

Research Article

Web Page Classification Using SVM and FURIA

¹P. Madhubala and ²K. Murugesan

¹Department of Computer Science and Engineering, Tagore Institute of Engineering and Technology, Tamilnadu, India

²Department of Electronics and Communication Engineering, Sree Sastha Institute of Engineering and Technology, Tamilnadu, India

Abstract: Text Classification classifies a document, under a predefined category. Mostly, an automatic text classification is an important application taken as a research topic, since the inception of digital documents. In this study, Hypernyms, superordinate words are identified in web and clubbed with entailment rule acquisition. Available tree of hyponym words in the document has been created and used with dependency tree. Features extraction is performed with weighted Term Frequency-Inverse Document Frequency (TF-IDF) where the weight of the word can be computed based on the number of hyponyms present in the radix tree. Performance evaluation is done using Support Vector Machine (SVM) classifier and Fuzzy Unordered Rule Induction Algorithm (FURIA) classifier.

Keywords: Hypernym, hyponym, radix tree, Support Vector Machine (SVM) and Fuzzy Unordered Rule Induction Algorithm (FURIA), Term Frequency-Inverse Document Frequency (TF-IDF)

INTRODUCTION

Webpage classification techniques use text in the page, the link structure, hyperlink structure or anchor text information to classify a target page. Uniform Resource Locator (URL) is another most commonly used information in web classification as it is the least expensive and most informative source (Kan and Thi, 2005). The amount of web pages on the web is more than 1 billion and there is an exponential increase in the amount of data available on the web. The addition of this enormous amount of data along with interactive and content rich nature of the web made it very popular. But these pages vary to a great extent in both the content and quality of information.

Knowledge is information used in an expert system to behave intelligently. Knowledge acquisitions are from human experts, books, documents, or files. The knowledge varies and is specific to the problem domain or it may be general knowledge such as knowledge about business, or it may be meta-knowledge. Knowledge Acquisition can be defined as the process of extracting, structuring and organizing knowledge from several sources (Nasuti, 2000).

Semantic Web relies on the formal ontologies for the comprehensive and transportable machine understanding purpose and so success of the semantic web depends strongly on the proliferation of ontologies. Conceptual structures are defined on ontology,

appropriate to the idea of machine process-able data on the semantic web. Ontologies are data schemas; provide a controlled vocabulary of concepts in which each includes an explicitly defined machine process-able semantics. Shared and common domain theories are explained so that ontologies help both people and machines to communicate concisely thereby supporting the exchange of semantics and not only syntax.

Ontology learning is used to extract an ontological element (conceptual knowledge) from input and building ontology from them. Many works are focused at constructing the ontologies automatically from a given corpus with a limited human input (Hazman *et al.*, 2011). It consists of techniques to build ontology from scratch, or improve/adapt an existing ontology in a semiautomatic fashion. Ontology learning methods require diverse techniques from different fields like knowledge acquisition, database management, natural-language processing, information retrieval, artificial intelligence and machine learning.

Objective of Ontology Learning (OL) is the integration of a multitude of disciplines to facilitate the construction of ontologies (ontology engineering and machine learning). Since the fully automatic knowledge acquisition by machines remains in the distant future, the overall process is considered as a semi-automatic with human intervention. OL relies on the “balanced cooperative modeling” paradigm (Maedche and Staab, 2004) which describes a coordinated interaction

Corresponding Author: P. Madhubala, Department of Computer Science and Engineering, Tagore Institute of Engineering and Technology, Tamilnadu, India

This work is licensed under a Creative Commons Attribution 4.0 International License (URL: <http://creativecommons.org/licenses/by/4.0/>).

between human modeler and learning algorithm for ontology construction. In most semantic web applications, ontology knowledge is an essential part. Yet ontology is not sufficient for representing an inferential knowledge (Soumya and Swathi, 2013) as ontology-based reasoning has limitations compared with rule-based reasoning.

Text mining is used to retrieve documents related to ontology's domain from search engines, online digital libraries and scholarly search engines where text classification identifies them from collected documents, those likely to be related to ontology's domain. Many text mining techniques are presented to extract information for ontology enrichment from existing documents. Textual entailment recognition is the deciding task with two given text fragments, whether the meaning of one text is entailed from another text. This task captures a broad range of relevant inferences for multiple applications. Rule acquisition is an essential one but forms bottlenecks during deployment and is also time consuming and laborious as it requires knowledge and domain experts. A communication problem occurs between them if they acquire rules from several sites of the same domain (Dagan *et al.*, 2010).

