Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-135

Toward Building a Legal Knowledge-Base of Chinese Judicial Documents for Large-Scale Analytics

Amarnath GUPTA [a,1], Alice Z. WANG [b], Kai LIN [a], Haoshen HONG [a], Haoran SUN [a], Benjamin L. LIEBMAN [b], Rachel E. STERN [c], Subhasis DASGUPTA [a], and Margaret E. ROBERTS [a,2]

[a] *University of California San Diego, USA*
[b] *Columbia Law School, USA*
[c] *University of California Berkeley, USA*

Abstract. We present an approach for constructing a legal knowledge-base that is sufficiently scalable to allow for large-scale corpus-level analyses. We do this by creating a polymorphic knowledge representation that includes hybrid ontologies, semistructured representations of sentences, and unsupervised statistical extraction of topics. We apply our approach to over one million judicial decision documents from Henan, China. Our knowledge-base allows us to make corpus-level queries that enable discovery, retrieval, and legal pattern analysis that shed new light on everyday law in China.

Keywords. legal ontology, information extraction, knowledge representation, topic model, Chinese legal documents, text analytics

1. Introduction

In recent years, governments around the world have moved to make information about their legal systems more transparent in order to hold courts accountable to the public and inform legal participants of past court behavior. In Europe, the `OPENLAWS.eu` Consortium is developing an open platform where laws, cases and legal literature from all member states will be made publicly available [24]. In China, the court system recently began mandating that courts upload decision documents to the public website run by the Supreme People's Court (SPC) [17]. While millions of documents are available in each of these contexts, much of the information in the documents is unstructured, and therefore not useful in aggregate for the public. As larger and increasingly more complete collections of legal data become available, there is a corresponding need to construct *publicly available legal knowledge-bases* – formal representations of legal information – from these documents to facilitate their analysis.

The idea of creating legal knowledge-bases, and more generally knowledge-based systems, is not new [23,12,5]. Legal knowledge-bases have been developed in the past

[1]Corresponding Author, E-mail: a1gupta@ucsd.edu
[2]Corresponding Author, E-mail: meroberts@ucsd.edu

for diverse tasks like citation analysis [11], e-governance [12], criminal law analysis [8] and legal advice systems [25]. However, the scope of these tasks has mostly been confined within a single document and in some cases, to small databases: understanding the provisions of a particular law, the argumentation structure of a particular legal case, or the logical reasoning of a particular court procedure. While these analyses can be useful for specific problems, much can be gained by building knowledge-bases to support large-scale analyses that inform legal researchers about the deep characteristics of the complete collection. The goal is to enable "reading at a distance" [13], by capturing knowledge that helps a researcher uncover patterns and emerging trends that can only be mined from a large legal corpus. These analyses can also be reintegrated as a statistically derived knowledge-item into the knowledge base to be reused in subsequent analyses. We call this form of analyses *legal pattern analyses* (LPA).

In this paper, we present an approach for constructing a logically sound legal knowledge-base that allows for large-scale analyses. We apply this approach to a corpus of over 1.1 million judicial records from Henan China.[3] On this example corpus, the purpose for the knowledge-base is to enable a user perform such tasks as:

- *Knowledge-based Retrieval:* "Retrieve the most common defendants in administrative cases." Administrative litigation cases in Chinese law are those where individuals are most likely to challenge the government and therefore are of interest to political scientists studying citizen activism in China [19]. What types of government entities are the most common targets of these cases?
- *Knowledge-based Discovery:* "Discover the issues of dispute for divorce-related cases where the plaintiff is female." Women are known to be disadvantaged in divorce cases under Chinese law [10,16,14]. In what circumstances do they use the legal system to file complaints?
- *Knowledge-based Pattern Analysis:* "Calculate the major differences between cases where plaintiffs file individually versus collectively in administrative cases against the government." Collective action against government entities is viewed as politically sensitive because it could spill over into protest [6], and courts sometimes break up collective claims into individual lawsuits for this reason. On what issues is the government sued by a collection of individuals in the Chinese legal system?

