A Research-oriented Survey and Current Status on Feature Extraction, Ontology Construction towards Natural Language Processing

Dhanasekaran.K[1] and Rajeswari.R[2]

[1]Dept.of Computer Science and Engineering, Info Institute of Engineering, Anna University of Technology Coimbatore, Tamilnadu, India
[2]Dept.of Electrical and Electronics Engineering, Govt.College of Technology, Anna University of Technology Coimbatore, Tamilnadu, India

Abstract

Due to the increasing amount of information in the World Wide Web, there is a need to develop an efficient approach which can process wide collection of data and convert those data into meaningful knowledge to the user. With this in mind, the researchers have focused on improving the feature extraction method in natural language texts that are available in the Web in various forms such as news, articles, journals, design documents, etc.Over the past few years, the problem has been identified from many areas, in many domains and is solved giving some satisfactory result, which gives us future direction for our research. In this paper, we review some existing approaches, issues and advantages of it to support the future research in the related areas of text mining, information retrieval, semantic web, web mining and natural language processing.

Keywords:*Knowledge Extraction,Natural Language, OntologyLearning,GeneClassification,Term Extraction*

1. Introduction

Feature Extraction method has been a focus for the research community over a decade irrespective of disciplines. The method plays a vital role in engineering, education, research, Internet, finance, multimedia, banking, fraud detection, and other domains as well.

It is observed that the Feature Extraction is a special form of dimensionality reduction. If input data is too large, then it may have redundant data so that the input data should be transformed into a reduced representation, say, a set of features called feature vector [1].

The feature set will be formed by extracting the relevant information from the input data. When large numbers of variables are used, then it requires a large amount of memory and time or a classification algorithm which overfits the training sample [2].In computer vision, the extraction of visual features consists of mathematical transformations computed on the pixels of a digital image. The visual features are used in the applications such as object detection or image search by content. The other key areas of application of feature extraction are: Natural Language Processing [4], audio processing.

The Feature Extraction method plays a very important role in Web Mining as well. The definition of web mining is "A methodology of information retrieval tools using data mining to extract information about both the content of the pages, its structure of relations and the navigation log of users.

Related to Web Mining, the feature extraction can be applied on the following fields:

1. Web Content Mining
2. Web Structure Mining
3. Web Usage Mining

In Text Mining, the feature extraction method can be applied with respect to various kinds of texts. Some methods and proposals can be considered as text types that are mainly aimed at grouping and classifying the text language based on common characteristics [2][3].

The issue of text types is approached from disciplines and learning as discourse analysis and text linguistic. For the text with common characteristics, it is difficult to establish the systematic classification, because of the diversity of texts and their variable length. The conventional classification of texts includes scientific texts, Legal texts, Journalistic texts, and Humanistic texts, literary texts, advertising copy and digital texts.

The source texts for feature extraction can also be classified according to their role in communication. They are:

a. Informational text: It reports something to the user.
b. Text management: It encourages the listener to take action.

c. Expressive texts: It reveals the subjectivity of the speaker.

The text sequences are of four categories:

Narrative: It describes an account of events developed in a certain place and it will be carried out in some order.

Descriptive texts: It has the characteristics of an object, without lapse of time. Some examples are: Scientific texts, technical texts and social texts.Similarly, other types of texts such as **Argumentative texts, Informational texts** such as educational texts, specialized texts or arguments can also be used as source for presenting some useful knowledge to the user with the help of an efficient querying technique and some text processing technique.

In recent years, an interesting research is started in the field of machine translation and language processing. The difficulties in natural language processing are analyzed at the various levels to propose some new idea for solving the problems. The levels are as follows:

a. Standard lexicon level
b.Reference level
c.Structural and semantic level
d.Pragmatic level

The feature extraction method and text classification approaches can be used to solve problems at the various levels and is possible to derive rules for knowledge representation from the syntactic structure in order to ensure that the interesting knowledge is presented to the user. In this survey, some existing techniques related to feature extraction, Ontology construction in NLP and text classification are to be discussed giving some future direction.

\

2. A Review Of Existing Approaches And Current Status

Michele Carenini et.al. have published their paper on "Improving Communication in E-democracy Using Natural Language Processing". Their work focused on improving communication between public administrators and their citizens [4]. This supports friendly and intuitive information access to the citizens.

These authors observed that no efficient scheme exist to support real interaction with real users on real repositories of information.NLP can be a powerful technique to access structured and nonstructured information and to improve human-computer interaction so that authors aimed for discovering a NLP approach which improves e-democracy by increasing the citizens' participation in the decision-making process.

In this paper the goal was to test satisfiability of the requirements and to test robustness and the usability of the augmented phrase structure grammars in a highly sensitive environment. The authors also aimed at developing two toolsets to improve the communication of users in the context of urban planning.

An important assumption of this approach is that the language fragment shows some (minimum) criteria of grammaticality. Many citizens don't care about syntactic mistakes so that NLP approach is used to analyze a large corpus of the language and enable the system to handle linguistic objects that fall outside the field of interest.

