Forecasting of Trends in Legal Spend Management

Pragati Awasthi, Jerzy Bala, Sebastian Carter
Bottomline Technologies Inc.
325 Corporate Drive
Portsmouth, NH 03801, USA
pragati.awasthi/jbala/Sebastian.Carter/@bottomline.com

***Abstract*—The paper describes a framework for forecasting narrative trends (text-based description of cost items) in legal spending. This is based on the application of topic discovery and time series forecasting. The algorithm presented in this paper discovers a number of abstract topics in a corpus based on clusters of words that are found in each line item spending document, along with the respective frequency of those words. Specifically, Latent Semantic Analysis transforms a sequence of cost descriptions into a set of numerical Topic-based univariate time series. The resulting set of time series is used to forecast future trends using the ARIMA (AutoRegressive Integrated Moving Average) approach. This type of semantic forecasting of spending trends can facilitate the discovery of counterparty intent(s) and proactively adjust the litigation strategy (prove/disapprove a claim, counterclaim, etc.).**

Keywords— litigation case prediction, legal spend trending, topic discovery, semantic forecasting

I. INTRODUCTION

Lawyers' judgments about case outcomes (e.g., settle, go to trial, etc.) are often formed on complex cognitive processing steps that are based on (and limited to) intuition and experience in the law profession. As such, these are difficult to define and frame abstractly (e.g., as decision-making business intelligence human elucidated rules). On the other hand, in the case of Legal Spending, it is easy to demonstrate past cost items in specific case scenarios (as the temporal sequence of line item spending) by using historical data that is observed/collected and associated with such scenarios to automatically generate a topic-based description, and use some number of topics (i.e., topics that are strongly forecastable as validated using only the historical time-series data) as the univariate time series. Such time series represent historical trending data that can be used to forecast future trends.

Forecasting of future legal spending cost narrative trends provides Legal Spend Management (LSM) a mechanism that uses textual data (i.e., invoice narratives in this study) to judge the direction, with some level of confidence, of the case progression. A forecast series is able to constitute a trend prediction system that facilitates course-of-action decision-making. Some practical applications of the presented approach include:

- Forecasting escalation expenditures in future line item costs,
- Case comparison (i.e., comparing the forecasted types of line item costs with the previous cases),
- Allocation of resources for invoice processing,
- Estimating the length of litigation cases,
- Using forecasted line item cost types to discover evidences in order to understand claims, counterclaims, or some specific lines of defense.

II. RELAVANT WORK

The relevant work on the use of predictive analytics falls into representing general scenarios of the use of predictive analytics for claims management, as well as describing specific techniques. The following paper represents the two groups:

Lentz [1] identifies the following areas where predictive modeling is used to enhance the claim management process:

- Allocation of Resources
- Reserving/Settlement Values
- Recognition of Potentially Fraudulent Claims
- Identification of Potentially High Value Losses
- Expense Management
- Trend Analysis

The above paper identifies the concept of the early warning of potential "outliers": claims that appear routine but eventually develop into high value losses.

Brüninghaus and Ashley [2] present a multi-strategy algorithm called IBP (issue based prediction) that combines case-based and model-based reasoning for an interpretive CBR (case-based reasoning) application to predict the outcome of legal cases. It uses an ad-hoc model of the domain to identify the issues raised in the case (called a "weak model"). In the second step, it reasons with cases to resolve conflicting evidence related to each issue. IBP reasons symbolically about the relevance of cases and uses evidential inferences. Experiments with a collection of historic cases show that IBP's predictions are better than those made with its weak model or with cases alone. The authors claim that their approach has higher accuracy compared to standard inductive and instance-based learning algorithms.

Tarek and Kandil [3] describes the use of machine learning in construction litigation cases. They propose an automated litigation outcome prediction method for differing site condition (DSC) disputes through machine learning (ML) models. To develop the proposed method, this paper compares the performance of three ML techniques, namely: support vector machines, naïve Bayes, and rule induction and neural network classifiers (decision trees, boosted decision trees, and the projective adaptive resonance theory). The models were trained and tested using 400 DSC cases filed in the period from 1912 to 2007. Model predictions are on the basis of significant legal factors that govern verdicts in DSC disputes in the construction industry.

III. FORECASTING APPROACH

A. Approach

Figure 1 depicts the proposed forecasting approach.