Challenges. There are some inherent challenges in creating a knowledge-base that is conducive to a general set of corpus-level analyses that this paper seeks to address.
Linguistic Variability. Unlike knowledge-bases over formally-written legislation, judicial decision documents (JDDs) have a variable format. For example, arrest records of the defendant in a criminal case or decisions show wide variations in structure and level of detail. Hence, linguistic processing of JDDs for knowledge extraction is more complex, particularly when extracting the same information from the entire corpus.
Need for Heterogeneous Representation. No uniform knowledge representation technique can practically capture all requirements of the knowledge-base. This problem has been reported in prior research. For example, [1] uses description logic for facts and a logic programs for rules, while [2] uses a hybrid rule-based/case-based model for divorce dispute resolution. Our knowledge-base must satisfy retrieval, discovery, and large-scale analytics, each requiring different inputs.

[3] More information about this corpus and what it represents is in [17].

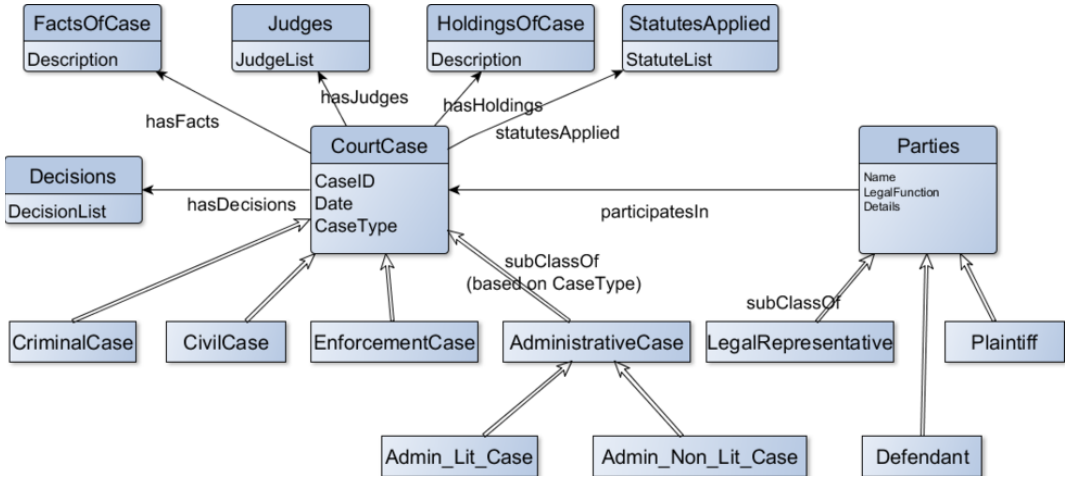

Figure 1. The basic schema of the Chinese Judicial Decision Documents. Several attributes of the CourtCase entity have been omitted for clarity. A simple arrow represents a relationship while a double-shaft arrow represents a subclass relationship. Here "admin_non_lit" stands for "Non-litigation administrative enforcement".

Lack of Completeness. "Knowledge" in the knowledge-base is always incomplete and sometimes inconsistent. Legal ontologies such as Core Legal Ontology (CLO)[4] do not capture concepts and relations that are in the documents but have not been formalized. The challenge is to be able to relate unstructured elements of the text to a legal ontology.

2. The Anatomy of a Chinese Judicial Decision Document

A JDD is written as unstructured text; however, because legal texts are formulaic, sections of the texts can be parsed. We take a data-centric approach to the problem [4] and apply an initial parsing [17] to extract a roughly relational structure, whose extended entity-relationship diagram [9] is shown in Figure 1.[5] This structure is rough because the exact content of a JDD depends on the case type it represents; for example, a criminal case includes information about prior criminal history, whereas an administrative case does not.

Figure 1 shows the schema of a case after initial parsing. "Parties" contains a list of individual party members, including a text description of each party, its role in the court case (e.g., plaintiff), and when applicable its relationship to other parties (e.g., the guardian of a minor defendant). The "factList" contains unstructured text of the legal facts of the case, a summary of primary arguments by the two sides as well, and the facts as established by the court. The "holdingList" contains the legal reasoning and analysis of the judge, which applies the law to the facts of the case. The "decisionList" specifies the legal verdict of the court including any judgments or case-dismissal statements. Both court cases and parties have subtypes. We only show some of the subtypes in Figure 1, and point out that while the subtypes of a court case can be syntactically recognized from the case identifier, the subtypes of the parties can be recognized only through text processing (see Section 3.2).

[4] `www.loa.istc.cnr.it/ontologies/CLO/`

[5] All analysis is done in the original language of the legal documents, Chinese.

