CAIT-UTC-031

Analyzing Asset Management Data Using Data and Text Mining

FINAL REPORT
July 2014

Submitted by:

Trefor Williams
Professor
Department of Civil & Environmental Engineering
Rutgers, The State University of New Jersey
Piscataway NJ 08854

Marv Halling
Associate Professor
Department of Civil & Environmental Engineering
Utah State University
Logan UT 84332

External Project Manager
Frank Otero,
Paco Technologies

In cooperation with

Rutgers, The State University of New Jersey
And
State of New Jersey
Department of Transportation
And
U.S. Department of Transportation
Federal Highway Administration

Disclaimer Statement

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No. CAIT-UTC-031	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle

Analyzing Asset Management Data Using Data and Text Mining

5. Report Date
July 2014

6. Performing Organization Code
CAIT/Rutgers, The State University of New Jersey

7. Author(s)
Trefor Williams, Marv Halling

8. Performing Organization Report No.
CAIT-UTC-031

9. Performing Organization Name and Address
Department of Civil & Environmental Engineering
Rutgers, The State University of New Jersey
Piscataway NJ 08854

10. Work Unit No.

11. Contract or Grant No.
DTRT12-G-UTC16

12. Sponsoring Agency Name and Address

Center for Advanced Infrastructure and Transportation
Rutgers, The State University of New Jersey
100 Brett Road
Piscataway, NJ 08854

13. Type of Report and Period Covered
Final Report
8/1/12-9/30/13

14. Sponsoring Agency Code

15. Supplementary Notes
U.S. Department of Transportation/Research and Innovative Technology Administration
1200 New Jersey Avenue, SE
Washington, DC 20590-0001

16. Abstract

Predictive models using text from a sample competitively bid California highway projects have been used to predict a construction projects likely level of cost overrun. A text description of the project and the text of the five largest project line items were used as input. The text data were converted to numerical attributes using text-mining algorithms and singular value decomposition. Two models were produced. The first used only the text description as input, while the second combined the text data with the numeric value of the low bid. Classification models were produced using the K-Star classification algorithm. Modeling results indicated information in the textual descriptions is related to the projects level of cost overrun.

17. Key Words
Pavement; Computer-Visions; GIS; Pavement Monitoring; Bidimensional Empirical Mode Decomposition

18. Distribution Statement

19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 10	22. Price

Contents

Abstract: Predictive models using text from a sample of competitively bid California highway projects have been used to predict a construction projects likely level of cost overrun. A text description of the project and the text of the five largest project line items were used as input. The text data were converted to numerical attributes using text-mining algorithms and singular value decomposition. Two models were produced. The first used only the text description as input, while the second combined the text data with the numeric value of the low bid. Classification models were produced using the K-Star classification algorithm. Modeling results indicated information in the textual descriptions is related to the projects level of cost overrun.

Keywords: text mining, construction, cost, prediction.

1. INTRODUCTION

1.1 Text Mining and Prediction of Competitively Bid Project Costs

Cost increases often occur in competitively bid construction projects. Potentially, useful indicators are available at the time of the bid opening that may provide an indication of a projects likelihood to experience large cost increases. Knowledge of the projects characteristics can allow owner organizations to better plan for contingency funding and indicate that increased scrutiny of a project is warranted to avoid large cost increases. At the time of the bid opening numeric data available include the low bid, the engineer's estimate and the number of bidders. Text data is also available that describes what is to be constructed, the magnitude of the project, and the various unit price pay items that describe the materials to be installed. In the past, it was not possible to use the text data as part of a predictive model but with the emergence of text mining algorithms and software it is now possible to incorporate the information from textual descriptions of a project in predictive models.

Text mining can be defined as the automatic discovery of previously unknown information from unstructured text data. Text mining involves extracting information of interest from text documents and then the use of data mining to discover new associations among the extracted information (Karamanis 2007). The purpose of this paper is to show how text and data mining techniques can be used to produce predictions of a projects anticipated level of cost overrun using text and numeric data. This paper will demonstrate how text about a construction project can be converted to a format that can be used by the KS-tar data mining classification algorithm to predict the level of cost increase expected during the project. The paper will also explore how textual and numeric data can be combined to increase the accuracy of the prediction.