A hyponym is a word describing more specific things. Proper nouns are good examples of hyponyms. Also, for a concept, hyponym can be described. For example Niagara falls and Ford for the concept of waterfall and car, respectively. Hypernyms refers to broad categories or general concepts. For example, car or fruit is hypernyms for more precise terms like Ford, or Banana. The relation of hyponyms capture the 'is a kind of' concept. Hyponyms refer to all the words and not only for 'things'. Words such as punch shoot and stab describes the 'actions' that is treated as co-hyponyms of the super-ordinate term injure.

This study identifies the web page Hypernyms (superordinate words) and clubbed it with entailment rule acquisition to classify the web documents.

LITERATURE REVIEW

Rios-Alvarado *et al.* (2011) introduced an approach to find hypernym, a relation between terms belonging to a specific knowledge domain. WordNet synsets and context information were combined to build an extended query set and this query set has been sent to a web search engine to retrieve the most representative hypernym for a term.

Ranwez *et al.* (2012) proposed an efficient algorithm to identify the additional key concepts based on closure of two common graph operators such as Least Common-Ancestor (LCA) and Greatest Common Descendant (GCD). Results show that the produced ontology excerpts focused on a set of concepts of interest and fast enough to use in interactive environments. For instance, OntoFocus ([http:// www.ontotoolkit.mines-ales.fr/](http://www.ontotoolkit.mines-ales.fr/)), was used to restrict, the

large Gene Ontology to a sub ontology focused on concept of annotating a gene related to breast cancer.

Yildiz and Yildirim (2012) proposed a method for the automatic acquisition of hypernym/hyponymy relations from a Turkish raw text. After prospective hyponyms has been extracted by using lexico-syntactic patterns, an Apriori algorithm was applied to eliminate faulty hyponyms and increase precision. A model based on a particular lexico-syntactic pattern and association rules for Turkish language retrieved many is-a relation with high precision.

Koirala and Rasheed (2008) made a comparative study on the effectiveness of using morphological and ontological information for text categorization. The results show that stemming-based text representation achieved better performance than hypernym-based text representation. In proposed method the stemming based representation has been combined with hypernym based representation. The combined representation does not produce an improved performance.

Vigneshwari and Aramudhan (2012) proposed a novel approach for creation of User Profiling Ontologies (UPO) which is used for personalizing the web. A cross ontology mechanism has been applied where the user queries were continuously monitored. Stemming process was performed so that the important keywords were extracted and the frequently visited web pages were ranked. Based on this, UPO was automatically generated. Ontology mapping was performed for mapping the UPO with global ontology to personalize the web pages. Data samples were collected and results proved a good improvement for the proposed approach.

Inyaem *et al.* (2009) studied and compared several machine learning methods to implement a Thai terrorism event extraction system to extract information from Thai news articles. Three named entity feature selection techniques were compared for entity recognition provided by terrorism gazetteer, terrorism ontology and terrorism grammar rules. The machine learning algorithms such as Naive Bayes (NB), K-Nearest Neighbor (KNN), Decision Tree (DTREE) and Support Vector Machines (SVM) were used for event extraction. Each term feature was weighted by the Term Frequency-Inverse Document Frequency (TF-IDF). Finite state transduction was applied for learning feature weights. Results show that the SVM algorithm with a terrorism ontology feature selection achieved best performance with 69.90% for both precision and recall.

Khan *et al.* (2010) proposed semantics base feature vector using Part of Speech (POS). This method used to extract the concept of terms using WordNet, co-occurring and associated terms. The proposed method was applied on small documents dataset which outperformed than term frequency/inverse document

frequency with BOW feature selection method for text classification.

Yoo (2011) suggested a modified knowledge acquisition framework that focused on the autonomous acquisition of knowledge in ordinary dialogues. This then underpinned by the functionality of SVM which was demonstrated to identify the topic of knowledge in a most accurate and efficient way. Feasibility of the proposed method has been validated by Context-based Knowledge Acquisition System (CKAS).

Liu *et al.* (2009) proposed a novel and effective method to extend an ontology instances from Chinese free text, which was achieved using SVM classification. Classification features were extracted from the training texts and from new texts based on an existed Chinese ontology. Then ontology has been changed as tree hierarchical structure and used as the training and learning strategy of SVM classifier. At last, new ontology instances were extracted from the new texts based on training results. Proposed method was better in terms of the semantic of ontology elements and instances extraction. Classification was completed in the identical procedure at same time. Experimental results showed that the average accuracy of instances extraction and classification achieved was 86.6%.