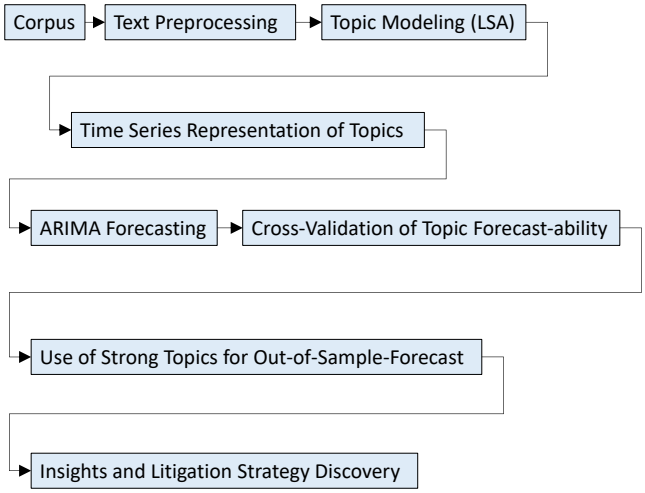


Fig. 1. Forecasting approach

1. Corpus preparation. The Corpus process reads text corpora from files and sends a corpus instance to its output channel.
2. Text Preprocessing. This process splits textual narratives into smaller units (tokens), filters them, and then runs them through normalization (Porter stemming).
3. Topic Modeling. This process discovers abstract topics in a corpus based on clusters of words found in each document along with their respective frequency. A document typically contains multiple topics in different proportions. The process also reports on the topic weight per document (stored in a data table).
4. Time series representation of topics. This process reinterprets any data table as a time series. The time series sequence is implied by instance order.
5. Time series forecasting. The ARIMA modeling approach is used in this study (the following section describes this modeling approach).
6. Evaluation of time series. This process evaluates different time series' topics by comparing the errors they make (e.g., mean absolute percent error) by using a number of folds in the time series cross-validation schema.
7. Use of forecastable topics for out-of-sample discovery of strong trends. This process outputs forward looking forecasts (topically we used 10 forward looking steps).
8. Discovery of trends**.** Forecasted steps are used to analyze upward and downward trends for each topic, together with the strength.

B. Line Item Spending Example

The following is the spent line item example list (i.e., the initial line items in the spending series, where typically the lists contain about 400 cost narratives).

• Review/analyze prior correspondence and discovery to identify key issues to address in pre-arbitration report
• Draft Civil Subpoena for Attendance by Telephone for Arbitration.
• Draft pre-arbitration report outlining defenses and likely outcome at arbitration.
• Begin drafting pre-hearing statement of proof for arbitration hearing.
• Telephone conference with a fact witness regarding facts of the subject accident prior to the upcoming arbitration
• Draft the Declaration to be used at arbitration in this matter.
• Revise the Civil Subpoena for Telephonic Attendance at Arbitration.
• Copying

C. Topic Modeling

Topic Modelling is a discovery process of clustering words from each document into a corpus based on the frequency of those words. A document might typically contain multiple topics with their different distributions. Topic Modeling is used to segment the textual data set into semantically coherent parts. Once topics are discovered in a document corpus, a measure of membership (weight) is generated that represents how closely a given case (i.e., a textual narrative representing line item cost) is matched to a given topic.

LSA[https://en.wikipedia.org/wiki/Latent_semantic_analysis] is one the common techniques in natural language processing, namely in distributional semantics. LSA discovers relationships in a corpus between documents and the words they contain and generates topic models that relate the documents to the words.

This mechanism assumes that words that are semantically related should be found in portions of the text. This is called the distributional hypothesis assumption. LSA maintains a matrix containing word counts per document, i.e., words (rows) by documents (columns). Once such matrix is constructed by a singular value decomposition (SVD) process, which is performed to reduce the number of rows (i.e., words) while maintaining similarities among columns. Columns are then compared using the cosine of the angle between two vectors. The results of this process appear as a numerical value between 0 (strong dissimilarity between documents) and 1 (strong similarity between documents).

The following is an example of the Ten-Topic representation.

1: letter, draft, send, cover, counsel, deposition, service, subpoena, documents, file
2: draft, deposition, cover, plaintiff, file, review, court, defendant, certify, letters
3: draft, subpoena, file, records, cover, opposing, answer, plaintiff, certificate, documents
4: court, defendant, counsel, review, documents, notice, reporter, deposition,
5: file, defendant, adjuster, deposition, letter, reporter, counsel
6: motion, subpoena, service, counsel, support, documents, court, order, case, send
7: review, letter, answer, defendant, records, c, court, file, plaintiff, documents
8: service, review, adjuster, case, motion, deposition, client, depo, certificate,
9: court, opposing, production, send, draft, new, motion, copy, review, service
10: interrogatories, file, answers, copy, records, subpoena, defendant, responses

LSA provides both positive and negative weights per topic. A positive weight means the word is highly representative of a topic, while a negative weight means the word is highly unrepresentative of a topic (the less it occurs in a text, the more likely the topic). Positive words are colored green and negative words are colored red. For example, the first topics might refer to the cost of preparing a deposition document that is set up by the attorney to address a subpoena (i.e., a court-ordered demand for a document).

