RESEARCH ARTICLE OPEN ACCESS

Ontology Based Concept Hierarchy Extraction of Web Data

S. Suresu [1], M. Elamparithi [2]
Research Scholar [1], Assistant Professor [2]
PG Department of Computer Applications
Sree Saraswathi Thyagaraja College, Pollachi
Tamil Nadu – India

ABSTRACT

Markov Logic Networks (MLNs) constitute a methodology for measurable social discovering that consolidates first request rationale with Markov arbitrary fields. A MLN is a first request rationale learning base with weights that can be either positive or negative, related to every equation. This exploration coordinating syntactic structure highlights proposed a technique for OLMLN semantic relations extraction taking into account Markov rationale systems (MLNs) and the object of connection extraction. The analyses demonstrate that the impact of OLMLN semantic connection extraction is better in the wake of coordinating into the syntactic structure highlights, in the near trial, the consequence of OLMLN semantic connection extraction in view of the MLNs is superior to the current methodologies and demonstrates that the proposed strategy has a decent impact. The Ontology Based Concept Hierarchy Extraction of Web Data (OBCHED) depicts the procedure of idea chain of command extraction. Idea pecking order is the procedure that contains sub procedure of idea distinguishing proof and idea extraction. The most existing system of order extraction was produced from the formal idea investigation and Markov Logic Networks. In OLMLN, chain of command extraction in light of connection forecast and web information recovery is proposed. Join forecast is utilized to anticipate the terms in learning process. The Ontology Based Concept Hierarchy Extraction of Web Data (OBCHED) comprises of Pre-Processing, Concept Identification and Concept Hierarchy Extraction.

Keywords:- Markov Logic Networks, Concept hierarchy extraction, Web data extraction.

I. INTRODUCTION

Ontologies are effectively formal and explicit specifications, in the form of concepts and relations, of shared conceptualizations. Ontologies may contain axioms for validation and enforcing constraints. A concept hierarchy constitutes the backbone of an ontology and many techniques have been proposed for extracting concept hierarchies from text. Ontology Based Concept Hierarchy Extraction of Web Data (OBCHED) is one such technique for extracting concept hierarchies from text. The Ontology Based Concept Hierarchy Extraction of Web Data (OBCHED) describes the process of concept hierarchy extraction. Concept hierarchy is the process that contains sub process of concept identification and concept extraction. The most existing technique of hierarchy extraction was developed from the formal concept analysis and Markov Logic Networks. In this research, concept hierarchy extraction is based on link prediction and web data retrieval. Link prediction is used to predict the terms in learning process. The OBCHED techniques have some types. There are Pre-Processing, Concept Identification and Concept Hierarchy Extraction. The Ontology Based Concept Hierarchy Extraction of Web Data (OBCHED) process is shown in Figure-1.

II. PRE-PROCESSING

A first step of the pre-processing phase is to tokenization. A token in a sentence is typically associated with a bag of features obtained via one or more of the following criteria:

- The string representing the token.
- Orthography type of the token that can take values of the form capitalized word, small case word, mixed case word, number, special symbol, space, punctuation, and so on.
- The Part of speech of the token.
- The list of dictionaries in which the token appears. Often this can be further refined to indicate if the token matches the start, end, or middle word of a dictionary. For example, a token like “New” that matches the first word of a dictionary of city names will be

associated with a feature, “Dictionary-Lookup = start of city.”

- Annotations attached by earlier processing steps.

After that, the tokens are annotated with Part of Speech (POS) tags and their lemmas. After the lemmatization, the chunking step takes place. The goal of this phase is to discover sets of words that, together, form a syntactic unit. After that, the syntactic analysis takes place in order to extract the syntactic dependencies between words. The syntactic dependencies are relationships that words hold within a sentence. They indicate, for instance, who are the subject and the object of a given verb or which noun is modified by a given adjective. The syntactic dependencies considered in this work are represented according to the Stanford dependencies. Then, the tokens containing terms from a stop list are removed. The remaining terms are weighted using TF-IDF scores only terms with weights above a certain value are selected. At lost, the WordNet is used for extracting hypernym relations between the selected terms.

