03. Stroudsburg, PA, USA : Association for Computational Linguistics.
- Tumonis, Vitalius. 2012. Legal realism & judicial decision-making. *Jurisprudencija*, **19**(4).
- Ulmer, S Sidney. 1963. Quantitative analysis of judicial processes : Some practical and theoretical applications. *Law and contemporary problems*, **28**(1), 164–184.
- Vapnik, Vladimir N. 1995. The nature of statistical learning. *Theory*.
- Waltl, Bernhard, Matthes, Florian, Waltl, Tobias, & Grass, Thomas. 2016. LEXIA - A Data Science Environment for Semantic Analysis of German Legal Texts. In : *Iris : Internationales rechtsinformatik symposium*. Salzburg, Austria.
- Waltl, BERNHARD, Landthaler, Jörg, Scepankova, Elena, Matthes, FLORIAN, Geiger, THOMAS, Stocker, CHRISTOPH, & Schneider, CHRISTIAN. 2017a. Automated extraction of semantic information from German legal documents. In : *Iris : Internationales rechtsinformatik symposium. association for computational linguistics*.
- Waltl, Bernhard, Bonczek, Georg, Scepankova, Elena, Landthaler, Jörg, & Matthes, Florian. 2017b. Predicting the outcome of appeal decisions in germany's tax law. Pages 89–99 of : *International conference on electronic participation*. Springer.
- Wyner, Adam, & Peters, Wim. 2010. Lexical Semantics and Expert Legal Knowledge towards the Identification of Legal Case Factors. Pages 127–136 of : *JURIX*, vol. 10.
- Wyner, Adam, Mochales-Palau, Raquel, Moens, Marie-Francine, & Milward, David. 2010. Approaches to text mining arguments from legal cases. Pages 60–79 of : *Semantic Processing of Legal Texts : where the Language of Law Meets the Law of Language*. Berlin, Heidelberg : Springer-Verlag.
- Wyner, Adam Z. 2010. Towards annotating and extracting textual legal case elements. *Informatica e diritto : special issue on legal ontologies and artificial intelligent techniques*, **19**(1-2), 9–18.

- Yadav, Vikas, & Bethard, Steven. 2018. A survey on recent advances in named entity recognition from deep learning models. *Pages 2145–2158 of: Proceedings of the 27th international conference on computational linguistics.*