ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 3, Issue 3 (July - Sept. 2015)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2015)

Construction Cost Overturn Prediction Using Porter on Text and Numerical Data

[I]Payal Singla, [II]Er. Mohit Kakkar
[I]Student, M.Tech CSE, [II]Assistant Professor, Dept. of Computer Science
[I,II]Desh Bhagat University, MandiGobindgarh, Punjab, India

Abstract
In this paper we have studied construction cost overturn prediction using porter on text and numerical data in a NLP to produce a prediction of the level of cost overrun using data mining classification algorithms. This model shows an average accuracy of 43.72%. This paper discusses how text can be processed, combined with numeric values. The pros and cons of the technique are discussed and a conclusion is devised.

Keywords
Feature extraction, Feature selection, Natural language process, pre-processing.

I. Introduction

Natural language processing is the process of extracting the information from the language text such as English into the meaningful result by applying the specialized artificial intelligence methods. When the text is gathered it may be loosely organized and may be interpreted as instructed text or missing information. If the text is not scanned properly then text mining leads to the garbage in out phenomena which affects the accuracy of the output. Therefore preprocessing phase organizes and guarantees successful implementation text analysis into different categories. There are two methods of pre-processing:

A. Feature Extraction

It is based upon morphological analysis (MA), syntactical analysis (SA), and semantic analysis (SA). MA deals with single words, tokenization, remove stop words and stemming-word. SA provides knowledge about the grammatical structure of a language known as syntax analysis. Semantic analysis deals with the meaning of the words based on Word Net-Affect.

B. Feature selection

The main aim of FS is to eliminate irrelevant and repeated information from text. It is based upon frequency based (FB), latent semantic indexing (LSI), and random mapping (RM). Frequency based FS deals with number of occurrences of a term related with the topic. LSI tends to improve the lexical matching by adopting approach. RM creates a map through the contents of a large document set.

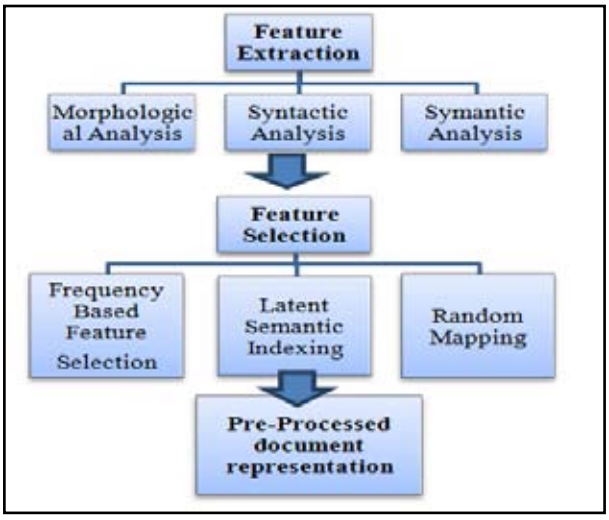

Fig.1: Pre-processing

C. Text mining using classification

It is a process of learning set of rules from a set of examples in a training set. It classifies each text to a certain category that is machine learning based and ontology based text classification.

D. Construction cost overturn prediction

Many factors affect construction cost overruns. It has now become possible to combine various text and data mining methods to make predictions to experience cost overruns. Text mining is automatic discovery of previously unknown information then with the use of data mining to discover new association. It has been found to give best result for some predictive applications.

II. Literature Survey

1.Williams, Trefor P.et. al. [1] has worked Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. This paper discusses how text can be combined with numerical data to produce a prediction of cost overrun using data mining classification algorithms. The stacking model had an average accuracy of 43.72% for five model runs. It was found that a stacking model that used only numerical data produced predictions with lower precision and recall. A potential application of this research is tobudget sufficient amount to complete a construction project.

2.Poria, Soujanyaet. al. [2] has proposed the use of Sentic patterns for the purpose of the sentiment analysis from the social data. The authors proposed the use of dependency-based rules for concept-level sentiment analysis. In this work, the authors have introduced a novel paradigm to concept-level sentiment analysis common-sense computing, and machine learning for improving the accuracy of tasks.

3.Narayanan, Viveket. al. [17] has worked on a fast and accurate sentiment classification using an enhanced Naive Bayes model. The authors have explored different methods of improving the accuracy of a Naive Bayes classifier for sentiment analysis. They have also observed that a combination of methods like effective negation handling, word n-grams and feature selection by mutual information results in a significant improvement in accuracy.