If we loosen certain rules of the grammars it may improve the performance for a system that deals with analyzing meaningful fragments in ungrammatical sentences.

Some problems require skilled technical people who adopts on the system architecture but it is a difficult task so that the system may be designed with efficient NLP technique based on Dirty NLP language that can manage an imperfect sentences.

In the same period Topon Kumar Paul and Hitoshi Iba published paper on "Prediction of Cancer Class with Majority Voting GP classifier Using Gene Expression Data". Most of the existing methods are overfitting due to a very small number of training samples compared to the huge number of genes and class imbalance. So they developed a Majority Voting Genetic Programming Classifier (MVGPC) for the classification of microarray data [2].

Instead of single rule or a single set of rules they used multiple rules with GP and then applied those rules to test samples to determine their labels with majority voting technique. Their experiments on four different public cancer data sets, including multiclass data sets showed that the test accuracies of MVGPC are better than those of other methods.

It is observed that individually those rules or sets of rules classify the test samples very poorly but as a group of rules they classify the samples very accurately. Moreover they seen that scaling didn't affect the accuracy and the system used some of the more frequently occurred genes in the evolved rules of MVGPC as the potential biomarkers of the types of cancers.

In future work they suggested that some issues such as., finding the existence of the quantitative relationship among the more frequently selected genes of MVGPC may be resolved and its performance on other multiclass data sets may be tested.

Wang Wei,Payam Barnaghi proposed "Probabilistic Topic Models for Learning Terminological Ontologies" and introduced the idea of developing models and utilizing it for document modeling and topic extraction in Information Retrieval[5]. The new approach focused on automatic learning of terminological ontologies from text corpus.

In that model the topic models were used as efficient dimension reduction techniques, which can capture semantic relationships between word-topic and topic-document interpreting in terms of probability distributions.

They have proposed two algorithms for learning terminological ontologies using the principle of topic relationship and exploiting information theory with the learned models.

While experiment the learned ontology statements were evaluated by the domain experts. When they compared their method with two existing concept hierarchy learning methods on the same data set, their method outperformed in terms of recall and precision measures.

For browsing, navigation, and information search and retrieval in digital libraries the precision level of the learned ontology was sufficient for deployment.

An automatic ontology learning approach cannot be used for direct formal reasoning but they can be used in applications where a certain error rate is tolerable, such as information retrieval, browsing, and navigation. Moreover, such ontologies can significantly reduce time and effort in the ontology engineering process involving some experts' judgment.

These authors have used two probabilistic models of the PLSA and LDA and their new method learns terminological ontology with respect to the SKOS model from text corpus.

The contributions includes 1) use of topic models for the purposes of ontology learning.2) using Kullback-leiber divergence as a probabilistic proxy for learning ontological relationships. They have proposed "Information Theory Principle for Concept Relationship which formally defines how to establish "broader" relationship between two topics.

The improvements have shown by augmenting the "broader" relationship with learning of "related" relationship. Developing two algorithms (local similarity hierarchy learning and global similarity hierarchy learning) authors have shown that ontologies learned using LDA are superior to the PLSA based method in terms of the recall,precision,and F1 measures.

This research was domain independent. As they described, in future, the approach can be generalized to learn simultaneously several subtrees in the topic hierarchy and the performance may be tested in other data sets because they used only computer science publication data set.

Yanhong Zhai and Bing Liu, have published a paper on "Structured Data Extraction from the Web Based on Partial Tree Alignment", which studies the problem of structured data extraction from arbitrary Web pages [6]. The objective of their proposed research is to automatically segment data records in a page, extract data items/fields from these records, and store the extracted data in a database.

Existing methods addressing the problem can be classified into three categories. The first category provides some languages to facilitate the construction of data extraction systems. The second category use machine learning techniques to learn wrappers (which are data extraction programs) from human labeled examples. Because manual labeling is time-consuming and is hard to scale to a large number of sites on the Web, methods in the third category are based on the idea of automatic pattern discovery.

However, multiple pages that conform to a common schema are usually needed as the input. In this paper, they propose a novel and effective technique (called DEPTA) to perform the task of Web data extraction automatically.

The method consists of two steps: 1) identifying individual records in a page and 2) aligning and extracting data items from the identified records.

In step 1, a method based on visual information and tree matching is used to segment data records. In step 2, a novel partial alignment technique is proposed.

This method aligns only those data items in a pair of records that can be aligned with certainty, making no commitment on the rest of the items. Experimental results obtained using a large number of Web pages from diverse domains show that the proposed two-step technique is highly effective.

In this work, the authors proposed a new approach to extract structured data from Web pages. Although the problem has been studied by several researchers, existing techniques either are inaccurate or make several assumptions. This method does not make these assumptions.

This only requires that the page contains more than one data record. This technique consists of two steps: 1) identifying data records without extracting data items in the records and 2) aligning corresponding data items from multiple data records and putting the data items in a database.