Process we were taking as part of portions by GATE tool. Using GATE tool we have to do some kinds of processes. There are tokenizations, POS tagging, chunking and syntactic analysis. Tokenization is the main part of learning process, because every word in the sentence named as tokens. The result of this portion displays the tokens of each and every sentence. Most of the application we were using POS tag to getting grammatical tags in the sentence or word corpus.

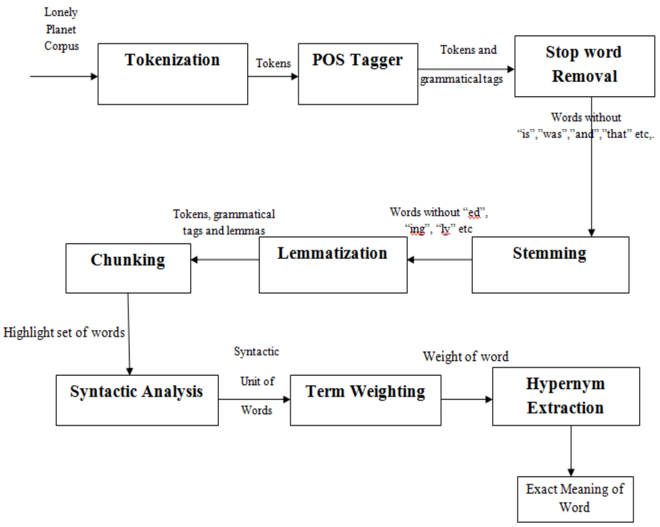

Figure-1: Pre-Processing

The author of Jone Correia and Rosario Girardi explains the portion of tagging in some way. The way of getting input as corpus and transform the input into a model that can be processed by computationally. POS tag getting over move onto the process stop-word removal, stemming and lemmatization. Stop-Word Removal is the process of removing unwanted letters in each and every word in ontology learning. Here, we had to remove the letters like “is”, “was”, “and”, “that” etc., in stop-word removal performing the taxonomic relationship based removal. Completion of removal we go the process of stemming. In stemming the stemming algorithm proper, retrieves the stem of a word by removing its longest possible ending which matches one on a list stored in the computer. Next, handles "spelling exceptions," mostly instances in which the "same" stem varies slightly in spelling according to what suffixes originally followed it. Stemming process should be remove the letters placed in the words are “ed”, “ly”, “ing”. Usually English words constitute some morphological paradigm to assigning the lemmas.

Lemmatization progress may be of grouping the words that belongs to the same inflectional paradigm and assigning to each paradigm its corresponding canonical form called lemma. Lemmatization process performs four steps. There are,

1. Removal of suffix of length
2. Addition of new lemma suffix
3. Removal of prefix of length
4. Addition of new lemma prefix

After performing the part of speech tagging the GATE tool do the process of chunking. These processes are shown in Figure-1.

The results came from the lemmatization process having tokens, grammatical tags and lemmas. Here after for our contribution we have to do the operation of chunking. The reason for performing this operation is to highlight the set of forming words and according to these words co-ordinate the integration of web data. In chunking, set of words could be formed and display in highlight manner. GATE tool performed this operation in our implementation side. Then every part in the ontology learning progress wants to know the syntactic meaning of the words. For this purpose we examine the syntactic unit of every word. This method is very useful but is not always easy to manipulate. The options for modification provide another way to identify the categories that are relevant for both word formation (morphology) and phrase formation (syntax).

The final stage of pre-processing is to calculate the term weight and also hypernym extraction. In term weighting option we have to find out the weight of every word. Term weight is calculated by the scores of TF-IDF calculation that is the Term Frequency and the inverse document frequency. Finally, we perform the process of hypernym extraction. Hypernym extraction performed with the help of hearst pattern. To determine the possible hypernym of particular noun we use the same parsed text. Then construct the vector of each hypernym. It would be useful for identify terms made up of multiple words rather that just using the head nouns of the noun phrases.