Schema Element	Ontology Concept
CourtCase	'Legal case' `CLO:CoreLegal.owl#LegalCase`
CourtCase.caseType	`subClassOf LegalCase`
Parties	`is-participant-in some CLO:CoreLegal.owl #LegalFunction`
Holding	`LegalAnalysisDescription` ⊓ `analyzedBy some Court`
Decision	'Judicial Decision' `CLO:CoreLegal.owl#JudicialDecision`
Fact	'Legal fact' `CLO:CoreLegal.owl#LegalFact`
Statute	'Law' `rdf:type CLO:CoreLegal.owl#Law`
Judge	`dbpedia.org/ontology/Judge`

Table 1. Ontology to Schema mapping in our knowledge-base

3. Our Approach

To build the knowledge-base, we take the following approach. We start with an existing initial ontology, which, although incomplete, maps well to the basic EER diagram in Figure 1. We adjust this ontology to ensure alignment with the all elements of the EER schema. Next, we extract semi-structured information from the JDDs with two different techniques. Then, we conduct a two-way annotation process from the ontology to the JDDs and from the JDDs back to the ontology. The annotated ontology and documents are stored in a scalable polystore system [7]. Finally, we compute a family of topic models on the data to create statistical representations of the remaining unstructured text.

3.1. An Initial Ontology

Our initial ontology is derived from two well-known ontologies in the domain of legal knowledge representation. The upper ontology is DOLCE+DnS Ultralite (DUL) ontology,[6] which was chosen primarily because of its elaborate coverage of the concept space including `social objects`, `Conceptual Objects` (called `concepts`) and `situations`. The domain ontology is adapted from the Core Legal Ontology (CLO) which, in turn builds on DUL and the Information Object Ontology Lite.[7] The CLO introduces the basic concepts of jurisprudence including `law, legalFunction, legalDescription, crime` and `legallyRelevantCircumstance`.

Schema Alignment. The schema elements of Figure 1 are first mapped to the ontological concepts in Table 1. Next, we directly relate the `caseType` attribute to the subclasses of the `LegalCase` concept. The mapping for `Parties` implies that every party in the list of parties plays the role of a `legal function` as specified in the CLO, which designates plaintiffs, defendants, attorneys, etc. as legal functions that are fulfilled by concrete instances of `NaturalPerson` entities. Similarly, a statute is interpreted as an individual instance (`rdf:type`) of the concept of `law`. The `holding` is mapped to our extension of the CLO which admits the concept of `LegalAnalysisDescription` ⊏ DUL.`Description`. This common structure of cases can be extracted from each case fairly easily, but still draws only basic information from each decision.

Initial Ontology Augmentation. In CLO, the concept of `law` (corresponding to a statute in the schema) is the subclass of a `legal description`, which is the subclass of the

[6] `http://www.ontologydesignpatterns.org/ont/dul/DUL.owl`

[7] `www.ontologydesignpatterns.org/ont/dul/IOLite.owl`

CLO concept `description`. CLO also defines the concept of `legal case` as a derived subclass of the DUL concept of `legal fact` which depicts situations depending on legal norms. For example, the legal case called `crime` satisfies norms of incrimination. But how does a concrete crime type such as arson ("setting fire" – 放火) relate to the concept of `law`? We extend CLO by creating a subclass tree under `law`. The tree is derived from the case classification documentation issued by the Supreme People's Court in China. This subtree categorizes laws at a level of detail that can be more effectively correlated with the judicial decisions documents. For example, the `tort liability law` (侵权法) is a subclass under the concept `civil laws` (民法) branch. We also introduce a new subclass hierarchy under `legal case` to represent a hierarchy of legal case types (e.g., `Product transporter responsibility dispute`, 产品责任纠纷, is a superclass of 产品运输者责任纠纷). The rest of the ontology classes where assigned to the JDDs based on existing categories available from the SPC website.

The resulting ontology is checked for consistency with Protégé's *Hermit reasoner* and then stored in a graph database system (Neo4J) through *SciGraph*,[8] an ontology manager developed over Neo4J. SciGraph uses the OWL API to decompose each axiom and a model conversion algorithm to re-represent them as graph. The graph nodes are typed and can represent concepts, individuals and anonymous classes; the edges represent subClassOf, equivalence, union (⊔), intersection (⊓) that are used in the ontology. SciGraph is a lossless representation of the asserted ontology – its edges capture quantifiers (i.e., `some, only`, ...) and edge properties like transitivity. Simple inference procedures like transitive closure computation are implemented through graph-based operations. This implementation supports knowledge-based querying (Section 4).

