International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 9, Issue 3 (December 2013), PP. 42-48

Information Extraction andRule Prediction Using DiscoTEX

Sonal D. Raut[1], Ms. Prashasti Kanikar[2]
[1]*Mtech (CE) StudentComputer Engineering DepartmentSVKM's NMIMS University*
Mumbai, Maharashtra.
[2]*Asst. ProfessorComputer Engineering DepartmentSVKM's NMIMS University*
Mumbai, Maharashtra.

Abstract:-Text Mining is a process of deriving high quality information from the text. Text mining usually involves processing of the unstructured document, extracting features from the document and storing it in the database and finally using KDD techniques rules are evaluated. IE module is use to convert the corpus text into a structured database from which rules are mined. This paper describes a system called DiscoTEX (Discovery from Text Extraction) which integrates IE(information extraction) module with the KDD module. For DiscoTEX an IE learning system Rapier (Robust Automated Production of Information Extraction Rules) is used which automatically creates the structured database.

Keywords:- Text Mining, KDD, DiscoTEX, Rule Mining, Rule Induction, Information Extraction(IE).

I. INTRODUCTION

The problem of text mining from unstructured text is attracting increasing attention. This paper suggests a new framework for text mining based on the integration of Information Extraction (IE) and traditional Knowledge Discovery from Databases (KDD). Traditional data mining assumes that the information to be mined is already in the form of a relational database.

But unfortunately, for many applications, electronic information is available in the form of unstructured natural-language documents rather than structured databases. Information extraction plays an important role in creating a structured database which is then provided to the KDD module for discovering rules. Fig. 1 shows the simple diagram for text mining using IE module and KDD module.

Constructing an IE system is a difficult task. By manually annotating a small number of documents with the information to be extracted, a reasonably accurate IE system can be induced from the labeled corpus and then the IE system is applied to a large corpus of text to construct a database

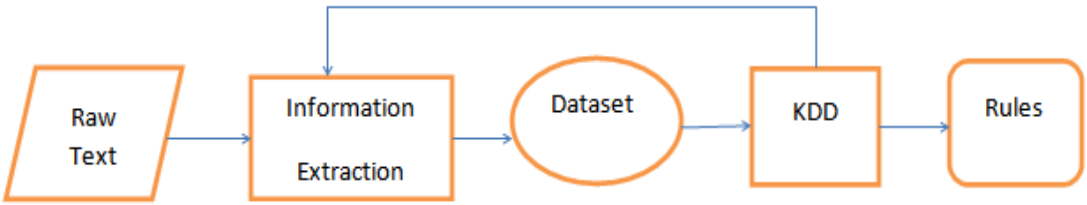

Fig. 1: Overview of IE-based Text Mining Framework [1]

The remainder of the paper is organized as follows. Section II describes a system called DiscoTEX (Discovery from Text Extraction) that combines IE and KDD technologies to discover prediction rules.
Section III describes performance measure metrics for DiscoTEX. Section IV, Section V, Section VI talks about the inferences, future scope, conclusion respectively.

A. Information Extraction

The process of information extraction consists of two main steps:

1. Extracting textual information- First three blocks
2. Compression of extracted information- Last three blocks

Extracting textual information is composed of three steps. First step is of Tokenization. Next all the special characters are removed after which all the stop words are removed. After extracting the textual

information the extracted information needs to be compressed. For compressing similar words (synonyms) are combined. Then Stemming is performed and finally the most frequently occurring words are extracted.