III. CONCEPT IDENTIFICATION

Concept Identification is an important portion covered in our proposed system. Concept identification is performed by the technique of MLN. Using MLN we have to perform the process of learning weight and inference. Figure 4.2 describes the process of concept identification. For performing the learning weight we have to use the method of MLN. To find the weights in a database we have to use the Maximum a Posteriori (MAP) weight method. This means the weights that maximize the product of their prior probability and the data likelihood. Pseudo-likelihood is that the product of the conditional chance of every variable given the values of its neighbors within the data. Whereas economical for learning, it will offer poor results once long chains of inference are needed at enlarging time. Pseudo-likelihood is systematically outperformed by discriminative coaching, it minimizes the negative conditional probability of the question predicates given the evidence ones. This learning weight can be performed by four methods. First, progress based on voted Perceptron. Here, using gradient descent algorithm use the gradient named as g, scaling based learning rate η, and to update the weight vector w, it can be represented by,

$$W_{t+1} = W_t - \eta g$$

The spinoff of the negative conditional log-likelihood (CLL) with relevancy a weight is that the distinction of the expected range of true groundings of the corresponding clause and therefore the actual range in step with the information.

$$\frac{\partial}{\partial W_i} - \log P(Y{=}y \mid X{=}x) = E_w[n_i] - n_i$$

Where y is the state of the non-evidence atoms in the data, and x is the state of the evidence.

The second process is the contrastive divergence. In contrastive divergence we use MCMC algorithm. The MCMC algorithmic program usually used with contrastive divergence is Josiah Willard Gibbs sampling, expect for MLNs a lot of quicker various method MC-SAT is offered. Because ordered sample in MC-SAT square measure a lot of less related to than ordered sweeps in Josiah Willard Gibbs sampling, they carry additional data and square measure doubtless to yield a better descent direction. Specially, the various samples square measure doubtless to be from completely different modes, reducing the error and potential instability related to choosing one mode.

The third progress is per-weight learning rates. To modify each algorithms to its own a distinct learning rate for each weight. Since standardization of each learning rate individually is impractical, we use an easy heuristic to assign a learning rate to every weight.

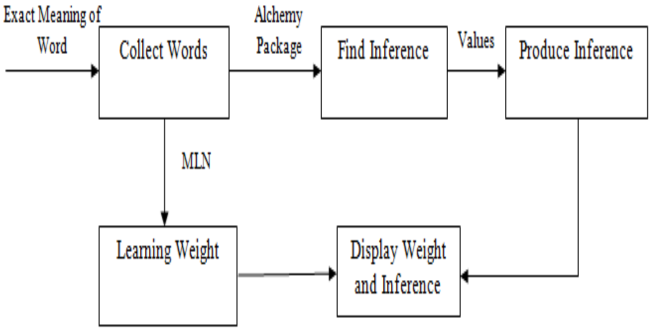

Figure-2: Concept Identification

$$\eta_i = \frac{\eta}{n_i}$$

Where η is the user-specified global learning rate and n_i is the number of true groundings of the ith formula. These values are being fixed, so it cannot be contribute to the variance. The final process in the series is Diagonal Newton. Diagonal Newton is just multiplying the gradient, g, by the inverse Hessian, H inverse.

$$W_{t+1} = W_t - H^{-1}g$$

Diagonal Newton (DN) methodology uses the inverse of the diagonoized jackboot insitu of the inverse jackboot. DN typically uses a smaller step size than the total Newton methodology. The main aim of this method is to found the step size. In each iteration, we take a step in the diagonalized Newton direction

$$W_i = W_i - \alpha \frac{E_w[n_i] - n_i}{E_w[n_i^2] - (E_w[n_i])^2}$$

Then we compute the step size,

$$\alpha = \frac{-d^T g}{d^T H d + \lambda d^T d}$$

Where d is the search direction. For a quadratic function and $\lambda = 0$, this step size would move to the minimum function value along d.

