© 2016 IJSRSET | Volume 2 | Issue 6 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Design of Ontology Learning Model Based on Text Classification for Domain Concept Taxonomy

Monica Sankat [*1], R. S. Thakur[2], Shailesh Jaloree [3]

[1]Department of Applied Mathematics and Computer Application, SATI, Vidisha, Madhya Pradesh, India
[2]Department of Computer Application, MANIT, Bhopal, Madhya Pradesh, India
[3]Department of Applied Mathematics and Computer Application, SATI, Vidisha, Madhya Pradesh, India

ABSTRACT

In this paper we take the approach that constructed the Domain Concept Taxonomy which attempted to take a method that extremely beneficial for the knowledge acquisition task. This work is the integration of knowledge acquisition with machine learning techniques to increase the ontology creation effect, including taxonomy relation Generation, non-taxonomy relation Generation. In this work, the related techniques of machine learning and statistical natural language processing attempt to construct the Domain Concept Taxonomy.

Keywords : Ontology, Domain Concept Taxonomy, WEKA, Matthews Correlation Coefficient, Precision Recall Curve, Receiver Operating Characteristics, SVM

I. INTRODUCTION

According to Gruber [1], "Ontologies are considered as taxonomic hierarchies of classes", it can be said that the key module in ontology is the taxonomy. Such taxonomies, as the main component of ontology provide an organizational model for a domain. Learning taxonomy is a complex task. For building taxonomy, the identification of hyperonomy/hyponomy relations between terms (terms) is compulsory. Hyponymy can be defined as: an expression A is a hyponym of a B if the meaning of B is part of the meaning of A and A is a subordinate of B. By contrast, B is a hyperonym of A if B includes the meaning of A and B is a superior to A. For instance, Mercury, Jupiter, and Mars are hyponyms of Planets, in the contrast Planet are hyperonym of Mercury, Jupiter, and Mars. Other names for the hyponym relationship are is-a, parent-child, or broader-narrower relationships [2].A supervised method to find hypernyms relations between the terms into a knowledge domain is proposed. This is, given a corpus of text and the group of related terms (called concepts), a combination of lexical patterns with supervised information and context information is applied. This paper covers the major aspects of taxonomy relation Generation, non-taxonomy relation Generation.

II. METHODS AND MATERIAL

A. Related Work

Boyce [15] presented a method for domain experts to develop ontologies for use in the delivery of courseware content. They focused in particular on relationship types that allow us to represent rich domains sufficiently.

Fortuna [16] proposed a semi-automatic and data-driven ontology editor called OntoGen, focusing on editing of topic ontologies .The system combines text data mining techniques with an efficient user interface to decrease the time spent and complexity.

Fortuna [17] presents a new version of OntoGen system. The system integrates machine learning and text data mining algorithms into an efficient user interface making ease of use for users who are not ontology engineers.

Mei-ying Jia et al. [18] has proposed automated ontology construction method. The method is not pure auto-mated. It uses existing thesaurus and database of Military Intelligence. The thesaurus provides classes information for the ontology and the database provides

the instances. Here, only three types of relationships are used between concepts of constructed ontology.

Bhowmick [19] present a framework for manual ontology engineering in education domain for managing learning content of the syllabus related requirements of school students. In this paper, a multilingual framework for management of knowledge structures of such domains.

To reduce the effort of manual ontology building, Choudhary propose a methodology for building ontology in semi-automatic manner. In his paper algorithms are developed for automatic discovery of concepts from Web for building domain ontology. Relationships among the concepts are assigned in semi-automated manner [20].

Navigli [21] in his paper presented a methodology for automatic ontology enrichment and document explanation with concepts and relations of an existing ontology. They defined Natural language definitions from available taxonomies in a given domain are processed. These regular expressions are useful to identify general-purpose and domain-specific relations.

B. State of the Art

Ontology learning systems have different purposes; they mainly extract concepts and relationships from a collection of documents related to specific domain in order to construct ontology. There are different methods applied to learn certain ontology primitives according to the follow tasks:

- Extracting the relevant domain terminology and synonyms from a text collection.
- Discovering concepts which can be regarded as abstractions of human thought.
- Deriving a concept hierarchy organizing these concepts.
- Extending an existing concept hierarchy with new concepts.
- Learning non-taxonomic relations between concepts.
- Populating the ontology with instances of relations and concepts.
- Discovering other axiomatic relationships or rules involving concepts and relations.

C. Proposed Methodology for Text Classification using Supervised Learning

The goal of our system is to build a compact model of the free Text so that new unlabeled concepts can be reliably categorized. We induce a classification model from a training collection that includes a mix of labelled Classes from various categories as shown in Figure 1.