1.2 Literature Related to Text Mining and Cost Prediction in Construction

There have been several applications of the use of classification and data mining to construction management problems. A prototype system that automatically classifies construction documents according to project components using data mining techniques was proposed by Caldas et al. (2002). Soibelman and Kim (2002) addressed the need for data mining in the construction industry, and the possibility to identify predictable patterns in construction data that were previously thought to be chaotic. In that study, a prototype knowledge discovery and data mining (KDD) system was developed to find the cause of activity delays from a U.S. Army Corps of Engineer’s database called the Resident Management System. Soibelman et al. (2008) have addressed the need to develop additional frameworks that allows the development of data warehouses from complex construction unstructured data and to develop data modeling techniques to analyze common construction data types. Existing construction text mining research has focused on methods of classifying documents and extracting information from databases. Recent research in cost prediction includes Son et al. (2012) who have developed a model using Principal Component Analysis and Support Vector Regression using 64 variables to predict cost performance on building projects.

2. METHOD

2.1 Data for Analysis

Data for this analysis was collected from websites of the California Department of Transportation. Data from 1221 competitively bid highway projects were collected. The bid opening data and the completed cost after change orders were collected. The project data collected varied widely in cost magnitude and type of construction. Projects varied from small maintenance projects to major rehabilitations or large new construction projects. Through experimentation, it was found that trimming large outliers from the data set produced better predictions, therefore 47 projects that were completed at a cost more then 25% lower then the original bid price, were excluded from the analysis. The maximum cost overrun in the data set was 74% greater then the original low bid amount.

Table 1. Example Project Text

Project	Text Description
1	CONSTRUCT TRUCK CLIMBING LANE ASPHALT CONCRETE (TYPE A) CLASS 2 AGGREGATE BASE TEMPORARY RAILING (TYPE K) MOBILIZATION
2	SEAL COAT SLURRY SEAL TRAFFIC CONTROL SYSTEM ASPHALTIC EMULSION (POLYMER MODIFIED) 100 MM THERMOPLASTIC TRAFFIC STRIPE
3	GRADE CHANNEL AND REPLACE CULVERTS 730 MM X 1150 MM OVAL SHAPED REINFORCED CONCRETE PIPE (CLASS IV) 450 MM REINFORCED CONCRETE PIPE (CLASS V) RECONSTRUCT FENCE MINOR CONCRETE (BACKFILL)
4	REALIGN ROADWAY STRUCTURAL CONCRETE, BRIDGE (SEGMENTAL BOX GIRDER) STRUCTURAL CONCRETE, BRIDGE MOBILIZATION TIME-RELATED OVERHEAD

The text data collected were taken from the bid summary for each project. The bid summary information contains a short one-sentence description of the project work to be performed. Table 1 shows a sample of the project descriptions. To provide additional information, the textual descriptions of the five largest project line items by dollar amount in the bid summary form were appended to the project description for each project. This was done to provide more information that could allow various project types to be differentiated by the data mining algorithms.

2.2 Modeling Process

The models were constructed using Rapid Miner. The Rapid Miner software is an open-source data and text mining toolbox that is widely used to build data and text mining models. Software for many different types of data mining algorithms is available for experimentation in Rapid Miner. Rapid Miner uses a building blocks approach that allows models to be developed without the need for extensive programming.

Two different models were developed. The modeling steps for the first model are shown in Figure 1. This model uses only text as input. Figure 2 shows the modeling steps for the second model that combines the text data with the low bid value. The model is generally the same as the first model except that the numeric value of the low bid for each project is joined to the output of the text mining data and then submitted to the K-Star classification algorithm for processing. For both models, after the data set is read into the Rapid Miner model it is split with 50% of the data set aside to be used as a testing set and 50% used to train the model.

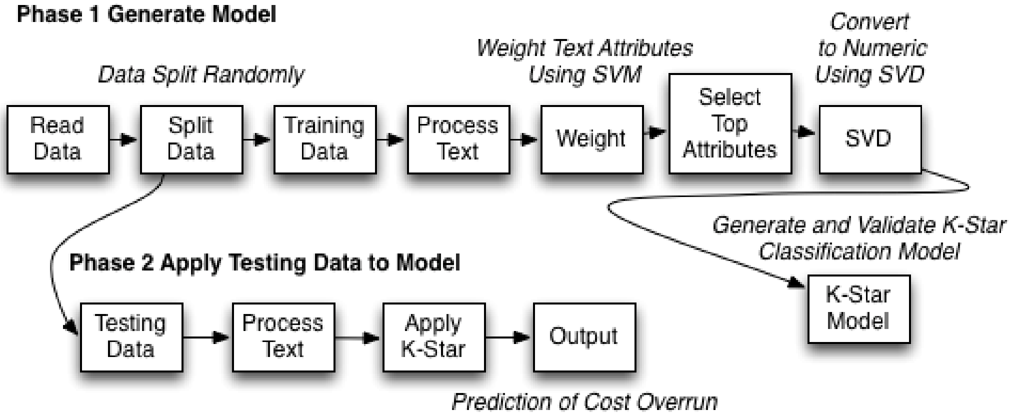
Phase 1 Generate Model
Data Split Randomly
Weight Text Attributes Using SVM
Convert to Numeric Using SVD
Read Data
Split Data
Training Data
Process Text
Weight
Select Top Attributes
SVD
Generate and Validate K-Star Classification Model
Phase 2 Apply Testing Data to Model
K-Star Model
Testing Data
Process Text
Apply K-Star
Output
Prediction of Cost Overrun