Regarding inference we have to perform the task of finding inference using alchemy software we have to finalize the inference values of each word in the schema. Before enter into the process of concept extraction we have to know about the PLSA process. Using this method we can derive the meaningful words in the corpus. So, using PLSA we find the synonym of the word.

IV. CONCEPT HIERARCHY EXTRACTION

In Concept Hierarchy Extraction we have to do three kinds of processes. There link prediction, hierarchy extraction using FCA and hierarchy extraction using Hearst Pattern. In that progress we have to include some additional processes. In first, link prediction task can be done with the help of MLN. The Markova logic network can extract the words in different manner. For that purpose predicts the links in an efficient manner. The link prediction downside could be relevant to variety of fascinating current applications of social networks. Progressively, for example, researchers in artificial intelligence and data processing have argued that oversized organizations, such as an organization, will benet from the interactions inside the informal social network among its members. These can be serving to supplement the official hierarchy obligatory by the organization itself. Effective ways for link prediction may be wont to analyze such a social network and recommend promising interactions or collaborations that have not nevertheless been utilized inside the organization. Figure-3 can explains the process of hierarchy extraction.

Formal Concept Analysis (FCA) is a method mainly used for the analysis of data in hierarchy manner. FCA can be seen as a conceptual clustering technique as it also provides intentional descriptions for the abstract concepts or data units it produces. Concept Hierarchies represent a conceptualization of a website with regard to a given corpus within the sense that they represent the relations between terms as they are utilized in the text. However corpora represent a really restricted read of the globe or a certain domain thanks to the very fact that if one thing is not mentioned, it does not mean that it is not relevant, however merely that it's not a difficulty for the text in question. The learned construct hierarchies have to be compelled regarded as approximations of the conceptualization of an exact domain. In concept extraction process we move onto the step of hierarchy extraction using hearst pattern. Usually hearst patterns are used to discover the hypernym in the learning process. This will produce excellent result of discovering meaningful words. The main goal of our work is to automatically identify lexico-syntactic patterns indicative of hypernym. According to the taxonomic relationships we have to do the extraction process in an efficient manner.

