# Predicting and Analyzing Factors in Patent Litigation

**W. M. Campbell, L. Li, C. K. Dagli, K. Greenfield, E. Wolf, J. P. Campbell**
Human Language Technology Group
MIT Lincoln Laboratory
Lexington, MA 01740
wcampbell@ll.mit.edu

## Abstract

Patent litigation is an expensive and time-consuming process. To minimize its impact on the participants in the patent lifecycle, automatic determination of litigation potential is a compelling machine learning application. In this paper, we consider preliminary methods for the prediction of a patent being involved in litigation using metadata, content, and graph features. Metadata features are top-level easily-extractable features, i.e., assignee, number of claims, etc. The content feature performs lexical analysis of the claims associated to a patent. Graph features use relational learning to summarize patent references. We apply our methods on US patents using a labeled data set. Prior work has focused on metadata-only features, but we show that both graph and content features have significant predictive capability. Additionally, fusing all features results in improved performance. We also perform a preliminary examination of some of the qualitative factors that may have significant importance in patent litigation.

## 1 Introduction

Patent litigation is a potentially lengthy and expensive situation. Throughout a patent's lifecycle, there are many opportunities to circumvent issues that may arise in the process. Initially, when companies or individuals prepare patents, due diligence on prior art and claims is critical for successful filing. Later, when the US Patent Office (USPTO) receives the patents, indicators of litigation related issues could provide good leverage for examiners to analyze and triage patents for potential problems.

In this paper, we present our preliminary study on methods for predicting if a patent will likely be involved in litigation. We focus on three straightforward approaches. First, metadata from patent filings such as assignee, number of claims, category, etc., can be used directly to predict a litigation posterior probability. Second, the content of the claims in a patent can be used to predict litigation using text analytics. Third, we construct a citation graph and use relational learning [1] to predict the litigation potential. In addition, we attempt to understand some explanatory factors that are likely to be involved in the litigation prediction.

Multiple prior efforts related to patent litigation prediction have focused on examining the metadata to find indicators for patent litigation. In [2], the USPTO does an analysis of patent- and examination-related characteristics to study the likelihood of subsequent patent litigation and the filing of a petition for inter partes review (IPR). The analysis focuses primarily on metadata features. In [3], Chien et al. have examined both the intrinsic and acquired traits (features). Some example features are number of claims, number of foreign counterpart patents, recorded assignments, recorded transfers, etc. Logistic regression is used to predict the litigation outcome. Our current work differs from these methods

*This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

in that we add content and graph structure to our prediction process. The content feature may help diagnose problematic language in patents. And graph-based methods leverage related (potentially problematic) patents in order to quantify litigation risk.

The outline of this paper is as follows. In Section 2, we outline the problem of patent litigation prediction and describe our approach. In Section 3, we discuss metadata features for each patent. In Section 4, we present methods for using content for litigation prediction. Section 5 discusses approaches for graph construction from patents and details our approach. Finally, Section 6 demonstrates and qualitatively analyzes our methods for litigation prediction.

## 2 Problem Description

In this paper, we seek to illuminate the root of why certain patents are litigated based on characteristics of the patent, textual content and related patents. In particular, we want to focus on two main tasks. The first task is to investigate whether we can predict which patents will result in litigation. The second task is to identify factors that are more likely to be associated with litigated patents. A system that implements these two tasks will enable an analyst to identify which patents are potentially litigious and to understand why those patents have been classified as such.

To achieve the goals above, we propose a system which fuses the decisions (posterior probabilities) of multiple classifiers of the form seen in Figure 1 below.

