Semantically Enriched Knowledge Extraction With Data Mining

Anuj Tiwari
Department of Civil Engineering
Indian Institute of Technology-Roorkee
India

P. Srujana
Computer Science and Engineering
CMR Technical Campus-Hyderabad
India

K. Rajesh
Computer Science and Engineering
CMR Technical Campus-Hyderabad
India

Abstract — **while data mining has enjoyed great popularity and success in recent years, Semantic web is shaping up as a next big step in the evolution of World Wide Web. It is the way web is growing as a smarter cyberspace. In field of Information and communication technology huge amount of data is available that need to be turned into knowledge. On the one side Data Mining is a nontrivial extraction of implicit, previously unknown and potentially useful knowledge from data in databases and on the other side Semantic web developing new platform to represent extracted knowledge in both the machine and human understandable format. The aim of this paper is to explore the concept of data mining in the context of semantics. Paper uses a basic input dataset with an open source software WEKA and a commercial one SAS for knowledge discovery; further this knowledge is represented in human understandable format with NLP (Natural Language Processing library) and in machine understandable format (RDF) with an indigenous algorithm implemented with java.**

Keywords— Data Mining; Semantic Web; Ontology; Knowledge; RDF; WEKA; SAS.

1. INTRODUCTION

We all are surrounded with a lot of data. Every time we watch television, we do any type of search on the internet, we swipe our ATM card etc more and more data is generated. In order to explore, analyze and discover valid, implicit, novel, understandable, potentially useful patterns, associations or relationships in large quantities of data a number of analytical tools are required that allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Data mining is the collection of methods that analyze data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both.

Valid: The patterns hold in general.
Novel: We did not know the pattern beforehand.
Useful: We can devise actions from the patterns.
Understandable: We can interpret and comprehend patterns.

Data mining deals with what kind of patterns can be mined. In this world of information and communication technology data is continuously growing like anything. This flood of data and sophisticated tools of data mining together very productive for business purpose where companies are interested in various patterns like purchase, educational, traffic, habits etc.

Data mining tasks are generally divided into two major categories. The objective of predictive tasks is to predict the values of a particular attributes based on the values of other attributes while Descriptive tasks derive patterns (correlations, trends, clusters, trajectories and anomalies) to summarize the underlying relationships in data.

Evolution of internet in last couple of decades brought a remarkable growth in the development of new technologies and applications, contributing to a historic transformation in the way we work, communicate, socialize, learn, create and share information, and organize the flow of people, ideas, and things around the globe. Being an extension of the existing web technology 'Semantic Web' is well recognized now as an effective infrastructure to enhance visibility of knowledge on the Web for humans and computers alike [1]. 'Semantic Web' enables the description of contents and services in machine-readable form, and enables annotating, discovering, publishing, advertising and composing services to be automated. It was developed based on Ontology, which is considered as the backbone of the Semantic Web [2]. Jasper and Uschold identify three major uses of semantic web and ontologies [3]:

(i) To assist in communication between human and computers,
(ii) To achieve interoperability (communication) among software systems, and
(iii) To improve the design and the quality of software systems.

2. METHODOLOGY

Methodology adopted for RDF generation is as follows:

Step 1: Input Data is retrieved from the DBMS.
Step 2: Open source data mining tool WEKA is used for data preprocessing and Nominal data set preparation.

Step 3: Nominal data is passed to commercial data mining tool SAS, regression is applied on that data.
Step 4: Output dataset is saved in database.
Step 6: Open NLP is a tool which is used to parse the data which is retrieved from the database.
Step 7: Triples are extracted from the sentences using Stanford NLP parser tool.
Step 8: RDF is generated using APACHE JENA.

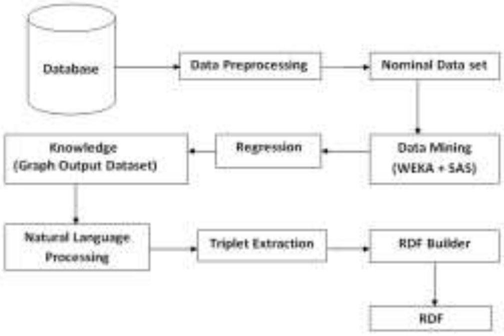


Fig. 1. Flow digram dipecting adopted methodology.