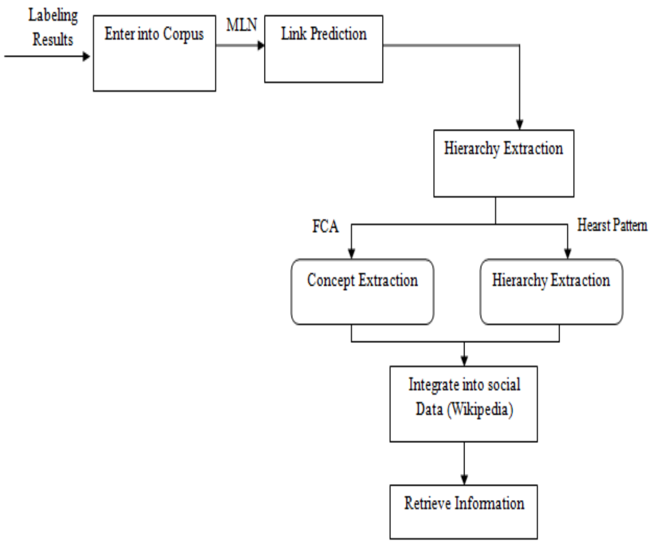

Figure-3: Concept Hierarchy Extraction

Our first super-ordinate classifier is predicted on the intuition that unseen noun pairs square measure additional likely to be a super ordinate try if they occur within the check set with one or additional lexico-syntactic patterns found to be indicative of super ordination. We have a tendency to then produce a feature count vector for every such noun try. Our process is to discovering that dependency ways would possibly prove helpful options for our classifiers. Dependency ways consisting of every dependency path that occurred between a minimum of five distinctive noun pairs in our corpus. To evaluate these options, we tend to make a binary classifier for every pattern that merely classifies a noun combine as hypernym/hyponym if and providing the specific pattern happen a minimum of once for that noun combine. In our process we decide to integration of our hierarchy process into social web data. Social web data could produce the results of like Wikipedia. Wikipedia can efficiently perform the task of information retrieval. So like that our hierarchy progress should display the information about web data. The information retrieval can be integrated into the ontology learning. The social web data can produce the large amount of information of tourism places. Using these three methods such as MLN, FCA and also hearst pattern we have to produce the result in a hierarchy manner. The hierarchy order will take the order of country, organization, date, person, location, money then vehicle. The reason for choosing this order is based on the corpus. Because here, we were chosen the corpus of lonely planet. This lonely planet can contain the details of tourism.

V. RESULT AND DISCUSSIONS

We accomplish the wide-ranging set of experiments to scrutinize the performance of the proposed method of OBCHED framework by comparing it into the state-of-art method. The proposed OBCHED technique consists of three processes. There are Pre-Processing, Concept Identification and Concept Hierarchy Extraction. That technique can provide better results of extraction. Using this proposed technique we have to acquire the accuracy of words involved in the progress. The extraction results and some other dataset results are should be given below.

The OBCHED technique consists of the steps of Pre-processing, Weight Learning, Inference and Concept Hierarchy Extraction. Here the pre-processing steps contains tokenization, POS tagging, chunking and syntactic analysis these all are performed by the GATE tool. Then Stemming, Stop words, Lemmatization, term weighting these can be have performed by the java code. Finally we calculate the Term weight for the pre-processing process. Then hierarchy extraction should integrate the social web data. These all process is covered by the OBCHED techniques. The experiment results can be compared with the technique of PREHE and OBCHED. Figure-5 and Figure-6 can show the results of comparison of both techniques. The results can be varying with the parameter of accuracy. Our process should produce more accuracy when compared to previous technique. The accuracy can be calculated with the help of the values of precision, recall and alsoF1 measure. Our proposed OBCHED technique can have the best precision, recall and also F1 measure. According to those values the accuracy of every technique shall be calculated.

