Data Science Techniques for Law and Justice

Current State of Research and Open Problems

Alexandre Quemy[1,2]

September 14, 2017

[1]: IBM Analytics, Cracow Software Lab
[2]: Faculty of Computing, Poznan University of Technology

Context & Problems

Litterature Review

Hypergraph CBR

CONTEXT & PROBLEMS

Initial observation

Law is complex:

1. Access to knowledge from different sources.
2. Collect, connect, and exploit knowledge .
3. *Messy concept*:
    - grey areas of interpretation,
    - many exceptions, non-stationarity,
    - deductive,
    - inductive reasoning,
    - non-classical logic,
    - …

The problems

Even for Judges, Lawyers & Legal experts !

1. More and more law texts include quantitative critera.
2. What is a good or bad justice decision?
3. How is really taken a justice decision?

Lot of philosophical and sociological work but can data science help?

The problems

1. Predicting the outcome of a case given the legal environment. (Prediction)
2. Building a legal justification, given some facts, a set of law texts with the jurisprudence and an outcome. (Justification)
3. Taking the *best* decisions w.r.t. the legal environment dynamics and some criteria. (Decision)
4. Modifying the legal environment dynamics to match some criteria. (Control)

Remarks:

1. Why "Prediction" exists since "Justification" can provide an outcome?
2. "Decision" and "Control", two sides of the same medal.
3. The litterature mostly study the "Prediction".

"*The level of a fine must be sufficiently high both to punish the firms involved and to deter others from practices that infringe the competition rules. [...] The basic amount is calculated as a percentage of the value of the sales connected with the infringement [...]. The percentage of the value of sales is determined according to the gravity of the infringement (nature, combined market share of all the parties concerned, geographic scope, etc.) and may be as much as 30 %.*"[1]

---

[1] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV%3Al26118

How hard is Prediction?

Best legal experts on SCOTUS [?]:

1. 67.4% correct prediction for judges
2. 58% correct prediction for cases

MARKET LANDSCAPE FOR LEGAL ANALYTICS

Research to Business, why does it matter ?

In US:

1. 1,300,000 licensed attorneys in the United States.
2. 58 million consumers in the U.S. sought an attorney.
3. 200 law schools.

In France:

- 60000 lawyers, **+41% in 10 years**, 8355 judges,
- in 2014, **791.448 basic missions** for juridical help.
- around 50 law universities
- **legal analytic is a priority axis of development**

As February 2016:
"The total addressable market for legal software – both corporate law departments and law firms—is **15.9 billion annually**; the market spends $3 billion each year; law departments spend $1.5 billion annually on 11 types of software—from matter management to compliance to legal analytics – in a market with a **$6.5 billion potential** and; while all technology segments are growing." — InsideCounsel

MARKET SIZE: CORPORATE LEGAL SOFTWARE

Corporate Legal Software	2015	2019	CAGR	Total Addressable Market (with 100% penetration)	2015 Penetration
e-Billing	$202m	$235m	4%	$524m	39%
Matter management	$195m	$279m	9%	$611m	32%
Contracts management	$187m	$346m	17%	$769m	24%
Governance & compliance	$147m	$270m	16%	$825m	18%
IP management	$140m	$194m	8%	$768m	18%
Legal hold	$129m	$158m	5%	$465m	28%
Document management	$127m	$183m	10%	$437m	29%
Legal project management	$102m	$198m	18%	$649m	16%
Knowledge management	$99m	$259m	27%	$657m	15%
Legal analytics	$73m	$145m	19%	$488m	15%
Collaboration tools	$61m	$92m	11%	$366m	18%

LITTERATURE REVIEW

Plenty of approaches:

1. Stochastic Block Model [?]: 77%
2. NLP + SVM [?]: 79%
3. Random Forest [?]: 69.7%
4. ...

Remarks:

1. Two categories: LK vs non LK.
2. Realism conforted in both!
3. Not a single player game! [?, ?]

All "predictions" are not equals!

1. General
2. Robust
3. Fully predictive

And the winner is... Random Forest [?]

CBR cycle [**?**]:

1. Search for the most related past cases, either by filtering the irrelevant cases or selecting the closest ones depending on a metric and a KNN algorithm.

2. Adapt the best case solution to the new case.

3. Evaluate and revise the proposed solution, including at least why the solution is not satisfying.

4. Integrate the solution to the database.

Perform better than Rule-based [**?**] but:

1. similarity and relevance of precedent cases are dynamic,
2. non-stationary as social and governmental laws evolve

Novel approach: learning rules a set of similar cases, then predict and justify with them [**?**].

AA [**?**] = Toolbox for non-monotonic reasoning.
AA cycle[**?**]:

1. Defining the arguments and the relation(s) between them.
2. Valuating the arguments, etc.
3. Selecting some arguments using some *semantic*.

1. Promising but very normative.
2. Never applied on real data!
3. Good at providing explanation.

AA-CBR

1. Arguments as past cases [**?**, **?**],
2. Rule learnt from past cases [**?**]

	Information	LK	General	Robust	Fully Pred.	Extra Data	Just.
Pred. Models							
[34, 42]	Data-driven	no	no	no	yes	past cases	no
[19]	Data-driven	no	yes	yes	no	past votes	no
[27]	Data-driven	no	yes	yes	yes	past cases	no
[2]	Data-driven	no	yes	no	yes	past cases	no
Ideal Point							
[43, 44]	Data-driven	no	yes	no	yes	non-legal	no
[33, 38, 39]	Data-driven	no	yes	yes	yes	past votes	no
[30, 45]	Data-driven	no	yes	no	yes	amicus	no
[23]	Data-driven	no	yes	no	yes	opinions	no
CBR							
[3]	Rule-based	yes	no	no	yes	past cases & legal factors	yes
AA							
[8, 9]	Both	yes	no	no	yes	past cases & legal factors	yes
[35, 36]	Both	yes	yes	yes	yes	norm	yes
[14]	Rule-Based	no	yes	yes	yes	-	yes
[6, 40]	Both	no	yes	no	yes	-	-
[31]	Rule-based	no	yes	no	yes	-	-
AA-CBR							
[50]	Both	no	no	no	yes	past cases	partly
[5]	Both	no	yes	yes	yes	past cases	partly
[37]	Both	no	yes	yes	yes	past cases	partly

HYPERGRAPH CBR

THE END

Thank you for your attention!
Questions?