Classifying Legal Expertise from Structured and Unstructured Data

Chase Basich
Stanford University
Stanford, CA 94305
cdbasich@stanford.edu

Austin Chambers
Stanford University
Stanford, CA 94305
austin1@stanford.edu

1 Introduction

1.1 Problem

One of the biggest problems law firms face is in locating lawyers with select expertise or past legal work relevant to a given situation. Without the ability to locate the right lawyers, the established legal matter teams may not be competitive. Without the ability to find past work that is related and relevant, firms deprive themselves of the ability to learn from their experiences and to best pitch for new business within a particular area. This results in a tremendous amount of wasted opportunity and time, leading law firms and vendors to invest time in finding a solution to this problem.

1.2 Firm's Attempts

The traditional approach to this problem is to invest in (1) developing a taxonomy to describe a valuable dimension, and (2) to develop internal processes to manually classify data into this taxonomy. Unfortunately, these "knowledge management" processes rarely work. This is because because the rate at which new information comes in is greater than the rate at which it can be classified, resulting in a situation where the value of this undertaking is never fully realized. Any further business processes or tools introduced to help facilitate this effort then become are at odds with an organizational culture that fights for the bottom line. Nevertheless, it is the past knowledge management efforts to apply structure to data that deliver a great amount of value.

1.3 Proposed Solution and Goal

We believe that the combination of vast unstructured data and low-quality structured data will be sufficient for us to solve the legal expertise problem with accuracy. Since the private data within a law firm is unavailable, we will use the public marketing articles about past work (matter narratives) and attorney profile pages as a proxy. To test our hypothesis, we develop a supervised learning algorithm and test the accuracy against the data from two different law firms. If our accuracy is sufficient for both law firms, we believe this approach will extend to the private data that exists within firms.

2 Datasets

The attorney profiles and matter narratives from the two datasets have a marketing and sales orientation. It has been made publicly available so that potential customers can learn about the excellent attorneys or past work that may be relevant to the work at hand. Despite the marketing bias of this information, the data is ideal for our purposes because it meets the following criteria:

1. The data is relevant to the topic of expertise classification.
2. The data contains contains unstructured data in the form of narrative summaries or attorney biographies.

3. The data contains structured data in the form of tags and links, but this data is inconsistent and occasionally incorrect.

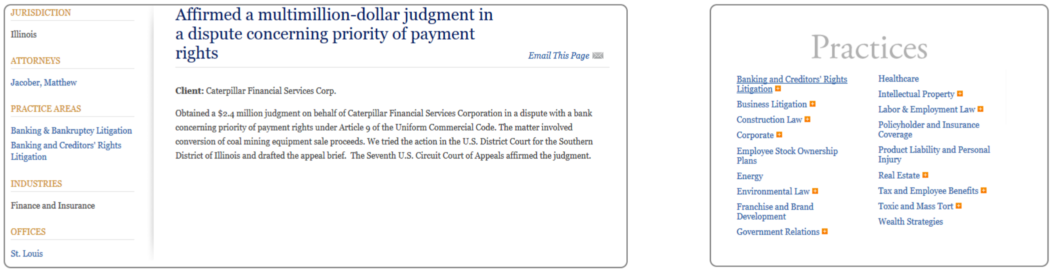
Accordingly, this data should serve as an excellent proxy for similar information relevant to expertise that exists as structured and unstructured data within a firm.

2.1 Dataset Retrieval

The datasets for both firms were retrieved by accessing the website of both law firms and manually exporting search results containing links to attorney profiles and matter narratives. The URLs from these links were extracted, loaded into relational database, and C# application was developed to created to retrieve the HTML page for each URL. The C# application then used the consistent XML structure of the HTML to segment the data using XPATH into attorney, practice, and narrative tables.

2.2 Dataset Comparison

The first dataset contained 520 legal matter narratives and 143 attorney profiles that we have scraped from the Lathrop and Gage website (http://www.lathropgage.com/experience.html). Each narrative includes a title, a summary, a list of involved attorneys, practice areas, offices of the law firm, and the related industry. Occasionally, the associated client involved in the matter can also be identified.



JURISDICTION

Illinois

ATTORNEYS

Jacober, Matthew

PRACTICE AREAS

Banking & Bankruptcy Litigation
Banking and Creditors' Rights Litigation

INDUSTRIES

Finance and Insurance

OFFICES

St. Louis

Affirmed a multimillion-dollar judgment in a dispute concerning priority of payment rights

Email This Page

Client: Caterpillar Financial Services Corp.

Obtained a $2.4 million judgment on behalf of Caterpillar Financial Services Corporation in a dispute with a bank concerning priority of payment rights under Article 9 of the Uniform Commercial Code. The matter involved conversion of coal mining equipment sale proceeds. We tried the action in the U.S. District Court for the Southern District of Illinois and drafted the appeal brief. The Seventh U.S. Circuit Court of Appeals affirmed the judgment.