D. *ARIMA Forecasting*

The specific forecasting approach used in the study is based on the ARIMA (AutoRegressive Integrated Moving Average) forecasting algorithm [https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average]. The algorithm captures temporal structures in a time series to find the best fit for historic data flow. An ARIMA model is denoted as an "ARIMA (p, d, q)", where:

- *p* is the number of autoregressive terms,
- *d* is the number of differences needed for stationarity, and
- *q* is the number of lagged forecast errors in the prediction equation.

For a given time series, X(t), two models are invoked – auto-regression and moving average. In the regression model, X(t) can be explained by some function of its previous value, X(t-1), plus some unexplainable random error, E(t).

X(t) = A(1) * X(t-1) + E(t)

where X(t) = time series under investigation

A(1) = the autoregressive parameter of order 1

X(t-1) = the time series lagged 1 period

E(t) = the error term of the model

Higher orders in the regression model can be used. For example the second order follows the following equation (two autoregressive parameters, A(1) and A(2) are used here).

X(t) = A(1) * X(t-1) + A(2) * X(t-2) + E(t)

The moving average model expresses X(t) at time t as the function random errors that occurred in past time periods. The moving average of order 1 is given by the following function:

X(t) = B(1) * E(t-1) + E(t)

In searching the parameter space of the two ARIMA modeling approaches, the Akaike Information Criteria (AIC) is as a measure statistical model fit [https://en.wikipedia.org/wiki/Akaike_information_criterion (AIC)]. When comparing two models, the one with the lower AIC is chosen.

Fitting an ARIMA model requires the series to be stationary, i.e., mean and variance are time invariant. Automated ARIMA generate a set of optimal (p, d and q) parameters using the auto.arima function and picks the set that optimizes the model fit criteria. The Akaike Information Criterion is used as a measure of a statistical forecasting model fit.

IV. Forecasting Example

A. *Validation of ARIMA Forecasting Models*

Cross-validation is a popular technique for tuning hyperparameters and for producing robust measurements of model performance. Two of the most common types of cross-validation are *k*-fold cross-validation and hold-out cross-validation. In time series cross validation, the training set consists only of observations that occurred prior to the observations that formed the test set. Thus, no future observations can be used in constructing the forecast. When dealing with time series data, traditional cross-validation (like k-fold) should not be used because of temporal dependencies.

Error estimation using n-fold validation for time-series.

Mean Absolute Percentage Error:

$$\mathrm{M} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

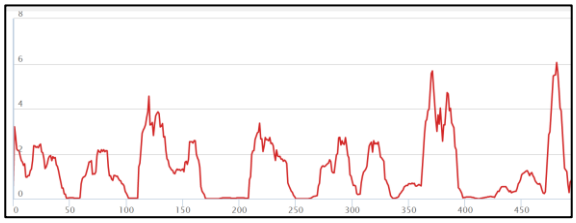
where A_t is the actual value and F_t is the forecast value. The difference between A_t and F_t is divided by the actual value A_t again. The absolute value in this calculation is summed

for every forecasted point in time and divided by the number of fitted points *n.*

B. Experimental Result Example

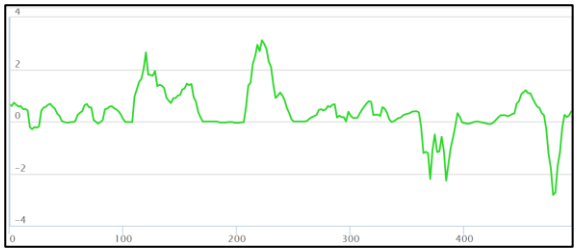
The following is the example of the time series from the Ten-Topic representation, where X=sequence of line items and Y=weight assigned by LSA to a specific topic (Fig 2). A positive weight means the word is highly representative of a topic, while a negative weight means the word is highly unrepresentative of a topic.