4.2015. Khan,Atifet. al. [4] has proposed the framework for multi-document abstractive summarization based on semantic role labeling. The authors have proposed a framework for abstractive summarization of multi-documents, which aims to select contents of summary not from the source document sentences but from the semantic representation of the source documents. Content selection for summary is made by ranking the predicate argument structures

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

based on optimized features, and using language generation for generating sentences from predicate argument structures. Our proposed framework differs from other abstractive summarization approaches in a few aspects.

5.Ramnath, Krishnanet. al. [5] has been proposed the model known as AutoCaption. The outputs of the modules are combined to generate a large set of candidate captions, which are returned to the phone. The phone client includes a convenient user interface that allows users to reorder, add, or delete words to obtain the grammatical style they prefer.

6.C.G. Wilmotet.al.[13] has proposed the model known as highway construction. The most influence factors were found to be the cost of material, labor and equipments used in constructing the facility.

7. 2013 J. Zhanget.al.[6] has proposed semantic modeling, semantic natural language processing techniques .It used six phase iterative approach that contains text classification, information extraction, information transformation, compliance reasoning. Algorithms used for discussing and presenting information transformation which transform the extracted information.

III. Issues in Existing Approaches

- Existing system does not analyze grammatical rules in order to get the clear meaning of the sentence which adds misinterpretation of the text.
- Existing system does not use branching(reverse stemming) to fix grammatical errors which leads to the low performance of the system.
- Existing system does not automatically decide N in N-Gram text which sometimes makes it difficult to differentiate the phrase.
- Existing system offer very low accuracy which is less than 50%.

References

[1] *Williams, Trefor P., and Jie Gong. "Predicting construction cost overruns using text mining, numerical data and ensemble classifiers." Automation in Construction 43 (2014): 23-29.*

[2] *Poria, Soujanya, Erik Cambria, GrégoireWinterstein, and Guang-Bin Huang. "Sentic patterns: Dependency-based rules for concept-level sentiment analysis." Knowledge-Based Systems 69 (2014): 45-63.*

[3] *Phu, Vo Ngoc, and PhanThiTuoi. "Sentiment classification using Enhanced Contextual Valence Shifters."In Asian Language Processing (IALP), 2014 International Conference on, pp. 224-229. IEEE, 2014.*

[4] *Khan, Atif, NaomieSalim, and Yogan Jaya Kumar. "A framework for multi-document abstractive summarization based on semantic role labelling." Applied Soft Computing 30 (2015): 737-747.*

[5] *Ramnath, Krishnan, Simon Baker, Lucy Vanderwende, Motaz El-Saban, Sudipta N. Sinha, AnithaKannan, Noran Hassan et al. "AutoCaption: Automatic caption generation for personal photos." In Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, pp. 1050-1057. IEEE, 2014.*

[6] *J. Zhang, N.M. El-Gohary, Information transformation and automated reasoning for automated compliance checking in construction, Proceedings of the ASCE International Workshop on Computing in Civil Engineering, Los Angeles, CA, 2013, pp. 701–708.*

[7] *H. Son, C. Kim, C. Kim, Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables, Autom. Constr. 27 (2012) 60–66.*

[8] *C. Gkritska, S.S. Labi, Estimating cost discrepancies in highway contracts: multistep econometric approach, J. Constr. Eng. 134 (12) (2008) 953–962.*

[9] *S.M. Trost, G.D. Oberlender, Predicting accuracy of early cost estimates using factor analysis and multivariate regression, J. Constr. Eng. 129 (2) (2003) 198–204.*

[10] *K.M. Nassar, W.M. Nassar, M.Y. Hegab, Evaluating cost overruns of asphalt paving project using statistical process control methods, J. Constr. Eng. 131 (11) (2005) 1173–1178.*

[11] *K. Petroutsatou, E. Georgopoulos, S. Lambropoulos, J.P. Pantouvakis, Early cost estimating of road tunnel construction using neural networks, J. Constr. Eng. 138 (6) (2011) 679–687.*

[12] *T.P. Williams, Bidding ratios to predict highway project costs, Eng. Constr. Archit. Manag. 12 (1) (2005) 38–51.*

[13] *C.G. Wilmot, G. Cheng, Estimating future highway construction costs, J. Constr. Eng. 129 (3) (2003) 272–279.*

[14] *I. Mierswa, et al., Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Rapid prototyping for complex data mining tasks, Yale, 2006. pp. 935–40.*

[15] *M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor. Newsl. 11 (1) (2009) 10–18.*

[16] *M.A. Hearst, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Untangling Text Data Mining, 1999, pp. 3–10.*

[17] *Narayanan, Vivek, IshanArora, and Arjun Bhatia. "Fast and accurate sentiment classification using an enhanced Naive Bayes model." In Intelligent Data Engineering and Automated Learning–IDEAL 2013, pp. 194-201. Springer Berlin Heidelberg, 2013.*