Luong *et al.* (2009) presented a complete framework for ontology learning that enabled to retrieve documents from the Web using focused crawling and used SVM classifier to identify domain-specific documents and performed text mining to extract useful information for the ontology enrichment process. Wang and Lu (2011) researched on ontology auto-extension. Technology based on information processing was applied to automatically extend ontology instances from free text. Three tasks were identified after analyzing on ontology auto-extension, a new Binary Tree (BT) was constructed based on the ontology taxonomy and then it was used as the training and learning strategy of SVM classifier. Main advantage of Onto-BT-SVM model was that the semantic of ontology is fully used. Different multi-class classification strategies were compared to verify their impacts on classification effects. Experiment results show that the recall and accuracy obtained was 90.5 and 93.5%, respectively.

Bai and Li (2009) proposed an improved Naive Bayesian Web text classification algorithm. Common bayesian classifier assumed equal importance on all the items while the terms in each title were considered to be more important than others. Experiments show that the improved Naive Bayesian algorithm was more precise in the text classification.

Xu *et al.* (2011) presented a web page classification algorithm named as Link Information Categorization (LIC). Based on the K nearest neighbor method, proposed method combined information on the

website features, to implement the Web page link with information classification. Experiments show that the algorithm achieved higher efficiency and accuracy on Web page classification.

Bo *et al.* (2009) studied some feature selection methods such as ReliefF and Symmetrical Uncertainty (SU) to address the high dimensional text vocabulary space of web pages. Hidden Naive Bayes, Complement class Naive Bayes and other traditional techniques were used for web page classification. Results show that the abilities of HNB performed better than other methods and SU was more competitive than ReliefF in web pages categorization.

METHODOLOGY

In this study, dataset namely, Barnes and Noble is used. Features are extracted using weighted TF-IDF where the weight of the word is computed depending on the number of hyponyms present in the radix tree. Performance can be evaluated using SVM classifier and FURIA classifier.

Entailment rules: An entailment rule ‘L→R’ is a directional relation between two templates, L and R. For example, ‘X acquires Y→X own Y’ or ‘X beat Y →X play against Y’. Templates correspond to text fragments with variables are either linear phrases or parse sub-trees. Goal of entailment rules is to help applications infer one text variant from another. This rule is applied to a given text only when L can be inferred from it, with appropriate variable instantiation. Entailment rules must apply only in specific contexts, namely relevant contexts. For example, the rule ‘X acquires Y→X buy Y’ can be used in the context of ‘buying’ events.

Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF is obtained from the theory of language modeling that the terms in a given document is divided into with and without the property of eliteness (i.e., the term is approximately the topic of given document or not) (Zhang *et al.*, 2008). Eliteness of a term is evaluated by TF and IDF that measures the importance of this term in the collection. The standard evaluation of TF-IDF used for term weighting is as in Eq. (1):

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

where,

$w_{i,j}$: The weight for i^{th} term in j^{th} document

N : The number of documents in the collection

$tf_{i,j}$: The term frequency of i^{th} term in j^{th} document

df_i : The document frequency of i^{th} term in the collection

TF-IDF is a common metric in text categorization; it has been used as a unigram feature weight

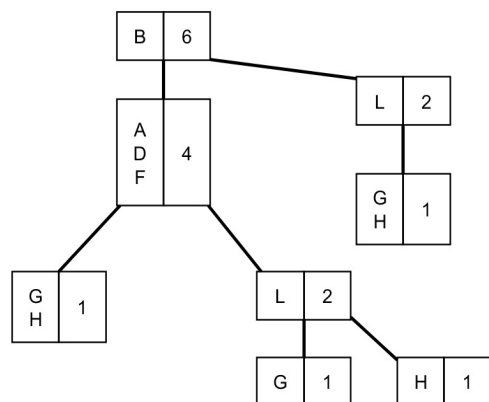


Fig. 1: Example of radix-tree structure

(Jotheeswaran and Kumaraswamy, 2013). TF-IDF consists of two scores, term frequency and inverse document frequency. TF counts the number of times a term occurs in a document, whereas IDF divides the total documents by documents where a specific word appears repeatedly. Multiplication of these values gives a high score for words that repeatedly appear in limited documents. Terms appearing frequently consist of a low score.