Figure 1. Model Flow Chart Using Only Text as Input

In addition to the text collected for each project, the percentage of project cost overrun was calculated. This was the percentage increase above the original bid amount for the completed project. Each project was assigned to one of three cost overrun groups. The output of the model is a prediction of which of the three levels of overrun a project will to have. Projects categorized as having large cost overruns are projects with a cost overrun greater then 6%. Projects categorized as being completed near the original low bid had overruns between + 5% and -3% under runs. Projects categorized as under run projects were all projects with under runs less then -3%.

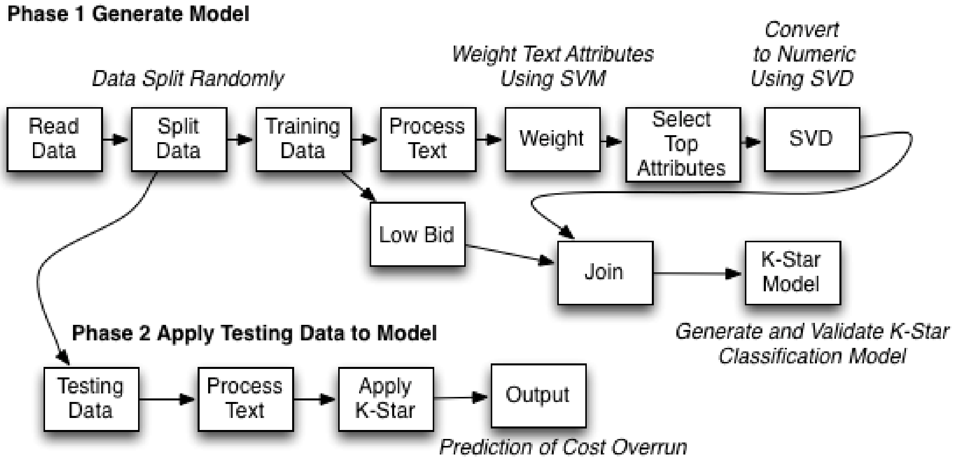

Figure 2. Second Model Flow Chart with Low Bid numeric variable added

2. 3 Text Mining

(1) Processing the Text

To be useable by the data mining algorithms in Rapid Miner the text must be processed and converted to a numeric value. For each project in the database of California projects, the available text was transformed into a vector that provides a numerical representation of the information in the text. Information was extracted from text describing different types of projects conducted and the relationship of the text describing a project to cost overruns. The text for each project must be transformed into a numerical vector that is suitable for use with a data-mining algorithm. There are several steps that are necessary to transform the unstructured text for each project into a standardized numeric vector. These steps are tokenization, stopping, stemming, normalization and vector generation (Miner et al. 2012). The data transformation steps applied to the project description text performed by the

Rapid Miner software were (Weiss et al. 2010):

1. Transform Cases. Uppercase letters were removed from the text.
2. Tokenize. The unstructured text is transformed into a sequence of tokens. Tokens can take on different forms, however in this model the tokens were equivalent to single words. Term frequency was used to scale a tokens value.
3. Stemming. In this data transformation related word tokens are normalized into a single form. For example “walking” would be transformed to “walk” (Miner et al. 2012). Through experimentation, stemming was found to increase the prediction accuracy.
4. Filter Stopwords. Common words like “and” and “but” are removed by removing words on a predefined list.
5. Filter Token Length. This filter removes words that are less then three characters long.
6. Generate n-gram terms. In this process, Rapid Miner has been set to allow two word terms to be entered in the term-project matrix that is generated for the text. In other words, word pairs can be used to produce additional tokens.

Table 2. Words and Word Pairs Frequently Indicating an Overrun Project

Word	High Overrun	Near Low Bid	Under Run
bridg	165	118	21
signal	92	71	18
overhead	43	29	3
concret_bridg	42	30	4
bridg_remov	16	7	1
polyest_concret	16	7	2
excav_asphalt	14	3	1

After the text has been processed it is possible to examine the word tokens and to see the relationship of various tokens to the level of cost overrun. Table 2 shows some of the word tokens that are frequently associated with cost overrun projects. Note that the words have been stemmed to their root.