They proposed an enhanced method based on visual cues for step 1. In step 2, they proposed a novel partial tree alignment technique to align corresponding data fields of multiple data records. Empirical results using a large number of Web pages demonstrated the effectiveness of the proposed technique.

2.1 Experimental Results

Table 1 shows the results of DEPTA on data record extraction and data item alignment.

The pages are divided into five categories according to experimental results and are listed in the first column. For the 108 pages in Category 1, both MDR and DEPTA identify all the data records correctly. DEPTA also aligns all the data items correctly. For the 37 pages in Category

2, DEPTA identifies all the data records whereas MDR does not.

TABLE 1Experimental Results

Cate-gory	No. of Pages	Data Record Extraction							Data Item Alignment			
			MDR			DEPTA (MDR-2)			DEPTA			
		ACT	COR	WRG	MISS	COR	WRG	MISS	ACT	COR	WRG	MISS
1	108	1447	1447	0	0	1447	0	0	9498	9498	0	0
2	37	534	337	4	193	534	0	0	3380	3380	0	0
3	12	210	210	0	0	210	0	0	1416	1362	54	0
4	8	112	68	2	42	112	0	0	1344	1328	16	0
5	35	659	522	142	35	546	142	35	3372	3185	1278	187
Total	**200**	**2962**	**2584**	**148**	**270**	**2849**	**142**	**35**	**19019**	**18753**	**1348**	**187**
Recall			87.24%			96.18%			98.60%			
Precision			94.55%			95.25%			93.29%			

Data items are aligned correctly by DEPTA.For the 12 pages in Category 3, both MDR and DEPTA identify all the data records correctly, but a small number of items are incorrectly aligned by DEPTA. For the eight pages in Category 4, DEPTA identifies all the data records whereas MDR does not. Some data items are aligned incorrectly.

For the 35 pages in Category 5, both MDR and DEPTA miss some data records or identify incorrect data records, and, consequently, their data items are either not extracted or extracted wrongly. The two columns marked with “ACT” show the number of actual data records and the number of actual data items in each Web page, respectively.

For comparison purposes, these numbers are manually counted. Note that they only count the data records which contain valued information. Some areas such as navigation bars in a page also contain data with regular patterns and they are also identified by the proposed technique. DEPTA takes into consideration the visual information of the identified data records to decide whether to output them or not. The criteria used to decide the importance of a list of data records is similar to the ones used in [30].

The three columns marked with “COR” shows the numbers of correct data records extracted by MDR and DEPTA and the number of data items correctly aligned and extracted by DEPTA, respectively. The meaning of a data record is clear. The meaning of a data item needs some explanation.

In this work, they assume that data items are segmented by HTML elements. Thus, a data item is correctly aligned and extracted if the text fragment is enclosed in an HTML element and aligned correctly with the same type of information in other data records. Note that the data items may not be values of attributes as we recognize semantically.

For example, for Price: $20, the extraction of “Price: $20” as one data item is considered correct if the alignment is also correct. Their system does not further split “Price” and “$20” into two items and assign “$20” as the value of the attribute “Price.”An another example, for Price: $20, “Price” and “$20” are treated as two separate data items.

They have not studied text segmentation and data labeling in this work. They have planned to investigate these issues in the future.

A correct alignment also needs some explanation. Basically, we take the best result in the alignment of a set of data items. For example, there are five items representing the same type of information from five different data records, i.e., they should be aligned together.

If all the five items are aligned (in one alignment), then they are all corrected. However, if they are not aligned together, errors are produced. For instance, two alignments (instead of one) are produced, one with three items and the other with two items, then the three items are aligned correctly, and the two items are not.

The three columns marked by “WRG” show the numbers of wrong data records extracted by MDR and DEPTA and the number of data items incorrectly aligned by DEPTA, respectively.

For a data record, a wrong extraction means that only part of the content of the data record is extracted, or information outside of the data record boundary is extracted and enclosed in it. For a data item, incorrect alignment usually means items of the same attribute are placed into different columns, or items of different attributes are placed into the same column.

The three columns marked by “MISS” show the numbers of data records and data items that were not identified or extracted by MDR and DEPTA, respectively.

The last three rows of Table 1 give the total of each column, the recall and precision of each system. For data record extraction, the recall and precision are computed based on the total number of correct data records found in all pages and the actual number of data records in these pages. For data item alignment, the precision and recall computation has considered all wrongly extracted or missing data records introduced in step 1 of DEPTA.

The conflict resolution method helps to improve the alignment of some data items in 30 pages out of 34 pages which have ambiguities in alignments.

Also, there are about 7 percent of the sites that contain noncontiguous data records. The comparison of data item alignment of DEPTA (the second step) and RoadRunner is shown in Table 2. For an accurate comparison, they recommended not to consider the erroneous/ missing records introduced by the first step of DEPTA (MDR-2). That is, we only use the data records that are correctly identified by MDR-2 in the comparison.