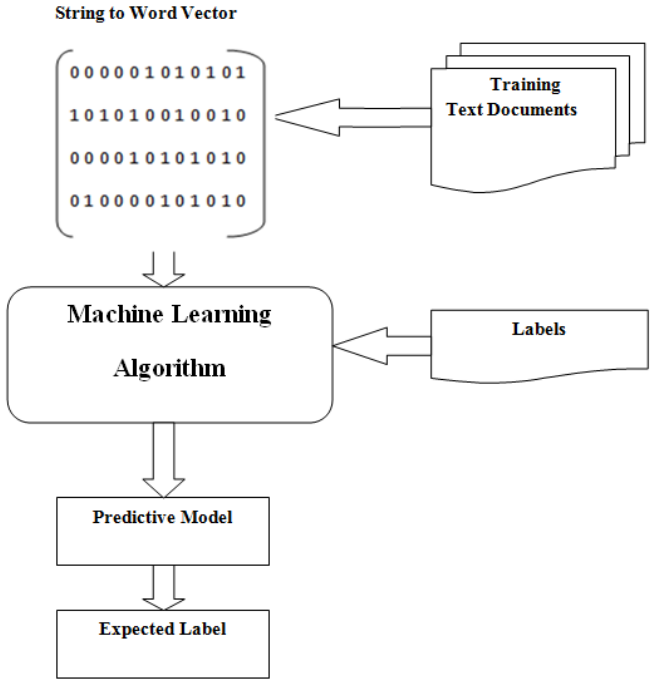

Figure 1. Classification model for Training Collection

There are many techniques which are used for text classification. Following are some techniques:

- Decision Tree Induction
- Bayesian Classification
- Support Vector Machine
- Classification using Neural Network

1) Decision Tree Induction

Decision trees [3] [4] are the most extensively used inductive learning methods. Decision tree classification is the learning of decision trees from labelled training documents. ID3 is one of the most well known decision tree algorithms and its successor C4.5.A decision tree is a flowchart like tree structures, where each internal node denotes a test on document, each branch denotes a result of the test, and each leaf node holds a class label. It is a top-down method which iteratively construct decision tree classifier.

2) Bayesian Classification

The Naïve Bayes Classifier [5] [7] [8] is the straightforward probabilistic classifier used to classify the text documents. It rigorous assumption that each feature word is independent of other feature words in a document. The basic scheme is to use the shared probabilities of words and categories to estimate the class of a given document. Given a document, the probability with mi each class is calculated as rj.

$$P(rj \mid mi) = \{P(mi \mid rj).P(rj)\}/ P(mi)$$

As P (mi) is the same for all class .then; label (mi) the class (or label) of mi can be determined by

$$Label(mi) = \arg Max\ rj\ \{P(rj \mid mi)\} = \arg Max\ rj\ \{P(mi \mid rj).\ P(rj)\}$$

3) Support Vector Machine

The Support Vector Machine (SVM) [5] [9] [10] technique is a popular and highly accurate machine learning method for classification problems. SVM try to find an optimal hyperplane within the input space so as to correctly classify the multi-class classification problem. For linearly separable space, the hyperplane is written as

$$\mathbf{v.X+a=0}$$

Here X is an arbitrary object to be classified; the v vector and constant a are learned from a training set of linearly separable objects.

4) Text Classification using Neural Network

Neural networks [11] [12] [13] [14] have emerged as a significant tool for classification. Neural networks are data driven self-characterize methods in that they can adjust themselves to the data without any explicit specification of functional form for the primary model. For classifying a given test document, its term weights are loaded into the input units; the activation of these units is generated forward through the network, and the value of the output unit(s) determines the categorization decision(s).

III. RESULTS AND DISCUSSION

A Classification techniques are implemented over dataset to analyze the Classification Barkatullah University of data set on WEKA (THE Waikato Environment For Knowledge Analysis) open source software which consists of a collection of machine learning algorithms .For classification data are taken from Barkatullah University, the experiments are implemented on Intel(R) Core (TM) 2 Duo computing machine, with CPU 2.20 GHZ and 3GB RAM.

Table 1 shows the confusion matrix of different classifiers implemented for the analysis purpose. Table 2 presents the various performance measures like True positive(TP) which means the number of positive examples that are correctly predicted as positive, False positive(FP) which means the number of positive examples that are actually negative ,Precision is the fraction of those positive predicted are actually positive, Recall is known as true positive rate, F-Measure is the harmonic mean of Precision and Recall, there is always trade-off between recall and precision to best judge the accuracy, Matthews Correlation Coefficient(MCC) is a correlation, Receiver Operating Characteristics(ROC) area and Precision Recall Curve(PRC) area.

Error Measure like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are calculated in Table 3.

Table 1. Confusion Matrix for Classifiers

Confusion Matrix for SVM		
a	b	Classified as
248	2	a=Academic
0	50	b= Administrative

Confusion Matrix for NB		
a	b	Classified as
248	2	a=Academic
10	40	b= Administrative

Confusion Matrix for J48		
a	b	Classified as
248	2	a=Academic
28	22	b= Administrative

Confusion Matrix for NN		
a	b	Classified as
248	2	a=Academic
0	50	b= Administrative

