A Mutually Beneficial Integration of Data Mining and Information Extraction

Un Yong Nahm
Department of Computer Sciences,
University of Texas, Austin, TX 78712-1188

Raymond J. Mooney
pebronia@cs.utexas.edu
mooneyg@cs.utexas.edu

Extended abstract of the paper-**'A Mutually Beneficial Integration of Data Mining and Information Extraction'** by:
Noopur Nishikant Zambare(B20ME051)

1 Introduction

This paper aims to benefit the process of Information Extraction with the help of Data Mining techniques. Information from the documents is collected as a database and is sent for the flattening into more composed and structured set via the process of Information Extraction. Data mining algorithms are used for the structured data. DISCOTEX stands for DISCOvery from Text Extraction.
KDD creates relationships between the various slot filters that gives the additional clues about what text is actually useful from the document. Hence this data goes through RAPIER system and then DiscoTex. Applying internal filters is however optional. Data mining techniques generates set of rules from the extracted text. These set of rules can be learned and even can be used for the future predictions.

2 Method

This paper focuses on the implementation of Data Mining techniques in order to benefit the Information Extraction process. The work of Information Extraction is to convert unstructured data into a structured database. Data mining deals with extracting the useful and meaningful data out of the large data sets. It discloses the relations from the hidden patterns and generates new set of rules based on it. It consists of tools that can predict the behaviors and could even predict the future trends.
KDD holds the meaning Knowledge Discovery in Databases. It refers to the broad procedure of discovering knowledge in data and emphasizes on the applications of domain specific Data Mining techniques. It actually implements the data specific techniques. Data Mining is the root of the KDD procedure, consisting of algorithms to judge the data and create a model by discovering the relationships between the texts.
This paper has described how KDD could be used in IE in order to improve the performance. Natural-language information extraction methods can change the unstructured collection of textual documents into a more structured and organized database.

- Standard KDD methods can then be implemented over this structured database in order to discover and generate novel relationships.
- The predictive relationships between different slot fillers discovered by KDD can provide additional clues about what information should be extracted from a document.
- Thus it helps in increasing the predicting accuracy of the model.

3 Experiment And Evaluation

It has been experimented over the data set consisting the 600 job announcement templates along with 4000 additional input to generate the rules. Fillers have been used in the slots of languages, platforms, applications and Areas as they are usually filled with multiple value possibilities.

To evaluate the model, precision and recall were calculated. Precision is the measure of how much predictions were actually correct out of the positive identifications. Whereas recall gives measure of how much of the proportion were actual positives, correctly identified.

$$precision = \frac{CorrectFillersExtracted}{FillersExtracted}$$

$$recall = \frac{CorrectFillersExtracted}{FillersInCorrectTemplates}$$

$$F-easure = \frac{2PrecisionRecall}{Precision+Recall}$$

4 Future Research And Scope

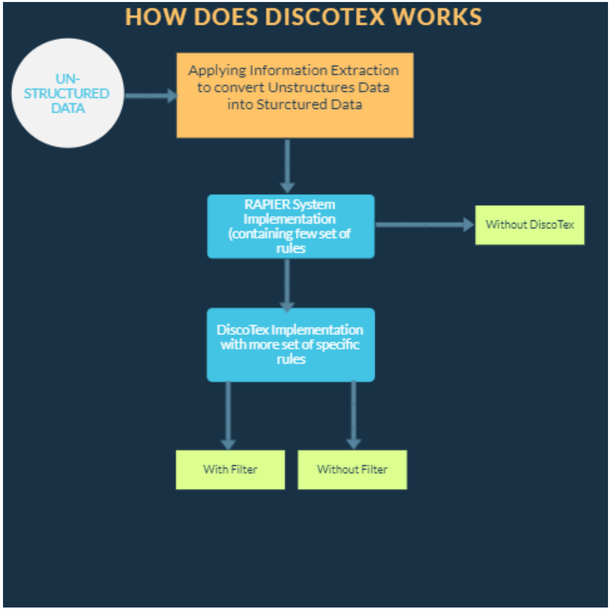
It can be used be used in studying Bussiness Analytics, for filtering emails and the related documents. Could be used to judge and extract the domain specific information from the large data sets.

5 Limitations

There are some bounds to this experiment. It was performed under limited data set hence there is possibility of over fitting issues within the model. Also the attributes used in the job titles were less in number that reduced the dimensions of the data.

6 Conclusion

This model was experimented for three different models consisting one just without DiscoTex, another with DiscoTex over RAPIER without filter and DiscoTex over RAPIER with filter. For the comparison, taking each training set size, each pair of systems were compared to determine if their differences in recall and F-measure were found to statistically significant (p < 0:05).
From the graphs given in the paper, it could be clearly stated that both precision and F-measure are dependent on the training data size and holds a direct proportionality


HOW DOES DISCOTEX WORKS
UN-STRUCTURED DATA
Applying Information Extraction to convert Unstructures Data into Sturctured Data
RAPIER System Implementation (containing few set of rules
Without DiscoTex
DiscoTex Implementation with more set of specific rules
With Filter
Without Filter

A model for retrieving information