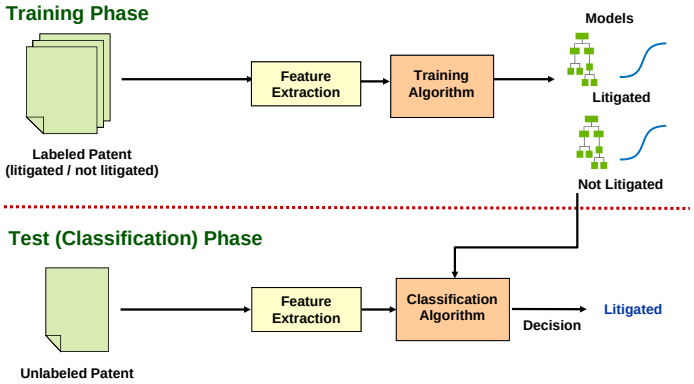



Figure 1: The patent litigation classification system diagram

We adopt a standard train/test classification pipeline using labeled training data that include both the litigated and non-litigated patents. For each of the classifiers, we extract features specific to patent metadata, textual content or related patents. The corresponding feature extraction is detailed in Sections 3, 4, 5, respectively. Once features are extracted, a classification model is chosen and trained using the labeled training data. For categorical and relational features, a random forest classifier is chosen. For the content feature, a logistic regression classifier is chosen. Each trained classifier emits a posterior probability which we use to train a logistic regression fuser. We now describe in detail the various feature extractors.

## 3 Metadata Features

To perform basic litigation prediction, we first look at various metadata features in patents. As shown in previous work [3], this mix of categorical and numerical data can often provide basic indicators of potentially litigious patents. For example, one may expect certain industries to be more litigious than others and that larger corporations may be more susceptible to litigation activities than smaller companies. Additionally, it may be reasonable to expect the more claims and references a patent cites, the more intellectual property is at risk of potential violation. The specific metadata features used in this work are given in Table 1.

## 4 Content Features

For content-based litigation prediction, we use the standard method for text classification based on the word usage [4]. For each patent, we extract the claims, perform text normalization including removing numbers and lower-casing, and compute word counts $\text{count}(w_i|P)$. Here, $w_i$ is the $i$th

Table 1: Meta-Data Features used for Patent Litigation Prediction

| Meta-Data Feature | Example |
|---|---|
| Assignee Information | `Microsoft Corporation` |
| Application/Grant Date | `05-29-07/05-03-11 (1434 days)` |
| Invention Title | `Focal Length Estimation for Panoramic Stitching` |
| Number of Applicants, Claims, Figures & References | `2, 19, 3, 33` |
| Patent Type | `Utility Patent` |
| Publication Country | `United States of America` |

word in the dictionary of all possible words, and $P$ represents the patent. As in the standard text classification, word counts are converted to a vector $\mathbf{v}$ with weighted probabilities

$$v_i = c_i p(w_i|P) \tag{1}$$

where $p(w_i|P) = \frac{\text{count}(w_i|P)}{\sum_j \text{count}(w_j|P)}$. We use a log weighting of $c_i = \log\left(\frac{1}{p(w_i|\text{all})}\right) + 1$, where $p(w_i|\text{all})$ is the probability of the word $w_i$ across all claims in the patent set.

In addition, we train a two class model using logistic regression where the labels are litigated and not-litigated. A Platt scaling is applied to the output to obtain a posterior probability $p(L|P)$.

## 5 Graph Features

For graph-based patent litigation, two steps are performed. First, a citation graph is constructed from the patent documents. Second, a relational learning method is applied to predict the litigation posterior using the 1-hop neighborhood.

Formally, let $G = (V, E)$ be a graph with node set $V$ and edge set $E$. For graph construction, $V$ is the set of all patents in the corpus. We form a directed graph by having an edge from patent $i$ to patent $j$ if patent $i$ cites patent $j$ in the references section. Note that in this work, we ignore other references such as papers or applications in the graph construction. We denote the set of patents cited by patent $i$ (neighbors of node $i$) as $\mathcal{N}(i)$.