(2) Weighting Relevance of Tokens Using SVM

The output of the text mining process is a very large, sparse matrix where each row is a project and each column is the word tokens that have been identified. For the California project data, approximately 7000 word tokens were identified. Features used for text mining are often high dimensional in nature. These cause over-fitting when training data is not sufficient. Dimensionality reduction leads to performance enhancement in such situations (Saha et al. 2012). To reduce the dimensionality for this problem a Support Vector Machines (SVM) algorithm was used to weight the importance of the word token attributes. Support vector machines are a group of supervised learning methods that can be applied to classification or regression (Ivanciuc 2007). After the SVM weighting was completed, the 500 highest weighted word token attributes were selected for use in building the K-Star model.

(3) Transformation to Numeric Values Using SVD

After these transformations were completed a matrix of projects and terms is created. Then, Singular Value Decomposition (SVD), a dimensionality reduction method, was used to transform the matrix of projects and terms into two numeric values for each project. SVD provides a convenient way for breaking the large matrix of projects and words output from the text processing models, into simpler, meaningful pieces. The values generated by the SVD algorithm represent the information derived from the text for each project.

2.4 The K-Star Classification Algorithm

After processing the transformed text training data, now in a numeric format, was submitted to the K-Star classification algorithm. The purpose of the K-Star model that is generated in this step is to classify a project as having a “high overrun”, “near” or “under run” level of completed project cost. The K-Star algorithm is an instance based learning scheme developed by Cleary et al (1995). Witten and Frank (2005) describe the K-Star algorithm as a lazy classifier where the “…training instances are stored and do no real work until the classification time.” They also state that K-Star is a nearest neighbour method with a generalized distance function. It has been shown that nearest neighbour methods work well when combined with the pruning of noisy exemplars.

2.5 Bootstrapping

Validation of the K-Star model generated was done using a boot strapping technique. Bootstrapping is a resampling process where a procedure, in this case the generation of the K-Star classification, is performed several times with a different random sampling of the training data to produce more accurate predictions. A scheme was employed where 70% of the training cases are used to train the model and 30% are used to test the model. The model is trained and predictions are produced using the selected 30% of the cases as inputs. This process is repeated 10 times with a different set of training and test cases selected randomly from the training data set for each model run to. The bootstrapping technique output is the averaged predictions for the 10 runs. This procedure serves to validate that the generated model is generalizable and not over fit to the training data.

2.6 Predictions Using the Training Data

After the K-Star model was built, a Rapid Miner model was used to predict the level of cost overrun for the testing set of project data that was not used in the training process. The raw testing data were transformed using the same text transformation techniques to convert the text to numeric data that can be accepted by the trained K-Star model. After transformation the testing data were submitted to the trained K-Star model and prediction results were produced.

Table 3. Predictions Using Text Only

		Class Precision			Class Recall		
Run	Accuracy	High Overrun	Near	Under run	High Overrun	Near	Under run
Run 1	38.91%	35.58%	42.70%	0.00%	50.92%	45.00%	0.00%
Run 2	40.10%	37.67%	46.79%	0.00%	74.31%	28.08%	0.00%
Run 3	42.32%	39.37%	49.42%	0.00%	74.77%	32.69%	0.00%
Run 4	42.15%	39.64%	45.56%	0.00%	61.47%	43.46%	0.00%
Run 5	45.39%	44.79%	45.51%	0.00%	19.72%	85.77%	0.00%

3. RESULTS

Table 3 summarizes the predictions made using only text inputs. Five model runs were made. Each run used used different groups of projects for the training and testing cases. The prediction accuracy of the models ranged from 38.91 % to 45.39%. Precision is a measure of the accuracy of a cost overrun prediction. That is, if the model predicts a large cost overrun, the precision is the percentage of large cost overrun predictions that are correct. Recall is a measure of the models capability to select the correct level of cost overrun for a test project. It is the ratio of correct predictions for a certain level of cost overrun and the total number of projects that were predicted to be at that level of overrun. The models performed best in predicting projects with large cost overruns and projects completed near the low bid amount. The model was unable to correctly categorize any of the under run testing projects and only categorized projects as “High Overrun” or “Near”

Table 4 shows the results of the model that combined the text data with the numeric value of the low bid amount. The addition of the low bid amount appears to generate models that provide similar or slightly better results then the text only model. For the five models produced overall accuracy varied from 38.40% to 44.20%. Additionally, this model was able to correctly predict from 20-25% of the under run projects.