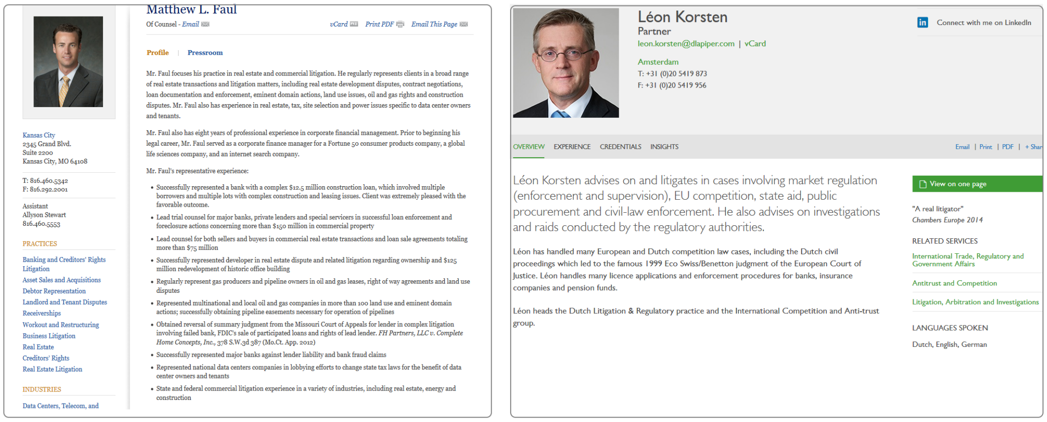
Practices

Banking and Creditors' Rights Litigation
Business Litigation
Construction Law
Corporate
Employee Stock Ownership Plans
Energy
Environmental Law
Franchise and Brand Development
Government Relations

Healthcare
Intellectual Property
Labor & Employment Law
Policyholder and Insurance Coverage
Product Liability and Personal Injury
Real Estate
Tax and Employee Benefits
Toxic and Mass Tort
Wealth Strategies

Figure 1: Matter Narratives and Practices in the first dataset (Firm A)

The second dataset contained 3659 attorney profiles without any matter narrative data from DLA Piper (https://www.dlapiper.com/en/us/). The depth of the narrative on this data set was less detailed, but there were far more samples in this set. Although the attorney biography data included same type of information across firms, the labels that are associated with each attorney vary greatly. Most attributes assigned to a specific attorney are not assigned across all attorneys (e.g., the "languages" attribute).



Matthew L. Faul

Of Counsel - Email

vCard Print PDF Email This Page

Profile | Pressroom

Mr. Faul focuses his practice in real estate and commercial litigation. He regularly represents clients in a broad range of real estate transactions and litigation matters, including real estate development disputes, contract negotiations, loan documentation and enforcement, eminent domain actions, land use issues, oil and gas rights and construction disputes. Mr. Faul also has experience in real estate, tax, site selection and power issues specific to data center owners and tenants.

Mr. Faul also has eight years of professional experience in corporate financial management. Prior to beginning his legal career, Mr. Faul served as a corporate finance manager for a Fortune 50 consumer products company, a global life sciences company, and an internet search company.

Mr. Faul's representative experience:

- Successfully represented a bank with a complex $12.5 million construction loan, which involved multiple borrowers and multiple lots with complex construction and leasing issues. Client was extremely pleased with the favorable outcome.
- Lead trial counsel for major banks, private lenders and special servicers in successful loan enforcement and foreclosure actions concerning more than $150 million in commercial property
- Lead counsel for both sellers and buyers in commercial real estate transactions and loan sale agreements totaling more than $75 million
- Successfully represented developer in real estate dispute and related litigation regarding ownership and $125 million redevelopment of historic office building
- Regularly represent gas producers and pipeline owners in oil and gas leases, right of way agreements and land use disputes
- Represented multinational and local oil and gas companies in more than 100 land use and eminent domain actions; successfully obtaining pipeline easements necessary for operation of pipelines
- Obtained reversal of summary judgment from the Missouri Court of Appeals for lender in complex litigation involving failed bank, FDIC's sale of participated loans and rights of lead lender. *FH Partners, LLC v. Complete Home Concepts, Inc.*, 378 S.W.3d 387 (Mo.Ct. App. 2012)
- Successfully represented major banks against lender liability and bank fraud claims
- Represented national data centers companies in lobbying efforts to change state tax laws for the benefit of data center owners and tenants
- State and federal commercial litigation experience in a variety of industries, including real estate, energy and construction