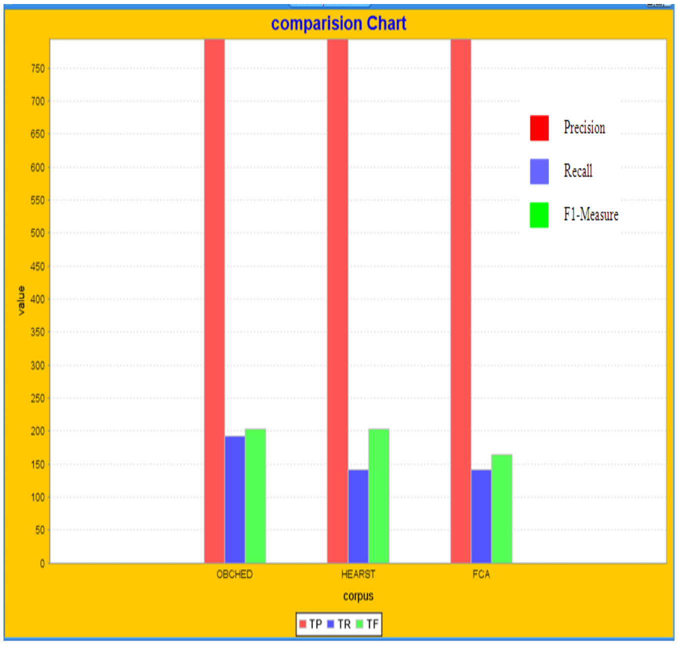

Figure-5: Lonely planet corpus experimental results

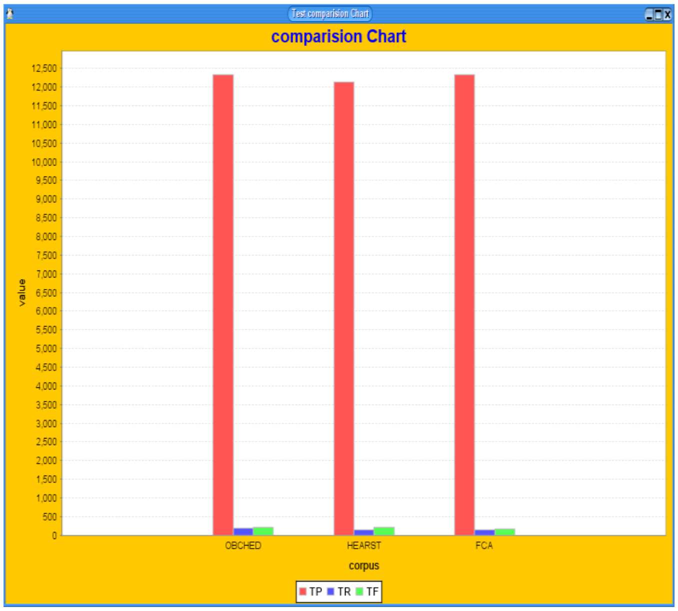

Figure-6: Lonely planet corpus experimental results

VI. CONCLUSION

Pre-processing consists of many activities. These are all placed in ontology learning progress. Pre-processing is used for extracting meaningful words from the corpus. The pre-processing activities are could be performed by tools and languages. In our process we use GATE tool for performing operations of tokenization, POS tagging, chunking and syntactic analysis. After that we had to do the activities of stop-word removal, stemming, lemmatization, term weighting and also hypernym extraction. These all are done by using Java language. Second process is concept identification. In concept identification we use MLN method to learning the weight of words. For that purpose we can use the simple weight learning method to produce the good results. The main progress include in that is to find the inference values. For finding the inference we could use alchemy process. It may be the software to produce the optimized values of every word in the corpus.

Alchemy Packages also used for implement the concept identification process. Alchemy packages are used for make the perfect inference process. For implementing all this performance, we use the dataset of Lonely Planet. The final process of our OBCHED technique is the concept hierarchy extraction. In that situation, need to do the efficient way of hierarchy extraction. For this reason we predict the links in the social networks. The link prediction process to be done with method of Markov Logic Network (MLN). Then hierarchy extraction shall be processed using formal concept analysis (FCA). In this period, we have to extract the words with meaningfully. After that perform the hierarchy process in Hearst pattern manner. It could provide the hierarchy results in proper manner. For this purpose we should use the wordnet tool for extracting the hypernym words which means exact meaning. Our OBCHED technique can do the process of integrating the social data into our hierarchy manner. It could produce the efficient way of retrieving the content from the web.

REFERENCES

[1] Faure, D., Nedellec C., and Rouveirol, C., Acquisition of Semantic Knowledge using Machine Learning Methods: The System ASIUM, Technical Report number ICS-TR-88-16, Laboratoire de Recherche en Informatique, Inference and Learning Group, Universite Paris Sud, 1998.

[2] Yamaguchi, T., Acquiring Conceptual Relations from domain-Specific Texts, Proceedings of the IJCAI 2001 Workshop on Ontology Learning (OL'2001), Seattle, USA, 2001.

[3] Shamsfard M., and Barforoush, A. A., (a) An Introduction to HASTI: An Ontology Learning System, Proceedings of 6th Conference on Artificial Intelligence and Soft Computing (ASC'2002), Banff, Canada, June, 2002.

[4] Hahn U., Romacker, The SYNDIKATE Text Knowledge Base Generator, Proceedings of the 1st International Conference on Human Language Technology Research, San Diego, USA, 2001.