3.2. Information Structuring with Text Analytics

While the ontology and its mapping to the JDD schema creates a preliminary connection between them, much of the information content of the JDD is still buried inside its unstructured content. We will describe two methods by which we extract information from text into a *semistructured* (JSON) representation. This semistructured (labeled, ordered trees) model provides an additional advantage that the extracted information can be stored in a scalable semistructured database like MongoDB.

Term-Anchored Context-free Grammar. Our first approach of knowledge extraction from text applies context-free grammar rules to segments (e.g., `Parties`) of a JDD where "anchor terms" from a large but fixed vocabulary must appear . Our intention is to extract a complex set of properties of entities mentioned in the document, and the complex relationships between these entities. To see why this is important, consider the analyst's question: "Do repeat offenders get harsher sentences?" To determine whether a defendant is a "repeat offender" one has to extract the criminal record from the description of the defendants. In our example collection of JDDs, most criminal defendant descriptions present a history of their criminal record, although these descriptions are not standardized.

We take a grammar-based approach to information extraction from text. We argue that since the description of criminal records is "stylized" natural language, its grammar falls somewhere between a pure, context-free grammar (CFG) and an arbitrary context-sensitive grammar. We postulate that if we recognize a handful of *anchor terms* in the

[8] `https://github.com/SciGraph/SciGraph`

Action Prohibition	Original Judge Affirmation	Remand
Case Withdrawal	Custody of Child	Monetary Compensation
Confirm Illegality	Punishment Announcement	Property Distribution
nolle prosequi	Judgment Revocation	Penalty Abatement
Confiscation	Compulsory Execution	Divorce Approval

Figure 2. Examples of the 38 sentence categories parsed by our sentence modeling scheme

text, then the rest of the text can indeed be treated as though it satisfies a CFG grammar. The anchor terms are identified through different dictionaries such as the dictionary of law enforcement actions and the dictionary of charges that can be brought by the police. The grammar rules are centered around terms in these dictionaries, such as "imprisonment" or "drug possession," then a context-free rule can correctly extract the prior record. A preliminary evaluation shows that parties are correctly assigned in 85% of the cases. The errors primarily occur due to complex unparsed sentences, and in documents where there is no specific party section but the case title carries the information about parties.

Judicial Sentence Models. Our second approach to information extraction relies on sentence modeling and is applied toward understanding the court decisions. The first step in this approach uses the output of the CFG party extraction described before to instantiate participant names, aliases, and their roles in the decision section of the document. A Jaccard coefficient based scoring method is used for inexact matches and abbreviations.

The second step creates a classification of the types (Figure 2) of verdict sentences through a series of matching rules. As a simple example, the sentence "Dissolving the plaintiff Xu Shouzhen's and the defendant Liu Weihong's marriage relationship" can be easily categorized as a *marriage dissolution* verdict because it has the sentential pattern `removing <plaintiff-phrase> and <defendant-phrase> marriage relationship`. Similarly, a pattern `Criminal + <name> + commit + ... + crime` classifies it as a *Punishment Announcement* verdict. The complexity of the classification rules arises from the syntactic variations in the sentence structure and the context sensitive nature of the text. The recognizer of a compensation case may use synonyms and expression variants like "to ... compensate ..RMB." (向...赔偿...元..) In other cases, the classification rules must look at multiple consecutive sentences to provide adequate context.

Once a sentence is classified into one or more of 38 classes, we reanalyze the sentence to identify model parameters. For example, a compensation case will have payer(s), a set of payee(s), and a compensation amount for each payer-payee combination. When the compensation amounts are explicitly specified, we record them; when clauses like "equally paid" are used, they are specifically interpreted to determine the actual compensation amount. Often verdicts have additional clauses such as "payable once every year by October 1" – these clauses are captured within a "comment" node in the resulting tree.