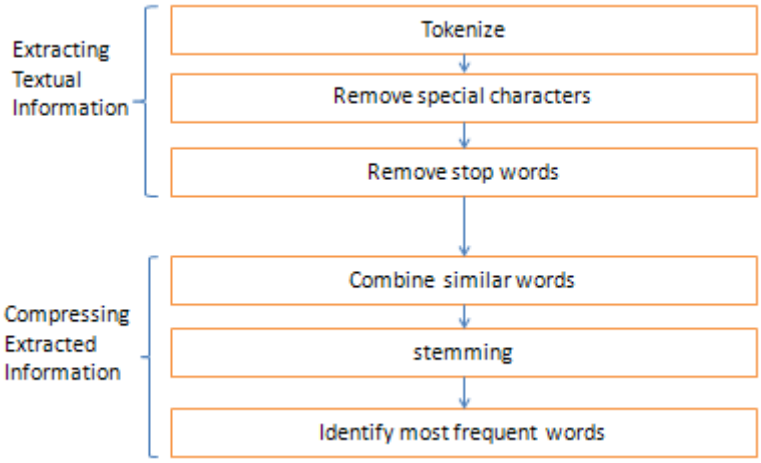

Fig. 2: Information Extraction

As shown in Fig.2, in the first step all the Keywords are extracted from each document. This process is called tokenizing. Secondly, the words are converted to lower case and all numbers and special characters are removed.

1) Stop Words And Stemming: Stop Words are those wordswhich are filtered out to reduce the size of the document which then helps in searching.Any group of words can be chosen as the stop words such as "the", "is", "at", "which", "and", "to" and so on. Overall document contains about 400-500 stop-words so removing these stopwords greatly reduces the size of the document.In stemming words are deconstructed to their basic form. All the related words will map to the same stem. For example words like "stemmer", "stemming", "stemmed" will be reduced to the stem.Similarly words like "fishing", "fisher", "fished" will be reduced to the root word fish. Finally, the most frequently used words descriptions are extracted.

II. THE DISCOTEX SYSTEM

A. Information Extraction

The main job of the IE system is to extract specific features from the corpus text(unstructured document) and store it in the database. Then various KDD techniques (C4.5 or ID3) can be applied to the database to generate the set of rules. IE extracts templates from the corpus text which contains the slots and the corresponding filler which is a substring of the document.

Sample Document

Title: Web Development Engineer
Location: Chicago, Austin

This candidate is responsible for design and implementationof the web-interfacing components of the (AccessBase)serverand development of the back-end duties.A candidate should have experience that includes:
One or more: Solaris, Linux plus Windows, IBM AIX
Programming: C/C++, Java, .Net
For Database access and integration: Oracle, ODBC
For CGI and scripting: Javascript, VBScript, Perl, PHP
Exposure to the following is a pluspoint: JDBC, Flash, JSP
Experience: more than 3 years of experience is required.

Filled Template To Be Extracted

Title: Web Development Engineer
Location: Chicago, Austin
Languages: C/C++, Java, Javascript, VBScript, PHP, ASP, .Net,Perl
Platforms: Solaris, Linux, Windows, IBM AIX,
Applications: Oracle, Flash, ODBC, JSP, JDBC
Areas: Database, scripting,CGI
Years of experience: more than 3 years

In this paper, we consider the task of extracting a database from postings to the USENET newsgroup, austin.jobs. Above is a sample message from the newsgroup and the filled computer-science job template where several slots may have multiple fillers. From the document the slots extracted are title, salary, city, language, platform, application, area, experience, degree.

Some of the slots have multiple slot fillers like language, area, application, while some have only one slot fillers like city, degree, and salary.As austin.jobs is not a moderated newsgroup as before posting on this newsgroup, posts are not approved by one or more individuals so many documents are posted. Not all posted documents are relevant to our task. So, before constructing a database using an IE system, irrelevant documents can be filtered out from the newsgroup using a trained text categorizer such as Naive-Bayes text categorizer. Naive-Bayes text categorizer [6] [7] was trained on this data to identify relevant documents. Rapier a machine learning IE system is use for extracting the information from the corpus text and it constructs an IE module for DiscoTEX.

B. Rule Induction

After constructing an IE system that extracts the desired set of slots for a given a sample document, a database is constructed from a corpus of texts by applying the extractor to each document to create a set of structured records (database). Standard KDD techniques are used to discover relationships. Various rule induction algorithm exist like local rule induction, global rule induction. DiscoTEX uses C4.5 rules. C4.5 is a decision tree algorithm which extracts rule by creating a decision tree. Before extracting the rules similar slot fillers are first collapsed into a standard predefined term. For example, "C++" is a popular filler for the platform slot, but it often appears as "C ++", "C+ +".Similarly ActiveX and Active X will be assigned to the same slot. All the similar fillers are then assigned to a unique slot. Table I: shows some of the fillers that are assigned to a unique slot after which rules are rules are induced using rule induction.