Out of the 200 pages, there are 174 pages from which all data items are aligned correctly by DEPTA and 26 pages from which 92 percent data items are aligned correctly. In comparison, there are only 110 pages from which all data items are aligned correctly by Road-Runner, 28 pages from which only 58 percent data items are aligned correctly, and 62 pages from which data items cannot be aligned by RoadRunner, which simply returns each data record as an item. There is no error reported by the system. It also ran the Tidy program

TABLE 2 Comparison of Data Item Alignment of DEPTA and RoadRunner

	ACT	ALG	COR	Precision	Recall
DEPTA	18290	18290	18203	93.29%	98.60%
RoadRunner	18290	13241	10476	79.12%	57.28%

The data item alignment results of the two systems are summarized in Table 2. “ALG” means aligned and “ACT” and “COR” have the same meanings as above. The results show that DEPTA outperforms RoadRunner significantly on both precision and recall. Note that the precision and recall are the same for DEPTA because all items are aligned (nothing lost).

In SEPTEMBER 2006, Rile Hu, Chengqing Zong and Bo Xu, have published paper on “An Approach to Automatic Acquisition of Translation Templates Based on Phrase Structure Extraction and Alignment”. In this paper, they proposed a new approach for automatically acquiring translation templates from unannotated bilingual spoken language corpora [8]. They adopted two basic algorithms: a grammar induction algorithm, and an alignment algorithm using bracketing transduction grammar.

The approach is unsupervised, statistical, and data-driven, and employs no parsing procedure. The acquisition procedure consists of two steps. First, semantic groups and phrase structure groups are extracted from both the source language and the target language. Second, an alignment algorithm based on bracketing transduction grammar aligns the phrase structure groups.

The aligned phrase structure groups are post-processed, yielding translation templates. Preliminary experimental results show that the algorithm is effective. Based on the grammars, the phrase structures are aligned using BTG. Finally, the aligned structures are treated as translation templates.

This method needs fewer resources than the method which is called as “parse-parse-match.”And it can get comparable results as those of the “parse-parse match” method. These results of the preliminary experiments show that this approach is viable.

However, it still faces many difficult tasks, including the improvement of grammar induction and alignment. As they said, in future, more information such as some dictionary information (including a synonym dictionary) and some additional preprocessing may be introduced.

Qinbao Song, Jingjie Ni and Guangtao Wang, have published a paper on “A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data”. Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features [7].

A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features.

Based on these criteria, a fast clustering-based feature selection algorithm, FAST, is proposed and experimentally evaluated in this paper. The FAST algorithm works in two steps.

In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features.

To ensure the efficiency of FAST, they adopt the efficient minimum-spanning tree clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study.

The results, on 35 publicly available real-world high dimensional image, microarray, and text data, demonstrate that FAST not only produces smaller subsets of features but also improves the performances of the four types of classifiers.

In their paper, they have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

They have compared the performance of the proposed algorithm with those of the five well-known feature selection algorithms FCBF, ReliefF, CFS, Consist, and FOCUS-SF on the 35 publicly available image, microarray, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record.

Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime, and the best classification accuracy for Naive Bayes, C4.5, and RIPPER, and the second best classification accuracy for

IB1. The Win/Draw/Loss records confirmed the conclusions.
They also shown that FAST obtains the rank of 1 for microarray data, the rank of 2 for text data, and the rank of 3 for image data in terms of classification accuracy of the four different types of classifiers, and CFS is a good alternative.At the same time, FCBF is a good alternative for image and text data.
Moreover, Consist and FOCUSSF are alternatives for text data. In future work, different types of correlation measures can be explored, and some formal properties of feature space may be studied.
Hong Huang and Hailiang Feng, investigated on "Gene Classification Using Parameter-free Semi-supervised Manifold Learning" which includes a new manifold learning method, called parameter-free semi-supervised local fisher discriminant analysis (*p*SELF), and is proposed to map the gene expression data into a low dimensional space for tumor classification [9]. A new difference-based optimization objective function with unlabeled samples has been designed.
The proposed method preserves the global structure of unlabeled samples in addition to separating labeled samples in different classes from each other. The semi-supervised method has an analytic form of the globally optimal solution, which can be computed efficiently by eigen decomposition.
Experimental results on synthetic data and SRBCT, DLBCL and Brain Tumor gene expression datasets demonstrate the effectiveness of the proposed method. The application of machine learning to data mining and analysis in the area of microarray analysis is rapidly gaining interest in the community.
The large number of gene expressions coupled with analysis over a period of time, provides an immense space of genomic dimensionality reduction and selection. This method exploits both statistically uncorrelated and parameter-free characteristics.*p*SELF can preserve the global structure of unlabeled samples in addition to separating labeled samples in different classes from each other, so it efficiently extracts the discriminant information in the low dimensional embedding space and addresses the semi-supervised learning problem for gene expression classification.
Experimental results on synthetic data and three well-known gene expression datasets demonstrate the effectiveness of the proposed *p*SELF algorithm. In this paper, the intrinsic structure preserved by *p*SELF is only the global structure of samples. Investigating that whether *p*SELF can preserve local structures together with unlabeled samples is an interesting future work.