Topic 1



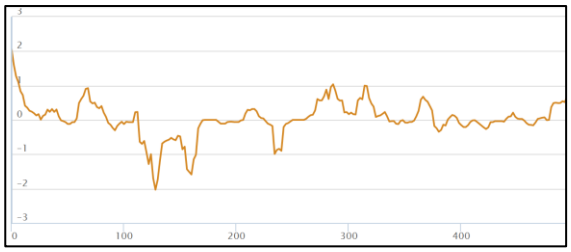
M=14% error (5 cross-validation folds)

Topic 2



M=16% error (5 cross-validation folds)

Topic 3



M=22% error (5 cross-validation folds)

Fig. 2. Example of three topic time series

The forecasted trends can be depicted in the qualitative trend insight table. The strength of the trend is represented by the mean absolute error that is computed by using cross-validation of historical time series. In most of the experiments in this study three folds and ten forecast steps were used to find strong forecastable topics. For the first three topics the average Mean Absolute Error is 17%.

Fig 3 depicts forward-looking forecast (5 steps) calculated using the ARIMA model. Topic 1 has the strongest trend. It refers to the cost of preparing a deposition document that is set up by the attorney to address a subpoena.

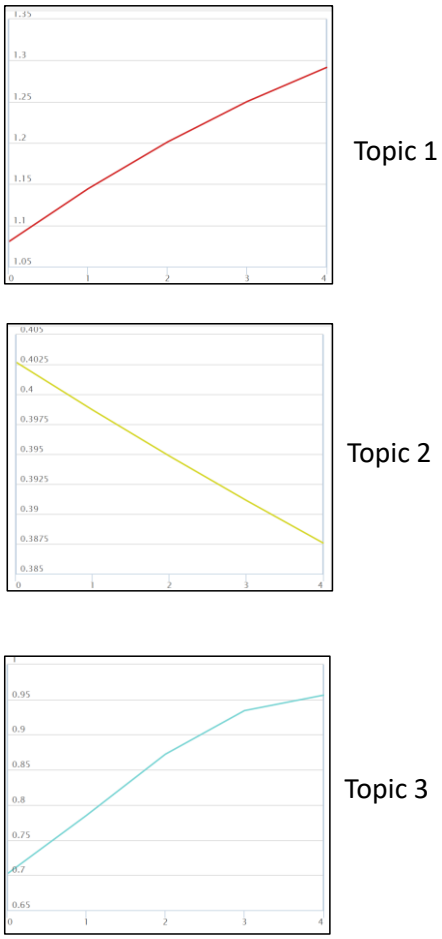


Fig. 3. Example of forward-looking forcast (X=forward steps, Y= topic weights)

	Trend Slope	Strength (M)
Topic 1	Strong Up	14
Topic 2	Low Down	16
Topic 3	Medium Up	22

Table 1. Example of forward looking trends

V. Conclusions

With an ever-increasing focus on data, analytics, and controlling costs, carriers and claims organizations are seeking innovative ways to improve business results. Many organizations are still stuck in the past, using outdated reporting tools to simply examine what has already happened. More advanced organizations are utilizing sophisticated modeling and Predictive Analytics to focus not only on what happened in the past, but what is likely to happen in the future. With analytics solutions such as the one described in the paper, these organizations can reorganize and transform their Litigation Management processes. Specifically, the framework presented in the paper for forecasting narrative trends (text based description of cost items) in legal spending based on the application of topic discovery and time series forecasting can facilitate the discovery of counterparty intent(s) and proactively

adjust the litigation strategy (prove/disapprove a claim, counterclaim, etc.).

The initial results described in the paper represent a work in progress. Future work will include the application of other topic modeling techniques (e.g., Latent Dirichlet Allocation), extensive evaluation of sets represented by a large number of topics, and more insightful representation of the system output (e.g., an ontology based semantic model explaining forecasted trends and their relationships).

REFERENCES

[1]. Willaim Lentz, Predictive Modeling - An Overview of Analytics in Claims Management Issue: November 2013 | P/C General Industry, Auto/Motor, General Liability, Marine, Property, Workers' Compensation.

[2]. Stefanie Brüninghaus and Kevin D. Ashley, "Combining Case-Based and Model-Based Reasoning for Predicting the Outcome of Legal Cases" International Conference on Case-Based Reasoning, ICCBR 2003.

[3]. Tarek Mahfouz and Amr Kandil, "Litigation Outcome Prediction of Differing Site Condition Disputes through Machine Learning Models", Journal of Computing in Civil Engineering. Volume 26 Issue 3 - May 2012.