Table I: Synonym Dictionary[1]

Standard Term	Synonyms
"Access"	" MS Access", " Microsoft Access"
"ActiveX"	"Active X"
"AI"	"Artificial Intelligence"
"ATM"	"ATM Svcs"
"C"	"ProC", " Objective C"
"C++"	"C ++", "C+ +"
"Client/Server"	"Client / Server", " Client-Server"

Discovered knowledge are written in the form of production rules. Suppose we have a condition that when area is database the language is Mysql, then it can be written as Database ε *area*→Mysql ε *language*.C4.5 rules can be used for building decision trees from which the rules are created.In order to discover prediction rules, each slot-value pair in the extracted database is treatedas a distinct binary feature, such as "graphics ε *area*" and then learn rules for predicting each of the binary feature from all other features. Following shows a sample of rules mined for the computer science job postings.

Sample

1. XML ε*language* → HTML ε*language*
2. Oracle ε*application* → SQL ε*language*
3. HTML ε*language* → Database ε*area*
4. ODBC ε*application* → JSP ε *language*
5. Java ε*language* → Web ε*area*

C. System Architecture

First, documents annotated by the user (user labeled documents) are provided to Rapier as training data. Then the rules are induced from this training data and placed in the rule base set. System architecture is as shown in Fig. 3.

The learned IE system(Rapier) then takes unlabeled texts and transforms them into a database of slot values, which is provided to the KDD component (i.e. C4.5or Ripper) as a training set for constructing a knowledgebase of prediction rules. The training data for KDD can include a larger IE-labeled set automatically extracted from raw text as well as user-labeled documents which are used for training IE. Rule Induction

algorithm used is C4.5. These prediction rules are then used during testing to add additional slot fillers whose presence in the document are confirmed before adding them to final extraction template.

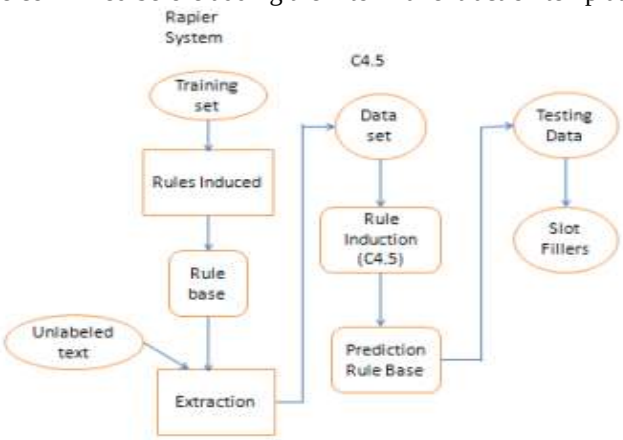

Fig. 3: System Architecture Of DiscoTEX

D. Proposed Methodology

Fig. 4: shows the process of Rule Mining. All the rules that are extracted may not be relevant, so a threshold value T is set for rule validation. A database is created by applying IE to a corpus of document. A slot value pair is selected from the database and treated as a binary feature. Then rules are produced and stored in the prediction rule base.

In order to check the validity of the rules each of the prediction rules are verified on training and validation data. In this process some of the rules may get filter out.A final step shown in the figure is filtering the discovered rules on both the training data and a validation data in order to retain only the most accurate of the induced rules. The rules that make incorrect predictions on either training or validation set are discarded.