In 2009 Elena.P Sapozhnikova has published a paper on "Multi-classification Approach with ART Neural Networks". They are investigated a novel method to solve a Multi-Classification (MC) task by using an Adaptive Resonance Theory (ART) neural network.
They applied a modified Fuzzy ARTMAP Algorithm called Multi-label-FAM (ML-FAM) to classify multi-label data. The instances of the tasks are labeled by multiple classes. This task is important in the field of gene function prediction or Web mining.
In early approaches, the MC task were based on learning independent binary classifiers for each class and combining outputs in order to obtain multi-label predictions. In this case, the labels in the label set are treated as independent that can significantly reduce classifier performance.
The authors have suggested that an alternate approach can directly train the data to predict a label set of an unknown size for each unseen instance.
Many real world problems produce more complex data sets. So, a particular text document could be assigned to multiple topics.Multi-classification becomes an issue when an instance belong to more than one class. The most current approaches such as., Support Vector Machine (SVM),-Nearest Neighbor (KNN) or Neural Networks can achieve high accuracy.
They focus on predictive performance rather than knowledge extraction. The standard classifiers cannot be directly applied to MC problem, because standard algorithms assume mutually exclusive class labels and standard performance measures are not suitable for evaluation of classifiers.
The authors have proposed an interpretable classifier which is belonging to fuzzy rule extraction method and it is not better due to the use of one-to-one Map field mapping in many situation.
In 2003 Xia Hong, Chris J.Harris proposed "NN Knowledge Extraction and Extended Gram-Schmidt Algorithm for Model Subspace Decomposition" which introduced a one to one mapping mechanism between a fuzzy rule base and a model matrix feature subspace using an inference mechanism. They constructed a model of a priori unknown dynamic systems using fuzzy rules.
The proposed method is used to decompose the model into submodel to avoid complexity in multiple dimensionality problems.
The drawback of the most current neurofuzzy learning algorithm is: Learning is based on a set of ID regressors or basis function; they are not based on a set of fuzzy rules for multidimensional input variables which can generate efficient model during learning stage.
The author noticed that the number of rules increased due to the increase in dimension. To solve this dimensionality problem, it is important to propose a variable reduction approach. This paper suggests us to develop a method in order to increase the model transparency.

In 2009 Heng Ji and Dekang Lin have presented a paper on "Gender and Animacy Knowledge discovery from Web scale N-grams for unsupervised person mention detection".
They have pointed out that this method can learn noun-gender and noun-animacy pair from web-scale n-grams using lexical pattern and then the confidence estimation metric is applied to filter noise.
This method is based on unsupervised learning that detects person mentions from raw texts using the selected pairs. An important resource for person mention is to discover automatically a large knowledge base of gender and animacy properties for all possible Noun-Phrase.
They raised a question for nominal mentions with more missing error than other types. The question is: Is it possible to automatically discover mentions from very large data set by effective semantic constraints? An attempt has been made by them to explore various effective confidence estimation metrics.
The traditional supervised learning methods are based on limited and static semantic resources which will not support effective identification of more rare mentions. In this paper, the future work is to develop a method for event extraction for more complicated IE tasks.
In 2011Animesh Kar, Deba Prasad Mandal have developed the methodology for "Finding opinion strength using Fuzzy Logic on Web reviews". As per their views, the review sentences are decomposed and individual characteristics of a product are evaluated. At the first step, mining is applied to extract product features that have been commented by customers.
Secondly, the function is used to identify opinion sentences in each review and also the opinion phrases are extracted in each opinion sentence.Finally, the summarized results are generated measuring the strength of opinion phrases.
In this paper, a supervised opinion orientation detection system called FOM (Fuzzy Opinion Miner) is introduced to build a model with important features by mining reviews. The authors have observed that the current opinion mining systems cannot capture the pragmatic meaning of customer evaluations.
They proposed fuzzy approximation to estimate the polarity intensity of the opinion and then quantified the sentiment into a score. This supervised approach combined existing text mining approaches with fuzzy approximation technique to mine customer reviews.
In this paper, they have stated that they are going to improve this technique by grouping features based on the usefulness of the texts. Thus there is a need to develop an efficient technique to improve the feature extraction and the subsequent summarization for processing natural language texts.