Kansas City
2345 Grand Blvd.
Suite 2200
Kansas City, MO 64108

T: 816.460.5342
F: 816.292.2001

Assistant
Allyson Stewart
816.460.5553

PRACTICES

Banking and Creditors' Rights Litigation
Asset Sales and Acquisitions
Debtor Representation
Landlord and Tenant Disputes
Receiverships
Workout and Restructuring
Business Litigation
Real Estate
Creditors' Rights
Real Estate Litigation

INDUSTRIES

Data Centers, Telecom, and

Léon Korsten
Partner
leon.korsten@dlapiper.com | vCard

Amsterdam
T: +31 (0)20 5419 873
F: +31 (0)20 5419 956

Connect with me on LinkedIn

OVERVIEW EXPERIENCE CREDENTIALS INSIGHTS

Email | Print | PDF | + Shar

Léon Korsten advises on and litigates in cases involving market regulation (enforcement and supervision), EU competition, state aid, public procurement and civil-law enforcement. He also advises on investigations and raids conducted by the regulatory authorities.

Léon has handled many European and Dutch competition law cases, including the Dutch civil proceedings which led to the famous 1999 Eco Swiss/Benetton judgment of the European Court of Justice. Léon handles many licence applications and enforcement procedures for banks, insurance companies and pension funds.

Léon heads the Dutch Litigation & Regulatory practice and the International Competition and Anti-trust group.

View on one page

"A real litigator"
Chambers Europe 2014

RELATED SERVICES

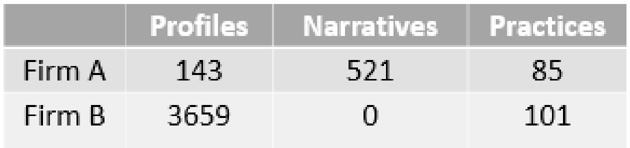
International Trade, Regulatory and Government Affairs
Antitrust and Competition
Litigation, Arbitration and Investigations

LANGUAGES SPOKEN

Dutch, English, German

Figure 2: Attorney Profiles for Firm A (Left) and Firm B (Right)

The practice lists for both firms were consistent in structure, being roughly the same and following the standard two-level practice hierarchy organization. However, the practices themselves varied greatly, with only a few common practices shared across the firms.



	Profiles	Narratives	Practices
Firm A	143	521	85
Firm B	3659	0	101

Figure 3: Examples of the Retrieved Datasets

3 Features and Methodology

Our classifier has three different sources of features, two different models, Multinomial Naive-Bayes and a Multiclass one-vs-rest SVM, and four main classification schemes. The goal in each classification scheme is to classify the data input as one of a set of legal practice areas.

3.1 Features

We use three different sources for features: the information for each matter narrative, the personal web page of each attorney, and short descriptions of each practice area. Each matter narrative consists of a short summary, a list of attorneys who worked on the matter and is labeled with 1-5 practice areas. A lawyer is represented by a biography and is labeled with 1-5 practice areas which may or may not be the same practice areas as the matters the lawyers is connected to. Each practice area is a short plain-text description.

To extract features from each source we parse and stem each word, removing all stop words. Names of attorneys and practice areas are left un-stemmed. The frequency of each word is used as the feature vector for the SVM, while the Naive Bayes uses a TF-IDF setup.

3.2 Models

We evaluated two different learning models, with the results of both presented in the results section. Our first model is a multinomial TF-IDF Naive-Bayes model. Our other model is a one-vs-rest SVM, consisting of one binary classifier for each practice area. For both models we used a standard cross-validation scheme, selecting a random 20% of the input to be reserved for testing for each run of the classifier.

Both models suffer from a small training corpus compared to the number of classes present and the resulting sparse matrix. We made two attempts to reduce the number of classes, but neither showed significant improvement in the classification success rate. The first was to cluster similar classes. To do so, we constructed a list of important words for each class by calculating $\sum_k log\frac{p(word_i|class_j)}{p(word_i|class_k)}$ for each class i, and selecting the top 10% of words and then calculated the cosine similarity between the resulting vectors for each class. Our next attempt was to use an explicit, pre-defined hierarchy of classes, where the classes are grouped by the firm. While both of these attempts dramatically reduced the sparsity of the matrix, they did not improve the classification rate.

3.3 Classification Schemes

We have four different measures for classification. Two are very basic: training and testing over only the lawyer data, and training and testing over only the matter narratives. In order to add more features, we also classify based on the joint feature sets. For example, if we are classifying narratives, in addition the the features of the narrative, we use the features of the attorneys associated with that particular narrative. When classifying attorneys, we additionally use the features of the narratives that attorney is associated with. When doing any joint classification we ensure that we split the cross-validation sets such that no attorney appears in both sets, even when classifying matter narratives. This ensures that our results can be generalized and are not overfitting the particular attorneys, as we found that doing otherwise results in the attorneys' names being the most important features identified by the models.