For relational learning, we use the aggregation method for combining neighborhood features [5]. That is, given a node feature $x_{i,\ell}$ where $i$ is the node and $\ell$ is the feature number, the aggregated feature for node $k$ is:

$$\hat{x}_{k,\ell} = \text{agg}\left[(x_{i,\ell}|i \in \mathcal{N}(k))\right]. \tag{2}$$

The aggregated feature $\hat{x}_{k,\ell}$ represents a summarization of the values in the neighborhood of a node $k$. A typical aggregator for categorical features is the mode, and a typical aggregator for a numerical feature might be the mean. Intuitively, an aggregator solves the problem of producing a single feature from a variable sized neighborhood.

For node features $x_{i,\ell}$, we use two broad categories of information. First, we retrieve metadata as in Section 3 for each of the nodes. For instance, aggregation on the number of claims might indicate that the patent and its references has a large or small number of related claims. Second, we compute the content-based posterior of litigation on all patent nodes in the graph using the method in Section 4. In this case, litigation of patents in a neighborhood may be a homophilous occurrence.

The aggregated features were combined into a feature vector at each node in the graph. This strategy turns relational classification into the standard method for patent litigation prediction—the aggregated features are just appended to the standard metadata and content feature vectors.

The choice of using patent citations as edges in the graph is a preliminary exploration. Many other possibilities exist. For instance, a richer set of nodes could include other types such as companies, applications, and invention categories. Edges could indicate more variety of relations between the resulting nodes. Additionally, similar content claims could be used to construct a $k$-nearest neighbor graph. In this case, attributes of similar content may provide additional insight into litigation prediction.

# 6 Experiments

## 6.1 Experimental Setup

For experiments, we utilized data scraped from the USPTO and Patexia. The USPTO data consisted of patents in XML format spanning a 12 year period from 2005-2016. The Patexia data consisted of approximately 64,000 lawsuits with filing dates from Jan. 2005 to Sept. 2016. A key point from the Patexia data is that it lists patents *related* to a lawsuit. Information on the exact patent (or patents) and exact claims disputed was not provided.

We examined the Patexia data and grouped by patent classifications. A sample of the top number of potential true trials for litigation by class is shown in Table 2. To ensure adequate training and testing data per patent class, we took patent classes with only a minimum number of patents ( 50). This resulted in 72 patent classes (patents involved in litigation). We split 70/30 for train and test, respectively. The total number of training instances was then 7042, and the total number of true testing instances was 3068.

Table 2: Number of instances of top ten patent classes for litigated patents

| Classification | # of instances |
|---|---|
| Drug, bio-affecting and body treating compositions | 694 |
| Data processing: financial, business practice, management, or cost/price determination | 690 |
| Telecommunications | 512 |
| Electrical computers and digital processing systems: multicomputer data transferring | 499 |
| Drug, bio-affecting and body treating compositions | 458 |
| Multiplex communications | 351 |
| Communications: electrical | 329 |
| Data processing: database and file management or data structures | 264 |
| Chemistry: molecular biology and microbiology | 239 |
| Illumination | 226 |

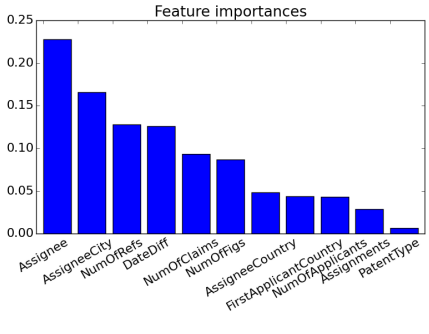
For constructing a false trial set, we used the fact that most patents are not litigated. We randomly sampled from our pool of all patents (2005-2016) in the 72 patent classes with the same prior as in the litigated patents case. Our resulting non-litigated patent set for training consisted of 18,854 examples, and the total number of testing instances was 7993 (a 70/30 split). We included more non-litigated examples to increase significance of testing results and provide adequate training examples.