In 2002 Hoda K.Mohamed, Ain Shams U have developed a method for "Automatic Document Classification", in that, they have investigated different parameter and design decision to reduce the complexity in capturing complex semantics of natural language.
They have used Neural Network and Weighting schema, and item vectors selected using combined techniques from stemmer algorithm and NLP.The different cases are classified based on the number of inputs to the classifier, weighting schema and effect of weighting words in the title. The authors have applied modified stemmer algorithm to the whole text file and NLP indexing technique is also used.
In 2010 William L.Sousan, Qiuming Zhu, Robin Gandki, William Mahoney and Anup Sharma have explored an approach "Using Term Extraction Patterns to discover Coherent Relationships from Open Source Intelligence". They have identified approach for syntax analysis of the word sequences in unstructured text documents (Web-based text/news articles) which allows us to extract Subject-Predicate-Object triples to form the basis for Term Extraction Patterns (TEP).
This method is used to discover domain-specific multi-word entities that can be classified based on their inter-relationships. Also it allows us to extract semantics from unstructured text in domain specific open source information and this can also be used to predict Cyber attack outbreaks.
The semi-automatic methodology followed in that paper is used to extract domain relevant terms using guidance of seed terms and user feedbacks.Finally, Coherent classification relationships can be executed to assist knowledge worker.
In 2009 Amal Zouaq and Roger NKambou have published a paper on "Evaluating the generation of domain Ontologies in the Knowledge Puzzle Porject".The author described the procedure to extract concept maps from texts that are followed by TEXCOMON,Knowledge Puzzle Ontology Learning tool.In this paper, they evaluated ontology in three dimension:structural,semantic and comparative. In structural evaluations, ontology is considered as graph based on a set of metrics. Semantic evaluation is carried out using human expert judgement.Finally comparative evaluation is done by comparing the output of current tools and new tools. This task has used the same set of documents for all cases.
They compared the ontological output in terms of concepts, attributes, hierarchical and non-taxonomic relationships. The method produced more interesting concepts and relationships but failed to avoid a lot of noise generation by lexico-syntactic patterns and their methods. They suggested developing method for improving the patterns.

Moreover, the OWL Java API of their project improved in terms of processing time. This paper consists of a future direction towards automating ontology evaluations in order to solve a number of problems such as., Ontology-learning, population, mediation and matching.

In 2004 Marta Sabou has published a paper on "Extracting Ontologies from Software Documentation: A Semi-automatic method and its evaluation". In his approach, he used software APIs to build domain ontology by extracting types of method functionalities. In that method, a small corpus is used for applying statistical techniques.

The author has described that there is a need to enhance the corpus and to develop a better extraction method that suits the small corpora. This method is encouraging towards building an Ontology extraction method from software APIs.

In 2010 B.Saleena, Dr.S.K. Srivatsa have published a paper on "A Novel Approach to develop a self-organized Domain Specific Search mechanism for Knowledge Acquisition using Ontology". The authors have created a search method for semantic web in order to design a self-organized system to retrieve information about a particular topic based on user interest in learning.

For this they created a knowledge library for DBMS domain using Ontology and Knowledge management technologies. They followed a strategy to group the relevant information for the user in a single search. The search is implemented based on keyword which is a time-consuming process.

The system has been developed in JAVA 2 API with OWL API for semantic web. It retrieves the inter-related contents, prerequisites and further readings needed to understand the topic as per user's interest. With the help of this, an e-learning framework is developed using Ontology based knowledge retrieval.

This paper included future work to enhance the system for various domains and to add a large set of functionalities to the UI screen to improve the user-friendliness. Also it has been suggested to develop a method for automatic extraction of information.

In 2009 Song Jun-feng, Zhang Wei-ming, Xiao Wei-dong, Xu Zhen-ning have carried out "Study on Construction and Integration of Military Domain Ontology, Situation Ontology and Military Rule Ontology for Network Centric Warfare". In this paper, they proposed approach to construct all three kinds of ontology mentioned above. They also addressed the integration approach using all these approaches. Then they have constructed scenario based knowledge infrastructure fragment using proposed approaches.

In future study; they are going to study what kinds of other component ontologies are needed for the knowledge infrastructure. They also planned to implement experimentation and at present, they would like to use protégé basic tool.

The conceptual graph of ontology language (OML) lacks precise semantics. OWL is a new synthesis of research on ontology language. The expressiveness of all the languages are very limited, and key inference problem have most complexity.So, there is a need for optimized reasoners.

In the same year, Zhang Rui-ling, XU Hong-Sheng have published a paper on "Using Bayesian Network and Neural Network Constructing Domain Ontology". In this paper, they have addressed that the current ontology construction methods have limitations. They are: 1) Requirement for human labor 2) Domain restrictions.

To avoid these problems, they developed an approach to construct ontology based on a novel method which contains Projective Adaptive Resonance Theory (PART) neural network and Bayesian Network Probability theorem.

Their system could acquire key terms automatically. Finally it reasons out the complete terms in the classification framework in order to construct domain ontology. The ontology is stored using a Resource Description Framework (RDF).

The Semantic Web can be deployed based on the rapid and efficient construction of the ontology. Some of the features of this work are: the PART architecture is included to overcome the lack of flexibility in clustering, and in the web page analysis, WordNet deals with the lack of knowledge acquisition.

Finally they said that there is a need to improve the precision of term location. Due to the accumulation of the number of documents in the ontology repository, the similarity calculation takes more time. This is unavoidable.So, if we build an approach to form a hierarchy of clusters, it will solve the problem.

The current methods can build only a partially automated classification of terms". This involves a time-consuming process and costly procedure.

In the same year, Yi Zhang, Li Tan, Jie Liu, ChangChang Yu have published a paper on "A Domain Ontology Construction method towards Healthy Housing". Due to the presence of various domains, there is no efficient framework for ontology construction. The authors have introduced an improved ontology construction method.