Finally, all four of the classification schemes use our third feature source, the text descriptions of each class. Because attorney biographies and matter narrative summaries use much of the same language as is found in the descriptions of the classes, including the relevant descriptions decreases the error rate.

4 Results

Our results were gathered by running a series of cross-validation trials, measuring accuracy by counting the number of correct guesses for each randomly-chosen testing dataset. Our first set of experimental results were used to determine the type of supervised learning algorithm to leverage. Accordingly, we confined our analysis to the first dataset and measured our Naive Bayes model against the SVM model.

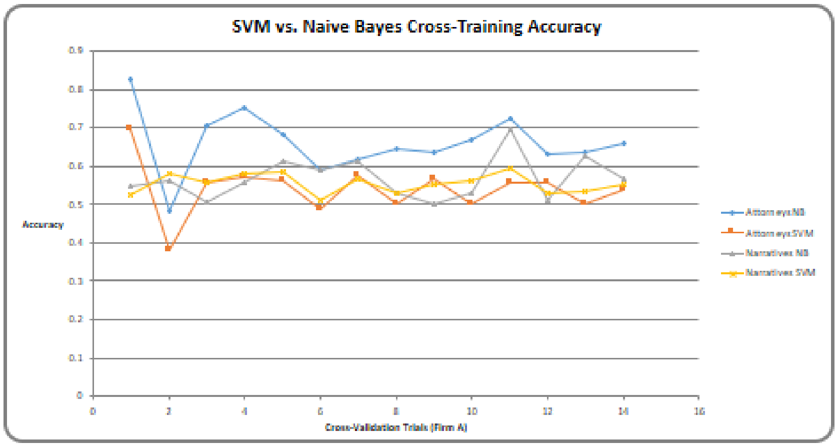


Figure 4: SVM vs. Naive Bayes Cross-Training Accuracy

The above results demonstrated an unexpectedly favorable outcome for our Naive Bayes classifier. Proceeding with Naive Bayes, we then compared the attorney classification success rate across the two datasets.

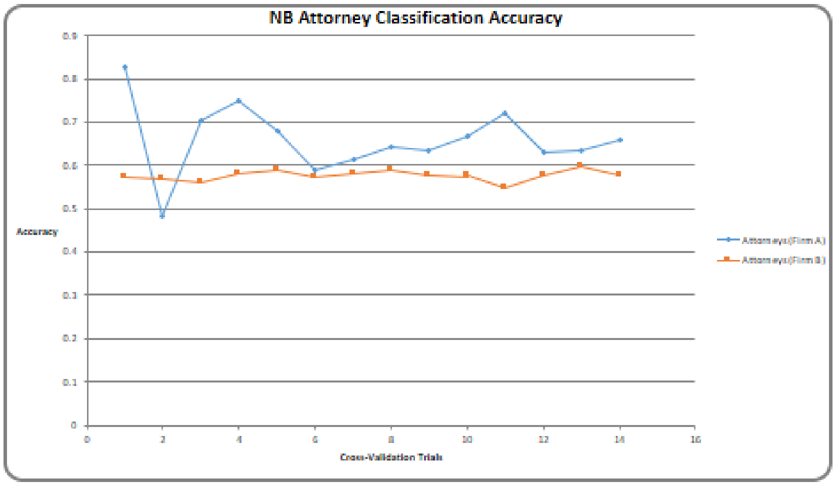


Figure 5: Naive Bayes Attorney Classification Accuracy

A quantitative measurement of our classification groupings is consistent with the above results. The results appear to be reasonable. For example, lawyers with a biography including patents and trademarks are classified into the Intellectual Property practice.

4.1 Analysis and Future Work

With a error rate of around 35%, despite over 80 different classes, our classifier could not replace detailed human labeling, but can performs well enough to provide valuable classification. Furthermore, We believe that it would improve a lot with a larger training size, as we currently only average fewer than 10 training examples per class.

The most promising aspect of this classifier is that it performs well classifying both the attorneys and the narratives. This suggests that it could be used as a powerful recommendation tool to be used by either firms or potential clients looking for an appropriate firm. A promising application would be to not only classify the attorneys and narratives, but to rank the likelihood of each class, as attorneys and matter narratives can have more than one class, and to calculate the similarities between attorneys and matters using the resulting vectors to suggest appropriate attorneys for a matter.

Although, we expected the SVM to outperform the Naive Bayes classifier, we have some initial thoughts on why this isn't the case. Primarily, we believe it is due to the number of classes compared to the number of training samples as the SVM is attempting to do a binary classification of one-vs-all despite many classes being very similar. Furthermore, even if the SVM achieved comparable rates to the Naive Bayes classifier, it would still be less useful as it takes an increasing amount of time for every class for it needs to construct a classifier. During our tests it averaged over 80 times as long to run as the Naive Bayes Classifier.