[5] Chalendar, G., Grau, B., SVETLAN' A System to Classify Nouns in Context, Proceedings of the ECAI 2000 Workshop on Ontology Learning (OL'2000), 2000.

[6] Sanchez, D., and Moreno, A. Creating ontologies from Web documents. In Recent Advances in Artificial Intelligence Research and Development. IOS Press, Vol. 113, pp.11-18, 2004.

[7] Sabou, M., Wroe, C., Goble, C., and Mishne, G. Learning Domain Ontologies for Web Service Descriptions: an Experiment in Bioinformatics. In Proceedings of the 14th International World Wide Web Conference (WWW2005), Chiba, Japan, 2005.

[8] Cimiano, P., Hotho, A., Staab, S. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. JAIR - Journal of AI Research, Vol. 24, pp. 305-339, 2005.

[9] Schmid, H. Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing,1994.

[10] Schmid, H. Lopar: Design and implementation. In Arbeitspapiere des Sonder for schungsbereiches, No. 149, 2000.

[11] Cimiano P., and Vaolker, J. Text2Onto – A Framework for Ontology Learning and Data-driven Change Discovery. In: Montoyo, A., Munoz, R., Metais, E. Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), Lecture Notes in Computer Science. Alicante, Spain: Springer, 2005.

[12] Ganter, B. and Wille, R. Formal Concept Analysis - Mathematical Foundations. Berlin:Springer-Verlag,1999.

[13] Frantzi, K., Ananiadou, S., and Tsuji, J. The cvalue/nc-value method of automatic recognition for multiword terms. In Proceedings of the ECDL .pp 585-604. 1998.

[14] Karoui, L., Aufaure, M., and Bennacer.N. Ontology Discovery from Web Pages: Application to Tourism. In ECML/PKDD 2004: Knowledge Discovery and Ontologies KDO-2004.

[15] Bennacer, N., and Karoui L. A framework for retrieving conceptual knowledge from Web pages. In Semantic Web Applications and Perspectives, Proceedings of the 2nd Italian Semantic Web Workshop, University of Trento, Trento, Italy, 2005.

[16] Davulcu, H., Vadrevu, S., and Nagarajan, S.OntoMiner: Bootstrapping Ontologies From Overlapping Domain Specific Web Sites. In: Poster presentation at the 13th International World Wide Web Conference May 17-22, New York, 2004.

[17] Assadi, H., Knowledge Acquisition from Texts: Using an Automatic Clustering Method Based on Noun-Modifier Relationship, Proceedings of 35th Annual Meeting of the Association for Computational Linguistics (ACL'97), Madrid, Spain, 1997.

[18] Finkelstein-Landau, M., Morin, E., Extracting Semantic Relationships between Terms: Supervised vs. Unsupervised Methods, Proceedings of International workshop on Ontological Engineering on the Global Information Infrastructure,71-80, Dagstuhl-Castle, Germany, 1999.

[19] Hahn U., and Marko, K. G., Ontology and Lexicon Evolution by Text Understanding, Proceedings of the ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering (OLT'2002), Lyon, France, 2002.

[20] Maedche, A., and Staab, S., (a), Ontology learning for the Semantic Web, IEEE journal on Intelligent Systems, Vol. 16, No. 2, 72-79, 2001.

[21] Hearst, M.A., Automatic Acquisition of Hyponyms from Large Text Corpora, Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France, July, 1992.

[22] 21.Assadi, H., Construction of a regional ontology from Text and its use within a Documentary System, Proceedings of Formal ontologies in Information Systems (FOIS'98), Italy, 1999.

[23] Sundblad, H., Automatic Acquisition of Hyponyms and Meronyms from Question Corpora, Proceedings of the ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering (OLT'2002), Lyon, France, 2002.

[24] Gamallo, P., Gonzalez, M., Agustini, A., Lopes, G., and de Lima, V. S., Mapping Syntactic Dependencies onto Semantic Relations, Proceedings of the ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering (OLT'2002), Lyon, France, 2002.