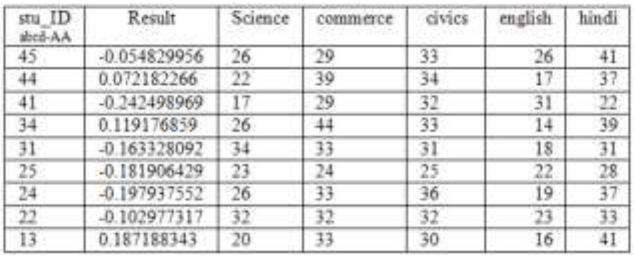
Standard database management system (like mySql) is used to save and mange a larger database. This larger database is preprocessed to generate nominal data set which in turn mined with open source data mining tool WEKA and commercial data mining tool SAS. WEKA generates graphical results and with SAS tabular results are obtained. Here regression is used as data mining method. Extracted knowledge is processed with Natural Language Processing (NLP) library and human interactive triplets are generated. RDF Builder step process these triplet and generate machine understandable RDF data set.

3. DESCRIPTION

A. Input Dataset

Data is stored in MS ACCESS database. It is retrieved as ".CSV" file. The data file is shown below:

Table 1. Input dataset.



stu_ID abcd-AA	Result	Science	commerce	civics	english	hindi
45	-0.054829956	26	29	33	26	41
44	0.072182266	22	39	34	17	37
41	-0.242498969	17	29	32	31	22
34	0.119176859	26	44	33	14	39
31	-0.163328092	34	33	31	18	31
25	-0.181906429	23	24	25	22	28
24	-0.197937552	26	33	36	19	37
22	-0.102977317	32	32	32	23	33
13	0.187188343	20	33	30	16	41

B. Data Mining Method

Input data is pre-processed using WEKA filters, here unused data is removed. The preprocess panel is shown in Fig 2. On this pre-processed data, regression is applied in order to derive the relation between independent and dependent variables. Independent variables are considered as the input values to the model and dependent variable is considered as output of the model. Regression finds the relationship between dependent and independent variables. In the dataset we considered, the dependent variable is "result" and subjects are the independent variables. Here we establish the relationship between the overall results of students to their performance in individual subjects.

Here, linear regression is used. The data we considered is linear in nature. Regression is given by the formula:

$$Y=a+bX$$

Where,

Y = Dependent variable
a= Intercept
b= Slope
X= Independent variable

C. Output

The output of the data pre-process using WEKA is shown in Fig 3.

Fig. 2. Image dipecting primary preprocessed results.

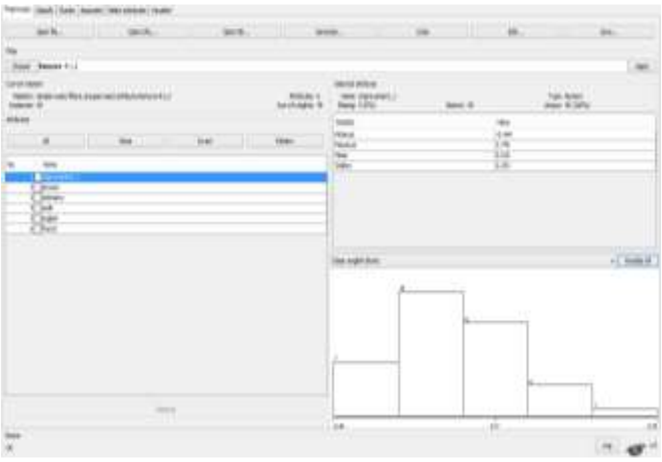
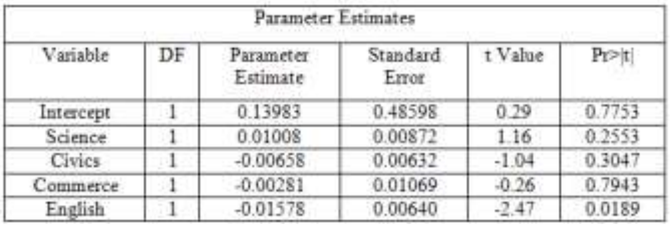


Table 2. Output dataset.



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.13983	0.48598	0.29	0.7753
Science	1	0.01008	0.00872	1.16	0.2553
Civics	1	-0.00658	0.00632	-1.04	0.3047
Commerce	1	-0.00281	0.01069	-0.26	0.7943
English	1	-0.01578	0.00640	-2.47	0.0189

4. TRIPLET EXTRACTION

Stanford CoreNLP provides a set of natural language analysis tools which can take raw text input and give the base forms of words, their parts of speech, whether they are names of companies, people, etc [6][7]., normalize dates, times, and numeric quantities, and mark up the structure of sentences in terms of phrases and word dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, etc.

