Analysis of US Patent Litigation Outcomes from 1963-2015: A Machine Learning Approach

Hunter Sporn ‘21, Nicholas Schmeller ‘21, Joseph Puryear III ‘19

PRINCETON UNIVERSITY

Motivation

- Thousands of patents are filed in the USA on an annual basis
- Most patent lawsuits costs the parties millions of dollars and years of litigations each
- The ability to improve the current modeling of patent litigation would greatly improve the performance of the process

Goal

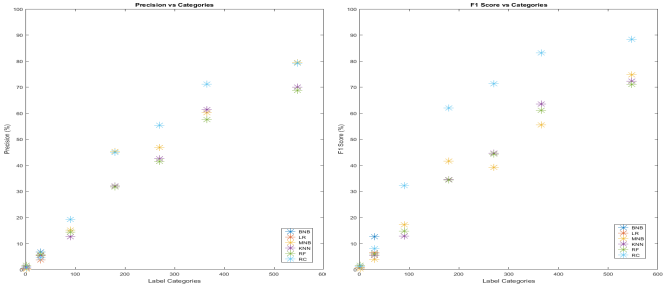
- Use U.S. Patent and Trademark Office data from 1963-2015 to develop a robust model to determine if a case will be a significant burden on the legal system

Related Work

- Little literature exists exploring the application of machine learning onto patent litigations
- Machine learning in law allows for “new techniques” like [9]:
  - “Help attorneys improve legal strategies”
  - “Conduct informed fact discovery”
  - “Provide testifying experts with the most complete set of relevant information”
  - “Prepare analyses at a previously unseen level of granularity”
- Ruger et al. explored the 2002 Term cases of the Supreme Court in 2004 [7]:
  - The statistical model was able to determine whether the court affirmed or reversed a decision 75% of the time
  - Meanwhile, law experts collectively performed at 59.1% accuracy
- Katz et al. expanded upon Ruger’s et al. initial research in 2017 by exploring the Supreme Court cases from 1816-2015 while applying machine learning [8]:
  - Their random forest classifier model was able to achieve 70.2% and 71.9% accuracy when predicting the case’s outcome and the justice’s individual vote, respectively
  - This model performs slightly lower than Ruger et al., but is more widely applicable across nearly two centuries of cases
- The Supreme Court covers various topics from tax law to patent law to administrative law to much [8]
- Pinheiro et al. discusses how machine learning could benefit patent litigation and assist in deciding whether parties should decide whether to negotiate a settle or engage in costly litigation [9]

Methods

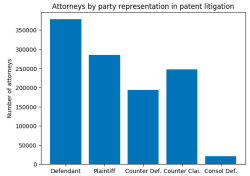
- Initial Dataset
  - Consists of ~74,000 cases from 1963-2015 with 5 different files containing the litigating parties, their attorneys, results, locations, and dates
- Case duration with respect to courts and judges:
  - The initial dataset was dropped to 7,288 cases after removing any cases that did not have the file date, closed date, court name, or judge assigned to them
  - Court names and judges were label encoded; The duration of each case was determined
  - Various models were applied with varying the categories; Categories were segmented into daily, monthly, 3 months, 6 months, 9 month\s, 1 year, and 1.5 years.



  - Precision and F1 both improve as the number of categories decrease

Case Study: “Simple” Data Analysis

- In order to understand how and why indicators such as court location, judge behavior, and textual analysis on case name match with “burden factors”, consider an example of correlation between two datasets that don’t necessarily use machine learning techniques
- Consider the connection between number of attorneys on a side in a case (“burden factors” in terms of personnel required) and role of the side



Results

- Clear signal for bag-of-words representation with minimum threshold of 2 occurrences
  - Diverse names suggest that if a word is involved in more than one case, it can be associated with a ”burden factor”
- Label encoding of courts and judges provides a roughly linear correlation between accuracy/precision and broader label categorization
  - The existence of a tradeoff in accuracy for higher precision indicates a strong signal in the court location and specific judges for burden factor
- Location of courts by circuit adds strong signal as well, correlation between circuit and case duration can be assumed to be partly influenced by individual court and judge behaviors, but we posit the existence of latent factors in circuit location
- The clearest indication for “burden factor” is a rough tie between number of documents involved and number of days spent on the case
  - Combining the two indicators seems promising but adds new layers of noise into the data which makes models difficult to understand and multimodal

Conclusions

- The relationship between publicly available metadata and burden factors is significant and worth studying further
- Companies whose business model relies on how expensive a case is should seek to understand this correlation
- Further work should focus on text analysis of initial case documents to predict early in a case what kind of resources will be spent on it

References

- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Pret-tenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Per-rot, and E. Duchesnay. Scikit-learn: Machine learning in Python.Journal of Machine LearningResearch, 12, 2011.
- What is patent litigation?Morningside Translations, 2018.
- US Patent Trademark Office. Detailed patent litigation data on 74k cases, 1963-2015. 2018.
- Alan C. Marco, Asrat Tesfayesus, and Andrew A. Toole. Patent litigation data from us districtcourt electronic records (1963-2015).USPTO Economic Working Paper No. 2017-06, 2017.
- Lei Mei and E. Robert Yoches. Unique aspects of u.s. patent litigation.Lexis Nexis China LegalReview, 2007.
- William C. Spence, Jason Wejnert, and Brian Beck. Global patent litigation strategy.IAMMedia, 2018.
- Theodore W. Ruger, Pauline Kim, Andrew D. Martin, and Kevin Quinn. The supreme courtforecasting project: Legal and political science approaches to predicting supreme court deci-sionmaking.Columbia Law Review, 104, 2004.
- Daniel Martin Katz, Michael J. Bommarito II, and Josh Blackman. A general approach forpredicting the behavior of the supreme court of the united states.PLOS ONE, 2017.
- Lisa B. Pinheiro, Jimmy Royer, Mihran Yenikomshian, Nick Dadson, and Paul E. Greenberg.Using machine learning in litigation.Analysis Group Economic, Financial, and Strategy Con-sultant, 2017.
- Special Thanks to Jacob Zimmer for the Poster Template.