In TF-IDF function weights, each vector component of each document on the following basis. First, word frequency in the document is incorporated (Soucy and Mineau, 2005) and the more a word appears in a document (e.g., its TF is high) the more it is estimated as significant one. IDF measures how infrequent a word is in the collection and is estimated using the whole training text collection at hand. For frequent word in the text collection, particular representative is not considered. Conversely if the word is infrequent in the text collection, then it is believed to be very relevant for that document.

Radix trees: The radix tree (Siragusa *et al.*, 2013) is a lexicographically ordered data structure representing a set of strings. It is built via radix sort in time and space linear in the total length of the strings. Properties of radix tree is the height (and complexity) that depends on the length of the keys but in general not on the number of elements in the tree (Leis *et al.*, 2013):

- Not required a rebalancing operations and all insertion orders result in the same tree
- The keys can be stored in lexicographic order
- Path to a leaf node represents the key of that leaf and so keys are stored implicitly and can be reconstructed from paths (Fig. 1)

A radix tree has two types of nodes:

- Inner nodes, maps the partial keys to other nodes
- Leaf nodes, stores the values corresponding to the keys

An inner node representation is as an array of 2s pointers. Structure of Radix-tree performs data compression with cluster nodes that share the same branch (Valêncio *et al.*, 2012). Radix trees share some disadvantages: can only be applied to strings of elements or elements with an efficiently reversible mapping (injection) to strings, nonexistence of full generality of balanced search trees. A reversible mapping to strings produces the required total ordering for balanced search trees, but not the other way around. This can be problematic if a data type provides a comparison operation, but not a (de) serialization operation.

Support Vector Machine (SVM): SVM is a classification and regression prediction tool and uses a machine learning theory to maximize predictive accuracy for automatically avoiding over-fit to the data. SVM is defined as systems which use hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. SVM, popularized initially with the NIPS community and now is an active part of the machine learning research all around the world.

SVM became famous when, pixel maps are used as input giving accuracy comparable to sophisticated neural networks with elaborated features in a handwriting recognition task (Jakkula, 2006). It is applied in many applications, such as hand writing analysis, face analysis and so forth, especially for pattern classification and regression based applications. SVM handles simple, linear and more complex classification tasks. Both separable and non-separable problems can be handled by SVMs in the linear and nonlinear case (Luts *et al.*, 2010). The idea is to map the original data points from the input space to a high dimensional, or even infinite-dimensional, feature space such that the classification problem becomes simpler in the feature space. The mapping can be performed by a suitable choice of a kernel function.

Fuzzy Unordered Rule Induction Algorithm (FURIA): FURIA is a fuzzy rule-based classification method. This is a modification and extension of the state-of-the-art rule learner RIPPER. Fuzzy rules can be obtained by replacing intervals by fuzzy intervals with trapezoidal membership function (Rahman and Davis, 2012):

$$I^F(v) = \begin{cases} 1 & \phi^{c,L} \leq v \leq \phi^{c,U} \\ \frac{v - \phi^{s,L}}{\phi^{c,L} - \phi^{s,L}} & \phi^{s,L} < v < \phi^{c,L} \\ \frac{\phi^{s,U} - v}{\phi^{s,U} - \phi^{c,U}} & \phi^{c,U} < v < \phi^{s,U} \\ 0 & \text{else} \end{cases} \quad (2)$$

where, $\phi^{c,L}$ and $\phi^{c,U}$ are the lower and upper bound of the membership of the fuzzy sets. For an instance $x = (x_1, \dots, x_n)$ the degree of the fuzzy membership can be found using the formula in Eq. (3):

$$\mu_r F(x) = \prod_{i=1 \dots k} i_i^F(x_i) \quad (3)$$

FURIA learns fuzzy rules and unordered rule set. This induces rules for each class separately with “one class-other classes” dividing strategy. When the classifier is trained using one class then other classes is taken into account. This helps to achieve a state when it is not main rule and the sequence of classes in the training processes are irrelevant (Gasparovica and Aleksejeva, 2011). However, this approach includes some shortcomings: if a record is equally covered by rules of two classes, then certainty factor has to be calculated. Improvements of RIPPER algorithm affect pruning modifications. However, the strength of this algorithm is the rule stretching method to solve the pressing problem of new records that is classified might be outside the space covered by the previously induced rules.

EXPERIMENTAL RESULTS

The performance of the proposed system was evaluated using FURIA and SVM. The results are shown in Table 1.