POSTaggerAnnotator class generates parts of speech annotation. Labels tokens with their POS tag.

D. Input Files

1. Total significance of results is 3.
2. R-Square represents fitness of the model.
3. The model is fitted by 27%.
4. English has the major affect on the Result.
5. The parameter estimate of Model is negative.

Loading POS Tagger model ... done (Total significance of results is 3. 2.553s)

```
Output:Total_JJ significance_NN of_IN results_NNS is_VBZ 3._VBG
Array list s3 is:[Total_JJ, significance_NN, of_IN, results_NNS]
Array list s4 is:[is_VBZ, 3._VBG]
Subject'0'=========significance
Subject'1'=========results
Predicate ===========[3.]
object is ========= []
```

```
R-Square represents fitness of the model
output:R-Square_DT represents_VBZ fitness_NN of_IN the_DT model_NN
Array list s3 is:[R-Square_DT]
Array list s4 is:[represents_VBZ, fitness_NN, of_IN, the_DT, model_NN]
Predicate ===========[represents]
object is ========= [model]
```

```
The model is fitted by 27%
output:The_DT model_NN is_VBZ fitted_VBN by_IN 27%_CD
Array list s3 is:[The_DT, model_NN]
Array list s4 is:[is_VBZ, fitted_VBN, by_IN, 27%_CD]
Subject'0'=========model
Predicate ===========[fitted]
object is ========= []
English has the major affect on the Result
output:English_NNP has_VBZ the_DT major_JJ affect_VBP on_IN the_DT Result_NN
Array list s3 is:[English_NNP]
Array list s4 is:[has_VBZ, the_DT, major_JJ, affect_VBP, on_IN, the_DT, Result_NN]
Subject'0'=========English
Predicate ===========[affect]
object is ========= [Result]
```

```
 The parametr estimate of model is negative
output:The_DT parametr_NN estimate_NN of_IN model_NN is_VBZ negative_JJ
Array list s3 is: [The_DT, parametr_NN, estimate_NN, of_IN, model_NN]
Array list s4 is: [is_VBZ, negative_JJ]
Subject'0'=========parametr
Subject'1'=========estimate
Subject'2'=========model
Predicate ===========[is]
object is ========= []
```

II. RESOURCE DESCRIPTION FORMAT

After Apache Jena API uses Java system for RDF providing support for manipulating RDF models, parsing RDF/XML.