In this method, they used graphic language to represent domain knowledge for research domain. The system evaluated and verified the correction of relationships and hierarchy for constructed ontology. The proposed method for Healthy Housing solves the problems such as lack of semantic expression and understanding of expert knowledge.

In future work, it is stated that the research can be done to utilize this Housing ontology in the Healthy Housing

Intelligent Synthesized Evaluation System by mapping and matching domain ontology. Then it will be possible to realize the interchange between natural language of professional field and conceptual-level ontology language which can be understood by machines.

In 2010 Zhang Dan, hang Li, Jiang Hao have published their paper "Research on Semi-automatic Domain Ontology Construction". They have applied Data Mining method and word partitioning technique to construct semi-automatic domain ontology.

In this paper, they said that the semi-automatic approach still poses a problem because of the difficulties in constructing a common tool. The reasons are: choosing data source is manual, extracting compound words without considering the characteristics of language, and analyzing the grammatical components of sentence to conclude the relations among concepts. The authors said that the methods can be tried to address these issue.

In 2011 Tauqeer Ahmad Usmani and Durgesh Pant have published a paper on "Intelligent Information Retrieval through Semantic Web Service Discovery Methods". In that paper they have stated that the Conventional Information retrieval system is lacking a uniform semantic description for information, so users cannot find more relevant information.

They said that the challenging tasks in Information retrieval system are: 1) how to make the managed resource that has a machine understandable meaning so as to find what users really want.2) how to realize the semantic searching by means of the domain knowledge.

They have mentioned that if we develop an intelligent information system, it improves the recall rate and precision rate and also allows the users to search information with natural language. In their future work they have stated that the intelligent system still poses updation problem so that an approach may be developed to overcome the shortcomings.

In 2011 Zhongwu Zhai,Bing Liu introduced "Product Feature Grouping for Opinion Mining Using Soft-Constraints and EM" which has been accepted for publication in IEEE Intelligent Systems"[3].

In opinion mining peoples can express their views for the same feature of the product by using different words and phrases.

We need to group these words and phrases under the same feature so that we can give the meaningful summary.

For this purpose these authors proposed a constrained semi-supervised learning method to solve the problem. They argued that same form of supervision is needed for the problem because its solution depends on the user application needs.

The experimental results showed promising results by using reviews from five different domains. Then their Expectation Maximization algorithm is used to solve the problem and improved the performance by considering the two soft constraints and the lexical similarity. In future these authors have planned to explore more natural language knowledge at the semantic level to improve the accuracy.

These authors have demonstrated the generality of the proposed method conducting experiments using reviews from 5 domains: *Hometheater*(HT), *Insurance*(I), *Mattress*(M), *Car*(C) and *Vacuum*(V).They obtained the data sets and the *gold standard* feature expressions and their groups from a company that provides opinion mining services. The details of the data sets and the gold standards are given in Table 1.

TABLE 3. Data sets and gold standards

	HT	**I**	**M**	**C**	**V**
#Sentences	6355	2446	12107	9731	8785
#Reviews	587	2802	933	1486	551
#Expressions	237	148	333	317	266
#Groups(*K*)	15	8	15	16	28

They carried out comparison between the results of SC-EM and the 16 baseline methods. All labeled data were selected randomly. These authors used 10%, 20%, 30%, 40%, and 50% of the feature expressions from the gold standard data as the labeled set L, and the rest as the unlabeled set U so as to see the effects of different numbers of labeled examples (seeds*)*.They reported the results by running the algorithms 30 times for each setting. They showed the detailed *purity*, *entropy* and *accuracy* results for 30% as the labeled data (70% as unlabeled).

For the other proportions of labeled data, each result is the average of the 5 data sets. For entropy, the smaller the value is, the better the result is, but for purity and accuracy, the larger the better. For these experiments, they used the window size $t = 5$. The proposed algorithm (***SC-EM***) outperforms all 16 baseline methods by a large margin on every dataset.

3. Conclusion

In this paper, various approaches, research issues, merits and demerits of techniques in the related areas of data mining, NLP, and ontology construction have been discussed. The tricks in the methodology followed by researchers and performance of methods are reviewed to support the future research. Since the semantic web has increasing collection of documents, the problems in information retrieval due to the complex syntactic and semantic structure need to be addressed to find the right solution for the problem. This paper will give the direction towards finding some efficient solution for the problems in these related areas using some machine learning techniques.

References

[1] Xin-fu LI, lei-lei ZHAO, li-hong WU,"A feature extraction method using base phrase and keyword in Chinese text", "in press" Proceedings of 2008 3rd International Conference on Intelligent System and Knowledge Engineering, Pages: 680-685, 2008 IEEE.

[2] Topon Kumar Paul and Hitoshi Iba,"Prediction Of Cancer Class With Majority Voting Genetic Programming Classifier Using Gene Expression Data", IEEE/ACM transactions on computational biology and bioinformatics,vol.6,April-June 2009,Pages:353-367.