## 6.2 Metadata Features

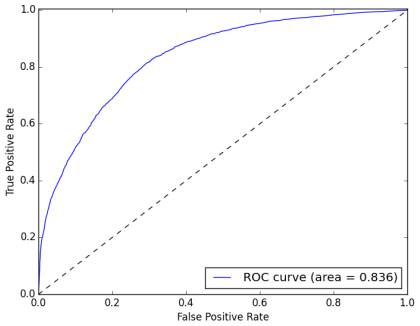
We analyze the contribution of each metadata feature in predicting patent litigation. We first extract all the metadata-based features, as described in Section 3. Then we combine these features by training a random forest model, which is an ensemble learning method that constructs a number of decision trees and averages the probabilities of an outcome from these decision tree. The importance of features is measured in terms of the mean decrease impurity. As described in [6], it is the total decrease in node impurity averaged over all trees. From Figure 2(a), we can see that the two most important features are Assignee and AssigneeCity. And the least significant feature is the PatentType. All other features also show fairly significant importance. Using the trained the mode, we obtain a posterior probability $p(L|P)$ for each patent in the testing set. Figure 2(b) shows the performance.

## 6.3 Content Feature

We trained a logistic regression model of litigation and no litigation using the training set and the methods described in Section 4. Results for content-based litigation classification can be seen in Figure 3. The results in Figure 3(b) show that significant information is contained in the claims data for determination of litigation. Initial efforts to qualitatively understand the classification process can be obtained by looking at the words with large positive (litigation) or large negative (no litigation) weights. Figure 3(a) shows these words. We see that there are some classification indicators (organic, metal seminconductor), but there are also some words that are more functional. Although this gives initial clues, more detailed analysis is needed to understand the role of these words in the classification process.




(a) Feature importance

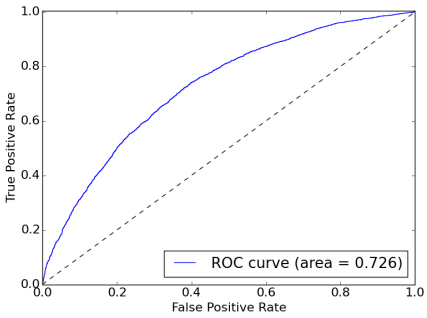



(b) ROC curve

Figure 2: Performance of metadata features for patent litigation prediction

| **Words Assoc. with Litigation:** |
|---|
| helical, cellular, occupant, connections, custom, metal-semiconductor, payment, price, claim |
| **Words Assoc. with No Litigation:** |
| organic, install, reflected, non-transitory, structural, higher, operably, writing, il-2, targeted |

(a) Important words associated with litigation and no litigation




(b) ROC curve

Figure 3: Performance of the content feature for patent litigation prediction

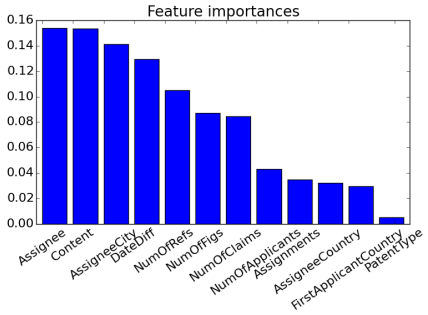
### 6.4 Graph Features

We use the pipeline of four stages to predict patent litigation using the graph features. We first construct a patent citation graph. For each node in the graph, we extract a list of metadata features and compute the content feature based on patent claims. For relational learning, we perform a simple feature aggregation using the 1-hop neighborhood of each node. The aggregated features are then combined using the random forest model. Figure 4 shows the contribution of each aggregated feature and the prediction performance. Observe from Figure 4(a) that the aggregated Assignee and AssigneeCity are still among the most significant features for litigation prediction. In addition, the aggregated Content feature is the second most significant feature while the aggregated PatentType is the least significant feature. Using the trained random forest model, we then obtain a posterior probability $p(L|P)$ for each available patent in the testing set. Figure 4(b) shows the performance based on the fusion of graph features; an AUC score of $0.835$ is achieved. It is to be noted that about half the patents are isolated nodes in the graph. Thus, they do not have graph features and they are not included in the evaluation shown in Figure 4.