Performance measurement on rules using FURIA: The proposed technique improved the precision by 6.83% when compared with entailment based rule acquisition for measurement on Rules using FURIA. It increases by 4.72% in measurement of antecedent and by 5.31% in measurement of consequent evaluation.

The proposed technique improved recall by 6.51% when compared with entailment based rule acquisition for measurement on Rules using FURIA. It increases by 15.08% in measurement of antecedent and by 10.53% in measurement of consequent evaluation (Table 2).

Performance measurement on rules using SVM: The proposed technique improved the precision by 6.43% when compared with entailment based rule acquisition for measurement on Rules using SVM. It increases by

Table 1: Precision

Techniques	Entailment based rule acquisition	Proposed technique
Performance measurement on rules using FURIA	89.88	96.47
Performance measurement on antecedent	92.25	96.83
Performance measurement on consequent	92.48	97.67

Table 2: Recall

Techniques	Entailment based rule acquisition	Proposed technique
Performance measurement on rules using FURIA	88.23	94.38
Performance measurement on antecedent	77.92	91.76
Performance measurement on consequent	79.98	89.40

Table 3: Precision

Techniques	Entailment based rule acquisition	Proposed technique
Performance measurement on rules using SVM	91.39	97.68
Performance measurement on antecedent	93.43	98.73
Performance measurement on consequent	93.64	98.92

Table 4: Recall

Techniques	Entailment based rule acquisition	Proposed technique
Performance measurement on rules using SVM	90.75	96.37
Performance measurement on antecedent	79.46	93.79
Performance measurement on consequent	81.67	91.17

5.36% in measurement of antecedent and by 5.33% in measurement of consequent evaluation (Table 3).

The proposed technique improved the recall by 5.83% when compared with entailment based rule acquisition for measurement on Rules using SVM. It increases by 15.27% in measurement of antecedent and by 10.42% in measurement of consequent evaluation (Table 4).

CONCLUSION

Text classification is a necessity due to the very large amount of text documents that have to be dealt with daily. In this study, TF-IDF is used for text document classification which represents each document as a “TF-IDF” vector in features space either word or phrase that was used for training documents. Also, identifies Hypernyms (superordinate words) clubbing them with entailment rule acquisition. Performance evaluation was performed by SVM and FURIA classifiers. Experimental results show that the proposed method achieves higher precision and recall when compared to entailment based rule acquisition.