Generated RDF for Triples

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:SASResult="http://sasresults.edu#" >
  <rdf:Description rdf:about="http://sasresults.edu#results">
   <SASResult:Subject>'results'</SASResult:Subject>
   <SASResult:Predicate>'Total'</SASResult:Predicate>
   <SASResult:Object>'null'</SASResult:Object>
</rdf:Description>
  <rdf:Description rdf:about="http://sasresults.edu#model">
   <SASResult:Subject>'model'</SASResult:Subject>
   <SASResult:Predicate>'null'</SASResult:Predicate>
   <SASResult:Object>'null'</SASResult:Object>
</rdf:Description>
  <rdf:Description rdf:about="http://sasresults.edu#model">
   <SASResult:Subject>'model'</SASResult:Subject>
   <SASResult:Predicate>'null'</SASResult:Predicate>
   <SASResult:Object>'fitted'</SASResult:Object>
</rdf:Description>
  <rdf:Description rdf:about="http://sasresults.edu#Result">
   <SASResult:Subject>'Result'</SASResult:Subject>
   <SASResult:Predicate>'major'</SASResult:Predicate>
   <SASResult:Object>'null'</SASResult:Object>
  </rdf:Description>
    <rdf:Description rdf:about="http://sasresults.edu#eng">
   <SASResult:Subject>'eng'</SASResult:Subject>
   <SASResult:Predicate>'negative'</SASResult:Predicate>
   <SASResult:Object>'null'</SASResult:Object>
  </rdf:Description>
</rdf:RDF>
```

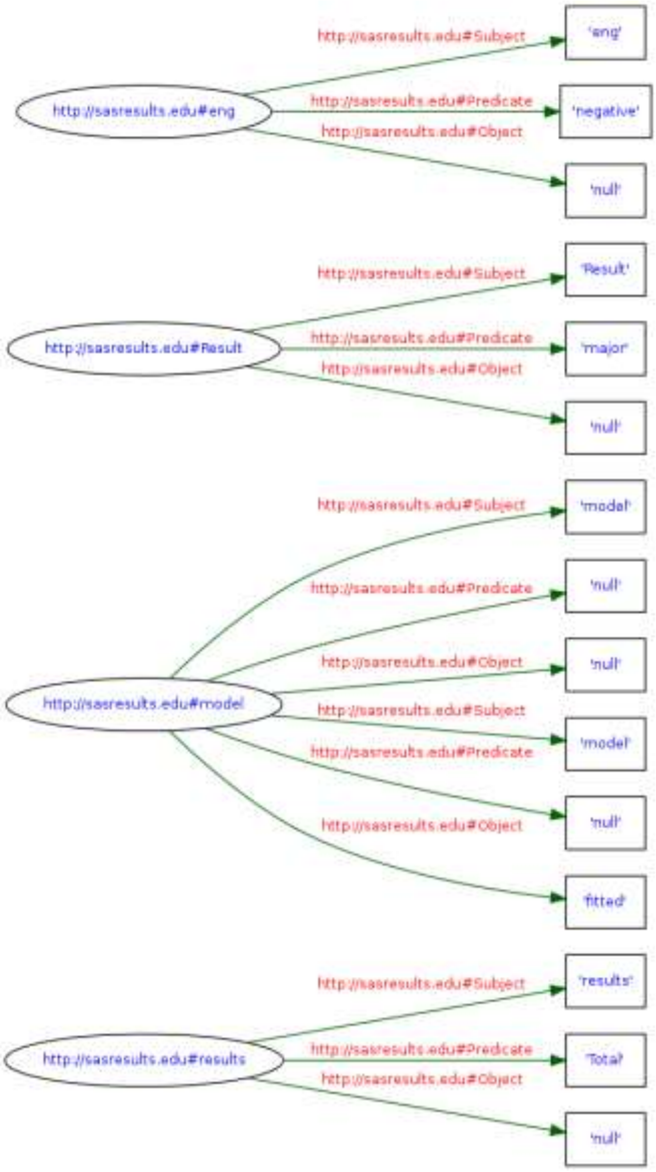

http://sasresults.edu#eng
http://sasresults.edu#Subject
'eng'
http://sasresults.edu#Predicate
'negative'
http://sasresults.edu#Object
'null'
http://sasresults.edu#Result
http://sasresults.edu#Subject
'Result'
http://sasresults.edu#Predicate
'major'
http://sasresults.edu#Object
'null'
http://sasresults.edu#model
http://sasresults.edu#Subject
'modal'
http://sasresults.edu#Predicate
'null'
http://sasresults.edu#Object
'null'
http://sasresults.edu#Subject
'model'
http://sasresults.edu#Predicate
'null'
http://sasresults.edu#Object
'fitted'
http://sasresults.edu#results
http://sasresults.edu#Subject
'results'
http://sasresults.edu#Predicate
'Total'
http://sasresults.edu#Object
'null'

Fig. 3. RDF graph model .

5. CONCLUSION

In this paper, we have proposed a new way to generate and present knowledge from large amounts of potentially heterogeneous and distributed data set. Resulted RDF aims at describing and formalizing entities from the domain of data mining and knowledge discovery. This system will help to build expert automated decision support system based on the data mining results.

6. REFRENCES

[1] Li Ding, Pranam Kolari, Zhongli Ding, and Sasikanth Avancha, Using Ontologies in the Semantic Web: A Survey, book-chapter in Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems, 2006.

[2] M. Taye , "Understanding Semantic Web and Ontologies: Theory and Applications", J. of Computing, Vol. 2(6), June 2010, NY, USA, ISSN 2151-9617.

[3] R. Jasper and M. Uschold. A framework for understanding and classifying ontology applications. In Proceedings of the IJCAI99 Workshop on Ontologies and Problem-Solving Methods(KRR5), 1999.

[4] Mizoguchi, R.: Tutorial on ontological engineering - part 3: Advanced course of ontological engineering. New Generation Comput 22(2) (2004)

[5] Smith, B.: Ontology. In: Blackwell Guide to the Philosophy of Computing and Information, pp. 155–166. Oxford Blackwell, Malden (2003).

[6] Gruber, T.R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5, 199-220.

[7] Ganter, B.; Stumme, G.; Wille, R. (Eds.) (2005). Formal Concept Analysis: Foundations and Applications. Lecture Notes in Artificial Intelligence, no.3626, Springer-Verlag. ISBN 3-540-27891-5.

[8] Anuj Tiwari, Dr. Kamal Jain (2014), “Ontology Driven Architecture for Web GIS”, India Geospatial Forum 2014, February 5–7, 2014, 60, Hyderabad, India.