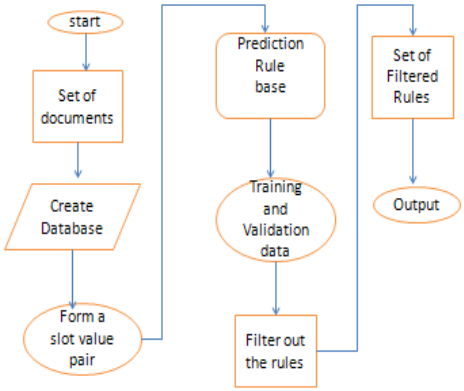

Fig. 4: Rule Mining

KDD process intern helps information extraction process to discover additional slot fillers, thus improving the recall. Fig. 5 shows the flowchart for extracting additional slot fillers.

RB – Set of discovered rules
d – Set of documents
D – Labeled example in database
F- Set of additional slot fillers

For every labeled document in the database extract fillers and store it in F. The final decision of whether or not to extract predicted slot filler is based on whether the filler or any of its synonyms occurs in the document as a part of the string. If the filler is found, then the extractor considers its predicted slot filler appropriate and extracts the filler.

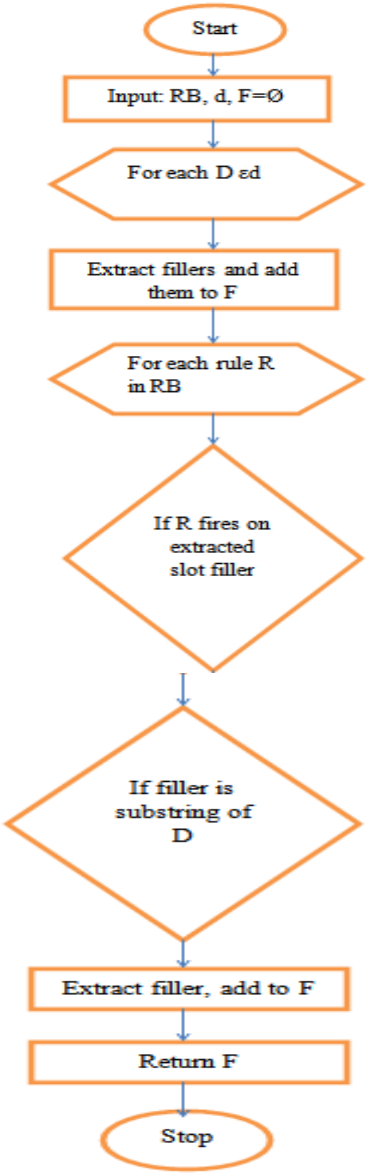

Fig. 5:IE For Additional Slot Fillers.

III. PERFORMANCE MEASURE

We evaluate the performance of DISCOTEX using IE performance metrics of precision, recall and F-measure with respect to predicting slot fillers.

Recall is defined as the ratio of the number of relevant records retrieved to the total number of relevant records in the database as shown in Fig.6. For example to search "Slot Fillers" in the document, if out of 80 slot fillers only 30 are retrieved then the recall is 0.375. The remaining 50 terms that are relevant but not retrieved are called as silences.

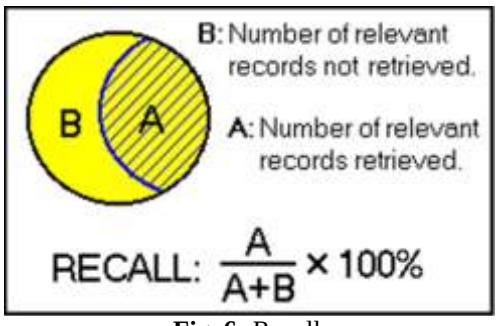

Fig. 6: Recall

Precision is defined as the ratio of the number of relevant records retrieved to the total number of both relevant and irrelevant records retrieved. For example to search "Slot Fillers" in the document, if there are 100 candidates out of which 70 are slot fillers then the precision is 0.70. The remaining 30 terms which are irrelevant and not retrieved are called as Noise.Fig.7 shows the diagrammatic representation of precision. Precision and recall are inversely proportional to each other. Precision increases when the noise is less and as the noise is less, the number of relevant terms are more, which means recall is less, as the denominator for recall is the total number of relevant terms. So it is very difficult to achieve high recall and high precision simultaneously. So it is preferable to maintain a higher degree of precision or recall. Generally the proportion of noise and silences also vary inversely.