REFERENCES

- Bai, P. and J. Li, 2009. The improved naive Bayesian WEB text classification algorithm. Proceeding of International Symposium on Computer Network and Multimedia Technology (CNMT, 2009), pp: 1-4.
- Bo, S., S. Qiurui, C. Zhong and F. Zengmei, 2009. A study on automatic web pages categorization. Proceeding of IEEE International Advance Computing Conference (IACC, 2009), pp: 1423-1427.
- Dagan, I., B. Dolan, B. Magnini, D. Roth, I. Dagan, B. Dolan and P. Pantel, 2010. Recognizing textual entailment: Rational, evaluation and approaches-erratum. Nat. Lang. Eng., 16(1): 105.
- Gasparovica, M. and L. Aleksejeva, 2011. Using fuzzy unordered rule induction algorithm for cancer data classification. Proceeding of 17th International Conference on Soft Computing, MENDEL 2011. Czech Republic, Brno, pp: 141-147.
- Hazman, M., S.R. El-Beltagy and A. Rafea, 2011. Survey of ontology learning approaches. Int. J. Comput. Appl., 22(9).
- Inyaem, U., P. Meesad and C. Haruechaiyasak, 2009. Named-entity techniques for terrorism event extraction and classification. Proceeding of 8th International Symposium on Natural Language Processing (SNLP'09), pp: 175-179.
- Jakkula, V., 2006. Tutorial on support vector machine (svm). School of EECS, Washington State University.
- Jotheeswaran, J. and Y. Kumaraswamy, 2013. Opinion mining using decision tree based feature selection through Manhattan hierarchical cluster measure. J. Theor. Appl. Inform. Technol., 58(1).
- Kan, M.Y. and H.O.N. Thi, 2005. Fast webpage classification using URL features. Proceeding of the 14th ACM International Conference on Information and Knowledge Management, pp: 325-326.
- Khan, A., B. Baharudin and K. Khan, 2010. Semantic based features selection and weighting method for text classification. Proceeding of the 2010 International Symposium in Information Technology (ITSim), 2: 850-855.
- Koirala, C. and K. Rasheed, 2008. Comparison of the effects of morphological and ontological information on text categorization. Proceeding of the 7th International Conference on Machine Learning and Applications (ICMLA'08), pp: 783-786.
- Leis, V., A. Kemper and T. Neumann, 2013. The adaptive radix tree: ARTful indexing for main-memory databases. Proceeding of the IEEE 29th International Conference on Data Engineering (ICDE, 2013), pp: 38-49.
- Liu, J., G. Wang and Z. Jiang, 2009. Research on chinese ontology instance extension based on SVM. Proceeding of the International Symposium on Intelligent Ubiquitous Computing and Education, pp: 564-568.
- Luong, H.P., S. Gauch and Q. Wang, 2009. Ontology learning through focused crawling and information extraction. Proceeding of the International Conference on Knowledge and Systems Engineering (KSE'09), pp: 106-112.
- Luts, J., F. Ojeda, R. Van de Plas, B. De Moor, S. Van Huffel and J.A. Suykens, 2010. A tutorial on support vector machine-based methods for classification problems in chemometrics. Anal. Chim. Acta, 665(2): 129-145.
- Maedche, A. and S. Staab, 2004. Ontology Learning. In: Stab, S., *et al.*, (Eds.), Handbook on Ontologies. Springer-Verlag, Berlin, Heidelberg, pp: 173-190.
- Nasuti, F.W., 2000. Knowledge acquisition using multiple domain experts in the design and development of an expert system for disaster recovery planning. Ph.D. Thesis, Nova Southeastern University.
- Rahman, M.M. and D.N. Davis, 2012. Fuzzy unordered rules induction algorithm used as missing value imputation methods for K-mean clustering on real cardiovascular data. Proceeding of the World Congress on Engineering, Vol. 1.
- Ranwez, V., S. Ranwez and S. Janaqi, 2012. Subontology extraction using hyponym and hypernym closure on is-a directed acyclic graph. IEEE T. Knowl. Data En., 24(12): 2288-2300.
- Rios-Alvarado, A.B., I. Lopez-Arevalo and V. Sosa-Sosa, 2011. Discovering hypernyms using linguistic patterns on web search. Proceeding of the 7th International Conference on Next Generation Web Services Practices (NWeSP), pp: 302-307.
- Siragusa, E., D. Weese and K. Reinert, 2013. Scalable string similarity search/join with approximate seeds and multiple backtracking. Proceeding of the Joint EDBT/ICDT 2013 Workshops, pp: 370-374.
- Soucy, P. and G.W. Mineau, 2005. Beyond TFIDF weighting for text categorization in the vector space model. Proceeding of the 19th International Joint Conference on Artificial Intelligence (IJCAI, 2005), pp: 1130-1135.
- Soumya, S. and H. Swathi, 2013. Automatic repeated rule acquisition from similar web sites using rule ontology. Int. J. Comput. Appl., 66(6): 17-22.
- Valêncio, C.R., F.T. Oyama, P.S. Neto, A.C. Colombini, A.M. Cansian, R.C.G. de Souza and P.L.P. Corrêa, 2012. MR-Radix: A multi-relational data mining algorithm. Human-Centric Comput. Inform. Sci., 2(1): 1-17.
- Vigneshwari, S. and M. Aramudhan, 2012. A novel approach for personalizing the web using user profiling ontologies. Proceeding of 4th International Conference on Advanced Computing (ICoAC, 2012), pp: 1-4.

- Wang, X. and Q. Lu, 2011. Ontology auto-extension based on improved SVM algorithm. Proceeding of the International Conference on E-Business and E-Government (ICEE, 2011), pp: 1-4.
- Xu, Z., F. Yan, J. Qin and H. Zhu, 2011. A web page classification algorithm based on link information. Proceeding of the 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES, 2011), pp: 82-86.
- Yildiz, T. and S. Yildirim, 2012. Association rule based acquisition of hyponym and hypernym relation from a Turkish corpus. Proceeding of the International Symposium on Innovations in Intelligent Systems and Applications (INISTA, 2012), pp: 1-5.
- Yoo, K., 2011. SVM-based knowledge topic identification toward the autonomous knowledge acquisition. Proceeding of the IEEE 9th International Symposium on Applied Machine Intelligence and Informatics (SAMI), pp: 149-154.
- Zhang, W., T. Yoshida and X. Tang, 2008. TFIDF, LSI and multi-word in information retrieval and text categorization. Proceeding of the IEEE International Conference on Systems, Man and Cybernetics (SMC, 2008), pp: 108-113.