If a sentence corresponds to multiple possible models, a *conflict resolution* process is applied. For example, a verdict that affirms the original judgment always includes rejecting other requests. This verdict will be identified as both "Affirm Original Judgment" and "Reject Requests" types in model selection stage. In this case we order the verdict types by their frequency of occurrence, and select the top scoring model. Our preliminary evaluation shows that the sentence classification has over 95% accuracy for tweets

with correctly parsed parties; 5% error-rate is due to complex sentences in the decision section that could not be parsed properly.

3.3. Bidirectional Mapping

The next step is a two-way mapping from the ontology to the restructured JDDs and from the JDDs back to the ontology. The rationale for the two-way mapping comes from the observation that analysis of a JDD corpus yields new concepts, individuals and relationships that should be included in an application ontology that "hangs from" the domain ontology from Section 3.1. Simultaneously, the process creates an ontological annotation into the semistructured data that explicitly marks ontological concepts/individuals to the JDDs. For instance, the term "Entrusted Agent" is a new instance of `legalFunction`, "arrest record" is a new concept, and $civil-case \overset{describedBy(some)}{\longrightarrow} civilLaw$ is a new relationship that would be added to the ontology. In the other direction, we annotate the ontology with JDD-indexing mappings such as $attorney \overset{mapsTo}{\longrightarrow} entrustedAgent \overset{occursIn}{\longrightarrow}$ 100.*Parties*.3 where the first element `attorney` is a CLO concept, the second element `entrustedAgent` is a party type and the third element `100.Parties.3` represents the 3rd Parties element in document having ID 100. These are encoded in the JDD data as JSON element `mappedEntityType` added to every recognized instance of a concept.

3.4. Leveraging Unstructured Text

Last, we leverage unsupervised natural language processing to extract information from the remaining unstructured text. Topic models have been amply used for tasks related to legal document understanding as diverse as extracting domain and argument related words [20], legal document summarization [15], finding differences in decision patterns across courts [18] and shifts in the content of the case-law of international courts over time [21]. They identify "topics," or clusters of frequently co-occurring terms in a collection of documents [3].

In our setting, we estimate the Structural Topic Model (STM) [22] over the results of a query which subsets the data based on some conjunctive predicate P. The predicates may place conditions on metadata (date $>$ 1/1/2014), or document content (e.g., Facts.factList contains "pollution"), or on derived structures (e.g., verdict type = "Punishment Abatement") or any conjunction of the above. Further the topic model can be run on any subset of the parts of the document (e.g., only facts and decisions) – this subset is called the "scope" S of the model. This *PS conditioning* allows us to run multiple topic models on the same collection of legal documents, giving insight into the topics tailored to the analyst's interest. Each *PS* pair has a ranked topic-term list, and a ranked topic-document list. Further, if a term discovered in a topic belongs to the ontology, it is annotated by the ID of the ontology term. Ontological annotations can also be included in the estimation of the topic model by including them as covariates in the STM.

We illustrate the effect of *PS-conditioning* on the estimated topics by running a 30-topic STM on all civil cases, restricting the scope of the model to text in facts and holdings. Row 1 of Table 2 shows one interesting topic retrieved from the model related to medical care. Rows 2 and 3 show how this topic becomes more refined as increasingly restrictive predicates are added.

Predicate	Scope	Topic
(none)	facts, holding	Hospital, medical expenses, disability, compensation, care, plaintiff, calculation, cost, injury, transportation
contains(document, 'disability')	facts, holding	identification, calculation, disability, compensation, forensic, identification, hospitalization, mental
contains(document, 'disability') AND date > '1/1/2014'	facts, holding	Work injury, labor, payment, disability, company, work injury insurance, arbitration, subsidy, salary, disposable [income]

Table 2. Topic refinement under increasingly restrictive predicate conditioning.

4. Toward Knowledge-based Retrieval and Discovery

The query and discovery infrastructure of our legal knowledge-base is polystore called AWESOME [7] which is a data management system developed over multiple data management systems including Neo4J, AsterixDB (for JSON), PostgreSQL (for relational tables) and Apache Solr (for text indexing) together with SciGraph. A detailed description of the system is beyond the scope of this paper. Here we show three examples of how the implemented knowledge-base facilitates the tasks outlined in Section 1.