Table 4. Predictions with Text and Numeric Low Bid

		Class Precision			Class Recall		
Run	Accuracy	High Overrun	Near	Under Run	High Overrun	Near	Under Run
Run 1	44.20%	46.13%	48.13%	28.41%	55.74%	41.87%	23.81%
Run 2	40.78%	42.47%	46.04%	19.05%	48.42%	44.79%	15.09%
Run 3	43.17%	49.61%	47.40%	23.01%	27.04%	63.81%	27.08%
Run 4	38.40%	37.72%	44.23%	22.43%	28.77%	52.67%	28.86%
Run 5	42.66%	35.52%	50.76%	25.00%	30.37%	60.81%	19.19%

4. DISCUSSION

The model predictions had accuracies in the range of 38-46%. This initial work indicates that there is information contained in text data that may give indications of a projects likely cost performance. The results indicate that additions need to be made to the input data to boost the models predictive accuracy. The model that included both text and numeric inputs performed better then the text only model. This indicates exploration of other numeric descriptors of a project is warranted to boost model performance. There may also be subjective ratings of project complexity available from highway agency experts that could be included in the input to define as project as complex or routine. Possibly, a larger database of project text needs to be collected so that a larger pool of highly weighted word tokens can be identified.

The input data used for this analysis included only a very short description of the project. Perhaps, more detailed written summaries about a project may be available in the design documents. Potentially, using more detailed descriptions of the project could yield more accurate prediction by better describing any unusual features of a project.

It is widely recognized that many factors other then a description of the work to be performed affect the completed project cost. Shane et al. (2009) have identified internal and external factors that affect projects such as faulty execution, scope change, and unforeseen site conditions. These types of problems are not reflected in the text describing a construction projects scope and work tasks. Forecasting using text descriptions of a projects features will never achieve the highest levels of accuracy unless additional variables can be added to the input that reflect the construction contractor’s past performance and management skills, as well as metrics of a projects complexity.

5. CONCLUSIONS

This research has demonstrated how textual descriptions describing a construction project can be processed for use in a data mining system. The initial results indicate that there is often a relationship between the text that describes a construction project and the level of the cost increases that occur during construction. The combination of text and a numeric input produced better predictions The results indicate that additional work is warranted to collect more data and to expand the model inputs to increase the prediction accuracies. Future research in this area will include the collection of more data, the collection of data from other government agencies to see if model findings are generalizable to other jurisdictions, and the inclusion of more input variables to include contractor performance and project complexity metrics in the model.

ACKNOWLEDGMENTS

This research was supported by the Center for Advanced Infrastructure and Transportation at Rutgers University.

REFERENCES

Caldas, C.H. & Soibelman, L. (2003). Automating hierarchical document classification for construction management information systems, *Automation in Construction*, 12(4), 395-406

Cleary, J. G., & Trigg, L. E. (1995). K*: An Instance-based Learner Using an Entropic Distance Measure, *Machine Learning-International Workshop Then Conference,* San Francisco, USA, 108-114, Morgan Kaufmann Publishers, Inc.

Karamanis, N. (2007). Text Mining for Biology and Biomedicine. *Computational Linguistics*, 33(1), 135.

Ivanciuc, O., (2007). Applications of support vector machines in chemistry, *Reviews in computational chemistry*, 23, p. 291.

Miner, G., Elder IV, J., Hill, T., Nisbet, R., and Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

Saha, S.K., Mitra, P. & Sarkar, S. (2012) A comparative study on feature reduction approaches in Hindi and Bengali named entity recognition, *Knowledge-Based Systems*, 27, 322-32

Shane, J.S., Molenaar, K.R., Anderson, S. & Schexnayder, C. (2009). Construction project cost escalation factors, *Journal of Management in Engineering*, 25(4), 221-9

Soibelman, L., and Kim, H. (2002). Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering*, *16*(1): 39-48.

Soibelman, L., Wu, J., Caldas, C., Brilakis, I. & Lin, K.-Y. (2008). Management and analysis of unstructured construction data types, *Advanced Engineering Informatics*, 22(1), 15-27

Son, H., Kim, C. & Kim, C. (2012). Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables, *Automation in Construction*, 27, 60-6

Witten I.H., & Frank, E. (2005) *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

Weiss, S., Indurkhya, N. and Zhang, T. (2010). *Predictive text mining: a practical guide*. Springer-Verlag.