Table 2. Evaluation of Classifier with different Measures

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
SVM	0.992	0.000	1.000	0.992	0.996	0.977	0.996	0.999	Academic
	1.000	0.008	0.962	1.000	0.980	0.977	0.996	0.962	Administrative
NB	0.992	0.200	0.961	0.992	0.976	0.851	0.996	0.999	Academic
	0.800	0.008	0.952	0.800	0.870	0.851	0.996	0.973	Administrative
J48	0.992	0.560	0.899	0.992	0.943	0.593	0.718	0.899	Academic
	0.440	0.008	0.917	0.440	0.595	0.593	0.718	0.530	Administrative
NN	0.992	0.000	1.000	0.992	0.996	0.977	1.000	1.000	Academic
	1.000	0.008	0.962	1.000	0.980	0.977	1.000	1.000	Administrative

Table 3. Different Error Measure for Classifiers

Classifier	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
SVM	0.0067	0.0816
NB	0.0976	0.184
J48	0.716	0.2982
NN	0.0047	0.0449

IV. CONCLUSION

Text Classification classifies tons of text document manually is an expensive and time-consuming task. Text classifier is constructed using pre classified sample documents whose accuracy and time efficiency is much better than manual text classification. If the input to the classifier is having less noisy data, we obtain efficient results. Once patterns are identified we can classify given text or documents efficiently. Almost all the known techniques for classification such as decision trees, rules, Bayes methods, nearest neighbour classifiers, SVM classifiers, and neural networks have been extended to the case of text data.

V. REFERENCES

[1]. Gruber, T.: A translation approach to portable ontology specifications.Knowledge acquisition 5(2) (1993) 199-220

[2]. Cederberg, S., Widdows, D.: Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction.In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, Association for Computational Linguistics (2003) 118

[3]. Jiawei Han, Michelin Kamber, 2001, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, USA, 70-181.

[4]. Megha Gupta, Naveen Aggrawal, 19-20 March 2010, "Classification Techniques Analysis", NCCI 2010 -National Conference on Computational Instrumentation CSIO Chandigarh, INDIA, pp. 128-131.

[5]. Christoph Goller, Joachim Löning, Thilo Will and Werner Wolff, 2009, "Automatic Document Classification: A thorough Evaluation of various Methods", "doi=10.1.1.90.966".

[6]. Vishal Gupta, Gurpreet S. Lehal, August 2009 "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1.

[7]. B S Harish, D S Guru and S Manjunath, 2010, "Representation and Classification of Text Documents: A Brief Review", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" RTIPPR.

[8]. Jingnian Chen, Houkuan Huang, Shengfeng Tian and Youli Qu, 2009, "Feature selection for text

classification with Naïve Bayes", Expert Systems with Applications: An International Journal, Volume 36 Issue 3, Elsevier.

[9]. Wen Zhang, Taketoshi Yoshida and Xijin Tang, 2008, "Text classification based on multi-word with support vector machine", Journal: Knowledge Based Systems - KBS , vol. 21, no. 8, pp. 879-886, doi: 10.1016/j.knosys.2008.03.044, Elsevier.

[10]. Steve R. Gunn, 1998, "Support Vector Machines for Classification and Regression", University of Southampton.

[11]. Guoqiang Peter Zhang, November 2000, "Neural Networks for Classification: A Survey", IEEE Transactions on systems, man and cybernetics-Part C, Applications and Reviews, Vol. 30, NO. 4.

[12]. Larry Manevitz, Malik Yousef, 2007, "One-class document classification via Neural Networks", Neurocomputing 70, 1466–1481, Elsevier.

[13]. David Faraggi, Richard Simon, 1995, "The maximum likelihood neural network as a statistical classification model", Journal of Statistical Planning and Inference 46, 93-104, Elsevier.

[14]. Ali Selamat, Sigeru Omatu, 2004, "Web page feature selection and classification using neural networks", Information Sciences 158, 69–88, Elsevier.

[15]. Boyce, S., & Pahl, C. (2007). Developing Domain Ontologies for Course Content. Educational Technology & Society, 10 (3),275-288.

[16]. Fortuna Blaz, Grobelnik Marko & Mladenic Dunja(2006): Semi-automatic Data-driven Ontology Construction System. In: Proceedings of the 9th International multi-conference Information Society IS-2006, Ljubljana, Slovenia (2006).

[17]. Fortuna Blaz, Grobelnik Marko & Mladenic Dunja(2007), OntoGen: Semi-automatic Ontology Editor: Human Interface, Part II, HCII 2007, LNCS 4558, pp. 309–318, 2007.

[18]. Jia, M., Yang, B., Zheng D., Sun, W., Liu, Li., Yang, Jing.,(2009) "Automatic Ontology Construction Approaches and Its Application on Military Intelligence", Asia-Pacific Conference on InformationProcessing (APCIP), vol. 2, Pp. 348 – 351, 2009.

[19]. Bhowmick P.K., Roy D., Sarkar S. & Basu A.(2010), A Framework For Manual Ontology Engineering For Management Of Learning Material Repository,2010.

[20]. Choudhary Jaytrilok. & Roy Devshri.(2012) ,An Approach to Build Ontology in Semi-Automated way, Journal Of Information And Communication Technologies, Volume 2, Issue 5, May 2012.

[21]. Navigli, R., Gangemi, A., Velardi, P.(2003). Ontology learning and its application,2003