Retrieval. For the retrieval query, "Find the most common government entities that are defendants in administrative litigation cases," can be executed as follows: (a) from the ontology, find all cases that are type "administrative litigation cases", (b) extract all defendants from these cases, (c) remove the particular location of government from the party's name (d) return most frequent entities that are defendants, ordered by their frequency of occurrence. The results, shown in Table 3, provide a summary of the top 10 government bodies that are the target of contention in the Chinese legal system. In particular, levels of government dealing with land, family planning, benefits, and public security are most likely to appear as defendants in these cases.

1.	People's Government	2.	Public Security Bureau
3.	Human Resources and Social Security Bureau	4.	Land and Resources Bureau
5.	Housing and Urban Construction Bureau	6.	Real Estate Authority
7.	Population and Family Planning Commission	8.	Urban and Rural Planning Bureau
9.	Administration for Industry and Commerce	10.	Real Estate Authority

Table 3. Most common government entities that are defendants of administrative litigation cases, Henan.

Discovery. Our example query, "Discover the issues of dispute for divorce-related cases where the plaintiff is female" can be interpreted as follows: (a) from the ontology, find all subclasses of case types with the term "divorce" in them (ontology fragment), (b) based on the ontology IDs of these concepts, identify the JDDs that have been marked with these IDs (mapping fragment) , (c) filter those JDDs from (b) where the value of the party with `role: plaintiff` has `gender:female` (semistructured fragment), (d) with the facts and holdings sections from cases in (c), run the topic model with an increasing number of topics, and in case store the dominant topics. A topic is called "common" if the number documents supporting it is high, and the same topic, occurs across multiple topic counts. Rarer topics are discovered as the number of topics is higher, yielding finer

1. **Child Support**	"Maintenance", "care", "child", "child development", "daughter's marriage, "life", "paid", "daughter grow up healthy"
2. **Domestic Violence**	"Neck treatment", "beat", "pinch", "drinking", "relapse", "perforation of ear and eardrum", "threatening"
3. **Division of Property**	"Washing machine", "sofa", " Cabinet", "Haier color TV", "Dresser", "coffee table", "wall units", "water dispenser"
4. **Inadequacy of Alimony**	"Income", "education", "living expenses", "custody", "unreasonable demands", "visitation rights", "born out of wedlock", "usufruct"
5. **Reconciliation**	"Tolerant", "shortcoming", "mutual trust", "communication", "harmonious", "mutual understanding", "harmonious", "exchange"

Table 4. Common (1-4) and less common (5) issues for divorce cases where the plaintiff is female.

Topic	Prop. Individual Cases	Prop. Collective Cases
Withdrawals: Withdrawn, granted, withdrawn, charged, process, examined, halved, voluntarily, should	0.20	0.18
Public Security: Penalties, transcripts, decisions, inquiries, decisions, public security, law and order, detention, management, beating	0.04	0.01
Forest Rights: contract, forest warrants, trees, awarded, contract, Li Baowei, civil, signed, woodland, publicity	0.01	0.10
Land Use Rights: land, homestead, use, use certificate, issue, dispute, use rights, collective, land, area	0.10	0.15

Table 5. Topics most associated with individual and collective plaintiffs in administrative litigation cases.

topics. In Table 4 the first 4 topics are common, while the last topic appears only when the topic count > 35, and is supported by 343 JDDs.

Legal Pattern Analysis. Last, we show how the knowledge-based representation can be used to uncover patterns in legal cases on a corpus level. To do this, we turn to our example query "Calculate the major differences between cases where plaintiffs file individually versus collectively in administrative litigation cases." To do this we (a) select administrative litigation cases (b) retrieve the parties in all selected cases (c) distinguish between those with only one plaintiff from collective parties (d) run a topic model that estimates the relationship between the topics and the plaintiff type.

Table 5 shows the topics most associated with collective plaintiffs and those most associated with individual plaintiffs. Interestingly, one of the topics more associated with individual cases than collective cases is that of case withdrawal, thought to be a shortcoming of the Chinese legal system [19]. Land cases are most likely to be filed collectively, often overy land use rights or forest rights.

5. Conclusion

We have developed an initial approach for constructing a legal knowledge base that facilitates corpus-level analyses. By combining ontologies, semistructured representations of legal sentences, and unsupervised estimation of topics on remaining unstructured data, we allow for flexible analyses that retrieve, discover, and estimate patterns at the corpus level in Chinese legal documents. Our future work includes automatic assignment

of ontological classes to JDDs using the laws applied to a case. We hope our approach provides a framework for knowledge representation that facilitates our understanding of legal systems as a whole.