[3] Zhongwu Zhai†, Bing Liu‡, Jingyuan Wang†, Hua Xu† and Peifa Jia ,"Product Feature Grouping for Opinion Mining Using Soft-Constraints and EM", "in press" IEEE Intelligent Systems,2011.

[4] Michele Carenini,Angus Whyte,Lorenzo Bertorello, Massimo Vanocchi,"Improving Communication in E-democracy Using Natural Language Processing",IEEE Intelligent Systems,2007,Pages:20-27.

[5] Wang Wei, Payam Barnaghi and Andrzej Bargiela,"Probabilistic Topic Models for Learning Terminological Ontologies",IEEE Transactions on Knowledge and Data Engineering",July 2010,Pages:1028-1040.

[6] Yanhong Zhai and Bing Liu, "Structured Data Extraction from the Web Based on Partial Tree Alignment", IEEE Transactions on Knowledge And Data Engineering,December 2006,Pages:1614-1628.

[7] Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE Transactions on Knowledge And Data Engineering, VOL. X, NO.X 2011,Pages:1-14.

[8] Rile Hu, Chengqing Zong and Bo Xu,"An Approach to Automatic Acquisition of Translation Templates Based on Phrase Structure Extraction and Alignment", IEEE Transactions on Audio, Speech, And Language Processing,VOL. 14, NO. 5, 2006,Pages:1656-1663.

[9] Hong Huang and Hailiang Feng, paper "Gene Classification Using Parameter-free Semi-supervised Manifold Learning", IEEE/ACM Transactions on Computational Biology And Bioinformatics, VOL. X, NO. X 2011,Pages:1-13.

[10] Aree Thunkijjanukij, Asanee Kawtrakul, Supamard Panichsakpatana , Uamporn Veesommai , ," Lesson learned for ontology construction with Thai rice case study", "in press".World Conference on agricultural information and IT",2008,Pages:495-502.

[11] WEN Bi-long , HANG Li,"Method of building petroleum exploration and production domain ontology", "in press" Computer Engineering and Applications, 2009.45(034):p.1-3.

[12] Zhang Rui-ling, XU Hong-sheng,"Using Bayesian network and neural network constructing domain ontology", "in press" World Congress on Computer Science and Information Engineering, 2009, IEEE 2008, Pages: 116-231.

[13] Guarino.N, "Formal ontology in information systems", "in press"1998.Proceedings of IOS.

[14] Elena P. Sapozhnikova," Multi-label classification with art neural networks", "in press" Second International Workshop on Knowledge Discovery and Data Mining, 2009 IEEE, Pages: 144-147.

[15] Shubin Zhao Ralph Grishman," Extracting Relations with Integrated Information Using Kernel Methods", "in press" *Proceedings of the 43rd Annual Meeting of the ACL*, pages 419–426, Ann Arbor, June 2005.

[16] Hongbo Liu1, 3, Ajith Abraham2, and Benxian Yue3," Nature Inspired Multi-Swarm Heuristics for Multi-Knowledge Extraction", "in press" Advances in Machine Learning II, pp. 445–466, 2010.

[17] Weng,S.,Tsai,H,Liu,S.,and Hsu,C.,,"Ontology Construction for information classification ,Expert Systems with Applications",31(1),1-12.

[18] Gaoying Cui, Qin Lu, Wenjie Li, Yirong Chen. Corpus Exploitation from Wikipedia for Ontology Construction: 2125-2132.

[19] Chen-Huei Chou, Fatemeh Zahedi, Huimin Zhoa, 2008. Ontology for developing websites for natural disaster management: methodology and implementation.

[20] Antonio M. Rinaldi, 2009. An Ontology-Driven Approach for Semantic Information Retrieval on the Web. In ACM Transactions on Internet Technologies, Volume 9, Article 10.

[21] Ahmed Rafea, Hesham A. Hassan, Mohamed Yehia Dahab, 2006. TextOntoEx: Automatic Ontology Construction from Natural English Text. International conference of Artificial Intelligence and Machine Learning.

Dhanasekaran.K received the B.E degree in Computer Science from Mahendra Engineering College affiliated to Anna University Chennai in 2006.He received the M.E degree in computer science from K.S.R College of Engineering affiliated to the Anna University of Technology Coimbatore in June 2009.

Currently he is a research scholar&Assistant Professor in Info Institute of Engineering which is affiliated to Anna University of Technology,Coimbatore,Tamilnadu,India.His current research interests includes semantic Web, ontologies, machine learning,Data mining, semantic Web services, and information search and retrieval. He is a member of the MISTE.

Rajeswari.R received the B.E degree from Thiagarayar Engineering College in 1995.She received M.E degree from Thiagarayar Engineering College in 1998.She received her Ph.D in Power System Engineering from Anna University,Chennai,India in 2009.She is currently an Assistant Professor of Electronics and Instrumentation Engineering Department at the Government College of Technology,Coimbatore,India.Her research areas includes Power System Engineering,Power System Protection.She is a member of the ISTE.