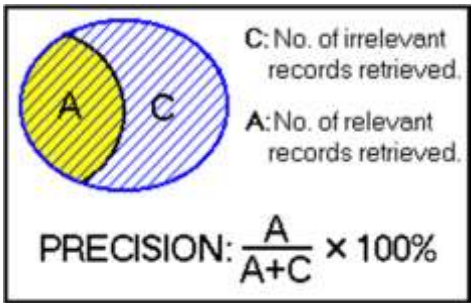

Fig. 7: Precision

When same weight is given to both precision and recall F-measure is computed. It is a harmonic mean of precision and recall.

$$\text{F-measure} = \frac{(2*\text{Precision}*\text{Recall})}{(\text{Precision} + \text{Recall})}$$

IV. INFERENCES

Text Mining could be either summarizing the document or extracting keywords in the form of rules. In this paper DiscoTEX system is used which integrates information extraction module with KDD module. Initially a template is prepared with the slot and the fillers which isto be extracted. Using this template a structured database is created. DscoTEX uses an automatically learned IE system (Rapier) to extract a structured database from a text corpus, and then mines this database with existing KDD tools.Before extracting rules similar words are assigned to a unique slot. KDD module intern helps extraction of the keywords with the help of the discovered rules.Rapier a learned machine system is use to perform information extraction and it constitutes an IE module for DiscoTEX system.

Paper1 and 2: IE system used is RAPIER. For generating rules C4.5 which is a decision tree algorithm is used.

Paper3: IE system used is BWI. For generating rules Soft-Apriorialgorithm is used.

V. FUTURE SCOPE

In futureDiscoTEX can be applied to larger text corpora. Also in this paper we have considered only one domain of database so in future DiscoTEX can be applied to other domains like medical, infrastructure and business. Also some other good metrics for evaluating the performance measure is required.

VI. CONCLUSION

Information Extraction (IE) enables the application of KDD to unstructured text corpora and KDD can discover predictive rules useful for improving further the performance of IE. There is increasing interest in the topic of text mining [8].Text Mining is relatively a new area of research for data mining, information extraction, machine learning and NLP (natural language processing). By properly integrating techniques from each of these areas new methods for extracting from large corpus of document can be discovered. Also results [2] demonstrate the knowledge obtained from an automatically extracted database is very close in accuracy to the knowledge extracted from the database that is manually constructed. Also the development of the effective text mining is becoming somewhat critical because of the growing interaction between computational linguistics and machine learning [4].

ACKNOWLEDGEMENT

This research was supported by Ms. Prashasti Kanikar. I would like to thank her for her support.

REFERENCES

[1]. Raymond J. Mooney and Un Yong Nahm, Department of Computer Sciences University of Texas, Text Mining with Information Extraction, Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.) pp.141-160, Van Schaik Pub., South Africa, 2005.

[2]. Un Yong Nahm, Raymond J. Mooney,Department of Computer Science, University of TEXAS, Using Information Extraction to Aid the Discovery of Prediction Rules from Text, Proceedings of the KDD(Knowledge Discovery in Databases)-2000Workshop on Text Mining, pp.51-58, Boston, MA, August 2000.

[3]. Un Yong Nahm and Raymond J.Mooney Department of Computer Sciences The University of Texas at Austin, Using Soft-Matching Mined Rules to Improve Information Extraction

[4]. C. Cardie and R. J. Mooney. Machine learning and natural language (Introduction to special issue on natural language learning). Machine Learning, 34:5–9, 1999.

[5]. M. W. Berry, editor. Proceedings of the Third SIAM International Conference on Data Mining(SDM-2003) Workshop on Text Mining, San Francisco, CA, May 2003.

[6]. A. K. McCallum. Bow: A toolkit for statisticallanguage modeling, text retrieval, classification andclustering, 1996.

[7]. T. Mitchell. Machine Learning. McGraw-Hill, NewYork, NY, 1997.

[8]. M. A. Hearst. Untangling text data mining. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), pages 3–10, College Park, MD, June 1999.