References

[1] M. Alberti, A. Gomes, R. Gonçalves, J. Leite, and M. Slota. Normative systems represented as hybrid knowledge bases. *Computational Logic in Multi-Agent Systems*, pages 330–346, 2011.

[2] M. Araszkiewicz, A. Łopatkiewicz, A. Zienkiewicz, and T. Zurek. Representation of an actual divorce dispute in the parenting plan support system. In *Proc. of the 15th Int. Conf. on AI and Law*, pages 166–170. ACM, 2015.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. of Machine Learning Research*, 3(Jan):993–1022, 2003.

[4] L. K. Branting. Data-centric and logic-based models for automated legal problem solving. *Artificial Intelligence and Law*, 25(1):5–27, 2017.

[5] A. Cernian, D. Carstoiu, O. Vasilescu, and A. Olteanu. Ontolaw-ontology based legal management and information retrieval expert system. *J. of Control Engg and Applied Informatics*, 15(4):77–85, 2013.

[6] X. Chen. *Social protest and contentious authoritarianism in China*. Cambridge University Press, 2012.

[7] S. Dasgupta, K. Coakley, and A. Gupta. Analytics-driven data ingestion and derivation in the awesome polystore. In *IEEE Int. Conf. on Big Data (Big Data)*, pages 2555–2564. IEEE, 2016.

[8] M. El Ghosh, H. Naja, H. Abdulrab, and M. Khalil. Towards a legal rule-based system grounded on the integration of criminal domain ontology and rules. *Procedia Computer Science*, 112:632–642, 2017.

[9] R. Elmasri and S. B. Navathe. *Fundamentals of Database Systems*. Pearson, 2015.

[10] L. H. Fincher. *Leftover women: The resurgence of gender inequality in China*. Zed Books Ltd., 2016.

[11] F. Galgani, P. Compton, and A. Hoffmann. Lexa: Building knowledge bases for automatic legal citation classification. *Expert Systems with Applications*, 42(17):6391–6407, 2015.

[12] T. F. Gordon. A use case analysis of legal knowledge-based systems. In *JURIX*, 2003.

[13] J. Grimmer and B. M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297, 2013.

[14] X. He and K. Ng. Pragmatic discourse and gender inequality in china. *Law & Society Review*, 47(2):279–310, 2013.

[15] R. Kumar and K. Raghuveer. Legal document summarization using latent dirichlet allocation. *Int. J. of Computer Science and Telecommunications*, 3:114–117, 2012.

[16] K. Li. "what he did was lawful": Divorce litigation and gender inequality in china. *Law & Policy*, 37(3):153–179, 2015.

[17] B. L. Liebman, M. Roberts, R. E. Stern, and A. Z. Wang. Mass digitization of chinese court decisions: How to use text as data in the field of chinese law. *Social Science Research Network Collection*, June 2017.

[18] M. A. Livermore, A. Riddell, and D. Rockmore. Agenda formation and the us supreme court: A topic model approach. 2016.

[19] N. Mahboubi. Suing the government in china. *Democratization in China, Korea, and Southeast Asia. Londres: Routledge*, pages 141–155, 2014.

[20] H. Nguyen and D. J. Litman. Extracting argument and domain words for identifying argument components in texts. In *Argument Mining at NAACL-Human Language Technology*, pages 22–28, 2015.

[21] Y. Panagis, M. L. Christensen, and U. Sadl. On top of topics: Leveraging topic modeling to study the dynamic case-law of international courts. In *JURIX*, pages 161–166, 2016.

[22] M. E. Roberts, B. M. Stewart, and E. M. Airoldi. A model of text for experimentation in the social sciences. *J. of the Amer. Stat. Assoc.*, 111(515):988–1003, 2016.

[23] A. Stranieri and J. Zeleznikow. The evaluation of legal knowledge based systems. In *Proc. of the 7th Int. Conf. on AI and Law*, pages 18–24, 1999.

[24] R. Winkels et al. The openlaws project: Big open legal data. In *Proc. of the 18th Int, Legal Informatics Symposium IRIS 2015*, pages 189–196, 2015.

[25] T. Zurek. Conflicts in legal knowledge base. *Foundations of Computing and Decision Sciences*, 37(2):129–145, 2012.