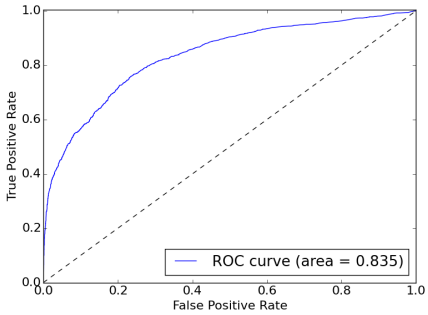
### 6.5 Fusion

The goal for fusion is to combine posterior probabilities estimated from different features to obtain better prediction performance and to analyze factors in patent litigation. The posteriors to be fused are from metadata-, content- and graph-based features. These posteriors are estimated using the methods described in the previous sections. For missing features, we impute an uninformative posterior of $0.5$.

Once all the missing values are imputed, we train a logistic regression model to fuse the metadata-, content- and graph-based posteriors from the patents in the training set. Figure 5(a) shows the regression coefficients for the three features. Observe that the metadata-based feature has the highest weight, followed by the graph-based feature. The content-based feature has the lowest coefficient weight. This shows that there is still significant room for improvement in terms of extracting better graph-based and content-based features. We then apply the trained logistic regression model to the testing set, the performance of fusion is shown in Figure 5(b) with an AUC score of $0.841$.

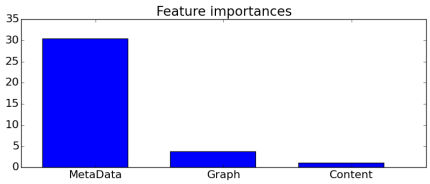



(a) Feature importance

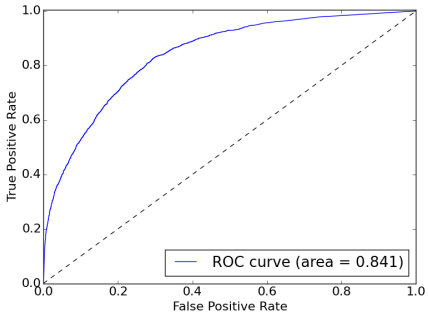



(b) ROC curve

Figure 4: Performance of graph features for patent litigation prediction




(a) Feature importance




(b) ROC curve

Figure 5: Performance of logistic regression fusion for patent litigation prediction

## 7 Conclusions

In this paper, we demonstrated metadata, content, and graph features for predicting and analyzing factors in patent litigation. For metadata features, we extracted a mix of categorical and numerical data and performed fusion over these features. The importance scores of these features were computed and ranked. For the content feature, we used the standard method for text classification based on word usage. Method for extracting relational graph features was also presented. Factor analysis over these relational graph features was performed and features were ranked based on the importance scores. Most of the features used in the metadata and graph analysis showed fairly significant importance. Fusion of the metadata-, content-, and graph-based features showed that the metadata feature has the highest weight in litigation prediction, while the content- and graph-based features showed relatively small impact compared to the metadata feature.

Future work includes analyzing and extracting more informative content and graph features. For content, creating domain specific analysis would add additional insight into litigation prediction. Another future work is to construct interpretations useful to subject matter experts.

## References

[1] Lise Getoor. *Introduction to statistical relational learning*. MIT press, 2007.

[2] Alan C. Marco, Richard D. Miller, Kathleen Kahler Fonda, Pinchus M. Laufer, Paul Dzierzynski, and Martin Rater. Patent litigation and USPTO trials: Implications for patent examination quality. `https://www.uspto.gov/sites/default/files/documents/Patent%20litigation%20and%20USPTO%20trials%2020150130.pdf`. Accessed: 2016-10-31.

[3] Colleen V Chien. Predicting patent litigation. *Texas Law Review*, 90:283, 2011.

[4] T. Joachims. *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers, 2002.

[5] Jennifer Neville, David Jensen, Lisa Friedland, and Michael Hay. Learning relational probability trees. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 625–630. ACM, 2003.

[6] Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.